Forcing the Whole Video as Background: An Adversarial Learning Strategy for Weakly Temporal Action Localization

Ziqiang Li
Chongqing University
ziqiangli@cqu.edu.cn

Jiaruo Yu
Chongqing University
jiaruoyu@cqu.edu.cn

Yongxin Ge*
Chongqing University
yongxin.ge@cqu.edu.cn

Zhongming Chen
Chongqing University
zhongmingcszmx@cqu.edu.cn

ABSTRACT

With video-level labels, weakly supervised temporal action localization (WTAL) applies a localization-by-classification paradigm to detect and classify the action in untrimmed videos. Due to the characteristic of classification, class-specific background snippets are inevitably mis-activated to improve the discriminability of the classifier in WTAL. To alleviate the disturbance of background, existing methods try to enlarge the discrepancy between action and background through modeling background snippets with pseudo-snippet-level annotations, which largely rely on artificial hypotheticals. Distinct from the previous works, we present an adversarial learning strategy to break the limitation of mining pseudo background snippets. Concretely, the background classification loss forces the whole video to be regarded as the background by a background gradient reinforcement strategy, confusing the recognition model. Reversely, the foreground(action) loss guides the model to focus on action snippets under such conditions. As a result, competition between the two classification losses drives the model to boost its ability for action modeling. Simultaneously, a novel temporal enhancement network is designed to facilitate the model to construct temporal relation of affinity snippets based on the proposed strategy, for further improving the performance of action localization. Finally, extensive experiments conducted on THUMOS14 and ActivityNet1.2 demonstrate the effectiveness of the proposed method.

CCS CONCEPTS
• Computing methodologies → Activity recognition and understanding.

KEYWORDS
temporal action localization, weakly supervised learning, adversarial learning, gradient enhancement, temporal enhancement

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM ’22, October 10–14, 2022, Lisbon, Portugal.
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9203-7/22/10...
https://doi.org/10.1145/3503161.3548300

1 INTRODUCTION

Temporal action localization, which aims to localize and classify actions in untrimmed videos, has attracted great interest in the computer vision community due to its potential application in various fields [25, 34, 36]. In this task, fully supervised methods, owning snippet-level labels, have triggered remarkable progress. However, with the rapid growth of videos on various websites, annotating the precise action instance boundary in untrimmed videos is expensive and time-consuming. Consequently, weakly supervised temporal action localization (WTAL) has drawn a lot of attention recently, where cheaper video-level labels are introduced. Different from the fully supervised counterpart, most existing WTAL methods adopt a "localization by classification" paradigm [42]. However, without snippet-wise labels, the classifier is prone to focus on salient features to boost its discriminability for classification [9]. Therefore, the background snippets with class-specific content in untrimmed videos would be misrecognized as foreground(action) by the classifier in WTAL methods, hindering the separation of background and actions [31]. Simultaneously, the capacity for action modeling of the classifier is weakened due to the preference for class-specific background snippets.

To alleviate the disturbance of the background, some classical approaches are proposed. Lee et al. [15] assume that background snippets could be dynamic and inconsistent, and model the background uncertainty by suppressing the feature magnitudes of pseudo background snippets with bottom-k scores. To further distinguish class-specific background (context) and action, ACSNet [21] introduces auxiliary categories for representing the class-specific background. Furthermore, CoLA [42] utilizes the fruitful temporal context relation to mine pseudo hard background and easy action snippets and conducts contrastive learning on them for better action modeling. All of the above works attend to explicitly model background snippets to further boost the discrepancy between background and actions, leading to precise localization. Nevertheless, these approaches largely rely on artificial hypotheticals for class-specific (hard) background snippets annotating, and this priority inevitably marks some snippets of action into pseudo annotations.
Let’s summarize the core challenges on WTAL: 1) preference for class-specific background blocks the robust capacity from action modeling, and 2) fully filtering out background snippets is impractical due to the absence of snippet-level labels. As mentioned above, the previous methods attempt to model background content by pseudo label annotating at the snippet level, as shown in Fig. 1(a). However, the representation of background occurs in the whole untrimmed videos. To illustrate it, we take the long jump in Fig. 1 as an example. Track-field and sandpit, as background content, appear both in the background and action snippets along the temporal dimension. Naturally, each snippet in untrimmed videos owns the tendency to be rated as background.

Inspired by the above analysis, distinct from the previous works that model background by pseudo-class-specific background annotating, we design a strategy to explicitly model action instances under the circumstance that the whole video is regarded as background. Concretely, motivated by GAN [6], we adopt an adversarial learning strategy based on the background and foreground classification losses. The background loss confuses the model by forcing the whole untrimmed video to tend to be background, while the foreground loss drives the model to focus on the action snippets. And competition between the two losses boosts the ability for action modeling to distinguish the foreground from the video-level background, reducing the mis-recognition caused by the class-specific background. Simultaneously, to complement the adversarial learning strategy, we also explicitly construct the temporal relation of affinity snippets, which consists of the base branch (BB), and the temporal continuity branch (TCB). Functionally, BB is the baseline of our method and TCB is an auxiliary one for it. Simply, the inputs of BB and TCB are the sequence pairs of affinity snippets. With the interaction of BB and TCB, the temporal continuity of the neighboring snippets is obtained by the model, which promotes the performance of the completeness modeling. Our full method is the combination of BGES and TEN. Extensive experiments conducted on THUMOS14 and ActivityNet demonstrate the effectiveness of the proposed method.

To sum up, the main contributions of our works are as follows: (1) Pioneeringly, we design the adversarial learning paradigm for WTAL to drive model focus on action under the encompassment of background. (2) A simple yet efficient strategy is introduced to guide the whole video to tend to be the background in the gradient backward process. (3) A novel temporal enhancement network (TEN) is proposed to explicitly perceive temporal relationships, aiming to model action completeness.

2 RELATED WORK

Fully Supervised Temporal Action Localization. The goal of temporal action localization (TAL) is to predict the precise start time and end time of action instances in an untrimmed video and recognize the corresponding action category. The fully supervised paradigm of TAL, relying on snippet-level annotations, has drawn lots of interest recently [40, 43], and it can be divided into two types. The methods [3, 16] of the first type generally adopt a two-stage pipeline that includes proposal generation and instance classification. These two-stage methods mainly aim to enhance the localization performance, such as improving temporal proposals searching [17, 18] and proposing robust classifiers [44]. Compared with the former type, the second type of works [22, 39] is more flexible and efficient, which applies a one-stage mode to generate proposals and classify categories simultaneously. Although the above methods make remarkable results, they are highly laborious-costing for the manual annotating timestamp of action instances.

Weakly Supervised Temporal Action Localization. Distinct from the fully supervised TAL, the weakly supervised methods [19, 30] mostly adopt video-level labels in training procedures, reducing the labor-intensive annotating cost. STPN [29] identifies a set of key segments with an attention mechanism and aggregates them by adaptive temporal pooling. To strengthen the confidence of foreground scores, Yang et al. [37] present an uncertainty-guided collaborative training strategy to mitigate the noise of pseudo action snippets and FAC-Net [10] takes the bilateral relations of foreground and action into consideration and guarantees the consistency of foreground and action. Moreover, some methods [20, 26] are proposed to explicitly model action instances. 3C-Net [28] introduces extensive information by counting the number of action instances, and TSCN [41] utilizes the consistency of RGB and optical flow.
modality to focus on action. Furthermore, Luo et al. [23] take into account the complexity of action and propose an Action Unit Memory Network. Additionally, some attempts [14, 31] are present to mitigate the background interference. ACSNet [21] introduces an extensive category to explicitly model background, and CoLA [42] proposes a snippet contrastive loss to refine hard background samples. Lee et al. [15] assume that background snippets could be dynamic and inconsistent and suppress the feature magnitudes of pseudo background snippets with bottom-k scores.

Adversarial Learning. Following the early work GAN [6], the idea of adversarial learning is widely applied in various fields [13, 33]. For example, Isola et al. [13] propose conditional adversarial networks (CGAN) as a general-purpose solution for image-to-image translation. Compared with CGAN, Circle-GAN [45] is proposed for translation learning in the absence of paired datasets. In addition, ADDA [33] adopts a novel unified framework for adversarial domain adaptation based on adversarial learning. Furthermore, Ganin et al. [4] present a gradient reversal layer (GRL) to minimize the discrepancy between feature distributions from the target and source domains. In our work, we force the whole video to be regarded as background in the gradient backpropagation procedure, driving the recognition model to focus on action semantic features by adversarial learning.

3 THE PROPOSED METHOD

Problem Definition. Assume that we have $N$ untrimmed videos $\{v_i\}_{i=1}^N$ with their corresponding video-level labels $\{\tilde{y}_i\}_{i=1}^N$, where $\tilde{y}_i \in \mathbb{R}_0^C$ and $\tilde{y}_i.e$ means the presence/absence of c-th action classes. Following recent methods [23, 41], for each video $v_i$, we divide it into non-overlapping snippets and feed them into a pre-trained EBD [2] to generate features as the input of our method. The features $X^r \in \mathbb{R}^{T \times D}$ and $X^o \in \mathbb{R}^{T \times D}$ are extracted from the RGB and the optical flow modality snippets, where $T$ and $D$ indicate the number of snippets and dimension of features. During inference, we attempt to generate a set of action instances $\{c_i, q_i, s_i, e_i\}$, where $c_i$ and $q_i$ are the action class and confidence score, $s_i$ and $e_i$ represent the start time and end time of action instances correspondingly.

Overview. We propose a novel temporal enhancement network (TEN) based on a background gradient enhancement strategy (BGES), which achieves the adversarial learning process to improve the model’s ability for action localization. BGES is applied in the backward procedure of background loss. In this manner, the background classification loss confuses the recognizing model by forcing the whole video to tend to be the background. On the contrary, the foreground classification loss guides the model to obtain action snippets under this condition. Consequently, competition between the two losses drives the model to self-boost the capacity for action modeling. Additionally, TEN is designed to explicitly enhance the temporal relation of affinity snippets for further improving the performance of the model in the adversarial learning procedure. For clarification, we first describe TEN and then introduce the BGES.

3.1 Temporal enhancement network

To model action completeness, Hide-and-seek [32] adopts a random hidden strategy on snippet sequences, and Liu et al. [19] propose a multi-branch network where each branch is utilized to focus on different action content of one snippet. However, all these works ignore the temporal properties of action instances. Consequently, the temporal enhancement network (TEN) is designed to promote the model to construct temporal relation of snippet sequences, which further boosts the capacity of the model for action modeling, leading to the localization completeness of action. As illustrated in Fig. 2, TEN is composed of two branches, namely, the base branch (BB) and the temporal continuity branch (TCB), and the two branches are built on the same parameter-shared model.

Base branch. BB is the baseline of our method, and here we introduce it to describe the model which consists of three parts, namely the feature embedding module, classifier, and temporal attention mechanism.

Feature embedding. Due to the features $X^r \in \mathbb{R}^{T \times D}$ and $X^o \in \mathbb{R}^{T \times D}$ are respectively extracted from the RGB and the optical-flow modality snippets, the distributions along the channel dimension of them are inconsistent. To bridge the modality domain gap, we leverage the independent temporal convolutional layers $\Phi_r$ and $\Phi_o$ to embed $X^r$ and $X^o$, for learning a new set of features that can be applied in our specific task. Simply, this process can be defined as:

$$
X^r_e = \Phi_r(X^r, w^r_o), \quad X^o_e = \Phi_o(X^o, w^o_r),
$$

(1)

where $X^r_e \in \mathbb{R}^{T \times E}$ and $X^o_e \in \mathbb{R}^{T \times E}$ are the embedding features from the RGB and the optical-flow modalities, $w^r_o$ and $w^o_r$ are learnable parameters, and $E$ represents the corresponding dimension of the feature. And $\Phi_r$ and $\Phi_o$ are implemented with single temporal convolutional layer following a ReLU function.

Classifier (CLS). As a localization-by-classification paradigm, each snippet in the untrimmed video is predicted a class score for applying thresholding to localize action instances. Therefore, given the embedding features $X^r_e$ and $X^o_e$, we utilize the classifiers to generate snippet-level class activation scores (CAS) $y_{r/o} \in \mathbb{R}^{T \times (C+1)}$ and $y_{r/o} \in \mathbb{R}^{T \times (C+1)}$ under the RGB and the optical-flow modalities. This procedure can be formulated as follows:

$$
y_{r/o} = X^r_e w^r_{cls} + b^r_{cls}, \quad y_{r/o} = X^o_e w^o_{cls} + b^o_{cls},
$$

(2)

where $w^r_{cls} \in \mathbb{R}^{E \times (C+1)}$ and $w^o_{cls} \in \mathbb{R}^{E \times (C+1)}$ are the weights of the classifiers, $b^r_{cls}$ and $b^o_{cls}$ are the biases, and the $C + 1$ represents the background class. Given the CASs of the RGB and the optical-flow modalities, we fuse them to get the final CAS $y$ of untrimmed videos, i.e. $y = (y_r + y_o) \ast 0.5$.

Temporal attention mechanism (TAM). Following the previous attempts [12, 31], we also introduce TAM to recognize action snippets. And TAM is considered as a foreground classifier to produce thresholding to localize action instances. Similar to the classifiers, the embedded features $X^r_e$ and $X^o_e$ are leveraged to calculate class-agnostic temporal attentions $a_r \in \mathbb{R}^T$ and $a_o \in \mathbb{R}^T$ of the two modalities:

$$
a_r = \sigma(X^r_e w^r_{att} + b^r_{att}), \quad a_o = \sigma(X^o_e w^o_{att} + b^o_{att}),
$$

(3)

where $w^r_{att} \in \mathbb{R}^E$ and $w^o_{att} \in \mathbb{R}^E$ are the learnable parameters of temporal attention mechanisms, $b^r_{att}$ and $b^o_{att}$ are the corresponding biases, and $\sigma (\cdot)$ represents the sigmoid activation function. And the final class-agnostic attentions $a$ is calculated by aggregating $a_r$ and $a_o$, i.e. $a = (a_r + a_o) \ast 0.5$.

Intuitively, the class-agnostic attention and action-class scores should be consistent, i.e. given the $i$-th snippet that belongs to the
c-th class of action, the values of \( a_i \) and \( y_{c,i} \) are similar. Naturally, we fuse the CAS and class-agnostic attention \( a \) to get the video-level action class probability \( P_f \in \mathbb{R}^{C+1} \) and the corresponding background probability \( P_b \in \mathbb{R}^{C+1} \):

\[
P_f = \tau(\sum_{i=1}^{T} y_i \cdot a_i), \quad P_b = \tau(\sum_{i=1}^{T} y_i \cdot (1 - a_i)),
\]

(4)

where \( \tau \) represents a SoftMax function, \( N_f = \sum_{i=1}^{T} a_i \) and \( N_b = \sum_{i=1}^{T} (1 - a_i) \) are the normalized factors. Notably, \( a_i \) is the foreground score of the i-th snippet and \( 1 - a_i \) represents the background one.

Finally, given the video-level probabilities of the action and the background categories, the cross-entropy losses are conducted to optimize the model:

\[
\mathcal{L}_f = -\sum_{c=1}^{C+1} \hat{y}_c \cdot \log(P_c^f), \quad \mathcal{L}_b = -\log(P_{b}^{C+1}),
\]

(5)

where \( c \) represents the action class, and \( \hat{y} \in \mathbb{R}^{C+1} \) is the ground-truth label. \( \mathcal{L}_f \) is the foreground classification loss to guide the model to focus on action snippets, and \( \mathcal{L}_b \) is the background classification one that makes the model distinguish the background from the action snippets.

Temporal continuity branch. Due to the temporal continuity of the snippet sequence, affinity snippets tend to be consistent in content and category. Therefore, TCB is introduced as an auxiliary unit to interact with BB, aiming to constrain the temporal consistency of the contiguous snippets. Since TCB and BB are built on the same model, we simplify the description of TCB and adopt \( X \) to denote both RGB and optical flow features.

As shown in Fig. 2, given the extracted feature sequence \( X = [X_1, X_2, \cdots, X_T] \in \mathbb{R}^{T \times D} \) of a video, we divide it into \( \frac{T}{2} \) segments by equal-interval and randomly select one snippet from each segment to get a sequence \( X = [X_1, \cdots, X_{T-K}] \in \mathbb{R}^{T \times D} \). Then, we uniformly repeat the sampled snippet in each divided segment until to recover the original length, and obtain a new sequence \( X^R \in \mathbb{R}^{T \times D} \). Similar to the sampling manner proposed in TSN [35], this way still makes \( X^R \) retain the temporal characteristic of the original sequence \( X \), and \( X \) and \( X^R \) are conducted as the sequence pair of affinity snippets. Thereafter, \( X^R \) is fed into the model to generate the corresponding action scores \( y^R \in \mathbb{R}^{T \times (C+1)} \) and temporal foreground attentions \( a^R \in \mathbb{R}^{T} \). Here, \((a^R, a^R)\) and \((y^R, y^R)\) are the score pairs of affinity snippets, and we conduct constraint losses on them to guide the model to learn the consistent information of the continuous snippets. For the temporal attention of BB and TCB, the \( L_1 \) normalized function is utilized to optimize them:

\[
\mathcal{L}_{att} = \frac{1}{T} \sum_{i=1}^{T} \left| a_i - G(a_i^R) \right| + \left| a_i^R - G(a_i) \right|, \quad (6)
\]

where \( G(\cdot) \) represents the gaussian function among temporal dimensions, and it is utilized to smooth attention scores. Motivated by the success in knowledge distillation [7], a KL diverse loss is applied on the class activate scores to transfer information of affinity snippets:

\[
\mathcal{L}_{KL} = \frac{1}{T} \sum_{i=1}^{T} (KL(y_i^R || y_i) + KL(y_i || y_i^R)), \quad (7)
\]

where \( KL(p||q) \) is the KL divergence of \( p \) from \( q \). Consequently, by the mutlual learning procedure, the above constraint losses facilitate the model to learn the consistency of continuous snippets, improving the model for modeling action continuity.

Joint training. The action and the background classification losses are solely conducted on BB, and we combine all the losses above to optimize our model:

\[
\mathcal{L}_{all} = \mathcal{L}_f + \lambda \cdot \mathcal{L}_b + \beta \cdot (\mathcal{L}_{KL} + \mathcal{L}_{att}), \quad (8)
\]

where \( \lambda \) and \( \beta \) represent the hyper-parameters to adjust the balance among the losses.
3.2 The background gradient enhancement strategy

The class-specific background tends to co-occur with the action snippets [31], hindering the removals of background snippets from localized action instances. Motivated by GAN, we aim to construct an adversarial learning procedure that the background classification loss $L_{bg}$ guides the model to regard the whole video as background in the backward procedure while the foreground classification losses $L_{fg}$ drive it to focus on action, as illustrated in Fig. 3. As a result, competition between $L_{bg}$ and $L_{fg}$ prevents the model from mis-activating class-specific background snippets that own similar static content to action snippets and promotes its ability for action modeling, simultaneously. Here, we propose a background gradient enhancement strategy applied in $L_{bg}$ to achieve the adversarial learning procedure. Similar to GRL [4], the designed strategy aims to modify the gradient in the backward procedure. To clarify our strategy, we first formulate the generalized gradient updating process:

$$w_{new} = w_{old} - \eta \cdot \frac{\partial L}{\partial w_{old}},$$  

(9)

where $w$ represents a learnable parameter, $\eta$ is the updating rate, and $L$ means optimization loss function. As mentioned in section 3.1, both the classifier and TAM are utilized to predict the action probability of each snippet. Naturally, the proposed background gradient enhancement strategy can be implemented from two perspectives, i.e. 1) modification of the weights of the classifier, and 2) revision of the parameters of TAM.

Modification of the classifier’s weights. Intuitively, we can adopt an additional background classification loss that optimizes the model by setting each snippet as background, enhancing the background gradient of the classifier. And the loss is formulated as:

$$L^u_{bg} = - \sum_{c=1}^{C+1} y^c_{bg} \cdot \log(\frac{1}{T} \sum_{i=1}^{T} y_i^c),$$  

(10)

where $y^c_{bg}$ is one-hot label of the background. With the background loss $L^u_{bg}$, the background gradient of classifier for each snippet is raising during the backward procedure (calculated by Eq. 9), increasing the difficulty of the model to recognize the action content.

Revision of the weights of TAM. TAM is utilized to actively predict the score that each snippet belongs to action or background, so the revision of the parameters of TAM can effectively achieve the background gradient enhancement strategy. Distinct from the manner adopted on the classifier, the proposed strategy based on the modification of the weight of TAM cannot be directly realized by introducing an auxiliary loss. For achieving the goal, we first analyze the gradient backward process of background classification loss $L_{bg}$, and then implement the background gradient enhancement strategy based on the analysis.

(1) Analysis. Take the attention weight $w_{att}^v$ in RGB modality as an example, we formulate the derivation result with $P_{bg}^{c+1}$ and $y_{r,i}^{C+1}$ as:

$$\frac{\partial L_{bg}}{\partial w_{att}^v} = \sum_{i=1}^{T} \left\{ \frac{(1 - P_{bg}^{c+1})}{2} \cdot \frac{(y_{r,bg}^{C+1} - y_{r,i}^{C+1})}{N_b} \cdot \alpha_i^v \cdot (1 - \alpha_i^v) \cdot X_{r,i}^{C+1} \right\} = \sum_{i=1}^{T} g_i,$$

(11)

where $y_{r,bg}^{C+1} = \frac{1}{N_b} \sum_{i=1}^{T} (1 - \alpha_i^v) \cdot y_{r,i}^{C+1}$ is the video-level background score, $X_{r,i}^{C+1}$ and $g_i$ represent the embedded feature and the gradient of the $i$-th snippet, respectively. The Eq. 11 indicates that the $\frac{\partial L_{bg}}{\partial w_{att}^v}$ accumulates with the gradient of each snippet, and the gradient direction of $g_i$ from the $i$-th snippet depends on the difference $y_{r,bg}^{C+1} - y_{r,i}^{C+1}$. Concretely, if the background score $y_{r,bg}^{C+1}$ of the $i$-th snippet is higher than the video-level one, i.e. $(y_{r,bg}^{C+1} - y_{r,i}^{C+1}) < 0$, the gradient $g_i$ is calculated as a positive by Eq. 11. Correspondingly, the temporal attention weight $w_{att}^v$ calculated by Eq. 9 is lower than it was before, leading to the decrease of foreground score $a_i^v$, and the rise of the background score $(1 - a_i^v)$ as Eq. 4.

(2) Implementation. Based on the above analysis, we attempt to change the difference $(y_{r,bg}^{C+1} - y_{r,i}^{C+1})$ to $(-y_{r,bg}^{C+1} - y_{r,i}^{C+1})$ in Eq. 11, for reducing the gradient of temporal attention weight $w_{att}^v$, which raises the background score $1 - a_i^v$. For clarity, we formulate the weight updating procedure before and after the gradient transformation of the $i$-th snippet as follows:

$$w_{new} = w_{old} + \eta \cdot \frac{val \cdot (y_{r,bg}^{C+1} - y_{r,i}^{C+1})}{2},$$

$$\rightarrow w_{new}' = w_{old} + \eta \cdot \frac{val \cdot (-y_{r,bg}^{C+1} - y_{r,i}^{C+1})}{2},$$

(12)

where $val > 0$ represents the remaining terms in Eq. 11, and the video-level background score is positive. From Eq. 12, we can learn that $w_{new} > w_{new}'$. And the manner, modifying $(y_{r,bg}^{C+1} - y_{r,i}^{C+1})$, increases the gradient that tends to background by $\eta \cdot val \cdot y_{r,bg}^{C+1}$ for the $i$-th snippet. Fortunately, with a clever construction procedure, the background gradient enhancement strategy represented as Eq. 12 can be approximated by setting the normalized factor $N_f$ as $N_f = \sum_{i=1}^{T} a_i$ in Eq. 4, and the new gradient formula is defined as:

$$\frac{\partial L_{bg}}{\partial w_{att}^v} = \sum_{i=1}^{T} \left\{ \frac{(1 - P_{bg}^{c+1})}{2} \cdot \frac{(-y_{r,bg}^{C+1} - y_{r,i}^{C+1})}{N_b} \cdot \alpha_i^v \cdot (1 - \alpha_i^v) \cdot X_{r,i}^{C+1} \right\},$$

(13)

where $y_{r,bg}^{C+1}$ is modified as $\frac{1}{N_f} \sum_{i=1}^{T} (1 - \alpha_i^v) \cdot y_{r,i}^{C+1}$, and the normalized factor of $P_{bg}$ is also set as $N_f$. Notably, we solely revise the

Figure 3: Illustration of the forward and the backward procedures of the model. And the background gradient enhancement strategy is applied to the background classification loss.
normalized factor \( N_b \) in Eq. 4, and the distribution tendency of the background probability is retained. As a result, this is an ingenious way to realize the proposed strategy, leading to the adversarial learning procedure for the model. Here, we adopt the strategy implemented by modifying the weights of TAM in our method, and the extensive experiments demonstrate its effectiveness.

**Discussion:** Notably, GRL [4] is applied to map the features of the target and the source domains into a similar distribution space by reversing the gradient. And our work is aiming at improving the model performance to enlarge the discrepancy between action and background, so GRL is not suitable for our task. Functionally, GRL solely revises the difference \( y_{r,bg} \) to \( y_{r,i} \) in Eq. 11. In this condition, if the \( i \)-th snippet owns highly background score \( y_{r,bg} \), the gradient direction of \( q_i \) would be toward to the foreground instead of the background, and vice versa. Consequently, the excessive confusion caused by GRL blocks the model from learning action information. Distinct from the GRL, the proposed strategy modifies the difference \( (y_{r,i} - y_{r,bg}) \) to \( (-y_{r,bg} + y_{r,i}) \), which impartially increases the background gradient of each snippet by the same magnitude, i.e. \( y_{r,bg} \). In this way, the background loss \( L_{bg} \) involves the model by forcing the gradient of each snippet to tend to be background in backward procedure. Under this condition, the model, driven by the foreground loss \( L_{fg} \), has the ability to focus on action information to mitigate the interference of background-gradient enhancement.

### 3.3 Action localization

In inference stage, we solely run a forward procedure with the base branch to get a set of action proposals for untrimmed videos in test dataset. From the base branch, we obtain the class-activate scores \( y \in \mathbb{R}^{T \times (C+1)} \) and temporal attentions \( a \in \mathbb{R}^T \). And a SoftMax operation is performed on \( y \) to get the normalized action probability \( \tilde{y} \). According to video-level label \( c \) (over a threshold \( p_{cls} \)) predicted by BB, we capture the class-specific score \( \tilde{y}^c \in \mathbb{R}^T \). Following the previous work [24], we calculate the final localization scores \( S_l \) by fusing \( \tilde{y}^c \) and \( a \):

\[
S_l = \varepsilon \cdot \tilde{y}^c + (1 - \varepsilon) \cdot a, \tag{14}
\]

where \( \varepsilon \) is a hyper-parameter. Next, we apply multi-thresholds on \( S_l \) to generate the set \( \{ s_i, e_i \} \) of each action instance. In addition, Non-Maximum Suppression (NMS) is utilized to filter out some redundancy proposals. For the confidence score \( q_i \), we compute it with the general operation mentioned in previous work [19].

### 4 EXPERIMENTS

#### 4.1 Dataset and evaluation

**Dataset:** We evaluate the proposed approach in two benchmarks, i.e. THUMOS14 [11] and ActivityNet1.2 [1]. As for THUMOS14, it contains 200 videos in the validation set and 213 videos in the test set with 20 categories. Following the previous [41], we utilized the validation set for training and evaluate our method in the test set. In particular, this dataset is particularly challenging as it includes multiple action instances even with different category labels in a long untrimmed video. ActivityNet1.2 consists of 4819 training videos and 2383 validation videos with 100 categories. By convention [15], we train our model with the training set and test on the validation set. In contrast to THUMOS14, it is a large-scale benchmark and generally contains an action instance in a video.

**Evaluation Metrics:** We follow the standard evaluation protocol that reports the performance of our method with the Mean Average Precision (mAP) values under various intersection over union (IoU) thresholds. And we utilize the evaluation code provided by ActivityNet to validate the performance in ActivityNet1.2.

#### 4.2 Implementation details

We adopt the I3D [2] network as our feature extractor, which is pre-trained on Kinetics. For a fair comparison, we have not fine-tuned the I3D network on THUMOS14 and ActivityNet1.2. We only fuse their corresponding CAS and temporal attention for training. Adam optimizer is adopted to optimize the networks with a weight decay of 1e-3. The learning rate is initialized as 1e-3 and decreased to 1e-4 after half of the training iterations. For hyperparameters, we set \( \varepsilon, \beta \), and \( k \) as 0.5, 0.1, and 4, respectively. And \( \lambda \) is set as 0.1 and 0.02 for THUMOS14 and ActivityNet1.2, respectively. During inference, the thresholds of the class probability and NMS are set as 0.1 and 0.5 respectively. All the experiments are conducted on Pytorch framework with one 3090 GPU.

### 4.3 Comparison with state-of-the-arts

We compare the proposed approach with recent weakly supervised state-of-the-art methods and several fully supervised. And the results are reported in Tab. 1 and Tab. 2. It can be seen that our method outperforms the mentioned weakly supervised approaches in terms of almost all metrics. Specifically, our method leads to 5.4% average mAP (0.3:0.7) improvement over AUMN [23], which is also designed to model the action of the video. Moreover, we improve the 1.9% average mAP (0.1:0.5) and 2.2% average mAP (0.3:0.7), compared with CO2-Net [8], in which the late fusion way is similar to ours. Compared with several fully supervised counterparts, our method still obtains competitive performance at low temporal IoU. Consequently, the above results demonstrate the effectiveness of the proposed method. ActivityNet1.2 is a large dataset, we also compare several recent works in Tab. 2. It can be seen that our method still achieves great performance at low IoU (0.5). However, our method does not achieve the same improvement on ActivityNet1.2 as it does on THUMOS14. We argue the reasons are that the background snippets on ActivityNet1.2 are easier to distinguish compared to THUMOS14 and the annotations of ActivityNet1.2 are coarser than those of THUMOS14 [8]. Notably, the proposed method mainly contributes to the mitigation of background disturbance. As a result, our method effectively finds out the action instance (i.e. high performance at low IoU) but can not achieve better localization completeness similar to the performance on THUMOS14.

#### 4.4 Ablation study

By convention [23], to look deeper into the proposed method, we conduct the corresponding ablation studies on THUMOS14.

**Effectiveness of each component.** To verify the effect of each design mentioned in the method section, we conduct corresponding ablation studies and report the results in Tab. 3. Here, we treat “BL” as the baseline of our method and set “BGES” as the background gradient strategy based on the modification of attention mechanism.
Table 1: Localization performance comparison with state-of-the-art methods on the THUMOS14 test set.

| Supervision | Method         | mAP(%)@IoU | AVG mAP(%) |
|-------------|----------------|------------|------------|
|             |                | 0.1 0.2 0.3 0.4 0.5 0.6 0.7 (0.1:0.5) (0.3:0.7) (0.1:0.7) |
| Fully       | SSN (2017) [44]| 60.3 56.2 50.6 40.8 29.1 - - | 47.0 28.7 37.8 |
|             | BSN (2018) [18]  | - - 53.5 45.0 36.9 28.4 20.0 - | 38.5 - |
|             | P-GCN (2019) [40]| 69.5 67.5 63.6 57.8 49.1 - - | 61.5 - - |
|             | TSCN (2020) [41]| 63.4 57.6 47.8 37.7 28.7 19.4 10.2 | 47.0 28.7 37.8 |
|             | UM (2021) [15]  | - - - 46.9 39.2 30.7 20.8 12.5 - | 30.0 - |
|             | ACSNet (2021) [21]| - - 51.4 42.7 32.4 22.0 11.7 - | 32.0 - |
|             | AUM (2021) [23] | 66.2 61.9 54.9 44.4 33.3 20.5 9.0 | 52.1 32.4 41.4 |
|             | CoLA (2021) [42]| 66.2 59.5 51.5 41.9 32.2 22.0 13.1 | 50.3 32.1 40.9 |
| Weakly      | D2-Net (2021) [27]| 65.7 60.2 52.3 43.6 34.0 - - | 38.5 - |
|             | TSCN (2020) [41]| 66.2 61.9 54.9 44.4 33.3 20.5 9.0 | 52.1 32.4 41.4 |
|             | UM (2021) [15]  | 66.2 59.5 51.5 41.9 32.2 22.0 13.1 | 50.3 32.1 40.9 |
|             | D2-Net (2021) [27]| 65.7 60.2 52.3 43.6 34.0 - - | 38.5 - |
|             | BSN (2018) [18]  | - - 53.5 45.0 36.9 28.4 20.0 - | 38.5 - |
|             | P-GCN (2019) [40]| 69.5 67.5 63.6 57.8 49.1 - - | 61.5 - - |
|             | TSCN (2020) [41]| 63.4 57.6 47.8 37.7 28.7 19.4 10.2 | 47.0 28.7 37.8 |
|             | UM (2021) [15]  | - - - 46.9 39.2 30.7 20.8 12.5 - | 30.0 - |
|             | ACSNet (2021) [21]| - - 51.4 42.7 32.4 22.0 11.7 - | 32.0 - |
|             | AUM (2021) [23] | 66.2 61.9 54.9 44.4 33.3 20.5 9.0 | 52.1 32.4 41.4 |
|             | CoLA (2021) [42]| 66.2 59.5 51.5 41.9 32.2 22.0 13.1 | 50.3 32.1 40.9 |

Table 2: Comparison of the proposed method with other approaches on the ActivityNet1.2 validation set. AVG indicates average mAP from IoU 0.5 to 0.95 with a 0.05 increment.

| Supervision | Method          | mAP@IoU | AVG mAP(%) |
|-------------|----------------|---------|------------|
|             |                | 0.5 0.75 0.95 AVG(0.5:0.7) |
| Fully       | SSN (2017) [44]| 41.3 27.0 6.1 26.6 |
|             | DGAM (2020) [31]| 41.0 23.5 5.3 24.4 |
|             | TSCN (2020) [41]| 37.6 23.7 5.7 23.6 |
|             | BS (2020) [14] | 38.5 24.2 5.6 24.3 |
| Weakly      | HAM-Net (2021) [12]| 41.0 24.8 5.3 25.1 |
|             | UM (2021) [15]  | 41.2 25.6 6.0 25.9 |
|             | ASCNet (2021) [21]| 40.1 26.1 6.8 26.0 |
|             | Ours           | 41.6 24.8 5.4 25.2 |

Table 3: Evaluation of the effectiveness of the designed component under multiple IoU thresholds from 0.3 to 0.7 with an increase of 0.1.

| Method      | mAP@IoU | AVG(0.3:0.7) |
|-------------|---------|--------------|
|             | 0.3 0.4 0.5 0.6 0.7 AVG(0.3:0.7) |
| BL          | 51.4 42.1 31.3 20.6 10.5 31.2 |
| BL+BGES    | 53.2 43.6 33.2 21.8 10.9 32.5 |
| TEN         | 53.7 46.1 35.6 24.8 13.2 34.7 |
| TEN+BGES   | 55.7 46.0 35.2 22.6 12.1 34.3 |

Table 4: Comparisons of different types of combinations for the gradient modification strategy and the designed component.

| Method      | mAP@IoU | AVG(0.3:0.7) |
|-------------|---------|--------------|
|             | 0.3 0.4 0.5 0.6 0.7 AVG(0.3:0.7) |
| BL          | 51.4 42.1 31.3 20.6 10.5 31.2 |
| BL+GRL     | 53.2 43.6 33.2 21.8 10.9 32.5 |
| TEN+GRL    | 55.7 46.0 35.2 22.6 12.1 34.3 |
| TEN+BVL    | 55.9 47.1 35.6 24.5 12.8 35.2 |
| TEN+BVL+BGES| 57.4 49.5 39.0 27.1 14.1 37.5 |

Table 5: Ablation studies of different kinds of object loss on attention (ATT) and class activate scores (CAS).

| Object | Loss | mAP@IoU |
|--------|------|---------|
|        | 0.1 0.2 0.3 0.4 0.5 0.6 0.7 AVG(0.3:0.7) |
| ATT    | MAE 68.1 64.0 56.5 46.5 35.3 24.5 13.2 44.0 |
|        | MSE 67.8 63.9 56.2 46.4 35.0 24.4 13.2 43.8 |
| CAS    | MAE 67.9 63.7 56.4 46.4 35.2 24.7 13.3 43.9 |
|        | MSE 68.0 64.0 56.3 46.5 35.1 24.5 13.3 43.9 |
|        | KL   69.4 64.0 56.9 48.6 38.1 26.8 14.3 45.5 |

Weights, and “TEN” represents the temporal enhancement network. From Tab. 3, we can find that BGES obtains a 3.3% improvement over the baseline, and TEN also surpasses the baseline by 3.5%. BGES mainly promotes the performance on the lower IoU (mAP@IoU 0.3), while TEN improves the effectiveness on the higher one (mAP@IoU 0.5:0.7). And the full mode by combining BGES and TEN largely surpasses the baseline by 6.6%, and it outperforms both variants, proving that the above two components are not redundant.

Effect of different manners for gradient modification. We compare the different gradient-modification ways mentioned in the approach section, and the related results are reported in Tab. 4. Note that BVL is the proposed strategy shown in Eq. 10. We first apply GRL to the base model, and it outperforms the baseline on the low IoU but is not significant. Furthermore, we assemble GRL to TEN, and this manner solely improves the performance on low IoU but worsens the results on the average mAP@IoU (0.3:0.7), which indicates GRL solely facilitates the model to focus on the...
Table 6: Evaluation of different numbers of the sampling interval $k$ in TEN in terms of average mAP under IoU thresholds (0.1:0.1:0.7).

| $k$ | mAP@IoU | AVG mAP(%) |
|-----|---------|------------|
|    | 0.1  | 0.2  | 0.3  | 0.4  | 0.5  | 0.6  | 0.7  | 0.1:0.5 | 0.3:0.7 |
| 2  | 70.0 | 65.4 | 58.5 | 48.1 | 36.7 | 26.1 | 13.8 | 55.7 | 36.6 |
| 3  | 69.7 | 64.0 | 57.3 | 49.1 | 38.8 | 27.0 | 14.5 | 55.7 | 37.4 |
| 4  | 68.5 | 63.7 | 57.7 | 49.4 | 39.4 | 27.1 | 13.9 | 55.7 | 37.5 |
| 5  |       |       |       |       |       |       |       |       |

Figure 4: Ablation results of the baseline with different background classification loss weights on the THUMOS14 dataset.

The influence of the weight $\lambda$ of background classification loss. To better verify the effectiveness of BGES, we also investigate the influence of $\lambda$ by setting it from 0.0 to 0.5 in the baseline method. The results are shown in Fig. 4. It can be seen that the performance of the network has dropped obviously when the value is larger than 0.3. Therefore, it is evident that simply increasing the value of the weight $\lambda$ cannot achieve the adversarial learning procedure similar to BGES for improving the performance of action localization.

The effectiveness of MAE, MSE, and KL. We adopt different losses on temporal attention (ATT) and class activate scores (CAS) in Tab. 5. It can be seen that all of the constraint losses on ATT or CAS improve the performance of action localization. As for the CAS, KL loss outperforms MAE and MSE by 1.6% on the average mAP (0.1:0.7), which indicates that KL is beneficial to mutual learning among the branches. As for the ATT, MAE loss shows a better performance than MSE loss.

The effectiveness of MAE, MSE, and KL. We adopt different losses on temporal attention (ATT) and class activate scores (CAS) in Tab. 5. It can be seen that all of the constraint losses on ATT or CAS improve the performance of action localization. As for the CAS, KL loss outperforms MAE and MSE by 1.6% on the average mAP (0.1:0.7), which indicates that KL is beneficial to mutual learning among the branches. As for the ATT, MAE loss shows a better performance than MSE loss.

4.5 Qualitative results

To better demonstrate the effectiveness of the proposed method, we visualize several examples of localized action instances in Fig. 5. For the first instance of the soccer penalty, compared with the baseline, the proposed strategy BGES and the full method by combining BGES and TEN can effectively alleviate the interference of the background (shown in blue rectangles). In the second example of clean-and-jerk, compared with the baseline and BGES, the proposed TEN successfully constructs the temporal relation of affinity snippets and makes the detected results more smooth. Consequently, all of the visual results verify the effect of the proposed method.

5 CONCLUSION

In this work, we explore an adversarial learning strategy for WTAL and propose a novel temporal enhancement network (TEN) based on the background gradient enhancement strategy (BGES). We utilized the background gradient enhancement strategy to achieve the adversarial learning procedure, aiming to relieve the background disturbance in the model learning process. Additionally, the temporal enhance network is designed to construct the temporal consistency of continuous snippets, improving the ability to localize complete action. Finally, we conduct extensive experiments to verify the effectiveness of the proposed method.

6 ACKNOWLEDGEMENTS

The work described in this paper was partially supported by the National Natural Science Foundation of China (Grant no.62176031) and the Fundamental Research Funds for the Central Universities (Grant no. 2021CDJQY-018).
