

With a Little Help from my Temporal Context: Multimodal Egocentric Action Recognition

Evangelos Kazakos¹
Evangelos.Kazakos@bristol.ac.uk

Jaesung Huh²
jaesung@robots.ox.ac.uk

Arsha Nagrani^{†2}
arsha.nagrani@gmail.com

Andrew Zisserman²
az@robots.ox.ac.uk

Dima Damen¹
Dima.Damen@bristol.ac.uk

¹ Dept. of Computer Science
University of Bristol
Bristol, UK

² Visual Geometry Group
University of Oxford,
Oxford, UK

Abstract

In egocentric videos, actions occur in quick succession. We capitalise on the action’s temporal context and propose a method that learns to attend to surrounding actions in order to improve recognition performance. To incorporate the temporal context, we propose a transformer-based multimodal model that ingests video and audio as input modalities, with an explicit language model providing action sequence context to enhance the predictions. We test our approach on EPIC-KITCHENS and EGTEA datasets reporting state-of-the-art performance. Our ablations showcase the advantage of utilising temporal context as well as incorporating audio input modality and language model to rescore predictions. Code and models at: <https://github.com/ekazakos/MTCN>.

1 Introduction

Action recognition in egocentric video streams from sources like EPIC-KITCHENS poses a number of challenges that differ substantially from those of conventional third-person action recognition – where training and evaluation is on 10 second video clips and classes are quite high-level [30]. Actions are fine-grained (e.g. ‘open bottle’) and noticeably short, often one second or shorter. Along with the challenge, the footage offers an under-explored opportunity, as actions are captured in long untrimmed videos of well-defined and at-times predictable sequences. For example the action ‘wash aubergine’ can be part of the following sequence – you first ‘take the aubergine’, ‘turn on the tap’, ‘wash the aubergine’ and finally ‘turn off the tap’ (Fig. 1). Furthermore, the objects (the aubergine and tap in this case) are persistent over some of the neighbouring actions.



Figure 1: Egocentric video demonstrating two temporal context windows (pink, green) centred around the action to be recognised. We can infer ‘wash aubergine’ with higher accuracy if we know that the tap was turned on before and turned off afterwards.

In this work, we investigate utilising not only the action’s temporal context in the data stream, but also the temporal context from the labels of neighbouring actions. We propose a model that attends to neighbouring actions¹. Concretely, we use the attention mechanism of a multimodal transformer architecture to take account of the context in both the data and labels, using three modalities: vision, audio and language. We are motivated by previous works that demonstrate the significance of audio in recognising egocentric actions [31, 64, 68], and thus include the auditory temporal context in addition to the visual clips. We also utilise context further, by training a language model on the sequence of action labels, inspired by the success of language models [8, 13, 69] in re-scoring model outputs for speech recognition [11, 24, 45] and machine translation [23] (Fig. 2).

Our main contributions are summarised as follows: First, we formulate temporal context as a sequence of actions surrounding the action in a sliding window. Second, we propose a novel framework able to model multimodal temporal context. It consists of a transformer encoder that uses vision and audio as input context, and a language model as output context operating on the action labels. Third, we obtain state-of-the-art performance on two datasets: (i) the large-scale egocentric dataset EPIC-KITCHENS, outperforming high-capacity end-to-end transformer models; and (ii) the EGTEA dataset. Finally, we include an ablation study analysing the importance of the extent of the temporal context and of the various modalities.

2 Related Work

Action Recognition. There is a rich literature of seminal works in action recognition innovating temporal sampling [31, 63, 73], multiple streams [17], spatio-temporal modelling [21, 41, 61] or modelling actions as transformations from initial to final states [65]. Our work is related to the more recent transformer-based approaches [1, 7, 9, 21, 48, 49, 51]. We compare our model with four recent works [1, 9, 48, 51]. [1] investigated spatio-temporal attention factorisation schemes, while in [9] the authors propose full-attention within a temporal window. [51] proposes temporal attention along trajectories with learnt tokens. [48] proposes fusion bottlenecks for cross-modal attention. These approaches require training on large-scale datasets and strong data augmentation for generalisation. Our model operates on pre-extracted features which are processed with a lightweight transformer and outperforms these works by relying on multimodal temporal context.

In egocentric action recognition, researchers propose a range of techniques to address its unique challenges [2, 31, 38, 47, 58]. [58] considered long-term understanding through an LSTM with attention to focus on relevant spatio-temporal regions, but the approach operates within the action clip solely. [2] shows that modelling hand-object interactions is beneficial,

¹Note that these action start/end times are readily available in labelled datasets for untrimmed videos and do not require additional labels. We just leverage these.

while in [38] the authors pre-train egocentric networks by distilling egocentric signals from large-scale third-person datasets. [31] shows that audio is key to egocentric action recognition due to sounds produced from interactions with objects. In [47], the authors propose a multimodal unsupervised domain adaptation approach to tackle the distribution shift between environments. We build on these works but propose the first approach to incorporate both audio-visual and language-model predictions.

Temporal Context for Video Recognition Tasks. A few works have considered temporal context to improve the models’ performance for action anticipation [18, 54], action recognition [10, 50, 67, 71], and object detection [3, 6]. [54] proposes a model that operates on multi-scale past temporal context for action anticipation while they also test on action recognition by modifying the architecture to consider both past and future context. In [71], a set of learnable query vectors attends to dense temporal context to identify events in untrimmed videos. In [50], an encoder-decoder LSTM is proposed for classifying sequences of human actions, effectively attending to the relevant temporal context of each action. Closest to our work is [67], where long-term features from the past and the future of an untrimmed video are utilised, to improve the recognition performance of the ongoing action. A key difference is that we exploit both the temporal bounds and predicted labels of neighbouring actions.

Multimodal transformers. The self attention mechanism of transformers provides a natural bridge to connect multimodal signals. Applications include audio enhancement [16, 62], speech recognition [25], image segmentation [62, 70], cross-modal sequence generation [20, 36, 37], video retrieval [19] and image/video captioning/classification [27, 28, 35, 43, 59, 60]. A common paradigm (which we also adapt) is to use the output representations of single modality convolutional networks as inputs to the transformer [19, 34]. Unlike these works, we use transformers to combine modalities in two specific ways – we first combine audio and visual inputs to predict actions based on neighbouring context, and then *re-score* these predictions with the help of a language model applied on the outputs.

Language models for action detection. There are a few works that incorporate a language model in action detection. The most relevant work is [53], which utilises a statistical n-gram language model along with a length model and an action classifier. [42] combines the Connectionist Temporal Classification (CTC) [22] with a language model to learn relationships between actions. Both of these works improve the results by considering the contextual structure of the sequence of actions, albeit relying on a statistical language model, whereas in this work we utilise a neural language model which has shown better performance [5, 15, 33].

3 Multimodal Temporal Context Network (MTCN)

Given a long video, we predict the action in a video segment by leveraging the *temporal context* around it. We define the temporal context as the sequence of neighbouring actions that precede and succeed the action, and aim to leverage that information, when useful, through learnt attention. We utilise multimodal temporal context both at the input and the output of our model. An audio-visual transformer ingests a temporally-ordered sequence of visual inputs, along with the corresponding sequence of auditory inputs. We use modality-independent positional encodings as well as modality-specific encodings. The language model, acting on the output of the transformer learns the prior temporal context of actions, i.e. the probability of the sequence of actions, using a learnt text embedding space.

Inspired by similar approaches [66], and instead of using a single summary embedding as in prior works for image [14] and action classification [1], the audio-visual model utilises two separate summary embeddings to attend to the action (i.e. verb) class and the object (i.e.

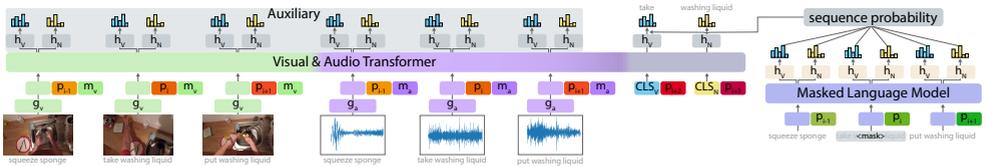


Figure 2: The Multimodal Temporal Context Network (MTCN). Visual and auditory tokens are tagged with positional and modality encodings. An audio-visual transformer encoder attends to the sequence. Verb and noun summary embeddings with independent positional encodings, predict the action at the centre of the window (‘take washing liquid’). The classifier also predicts the sequence of actions to train an auxiliary loss that enhances the prediction of the centre action. The language model filters out improbable sequences.

noun) class. This allows the model to attend independently to the temporal context of verbs vs objects. For example, the object is likely to be the same in neighbouring actions while the possible sequences of verbs can be independent of objects (e.g. ‘take’ \rightarrow ‘put’). Each summary embedding uses a different learnt classification token, and the classifier predicts a verb and a noun from the summary embeddings. The predictions of the audio-visual transformer are then enhanced by filtering out improbable sequences using the language model. We term the proposed model Multimodal Temporal Context Network (MTCN).

In the next three subsections, we detail the architectural components of MTCN, as well as our training strategy. An overview of MTCN is depicted in Fig. 2.

3.1 Audio-Visual Transformer

Let $X_v \in \mathbb{R}^{w \times d_v}$ be the sequence of visual inputs from a video, and $X_a \in \mathbb{R}^{w \times d_a}$ the corresponding audio inputs², for w consecutive actions in the video (i.e. the temporal context window), with d_v, d_a being the input dimensions of the two modalities respectively. X_v and X_a correspond to features extracted from visual and auditory networks, respectively. Our temporal window is centred around an action b_i with surrounding action segments, excluding any background frames. That is, each action b_j within the window, $i - \frac{w-1}{2} \leq j \leq i + \frac{w-1}{2}$ is part of the transformer’s input.

Encoding layer. Our model first projects the inputs X_v, X_a to a lower dimension D and tags each with positional and modality encodings. Then, an audio-visual encoder performs self-attention on the sequence to aggregate relevant audio-visual temporal context from neighbouring actions. Because all self-attention operations in a transformer are permutation invariant, we use positional encodings to retain information about the ordering of actions in the sequence. We use w learnt absolute positional encodings, shared between audio-visual features to model corresponding inputs from the two modalities. Modality encodings, $m_v, m_a \in \mathbb{R}^D$, are learnt vectors added to discriminate between audio and visual tokens.

A classifier predicts the action b_i , using two summary embeddings, acting on the learnt verb/noun tokens. We use the standard approach of appending learnable classification tokens to the end of the sequence but use two tokens, one for verbs and one for nouns, denoted as $\text{CLS}_V, \text{CLS}_N \in \mathbb{R}^D$, with unique positional encodings. To summarise, the encoding layer transforms the inputs X_v and X_a as follows:

$$\begin{aligned} X_{v_j}^e &= g_v(X_{v_j}) + p_j + m_v & X_{a_j}^e &= g_a(X_{a_j}) + p_j + m_a & \forall j \in [1, \dots, w] & \quad (1) \\ \text{CLS}_V^e &= \text{CLS}_V + p_{w+1} & \text{CLS}_N^e &= \text{CLS}_N + p_{w+2} & X^e &= [X_v^e; X_a^e; \text{CLS}_V^e; \text{CLS}_N^e], & \quad (2) \end{aligned}$$

²Video and audio inputs/features are synchronised, therefore they have the same length w .

where $[\cdot]$ denotes input concatenation and $p \in \mathbb{R}^{(w+2) \times D}$ are the positional encodings. $g_v(\cdot) : \mathbb{R}^{d_v} \mapsto \mathbb{R}^D$ and $g_a(\cdot) : \mathbb{R}^{d_a} \mapsto \mathbb{R}^D$, are fully-connected layers projecting the visual and audio features, respectively, to a lower dimension D . The input to the transformer is $X^e \in \mathbb{R}^{(2w+2) \times D}$. In appendix E, we compare the absolute positional encodings with relative [55] and Fourier feature positional encodings [28].

Transformer and classifier. We use a transformer encoder $f(\cdot)$ to process sequential audio-visual inputs, $Z = f(X^e)$. We share the weights of the transformer encoder layer-wise. In appendix E, we compare this to a version without weight sharing. Weight sharing uses $2.7 \times$ less parameters with comparable results. A two-head classifier $h(\cdot)$ for verbs and nouns then predicts the sequence of w actions from both the transformed visual and audio tokens $\hat{Y} = h(Z_{1:2w})$, and the action b_i from the summary embeddings $\hat{y} = h(Z_{2w:2w+2})$.

Loss function. Recall that our goal is to classify b_i , the action localised at the centre of our temporal context. Nevertheless, we can leverage the ground-truth of neighbouring actions within w for additional supervision to train the audio-visual transformer. Our loss is composed of two terms, the main loss for training the model to classify the action at the centre of our temporal context $i = \frac{w}{2}$, and an auxiliary loss to predict all actions in the sequence:

$$L_m = CE(Y_i^V, \hat{y}^V) + CE(Y_i^N, \hat{y}^N) \quad (3)$$

$$L_a = \sum_{j=1}^w (CE(Y_j^V, \hat{Y}_j^V) + CE(Y_j^V, \hat{Y}_{|w|+j}^V) + CE(Y_j^N, \hat{Y}_j^N) + CE(Y_j^N, \hat{Y}_{|w|+j}^N)) \quad (4)$$

$$loss = \beta L_m + (1 - \beta) L_a, \quad (5)$$

where $CE()$ is a cross-entropy loss, and $Y = (Y_1, \dots, Y_w)$ is the ground-truth of the sequence, while $\hat{Y}_1, \dots, \hat{Y}_w$ and $\hat{Y}_{w+1}, \dots, \hat{Y}_{2w}$ correspond to predictions from the transformed visual and auditory inputs respectively. We use β to weight the importance of the auxiliary loss.

3.2 Language Model

In addition to input context from the visual and audio domains, we introduce output context using a language model. Language modelling, commonly applied to predict the probability of a sequence of *words*, is a fundamental task in NLP research [8, 13, 52, 69]. Our language model predicts the probability of a sequence of *actions*. We use the language model to improve the predictions of the audio-visual transformer by filtering out improbable sequences.

We adopt the popular Masked Language Model (MLM), introduced in BERT [13]. We train this model independently from the audio-visual transformer. Specifically, given a sequence of actions $Y = (Y_{i-\frac{w-1}{2}}, \dots, Y_{i+\frac{w-1}{2}})$, we randomly mask any action Y_j and train the model to predict it. For example, an input sequence to the model (for $w = 5$) would be: ('turn on tap', 'wash hands', <MASK>, 'pick up towel', 'dry hands'). Without any visual or audio input, the language model is tasked to learn a high prior probability for 'turn off tap', which is masked. Note that the model is trained using the ground-truth sequence of actions. For input representation, we split the action into verb and noun tokens (e.g. 'dry hands' \rightarrow 'dry' and 'hand'), convert them to one-hot vectors and input them into separate word-embedding layers³. MLM takes as input the concatenation of verb and noun embeddings. The outputs are scores for verb and noun classes using a two-head classifier, and the model is trained with a cross-entropy loss per output.

³Learning a word embedding outperforms pretrained embeddings.

3.3 Inference

Given the the scores of the sequence, $\hat{S} = (\hat{y}_{i-\frac{w-1}{2}}, \dots, \hat{y}_{i+\frac{w-1}{2}})$, from the audio-visual model⁴, we apply a beam-search of size K to find the K most probable action sequences \hat{S}_b . Therefore, \hat{S}_b is of size $K \times w$. In inference, the trained language model takes as input \hat{S}_b , i.e. it operates on sequences predicted from the audio-visual transformer.

For each sequence l , we calculate the probability of the sequence $p_{LM}(\hat{S}_b^l)$ from the language model by utilising the method introduced in [56]. We mask actions, one at a time, and predict their probability. $p_{LM}(\hat{S}_b^l)$ is the sum of log probabilities of all actions in l . We also calculate the probability $p_{AV}(\hat{S}_b^l)$ by summing the log probabilities of all predicted actions in l by the audio-visual model. Then, we combine the probabilities of sequences of the audio-visual and language models:

$$p(\hat{S}_b^l) = \lambda p_{LM}(\hat{S}_b^l) + (1 - \lambda) p_{AV}(\hat{S}_b^l). \quad (6)$$

Sequences are then sorted in descending order by $p(\hat{S}_b^l)$. The score of the centre action from the sequence with the highest probability, is used as the final prediction.

4 Experiments

4.1 Implementation Details

Datasets. EPIC-KITCHENS-100 [12] is the largest egocentric audio-visual dataset, containing unscripted daily activities, thus offering naturally variable sequences of actions. There are on average 129 actions per video (std 163 actions/video and maximum of 940 actions/video). This makes the dataset ideal for exploring temporal context. The length of sequences of $w = 9$ actions (this is our default window length) is 34.4 seconds of video on average with an std of 27.8 seconds⁵. EGTEA [40] is another video-only egocentric dataset. There are 28 hours of untrimmed cooking activities with 10K annotated action segments. Although the dataset does not contain audio, it has sequential actions annotated within long videos, and we use it to train part of our approach (vision and language).

Visual features. For EPIC-KITCHENS, we extract visual features with SlowFast [17], using the public model and code from [12]. We first train the model with slightly different hyperparameters, where we sample a clip of 2s from an action segment, do not freeze batch normalisation layers, and warm-up training during the first epoch starting from a learning rate of 0.001. We note that this gave us better performance. All unspecified hyperparameters remain unchanged. For feature extraction, 10 clips of 1s each are uniformly sampled for each action segment, with a center crop per clip. The resulting features have a dimensionality of $d_v = 2304$. The SlowFast visual features are used for all the results in this paper, apart for the comparison with the state of the art in Table 3 where we additionally experiment with features from Mformer-HR [51]. These are extracted from the EPIC-KITCHENS pretrained model using a single crop per clip. The resulting features have a dimensionality of $d_v=768$. For EGTEA, see appendix F.

Auditory features. We use Auditory SlowFast [32] for audio feature extraction when present. Similarly to the visual features, we extract 10 clips of 1s each uniformly spaced for each action segment, with average pooling and concatenation of the features from the Slow and Fast streams, and the resulting features have the same dimensionality, $d_a = 2304$.

⁴These are temporally ordered predictions from the summary embeddings, and thus different from \hat{Y} .

⁵minimum of 3.4 seconds to a maximum of 720.2 seconds.

| w | Overall | | | | | | Unseen Participants | | | Tail-classes | | |
|-----|--------------------|--------------|--------------|--------------------|--------------|--------------|---------------------|--------------|--------------|--------------------|--------------|--------------|
| | Top-1 Accuracy (%) | | | Top-5 Accuracy (%) | | | Top-1 Accuracy (%) | | | Top-1 Accuracy (%) | | |
| | Verb | Noun | Action | Verb | Noun | Action | Verb | Noun | Action | Verb | Noun | Action |
| 1 | 67.93 | 52.29 | 41.30 | 90.53 | 76.47 | 61.52 | 61.13 | 44.60 | 32.58 | 42.05 | 27.42 | 21.48 |
| 3 | 69.80 | 55.24 | 43.52 | 91.30 | 79.04 | 63.25 | 61.41 | 46.48 | 33.71 | 39.09 | 32.58 | 23.06 |
| 5 | 70.38 | 56.16 | 45.13 | 91.67 | 79.47 | 64.14 | 61.97 | 46.95 | 34.74 | 43.12 | 32.53 | 24.54 |
| 7 | 70.43 | 56.19 | 45.01 | 91.23 | 79.13 | 63.52 | 62.63 | 47.14 | 34.84 | 41.31 | 32.79 | 24.12 |
| 9 | 70.60 | 56.26 | 45.48 | 91.14 | 79.06 | 63.06 | 63.76 | 47.14 | 35.87 | 41.36 | 32.84 | 24.70 |
| 11 | 70.55 | 55.74 | 44.68 | 91.18 | 79.23 | 63.02 | 62.91 | 46.57 | 34.74 | 41.82 | 33.58 | 24.44 |

Table 1: Analysis of temporal context extent for MTCN on EPIC-KITCHENS.

Architectural details. Both the audio-visual transformer encoder and the language model consist of 4 layers with shared weights, 8 attention heads and a hidden unit dimension of 512. In the audio-visual transformer, positional/modality encodings as well as verb/noun tokens have also dimensionality $D = 512$ and are initialised to $\mathcal{N}(0, 0.001)$. The layers $g_v(\cdot)$ and $g_a(\cdot)$ reduce the features to the common dimension $D = 512$. In the encoding layer, dropout is applied at the inputs of $g_v(\cdot)$ and $g_a(\cdot)$ as well as at X^e . In the language model, both verb and noun word-embedding layers have a dimension of 256, and positional encodings have a dimension of 512, while dropout is applied to its inputs.

Scheduled sampling. We modify the scheduled sampling from [4] to train the language model. At each training iteration, we randomly mask an action, predict it, and replace the corresponding ground-truth with the prediction.

Train / Val details. For EPIC-KITCHENS, the audio-visual transformer is trained using SGD, a batch size of 32 and a learning rate of 0.01 for 100 epochs. Learning rate is decayed by a factor of 0.1 at epochs 50 and 75. In the loss function, we set $\beta = 0.9$. For regularisation, a weight decay of 0.0005 is used and a dropout 0.5 and 0.1 for the encoding layers and transformer layers respectively. We use mixup data augmentation [72] with $\alpha = 0.2$. The language model is trained for the same number of epochs with a batch size of 64, Adam optimiser with initial learning rate of 0.001 and the learning rate is decreased by a factor of 0.1 when validation accuracy saturates for over 10 epochs. The values of dropout and weight decay are the same as those of the audio-visual model. For inference, we tune λ in Equation 6 on the validation set with grid-search from the set $\{0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3\}$, and we use a beam size of $K = 10$. For training the audio-visual transformer, we randomly sample 1 out of 10 features per action. For testing, we feed all 10 features per action to the transformer and we share the positional encoding corresponding to an action with all 10 features. We also tried single feature per action followed by averaging 10 predictions but observed no difference in performance. For the train/val details in EGTEA, please see appendix F.

Evaluation metrics. For EPIC-KITCHENS-100, we follow [12] and report top-1 and top-5 accuracy for the validation and test sets separately (Test results and results on EPIC-KITCHENS-55 in appendix A and B). We also follow [12] and report results for two subsets within val/test: unseen participants and tail classes. For EGTEA, we follow [29, 44, 46] and report top-1 accuracy and mean class accuracy using the first train/test split.

4.2 Analysis of temporal context length

We first analyse the importance of the temporal context extent by varying the size of w , i.e. the length of window of actions that our model observes. We use the validation set of EPIC-KITCHENS for this analysis as well as for the ablations in Sec. 4.3, and report EGTEA analysis in appendix G. We perform this analysis both for MTCN containing all modalities

| Overall | | | |
|--------------------|--------------|--------------|--------------|
| Top-1 Accuracy (%) | | | |
| w | Verb | Noun | Action |
| 1 | 19.32 | 3.74 | 0.82 |
| 3 | 38.08 | 45.56 | 23.85 |
| 5 | 42.15 | 50.36 | 29.48 |
| 7 | 42.93 | 50.35 | 29.91 |
| 9 | 43.06 | 50.22 | 29.41 |
| 11 | 41.89 | 49.96 | 29.14 |

Table 2: Analysis of the temporal context on LM.

as shown in Table 1, as well as for the language model as shown in Table 2. Varying the length of the window has a big impact on the model’s accuracy showcasing that MTCN successfully utilises temporal context. Using temporal context outperforms $w = 1$, i.e. no temporal context. As the window length increases the performance also increases. Overall top-1 accuracy increases up to $w = 9$ while top-5 up to a window $w = 5$. Performance on unseen participants and tail classes also increases up to $w = 9$.

In Table 2, experiments are conducted by masking the centre action and measuring how well the model predicts it. We use ground-truth for the other actions, as in this experiment we are interested in the maximum possible performance of the language model, i.e. assuming correct predictions from the audio-visual model. Here $w = 1$ corresponds to a language model that randomly guesses the masked action without any context. When varying the length of the window for the language model, verb performance increases when enlarging the temporal context from $w = 3$ to $w = 9$ while for nouns optimal temporal context is $w = 5$, and $w = 7$ for actions. Interestingly, the language model is performing particularly well for nouns, since the same object is often used for the entire sequence.

Finally, Fig. 3 shows the effect of temporal context extent on verb and noun accuracy of the individual modalities as well as of MTCN. For verbs, visual modality performance increases up to 7 actions and then decreases while for nouns, it steadily increases up to 11 actions, possibly because larger context is able to resolve ambiguities due to occlusion. Audio can better capture temporal context for verbs than the language model, showcasing that the progression of sounds conveys more useful information about the action. For nouns, the language model outperforms audio. Moreover, the language model performs better on nouns because of repetitions of the same object in the sequence. MTCN utilising all modalities (A-V-LM) benefits the most from larger temporal context, particularly for verbs. With this analysis completed, we fix w to 9 in all subsequent experiments.

4.3 Results and Ablations

We compare our approach with the state-of-the-art (SOTA) approaches on EPIC-KITCHENS-100 as shown in Table 3. MTCN significantly outperforms convolutional approaches [17, 31, 41, 63]. We outperform the audio-visual TBN [31] by 8% on top-1 actions, and SlowFast [17] by 6% (using the same visual features). We also outperform very recent transformer-based approaches [1, 9, 48, 51], reporting published results (only top-1 accuracy). Note, that MTCN consists of a lightweight transformer that operates on pre-extracted features, while [1, 9, 48, 51] are high-capacity models trained end-to-end.

Additionally, when we employ visual features from Mformer-HR [51], MTCN improves over all transformer-based approaches, including audio-visual fusion [48], by 3.5% on top-1

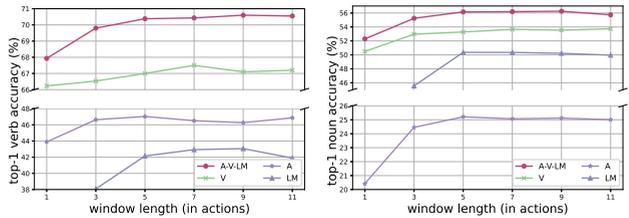


Figure 3: Effect of temporal context on verb (left) and noun (right) accuracy of individual modalities and for MTCN on EPIC-KITCHENS. Y-axis cut to emphasise details.

| Model | Overall | | | | | | Unseen Participants | | | Tail-classes | | |
|-----------------------------|--------------------|-------------|-------------|--------------------|-------------|-------------|---------------------|-------------|-------------|--------------------|-------------|-------------|
| | Top-1 Accuracy (%) | | | Top-5 Accuracy (%) | | | Top-1 Accuracy (%) | | | Top-1 Accuracy (%) | | |
| | Verb | Noun | Action | Verb | Noun | Action | Verb | Noun | Action | Verb | Noun | Action |
| TSN [63] | 60.2 | 46.0 | 33.2 | 89.6 | 72.9 | 55.1 | 47.4 | 38.0 | 23.5 | 30.5 | 19.4 | 13.9 |
| TBN [31] | 66.0 | 47.2 | 36.7 | 90.5 | 73.8 | 57.7 | 59.4 | 38.2 | 29.5 | 39.1 | 24.8 | 19.1 |
| TSM [41] | 67.9 | 49.0 | 38.3 | 91.0 | 75.0 | 60.4 | 58.7 | 39.6 | 29.5 | 36.6 | 23.4 | 17.6 |
| SlowFast [17] | 65.6 | 50.0 | 38.5 | 90.0 | 75.6 | 58.6 | 56.4 | 41.5 | 29.7 | 36.2 | 23.3 | 18.8 |
| ViViT-L/16x2 [1] | 66.4 | 56.8 | 44.0 | - | - | - | - | - | - | - | - | - |
| X-ViT (16x) [9] | 68.7 | 56.4 | 44.3 | - | - | - | - | - | - | - | - | - |
| Mformer-HR [51] | 67.0 | 58.5 | 44.5 | - | - | - | - | - | - | - | - | - |
| MBT [48] | 64.8 | 58.0 | 43.4 | - | - | - | - | - | - | - | - | - |
| MTCN - v.f. SlowFast [17] | 70.6 | 56.3 | 45.5 | 91.1 | 79.1 | 63.1 | 63.8 | 47.1 | 35.9 | 41.4 | 32.8 | 24.7 |
| MTCN - v.f. Mformer-HR [51] | 70.7 | 62.1 | 49.6 | 90.7 | 83.1 | 68.6 | 63.7 | 50.9 | 38.9 | 41.9 | 39.2 | 27.7 |

Table 3: Comparison with SOTA on EPIC-KITCHENS-100 using two visual features (‘v.f.’)

| V | | | | Overall | | | | | | Unseen Participants | | | Tail-classes | | |
|---|---|---|---|--------------------|--------------|--------------|--------------------|--------------|--------------|---------------------|--------------|--------------|--------------------|--------------|--------------|
| | | | | Top-1 Accuracy (%) | | | Top-5 Accuracy (%) | | | Top-1 Accuracy (%) | | | Top-1 Accuracy (%) | | |
| | | | | Verb | Noun | Action | Verb | Noun | Action | Verb | Noun | Action | Verb | Noun | Action |
| ✓ | ✗ | ✗ | ✓ | 67.10 | 53.54 | 41.49 | 90.62 | 78.22 | 62.32 | 58.87 | 43.29 | 30.99 | 40.97 | 30.47 | 22.35 |
| ✓ | ✗ | ✓ | ✓ | 67.84 | 54.08 | 42.05 | 90.63 | 78.20 | 60.68 | 59.06 | 44.23 | 31.36 | 39.77 | 31.32 | 22.38 |
| ✓ | ✓ | ✗ | ✓ | 70.23 | 55.82 | 45.00 | 91.13 | 79.06 | 64.58 | 63.29 | 46.38 | 35.02 | 41.76 | 32.26 | 24.41 |
| ✓ | ✓ | ✓ | ✗ | 69.31 | 55.46 | 43.81 | 91.19 | 79.76 | 62.35 | 61.13 | 46.01 | 33.90 | 39.55 | 30.74 | 22.74 |
| ✓ | ✓ | ✓ | ✓ | 70.60 | 56.26 | 45.48 | 91.14 | 79.06 | 63.06 | 63.76 | 47.14 | 35.87 | 41.36 | 32.84 | 24.70 |
| ✓ | ✓ | † | ✓ | 71.33 | 63.56 | 50.32 | 92.05 | 83.16 | 67.85 | 62.25 | 52.68 | 37.56 | 41.53 | 44.16 | 29.89 |

Table 4: Ablation on multimodal temporal context and auxiliary loss in EPIC-KITCHENS.

†: upper bound using ground-truth knowledge as input to the language model.

nouns and 5% on actions. We attribute the boost on nouns to the enhanced object recognition performance of the ViT backbone [14] in Mformer-HR and its large-scale pretraining. These results, however, further demonstrate the potential to boost other methods and features by utilising multimodal temporal context.

We compare MTCN without audio with the state-of-the-art on EGTEA, in Table 5. This model uses $w = 3$. Our model improves over previous approaches by 3% in both top-1 and mean class accuracy. Note that MTCN outperforms SlowFast [17] which we used to extract features. In appendix G, we ablate the temporal context and language model, where we showcase that the language model provides a bigger boost in performance, possibly due to the absence of the audio modality (also shown in Table 5: ‘V’ vs ‘V+LM’).

Multimodality Ablation. We offer an ablation to identify the performance impact of the components of our MTCN. Results are shown in Table 4. We remove audio and language modalities from our model’s input and output respectively to assess their importance. Multimodal context is important, as our proposed model enjoys considerable margins compared with the model trained only with visual context (line 1 in Table 4). Audio is beneficial, confirming the finding of prior works.

Although the language model provides smaller boost in performance than audio, it showcases that it is useful to model prior temporal context at the output level, and that its benefits are complementary to audio. We also include upper bound performance improvement

| Method | Top-1 (%) | MC(%) |
|---------------------|--------------|--------------|
| Li et al. [39] | - | 53.30 |
| Ego-RNN [57] | 62.17 | - |
| Kapidis et al. [29] | 68.99 | 61.40 |
| Lu et al. [44] | 68.60 | 60.54 |
| SlowFast [17] | 70.43 | 61.92 |
| MCN [26] | 55.63 | - |
| Min et al. [46] | 69.58 | 62.84 |
| MTCN (V) (Ours) | 72.55 | 64.86 |
| MTCN (V+LM) (Ours) | 73.59 | 65.87 |

Table 5: Comparative results on EGTEA. MC: Mean Class

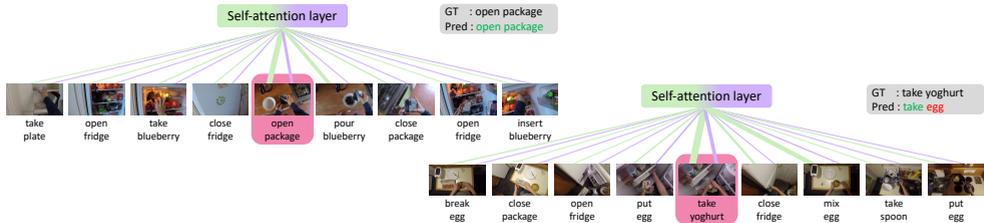


Figure 4: Qualitative results of attention weights along with the predictions of our model in EPIC-KITCHENS. Green and purple edges represent attention from the noun embedding to visual and auditory tokens, respectively. Thickness indicates attention weights magnitude to centre (bordered) and temporal context actions.

from the addition of the language model, where it takes as input the ground-truth preceding/succeeding actions rather than the predictions from the audio-visual transformer (last line in Table 4), effectively a language infilling problem. These results demonstrate the potential margin for improvement, particularly for nouns and accordingly action accuracy as well as tail classes. In appendix C, we report statistical significance over multiple runs for the language model, and compare it to alternative baselines.

Auxiliary loss. Also from Table 4, training MTCN with auxiliary loss boosts its performance almost in all metrics, confirming that utilising supervision from neighbouring actions can improve the performance of the action of interest, i.e. the action at the centre of the window.

While we focus on offline action recognition, we can evaluate our model for online recognition by predicting the last action in the sequence instead of the centre one. Results are included in appendix D for different values of w , where $w = 7$ provided best results for EPIC-KITCHENS. Best top-1 action performance using past context solely is 42.96% compared to 45.48% using surrounding context. However, we demonstrate that our model can leverage multimodal temporal context in this setting.

In Fig. 4, we visualise the auditory and visual temporal context attention on a correctly recognised sequence (left) – where subsequent actions of pouring and closing are particularly informative, and an incorrectly (right) predicted sequence – where both attention weights are high on actions containing the egg, causing the model to incorrectly predict the noun as egg. More qualitative examples are included in appendix H.

5 Conclusion

We formulate temporal context as a sequence of actions, and utilise past and future context to enhance the prediction of the centre action in the sequence. We propose MTCN, a model that attends to vision and audio as input modality context, and language as output modality context. We train MTCN with additional supervision from neighbouring actions. Our results showcase the importance of audio and language as additional modality context. We report SOTA results on two egocentric video datasets: EPIC-KITCHENS and EGTEA. An extension to MTCN would incorporate an actionness score of neighbouring frames, to distinguish background frames and learn from action sequences in untrimmed videos without temporal bounds during testing. This would bridge the problems of recognition and detection, utilising multimodality and temporal context. We will be exploring this in future work.

Acknowledgements. This work used public datasets and is supported by EPSRC Doctoral Training Program, EPSRC UMPIRE (EP/T004991/1) and the EPSRC Programme Grant VisualAI (EP/T028572/1). Jaesung Huh is funded by a Global Korea Scholarship.

References

- [1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2021.
- [2] Fabien Baradel, Natalia Neverova, Christian Wolf, Julien Mille, and Greg Mori. Object level visual reasoning in videos. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2018.
- [3] Sara Beery, Guanhang Wu, Vivek Rathod, Ronny Votel, and Jonathan Huang. Context r-cnn: Long term temporal context for per-camera object detection. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [4] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.
- [5] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *Journal of Machine Learning Research (JMLR)*, 3:1137–1155, 2003.
- [6] Gedas Bertasius and Lorenzo Torresani. Classifying, segmenting, and tracking object instances in video with mask propagation. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of International Conference on Machine Learning (ICML)*, 2021.
- [8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.
- [9] Adrian Bulat, Juan-Manuel Perez-Rua, Swathikiran Sudhakaran, Brais Martínez, and Georgios Tzimiropoulos. Space-time mixing attention for video transformer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [10] Alejandro Cartas, Petia Radeva, and Mariella Dimiccoli. Modeling long-term interactions to enhance action recognition. In *Proceedings of International Conference on Pattern Recognition (ICPR)*, 2021.
- [11] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

- [12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 2021.
- [13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.
- [14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021.
- [15] Yerai Doval and Carlos Gómez-Rodríguez. Comparing neural-and n-gram-based language models for word segmentation. *Journal of the Association for Information Science and Technology*, 70:187–197, 2019.
- [16] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics (TOG)*, 37:1–11, 2018.
- [17] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast Networks for Video Recognition. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2019.
- [18] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with Rolling-Unrolling LSTMs and modality attention. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2019.
- [19] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.
- [20] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.
- [21] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. Video action transformer network. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [22] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of International Conference on Machine Learning (ICML)*, 2006.

- [23] Caglar Gulcehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, and Yoshua Bengio. On integrating a language model into neural machine translation. *Computer Speech & Language*, 45:137–148, 2017.
- [24] Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. Deep speech: Scaling up end-to-end speech recognition. *CoRR*, abs/1412.5567, 2014.
- [25] David Harwath, Antonio Torralba, and James Glass. Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.
- [26] Yifei Huang, Minjie Cai, Zhenqiang Li, Feng Lu, and Yoichi Sato. Mutual context network for jointly estimating egocentric gaze and action. *IEEE Transactions on Image Processing (TIP)*, 29:7795–7806, 2020.
- [27] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *Proceedings of Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020.
- [28] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver: General perception with iterative attention. *CoRR*, abs/2103.03206, 2021.
- [29] Georgios Kapedis, Ronald Poppe, Elsbeth van Dam, Lucas Noldus, and Remco Veltkamp. Multitask learning to improve egocentric action recognition. In *Proceedings of International Conference on Computer Vision Workshops (ICCVW)*, 2019.
- [30] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017.
- [31] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2019.
- [32] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-fast auditory streams for audio recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.
- [33] Yoon Kim, Yacine Jernite, David Sontag, and Alexander Rush. Character-aware neural language models. In *Proceedings of AAAI Conference on Artificial Intelligence*, 2016.
- [34] Sangho Lee, Youngjae Yu, Gunhee Kim, Thomas Breuel, Jan Kautz, and Yale Song. Parameter Efficient Multimodal Transformers for Video Representation Learning. In *International Conference on Learning Representations (ICLR)*, 2021.
- [35] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. Entangled transformer for image captioning. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2019.

- [36] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. Learning to generate diverse dance motions with transformer. *CoRR*, abs/2008.08171, 2020.
- [37] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2021.
- [38] Yanghao Li, Tushar Nagarajan, Bo Xiong, and Kristen Grauman. Ego-exo: Transferring visual representations from third-person to first-person videos. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [39] Yin Li, Miao Liu, and James M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2018.
- [40] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2018.
- [41] Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2019.
- [42] Mengxi Lin, Nakamasa Inoue, and Koichi Shinoda. Ctc network with statistical language modeling for action sequence recognition in videos. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, 2017.
- [43] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [44] Minlong Lu, Danping Liao, and Ze-Nian Li. Learning spatiotemporal attention for egocentric action recognition. In *Proceedings of International Conference on Computer Vision Workshops (ICCVW)*, 2019.
- [45] Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2010.
- [46] Kyle Min and Jason J. Corso. Integrating human gaze into attention for egocentric activity recognition. In *Proceedings of Winter Conference on Applications of Computer Vision (WACV)*, 2021.
- [47] Jonathan Munro and Dima Damen. Multi-modal domain adaptation for fine-grained action recognition. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [48] Arsha Nagrani, Shan Yang, Anurag Arnab, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

- [49] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. *CoRR*, abs/2102.00719, 2021.
- [50] Yan Bin Ng and Basura Fernando. Human action sequence classification. *CoRR*, abs/1910.02602, 2019.
- [51] Mandela Patrick, Dylan Campbell, Yuki M. Asano, Ishan Misra Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F. Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
- [52] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.
- [53] Alexander Richard and Juergen Gall. Temporal action detection using a statistical language model. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [54] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.
- [55] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.
- [56] Joonbo Shin, Yoonhyung Lee, and Kyomin Jung. Effective sentence scoring method using bert for speech recognition. In *Proceedings of Asian Conference on Machine Learning (ACML)*, 2019.
- [57] Swathikiran Sudhakaran and Oswald Lanz. Attention is all we need: Nailing down object-centric attention for egocentric activity recognition. In *Proceedings of British Machine Vision Conference (BMVC)*, 2018.
- [58] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Lsta: Long short-term attention for egocentric action recognition. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [59] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *CoRR*, abs/1906.05743, 2019.
- [60] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2019.
- [61] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

- [62] Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Daniel PW Ellis, and John R Hershey. Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2021.
- [63] 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 *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.
- [64] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.
- [65] Xiaolong Wang, Ali Farhadi, and Abhinav Gupta. Actions ~ transformations. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [66] Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2019.
- [67] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [68] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *CoRR*, abs/2001.08740, 2020.
- [69] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.
- [70] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. Cross-modal self-attention network for referring image segmentation. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [71] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Temporal query networks for fine-grained video understanding. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.
- [72] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proceedings of International Conference on Learning Representations ICLR*, 2018.
- [73] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2018.