Multi-Resolution Audio-Visual Feature Fusion for Temporal Action Localization

Edward Fish
University of Surrey
edward.fish@surrey.ac.uk

Jon Weinbren
University of Surrey
j.weinbren@surrey.ac.uk

Andrew Gilbert
University of Surrey
a.gilbert@surrey.ac.uk

Abstract

Temporal Action Localization (TAL) aims to identify actions' start, end, and class labels in untrimmed videos. While recent advancements using transformer networks and Feature Pyramid Networks (FPN) have enhanced visual feature recognition in TAL tasks, less progress has been made in the integration of audio features into such frameworks. This paper introduces the Multi-Resolution Audio-Visual Feature Fusion (MRAV-FF), an innovative method to merge audio-visual data across different temporal resolutions. Central to our approach is a hierarchical gated cross-attention mechanism, which discerningly weighs the importance of audio information at diverse temporal scales. Such a technique not only refines the precision of regression boundaries but also bolsters classification confidence. Importantly, MRAV-FF is versatile, making it compatible with existing FPN TAL architectures and offering a significant enhancement in performance when audio data is available.

1 Introduction

Temporal Action Localization (TAL) is concerned with detecting the onset and offset of actions and their class labels in untrimmed and unconstrained videos. Recently, the combined use of transformer networks and Feature Pyramid Networks (FPN) [58, 63, 9, 55, 44] has led to a significant boost in the performance and efficiency of TAL tasks by leveraging multi-resolution visual features. However, there has not yet been a study on combining audio information in such network architectures for this task, specifically how to fuse audio information over different temporal resolutions. The challenge lies in integrating audio and visual data and determining the density of audio information required across different FPN channels for different actions. While some channels might require richer audio input to accurately identify action segments due to higher visual downsampling, others with more detailed visual cues might need less audio assistance. For instance, as shown in Fig 1, an action such as 'chopping' can be better located using high-resolution (i.e. less downsampled) audio features. In contrast, an activity such as 'washing up' may only require some low-resolution audio information. A final example could be for an action such as 'pick-up', which requires no audio input. With this in mind, a fusion method for audio TAL should accommodate multiple temporal audio resolutions while also including a mechanism to gate audio information in specific temporal pathways.

This paper presents a novel framework for Multi-Resolution Audio-Visual Feature Fusion (MRAV-FF) as a first step to solving these issues. Our methodology is rooted in a hierarchical gated cross-attention fusion mechanism that adaptively combines audio and visual features over varying temporal scales. Unlike existing techniques, MRAV-FF weighs the significance of each modality's features at

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

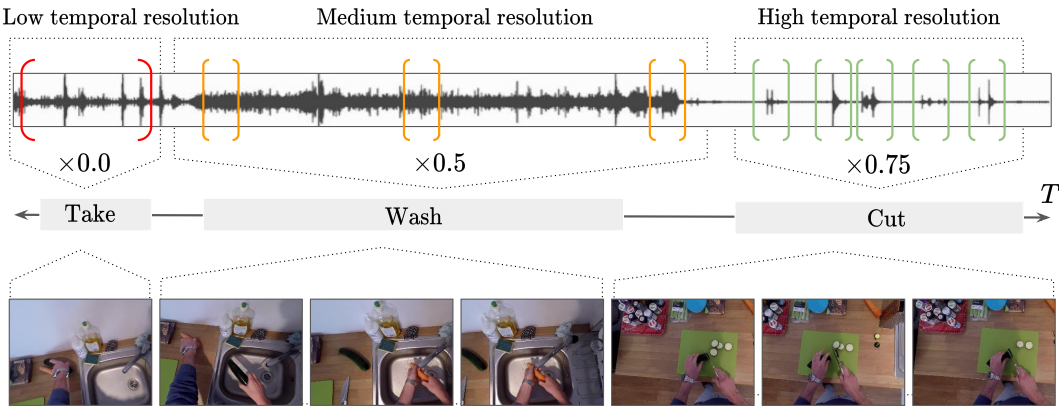

Figure 1: We use a Feature Pyramid Network (FPN) to encode audio-visual action features along different temporal resolutions. We then gate the fusion of the audio features depending on their application to the action classification and regression boundaries. For example, the action 'take' requires no audio, which is gated out. In contrast, the action 'chop' can be better localised by combining high-temporal resolution audio features with visual features. Our method learns both the temporal resolution and the gating values end-to-end.

various temporal scales to improve the regression boundaries and classification confidence. Furthermore, our method can be easily plugged into any FPN TAL architecture to boost performance when audio information is available.

2 Related Work

Temporal Action Localization (TAL). Methods in TAL can be separated as single and two-stage. Where single stage methods generate a large number of proposal segments which are then passed to a classification head[14, 5, 19, 28, 17, 66, 28, 27, 25, 8, 14, 32]. Single-stage methods include the use of graph neural networks[4, 57, 62, 57] and more recently, transformers [53, 7, 46]. Recent progress in single-stage TAL has shown improvements over two-stage methods in accuracy and efficiency, combining both action proposal and classification in a single forward pass. Works inspired by object detection [42, 31], saliency detection [26], and hierarchical CNN's [60, 26, 61] all combine proposal and classification. Current SOTA methods in TAL utilise transformer-based [51] feature pyramid networks (FPN's) [63, 9, 55, 44], which combine multi-resolution transformer features with classification and regression heads.

Audio-Visual Fusion. Audio-visual fusion via learned representations has been explored in several video retrieval and classification tasks [13, 1, 56, 54, 37, 24, 23, 24]. Audio-visual TAL has been less explored, with most approaches focused on audio-visual events in which the audio and visual events are closely aligned [48, 3]. Concurrent works exploring audio-visual fusion in TAL have adopted two-stage late fusion approaches. Recent works have also explored audio-visual cross-attention [41] but over a single temporal resolution and without any gated fusion control.

3 Method

Problem Definition Consider an untrimmed input video denoted as $\mathcal{X}$. The goal is to represent $\mathcal{X}$ as a set of feature vectors symbolized as $\mathcal{X} = \{x_1, x_2, \ldots, x_T\}$. Each x_t corresponds to discrete time steps, $t = \{1, 2, \ldots, T\}$. Notably, the total duration T is not constant and may differ across videos. For illustrative purposes, x_t can be envisaged as a feature vector extracted from a 3D convolutional network at a specific time t within the video. The primary objective of TAL is to identify and label action instances present in the input video sequence $\mathcal{X}$. These instances are collectively denoted as $\mathcal{Y} = \{y_1, y_2, \ldots, y_N\}$, where N signifies the total number of action instances in a given video. This value can be variable across different videos. Each action instance, y_i, is defined by the tuple

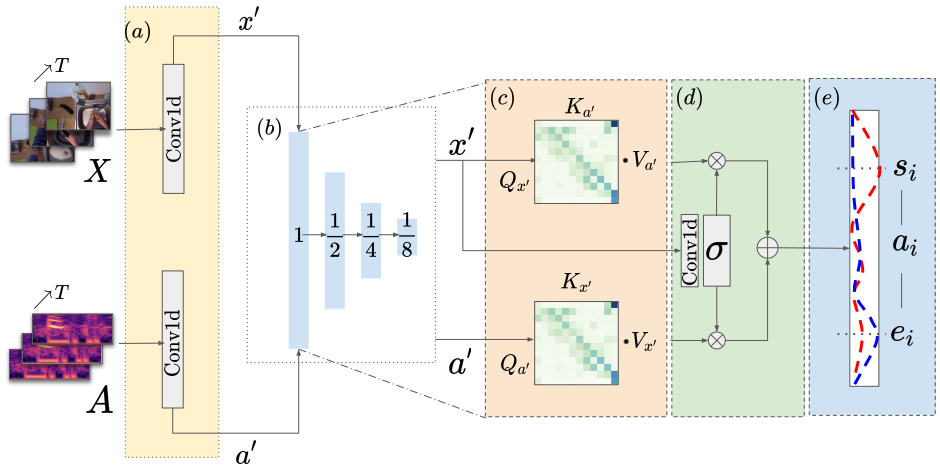

Figure 2: A high-level representation of our multi-resolution audio-fusion method. (a) Audio and visual features are projected to a shared dimension via a 1D convolution. (b) Max-Pooling is applied to downsample features. (c) Following downsampling, we apply multi-headed cross attention in each temporal layer between audio and visual features. (d) The video features are then used as context to scale audio and visual attended embeddings. (e) The concatenated embedding is then used for both regression and classification.

$y_i = (s_i, e_i, a_i)$, where s_i represents the starting time or onset of the action instance, e_i denotes the ending time or offset of the action instance, and a_i specifies the action category or label.

The parameters must adhere to the conditions: $s_i, e_i \in \{1, \ldots, T\}$, $a_i \in \{1, \ldots, C\}$ (with C indicating the total number of predefined action categories), and $s_i < e_i$, which ensures the starting time precedes the ending time for every action instance. Furthermore, alongside the visual feature set $\mathcal{X}$, we introduce an audio feature set $\mathcal{A}$. This set can be represented as $\mathcal{A} = \{a_1, a_2, \ldots, a_{T_{\text{audio}}}\}$, spanning up to T_{audio} time steps. Notably, the total duration T_{audio} may or may not align with T from the visual features, depending on the extraction mechanism and granularity of the audio features.

A significant challenge in TAL with multi-modal inputs is to devise an optimum method for fusing visual and audio features. This fusion aims to leverage complementary information from both modalities, enhancing the robustness and accuracy of action localization and classification.

Method Overview As depicted in Fig 2, our proposed method is structured around three core components. First, video and audio features are extracted from untrimmed videos using frozen, pre-trained encoders. These encoders provide a robust foundation for capturing the inherent characteristics of the media without additional training overhead. Post-extraction, these features are further refined via a shallow convolution layer. Subsequently, they are channelled into a feature pyramid network. This network's features experience iterative downsampling and are intricately fused through our novel cross attention mechanism. This mechanism ensures effective alignment and integration of features from diverse modalities and resolutions, facilitating the capture of complex temporal relationships. Finally, upon feature fusion, each temporal feature vector is processed by two dedicated decoders: one for regression, predicting action onsets and offsets, and the other for classification, identifying specific action class labels. This dual-decoder approach ensures accurate temporal localization and semantic identification of each detected action.

Audio-Visual Temporal Fusion: Given projected audio embeddings $\mathcal{A} = \{a_1, a_2, \ldots, a_{T_{\text{audio}}}\}$ and visual embeddings $\mathcal{X} = \{x_1, x_2, \ldots, x_T\}$ for each timestep, we can break down the process as follows:

Downsampling: For any feature set F, the downsampled feature F' is computed as:

$$F' = \text{MaxPool}(F, \text{stride} = 2) \tag{1}$$

Multi-Headed Cross Attention: The attention mechanism can be denoted for any feature f as:

$$\text{Attention}(f) = \text{Softmax}(fQf^TK)V \tag{2}$$

where Q, K, and V are the learned query, key, and value matrices, respectively. Given the downsampled video feature x' and audio feature a', the cross-modal projection for the video as query and audio as query is defined as:

$$P_x = x'Q_x(a'K_a)^TV_a \quad \text{and} \quad P_a = a'Q_a(x'K_x)^TV_x \tag{3}$$

where Q_x, K_x, V_x and Q_a, K_a, V_a are the respective learned matrices for the video and audio modalities.

Gated Audio-Visual Fusion: To further refine our fusion process, we introduce a gating mechanism which adaptively scales the contribution of audio and visual features based on the context of the visual content. For each downsampled visual feature x', we compute a gating scalar g using a sigmoid function:

$$g = \sigma(\text{FC}(x')) \tag{4}$$

where σ denotes the sigmoid activation function, ensuring g is in the range $[0, 1]$, and FC is a fully connected layer. Using the gating scalar, the cross-modal projections are adjusted as follows:

$$P_{x,\text{gated}} = g \cdot P_x \qquad\qquad P_{a,\text{gated}} = (1 - g) \cdot P_a \tag{5}$$

The combined feature representation after the gated cross-modal projection is then:

$$F_{\text{gated\_combined}} = \text{Conv1D}([P_{x,\text{gated}}; P_{a,\text{gated}}]) \tag{6}$$

Regression and Classification: Each temporal layer outputs gated features to the classification head and the regression head for action instance detection. The output of each instant t in feature pyramid layer l is denoted as $\hat{o}_t^l = (\hat{c}_t^l, \hat{d}_{st}^l, \hat{d}_{et}^l)$.

We use the same loss as described in [49, 64, 63]:

$$\mathcal{L} = \frac{1}{N_{pos}} \sum_{l,t} \mathbb{1}_{\{c_t^l > 0\}} (\sigma_{IoU} \mathcal{L}_{cls} + \mathcal{L}_{reg}) + \frac{1}{N_{neg}} \sum_{l,t} \mathbb{1}_{\{c_t^l = 0\}} \mathcal{L}_{cls} \tag{7}$$

Where σ_{IoU} is the temporal IoU between the predicted segment and the ground truth action instance, and $\mathcal{L}_{cls}$, $\mathcal{L}_{reg}$ is focal loss [29] and IoU loss [43]. N_{pos} and N_{neg} denote the number of positive and negative samples. The term σ_{IoU} is used to re-weight the classification loss at each instant, such that instants with better regression (i.e. of higher quality) contribute more to the training.

4 Evaluation

4.1 Dataset

EPIC-Kitchens 100 [10] is an egocentric dataset containing two tasks: noun localization (e.g. door) and verb localization (e.g. open the door). It has 495 and 138 videos, with 67,217 and 9,668 action instances for training and inference, respectively. The number of action classes for noun and verb are 300 and 97. We follow all other methods [27, 63, 9, 62, 47], and report the mean average precision (mAP) at different intersection over union (IoU) thresholds with the average mAP computed over [0.1:0.5:0.1] in Table 1.

We show the effectiveness of our audio-fusion method in increasing performance of unimodal models by adding our MRAV-FF to the best performing existing FPN networks. We show how our method improves the performance of both ActionFormer and TemporalMaxer by +0.9 mAP and +0.4 mAP for verbs and +0.9 and +0.7 for nouns.

Task	Method	tIoU					
		0.1	0.2	0.3	0.4	0.5	Avg
Verb	BMN [27, 11]	10.8	9.8	8.4	7.1	5.6	8.4
	G-TAD [57]	12.1	11.0	9.4	8.1	6.5	9.4
	ActionFormer [63]	26.6	25.4	24.2	22.3	19.1	23.5
	TemporalMaxer [47]	27.8	26.6	25.3	23.1	19.9	24.5
	ActionFormer + MRAV-FF	27.6	26.8	25.3	23.4	19.8	24.6
	TemporalMaxer + MRAV-FF	**28.5**	**27.4**	**26.0**	**23.7**	**20.12**	**25.1**
Noun	BMN [27, 11]	10.3	8.3	6.2	4.5	3.4	6.5
	G-TAD [57]	11.0	10.0	8.6	7.0	5.4	8.4
	ActionFormer [63]	25.2	24.1	22.7	20.5	17.0	21.9
	TemporalMaxer [47]	26.3	25.2	23.5	21.3	17.6	22.8
	ActionFormer + MRAV-FF	26.4	25.4	23.6	21.2	17.4	22.8
	TemporalMaxer + MRAV-FF	**27.4**	**26.2**	**24.4**	**21.8**	**17.9**	**23.5**

Table 1: The performance of our proposed method on the EPIC-Kitchens 100 dataset. [11]

Task	Method	tIoU					
		0.1	0.2	0.3	0.4	0.5	Avg
Verb	Concatenation	28.02	26.96	25.5	23.48	19.87	23.89
	Channel Pooling	25.63	24.59	23.09	21.14	17.95	23.06
	MRAV-FF	**28.5**	**27.4**	**26.0**	**23.7**	**20.12**	**25.1**
Noun	Concatenation	26.39	25.42	23.57	21.19	17.42	22.8
	Channel Pooling	25.7	24.53	22.95	20.52	17.04	22.21
	MRAV-FF	**27.4**	**26.2**	**24.4**	**21.8**	**17.9**	**23.5**

Table 2: Results for an ablation experiment on EPIC-Kitchens 100 [11] TAL task, where we replace the MRAV-FF module with existing approaches to feature fusion including concatenated projection and channel pooling. We observe that simple fusion methods hinder performance when compared with uni-modal FPN networks demonstrating the need for a more nuanced fusion strategy.

Task	Method	tIoU					
		0.1	0.2	0.3	0.4	0.5	Avg
Verb	Damen [12]	10.83	9.84	8.43	7.11	5.58	8.36
	AGT [38]	12.01	10.25	8.15	7.12	6.14	8.73
	OWL [41]	14.48	13.05	11.82	10.25	8.73	11.67
	MRAV-FF	**28.5**	**27.4**	**26.0**	**23.7**	**20.12**	**25.1**
Noun	Damen [12]	10.31	8.33	6.17	4.47	3.35	6.53
	AGT [38]	11.63	9.33	7.05	6.57	3.89	7.70
	OWL [41]	17.94	15.81	14.14	12.13	9.80	13.96
	MRAV-FF	**27.4**	**26.2**	**24.4**	**21.8**	**17.9**	**23.5**

Table 3: The performance of our proposed method on the EPIC-Kitchens 100 dataset [11] compared to existing approaches for audio-visual feature fusion on TAL. Our method demonstrates a large increase in performance jointly attributed to the addition of feature pyramid architecture and our fusion strategy.

4.2 Ablation Results

We perform initial ablation experiments to evaluate the performance of our proposed method and present the results in Tab 2. Each experiment is conducted on EPIC-Kitchens, where we edit the temporal fusion method in each temporal block. We first exchange our MRAV-FF temporal block for simple feature fusion in which we concatenate and project the audio-visual features at each temporal scale via a 1D-CNN. We notice that this actually harms network performance over unimodal features demonstrating the need for a gated approach to fusion. Similarly we also replace the block with a max-pooling layer inspired by [47] which pools channel-wise for feature fusion. Again this method has a negative impact on network performance.

Type	Model	Feature	tIoU↑						time(ms) ↓
			0.3	0.4	0.5	0.6	0.7	Avg.	
Two-Stage	BMN [27]	TSN [52]	56.0	47.4	38.8	29.7	20.5	38.5	483*
	DBG [25]	TSN [52]	57.8	49.4	39.8	30.2	21.7	39.8	—
	G-TAD [57]	TSN [52]	54.5	47.6	40.3	30.8	23.4	39.3	4440*
	BC-GNN [4]	TSN [52]	57.1	49.1	40.4	31.2	23.1	40.2	—
	TAL-MR [66]	I3D [6]	53.9	50.7	45.4	38.0	28.5	43.3	>644*
	P-GCN [62]	I3D [6]	63.6	57.8	49.1	—	—	—	7298*
	P-GCN [62] +TSP [2]	R(2+1)1 D [50]	69.1	63.3	53.5	40.4	26.0	50.5	—
	TSA-Net [17]	P3D [40]	61.2	55.9	46.9	36.1	25.2	45.1	—
	MUSES [32]	I3D [6]	68.9	64.0	56.9	46.3	31.0	53.4	2101*
	TCANet [39]	TSN [52]	60.6	53.2	44.6	36.8	26.7	44.3	—
	BMN-CSA [45]	TSN [52]	64.4	58.0	49.2	38.2	27.8	47.7	—
	ContextLoc [67]	I3D [6]	68.3	63.8	54.3	41.8	26.2	50.9	—
	VSGN [65]	TSN [52]	66.7	60.4	52.4	41.0	30.4	50.2	—
	RTD-Net [46]	I3D [6]	68.3	62.3	51.9	38.8	23.7	49.0	>211*
	Disentangle [68]	I3D [6]	72.1	65.9	57.0	44.2	28.5	53.5	—
	SAC [59]	I3D [6]	69.3	64.8	57.6	47.0	31.5	54.0	—
Single-Stage	A²Net [60]	I3D [6]	58.6	54.1	45.5	32.5	17.2	41.6	1554*
	GTAN [34]	P3D [40]	57.8	47.2	38.8	—	—	—	—
	PBRNet [30]	I3D [6]	58.5	54.6	51.3	41.8	29.5	—	—
	AFSD [26]	I3D [6]	67.3	62.4	55.5	43.7	31.1	52.0	3245*
	TAGS [36]	I3D [6]	68.6	63.8	57.0	46.3	31.8	52.8	—
	HTNet [22]	I3D [6]	71.2	67.2	61.5	51.0	39.3	58.0	—
	TadTR [33]	I3D [6]	74.8	69.1	60.1	46.6	32.8	56.7	195*
	GLFormer [18]	I3D [6]	75.9	72.6	67.2	57.2	41.8	62.9	—
	AMNet [33]	I3D [6]	76.7	73.1	66.8	57.2	42.7	63.3	—
	ActionFormer [63]	I3D [6]	82.1	77.8	71.0	59.4	43.9	66.8	80
	ActionFormer [63] + GAP [35]	I3D [6]	82.3	—	71.4	—	44.2	66.9	>80
	TemporalMaxer	I3D [6]	**82.8**	**78.9**	**71.8**	**60.5**	44.7	**67.7**	**50**
	TemporalMaxer + MRAVFF	I3D [6] + Audio [20]	82.2	78.2	71.5	59.9	**45.3**	67.4	60

Table 4: Performance of our method on the THUMOS dataset for TAL. We observe that audio-visual fusion on edited videos is much more challenging than the raw-video setting due to the addition of background music, narration, and audio-visual misalignment.

4.3 Further Results

Furthermore in Tab 3 we evaluate our method with other approaches to audio-visual fusion for TAL on EPIC-Kitchens. We show a large increase in performance, which can be attributed to both the effectiveness of the FPN structure for audio visual temporal pooling and also our MRAV-FF fusion module. The lack of available comparative methods for audio-visual fusion further illustrates the importance of updated baselines in this field.

Finally, we also evaluate the method on the THUMOS14 dataset which [21] contains 200 validation videos and 213 testing videos with 20 action classes. THUMOS14 presents a different challenge to ego-centric audio-visual fusion, since the videos are heavily edited and contain many actions that do not have audio-visual alignment. For example, many videos are of sporting events where there is no localized audio information, contain music, narration, or have no audio at all. Due to these challenges there are no existing TAL audio-visual fusion works to our knowledge that test their methods on THUMOS14.

Following previous work [27, 28, 57, 66, 63], we trained the model on the validation set and evaluate on the test set. Our results in Tab 4 demonstrate that our method struggles to handle this audio-visual disparity only improving on the 0.7 iou threshold.

5 Implementation

5.1 Feature Extraction

Visual Features: We use the features provided by existing works in TAL [63, 27, 57]. For EPIC-Kitchens features are extracted using a SlowFast network [15] pre-trained on EPIC-Kitchens [11]. During extraction we use a 32-frame input sequence with a stride of 16 to generate a set of 2304-D features.

Audio Features: For the audio preprocessing and feature extraction, we followed a series of well-established steps to derive meaningful representations:

1. **Resampling:** All audio data was resampled to a uniform rate of 16 kHz in mono.
2. **Spectrogram Computation:** We computed the spectrogram by extracting magnitudes from the Short-Time Fourier Transform (STFT). This utilized a window size of 25 ms, a hop size of 10 ms, and a periodic Hann window for the analysis.
3. **Mel Spectrogram Mapping:** The computed spectrogram was then mapped to a mel scale, producing a mel spectrogram with 64 mel bins that cover the frequency range from 125 Hz to 7500 Hz.
4. **Log Mel Spectrogram Stabilization:** To enhance the stability and avoid issues with the logarithm function, we calculated a stabilized log mel spectrogram as:

$$\text{Log-Mel} = \log(\text{Mel-Spectrogram} + 0.01)$$

   Here, the offset of 0.01 prevents the computation of the logarithm of zero.
5. **Framing:** Finally, the derived features were segmented into non-overlapping examples spanning 0.96 seconds each. Every example encapsulates 64 mel bands and 96 time frames, with each frame lasting 10 ms.

Following extraction, the features are projected to 128-D features via a VGG audio encoder network [20] pretrained on AudioSet [16]. The network outputs embeddings of shape $T \times 128$ where T is the temporal input dimension as defined in the paper.

6 Conclusion

We demonstrate an effective method for audio-visual fusion with Feature Pyramid Networks. Our drop-in method can be applied to any FPN architecture for temporal action localization and serves as a competitive benchmark for continued research in audio-visual fusion.

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911 AI4Media.

References

[1] Juan León Alcázar, Fabian Caba, Ali K Thabet, and Bernard Ghanem. Maas: Multi-modal assignation for active speaker detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 265–274, 2021.

[2] Humam Alwassel, Silvio Giancola, and Bernard Ghanem. Tsp: Temporally-sensitive pretraining of video encoders for localization tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3173–3183, 2021.

[3] Anurag Bagchi, Jazib Mahmood, Dolton Fernandes, and Ravi Kiran Sarvadevabhatla. Hear me out: Fusional approaches for audio augmented temporal action localization. *arXiv preprint arXiv:2106.14118*, 2021.

[4] Yueran Bai, Yingying Wang, Yunhai Tong, Yang Yang, Qiyue Liu, and Junhui Liu. Boundary content graph neural network for temporal action proposal generation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 121–137. Springer, 2020.

[5] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. Sst: Single-stream temporal action proposals. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2911–2920, 2017.

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

[7] Shuning Chang, Pichao Wang, Fan Wang, Hao Li, and Jiashi Feng. Augmented transformer with adaptive graph for temporal action proposal generation. *arXiv preprint arXiv:2103.16024*, 2021.

[8] Guo Chen, Yin-Dong Zheng, Limin Wang, and Tong Lu. Dcan: improving temporal action detection via dual context aggregation. In *AAAI*, 2022.

[9] Feng Cheng and Gedas Bertasius. Tallformer: Temporal action localization with long-memory transformer. *Eur. Conf. Comput. Vis.*, 2022.

[10] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018.

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision. *arXiv preprint arXiv:2006.13256*, 2020.

[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pages 1–23, 2021.

[13] 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. *arXiv preprint arXiv:1804.03619*, 2018.

[14] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Daps: Deep action proposals for action understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 768–784. Springer, 2016.

[15] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[16] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[17] Guoqiang Gong, Liangfeng Zheng, and Yadong Mu. Scale matters: Temporal scale aggregation network for precise action localization in untrimmed videos. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.

[18] Yilong He, Yong Zhong, Lishun Wang, and Jiachen Dang. Glformer: Global and local context aggregation network for temporal action detection. *Applied Sciences*, 12(17):8557, 2022.

[19] Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1914–1923, 2016.

[20] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[21] Haroon Idrees, Amir R Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017.

[22] Tae-Kyung Kang, Gun-Hee Lee, and Seong-Whan Lee. Htnet: Anchor-free temporal action localization with hierarchical transformers. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 365–370. IEEE, 2022.

[23] Evangelos Kazakos, Jaesung Huh, Arsha Nagrani, Andrew Zisserman, and Dima Damen. With a little help from my temporal context: Multimodal egocentric action recognition. *arXiv preprint arXiv:2111.01024*, 2021.

[24] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5492–5501, 2019.

[25] Chuming Lin, Jian Li, Yabiao Wang, Ying Tai, Donghao Luo, Zhipeng Cui, Chengjie Wang, Jilin Li, Feiyue Huang, and Rongrong Ji. Fast learning of temporal action proposal via dense boundary generator. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11499–11506, 2020.

[26] Chuming Lin, Chengming Xu, Donghao Luo, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Learning salient boundary feature for anchor-free temporal action localization. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3320–3329, 2021.

[27] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3889–3898, 2019.

[28] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[29] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[30] Qinying Liu and Zilei Wang. Progressive boundary refinement network for temporal action detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 11612–11619, 2020.

[31] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 21–37. Springer, 2016.

[32] Xiaolong Liu, Yao Hu, Song Bai, Fei Ding, Xiang Bai, and Philip HS Torr. Multi-shot temporal event localization: a benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12596–12606, 2021.

[33] Xiaolong Liu, Qimeng Wang, Yao Hu, Xu Tang, Shiwei Zhang, Song Bai, and Xiang Bai. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, 31:5427–5441, 2022.

[34] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Gaussian temporal awareness networks for action localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 344–353, 2019.

[35] Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Post-processing temporal action detection. *arXiv preprint arXiv:2211.14924*, 2022.

[36] Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Proposal-free temporal action detection via global segmentation mask learning. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part III*, pages 645–662. Springer, 2022.

[37] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *arXiv preprint arXiv:2107.00135*, 2021.

[38] Megha Nawhal and Greg Mori. Activity graph transformer for temporal action localization. *arXiv preprint arXiv:2101.08540*, 2021.

[39] Zhiwu Qing, Haisheng Su, Weihao Gan, Dongliang Wang, Wei Wu, Xiang Wang, Yu Qiao, Junjie Yan, Changxin Gao, and Nong Sang. Temporal context aggregation network for temporal action proposal refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 485–494, 2021.

[40] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.

[41] Merey Ramazanova, Victor Escorcia, Fabian Caba, Chen Zhao, and Bernard Ghanem. Owl (observe, watch, listen): Audiovisual temporal context for localizing actions in egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4879–4889, 2023.

[42] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[43] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[44] Dingfeng Shi, Qiong Cao, Yujie Zhong, Shan An, Jian Cheng, Haogang Zhu, and Dacheng Tao. Temporal action localization with enhanced instant discriminability. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2023.

[45] Deepak Sridhar, Niamul Quader, Srikanth Muralidharan, Yaoxin Li, Peng Dai, and Juwei Lu. Class semantics-based attention for action detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13739–13748, 2021.

[46] Jing Tan, Jiaqi Tang, Limin Wang, and Gangshan Wu. Relaxed transformer decoders for direct action proposal generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13526–13535, 2021.

[47] Tuan N Tang, Kwonyoung Kim, and Kwanghoon Sohn. Temporalmaxer: Maximize temporal context with only max pooling for temporal action localization. *arXiv preprint arXiv:2303.09055*, 2023.

[48] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018.

[49] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Int. Conf. Comput. Vis.*, 2019.

[50] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[52] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[53] Lining Wang, Haosen Yang, Wenhao Wu, Hongxun Yao, and Hujie Huang. Temporal action proposal generation with transformers. *arXiv preprint arXiv:2105.12043*, 2021.

[54] Weiyao Wang, Du Tran, and Matt Feiszli. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12695–12705, 2020.

[55] Yuetian Weng, Zizheng Pan, Mingfei Han, Xiaojun Chang, and Bohan Zhuang. An efficient spatio-temporal pyramid transformer for action detection. In *Eur. Conf. Comput. Vis.*, 2022.

[56] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020.

[57] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10156–10165, 2020.

[58] Ceyuan Yang, Yinghao Xu, Jianping Shi, Bo Dai, and Bolei Zhou. Temporal pyramid network for action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 591–600, 2020.

[59] Le Yang, Junwei Han, Tao Zhao, Nian Liu, and Dingwen Zhang. Structured attention composition for temporal action localization. *IEEE Transactions on Image Processing*, 2022.

[60] Le Yang, Houwen Peng, Dingwen Zhang, Jianlong Fu, and Junwei Han. Revisiting anchor mechanisms for temporal action localization. *IEEE Transactions on Image Processing*, 29:8535–8548, 2020.

[61] Min Yang, Guo Chen, Yin-Dong Zheng, Tong Lu, and Limin Wang. Basictad: an astounding rgb-only baseline for temporal action detection. *arXiv preprint arXiv:2205.02717*, 2022.

[62] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7094–7103, 2019.

[63] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 492–510. Springer, 2022.

[64] Shifeng Zhang, Cheng Chi, Yongqiang Yao, Zhen Lei, and Stan Z Li. Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[65] Chen Zhao, Ali K Thabet, and Bernard Ghanem. Video self-stitching graph network for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13658–13667, 2021.

[66] Peisen Zhao, Lingxi Xie, Chen Ju, Ya Zhang, Yanfeng Wang, and Qi Tian. Bottom-up temporal action localization with mutual regularization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16*, pages 539–555. Springer, 2020.

[67] Zixin Zhu, Wei Tang, Le Wang, Nanning Zheng, and Gang Hua. Enriching local and global contexts for temporal action localization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13516–13525, 2021.

[68] Zixin Zhu, Le Wang, Wei Tang, Ziyi Liu, Nanning Zheng, and Gang Hua. Learning disentangled classification and localization representations for temporal action localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3644–3652, 2022.