DEL: Dense Event Localization for Multi-modal Audio-Visual Understanding

Mona Ahmadian
University of Surrey
m.ahmadian@surrey.ac.uk

Amir Shirian
JPMorgan Chase
amirdonte15@gmail.com

Frank Guerin
University of Surrey
f.guerin@surrey.ac.uk

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

Abstract

*Real-world videos often exhibit overlapping events and intricate temporal dependencies, posing significant challenges for effective multimodal interaction modeling. We introduce **DEL**, a framework for dense semantic action localization, aiming to accurately detect and classify multiple actions at fine-grained temporal resolutions in long untrimmed videos. DEL consists of two key modules: the alignment of audio and visual features, which leverages masked self-attention to enhance intra-mode consistency, and a multimodal interaction refinement module that models cross-modal dependencies across multiple scales, enabling both high-level semantics and fine-grained details. We report results on multiple real-world Temporal Action Localization (TAL) datasets, UnAV-100, THUMOS14, ActivityNet 1.3, and EPIC-Kitchens-100. The source code will be made publicly available. These advances enable more accurate analysis of complex, real-world scenes, from surveillance to accessible media understanding.*

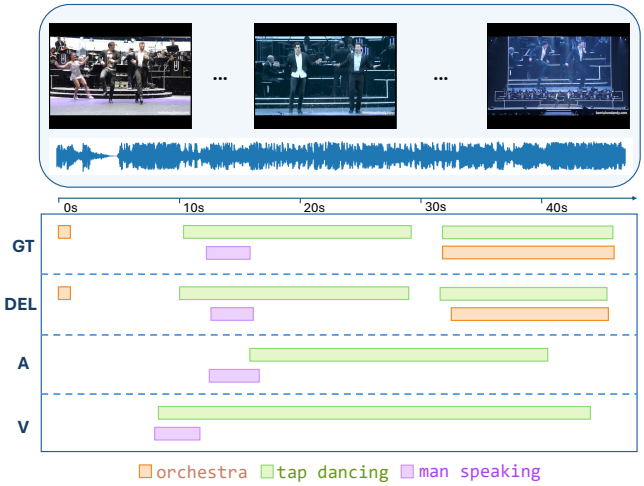

Figure 1. Real-world videos often feature overlapping events of different lengths, making localization difficult. This image compares ground-truth (GT) with predictions from DEL, an audio-only model (A), and a visual-only model (V). While A and V struggle with a specific category, **DEL** accurately detects both short and long events, even when overlapping.

1. Introduction

Temporal action localization (TAL) involves identifying and classifying action boundaries in untrimmed videos, a task made difficult by varying action durations and overlaps [28]. Real-world video understanding is inherently multimodal, requiring both visual and auditory cues [11, 21, 22]. For instance, distinguishing speech from silent mouthing is challenging using visuals alone, but can be resolved with audio input. Although audio and visual modalities are complementary, their fusion is non-trivial due to temporal misalignment, diverse event durations, and intricate cross-modal interactions [27]. Prior work in Audio-Visual Event Localization (AVE) has largely focused on trimmed videos with single events [7, 23, 29], whereas dense localization requires detecting all overlapping events across varying durations in untrimmed videos [1, 4, 8, 10]. Recent TAL models leverage transformers and Feature Pyramid Networks (FPN) for multi-scale visual reasoning [25, 31, 34, 37], but often neglect audio. A key challenge in audio-visual event localization is fusing multimodal information when events co-occur. Existing methods often process audio and visual streams independently or apply late fusion, limiting their ability to capture fine-grained temporal dependencies. Moreover, reliance on pre-trained feature extractors introduces misalignment due to domain gaps and differing objectives. While contrastive learning aids cross-modal alignment, most methods overlook intra-video structure, such as temporal coherence and cross-event correlations, which are vital for distinguishing similar events across time.

To address these issues, we propose **DEL**, a novel transformer-based framework that explicitly models cross-modal dependencies while preserving fine-grained temporal structure. DEL employs multi-scale fusion to support robust localization in densely overlapping scenarios. Two

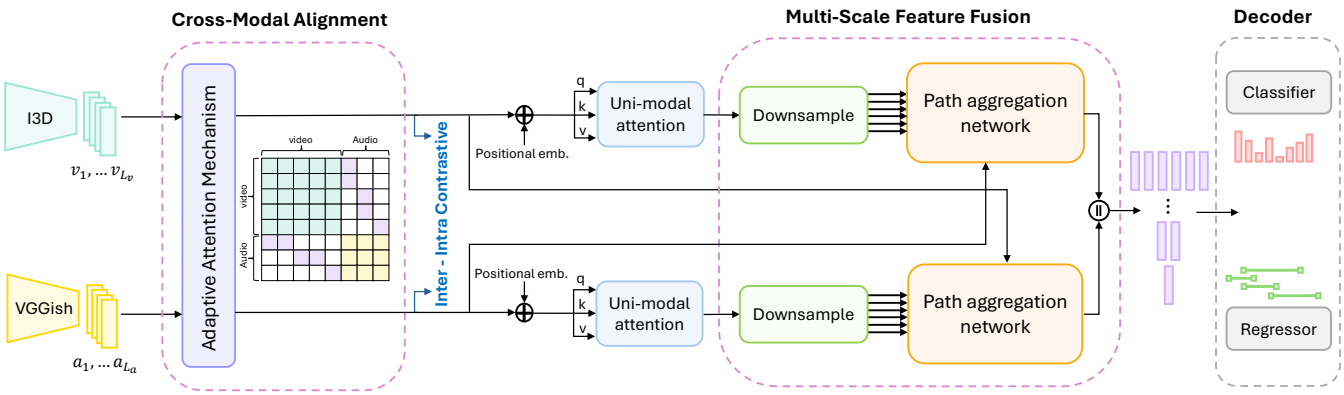

Figure 2. Overview of our proposed **DEL** framework. Our model integrates **(1)** an *adaptive attention mechanism* for aligning audio and visual features, **(2)** *inter- and intra-sample contrastive learning* to enhance event discrimination, and **(3)** a *multi-scale path aggregation network* for feature fusion. $\|$ represents the concatenation operation.

key modules underpin our approach: (1) a **multimodal adaptive attention** mechanism using masked self-attention to ensure temporal coherence and intra-modal consistency, and (2) a **path aggregation network** that captures both fine-grained and high-level temporal semantics. We further introduce a dual contrastive loss: intra-sample contrast enhances feature discrimination within modalities, while inter-sample contrast improves cross-modal alignment. A feature scoring mechanism automatically selects contrastive pairs, removing the need for manual sampling and improving training efficiency.

2. Related Works

Deep learning has driven advances in temporal action localization (TAL), enabling accurate detection of actions in untrimmed videos. TAL methods include two-stage approaches [14, 15], which first generate proposals, and single-stage methods that predict actions directly. Our work builds on the latter for efficiency. Within this context, anchor-based methods [3, 38] rely on predefined temporal regions, whereas anchor-free techniques [20, 35] directly regress event boundaries. Recent models integrate GNNs [36] and Transformers [30], with transformer-based FPNs [37] improving multi-scale temporal reasoning and localization. Audio-visual fusion has shown promise in video retrieval [12], but remains underexplored in TAL due to challenges like modality misalignment and asynchronous events [27]. Most methods employ late fusion, which limits fine-grained temporal modeling. While recent methods explore cross-attention [17, 33] they often fall short in modeling dynamic, multi-scale interactions critical for real-world event understanding. Contrastive learning [9] helps align modalities but typically ignores intra-video structure. In contrast, our **DEL** framework captures cross-modal dependencies via adaptive attention and multi-scale

3. Method Overview

Figure 2 illustrates the DEL framework for dense audio-visual event localization in untrimmed videos. Given tokenized audio-visual input, our method proceeds via three modules: (1) **Adaptive Attention** for dynamic cross-modal alignment; (2) **Score-based Contrastive Learning** to enhance feature discrimination; and (3) a **Path Aggregation Network** for robust multi-scale temporal fusion.

Problem Formulation

Given a video segmented into T audio-visual pairs $\mathbb{S} = \{(\mathbf{V_t}, \mathbf{A_t})\}_{t=1}^{T}$, where $\mathbf{V_t}$ and $\mathbf{A_t}$ represent visual and audio features at time t, we aim to predict localized events:

$$\hat{\mathbb{S}} = \{\hat{s}_t = (\delta_{start,t}, \delta_{end,t}, q(y_t))\}_{t=1}^{T},$$

where $q(y_t) \in [0,1]^{|\Lambda|}$ is the event classification probability for Λ, the set of all event classes, and $\delta_{start,t}, \delta_{end,t}$ are temporal offsets. The final predictions are:

$$\hat{\tau}_{start,t} = t - \delta_{start,t}, \quad \hat{\tau}_{end,t} = t + \delta_{end,t}, \quad \hat{\lambda}_t = \arg\max_{\lambda \in \Lambda} q(\lambda_t).$$

Adaptive Cross-Modal Attention

To effectively align temporally offset audio and visual signals, we employ an attention-based mechanism with learnable masking $\mathbf{M} \in \mathbb{R}^{(L_v+L_a)\times(L_v+L_a)}$, where L_v and L_a denote the lengths of the visual and audio sequences, respectively. Given concatenated input $\mathbf{X} = [\mathbf{V}|\mathbf{A}] \in \mathbb{R}^{(L_v+L_a)\times d}$, where d is the embedding dimension, we compute attention as:

$$\text{aat}_{i,j} = \frac{m_{i,j}\exp(\mathbf{Q}_i\mathbf{K}_j^\top/\sqrt{d})}{\sum_k m_{i,k}\exp(\mathbf{Q}_i\mathbf{K}_k^\top/\sqrt{d})},$$

with $\mathbf{Q}, \mathbf{K}$ derived from linear projections of $\mathbf{X}$. The mask guides both intra- and inter-modal alignment across temporally corresponding features.

Score-Based Contrastive Learning To improve event discrimination and modality alignment, we adopt a dual contrastive loss that operates both across and within video samples.

- **Inter-sample loss** aligns $[CLS_V]$ and $[CLS_A]$ tokens across paired samples.
- **Intra-sample loss** leverages token-level predictions—event score s_t and category c_t—to mine positive and hard-negative samples within a video.

The contrastive loss encourages alignment between correctly predicted segments and penalizes ambiguous ones:

$$\ell(z, z^+, z^-) = -\log\left(\frac{\exp(z^\top z^+/\tau)}{\exp(z^\top z^+/\tau) + \sum_k \exp(z^\top z_k^-/\tau)}\right),$$

where τ is a learnable temperature parameter.

Path Aggregation Network

To capture events of varying durations, we build a multi-scale feature pyramid. Modality-guided adapters integrate cross-modal cues:

$$V_l' = V_l \cdot \sigma\left(\max_j (V_l A_j^\top)\right)^\top, \quad A_l' = A_l \cdot \sigma\left(\max_k (A_l V_k^\top)\right)^\top,$$

where σ is the sigmoid activation function. These updated features are fused across scales and refined via multi-head attention (MHA):

$$V' = V + \mathrm{MHA}(V, \tilde{A}, \tilde{A}), \quad A' = A + \mathrm{MHA}(A, \tilde{V}, \tilde{V}),$$

where $\tilde{A}, \tilde{V}$ denote compact multi-scale tokens obtained via adaptive pooling.

Overall Objective Function The final loss combines contrastive and supervised objectives: $\mathscr{L}_{inter}$ and $\mathscr{L}_{intra}$, and the score cross entropy loss $\mathscr{L}_{score}$. Additionally, the classification head is trained using a cross-entropy loss, $\mathscr{L}_{\text{cls}}$, which ensures accurate event categorization, while the regression head is optimized with a smooth L1 loss, $\mathscr{L}_{\text{reg}}$, to refine the temporal boundaries of each detected event:

$$\mathscr{L}_{DEL} = \lambda_1 \mathscr{L}_{inter} + \lambda_2 \mathscr{L}_{intra} + \lambda_3 \mathscr{L}_{score} + \lambda_4 \mathscr{L}_{cls} + \lambda_5 \mathscr{L}_{reg},$$

with weights λ_i balancing each term.

4. Experiments

Dataset and Metrics We evaluate DEL on four benchmarks: THUMOS14 [10], ActivityNet-1.3 [1], EPIC-Kitchens-100 [4], and UnAV-100 [8]. Following standard practice, we report mean Average Precision (mAP) across multiple temporal IoU thresholds. To ensure statistical robustness, all results are averaged over five training runs.

Feature Encoder. For THUMOS14, ActivityNet, and UnAV-100, we adopt I3D [2], pretrained on Kinetics-400, for visual features, and VGGish [7], pretrained on AudioSet, for audio features. For EPIC-Kitchens-100, where fine-grained temporal resolution is critical, we follow [24, 37] in using SlowFast [5] pretrained on EPIC-Kitchens. All features are projected to a shared embedding space.

4.1. Main Results

THUMOS14. Tab. 1 shows that DEL achieves an average mAP of 71.9%, outperforming TriDet by +2.6%. DEL excels at higher tIoU thresholds, achieving 68.4% at 0.6 and 60.5% at 0.7, indicating strong temporal precision.

Method	0.3	0.4	0.5	0.6	0.7	Avg
MUSES [18]	68.9	64.0	56.9	46.3	31.0	-
ContextLoc [39]	68.3	63.8	54.3	41.8	26.2	50.9
A^2Net [35]	58.6	54.1	45.5	32.5	17.2	41.6
PBRNet [16]	58.5	54.6	51.3	41.8	29.5	-
AFSD [13]	67.3	62.4	55.5	43.7	31.1	52.0
TadTR [19]	62.4	57.4	49.2	37.8	26.3	46.6
Actionformer [37]	82.1	77.8	71.0	59.4	43.9	67.9
ASL [24]	83.1	79.0	71.7	59.7	45.8	66.8
TMaxer+MRAVFF [6]	82.2	78.2	71.5	59.9	45.3	67.4
TriDet [26]	**83.6**	**80.1**	**72.9**	62.4	47.4	69.3
DEL	81.0	78.0	71.8	**68.4**	**60.5**	**71.9**

Table 1. **Performance comparison on THUMOS14** We report mAP across multiple tIoU thresholds and compute the average mAP. Our method outperforms previous approaches on THUMOS14 with the same feature extraction.

ActivityNet-1.3. As shown in Tab. 2, DEL achieves the best overall performance with an average mAP of 38.0%, improving upon TriDet by +1.2%. The performance gains across all thresholds show that DEL generalizes well to diverse and long-form activities.

Method	0.5	0.75	0.95	Avg
MUSES [18]	50.0	35.0	6.6	34.0
ContextLoc [39]	56.0	35.2	3.6	34.2
VSGN [38]	52.3	35.2	8.3	34.7
A^2Net [35]	43.6	28.7	3.7	27.8
PBRNet [16]	54.0	35.0	9.0	35.0
AFSD [13]	52.4	35.3	6.5	34.4
TadTR [19]	49.1	32.6	8.5	32.3
Actionformer [37]	53.5	36.2	8.2	35.6
ASL [24]	54.1	37.4	8.0	36.2
TriDet [26]	54.7	38.0	8.4	36.8
DEL	**56.9**	**42.5**	**14.7**	**38.0**

Table 2. **Performance evaluation on ActivityNet 1.3.** We present mAP and average mAP results across various tIoU thresholds. Our approach surpasses previous methods with the same feature extraction.

EPIC-Kitchens-100. Tab. 3 presents results on this fine-grained kitchen activity dataset. Using I3D+VGGish, DEL achieves 27.1% (verb) and 25.2% (noun) average mAP. Using stronger features (VMAE2+ASlowFast), DEL achieves 30.5% and 28.1%, outperforming all baselines, including TIM. These gains highlight DEL's capability to handle densely overlapping, fine-grained actions—particularly those with subtle audio-visual interplay.

UnAV-100. In Tab. 4, DEL achieves 51.1% average mAP, outperforming the UnAV baseline by +3.3%. Its performance improves consistently with increasing tIoU thresholds, showing accurate boundary localization even in overlapping audio-visual scenarios.

4.2. Ablation Experiments

We conduct ablations on UnAV-100, a challenging dataset with dense, overlapping audio-visual events. We evaluate the impact of key components, feature extractors, fusion design, and input modalities.

Task	Method	Frozen Features	0.1	0.2	0.3	0.4	0.5	Avg
Verb	BMN [15]	I3D+VGGish	10.8	8.8	8.4	7.1	5.6	8.4
	G-TAD [32]	I3D+VGGish	12.1	11.0	9.4	8.1	6.5	9.4
	ActionFormer [37]	I3D+VGGish	26.6	25.4	24.2	22.3	19.1	23.5
	ASL [24]	I3D+VGGish	27.9	-	25.5	-	19.8	24.6
	ActionFormer + MRAV-FF [6]	I3D+VGGish	27.6	26.8	25.3	23.4	19.8	24.6
	TriDet [26]	I3D+VGGish	28.6	27.4	26.1	24.2	20.8	25.4
	TIM	VMAE2+ASlowFast	32.9	31.6	29.6	27.0	22.2	28.6
	DEL	I3D+VGGish	32.2	29.9	27.8	25.1	20.8	27.1
	DEL	VMAE2+ASlowFast	**35.1**	**33.6**	**31.5**	**28.8**	**23.5**	**30.5**
Noun	BMN [15]	I3D+VGGish	10.3	8.3	6.2	4.5	3.4	6.5
	G-TAD [32]	I3D+VGGish	11.0	10.0	8.6	7.0	5.4	8.4
	ActionFormer [37]	I3D+VGGish	25.2	24.1	22.7	20.5	17.0	21.9
	ASL [24]	I3D+VGGish	26.0	-	23.4	-	17.7	22.6
	ActionFormer + MRAV-FF [6]	I3D+VGGish	26.4	25.4	23.6	21.2	17.4	22.8
	TriDet [26]	I3D+VGGish	27.4	26.3	24.6	22.2	18.3	23.8
	TIM	VMAE2+ASlowFast	36.4	34.8	32.1	28.7	22.7	31.0
	DEL	I3D+VGGish	29.5	28.4	26.2	22.9	19.3	25.2
	DEL	VMAE2+ASlowFast	**33.1**	**31.3**	**29.3**	**26.1**	**20.8**	**28.1**

Table 3. **Performance on the EPIC-Kitchens-100 validation set** across multiple tIoU thresholds, with average mAP reported. Our method outperforms all baselines by a significant margin using the same feature extraction.

Method	0.5	0.6	0.7	0.8	0.9	Avg
VSGN [38]	24.5	20.2	15.9	11.4	6.8	24.1
TadTR [19]	30.4	27.1	23.3	19.4	14.3	29.4
ActionFormer [37]	43.5	39.4	33.4	27.3	17.9	42.2
UnAV [8]	50.6	45.8	39.8	32.4	21.1	47.8
DEL	**53.4**	**48.1**	**42.6**	**35.6**	**26.9**	**51.1**

Table 4. **Performance on the UnAV-100 test set**, showcasing our method's significant improvement over all baselines using the same feature extraction. We report mAP and average mAP at various tIoU thresholds.

Component Analysis. Table 5 shows the impact of removing our core modules: Adaptive Attention for Cross-modal Alignment (AAC), Score-based Contrastive Learning (SCL), and the Path Aggregation Network (PAN). Removing any component leads to a notable drop in performance, confirming their complementary contributions to robust localization.

AAC	SCL	PAN	0.5	0.6	0.7	0.8	0.9	Avg
×	✓	✓	51.1	45.7	41.0	34.5	25.8	49.6
✓	×	✓	51.5	44.7	38.7	33.3	25.8	49.7
✓	✓	×	51.3	45.0	39.4	33.8	25.3	49.5
✓	✓	✓	53.4	48.1	42.6	35.6	26.9	51.1

Table 5. **Component-wise ablation study**, evaluating the individual contributions of our proposed Adaptive Attention for Cross-Modal Alignment (AAC), Score-Based Contrastive Learning (SCL), and Path Aggregation Network for Multi-Scale Feature Fusion (PAN) modules.

Feature Pyramid Depth. Table 6 examines the number of temporal pyramid levels L. Six levels yield the best performance, capturing both fine- and coarse-scale cues. Fewer levels limit context modeling, while more introduce redundancy.

Feature Extractor Choice. In Tab. 7, we evaluate DEL with alternative encoders: DINOv2 for video and MERT for audio. This setup improves mAP by +1.4 (UnAV-100) and +1.3 (THUMOS14), showing that DEL generalizes across feature types.

L	0.5	0.6	0.7	0.8	0.9	Avg
1	47.5	42.7	37.3	30.6	22.0	45.5
2	48.5	43.2	37.7	31.2	23.0	46.4
4	48.4	43.5	38.2	32.0	23.5	47.2
6	53.4	48.1	42.6	35.6	26.9	51.1
7	51.0	45.9	40.1	33.9	24.6	49.0

Table 6. Ablation study on the design of the feature pyramid. L shows the number of layers for both audio and video.

Features	0.3	0.4	0.5	0.6	0.7	Avg
THUMOS14						
I3D+Vggish	81.0	78.0	71.8	68.4	60.5	71.9
DINOv2+MERT	81.5	79.1	73.1	70.8	64.6	73.3
UnAV-100						
I3D+Vggish	53.4	48.1	42.6	35.6	26.9	51.1
DINOv2+MERT	55.0	49.7	44.1	37.4	28.4	52.7

Table 7. Evaluation on THUMOS14 and UnAV-100 incorporating DINOv2 for video features and MERTv1 for audio features.

Modal Input Analysis. Table 8 compares DEL with visual-only and audio-only variants. Using both modalities improves average mAP by +13.2 (UnAV-100), +2.1 (THUMOS14), and +2.1 (EPIC-Verbs), confirming the critical role of audio-visual fusion.

Data	Video	Audio	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.75	0.8	0.9	0.95	Avg
2*THUMOS14	✓	×	-	-	79.9	76.8	70.7	67.6	59.7	-	-	-	-	70.8
	✓	✓	-	-	81.0	78.0	71.8	68.4	60.5	-	-	-	-	71.9
2*ActivityNet 1.3	✓	×	-	-	-	-	53.7	-	-	40.1	-	-	13.8	35.8
	✓	✓	-	-	-	-	56.9	-	-	42.5	-	-	14.7	38.0
2*EPIC-Kitchens-100 (Verb)	✓	×	30.9	28.7	26.6	24.1	20.1	-	-	-	-	-	-	26.0
	✓	✓	32.2	29.9	27.8	25.1	20.8	-	-	-	-	-	-	27.1
2*EPIC-Kitchens-100 (Noun)	✓	×	28.0	27.0	24.9	21.6	18.5	-	-	-	-	-	-	23.9
	✓	✓	29.5	28.4	26.2	22.9	19.3	-	-	-	-	-	-	25.2

Table 8. DEL performance with various modality combinations. Fusing audio and video yields the best results, emphasizing the importance of multi-modal input.

5. Conclusion

We presented DEL, a dense audio-visual event localization framework for untrimmed videos. By combining adaptive attention with a dual contrastive learning strategy, DEL effectively aligns audio and visual streams while modeling fine-grained temporal dependencies. A multi-scale path aggregation network further enhances cross-modal fusion. DEL achieves state-of-the-art results across four challenging benchmarks—UnAV-100, THUMOS14, ActivityNet 1.3, and EPIC-Kitchens-100—demonstrating its ability to localize overlapping events with high precision. DEL narrows the gap between controlled benchmark tasks and the complexity of real-world audiovisual scenarios, offering a strong foundation for downstream applications such as accessibility tools and intelligent video summarization.

References

[1] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 1, 3

[2] 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. 3

[3] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1130–1139, 2018. 2

[4] 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, 2022. 1, 3

[5] 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. 3

[6] Edward Fish, Jon Weinbren, and Andrew Gilbert. Multi-resolution audio-visual feature fusion for temporal action localization. *arXiv preprint arXiv:2310.03456*, 2023. 3, 4

[7] 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. 1, 3

[8] Tiantian Geng, Teng Wang, Jinming Duan, Runmin Cong, and Feng Zheng. Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22942–22951, 2023. 1, 3, 4

[9] Xixi Hu, Ziyang Chen, and Andrew Owens. Mix and localize: Localizing sound sources in mixtures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10483–10492, 2022. 2

[10] 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. 1, 3

[11] Licheng Jiao, Yuhan Wang, Xu Liu, Lingling Li, Fang Liu, Wenping Ma, Yuwei Guo, Puhua Chen, Shuyuan Yang, and Biao Hou. Causal inference meets deep learning: A comprehensive survey. *Research*, 7:0467, 2024. 1

[12] 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. 2

[13] 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. 3

[14] 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. 2

[15] 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. 2, 4

[16] Qinying Liu and Zilei Wang. Progressive boundary refinement network for temporal action detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11612–11619, 2020. 3

[17] Shuo Liu, Weize Quan, Chaoqun Wang, Yuan Liu, Bin Liu, and Dong-Ming Yan. Dense modality interaction network for audio-visual event localization. *IEEE Transactions on Multimedia*, 25:2734–2748, 2022. 2

[18] 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. 3

[19] 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. 3, 4

[20] Sauradip Nag, Xiatian Zhu, Yi-Zhe Song, and Tao Xiang. Proposal-free temporal action detection via global segmentation mask learning. In *European Conference on Computer Vision*, pages 645–662. Springer, 2022. 2

[21] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems*, 34:14200–14213, 2021. 1

[22] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 631–648, 2018. 1

[23] Arjun Prashanth, SL Jayalakshmi, and R Vedhapriyavadhana. A review of deep learning techniques in audio event recognition (aer) applications. *Multimedia Tools and Applications*, 83(3):8129–8143, 2024. 1

[24] Jiayi Shao, Xiaohan Wang, Ruijie Quan, Junjun Zheng, Jiang Yang, and Yi Yang. Action sensitivity learning for temporal action localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13457–13469, 2023. 3, 4

[25] Dingfeng Shi, Qiong Cao, Yujie Zhong, Shan An, Jian Cheng, Haogang Zhu, and Dacheng Tao. Temporal action localization with enhanced instant discriminability. *arXiv preprint arXiv:2309.05590*, 2023. 1

[26] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18857–18866, 2023. 3, 4

[27] 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. 1, 2

[28] Elahe Vahdani and Yingli Tian. Deep learning-based action detection in untrimmed videos: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4302–4320, 2022. 1

[29] Satvik Venkatesh, David Moffat, and Eduardo Reck Miranda. You only hear once: a yolo-like algorithm for audio segmentation and sound event detection. *Applied Sciences*, 12(7):3293, 2022. 1

[30] Lining Wang, Haosen Yang, Wenhao Wu, Hongxun Yao, and Hujie Huang. Temporal action proposal generation with transformers. *arXiv preprint arXiv:2105.12043*, 2021. 2

[31] Yuetian Weng, Zizheng Pan, Mingfei Han, Xiaojun Chang, and Bohan Zhuang. An efficient spatio-temporal pyramid transformer for action detection. In *European Conference on Computer Vision*, pages 358–375. Springer, 2022. 1

[32] 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. 4

[33] Cheng Xue, Xionghu Zhong, Minjie Cai, Hao Chen, and Wenwu Wang. Audio-visual event localization by learning spatial and semantic co-attention. *IEEE Transactions on Multimedia*, 25:418–429, 2021. 2

[34] 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. 1

[35] 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. 2, 3

[36] 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. 2

[37] Chen-Lin Zhang, Jianxin Wu, and Yin Li. Actionformer: Localizing moments of actions with transformers. In *European Conference on Computer Vision*, pages 492–510. Springer, 2022. 1, 2, 3, 4

[38] 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. 2, 3, 4

[39] 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. 3