Uncertainty-Guided Appearance-Motion Association Network for Out-of-Distribution Action Detection

Xiang Fang, Arvind Easwaran, Blaise Genest

Uncertainty-Guided Appearance-Motion Association Network for Out-of-Distribution Action Detection

Xiang Fang
IGP-ERI@N, NTU, Singapore
SCSE, NTU, Singapore
CNRS@CREATE, Singapore
xiang003@e.ntu.edu.sg

Arvind Easwaran
SCSE, NTU, Singapore
arvinde@ntu.edu.sg

Blaise Genest
CNRS@CREATE, Singapore
blaise.genest@cnrsatcreate.sg

***Abstract*—Out-of-distribution (OOD) detection targets to detect and reject test samples with semantic shifts, to prevent models trained on in-distribution (ID) dataset from producing unreliable predictions. Existing works only extract the appearance features on image datasets, and cannot handle dynamic multimedia scenarios with much motion information. Therefore, we target a more realistic and challenging OOD detection task: OOD action detection (ODAD). Given an untrimmed video, ODAD first classifies the ID actions and recognizes the OOD actions, and then localizes ID and OOD actions. To this end, in this paper, we propose a novel Uncertainty-Guided Appearance-Motion Association Network (UAAN), which explores both appearance features and motion contexts to reason spatial-temporal inter-object interaction for ODAD. Firstly, we design separate appearance and motion branches to extract corresponding appearance-oriented and motion-aspect object representations. In each branch, we construct a spatial-temporal graph to reason appearance-guided and motion-driven inter-object interaction. Then, we design an appearance-motion attention module to fuse the appearance and motion features for final action detection. Experimental results on two challenging datasets show that UAAN beats state-of-the-art methods by a significant margin, illustrating its effectiveness.**

***Index Terms*—uncertainty-guided appearance-motion association, out-of-distribution action detection**

I. Introduction

Deep neural networks (DNNs) [1]–[10] have achieved impressive success on the multimedia sample detection task under a closed-set assumption [11]–[17], where all the classes experienced during the test have been seen during training [18]–[25]. However, vanilla DNN-based methods compulsorily classify each sample into some of the known classes [26]–[35]. It will lead to irrecoverable losses when classifying some outliers into wrong classes in some high-risk scenarios, such as autonomous driving [36]–[42]. Thus, the out-of-distribution (OOD) detection task [43]–[46] is proposed to accurately detect these outliers from OOD classes and correctly classify the samples from in-distribution (ID) classes during testing. Although most OOD detection methods [47]–[52] have achieved remarkable performance, they only focus on static images shown in Fig. 1, which limits their multimedia applications, where multimedia data is dynamically varying and not static. For example, in the autonomous driving scenario, we have to respond when an unknown object appears or moves on the road. These autonomous driving videos are dynamic, while previous OOD detection works focus on image datasets under the static assumption. It is unrealistic to first divide each video into multiple images and then detect unknown objects/actions [53]–[57]. This dynamic OOD detection problem based on videos is largely unexplored in the studies.

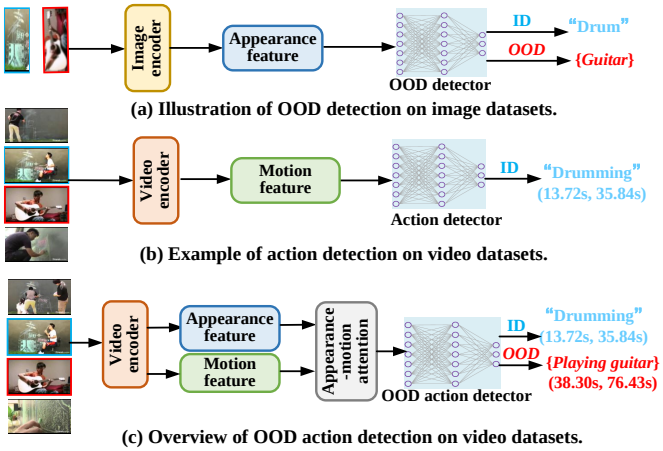

Fig. 1: (a) Illustration of out-of-distribution (OOD) detection that only detects static images. (b) Previous temporal action detection models only classify/localize the ID actions and cannot detect OOD actions. (c) Our target task: OOD action detection that can not only classify and localize ID actions, but also detect and localize OOD actions. For the boxes and text, we color ID actions as blue and OOD actions as red, and background is not colored. The output labels on OOD actions are only for illustration, and they are not the actual output of the model. Best viewed in color.

To handle dynamic OOD detection, we consider a challenging task: out-of-distribution action detection (ODAD) shown in Fig. 1. Given an untrimmed video, ODAD aims to classify ID actions, recognize OOD actions, and localize both ID and OOD actions. Although many action detection methods have been proposed to classify ID actions in an untrimmed video, few methods can correctly recognize OOD actions. An intuitive idea is to *combine the OOD detection task and the action detection task to solve the more challenging ODAD task*. We observe that there are some gaps between the two

This research is part of the programme DesCartes and is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

topics: 1) data gap, the OOD detection task focuses on images, while the action detection task targets videos; 2) feature gap, the OOD detection task extracts appearance features, while the action detection mainly extracts motion features; 3) output gap, the OOD detection task does not conduct localization, while the action detection task needs to localize the target action.

Therefore, we need to close the above gaps for the ODAD task shown in Fig. 1. An emerging issue is how to effectively integrate action and appearance knowledge for ODAD. To this end, we propose a novel Uncertainty-guided Appearance-motion Association Network (UAAN), which cleverly incorporates motion contexts into appearance-based object features for better detecting the actions among objects. Specifically, we first utilize the Faster R-CNN network [58] to extract the appearance-aware object features, and obtain the motion-aware object features by employing RoIAlign [59] on the I3D network [60], which fully mines the object visual information. Then, we design two separate branches with the same architecture to reason the appearance-guided and motion-guided object relations by graph convolutional networks (GCNs), respectively, so as to reason the inter-object interaction. Besides, we represent frame-level features by aggregating object features inside the frame. Finally, we design an appearance-motion association attention module to integrate appearance-aware object features and motion-aware object features for final detection.

In summary, our contributions are as follows: 1) We explore both appearance- and motion-aware object information for OOD action detection. Besides, we creatively propose a novel uncertainty-guided appearance-motion association network to reason inter-object interaction. 2) With the appearance and motion branch, we capture action-oriented and appearance-guided object relations via GCNs. By designing an appearance-motion association attention module, we integrate appearance and motion features from two branches for obtaining inter-object interaction. 3) Extensive experimental results on two challenging datasets show that our proposed model outperforms existing state-of-the-art approaches by a significant margin.

II. Related Works

Temporal action detection. Temporal action detection (TAD) focuses on localizing action instances and classifying their categories in untrimmed videos. Although these TAD methods have achieved decent success, they regard all the actions in this video as belonging to some of the pre-defined classes. Besides, they only rely on the motion feature, which captures the redundant background information and fails to perceive the fine-grained differences among video frames with high similarity. For example, two real-world actions "HighJump" and "LongJump" belong to different action classes. As shown in Fig. 2(a), the main difference between "HighJump" and "LongJump" is the appearance information: "high jump bar" in "HighJump" and "sandpit" in "LongJump". Since they only model the motion features to capture the same motion "jump", they cannot correctly distinguish the local details of different objects ("high jump bar" and "sandpit") in these frames,

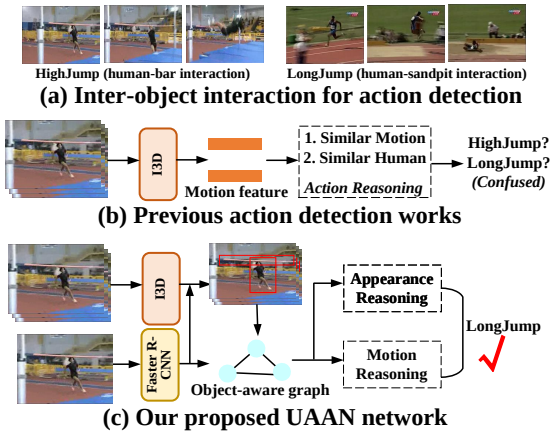

Fig. 2: Our motivation. (a) Example of inter-object interaction for action detection. (b) Existing action detection works only extracts frame-level motion information, and fails to distinguish similar motions "HighJump" and "LongJump". (c) We construct an object-aware graph to reason the inter-object interaction from the appearance and motion perspectives.

which will lead to wrong action detection results. Although some TAD methods utilize the object detection technologies for action detection, these methods focus on single-object detection setting so that they only detect one object "human" and ignore other relevant objects. Thus, these methods cannot achieve satisfactory performance, where supplementary file illustrates two qualitative results. Different from them, we creatively explore both appearance and motion features among different objects (*e.g.,* "high jump bar" and "sandpit") for action detection. By integrating appearance and motion features, we can correctly distinguish "HighJump" and "LongJump" via inter-object interaction.

Out-of-distribution detection. As a challenging multimedia task, out-of-distribution (OOD) detection targets to detect test samples from distributions that do not overlap with the training distribution. Different from previous OOD detection works that cannot handle video datasets, our proposed method can detect OOD actions in the challenging video dataset. Therefore, our method can be applied to more challenging scenarios than these image-based OOD detection methods.

More details of related works are in supplementary file.

III. Our Proposed UAAN

Given an untrimmed video, the ODAD task aims to first localize all actions with temporal boundary (t_s, t_e), then recognize OOD actions, and finally classify all ID actions. Given a training set $\{V_i, y_i, t_s^i, t_e^i\}_{i=1}^N$, where N denotes the video number, V_i denotes the i-th untrimmed video, $y_i \in \mathbb{R}^K$ is its multi-shot label denoting the action class that the action samples in V_i belong to, and K denotes the number of ID action classes. During training, the video data and the ID action labels are provided, while we cannot obtain the OOD action labels. During testing, for any untrimmed video V^t, we first distinguish if OOD actions are in V^t, then classify each ID action in V^t to one of the K classes, and finally localize the boundaries of all the actions.

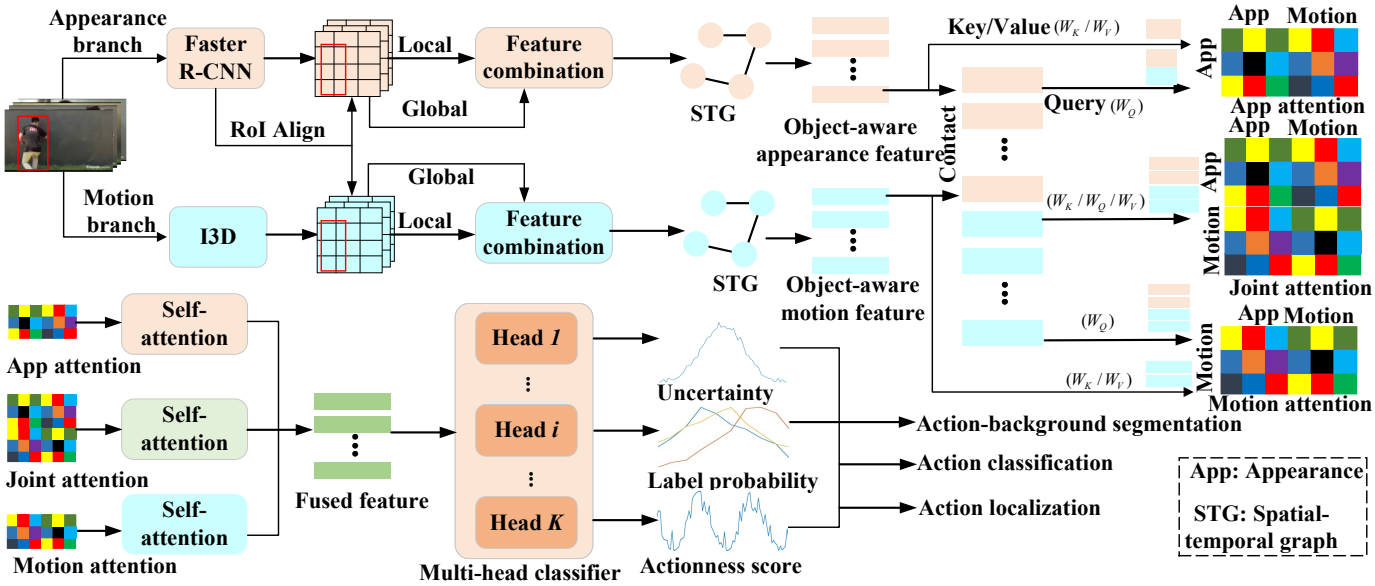

Fig. 3: Overview of the proposed model for the challenging ODAD task. We first utilize video encoder (Faster R-CNN and I3D) to extract appearance- and motion-aware object features. Then, we design separate appearance and motion branches to reason the spatial-temporal interaction between different objects. Besides, we design an appearance-motion attention module to fully integrate the appearance and motion features for final detection. Best viewed in color.

A. Video Representation

Unlike previous action detection methods that only utilize the I3D network to extract motion-aware features, we additionally consider extracting appearance-aware object features to explore inter-object interaction.

Appearance representation. For a video with T frames, we utilize Faster R-CNN applying RoIAlign to extract the region of interest from a ResNet backbone [61] and obtain S objects. Thus, we can obtain $S \times T$ objects in each video. These object-aware appearance features are represented as $V^a_{local} = \{o^a_{t,k}, b_{t,k}\}^{t=T,k=S}_{t=1,k=1}$, where $o^a_{t,k} \in \mathbb{R}^d$ and $b_{t,k} \in \mathbb{R}^4$ indicate the local appearance feature and bounding box location of k-th object in the t-th frame, respectively. However, only the object-aware appearance features are not enough to fully understand the entire video. For example, we cannot distinguish "HighJump" and "LongJump" only based on a jumper. Thus, by another ResNet model with a linear layer, we utilize another ResNet network to extract the frame-wise appearance representation $V^a_{global} \in \mathbb{R}^{T\times d}$.

Motion representation. We extract the features of video clips by the last convolutional layer in an I3D network. For local motion features, we also utilize RoIAlign to extract object-aware motion features: $V^m_{local} = \{o^m_{t,k}, b_{t,k}\}^{t=T,k=S}_{t=1,k=1}$, where $o^m_{t,k} \in \mathbb{R}^d$ and $b_{t,k} \in \mathbb{R}^4$. For global motion features, we extract the clip-aware features $V^m_{global} \in \mathbb{R}^{T\times d}$ by utilizing average pooling in the feature map and linear projection on the local motion features.

Position representation. In the challenging ODAD task, we consider both spatial and temporal positions of each object for reasoning about object-wise action relations. Therefore, a position encoding is added to object-level local features in both appearance and motion representations: $v^a_{t,k} = FFN([o^a_{t,k}; e^b; e^t])$ and $v^m_{t,k} = FFN([o^m_{t,k}; e^b; e^t])$, where e^t is obtained by position encoding according to each frame's index, $e^b = FFN(b_{t,k})$, and $FFN(\cdot)$ is the feed-forward network. After the above process, we denote $\overline{V}^a_{local} = \{v^a_{t,k}\}^{t=T,k=S}_{t=1,k=1}$ and $\overline{V}^m_{local} = \{v^m_{t,k}\}^{t=T,k=S}_{t=1,k=1}$.

Similarly, we add the position encoding e^t into global appearance and motion features: $\overline{V}^a_{global} = FFN([V^a_{global}; e^t])$ and $\overline{V}^m_{global} = FFN([V^m_{global}; e^t])$. Thus, we expand the above two global features from size $T \times d$ to $(T \times S) \times d$. To explore the context in objects, we concatenate object features with global features: $F^a = FFN([\overline{V}^a_{local}; \overline{V}^a_{global}])$, $F^m = FFN([\overline{V}^m_{local}; \overline{V}^m_{global}])$, where $F^a = \{f^a_{t,k}\}^{t=T,k=S}_{t=1,k=1} \in \mathbb{R}^{(T\times S)\times d}$ are the final encoded object-level appearance features, and $F^m = \{f^m_{t,k}\}^{t=T,k=S}_{t=1,k=1} \in \mathbb{R}^{(T\times S)\times d}$ are the final encoded object-level motion features.

B. Spatial-temporal Object Graph for Inter-object interaction

We observe that utilizing inter-object interaction can distinguish motion-similar actions (*e.g.*, "HighJump" and "LongJump"). As shown in Fig. 2, "HighJump" is the interaction between "human" and "high jump bar", while "LongJump" is the interaction between "human" and "sandpit". If we only use the common appearance information "human" and similar motion information "jump", we cannot correctly distinguish "HighJump" and "LongJump".

To explore the interaction between different objects for ODAD, we develop two separate branches based on appearance- and motion-aware object features to fully understand the video. In each branch, we reason about object relations by constructing a spatial-temporal graph.

Object graph construction. Since the detected objects have both spatial interactivity and temporal continuity, we capture spatial-temporal relations by constructing object graphs in each branch, respectively. Since the two branches share the

same architecture, we only describe the appearance branch's components as an example to avoid redundancy.

In the appearance branch, we define the object-wise appearance features $F^a = \{f^a_{t,k}\}^{t=T,k=S}_{t=1,k=1}$ including all objects in the whole video as nodes and build a fully-connected graph. Then, we adopt the graph convolution network (GCN) to extract the relation-aware object features via message propagation. In the object graph, we can obtain the adjacency matrix $A^a \in \mathbb{R}^{(T\times S)\times(T\times S)}$ to indicate the pairwise affinity between object features: $A^a = \sigma((F^aW^a_1)(F^aW^a_2)^\top)$, where W^a_1 and W^a_2 are learnable parameters, and $\sigma(\cdot)$ is the softmax function.

If two objects have strong semantic relationships, they will be highly correlated in the video, and they will have an edge with a high-affinity score in A^a. For deeper semantic reasoning, we introduce two-layer graph convolution with residual connection: $\hat{F}^a = LayerNorm(F^a + \tau(A^a\tau(A^aF^aW_3)W_4))$, where $\tau(\cdot)$ denotes the Relu function, W_3 and W_4 are weight matrices of the GCN layers, and $\hat{F}^a \in \mathbb{R}^{(T\times S)\times d}$ is the appearance-aware object features. Similarly, we can obtain the motion-aware object features $\hat{F}^m \in \mathbb{R}^{(T\times S)\times d}$.

Appearance and motion features association. After obtaining the object-wise features ($\hat{F}^a$ and $\hat{F}^m$), we integrate these features to model inter-object interaction. Thus, we first concatenate $\hat{F}^a$ and $\hat{F}^m$: $U = \begin{bmatrix}\hat{F}^a \\ \hat{F}^m\end{bmatrix}$, $U \in \mathbb{R}^{2(T\times S)\times d}$. To effectively integrate different types of visual features, we first introduce the dot-product attention: $Att(W_Q, W_K, W_V) = \sigma(\frac{W_QW_K^\top}{\sqrt{d}})W_V$, where W_Q, W_K and W_V are the query, key, and value matrices in the dot-product attention, respectively.

To enhance different features for inferring action contexts, we introduce the attention function $Att(\cdot)$: $X^a = LayerNorm(U + Att(U, \hat{F}^a, \hat{F}^a))$, $X^m = LayerNorm(U + Att(U, \hat{F}^m, \hat{F}^m))$, $X^{joint} = LayerNorm(U + Att(U, U, U))$, where $X \in \mathbb{R}^{(T\times S)\times d}$ is the enhanced features. Then, we integrate these enhanced features by:

$$X = Ave(X^a, X^m, X^{joint}), \tag{1}$$

where $Ave(\cdot)$ denotes the average pooling operation with $L2-$normalization, and $X = \{x_t\}^T_{t=1}$ is the fused feature.

C. Uncertainty-guided OOD Action Detection

Action-background segmentation. Real-world multimedia videos naturally mix background and actions (ID and OOD actions), which makes it insufficient to distinguish them only through classification and uncertainty. Therefore, we design an action detection module to indicate the likelihood of an frame belonging to an action. For convenience, we treat the action-background segmentation as a binary classification task.

To distinguish class-agnostic actions from backgrounds, we follow [62] to introduce the frame-aware actionness score: $\hat{a}_i \in [0, 1]$. For any two frames x_{t_1} and x_{t_2} ($t_1 \neq t_2$ and $x_{t_1}, x_{t_2} \in \{1, \cdots, T\}$), they have three relationships: 1) both x_{t_1} and x_{t_2} belong to the action; 2) both x_{t_1} and x_{t_2} are from the background; 3) one belongs to the action, while the other is from the background. Therefore, we introduce the cosine similarity $\cos(x_{t_1}, x_{t_2})$ and three thresholds ($\tau_{bb}, \tau_{aa}, \tau_{dif}$) to evaluate their relationship. Based on the annotated background frames and potential action frames, we can calculate the affinity loss $\mathcal{L}_{ABS}$ from three terms, *i.e.,* between two background frames $\mathcal{L}^{\text{bg}}$, between two action frames $\mathcal{L}^{\text{act}}$ and between the action-background pair $\mathcal{L}^{\text{dif}}$.

Particularly, we use the online hard example mining strategy [63] to constrain the training frame pair. We align two background frames into the same class by the following loss:

$$\mathcal{L}^{\text{bg}} = \sum\nolimits_{a_{t_1}\leq a_\tau}\sum\nolimits_{a_{t_2}\leq a_\tau} \max[\tau_{bb} - \cos(x_{t_1}, x_{t_2}), 0], \tag{2}$$

where a_τ is an actionness threshold; τ_{bb} is the similarity threshold between two frames from the background class. Also, we utilize a similar loss to align two action frames:

$$\mathcal{L}^{\text{act}} = \sum\nolimits_{a_{t_1}> a_\tau}\sum\nolimits_{a_{t_2}> a_\tau} \max[\tau_{aa} - \cos(x_{t_1}, x_{t_2}), 0], \tag{3}$$

where τ_{aa} is the similarity threshold between two frames from the action class. When x_{t_1} and x_{t_2} belong to action and background respectively, we can obtain the following loss:

$$\mathcal{L}^{\text{dif}} = \sum\nolimits_{a_{t_1}\leq a_\tau}\sum\nolimits_{a_{t_2}> a_\tau} [\cos(x_{t_1}, x_{t_2}) - \tau_{dif}, 0], \tag{4}$$

where τ_{dif} is the threshold to constrain the similarity between actions and backgrounds. The affinity loss jointly considers the above three terms and can be calculated as:

$$\mathcal{L}_{ABS} = \frac{1}{3}(\mathcal{L}^{\text{bg}} + \mathcal{L}^{\text{act}} + \mathcal{L}^{\text{diff}}). \tag{5}$$

Action classification. Different from previous OOD detection methods, we are required to detect OOD actions in an untrimmed video. In real-world multimedia scenario, the predicted likelihood for each action class follows a binomial distribution whose conjugate prior is a Beta distribution. We introduce evidential neural networks [64] based on Beta distribution to jointly formulate the multi-class classification and uncertainty modeling. We assume a Beta distribution $Beta(p_i|\alpha_i, \beta_i)$ over the action categorical probability $p_i \in [0, 1]$, where α_i and β_i are parameters indicating the positive and negative evidence of the i-th action, respectively. The evidence evaluates actions closest to the predicted ones in the feature space and is used to support the decision-making. Predicted labels should be the same as positive evidence, but different from negative evidence.

Based on α_i and β_i, we can design a subjective opinion $w_i = (b_i, d_i, u_i, a_i)$ for action $i \in \{1, 2, \cdots, K\}$, where belief $b_i \in [0, 1]$, disbelief $d_i \in [0, 1]$, uncertainty score $u_i \in [0, 1]$, and $b_i + d_i + u_i = 1$. For action i, the opinion w_i is obtained by α_i and β_i: $b_i = \frac{\alpha_i - 2}{\alpha_i+\beta_i}$, $d_i = \frac{\beta_i-2}{\alpha_i+\beta_i}$, $u_i = \frac{2}{\alpha_i+\beta_i}$. The expected belief probability $p_i = b_i + a_i \cdot u_i$, where a_i refers to a base rate representing prior knowledge without commitment, such as neither agree nor disagree.

For an object j, its positive and negative evidence are estimated as $\boldsymbol{\alpha}_j = s(h(\mathbf{x}_j; \boldsymbol{\theta})) + 1$ and $\boldsymbol{\beta}_j = s(h(\mathbf{x}_j; \boldsymbol{\theta})) + 1$, where $\boldsymbol{\alpha}_j = [\alpha_{1j}, \cdots, \alpha_{Kj}]^\top$ and $\boldsymbol{\beta}_j = [\beta_{1j}, \cdots, \beta_{Kj}]^\top$. $\mathbf{x}_j$ denotes the input video, $h(\mathbf{x}_j; \boldsymbol{\theta})$ represents the evidence vector predicted by the network for the classification and, $\boldsymbol{\theta}$ represents parameters for modeling. Here $s(\cdot)$ is the evidence function (*e.g.,* ReLU) to keep $\boldsymbol{\alpha}_j, \boldsymbol{\beta}_j \geq \mathbf{1}$.

To learn the above opinions, we define the Beta loss function by computing its Bayes risk for the action predictor. For the

binary cross-entropy loss for each action i over a batch of actors, we introduce the following loss:

$$\begin{aligned}\mathcal{L}_{Beta} =& \frac{1}{K}\sum\nolimits_{i=1}^{K}\sum\nolimits_{j=1}^{S}\int BC(y_{ij},p_{ij})Beta(p_{ij};\alpha_{ij},\beta_{ij})dp_{ij}\\ =& \frac{1}{K}\sum\nolimits_{i=1}^{K}\Big[y_{ij}\Big(\psi(\alpha_{ij}+\beta_{ij})-\psi(\alpha_{ij})\Big)\\ &+(1-y_{ij})\Big(\psi(\alpha_{ij}+\beta_{ij})-\psi(\beta_{ij})\Big)\Big],\end{aligned} \tag{6}$$

where K is the number of actions, and $BC(\cdot)$ is the binary cross-entropy loss, and $\psi(\cdot)$ is the *digamma* function. The log expectation of Beta distribution derives the last equality. $\mathbf{y}_j = [y_{1j},\cdots,y_{Kj}] \in \{0,1\}^K$ is the K-dimensional ground-truth action label for $\mathbf{x}_j$.

Action localization. As shown in Fig. 3, we can predict each action proposal $l_i = (e_s^i, e_e^i)$ and a refined stage to predict the temporal offset $\delta_i = (\hat{e}_s, \hat{e}_e)$ by our model. Since the boundaries of pre-defined proposals are coarse, we employ a boundary regression loss for calibrating the boundary. We calculate a regression loss for every positive sample:

$$\mathcal{L}_{reg} = \frac{1}{\mathcal{N}_p}\sum\nolimits_i \mathcal{L}_1(e_s^i,\hat{e}_s^i) + \mathcal{L}_1(e_e^i,\hat{e}_e^i), \tag{7}$$

where $\mathcal{L}_1$ denotes the smooth l_1 loss, $\mathcal{N}_p$ is the number of positive samples, e_s and e_e represent the truth error of proposal boundary, and $\hat{e}_s$ and $\hat{e}_e$ are the predicted errors.

To handle more complex scenarios, we additionally introduce the following Distance Intersection-over-Union (DIoU) loss [65]:

$$\mathcal{L}_{DIoU} = \frac{1}{\mathcal{N}_p}\sum\nolimits_{i=1}^{K}(1-DIoU(l_i,\delta_i)). \tag{8}$$

Therefore, we can obtain the final localization loss:

$$\mathcal{L}_{Local} = \mathcal{L}_{reg} + \gamma_0\mathcal{L}_{DIoU}, \tag{9}$$

where γ_0 is a parameter to balance two losses.

Training and inference. We train the proposed model by minimizing the following loss:

$$\mathcal{L}_{final} = \gamma_1\mathcal{L}_{ABS} + \gamma_2\mathcal{L}_{Beta} + \gamma_3\mathcal{L}_{Local}, \tag{10}$$

where γ_1, γ_2 and γ_3 are hyper-parameters to balance the importance between different losses.

During inference, we feed an untrimmed video into our proposed model, which generates proposals comprising of a classification label c_i, an uncertainty score u_i, an actionness score a_i and a predicted action location $l_i = (d_i^s, d_i^e)$. Therefore, by predefining an uncertainty threshold u_τ and an actionness threshold a_τ, we can directly determine whether the i-th action is an OOD action by the following process:

$$P(x_i) = \begin{cases}\text{OOD action}, & \text{if } u_i > u_\tau \;\&\; a_i > a_\tau,\\ \text{ID action}(\hat{y}_i), & \text{if } u_i \le u_\tau \;\&\; a_i > a_\tau,\\ \text{Background}, & \text{otherwise.}\end{cases} \tag{11}$$

IV. Experiments

Datasets: For a fair comparison with existing works, we utilize two challenging and popular video datasets: THUMOS14 [70] and ActivityNet1.3 [71]. THUMOS14 contains 200 validation videos and 212 testing videos from 20 labeled classes. ActivityNet1.3 has 19,994 videos with 200 action classes.

Evaluation metrics: To comprehensively evaluate the model performance, we follow [68] to utilize the following evaluation metrics: mAP, AUROC, AUPR, OSDR, FAR@95, where a smaller FAR@95 value means better performance, while larger values for the other metrics denote better performance.

Implementation details are available in supplementary file.

A. Comparison with State-of-the-Art

Compared methods. Following [68], we only compare the following state-of-the-art *open-source* methods for reproducibility: SoftMax, OpenMax [66], DEAR [67], OpenTAL [68] and TFE-DCN [69]. Since TFE-DCN cannot be directly used for the ODAD task, we embed TFE-DCN into OpenTAL by replacing the action detection module in OpenTAL. To distinguish it from the original model, we denote the embed model as TFE-DCN(+). Different from them, we explore the inter-object interaction for better understanding video actions.

Performance comparison. Tables I and II report performance comparison, where all the methods are tested using both the THUMOS14 unknown splits and the ActivityNet1.3 disjoint subset. For fair evaluation, we report all performance comparisons by averaging the results of each evaluation metric over the three THUMOS14 splits. The best results are in **bold**. Obviously, our UAAN surpasses all the compared methods by significant margins on all the metrics. Especially, our UAAN outperforms the best compared method TFE-DCN(+) over all the metrics. Specially, on THUMOS14 as the OOD set, UAAN achieves 4.55% absolute improvement over "OSDR" with "tIoU=0.6". For ActivityNet1.3 as the OOD set, our UAAN outperforms these compared methods by 3.90% over "AUPR" with "tIoU=0.75". The main reason is that most videos in both THUMOS14 and ActivityNet1.3 contain multiple significant objects, which includes inter-object interaction. However, previous methods ignore the interaction. Our UAAN can construct the spatial-temporal graph to model objects in videos, and then fully grasp the appearance and motion information to mine inter-object interaction. Another reason is that our utilized Beta-distribution assumption is more suitable for multi-object videos than the Dirichlet-distribution assumption.

ID action classification and ODAD. We further compare the performance of both ID action classification and OOD action detection in Fig. 4. Obviously, our UAAN significantly outperforms all compared methods over two evaluation metrics (AUROC and Accuracy), showing the effectiveness of our UAAN for both ID action classification and ODAD. **More results and analysis are in supplementary file.**

B. Ablation Study on THUMOS14

Main ablation study. We perform exhaustive ablation studies to analyze the effectiveness of each individual component. Table III shows the main ablation results on THUMOS14. Specifically, we remove each module to evaluate its significance. The result shows that our full model has precision improvement compared with each ablation model, which manifests that each above component provides a positive contribution. Compared with the first ablation model, our full model improves performance by 7.04% in terms of "FAR@95". It is because our

TABLE I: **Performance comparison (%) vs. Different tIoU Thresholds**. Models trained on the THUMOS14 closed set are tested by including the OOD action classes from THUMOS14 and ActivityNet1.3, respectively. Results are averaged over the two dataset splits.

Metrics	Methods	THUMOS14 as the OOD set						ActivityNet1.3 as the OOD set			
		tIoU=0.3	tIoU=0.4	tIoU=0.5	tIoU=0.6	tIoU=0.7	Mean	tIoU=0.5	tIoU=0.75	tIoU=0.95	Mean
AUROC(↑)	SoftMax	54.70	55.46	56.41	57.12	57.11	56.16	56.97	58.41	55.97	57.77
	OpenMax [66]	53.26	52.10	52.13	51.89	52.53	52.38	51.24	52.39	49.13	51.59
	DEAR [67]	64.05	64.27	65.13	66.21	66.81	65.29	62.82	66.23	67.92	65.69
	OpenTAL [68]	78.33	79.04	79.30	79.40	79.82	79.18	82.97	83.21	83.38	83.22
	TFE-DCN(+) [69]	79.12	79.56	79.69	80.34	80.41	79.85	79.82	83.73	84.02	83.13
	Our UAAN	**81.25**	**81.47**	**81.72**	**82.13**	**82.68**	**81.85**	**84.86**	**85.73**	**85.94**	**85.51**
AUPR(↑)	SoftMax	31.85	31.81	31.11	29.78	27.99	30.51	53.54	44.15	34.54	44.77
	OpenMax [66]	33.17	31.61	30.59	29.15	28.45	30.60	54.88	48.37	40.07	48.48
	DEAR [67]	40.05	39.45	38.05	37.58	36.35	38.30	53.97	47.22	45.59	48.46
	OpenTAL [68]	58.62	59.40	58.78	57.54	55.88	58.04	80.41	74.20	73.92	75.54
	TFE-DCN(+) [69]	58.86	60.72	58.84	57.35	58.22	58.80	80.74	76.28	74.22	77.08
	Our UAAN	**61.34**	**61.28**	**61.71**	**60.56**	**60.39**	**61.06**	**81.90**	**80.18**	**77.82**	**79.97**
OSDR(↑)	SoftMax	23.40	25.19	27.43	29.97	32.08	27.61	27.63	33.73	31.59	32.01
	OpenMax [66]	13.66	14.58	15.91	17.71	20.41	16.45	15.73	21.49	18.07	19.35
	DEAR [67]	36.26	37.58	39.16	41.18	42.99	39.43	38.56	43.72	42.20	42.18
	OpenTAL [68]	42.91	46.19	49.50	52.50	56.78	49.57	50.49	59.87	62.17	57.89
	TFE-DCN(+) [69]	43.65	46.30	48.72	52.87	57.02	50.11	50.83	60.24	63.71	58.26
	Our UAAN	**45.50**	**48.83**	**53.61**	**57.42**	**60.38**	**53.14**	**52.47**	**61.34**	**65.93**	**59.91**

TABLE II: ODAD performance (%), where we set the tIoU threshold as 0.3 on THUMOS14 and as 0.5 on the ActivityNet1.3. Models trained on THUMOS14 closed set are tested on the OOD sets by including the OOD classes from THUMOS14 and ActivityNet1.3, respectively. mAP is provided as the reference of the localization results on THUMOS14 ID set.

Methods	FAR@95(↓) on different OOD sets		mAP(↑)
	THUMOS14	ActivityNet1.3	
OpenMax [66]	90.34	91.36	36.36
SoftMax	85.58	85.05	55.81
DEAR [67]	81.42	84.01	52.24
OpenTAL [68]	70.96	63.11	55.02
TFE-DCN(+) [69]	68.80	62.54	53.27
Our UAAN	**66.43**	**58.35**	**57.63**

TABLE III: Main ablation study on THUMOS14, where we remove each key individual component to investigate its effectiveness. "OSG" is "object-aware spatial-temporal graph", "MAFA" is "motion and appearance features association", and "UOAD" is "uncertainty-guided OOD action detection".

OSG	MAFA	UOAD	FAR@95(↓)	AUROC(↑)	AUPR(↑)	OSDR(↑)
✗	✓	✓	73.47	75.88	57.12	39.80
✓	✗	✓	71.75	74.08	55.93	40.02
✓	✓	✗	72.82	74.43	56.20	41.38
✓	✓	✓	**66.43**	**81.25**	**61.34**	**45.50**

object-aware spatial-temporal graph can fully mine the inter-object interaction for action reasoning. For the second ablation model, our full model beats it by 7.17% over "AUROC". The main reason is that we can integrate both appearance and motion visual information of each object for action detection. Besides, our full model outperforms the third ablation model by 6.39%. The significant improvement is because we can segment action and background by $\mathcal{L}_{ABS}$ and leverage inter-object interaction for action classification by $\mathcal{L}_{Beta}$.

Hyper-parameters analysis. We conduct the ablation studies on the hyper-parameters $a_\tau, u_\tau, \gamma_0, \gamma_1, \gamma_2, \gamma_3$ in Fig. 5. For each ablation study, we change one hyper-parameter by fixing the others. We can obtain the best performance when $a_\tau = 0.5, u_\tau = 0.6, \gamma_0 = 0.8, \gamma_1 = 0.15, \gamma_2 = 0.3, \gamma_3 = 0.2$.

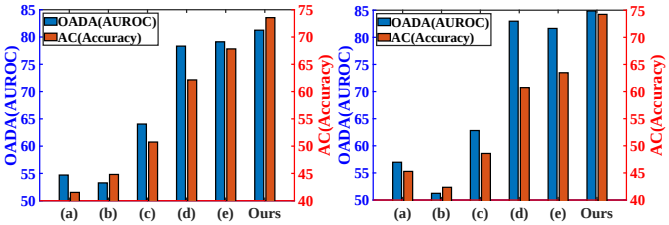

Fig. 4: Performance comparison for ID action classification and ODAD on THUMOS14 (left) and ActivityNet1.3 (right), where "ODAD(AUROC)" means "AUROC for ODAD" and "AC(Accuracy)" means "Classification accuracy for ID action classification". (a) is SoftMax, (b) is OpenMax [66], (c) is DEAR [67], (d) is OpenTAL [68], and (e) is TFE-DCN(+) [69]. Best viewed in color.

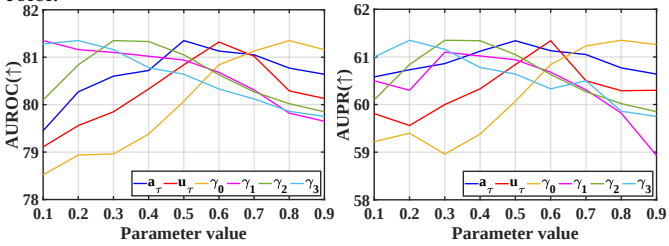

Fig. 5: Analysis on hyper-parameters. Best viewed in color.

V. Conclusion

In this paper, we target a challenging task: ODAD, which is an extension of OOD detection and action detection. To this end, we propose the novel UAAN method to explores both appearance features and motion contexts to reason spatial-temporal inter-object interaction. Experimental results confirm the effectiveness of our proposed UAAN.

References

[1] C. Sun, S. Jia, S. Hou, and S. Lyu, "Ai-synthesized voice detection using neural vocoder artifacts," in *CVPR*, 2023.

[2] C. Tian, L. Fei, W. Zheng, Y. Xu, W. Zuo, and C.-W. Lin, "Deep learning on image denoising: An overview," *Neural Networks*, 2020.

[3] C. Chen and A. Ross, "An explainable attention-guided iris presentation attack detector," in *CVPR*, 2021.

[4] F. Clement, A. Kaur, M. Sedghi, D. Krishnaswamy, and K. Punithakumar, "Interactive data driven visualization for covid-19 with trends, analytics and forecasting," in *IV*, 2020.

[5] H. Cheng, J. T. Zhou, W. P. Tay, and B. Wen, "Attentive graph neural networks for few-shot learning," in *MIPR*, 2022.
[6] H. Roy, S. Chaudhury, T. Yamasaki, and T. Hashimoto, "Image inpainting using frequency-domain priors," *JEI*, 2021.
[7] H. Liu, B. Jou, T. Chen, M. Topkara, N. Pappas, M. Redi, and S.-F. Chang, "Complura: Exploring and leveraging a large-scale multilingual visual sentiment ontology," in *ICMR*, 2016.
[8] I. Ide, H. Mo, N. Katayama, and S. Satoh, "Topic threading for structuring a large-scale news video archive," in *CIVR*, 2004.
[9] J. Li, R. Xu, X. Liu, J. Ma, Z. Chi, J. Ma, and H. Yu, "Learning for vehicle-to-vehicle cooperative perception under lossy communication," *TIV*, 2023.
[10] J. P. Robinson, C. Qin, Y. Henon, S. Timoner, and Y. Fu, "Balancing biases and preserving privacy on balanced faces in the wild," *TIP*, 2023.
[11] D. Huo, M. A. Kastner, T. Komamizu, and I. Ide, "Action semantic alignment for image captioning," in *MIPR*, 2022.
[12] A. Long, W. Yin, T. Ajanthan, V. Nguyen, P. Purkait, R. Garg, A. Blair, C. Shen, and A. van den Hengel, "Retrieval augmented classification for long-tail visual recognition," in *CVPR*, 2022.
[13] B. Niu, P. Martin, and W. Powley, "Towards autonomic workload management in dbmss," *JDM*, 2009.
[14] H. Baba and C. Shimizu, "The impact of apartment vacancies on nearby housing rents over multiple time periods: application of smart meter data," *IJHMA*, 2023.
[15] J. P. Robinson, M. Shao, H. Zhao, Y. Wu, T. Gillis, and Y. Fu, "Rfiw: Large-scale kinship recognition challenge," in *MM*, 2017.
[16] K. Koneripalli, S. Lohit, R. Anirudh, and P. Turaga, "Rate-invariant autoencoding of time-series," in *ICASSP*, 2020.
[17] K. Nakamura, N. Nitta, and N. Babaguchi, "Encryption-free framework of privacy-preserving image recognition for photo-based information services," *TIFS*, 2018.
[18] K. Yanai and Y. Kawano, "Food image recognition using deep convolutional network with pre-training and fine-tuning," in *ICMEW*, 2015.
[19] L.-W. Kang, C.-W. Lin, and Y.-H. Fu, "Automatic single-image-based rain streaks removal via image decomposition," *TIP*, 2011.
[20] K. Tsukuda, M. Hamasaki, and M. Goto, "Why and how people view lyrics while listening to music on a smartphone," *TIS*, 2023.
[21] L. Li, R. E. Sengpiel, D. L. Pascual, L. P. Tedesco, J. S. Wilson, and E. Soyeux, "Using hyperspectral remote sensing to estimate chlorophyll-a and phycocyanin in a mesotrophic reservoir," *IJRS*, 2010.
[22] M. A. Kastner, I. Ide, F. Nack, Y. Kawanishi, T. Hirayama, D. Deguchi, and H. Murase, "Estimating the imageability of words by mining visual characteristics from crawled image data," *MTA*, 2020.
[23] M. P. Queluz, "Content-based integrity protection of digital images," in *SWMC*, 1999.
[24] M. Ruan, X. Yu, N. Zhang, C. Hu, S. Wang, and X. Li, "Video-based contrastive learning on decision trees: From action recognition to autism diagnosis," in *ACM Multimedia Systems*, 2023.
[25] Y. Xie, R. Jiang, X. Guo, Y. Wang, J. Cheng, and Y. Chen, "mmeat: Millimeter wave-enabled environment-invariant eating behavior monitoring," *Smart Health*, 2022.
[26] M. Huang, X. Li, J. Hu, H. Peng, and S. Lyu, "Tracking multiple deformable objects in egocentric videos," in *CVPR*, 2023.
[27] N. Nitta and N. Babaguchi, "Automatic story segmentation of closed-caption text for semantic content analysis of broadcasted sports video." in *Multimedia information systems*, 2002.
[28] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia syst*, 2010.
[29] R. S. Aygün and A. Yazici, "Modeling and management of fuzzy information in multimedia database applications," *MTA*, 2004.
[30] R. Wang, L. Li, P. Wang, X. Tao, and P. Liu, "Feature-aware unsupervised learning with joint variational attention and automatic clustering," in *ICPR*, 2021.
[31] Y. Luo, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation," in *CVPR*, 2019.
[32] Y.-L. Wu, C.-Y. Tang, Y.-M. Yeh, and W.-C. Hung, "Using hilbert scan on statistical color space partitioning," *COMPUT ELECTR ENG*, 2013.
[33] Z. Si and H. Qi, "An exploration on temperature term in training deep neural networks," in *AVSS*. IEEE, 2019, pp. 1–7.
[34] Z. Ouyang, Y. Feng, Z. He, T. Hao, T. Dai, and S.-T. Xia, "Attentiondrop for convolutional neural networks," in *ICME*, 2019.
[35] J. Xu, Z. Bi, A. Singha, T. Li, Y. Chen, and Y. Zhang, "mmlock: User leaving detection against data theft via high-quality mmwave radar imaging," in *ICCCN*, 2023.
[36] O. Zendel, M. Schörghuber, B. Rainer, M. Murschitz, and C. Beleznai, "Unifying panoptic segmentation for autonomous driving," in *CVPR*, 2022.
[37] J. Li, R. Xu, J. Ma, Q. Zou, J. Ma, and H. Yu, "Domain adaptive object detection for autonomous driving under foggy weather," in *WACV*, 2023.
[38] H. Kurihata, T. Takahashi, I. Ide, Y. Mekada, H. Murase, Y. Tamatsu, and T. Miyahara, "Rainy weather recognition from in-vehicle camera images for driver assistance," in *IV*, 2005.
[39] M.-H. Wang, S.-Y. Wu, L.-H. Yen, and C.-C. Tseng, "Pathmon: Path-specific traffic monitoring in openflow-enabled networks," in *ICUFN*, 2016.
[40] V. Patil, P. Singh, S. Parikh, and P. K. Atrey, "Geosclean: Secure cleaning of gps trajectory data using anomaly detection," in *MIPR*, 2018.
[41] W. Li, G. Wang, L. Fidon, S. Ourselin, M. J. Cardoso, and T. Vercauteren, "On the compactness, efficiency, and representation of 3d convolutional networks: brain parcellation as a pretext task," in *IPMI*, 2017.
[42] Y. Ju, S. Jia, L. Ke, H. Xue, K. Nagano, and S. Lyu, "Fusing global and local features for generalized ai-synthesized image detection," in *ICIP*, 2022.
[43] B. Olber, K. Radlak, A. Popowicz, M. Szczepankiewicz, and K. Chachuła, "Detection of out-of-distribution samples using binary neuron activation patterns," in *CVPR*, 2023.
[44] Y.-C. Hsu, Y. Shen, H. Jin, and Z. Kira, "Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data," in *CVPR*, 2020.
[45] E. Zisselman and A. Tamar, "Deep residual flow for out of distribution detection," in *CVPR*, 2020.
[46] H. Wang, Z. Li, L. Feng, and W. Zhang, "Vim: Out-of-distribution with virtual-logit matching," in *CVPR*, 2022.
[47] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *ICLR*, 2018.
[48] W. Liu, X. Wang, J. Owens, and Y. Li, "Energy-based out-of-distribution detection," *NeurIPS*, vol. 33, pp. 21 464–21 475, 2020.
[49] Y. Sun, C. Guo, and Y. Li, "React: Out-of-distribution detection with rectified activations," *NeurIPS*, vol. 34, pp. 144–157, 2021.
[50] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," *NeurIPS*, 2018.
[51] S. Mohseni, M. Pitale, J. Yadawa, and Z. Wang, "Self-supervised learning for generalizable out-of-distribution detection," in *AAAI*, 2020.
[52] Y. Ming, Y. Sun, O. Dia, and Y. Li, "How to exploit hyperspherical embeddings for out-of-distribution detection?" in *ICLR*, 2023.
[53] R. Garg, A. Roussos, and L. Agapito, "A variational approach to video registration with subspace constraints," *IJCV*, 2013.
[54] R. Iqbal, S. Shirmohammadi, A. El Saddik, and J. Zhao, "Compressed-domain video processing for adaptation, encryption, and authentication," *IEEE MultiMedia*, 2008.
[55] R. Yan, J. Tang, X. Shu, Z. Li, and Q. Tian, "Participation-contributed temporal dynamic model for group activity recognition," in *MM*, 2018.
[56] T. Zhang, S. Ye, K. Zhang, J. Tang, W. Wen, M. Fardad, and Y. Wang, "A systematic dnn weight pruning framework using alternating direction method of multipliers," in *ECCV*, 2018.
[57] Y. Tonomura, A. Akutsu, Y. Taniguchi, and G. Suzuki, "Structured video computing," *IEEE multimedia*, 1994.
[58] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *NeurIPS*, 2015.
[59] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *ICCV*, 2017.
[60] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *CVPR*, 2017.
[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.
[62] 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 *CVPR*, 2021.
[63] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *CVPR*, 2016.
[64] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *NeurIPS*, 2018.
[65] 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 *AAAI*, 2020.
[66] A. Bendale and T. E. Boult, "Towards open set deep networks," in *CVPR*, 2016.
[67] W. Bao, Q. Yu, and Y. Kong, "Evidential deep learning for open set action recognition," in *ICCV*, 2021.
[68] W. Bao, Q. Yu, and Y. Kong, "Opental: Towards open set temporal action localization," in *CVPR*, 2022.
[69] J. Zhou and Y. Wu, "Temporal feature enhancement dilated convolution network for weakly-supervised temporal action localization," in *CVPR*, 2023.
[70] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes," http://crcv.ucf.edu/THUMOS14/, 2014.
[71] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *CVPR*, 2015.