One-Stage Open-Vocabulary Temporal Action Detection Leveraging Temporal Multi-scale and Action Label Features

Trung Thanh Nguyen^{1,2}, Yasutomo Kawanishi^{2,1}, Takahiro Komamizu^{3,1} and Ichiro Ide^{1,3}

¹Graduate School of Informatics, Nagoya University, Nagoya, Aichi 464-8601, Japan

²Guardian Robot Project, Information R&D and Strategy Headquarters, RIKEN, Seika, Kyoto 619-0288, Japan

³Mathematical and Data Science Center, Nagoya University, Nagoya, Aichi 464-8601, Japan

Abstract—Open-vocabulary Temporal Action Detection (Open-vocab TAD) is an advanced video analysis approach that expands Closed-vocabulary Temporal Action Detection (Closedvocab TAD) capabilities. Closed-vocab TAD is typically confined to localizing and classifying actions based on a predefined set of categories. In contrast, Open-vocab TAD goes further and is not limited to these predefined categories. This is particularly useful in real-world scenarios where the variety of actions in videos can be vast and not always predictable. The prevalent methods in Open-vocab TAD typically employ a 2-stage approach, which involves generating action proposals and then identifying those actions. However, errors made during the first stage can adversely affect the subsequent action identification accuracy. Additionally, existing studies face challenges in handling actions of different durations owing to the use of fixed temporal processing methods. Therefore, we propose a 1-stage approach consisting of two primary modules: Multi-scale Video Analysis (MVA) and Video-Text Alignment (VTA). The MVA module captures actions at varying temporal resolutions, overcoming the challenge of detecting actions with diverse durations. The VTA module leverages the synergy between visual and textual modalities to precisely align video segments with corresponding action labels, a critical step for accurate action identification in Open-vocab scenarios. Evaluations on widely recognized datasets THUMOS14 and ActivityNet-1.3, showed that the proposed method achieved superior results compared to the other methods in both Open-vocab and Closed-vocab settings. This serves as a strong demonstration of the effectiveness of the proposed method in the TAD task.

I. INTRODUCTION

Humans' ability to recognize unseen objects or actions with just their name or simple explanations stems from their capacity to apply accumulated relevant knowledge from past experiences. In recent years, advancements in large-scale Vision-and-Language (V&L) models have realized this capability on computers. These models learn shared representations among images and texts by leveraging diverse imagetext pairs through Contrastive Learning techniques [16]. As a result, these models can effectively extract valuable information from textual descriptions and visual representations for various tasks such as multimodal understanding [7], [27], [38], semantic comprehension [13], [18], [19], fewshot learning [1], [29], [31], and zero-shot learning [8], [20], [47]. This capability is attained without extensive training, allowing them to exhibit excellent performance across these tasks. Consequently, this progress has inspired researchers to explore the utilization of V&L models in various domains, including object detection [10], [48], action recognition [34], [45], and temporal action detection [15], [37].





(b) Illustration of the 1-stage approach (Proposed method).

Fig. 1. Comparing the 2-stage approach with the proposed method. (a) The 2-stage approach involves localizing temporal actions through proposal generation and utilizing the identified intervals for action identification in the alignment stage. (b) The proposed method leverages multi-scale features for both video-text alignment and action localization.

The Closed-vocabulary Temporal Action Detection (Closed-vocab TAD) task entails localizing temporal actions within videos and classifying the corresponding action classes, assuming the action classes are defined in advance. In contrast, Open-vocabulary Temporal Action Detection (Open-vocab TAD) requires localizing and identifying temporal actions absent from the training set. This means that Open-vocab TAD aims to accurately detect actions that have not been previously encountered. This poses a more challenging task, requiring the model to handle novel and unanticipated actions during the localization and identification processes. In existing studies [15], [37], a 2-stage TAD approach is adopted for Open-vocab TAD. In the first stage, the temporal actions are localized, and then, in the second stage, the identified interval of each action is utilized for action identification. However, errors in the first stage can influence the accuracy of the action identification. Moreover, existing Open-vocab TAD approaches face challenges in handling actions of different durations owing to the use of fixed temporal processing methods. This limitation arises when confronted with unseen actions, as the fixed temporal processing approach may not adequately

capture the temporal characteristics of these novel actions.

Figure 1 illustrates the comparison between the conventional and proposed methods. To address the limitations of the current Open-vocab TAD approach, we propose a 1stage approach that combines temporal action localization and identification for the Open-vocab TAD task. Moreover, the proposed method aims to overcome the challenges of different durations of actions by offering a solution that implements a multi-scale component.

The contributions of this study are summarized as follows:

- We propose a 1-stage approach for Open-vocab TAD that consists of two main modules: Multi-scale Video Analysis (MVA) and Video-Text Alignment (VTA), which are effective in addressing the challenges associated with Open-vocab scenarios and enabling accurate detection of a wide range of actions.
- We introduce a novel fusion strategy that combines temporal multi-scale features extracted from videos with action label features. This integration enhances the performance of action detection by effectively capturing actions of various lengths, thereby improving the accuracy and robustness of the model.
- We conduct extensive evaluations on widely used TAD datasets, including the THUMOS14 [12] and ActivityNet-1.3 [9] datasets. Through these evaluations, we demonstrate the effectiveness of the proposed MVA and VTA modules in achieving superior performance in both Open-vocab and Closed-vocab settings.

This paper is organized as follows: First, we briefly summarize the relevant literature in Section II. Details of the proposed method are presented in Section III, followed by an evaluation in Section IV. Finally, Section V concludes the paper and discusses future directions.

II. RELATED WORK

A. Closed-vocabulary Temporal Action Detection (Closed-vocab TAD)

Closed-vocab TAD focuses on action detection from untrimmed videos. Its approaches can be broadly classified into two types: 1-stage approaches [24], [25], [50], which are trained end-to-end and directly predict and classify action segments, and 2-stage approaches [35], [46], [51], which employ a range of techniques to predict candidate segments and subsequently identify them using action classifiers. In the context of the 1-stage method, a hierarchical architecture is constructed using a combination of Convolutional Neural Networks (CNNs) and Graph Neural Networks (GNNs). On the other hand, in the 2-stage approach, most previous studies emphasize the proposal generation phase, involving the prediction of action boundary probabilities and dense matching of start and end instants based on prediction scores. However, these approaches rely on a predefined set of actions for both the training and inference stages, requiring careful consideration of the completeness of the action annotations.

B. Transformer-based Closed-vocab TAD

In recent years, there has been a notable trend in Closedvocab TAD to harness the power of transformers, driven by the remarkable success of transformers in various domains like machine translation. Several recent studies [25], [40], [43] have embraced the attention mechanism inherent in transformers to enhance the performance of action detection. Specifically, DEtection TRansformer (DETR) [4] introduces a Transformer-based approach for image detection, where it learns shared decoder input features for all input videos and detects a fixed number of outputs. Building upon this, Liu et al. [25] proposes an end-to-end framework for Closed-vocab TAD. This training paradigm is known for its high efficiency and rapid prediction capabilities. Furthermore, Zhang et al. [50] employ a transformer-based encoder to extract video representations. In our work, we also utilize the capabilities of a Transformer-based encoder to extract video features, incorporating multi-scale features into our approach.

C. V&L models

In recent years, there has been growing interest in V&L models. They can learn unified representations for both images and texts, enabling the successful completion of diverse tasks that were previously considered challenging. Notable V&L models include Contrastive Language-Image Pre-training (CLIP) [36], A Large-scale ImaGe and Noisytext embedding (ALIGN) [14], and Batch, dAta and model SIze Combined scaling (BASIC) [33]. The shared embedding space learned by these models from large-scale Internet datasets has enabled highly accurate Open-vocab classification tasks without fine-tuning and expensive training processes. In the domain of videos, V&L models have been utilized for action classification tasks [15], [44] and videotext retrieval tasks [28]. By combining V&L models, these approaches can recognize new actions in unseen videos containing novel scenes. This advancement paves the way for recognizing unseen actions in real-world scenarios.

D. V&L for TAD

Recent research in the field of V&L applied to TAD has aimed to tackle two significant challenges: Open-set Temporal Action Detection (Open-set TAD) [2], [6] and Open-vocab TAD [15], [37]. Bao et al. [2] and Chen et al. [6] have proposed a framework for Open-set TAD by introducing an "Unknown" class to handle missing classes, whereas Open-vocab TAD focuses on dealing with an openended vocabulary of action classes. Both Ju et al. [15] and Rathod et al. [37] focus on a general Open-vocab TAD and employ a 2-stage model that replaces the supervised classifier in V&L feature comparisons. Our work differs from those in [15], [37] in several aspects. First, to avoid error propagation, a 1-stage model is adopted for both action localization and identification. Furthermore, to adapt to the diverse lengths of each action, multi-scale features along the temporal axis are utilized to improve the performance of Open-vocab TAD.



(b) Each layer in Multiscale component.

Fig. 2. (a) Overview of the proposed method. Each video frame is passed through a Pre-trained Image/Video Encoder, followed by Multi-scale Video Analysis and a Decoder for segment detection. Text labels are embedded using a Pre-trained Text Encoder and aligned with the video features to comprehend the relationship and accurately determine the action labels. (b) Each layer in the Multi-scale component utilizes a Transformer Encoder for feature extraction, followed by depthwise 1D convolution for downsampling between layers.

III. PROPOSED METHOD

For given input video frames $I = (I_1, I_2, \ldots, I_T)$, our objective is to estimate a series of temporal actions and corresponding classes $Y = \{y_1, y_2, \dots, y_N\}$ within the context of Open-vocab TAD $(D_{\text{train}} \cap D_{\text{test}} = \emptyset)$, which signifies that the actions present in the test set (D_{test}) do not appear in the training set (D_{train}) . Here, T is the length of the video, and N is the number of temporal actions in I. We extract a feature vector $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T)$ from the input video frames I, where the maximum value T of frames varies depending on the length of the video. The input text labels associated with actions are also extracted as features when combined with a prompt $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_M\}$, where M is the number of temporal actions in the training set (D_{train}) . These text features are integrated with video features to predict the labels of temporal actions. The output of the model is represented as a set of $\mathbf{y}_i = (s_i, e_i, a_i)$, where s_i and e_i denote the start and end times, respectively, and $a_i \in D_{\text{train}} \cup D_{\text{test}}$ is one of the action labels used for training or testing.

To elucidate the proposed method, we begin with an overview in Section III-A, followed by detailed explanations of its critical components in Sections III-B and III-C. Finally, we elaborate on learning objectives in Section III-D, which play crucial roles in enhancing the overall performance.

A. Overview

Figure 2 illustrates an overview of the proposed method. We adopt a 1-stage detection approach for temporal action detection in Open-vocab TAD. The proposed method consists of two key components: (1) Multi-scale Video Analysis (MVA) and (2) Video-Text Alignment (VTA) modules. The former determines the start and end times of actions, as well as determines whether that time frame contains actions or not. On the other hand, the latter is responsible for learning the relationship between video and text features to determine the labels for the actions.

B. Multi-scale Video Analysis (MVA) Module

The input for the MVA module is the video frames I, which are then encoded through a pre-trained image/video encoder to obtain feature sequences $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T) \in \mathbb{R}^{T \times D}$. The encoded information is then used to construct a multi-scale feature representation $\mathbf{Z}^* = {\mathbf{Z}^1, \mathbf{Z}^2, \dots, \mathbf{Z}^L}$, where L represents the number of scales corresponding to the hierarchical levels of the network. Finally, the feature representation is decoded to determine the start and end times of actions and assess the presence or absence of the action.

1) Projection layer: At the beginning, a shallow neural network is employed as a projection function $E : \mathbb{R}^D \to \mathbb{R}^{D'}$ to embed each input feature \mathbf{x}_t into a D'-dimensional space, resulting in $\mathbf{Z}^0 = (E(\mathbf{x}_1), E(\mathbf{x}_2), \dots, E(\mathbf{x}_T))$ as the output.

2) Multi-scale layer: The feature vector \mathbf{Z}^0 serves as the input for the multi-scale component. The embedded features \mathbf{Z}^l are then transformed into the feature representation \mathbf{Z}^{l+1} of the next scale layer using the Transformer Encoder [41] function f_l^{l+1} as:

$$\mathbf{Z}^{l+1} = f_l^{l+1}(\mathbf{Z}^l).$$

It is important to note that any function, such as a CNN [17], GNN [49], can be used for this transformation. Figure 2(b)

shows the architecture of the Transformer Encoder. The idea behind using the Transformer Encoder for the Open-vocab TAD task is to leverage the self-attention mechanism, which enables the calculation of frame similarity and the creation of a weight matrix for video frames. This allows automatic extraction of important frames, focusing on relevant action-containing segments and eliminating irrelevant noise. Within the Transformer Encoder, the Multi-head Attention mechanism captures diverse temporal dependencies between frames, while the MLP (Multi-Laver Perceptron) further refines the attended features, extracting discriminative information for precise action detection within video sequences. In this study, we utilize a strided depthwise 1D convolution after the Transformer Encoder for downsampling between layers. This convolution operation reduces the feature size by half in each subsequent layer. This process is applied for a total of L layers, resulting in a multi-scale feature representation denoted as $\mathbf{Z}^* = {\mathbf{Z}^1, \mathbf{Z}^2, \dots, \mathbf{Z}^L}.$

3) Decoder: The decoder D predicts sequence labeling for each frame using the multi-scale feature \mathbb{Z}^* . It estimates the probabilities of action occurrence and their corresponding time intervals, denoted as $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_N\}$. For each $\hat{y} = \{d_t^s, d_t^e, p(a_t)\}$, where $d_t^s = s_t + t$ represents the time difference from frame t to start time $s_t, d_t^e = e_t - t$ represents the time difference from frame t to end time e_t , and $p(a_t)$ indicates the probability of the presence of an action. The decoder incorporates a lightweight convolutional network with two heads: the Boundary Regression head and Background Classification head. The former estimates each action's start and end times within the video frames (d_t^s, d_t^e) , while the latter predicts the probabilities of action occurrence for each frame $p(a_t)$.

C. Video-Text Alignment (VTA) Module

The VTA module receives text labels of actions as input and integrates them with a prompt. These texts are transformed into text features using a pre-trained text encoder and are denoted as $\mathbf{A} = \{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_M\}$, where Mcorresponds to the number of classes in the training set (D_{train}) . Subsequently, the text features \mathbf{A} are combined with the length-aware pooled video features \mathbf{Z}_p^0 and $\mathbf{Z}_p^l \in \mathbf{Z}_p^*$ to align the text and video representations.

1) Length-aware pooling: In the proposed method, we take action-related features to establish alignment with textual information. To achieve this, action representations are extracted from the video features in projection layer \mathbf{Z}^0 and each layer $\mathbf{Z}^l \in \mathbf{Z}^*$ based on ground-truth segments. Following that, average pooling is performed on the features extracted from each action interval. This pooling process combines the features within the intervals and generates representative features denoted as \mathbf{Z}_p^0 and each layer $\mathbf{Z}_p^l \in \mathbf{Z}_p^*$. These representative features are then utilized to align with the corresponding text, facilitating the synchronization between video and text information.

2) Video-text alignment: The text features A are aligned with the video features extracted from \mathbf{Z}_p^0 and corresponding scale layer $\mathbf{Z}_p^l \in \mathbf{Z}_p^*$. We calculate the similarity between these features using the dot product, represented as $\mathbf{Z}_p^0 \cdot \mathbf{A}^{\top}$ and $\mathbf{Z}_p^l \cdot \mathbf{A}^{\top}$, where \top denotes the transpose operation. The dot product measures the similarity based on the magnitude and direction of their components, indicating the level of alignment between the text and video features. By applying the dot product, we can quantify the similarity and establish meaningful associations between textual and visual information in video-text alignment.

D. Learning Objectives

1) Objective function for MVA: This module utilizes two loss functions to the Boundary Regression head and Background Classification head for every frame t $(1 \le t \le T)$. The former employs Distance Intersection over Union (DIoU) loss [53] to accurately regress the distances to the boundaries of the actions, aiming for precise localization of the actions as:

$$L_{\rm BR}^t = 1 - \rm{IoU} + \mathcal{R}_{\rm DIoU}, \tag{1}$$

where IoU represents the Intersection over Union between the predicted and ground-truth action interval. The term \mathcal{R}_{DIoU} measures the inconsistency between the predicted interval and the ground-truth interval using the DIoU metric. The latter employs Focal loss [23] to effectively handle imbalanced samples between background and action as:

$$L_{\rm BC}^t = -\alpha_t (1 - p_t)^{\gamma} \log(p_t), \qquad (2)$$

where p_t is the predicted probability of an action occurrence, α_t is the balancing factor to address the class imbalance, and γ is a modulating factor that focuses on hard samples. The overall loss function of the MVA module (L_{MVA}) is defined as the sum of the above losses (Eqn. 1 and Eqn. 2) for each time step as:

$$L_{\rm MVA} = \sum_{t}^{T} \left(L_{\rm BR}^{t} + \lambda_1 L_{\rm BC}^{t} \right), \tag{3}$$

where λ_1 is a balancing coefficient.

2) Objective function for VTA: This module aims to model the cross-modal relationship between two modalities: video features and text features. The principle is to minimize the distance between representations of corresponding videotext pairs, while encouraging those of non-corresponding pairs. The learning objective L_{VTA} consists of two contrastive terms: $L_{\mathbf{Z}^0 \to \mathbf{A}}$ (Eqn. 4) and $L_{\mathbf{Z}^* \to \mathbf{A}}$ (Eqn. 5). The former aligns the video projection features with text features, while the latter aligns the video multi-scale features with text features. Below are the details of the VTA loss:

$$L_{\mathbf{Z}^0 \to \mathbf{A}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp((\mathbf{Z}_p^0 \cdot \mathbf{A}^\top)^+ / \tau)}{\sum_{j=1}^{M} \exp((\mathbf{Z}_p^0 \cdot \mathbf{A}^\top)^- / \tau)}, \quad (4)$$

$$L_{\mathbf{Z}^* \to \mathbf{A}} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{l=1}^{L} \log \frac{\exp((\mathbf{Z}_p^l \cdot \mathbf{A}^\top)^+ / \tau)}{\sum_{j=1}^{M} \exp((\mathbf{Z}_p^l \cdot \mathbf{A}^\top)^- / \tau)},$$

$$L_{\text{VTA}} = L_{\mathbf{Z} \to \mathbf{A}} + \lambda_2 L_{\mathbf{Z}^* \to \mathbf{A}},$$
(6)

where N is the number of temporal actions, M is the number of text classes, L is the number of multi-scale layers, "+" and "-" represent a pair of samples that are corresponding and non-corresponding, respectively, τ is a temperature hyperparameter controlling the impact of penalties on hard negative samples, and λ_2 is a balancing coefficient.

3) Overall objective function: The overall loss function of the proposed method is defined as the sum of the MVA loss (Eqn. 3) and VTA loss (Eqn. 6), multiplied by a balance coefficient λ_3 as:

$$L_{\text{Total}} = L_{\text{MVA}} + \lambda_3 L_{\text{VTA}}.$$
 (7)

IV. EVALUATION

In this section, we evaluate the proposed method extensively through a series of comprehensive experiments. In addition, we perform in-depth ablation studies to gain clearer insights into the key characteristics of the proposed method.

A. Experimental Conditions

1) Datasets: We perform evaluations on the THUMOS14 [12] and ActivityNet-1.3 [9] datasets, which are widely used for TAD.

- THUMOS14 dataset contains 413 videos with 20 action categories, with an average of 15 instances per video.
- ActivityNet-1.3 dataset is a large-scale action dataset, consisting of 200 activity classes and approximately 20,000 videos with more than 600 hours.

2) Data preparation: We decompose these datasets into train-test splits by following the Open-vocab data split strategy [37]. This approach randomly splits action categories into specific ratios, with corresponding videos associated with these splits. According to [37], we choose two ratios, namely, "75/25 split" and "50/50 split". The former involves 10 random splits, selecting 75% of action categories and corresponding videos as a training set and the rest as a test set. Similarly, the latter involves 5 random splits, selecting 50% for a training set and the rest as a test set. In addition, with the Acitivity-Net1.3 dataset, we employ the "Smart split" strategy described in [37], which leverages the hierarchy of action categories, with 25% of the labels allocated to the test set. The smart split is constructed by selecting pairs of neighboring leaf nodes from the hierarchy, designating one node for evaluation, and including the other in the training set. The selection process considers the perceptual similarities between classes to ensure an effective split.

3) Evaluation metrics: To evaluate the final output of the model, we utilize Soft Non-Maximum Suppression (Soft-NMS) [3] to eliminate duplicate detections, considering the potential overlap among the estimated action candidates. Subsequently, the results are evaluated based on the mean Average Precision (mAP), which measures the percentage of correctly estimated actions using threshold processing on the temporal Intersection over Union (tIoU) with ground truth.

4) *Comparison methods:* We compare the proposed method with the following methods:

- Baseline model: An 1-stage approach similar to the proposed model but utilizes a single scale within the multi-scale component to observe the effectiveness of the multi-scale component of the proposed method.
- OV-TAD [37]: The current state-of-the-art Open-vocab TAD setting, operating in a 2-stage TAD approach.
- STALE [30] and EffPrompt [15]: 1-stage and 2-stage Open-vocab TAD methods that incorporate the classification score, respectively, with the weakly-supervised action recognition model [42]. We also utilize these scores for comparison when comparing with these methods.

5) Implementation details: The proposed method was implemented according to the detailed description in Section III, utilizing six layers in the MVA module. For the video encoder, we utilize a two-stream Inflated 3D (I3D) ConvNet pre-trained on the Kinetics dataset [5] to extract video features. For the text encoder, we utilize a pre-trained CLIP model [36] to extract text features. Following Rathod et al. [37], we do not use a prompt in the main experiment. The text prompt is analyzed in Section IV-C.3. In the projection layers, we use fully connected layers with Gaussian Error Linear Unit (GELU) activation [11]. The balance coefficients in Eqns. (3), (6), and (7) are set as 1. We investigate the impact of the coefficients in Section IV-C.4. For optimization purposes, we utilize AdamW [26]. To attain optimal results, we carefully considered each dataset's model complexity and available training data, ensuring the appropriate selection of hyperparameters.

B. Experimental Results

Table I presents the performance of the proposed method with other comparison methods. We report the mAP at different tIoU thresholds, with the average calculated in the range of mAP@[0.30:0.10:0.70] for THUMOS14 and mAP@[0.50:0.05:0.95] for ActivityNet-1.3 datasets. The results are divided into two tables; Table I(a) shows the results in a completely Open-vocab setting, while Table I(b) shows the results with methods using fusion classification scores from a weakly-supervised action recognition model. In Table I(a), the results are split into two groups using pre-trained text features from CLIP base and CLIP large models.

1) THUMOS14: The proposed method consistently attained the highest results for both groups shown in Table I(a), surpassing other methods by a considerable margin. The results indicated a significant advantage of the proposed method, particularly at mAP@0.7, where it outperformed other methods nearly threefold. This demonstrated the substantial contribution of the multi-scale feature in achieving more accurate unseen action detection. In Table I(b), since the classes of actions in a THUMOS14 dataset video are mostly singular, the average mAP of the proposed method using fusion classification scores was more than twice that without using them. Furthermore, the proposed method's TABLE I

Results on THUMOS14 and ActivityNet-1.3 datasets. MAP (\uparrow) at different tIoU thresholds are reported. The average MAP in the range of [0.30:0.10:0.70] is reported for THUMOS14, while [0.50:0.05:0.95] is reported for ActivityNet-1.3. Best results within each group are highlighted in **BOLD**, while the overall best results are <u>underlined</u>.

	Image Feature				,	ГНИМО	S14 [12]]			Acti	ActivityNet-1.3 [9]	
Model		Text Feature	75/25			50/50			Smart	75/25	50/50		
			0.3	0.5	0.7	Avg.	0.3	0.5	0.7	Avg.	Avg.	Avg.	Avg.
OV-TAD [37]	CLIP B/16	CLIP B/16	21.8	13.2	3.7	12.9	18.0	8.9	2.2	9.5	23.4	21.4	19.5
	I3D	CLIP B/16	27.8	15.3	4.3	15.6	23.6	12.8	3.2	12.9	24.1	22.1	20.1
	CLIP B/32	CLIP B/32	21.4	11.8	3.5	12.0	15.4	7.7	1.9	8.0	22.6	19.4	17.3
	I3D	CLIP B/32	25.1	13.9	4.0	14.1	21.0	11.3	3.0	11.5	23.9	20.2	18.2
Baseline	I3D	CLIP B/16	24.3	16.6	5.6	15.8	15.5	10.7	3.8	10.2	21.7	16.7	14.5
Proposed	I3D	CLIP B/16	32.1	25.3	<u>13.6</u>	24.0	18.6	15.2	9.3	14.6	28.2	22.6	20.8
OV-TAD [37]	CLIP L/14	CLIP L/14	28.6	15.4	4.2	15.8	21.0	9.8	2.0	10.5	28.7	24.6	22.4
	I3D	CLIP L/14	30.1	16.8	4.7	17.0	<u>26.1</u>	14.3	3.6	14.5	28.1	24.8	<u>22.8</u>
Baseline	I3D	CLIP L/14	29.5	19.7	5.6	18.8	18.3	11.2	3.1	10.9	21.0	17.4	14.4
Proposed	I3D	CLIP L/14	<u>35.3</u>	<u>26.6</u>	<u>13.6</u>	<u>25.5</u>	18.7	<u>15.8</u>	<u>9.6</u>	<u>15.0</u>	<u>28.8</u>	<u>24.9</u>	21.1
(b) Results of methods that incorporate the classification score with the weakly-supervised action recognition model [42].													
EffPrompt [15]	I3D	CLIP B/16	39.7	23.0	7.5	23.3	37.2	21.6	7.2	21.9	_	23.1	19.6
STALE [30]	I3D	CLIP B/16	40.5	23.5	7.6	23.8	38.3	21.2	7.0	22.2	_	24.9	20.5

29.5

44.8

7.1

24.3

44.2

51.9

24.5

38.1

(a) Results in a completely Open-vocab setting using text features from pre-trained CLIP base and pre-trained CLIP large.

performance was nearly 1.5 to 2.0 times higher compared to the STALE [30] and EffPrompt [15] methods.

CLIP B/16

CLIP B/16

51.6

60.4

29.7

47.4

I3D

I3D

Baseline

Proposed

2) ActivityNet-1.3: In both groups in Table I(a), the proposed method consistently outperformed other methods in most of the split scenarios. Particularly in the Smart split, the proposed method demonstrated performance gains ranging from 4.1% to 6.5% in the pre-trained CLIP base group and from 0.1% to 7.8% in the pre-trained CLIP large group. Furthermore, the proposed method also demonstrated notable growth in the 75/25 split and 50/50 split. In contrast, the baseline model exhibited suboptimal performance on this dataset, which can be attributed to the dataset's large size and diverse nature, including a significant number of labels that were not sufficiently trained during the training process. These results underscore the significant contribution of the multi-scale feature in enabling the proposed method to overcome these challenges and achieve superior results. In Table I(b), the proposed method was also superior to other benchmarks when considering the fusion classification score.

C. Ablation Study

We conducted a series of excision experiments as part of an ablation study to evaluate the effectiveness of the proposed method on the THUMOS14 dataset using the "75/25 split" and "50/50 split" evaluation settings.

1) Feature extraction in the MVA module and fusion strategy in the VTA module: Table II shows the effects of modifying components within the MVA and VTA modules. Specifically, when we replaced Transformer layers (Trans Layer)

TABLE II

24.8

36.5

5.7

17.9

25.3

36.0

23.0

34.1

21.9

33.6

Effect of Changes of feature extraction in MVA module and fusion strategy in VTA module. Average mAP (\uparrow) in the range of [0.30:0.10:0.70] is reported.

	CLIP B/16		CLIP L/14	
	75/25	50/50	75/25	50/50
Proposed	24.0	14.6	25.5	15.0
Trans Layer \rightarrow Conv Layer	22.4	12.8	23.2	12.9
(w/o) Projection Feature (\mathbf{Z}_p^0)	21.7	12.3	23.0	14.4
(w/o) Multi-scale Feature (\mathbf{Z}_p^*)	22.5	12.2	24.7	14.0

with Convolutional layers (Conv Layer), we observed that the Trans Layer outperformed the Conv Layer, highlighting the beneficial impact of the attention mechanism. Additionally, we investigated altering the fusion strategy between text and video features. Utilizing either the individual projection feature (\mathbf{Z}_p^0) or the multi-scale feature (\mathbf{Z}_p^*) independently resulted in reduced performance compared to integrating both with the text feature. These results underscore the importance of harnessing the combined strength of temporal multi-scale features and action label features to enhance performance significantly.

2) Number of layers in the MVA module: Table III shows the performance impact of varying the number of layers in the MVA module. The results indicate that increasing the number of layers consistently improved performance, as evidenced by the increase in mAP. However, the model

TABLE III

Effect of Number of layers in the multi-scale feature of the MVA module. Average mAP (\uparrow) in the range of [0.30:0.10:0.70]and inference time are reported.

Number	CLIP	B/16	CLIP	L/14	Inference Time [s]	
of layers	75/25	50/50	75/25	50/50	$(B/16 \sim L/14)$	
1	17.5	10.5	18.3	10.0	$0.1038 \sim 0.1202$	
3	21.0	11.4	23.4	11.8	$0.1280 \sim 0.1428$	
5	23.5	13.6	24.8	15.1	$0.1774 \sim 0.1813$	
6	24.0	14.6	25.5	15.0	$0.1786 \sim 0.1878$	
7	24.3	14.2	26.8	14.8	$0.2013 \sim 0.2458$	

TABLE IV Impact of text prompts. Average mAP (\uparrow) in the range of [0.30:0.10:0.70] is reported.

Prompt	CLIP	B/16	CLIP L/14		
l lompt	75/25	50/50	75/25	50/50	
[CLASS]	24.0	14.6	25.5	15.0	
A video of action [CLASS]	24.3	12.8	26.8	13.5	
[CLASS] + [DESCRIPTION]	28.7	16.1	29.6	16.2	
[DESCRIPTION]	25.0	13.3	27.2	14.5	

tended to converge in terms of mAP at layers 5, 6, and 7, where the performances were relatively similar. Adding more layers beyond a certain point does not yield significant performance improvements. There was a slight trade-off with the inference time as the computational complexity increased with more layers. Specifically, seven layers required more than twice the computation of one layer. Therefore, selecting an optimal number of layers in the MVA module is crucial to balance performance and computational efficiency.

3) Text prompts: We evaluated the impact of text prompts on the Open-vocab TAD accuracy. The results, as shown in Table IV, indicate that a basic prompt such as "A video of action [CLASS]" yields negligible improvements. The THU-MOS14 dataset contains simplistic labels like Shot Put, High Jump, and Long Jump, which lack detailed contextual information. The experiments with [CLASS] + [DESCRIPTION] prompts, with descriptions generated from ChatGPT [32] like "High Jump: Athlete jumps over a horizontal bar", significantly enhanced the detection performance by approximately 4% in the 75/25 split and 1.5% in the 50/50 split. However, the results also underscore that merely adding descriptions without the action class is insufficient to substantially increase the accuracy of action detection, pointing to the necessity of a nuanced approach in employing text prompts for video analysis.

4) λ coefficient in the overall loss function: We experimented on the interplay between MVA Loss and VTA Loss as formulated in Eqn. 7. The outcomes of this analysis are visually represented in Fig. 3. The results demonstrate that varying the value of λ_3 from $0.2 \rightarrow 2.0$ only resulted in a deviation of no more than ± 0.5 compared to that when λ_3 was set to 1. This indicates a relatively stable performance



Fig. 3. Change of λ coefficient in the overall loss function.

of the loss function across a range of λ coefficient values.

D. Open-vocab TAD Results Visualization

We present the visualization of outcomes for several actions in Fig. 4. In Fig. 4(a), the results for the "Shot Put" action were identified almost accurately across varying action durations. However, the third action scene showed an error in action localization where the video frames predominantly displayed the athlete performing preparatory motions similar to the "Shot Put" action, leading to a wrong detection. Similarly, in Fig. 4(b), the "High Jump" action was also detected nearly accurately in the first two scenes, with varying lengths of the action. The third scene wrongly identified where it should have been "High Jump" was as "Long Jump". This error occured because most of the frames in this sequence showed the athlete running, which was a characteristic more associated with "Long Jump", leading to the misidentification of the action. These visualizations highlight the effectiveness and limitations of Open-vocab TAD in accurately localizing and identifying actions within a given video, underscoring the importance of nuanced differentiation between similar actions and the challenges posed by actions with similar preparatory movements.

E. Closed-vocab Temporal Action Detection Setting

1) Setting: In this section, we evaluate the Closed-vocab setting ($D_{test} \subset D_{train}$), which refers to the common context in which the model undergoes training and evaluation using the same action categories. It is important to note that the common context utilizes only videos and lacks the ability to detect unseen actions. In contrast, the Open-vocab setting employs both video and text, enabling evaluation of seen and unseen actions. To ensure a fair and consistent comparison, we use the same dataset splits as those utilized in previous studies and evaluate them on the THUMOS14 dataset.

2) Comparison methods: We considered the following methods for conducting comparisons with the proposed method. In the common context, some of the modern models for TAD in recent years have utilized the Inflated 3D (I3D) ConvNet and Temporal Segment Network (TSN) encoder backbone. In the Open-vocab setting, we conducted



(a) Illustrating action detection of the "Shot Put" event.

(b) Illustrating action detection of the "High Jump" event.

Fig. 4. Illustrating the results of Open-vocab TAD using CLIP B/16 as the text feature.

TABLE VComparison in a Closed-vocab setting using the THUMOS14dataset. Average MAP (\uparrow) in the range of [0.30:0.10:0.70] isREPORTED.

Model	Setting	Image Feature	Text Feature	mAP @Avg.	
BMN [22]		TSN	_	38.5	
TAL-MR [52]		I3D	_	43.3	
VSGN [51]		TSN	_	50.2	
AFSD [21]	Video	I3D	_	52.0	
TadTR [25]		I3D	I3D —		
ActionFormer [50]		I3D —		66.8	
TriDet [39]		I3D	_	69.3	
EffPrompt [15]		CLIP B/16	CLIP B/16	34.5	
STALE [30]		CLIP B/16	CLIP B/16	44.4	
STALE [30]		I3D	CLIP B/16	52.9	
OV-TAD [37]	Video	CLIP B/32	CLIP B/32	26.6	
OV-TAD [37]	Text	CLIP B/16	CLIP B/16	29.0	
OV-TAD [37]		CLIP L/14	CLIP L/14	32.6	
Baseline		I3D	CLIP B/16	39.9	
Proposed		I3D	CLIP B/16	59.5	

comparisons with methods using various pre-trained textimage models. The baseline model remains unchanged, as previously mentioned.

3) Results: The results presented in Table V show that the proposed method achieved the highest mAP@Avg of 59.5% in the video & text setting (Open-vocab methods). This performance is particularly notable compared to leading methods using only video settings (Closed-vocab methods) such as "ActionFormer" and "TriDet". The results of the

proposed method were closely competitive, and it outperformed approximately two-thirds of the existing methods. Meanwhile, the baseline model only achieved 39.9%, underscoring the significance of the 19.6% performance gap due to its lack of multi-scale component. These results highlight that the proposed method excels among the Open-vocab methods and performs well among the Closed-vocab methods.

V. CONCLUSION

We proposed a method for the Open-vocab TAD task that leveraged temporal multi-scale and action label features. The proposed 1-stage approach consists of a Multi-scale Video Analysis (MVA) module and a Video-Text Alignment (VTA) module. We also introduced a fusion strategy that combined temporal multi-scale features and action label features to improve the accuracy and robustness of action detection. A series of comprehensive experiments on THUMOS14 [12] and ActivityNet-1.3 [9] datasets indicated that the MVA module's multi-scale feature with the attention mechanism and the VTA module were instrumental in boosting performance. The number of layers in the MVA module significantly affected the experimental outcomes, necessitating a careful selection to balance performance with computational complexity. Furthermore, the use of text prompts within the VTA module impacted action identification results, highlighting the need for detailed analysis in Open-vocab setting.

In future work, we plan to investigate further methods for incorporating contextual information and higher-level scene understanding to enhance the performance of the action detection system. We also aim to address the accurate detection of actions' start and end times, as this significantly impacts the precise determination of those actions. Additionally, we aim to explore techniques for handling complex and dynamic scenes in cluttered or occluded environments.

ACKNOWLEDGMENT

This work was partly supported by JSPS KAKENHI JP21H03519 and JP24H00733. The computation was carried out using the General Projects on supercomputer "Flow" at Information Technology Center, Nagoya University.

REFERENCES

- [1] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan. Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.
- [2] W. Bao, Q. Yu, and Y. Kong. OpenTAL: Towards open set temporal action localization. In *Proceedings of the 2022 IEEE/CVF Conference* on Computer Vision and Pattern Recognition, pages 2979–2989, 2022.
- on Computer Vision and Pattern Recognition, pages 2979–2989, 2022.
 [3] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis. Soft-NMS— Improving object detection with one line of code. In Proceedings of the 16th IEEE International Conference on Computer Vision, pages 5561–5569, 2017.
- [4] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *Proceedings of the 16th European Conference on Computer Vision*, volume 1, pages 213–229, 2020.
- [5] J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299– 6308, 2017.
- [6] M. Chen, J. Gao, and C. Xu. Cascade evidential learning for openworld weakly-supervised temporal action localization. In *Proceedings* of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14741–14750, 2023.
 [7] W. Dai, L. Hou, L. Shang, X. Jiang, Q. Liu, and P. Fung. En-
- [7] W. Dai, L. Hou, L. Shang, X. Jiang, Q. Liu, and P. Fung. Enabling multimodal generation on CLIP via vision-language knowledge distillation. *Computing Research Repository arXiv Preprints*, arXiv:2203.06386, 2022.
- [8] V. S. Dorbala, G. Sigurdsson, R. Piramuthu, J. Thomason, and G. S. Sukhatme. CLIP-Nav: Using CLIP for zero-shot vision-andlanguage navigation. *Computing Research Repository arXiv Preprints*, arXiv:2211.16649, 2022.
- [9] C. H. Fabian, E. Victor, G. Bernard, and C. N. Juan. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.
 [10] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui. Open-vocabulary object
- [10] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui. Open-vocabulary object detection via vision and language knowledge distillation. *Computing Research Repository arXiv Preprints*, arXiv:2104.13921, 2021.
 [11] D. Hendrycks and K. Gimpel. Gaussian Error Linear Units (GELUs).
- [11] D. Hendrycks and K. Gimpel. Gaussian Error Linear Units (GELUs). Computing Research Repository arXiv Preprints, arXiv:1606.08415, 2016.
- [12] H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah. The THUMOS challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, 2017.
- [13] Y. Ji, R. Tu, J. Jiang, W. Kong, C. Cai, W. Zhao, H. Wang, Y. Yang, and W. Liu. Seeing what you miss: Vision-language pre-training with semantic completion learning. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6789– 6798, 2023.
- [14] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings* of the 2021 International Conference on Machine Learning, pages 4904–4916, 2021.
- [15] C. Ju, T. Han, K. Zheng, Y. Zhang, and W. Xie. Prompting visuallanguage models for efficient video understanding. In *Proceedings of the 17th European Conference on Computer Vision*, volume 35, pages 105–124, 2022.
- [16] P. H. Le-Khac, G. Healy, and A. F. Smeaton. Contrastive representation learning: A framework and review. *IEEE Access*, 8:193907– 193934, 2020.

- [17] C. Lea, M. D. Flynn, R. Vidal, A. Reiter, and G. D. Hager. Temporal convolutional networks for action segmentation and detection. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 156–165, 2017.
 [18] C. Li, M. Yan, H. Xu, F. Luo, W. Wang, B. Bi, and S. Huang. SemVLP:
- [18] C. Li, M. Yan, H. Xu, F. Luo, W. Wang, B. Bi, and S. Huang. SemVLP: Vision-language pre-training by aligning semantics at multiple levels. *Computing Research Repository arXiv Preprints*, arXiv:2103.07829, 2021.
- [19] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, C. Yejin, and G. Jianfeng. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the* 16th European Conference on Computer Vision, volume 30, pages 121–137, 2020.
- [20] X. Liang, L. Ma, S. Guo, J. Han, H. Xu, S. Ma, and X. Liang. MO-VLN: A multi-task benchmark for open-set zero-shot vision-andlanguage navigation. *Computing Research Repository arXiv Preprints*, arXiv:2306.10322, 2023.
- [21] C. Lin, C. Xu, D. Luo, Y. Wang, Y. Tai, C. Wang, J. Li, F. Huang, and Y. Fu. Learning salient boundary feature for anchor-free temporal action localization. In *Proceedings of the 2021 IEEE/CVF Conference* on Computer Vision and Pattern Recognition, pages 3320–3329, 2021.
- on Computer Vision and Pattern Recognition, pages 3320–3329, 2021.
 [22] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen. BMN: Boundary-matching network for temporal action proposal generation. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, pages 3889–3898, 2019.
 [23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for
- [23] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the 16th IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.
- [24] Q. Liu and Z. Wang. Progressive boundary refinement network for temporal action detection. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, number 07, pages 11612–11619, 2020.
 [25] X. Liu, Q. Wang, Y. Hu, X. Tang, S. Zhang, S. Bai, and X. Bai. End-
- [25] X. Liu, Q. Wang, Y. Hu, X. Tang, S. Zhang, S. Bai, and X. Bai. End-to-end temporal action detection with transformer. *IEEE Transactions on Image Processing*, pages 5427–5441, 2022.
 [26] I. Loshchilov and F. Hutter. Decoupled weight decay regularization.
- [26] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. Computing Research Repository arXiv Preprints, arXiv:1411.2539, 2019.
- [27] J. Lu, C. Clark, R. Zellers, R. Mottaghi, and A. Kembhavi. Unified-IO: A unified model for vision, language, and multi-modal tasks. *Computing Research Repository arXiv Preprints*, arXiv:2206.08916, 2022.
- [28] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022.
 [29] J. Mu, P. Liang, and N. Goodman. Shaping visual representations with
- [29] J. Mu, P. Liang, and N. Goodman. Shaping visual representations with language for few-shot classification. *Computing Research Repository* arXiv Preprints, arXiv:1911.02683, 2019.
- [30] S. Nag, X. Zhu, Y.-Z. Song, and T. Xiang. Zero-shot temporal action detection via vision-language prompting. In *Proceedings of the 17th European Conference on Computer Vision*, volume 3, pages 681–697, 2022.
- [31] I. Najdenkoska, X. Zhen, and M. Worring. Meta learning to bridge vision and language models for multimodal few-shot learning. *Computing Research Repository arXiv Preprints*, arXiv:2302.14794, 2023.
- [32] OpenAI. ChatGPT 3.5, 2022. URL: https://chat.openai.com/. Accessed: 2024-January.
- [33] H. Pham, Z. Dai, G. Ghiasi, K. Kawaguchi, H. Liu, A. W. Yu, J. Yu, Y.-T. Chen, M.-T. Luong, Y. Wu, M. Tan, and Q. V. Le. Combined scaling for zero-shot transfer learning. *Computing Research Repository* arXiv Preprints, arXiv:2111.10050, 2021.
- [34] R. Qian, Y. Li, Z. Xu, M.-H. Yang, S. Belongie, and Y. Cui. Multi-modal open-vocabulary video classification via pre-trained vision and language models. *Computing Research Repository arXiv Preprints*, arXiv:2207.07646, 2022.
 [35] Z. Qing, H. Su, W. Gan, D. Wang, W. Wu, X. Wang, Y. Qiao, J. Yan,
- [35] Z. Qing, H. Su, W. Gan, D. Wang, W. Wu, X. Wang, Y. Qiao, J. Yan, C. Gao, and N. Sang. Temporal context aggregation network for temporal action proposal refinement. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 485–494, 2021.
 [36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal,
- [36] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 2021 International Conference on Machine Learning*, pages 8748–8763, 2021.
- [37] V. Rathod, B. Seybold, S. Vijayanarasimhan, A. Myers, X. Gu, V. Birodkar, and D. A. Ross. Open-vocabulary temporal action detection with off-the-shelf image-text features. *Computing Research*

Repository arXiv Preprints, arXiv:2212.10596, 2022.

- [38] E. Salin, B. Farah, S. Ayache, and B. Favre. Are vision-language transformers learning multimodal representations? A probing perspective. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, number 10, pages 11248–11257, 2022.
- [39] D. Shi, Y. Zhong, Q. Cao, L. Ma, J. Li, and D. Tao. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings* of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18857–18866, 2023.
- [40] D. Shi, Y. Zhong, Q. Cao, J. Zhang, L. Ma, J. Li, and D. Tao. React: Temporal action detection with relational queries. In *Proceedings of the 17th European Conference on Computer Vision*, volume 10, pages 105–121, 2022.
- [41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:6000–6010, 2017.
- [42] L. Wang, Y. Xiong, D. Lin, and L. Van Gool. UntrimmedNets for weakly supervised action recognition and detection. In *Proceedings* of the 2017 IEEE conference on Computer Vision and Pattern Recognition, pages 4325–4334, 2017.
 [43] L. Wang, H. Yang, W. Wu, H. Yao, and H. Huang. Temporal
- [43] L. Wang, H. Yang, W. Wu, H. Yao, and H. Huang. Temporal action proposal generation with transformers. *Computing Research Repository arXiv Preprints*, arXiv:2105.12043, 2021.
- [44] M. Wang, J. Xing, and Y. Liu. ActionCLIP: A new paradigm for video action recognition. *Computing Research Repository arXiv Preprints*, arXiv:2109.08472, 2021.
- [45] Z. Weng, X. Yang, A. Li, Z. Wu, and Y.-G. Jiang. Transforming CLIP to an open-vocabulary video model via interpolated weight optimization. *Computing Research Repository arXiv Preprints*, arXiv:2302.00624, 2023.
- [46] M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. Ghanem. G-TAD: Sub-graph localization for temporal action detection. In *Proceedings* of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10156–10165, 2020.
- [47] Z. Xu, Y. Zhu, T. Deng, A. Mittal, Y. Chen, M. Wang, P. Favaro, J. Tighe, and D. Modolo. Challenges of zero-shot recognition with vision-language models: Granularity and correctness. *Computing Research Repository arXiv Preprints*, arXiv:2306.16048, 2023.
- [48] A. Zareian, K. D. Rosa, D. H. Hu, and S.-F. Chang. Openvocabulary object detection using captions. In *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021.
- [49] R. Zeng, W. Huang, M. Tan, Y. Rong, P. Zhao, J. Huang, and C. Gan. Graph convolutional networks for temporal action localization. In *Proceedings of the 17th IEEE/CVF International Conference on Computer Vision*, pages 7094–7103, 2019.
- [50] C.-L. Zhang, J. Wu, and Y. Li. ActionFormer: Localizing moments of actions with transformers. In *Proceedings of the 17th European Conference on Computer Vision*, volume 4, pages 492–510, 2022.
 [51] C. Zhao, A. K. Thabet, and B. Ghanem. Video self-stitching graph
- [51] C. Zhao, A. K. Thabet, and B. Ghanem. Video self-stitching graph network for temporal action localization. In *Proceedings of the* 18th IEEE/CVF International Conference on Computer Vision, pages 13658–13667, 2021.
- [52] P. Zhao, L. Xie, C. Ju, Y. Zhang, Y. Wang, and Q. Tian. Bottom-up temporal action localization with mutual regularization. In *Proceed*ings of the 16th European Conference on Computer Vision, volume 8, pages 539–555, 2020.
- [53] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren. Distance-IoU loss: Faster and better learning for bounding box regression. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, number 07, pages 12993–13000, 2020.