

Temporal Action Proposal Generation with Transformers

Lining Wang*

Harbin Institute of Technology

Haosen Yang*

Harbin Institute of Technology

Wenhuo Wu*

Department of Computer Vision
Technology (VIS), Baidu Inc.Hongxun Yao[†]
Harbin Institute of TechnologyHujie Huang
Harbin Institute of Technology

ABSTRACT

Transformer networks are effective at modeling long-range contextual information and have recently demonstrated exemplary performance in the natural language processing domain. Conventionally, the temporal action proposal generation (TAPG) task is divided into two main sub-tasks: boundary prediction and proposal confidence prediction, which rely on the frame-level dependencies and proposal-level relationships separately. To capture the dependencies at different levels of granularity, this paper intuitively presents an unified temporal action proposal generation framework with original Transformers, called **TAPG Transformer**, which consists of a *Boundary Transformer* and a *Proposal Transformer*. Specifically, the Boundary Transformer captures long-term temporal dependencies to predict precise boundary information and the Proposal Transformer learns the rich inter-proposal relationships for reliable confidence evaluation. Extensive experiments are conducted on two popular benchmarks: ActivityNet-1.3 and THUMOS14, and the results demonstrate that TAPG Transformer outperforms state-of-the-art methods. Equipped with the existing action classifier, our method achieves remarkable performance on the temporal action localization task. Codes and models will be available.

CCS CONCEPTS

- Computing methodologies → Activity recognition and understanding;

KEYWORDS

transformers, temporal action proposal generation, untrimmed video, temporal action localization

ACM Reference format:

Lining Wang*, Haosen Yang*, Wenhuo Wu*, Hongxun Yao[†], and Hujie Huang. 2021. Temporal Action Proposal Generation with Transformers. In *Proceedings of Woodstock '18: ACM Symposium on Neural Gaze Detection, Woodstock, NY, June 03–05, 2018 (Woodstock '18)*, 10 pages.

DOI: 10.1145/3394171.3413860

* Co-first authorship: the order of first authors was randomly selected. [†] 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.

Woodstock '18, Woodstock, NY

© 2021 ACM. 978-1-4503-7988-5/20/10...\$15.00

DOI: 10.1145/3394171.3413860

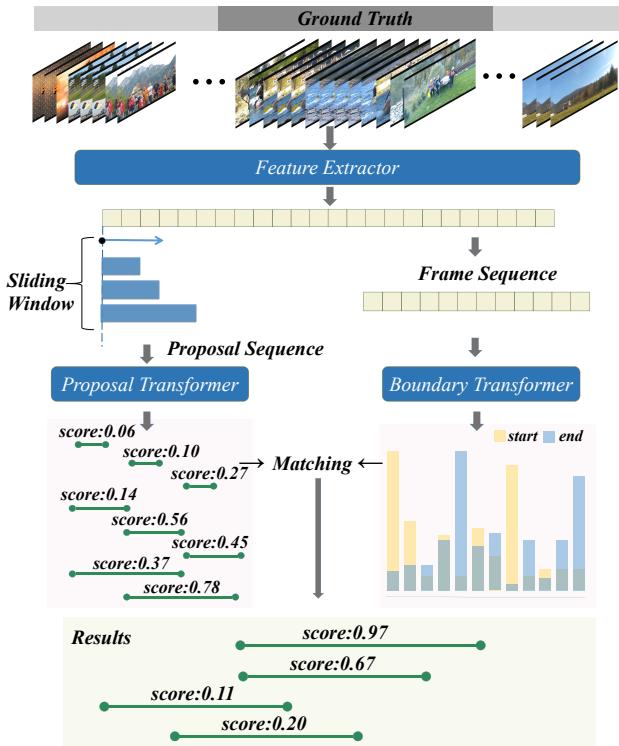


Figure 1: Overview of TAPG Transformer. Given an untrimmed video, Boundary Transformer captures inter-frame relation and Proposal Transformer encodes inter-proposal relation. We further match the boundary probabilities with proposal confidence sequence to generate the final proposals.

1 INTRODUCTION

With the rapid development of the mobile devices and Internet, hours of video were uploaded to the Internet every second. Huge video information has far exceeded the processing capacity of the conventional manual system, thus video content analysis methods have attracted increased interest from both academia and industry.

Video temporal action detection, is one of the most active research topics in video understanding, which focuses both on classifying the action instances present in an untrimmed video and on localizing them with temporal boundaries. Similar to object detection, temporal action detection task is composed of two sub-tasks: temporal action proposal generation (TAPG) and action recognition.

Recently, significant progress has been achieved in action recognition ([12, 24, 36, 45]) following the deep convolutional network paradigm. However, the performance of such two-stage temporal action detectors still have much room for improvement in mainstream benchmarks [4, 19], which is largely determined by the proposal quality from TAPG. Hence, many recent works focus on generating the high quality temporal action proposals which have flexible duration, precise boundaries and reliable confidence scores.

In general, there are two main categories in the existing TAPG methods: *anchor-based* regression methods and the *boundary-based* regression methods. These two kinds of regression methods usually adopt context information in video from different aspects. *anchor-based* methods [7, 10, 13–15, 27, 35] generate action proposals based on multi-scale and dense pre-defined anchors, in this way, we can evaluate confidence scores of proposals with rich proposal-level context information. Therefore, these methods can obtain reliable confidence scores, but are inflexible and usually have imprecise boundaries. Recently, *boundary-based* methods [23, 25, 26] utilize the frame-level context information around the boundaries to predict boundaries. Thus, compared with *anchor-based* methods, they can generate proposals with more flexible durations and more precise boundaries. Meanwhile, they are more sensitive to noises and fail to consider the rich proposal-level context for confidence evaluation. Based on these above analyses, we attempt to take fuller advantage of both frame-level context and proposal-level context for temporal proposal generation.

In this paper, we focus on inter-frame relation modeling and inter-proposal relation modeling for temporal proposal generation. Inspired by the successful application of Transformers [43] in sequence prediction task, we intuitively take advantage of Transformers in modeling long-range contextual information. Therefore, we propose an unified temporal action proposal generation framework with Transformers, called TAPG Transformer. As shown in Figure 1, our TAPG Transformer mainly contains two following modules: ***Boundary Transformer*** and ***Proposal Transformer***. ***Boundary Transformer*** aims to locate precise action boundaries by capturing the rich long-term temporal relationships between local details and global dependencies. To do so, the sequence of video features is provided as an input to the Transformer and output of the module is boundary probabilities. Then, ***Proposal Transformer*** is proposed to capture the potential inter-proposal relationships for confidence evaluation. Before performing inter-proposal relations, we propose a *Sparse Sampling Mechanism* to generate the sparse proposal sequence instead of the densely distributed proposals which will bring imbalanced data distribution between positive/negative proposals and more computational burdens. Then we feed them into the Proposal Transformer which discovers the complex relationships between proposals with different scales, and the Transformer outputs proposal confidences. Finally, we take the boundary probabilities and proposal confidences as input to *Fuzzy Matching* mechanism then get the matching proposals. Followed by the post processing, we obtain the final proposal set.

In summary, our work has three main contributions as follows:

- We propose a Boundary Transformer to capture the long-term frame-level dependencies for accurate temporal boundary prediction.

- We propose a Proposal Transformer with a sparse proposal sampling mechanism, which can learn the proposal-level context for proposal confidence evaluation. Besides, sparse sampling can significantly reduce the impact brought by the densely distributed proposals.
- Extensive experiments demonstrate that our method outperforms the existing state-of-the-art methods on THU-MOS14 and achieves comparable performance on ActivityNet-1.3, in both temporal action proposal generation task and temporal action detection task.

2 RELATED WORK

2.1 Video Action Recognition

Action recognition is a fundamental and important task of the video understanding area. Currently, the end-to-end action recognition methods can be mainly divided into two types: 3D CNN-based methods and 2D CNN-based methods. 3D CNNs are intuitive spatio-temporal networks and natural extension from their 2D counterparts, which directly tackle 3D volumetric video data [6, 40] but have a heavy computational cost. Recent studies have shown that factorizing 3D convolutional kernel into spatial (e.g., $1 \times 3 \times 3$) and temporal components (e.g., $3 \times 1 \times 1$) is preferable to reduce complexity as well as boost accuracy, e.g., P3D [33], R(2+1)D [42], S3D-G [50], etc. Thus, 3D channel-wise convolution are also applied to video recognition in CSN [41] and X3D [11]. To capture temporal information with reasonable training resources, the other alternative 2D CNN based architectures are developed i.e., TSM [24], TEI [28], TEA [22], MVFNet [46], etc. These methods aim to design efficient temporal module based on existing 2D CNNs to perform efficient temporal modeling. There is also active research on dynamic inference [48], adaptive frame sampling techniques [20, 47, 49], which we think can be complementary to the end-to-end video recognition approaches.

2.2 Temporal Action Proposal Generation

Current temporal action proposal generation methods can be mainly divided into anchor-based and boundary-based methods. The anchor-based methods refer to the temporal boundary refinements of sliding windows or pre-defined anchors. [19, 31] use temporal sliding window and [35, 57] use snippet-wise probability to generate candidate proposals. TURN [15], CTAP [13] and Rap [14] adopt pre-defined anchors to generate proposals. These methods can evaluate confidence scores of proposals with rich proposal-level context information and obtain reliable confidence scores. Boundary-based methods generate the boundary probability sequence, then apply the Boundary Matching mechanism to generate candidate proposals. BSN [26], BMN [25] and BSN++ [38] directly predict the probability of boundaries and cover flexible temporal duration of action instances. Also, the ideas of GCN have been adopted in some works, G-TAD [53] extract temporal context and semantic context by GCN, but applies conventional temporal convolutional layers for actual implementation in temporal context extraction. They can generate proposals with more flexible durations and more precise boundaries. MGG [27] and TCANet [32] combines a boundary-based and anchor-based advantages to generate proposal. However, they have limited capability of modeling dependencies from the local context,

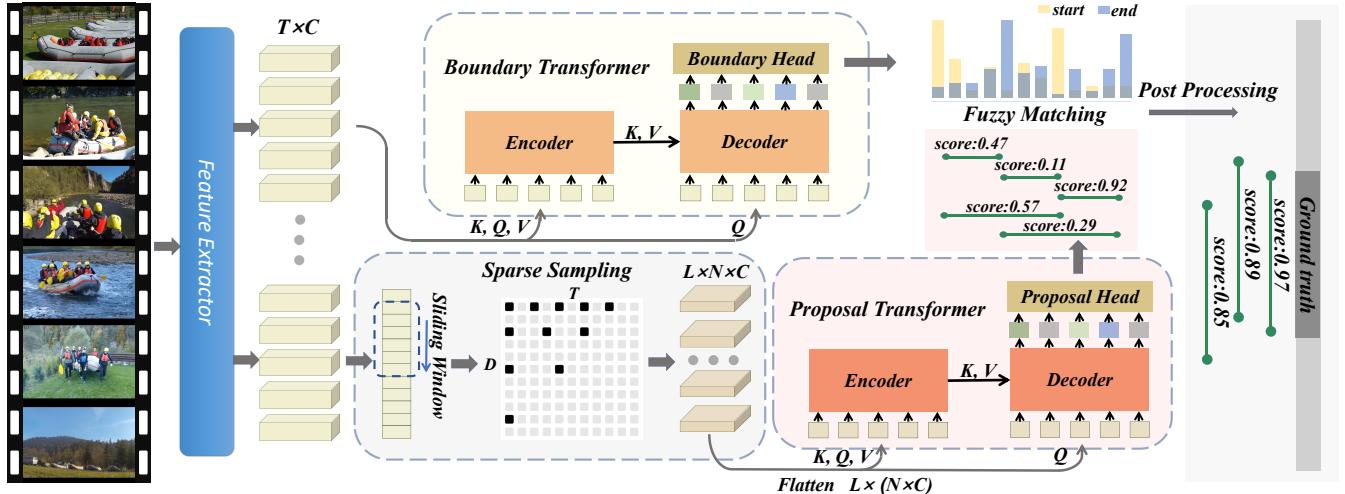


Figure 2: Illustration of the proposed TAPG Transformer framework. First we apply the feature extractor to encode video frames. The Boundary Transformer takes the feature sequence as input, and outputs boundary probability sequence. For Proposal Transformer, we apply sliding window and sparse sampling on the feature sequence to generate sparse multi-scale proposals, which are then fed into Proposal Transformer to generate confidence sequence. Finally, we construct proposals based on boundary probabilities sequence and obtain the corresponding confidence score from confidence sequence.

leading to performance penalty in the complex cases. In our work, we aim to make fuller use of inter-frame relation and inter-proposal relation for boundary prediction and the proposal-level confidence evaluation, respectively.

2.3 Transformers in computer vision

Transformer architectures are based on a self-attention mechanism that learns the relationships between elements of a sequence. The breakthroughs from Transformer networks [43] in language domain has sparked great interest in the computer vision community to adapt these models for vision. ViT [9] proposes to use NLP Transformer encoder directly on image patches. In order to apply Transformer model, DETR [5] treats object detection as a set prediction problem. Transformers are also adopted for Super resolution in [55], Image Colorization in [21], Tracking in [8, 54, 58], Pose estimation in [29], etc. Besides, for video understanding, there are also recent approaches seek to resolve this challenge using the Transformer networks. Video Transformer Network (VTN) that first obtain frame-wise features using 2D CNN and apply a Transformer encoder (Longformer [30]) on top to learn temporal relationships. Timesformer [1] presents a convolution-free approach to video classification built exclusively on Transformers over space and time. Our work is most closely related to the works which apply Transformers for temporal action proposal generation. RTD-Net [39] adapts the Transformer-like architecture for action proposal generation in videos. However, there are substantial differences between RTD-Net and our method. RTD-Net uses decoder-only Transformer to exploit the global temporal context and devises a three-branch detection head for training and inference. On the contrary, in our paper we propose to adopt two original Transformers (Boundary Transformer, Proposal Transformer) to capture both inter-frame

and inter-proposal relationships for precise boundary prediction and proposal confidence evaluation, respectively. On both two challenge benchmarks (*i.e.*, THUMOS14 and ActivityNet13), our method outperforms RTD-Net with a clear margin.

3 APPROACH

In this section, first, we present the problem definition for the temporal action proposal generation task, and then we describe the basic Transformer architecture [43] that applied attention mechanisms to make global dependencies in a sequence data. In Sec. 3.3, we present an overview of our **TAPG Transformer**. The architecture of TAPG Transformer is composed of two main parts: a **Boundary Transformer** (Sec. 3.4) and a **Proposal Transformer** (Sec. 3.5). Finally, the training and inference details of our TAPG Transformer are introduced in Sec 3.6 and Sec 3.7, respectively.

3.1 Problem Definition

An untrimmed video U can be denoted as a frame sequence $U = \{u_t\}_{t=1}^{l_v}$ with l_v frames, where u_t indicates the t -th RGB frame of video U . The temporal annotation set of U is composed by a set of temporal action instances as $\Psi_g = \{\phi_n^g\}_{n=1}^{N_g}$ and $\phi_n^g = (t_{s_n}, t_{e_n})$, where N_g is the number of ground-truth action instances, t_{s_n} and t_{e_n} are the starting and ending time of the action instance ϕ_n^g respectively. During training phase, the Ψ_g is provided. While in the testing phase, the predicted proposal set Ψ_p should cover the Ψ_g with high recall and high temporal overlapping.

3.2 Preliminary: General Transformer

The original Transformer [43] has an encoder-decoder structure. The **encoder** is composed of six identical blocks, with each block

having two sub-layers: a multi-head self-attention layer and a simple position-wise fully connected feed-forward layer. Similar to the encoder, the **decoder** in the Transformer model consists of six identical blocks. In addition to the two sub-layers in each encoder block, the decoder inserts a third sub-layer, which performs multi-head attention on the output of the corresponding encoder block.

A key feature of the Transformer architecture is the so-called self-attention mechanism, which explicitly models the interactions between all entities of a sequence for structured prediction tasks. Formally, let's denote a sequence of n entities by $F \in \mathbb{R}^{n \times d}$, where d is the embedding dimension to represent each entity. The input sequence F is projected onto three learnable linear transformations to get queries Q , keys K , and values V . Then, we compute the output of the self-attention mechanism as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V. \quad (1)$$

3.3 TAPG Transformer

Temporal action proposal generation task is generally divided into two sub-task: boundary prediction and proposal confidence evaluation. The long-range frame-level dependencies are desirable in precise boundary prediction. Moreover, modeling the rich inter-proposals relationships play a crucial role in confidence regression. Motivated by the success of Transformer models in the language domain, similar to how a Transformer operates on a sentence, it could naturally be applied to a sequence of frames or proposals. Therefore, we propose to solve each task sequentially by a different Transformer network which enables capturing the long-term information and dependencies between sequence elements. Figure 2 illustrates the architecture of our proposed framework, which called TAPG Transformer.

Given an untrimmed video U which contains l_v frames, we process the input video in a regular interval σ for reducing the computational cost. Then we utilize the feature extractor to encode the visual feature from video frames. We concatenate the output of the last fully connected layer in the two-stream network [36] to form the feature sequence $F = \{f_i\}_{i=1}^T$, where $T = l_v/\sigma$. Finally, the feature sequence F is used as the input of our framework.

Our model mainly contains two main modules: **Boundary Transformer** and **Proposal Transformer**. Boundary Transformer aims to capture the long-range dependencies between the frame-level features then output the boundary probabilities sequence. Similar to the Boundary Transformer which learns the relationships between frames, we propose a Proposal Transformer to model the relation between proposals with different scales and evaluate the confidence of proposals. Before performing inter-proposal relations, we need to generate the proposal features as the input to the Proposal Transformer. To do so, the previous works [25, 38] usually utilize dense sampling, which will generate a lot of redundant proposals and cause an imbalance data distribution between positive/negative proposals and temporal durations. Therefore, we propose a sparse sampling mechanism to generate the candidate proposals, then feed them into the Proposal Transformer which encodes the potential relationship between proposals, then output the corresponding confidence score from proposal confidence sequence. We take the boundary probabilities and proposal confidences as input to *Fuzzy*

matching mechanism then get the matching proposals. Followed by the post processing, we obtain the final proposal set. We will describe the details of each module in the following sections.

3.4 Boundary Transformer

Long-term temporal dependency modeling is essential for boundary regression. Previous methods usually conduct stacked temporal convolutions [23, 25, 26] or global temporal pooling [14] to capture temporal relationship. However, 1D temporal convolution can effectively capture local temporal relation between frames but lacks the capacity of modeling long-term temporal structure limited by the kernel size. On the contrary, the global temporal pooling performs average pooling to capture the global features over frames but fails to capture finer temporal structure and may introduce unnecessary noise. To this end, we propose the Boundary Transformer to model both local temporal relation and long-range temporal dependency structure with a Transformer structure. The architecture of the Transformer is the same as the original Transformer model [43], which has been described in Sec. 3.2. We adopt a M -layer Transformer and each layer has an encoder-decoder structure as follows:

The **Encoder** is applied to estimate the relevance of one frame to other frames. First, we obtain the frame-wise feature sequence $F \in \mathbb{R}^{T \times C}$ using two-stream model [36], where T is the length of the feature sequence and C is the feature dimension. Then we feed them to the Transformer encoder and generate an augmented feature sequence $F_g \in \mathbb{R}^{T \times C}$ with the global context. The encoder has a standard architecture, which consists of a multi-head self-attention module and a feed forward network (FFN). The outputs F_g of the encoder along with the embeddings are given as input to the Transformer decoder.

The **Decoder** contains three sub-layers, first two (multi-head self-attention and feed forward network) are similar to the encoder, while the third sub-layer performs multi-head attention over the output of the encoder stack. The input to the first self-attention module is the feature sequence $F \in \mathbb{R}^{T \times C}$, which is the same as that of the encoder. Then the module output the feature sequence F_q . In particular, for the second self-attention module, the K, V are transformed from F_g and Q is transformed from F_q . In this way, the decoder further enhances relevant features while reducing the over-smoothing effect from the decoder.

The output representation of decoder is then used as the global representation for the boundary prediction task. Specifically, we append a **Boundary Head** which encode the output representation of decoder with multi-layer perceptron (MLP) network and followed by a *Sigmoid* layer to generate the predicted boundary probability sequence.

3.5 Proposal Transformer

For proposal confidence prediction, previous works [25] first generate densely distributed proposals then predict the confidence scores of them. Obviously, dense proposals will result in imbalanced data distribution between positive/negative proposals and more computational burdens. Meanwhile, the inter-proposal relationships modeling is a previously overlooked point, which plays a crucial role in proposal confidence evaluation. To relieve these issues, we

propose a sparse sampling mechanism and a Proposal Transformer as follows.

The **Sparse sampling mechanism** aims to maintain the sparse proposal sequence. Specifically, we use the Fibonacci sequence as a sliding window group $W = \{w_i\}_{i=1}^D$ to generate the proposals with different scales, where D is the group size. The step size of each sliding window is computed as:

$$\text{step} = \lfloor \sqrt{w_i} \rfloor + \lfloor \frac{w_i}{\gamma} \rfloor, \quad (2)$$

where γ is used to constraint the step size increases with the increase of the window size. For each proposal, we follow the BMN [25] and construct weight term $w_{i,j} \in \mathbb{R}^{N \times T}$ via uniformly sampling N points between the temporal region. Then we conduct dot product in temporal dimension between $w_{i,j} \in \mathbb{R}^{N \times T}$ and $F \in \mathbb{R}^{T \times C}$ to generate the proposal feature with the shape $C \times N$. Finally, we get the sparse proposal feature sequence $P \in \mathbb{R}^{L \times S}$ and these proposal features are subsequently flattened, where L is number of proposal and $S = C \times N$.

Moreover, we propose the Proposal Transformer to enhance the proposal feature representations for proposal confidence prediction. We adopt a M -layer Transformer, which is based upon the encoder and the decoder in original Transformer model [43]. The **Encoder** aims to capture the relationships between proposals in various scales. The sparse proposal feature sequence $P \in \mathbb{R}^{L \times S}$ is given as input to the encoder. Then the encoder outputs the feature sequence P_g with the global view. Similar to the Boundary Transformer, the decoder takes inputs from the encoder as well as the previous outputs to generate enhanced proposal feature sequence. Taking the proposal sequence from the decoder output, our **Proposal Confidence Generator** will generate two types of confidence $C_c, C_r \in \mathbb{R}^{L \times 1}$ with *Sigmoid* activation, where C_c, C_r are used for binary classification and regression loss function separately.

3.6 Training

The overall objective of our framework is defined as:

$$\mathcal{L} = \mathcal{L}_b + \mathcal{L}_p, \quad (3)$$

where \mathcal{L}_b and \mathcal{L}_p are the objective functions of the Boundary Transformer and the Proposal Transformer respectively.

Objective of Boundary Transformer. The Boundary Transformer generates the starting and ending probability sequence P_s, P_e . We can define training objective of the Boundary Transformer as a two-task loss function. The loss function consists of starting loss and ending loss:

$$\mathcal{L}_b = \mathcal{L}_{bl}(P_s, G_s) + \mathcal{L}_{bl}(P_e, G_e), \quad (4)$$

where G_s and G_e are the ground truth labels of boundary sequence, and \mathcal{L}_{bl} is the binary logistic regression loss function.

Objective of Proposal Transformer. The Proposal Transformer generates sparse proposal confidence sequence C_c and C_r . The loss function \mathcal{L}_p consists of binary classification loss and regression loss:

$$\mathcal{L}_p = \mathcal{L}_c(C_c, G_c) + \mathcal{L}_r(C_r, G_c), \quad (5)$$

where \mathcal{L}_c is a binary logistic regression loss function, \mathcal{L}_r is a smooth L_1 loss, and G_c is the ground truth label of sparse proposal sequence.

The details of label assignment will be described in the supplementary material.

3.7 Inference

As mentioned above, the Boundary Transformer generates boundary probability sequence and the Proposal Transformer generates the sparse proposal confidence sequence. Then we construct a coarse proposals set ψ_p^c based on boundary probabilities. Take the proposal $\varphi = [t'_s, t'_e] \in \psi_p^c$ as an example, **Fuzzy matching** is devised to compute IoU (temporal Intersection over Union) between proposal φ and sparse proposal sequence, then select matching proposal $p_m = [t_m^s, t_m^e]$. Next, we refine the proposal as:

$$[t_s, t_e] = \begin{cases} [\frac{t'_s+t_m^s}{2}, \frac{t'_e+t_m^e}{2}], & \text{if } p_m^{cc} > \alpha_1 \text{ and } p_m^{cr} > \alpha_2 \\ [t'_s, t'_e], & \text{others} \end{cases}, \quad (6)$$

where p_m^{cc} is the proposal classification score, p_m^{cr} is the regression confidence, α_1 and α_2 are the adjustment thresholds. Finally, we get a proposal set $\psi_p = \{\varphi_n = (t_s, t_e, p_{t'_s}^s, p_{t'_e}^e, p_m^{cc}, p_m^{cr})\}_{n=1}^N$, where $p_{t'_s}^s, p_{t'_e}^e$ are the starting and ending probabilities.

Following the previous practices [25, 26], we also perform score fusion and redundant proposal suppression to further obtain final results. Specifically, in order to make full use of various predicted scores for each proposal φ_n , we fuse its boundary probabilities and confidence scores of matching proposal by multiplication. The confidence score p^f can be defined as :

$$p^f = p_{t'_s}^s \cdot p_{t'_e}^e \cdot p_m^{cc} \cdot p_m^{cr}. \quad (7)$$

Hence, the final proposal set as

$$\psi = \{\varphi_n = (t_s, t_e, p^f)\}_{n=1}^N. \quad (8)$$

Moreover, we use the Soft-NMS [2] algorithm for **Post-processing** to remove the proposals which highly overlap with each other.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

We evaluate our method on two challenge benchmarks: **ActivityNet-v1.3** [4] and **THUMOS14** [19]. ActivityNet-v1.3 is a large-scale video dataset for action recognition and temporal action detection tasks. It contains 10K training, 5k validation, and 5k testing videos with 200 action categories, and the ratio of training, validation and testing sets is 2:1:1. Our models trained on the training set and evaluated with the validation set. For THUMOS14, we use the subset of THUMOS14 that provides frame-wise action annotations. The model is trained with 200 untrimmed videos from its validation set and evaluated using 213 untrimmed videos from the test set. This dataset is challenging due to the large variations of the frequency and duration of action instances across videos.

Evaluation Metrics. Temporal action proposal generation aims to produce high quality proposals that have high IoU with ground truth action instances and high recall. To evaluate quality of proposals, Average Recall (AR) is the average recall rate under specified IoU thresholds for measuring the quality of proposals. ActivityNetv1.3, these thresholds are set to [0.5:0.05:0.95]. On THUMOS14, they are set to [0.5:0.05:1.0]. By limiting the average number (AN) of proposals for each video , we can calculate the area under the

AR vs. AN curve to obtain AUC. We use **AR@AN** and **AUC** as our metrics to evaluate TAPG models. For the temporal action detection task, mean Average Precision (**mAP**) under multiple tIoU is the widely used evaluation metric. On ActivityNet-v1.3, the tIoU thresholds are set to {0.5, 0.75, 0.95}, and we also test the average mAP of tIoU thresholds between 0.5 and 0.95 with step of 0.05. On THUMOS14, these tIoU thresholds are set to {0.3, 0.4, 0.5, 0.6, 0.7}.

4.2 Implementation Details.

For feature extractor, following previous works [25, 53], we adopt the two-stream network [36], where ResNet [16] and BN-Inception network [18] are used as the spatial and temporal networks respectively. During feature extraction, the interval σ is set to 16 and 5 on ActivityNet-1.3 and THUMOS14 respectively. For ActivityNet-v1.3, each video feature sequence is rescaled to $T = 100$ snippets using linear interpolation. For THUMOS14, we crop each video feature sequence with overlapped windows of size $T = 256$ and stride 128. Besides, for the sparse sampling mechanism, the length of sparse proposal sequence L is 451 and 1225 on ActivityNet-1.3 and THUMOS14, the max sliding window size D is 100 and 64 for ActivityNet-v1.3 and THUMOS14, respectively, and the constraint hyper-parameter γ is 21. For each proposal, we sample points $N = 32$. For Fuzzy matching, we set adjustment thresholds $\alpha_1 = 0.9$ and $\alpha_2 = 0.8$. Due to the limit of computation resource, we apply 1D Conv for dimension reduction, then take the features as the input to the Boundary Transformer and Proposal Transformer. For temporal action detection task, following [53], we take the video classification scores from [51] and [44] and multiply them to yield the fused confidence score p^f for evaluation. We train our TAPG Transformer from scratch using the Adam optimizer and the learning rate is set to 10^{-4} and decayed by a factor of 0.1 after every 10 epoch. To preserve the positional information in the transformer, we adopt sine positional encoding for queries and keys. The number of Transformer layers $M = 3$.

4.3 Comparison with State-of-the-art Results

In this subsection, we compare our method with the existing state-of-the-art methods on ActivityNet-v1.3 and THUMOS14. For a fair comparison, we adopt the same two-stream features used by previous methods in our experiments.

Table 1: Performance comparison with state-of-the-art proposal generation methods on test set of THUMOS14 in terms of AR@AN.

Method	@50	@100	@200	@500	@1000
TAG [57]	18.6	29.0	39.6	-	-
CTAP [13]	32.5	42.6	52.0	-	-
BSN [26]	37.5	46.1	53.2	61.4	65.1
MGG [27]	39.9	47.8	54.7	61.4	64.6
BMN [25]	39.4	47.7	54.8	62.2	65.5
BSN++ [38]	42.4	49.8	57.6	65.2	66.8
TCANet [32]	42.1	50.5	57.1	63.61	66.9
RTD-Net [39]	41.1	49.0	56.1	62.9	-
Ours	43.9	53.5	60.2	66.7	68.8

Table 2: Performance comparison with state-of-the-art action detection methods on test set of THUMOS14, in terms of mAP (%) at different tIoU thresholds.

Method	0.3	0.4	0.5	0.6	0.7
SST [3]	-	-	23.0	-	-
CDC [34]	40.1		23.0	-	-
TURN-TAP [15]	44.1	34.9	25.6	-	-
SSN [56]	51.9	41.0	29.8	-	-
BSN [26]	53.5	45.0	36.9	28.4	20.0
TAL-Net [7]	53.2	48.5	42.8	33.8	20.8
MGG [27]	53.9	46.8	37.4	29.5	21.3
DBG [23]	57.8	49.4	39.8	30.2	21.7
BMN [25]	56.0	47.4	38.8	29.7	20.5
G-TAD [53]	54.5	47.6	40.2	30.8	23.4
BSN++ [38]	59.9	49.5	41.3	31.9	22.8
TCANet [32]	60.6	53.2	44.6	36.8	26.7
RTD-Net [39]	53.9	48.9	42.0	33.9	23.4
Ours	65.3	59.3	50.8	40.1	28.5

Results on THUMOS14. We present performance comparison of the proposed TAPG Transformer with the state-of-the-art methods on THUMOS14 in Table 1 and Table 2, where our method improves the performance significantly for both temporal action proposal generation and action detection. For the temporal action proposal generation task, results are shown in Table 1, which demonstrate that TAPG Transformer outperforms state-of-the-art methods in terms of AR@AN with AN varying from 50 to 1000. For the temporal action detection task, the proposed TAPG Transformer also achieves superior results, as shown in Table 2. The performance of our method exceeds state-of-the-art proposal generation methods by a big margin at different tIoU thresholds. The high mAP reflects that our model is able to predict the proposals with high scores, and reduce the number of false positives.

Table 3: Performance comparison with state-of-the-art proposal generation methods on validation set of ActivityNet-1.3 in terms of AUC and AR@AN.

Method	AR@1 (val)	AR@100 (val)	AUC (val)
CTAP [13]	-		73.2
BSN [26]	32.2		74.2
MGG [27]	-		75.5
BMN [25]	-		75.0
BSN++ [38]	34.3		76.5
TCANet [32]	34.6		76.1
RTD-Net [39]	32.8		73.1
Ours	34.9	76.5	68.3

Results on ActivityNet-v1.3. Table 3 and Table 5 shows comparison results of the proposed TAPG Transformer with other methods on ActivityNet-v1.3. For the temporal action proposal generation task, as shown in Table 3, the performance of TAPG Transformer again outperforms state-of-the-art proposal generation methods in terms of AR@AN with AN varying from 1 to 100 and AUC. Especially when AN equals 1, we achieve 34.9% regarding the AR metric, which indicates that top-1 proposal has high

Table 4: Ablation study on THUMOS14. We show mAP (%) at different tIoU thresholds.

(a) **Study on different combinations of the components in TAPG Transformer and BMN.** BTR denotes Boundary Transformer and PTR denotes Proposal Transformer.

Model	Module	0.3	0.4	0.5	0.6	0.7
BMN	TEM	41.9	36.4	27.8	21.6	12.7
BMN	TEM+PEM	56.0	47.4	38.8	29.7	20.5
Ours	BTR	43.8	37.7	29.9	20.8	12.9
Ours	BTR + PEM	60.6	54.5	46.3	37.6	27.1
Ours	TEM + PTR	62.9	57.6	47.6	37.5	25.3
Ours	BTR + PTR	65.3	59.3	50.8	40.1	28.5

(c) **Study on different query forms of transformer decoder.**

Query	0.3	0.4	0.5	0.6	0.7
Learnable parameters	62.1	54.1	45.8	34.8	23.2
Encoder output	64.2	57.4	48.7	39.1	26.8
Original Feature	65.3	59.3	50.8	40.1	28.5

(b) **Study on sparse sampling mechanism and different window groups.** [start:stride:end] means that the sliding window size varies from start to end with stride.

Model	Number	window group	0.3	0.4	0.5	0.6	0.7
BMN	16384	Densely	56.0	47.4	38.8	29.7	20.5
BMN	1225	Fibonacci	58.2	50.5	41.2	32.9	23.3
Ours	800	[2:5:62]	63.5	57.6	49.8	40.5	29.4
Ours	1218	[2:3:62]	64.6	58.5	50.0	40.4	28.1
Ours	1726	[2:2:64]	65.2	59.1	50.8	41.0	28.8
Ours	1225	Fibonacci	65.3	59.3	50.8	40.1	28.5

(d) **Study on different numbers of TAPG Transformer layers.**

M	0.3	0.4	0.5	0.6	0.7
1	64.8	59.0	50.5	40.0	27.9
3	65.3	59.3	50.8	40.1	28.5
6	64.2	57.6	49.2	38.9	26.4

Table 5: Performance comparison with state-of-the-art action detection methods on validation set of ActivityNet-1.3, in terms of mAP (%) at different tIoU thresholds and the average mAP.

Method	0.5	0.75	0.95	Average
Singh et al. [37]	34.5	-	-	-
SCC [17]	40.0	17.9	4.7	21.7
CDC [34]	45.3	26.0	0.20	23.8
R-C3D [52]	26.8	-	-	-
BSN [26]	46.5	30.0	8.0	30.0
BMN [25]	50.1	34.8	8.3	33.9
GTAD [53]	50.4	34.6	9.0	35.1
BSN++ [38]	51.3	35.7	8.3	34.9
TCANet w/ BSN [32]	51.9	34.9	7.5	34.4
RTD-Net [39]	46.4	30.5	8.6	30.5
Ours	53.2	35.5	10.1	35.4

quality. For the temporal action detection task, as summarized in Table 5, our method achieves notable improvements on mAP over other proposal generation methods such as BSN [26], BMN [25] and G-TAD [53] at all tIoU thresholds. When tIoU is 0.95, the mAP we obtain is 10.1%, indicating that the boundaries of the generated proposals by TAPG Transformer are more precise.

4.4 Ablation Study

In this section, we conduct ablation studies on THUMOS14 to verify the effectiveness of each module in TAPG Transformer.

Effectiveness of Transformers. We perform ablation studies under different architecture settings and verify the effectiveness of the Boundary Transformer (BTR) and Proposal Transformer (PTR), separately. Table 4a demonstrates that both modules improve the performance individually. Temporal Evaluation Module (TEM) and Proposal Evaluation Module (PEM) are two modules of BMN [25]. As can be seen, "BTR + PEM" outperforming "TEM + PEM" indicates that BTR achieves more accurate boundaries, as shown in the second and fourth rows of Table 4a, while "TEM + PTR"

outperforming "TEM + PEM" demonstrates that PTR improves the proposal confidence, as shown in the second and fifth rows of Table 4a. To summarize, the evaluation results shown in Table 4a demonstrate that: 1) Compared with the previous works which only explore the local details, Boundary Transformer effectively captures the long-term information for accurate boundary prediction. 2) Instead of treating the proposals separately, Proposal Transformer effectively learns the relation of the proposals on different scales to generate high quality proposal confidence.

Effectiveness of Sparse Sampling mechanism. We perform ablation studies to verify the effectiveness of sparse proposal sequence and different sampling methods. Generating dense BM map to predict proposal confidence as done in [25] causes unbalanced positive/negative proposals and resource waste. As shown in the first two rows of Table 4b, generating less proposals (about only 7.5% of the BM maps) still achieves performance improvement. Meanwhile we design two ways to generate window group. The first is using different window sizes with equal stride in each group. The third to fifth rows in Table 4b shows that generating sliding windows with the stride set to 5, 3, and 2 all obtain good results. The second is using different window sizes with different distances in each group, specifically, we apply the Fibonacci sequence to generate candidate proposals. Satisfying performance is achieved as shown in the last row of Table 4b. In summary, the proposed TAPG Transformer achieves better performance with fewer proposals.

Effectiveness of Query Form of Transformer Decoder. We perform ablation studies to explore the performance of Transformer Decoder with different query forms. Long-term modeling easily leads to the smooth boundary of action and introduces global noise. To solve this problem, we design novel query forms of transformer decoder that are different from original Transformer. Specifically, we study on three query forms including learnable parameters [5], encoder output and original feature. The experimental results shown in Table 4c demonstrate that the query form of original feature achieves the best performance. We believe the reason is that introducing original feature helps eliminate the effects of smooth boundary of action and the global noise.

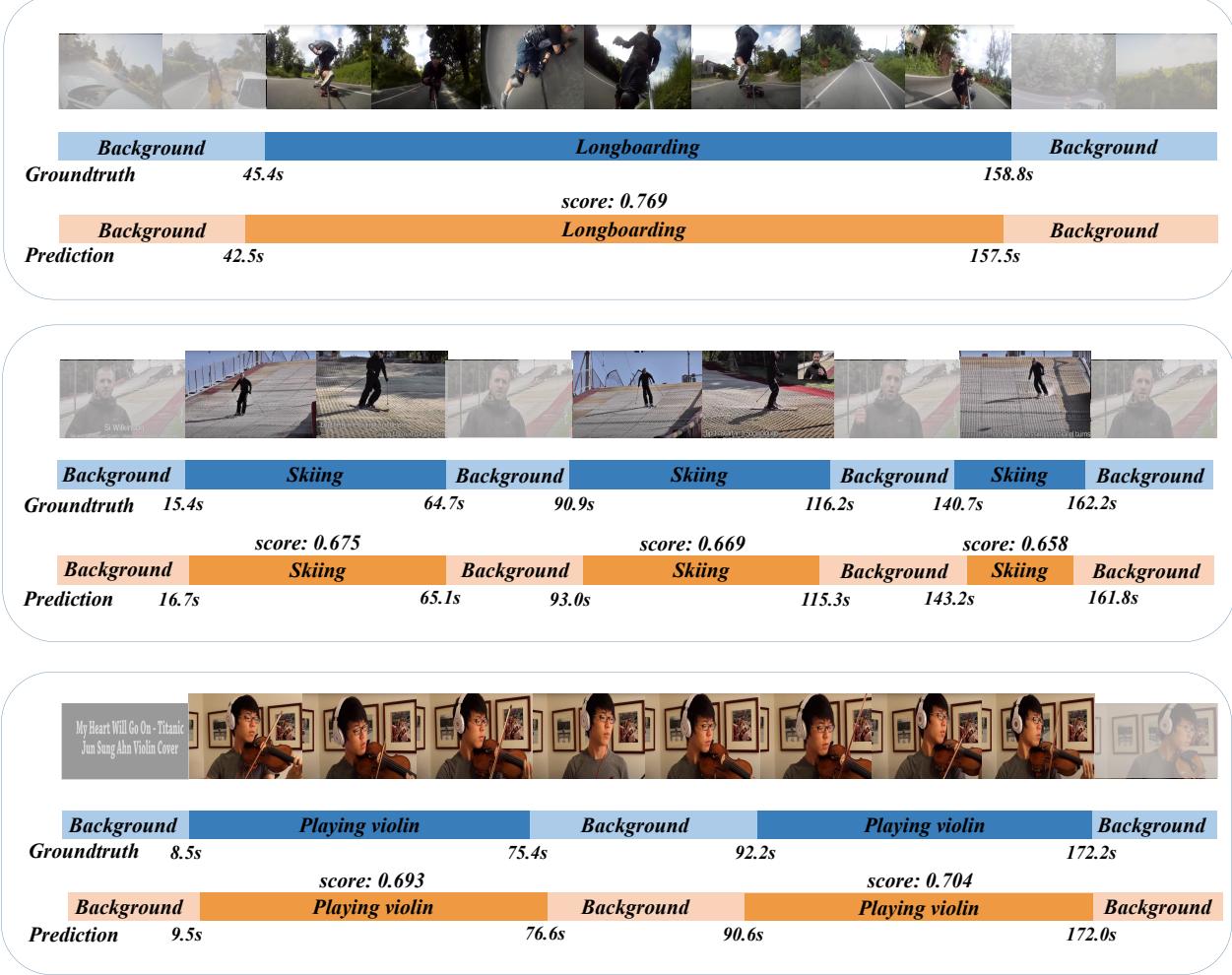


Figure 3: Visualization examples of generated proposals on ActivityNet1.3.

Effectiveness of Layers of Transformer. Stacking multi-layer transformers normally brings better performance, which has been proved in multiple tasks, e.g., NLP [43] and object detection [5]. We study on the different number of the stack of multiple layers in Boundary Transformer and Proposal Transformer. As shown in Table 4d, we achieve the best performance when stacking three Transformer modules. However, stacking an increasing number of layers does not necessarily lead to performance improvement, and the stack of excessive layers may lead to over-fitting as well, as shown in the last row of Table 4d.

4.5 Visualization

Figure 3 presents visualization examples of some representative and challenging cases on ActivityNet1.3. The proposals with the highest k scores are visualized in each video, where k is the number of ground truth. In the top video, although there exist scenes of falling and the first-person perspective, our method predicts the video correctly as a continuous action. The correct prediction shows that

our method understands the inter-frame relation and inter-proposal relation properly and does not overfit the noisy frames. The middle video includes multiple ground truth action instances, while our top-3 proposals match with them accurately, indicating the high quality of our generated proposals. In the bottom video, where the background frames and action frames are of a similar scene, our top-2 proposals are successfully aligned with the positions of ground truth action instances. Our method considers the high-level semantic information and can identify the noisy snippets even when they have a very similar appearance to the action. More examples and comparison results with the previous methods are shown in the supplementary material.

5 CONCLUSION

In this paper, we propose a novel temporal action proposal generation framework with transformers, which consists a Boundary Transformer and a Proposal Transformer. Boundary Transformer captures long-term temporal context and Proposal Transformer

models inter-proposal relation. Our TAPG Transformer can effectively enhance video understanding and noisy action instance localization. Extensive experiments show that our model achieves new state-of-the-art performance in temporal action proposal generation and action detection on THUMOS14 and ActivityNet1.3 datasets.

REFERENCES

- [1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is Space-Time Attention All You Need for Video Understanding? *arXiv preprint arXiv:2102.05095* (2021).
- [2] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. 2017. Soft-NMS—improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*. 5561–5569.
- [3] Shyamal Buch, Victor Escorcia, Bernard Ghanem, Li Fei-Fei, and Juan Carlos Niebles. 2017. End-to-End, Single-Stream Temporal Action Detection in Untrimmed Videos. In *BMVC 2017*.
- [4] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*. 961–970.
- [5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*. Springer, 213–229.
- [6] Joao Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*.
- [7] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. 2018. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1130–1139.
- [8] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. 2021. Transformer Tracking. *arXiv preprint arXiv:2103.15436* (2021).
- [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
- [10] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. 2016. Daps: Deep action proposals for action understanding. In *European Conference on Computer Vision*. Springer, 768–784.
- [11] Christoph Feichtenhofer. 2020. X3D: Expanding Architectures for Efficient Video Recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.* 203–213.
- [12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slow-fast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6202–6211.
- [13] Jiyang Gao, Kan Chen, and Ram Nevatia. 2018. Ctap: Complementary temporal action proposal generation. In *Proceedings of the European conference on computer vision (ECCV)*. 68–83.
- [14] Jialin Gao, Zhixiang Shi, Jian Li, Guanshuo Wang, Yufeng Yuan, Shimeng Ge, and Xi Zhou. 2020. Accurate Temporal Action Proposal Generation with Relation-Aware Pyramid Network. *arXiv preprint arXiv:2003.04145* (2020).
- [15] Jiyang Gao, Zhenheng Yang, Chen Sun, Kan Chen, and Ram Nevatia. 2017. Turn tap: Temporal unit regression network for temporal action proposals. In *ICCV*.
- [16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [17] F Caba Heilbron, Wayner Barrios, Victor Escorcia, and Bernard Ghanem. 2017. SCC: Semantic context cascade for efficient action detection. In *CVPR*.
- [18] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift.
- [19] Yu-Gang Jiang, Jingren Liu, A Roshan Zamir, George Toderici, Ivan Laptev, Mubarak Shah, and Rahul Sukthankar. 2014. THUMOS challenge: Action recognition with a large number of classes.
- [20] Bruno Korbar, Du Tran, and Lorenzo Torresani. 2019. Sscampler: Sampling salient clips from video for efficient action recognition. In *Int. Conf. Comput. Vis.* 6232–6242.
- [21] Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. 2021. Colorization transformer. *arXiv preprint arXiv:2102.04432* (2021).
- [22] Yan Li, Bin Ji, Xintian Shi, Jianguo Zhang, Bin Kang, and Limin Wang. 2020. TEA: Temporal Excitation and Aggregation for Action Recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.* 909–918.
- [23] Chuming Lin, Jian Li, Yabiao Wang, Ying Tai, Donghao Luo, Zipeng Cui, Chengjie Wang, Jilin Li, Feiyue Huang, and Rongrong Ji. 2020. Fast Learning of Temporal Action Proposal via Dense Boundary Generator.. In *AAAI*. 11499–11506.
- [24] Ji Lin, Chuang Gan, and Song Han. 2019. TSM: Temporal Shift Module for Efficient Video Understanding. In *Int. Conf. Comput. Vis.*
- [25] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. 2019. Bmn: Boundary-matching network for temporal action proposal generation. In *ICCV*. 3889–3898.
- [26] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. 2018. Bsn: Boundary sensitive network for temporal action proposal generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 3–19.
- [27] Yuan Liu, Lin Ma, Yifeng Zhang, Wei Liu, and Shih-Fu Chang. 2019. Multi-granularity generator for temporal action proposal. In *CVPR*. 3604–3613.
- [28] Zhaoyang Liu, Donghao Luo, Yabiao Wang, Limin Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Tong Lu. 2020. TEINet: Towards an Efficient Architecture for Video Recognition.. In *AAAI*. 11669–11676.
- [29] Weian Mao, Yongtao Ge, Chunhua Shen, Zhi Tian, Xinlong Wang, and Zhibin Wang. 2021. TFPose: Direct Human Pose Estimation with Transformers. *arXiv preprint arXiv:2103.15320* (2021).
- [30] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. 2021. Video transformer network. *arXiv preprint arXiv:2102.00719* (2021).
- [31] Dan Oneata, Jakob Verbeek, and Cordelia Schmid. 2014. The lear submission at thumos 2014. (2014).
- [32] Zhiwu Qing, Haisheng Su, Weihao Gan, Dongliang Wang, Wei Wu, Xiang Wang, Yu Qiao, Junji Yan, Changxin Gao, and Nong Sang. 2021. Temporal Context Aggregation Network for Temporal Action Proposal Refinement. *arXiv preprint arXiv:2103.13141* (2021).
- [33] Zhaofan Qiu, Ting Yao, and Tao Mei. 2017. Learning spatio-temporal representation with pseudo-3d residual networks. In *Int. Conf. Comput. Vis.* 5533–5541.
- [34] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. 2017. CDC: convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In *CVPR*.
- [35] Zheng Shou, Dongang Wang, and Shih-Fu Chang. 2016. Temporal action localization in untrimmed videos via multi-stage cnns. In *CVPR*.
- [36] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. In *NIPS*. 568–576.
- [37] Gurkirt Singh and Fabio Cuzzolin. 2016. Untrimmed video classification for activity detection: submission to activitynet challenge. *arXiv preprint arXiv:1607.01979* (2016).
- [38] Haisheng Su, Weihao Gan, Wei Wu, Junjie Yan, and Yu Qiao. 2020. BSN++: Complementary Boundary Regressor with Scale-Balanced Relation Modeling for Temporal Action Proposal Generation. *arXiv preprint arXiv:2009.07641* (2020).
- [39] Jing Tan, Jiaqi Tang, Limin Wang, and Gangshan Wu. 2021. Relaxed Transformer Decoders for Direct Action Proposal Generation. *arXiv preprint arXiv:2102.01894* (2021).
- [40] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Int. Conf. Comput. Vis.*
- [41] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. 2019. Video classification with channel-separated convolutional networks. In *Int. Conf. Comput. Vis.* 5552–5561.
- [42] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.*
- [43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.
- [44] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. 2017. Untrimmednets for weakly supervised action recognition and detection. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 4325–4334.
- [45] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*. Springer, 20–36.
- [46] Wenhao Wu, Dongliang He, Tianwei Lin, Fu Li, Chuang Gan, and Errui Ding. 2021. MVFNet: Multi-View Fusion Network for Efficient Video Recognition. In *AAAI*.
- [47] Wenhao Wu, Dongliang He, Xiao Tan, Shifeng Chen, and Shilei Wen. 2019. Multi-Agent Reinforcement Learning Based Frame Sampling for Effective Untrimmed Video Recognition. In *Int. Conf. Comput. Vis.*
- [48] Wenhao Wu, Dongliang He, Xiao Tan, Shifeng Chen, Yi Yang, and Shilei Wen. 2020. Dynamic Inference: A New Approach Toward Efficient Video Action Recognition. In *Proceedings of CVPR Workshops*. 676–677.
- [49] Zuxuan Wu, Caiming Xiong, Chih-Yao Ma, Richard Socher, and Larry S Davis. 2019. Adaframe: Adaptive frame selection for fast video recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.* 1278–1287.
- [50] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. 2018. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Eur. Conf. Comput. Vis.*
- [51] Yuanjun Xiong, Limin Wang, Zhe Wang, Bowen Zhang, Hang Song, Wei Li, Dahua Lin, Yu Qiao, Luc Van Gool, and Xiaoou Tang. 2016. Cuhk & ethz & siat

- submission to activitynet challenge 2016. *arXiv preprint arXiv:1608.00797* (2016).
- [52] Huijuan Xu, Abir Das, and Kate Saenko. 2017. R-c3d: Region convolutional 3d network for temporal activity detection. In *ICCV*.
 - [53] Mengmeng Xu, Chen Zhao, David S Rojas, Ali Thabet, and Bernard Ghanem. 2020. G-TAD: Sub-Graph Localization for Temporal Action Detection. In *CVPR*. 10156–10165.
 - [54] Yihong Xu, Yutong Ban, Guillaume Delorme, Chuang Gan, Daniela Rus, and Xavier Alameda-Pineda. 2021. TransCenter: Transformers with Dense Queries for Multiple-Object Tracking. *arXiv preprint arXiv:2103.15145* (2021).
 - [55] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. 2020. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5791–5800.
 - [56] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. 2015. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4694–4702.
 - [57] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. 2017. Temporal action detection with structured segment networks. In *ICCV*. 2914–2923.
 - [58] Tianyu Zhu, Markus Hiller, Mahsa Ehsanpour, Rongkai Ma, Tom Drummond, and Hamid Rezatofighi. 2021. Looking Beyond Two Frames: End-to-End Multi-Object Tracking Using Spatial and Temporal Transformers. *arXiv preprint arXiv:2103.14829* (2021).

Supplementary Material for Temporal Action Proposal Generation with Transformers

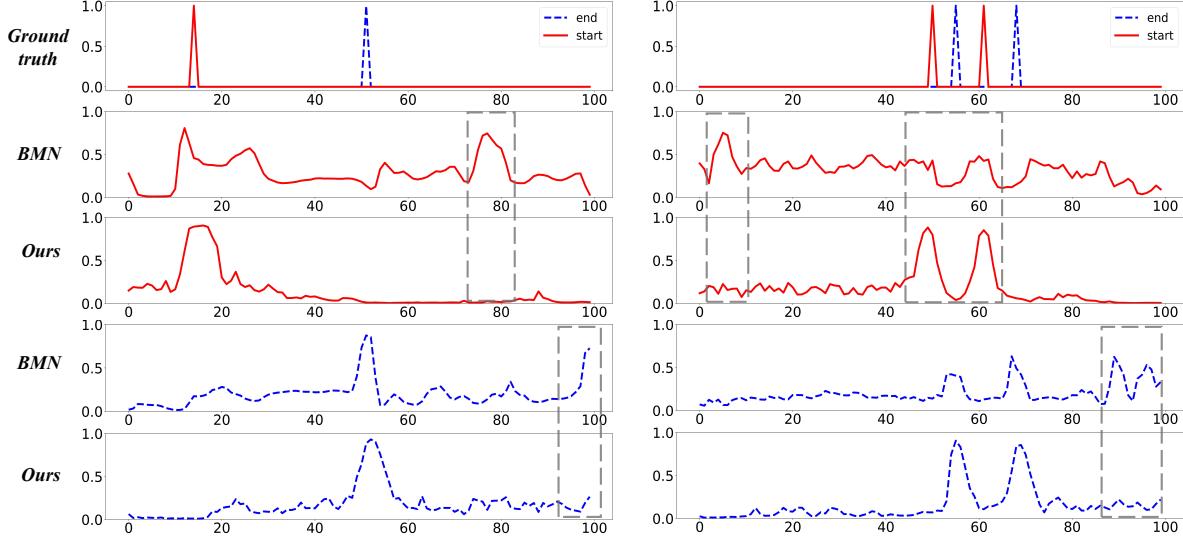


Figure 1: Visualization examples of starting and ending probability sequences on ActivityNet1.3.

1 LABEL ASSIGNMENT

The method of label assignment in each loss functions are described in detail below. For Boundary Transformer, we need to generate temporal boundary label sequence $G_S, G_E \in R^T$. Following [?], for a ground-truth action instance $\varphi_g = (t_s, t_e)$ with duration $d_g = t_e - t_s$ in annotation set Ψ_w , we denote its starting and ending regions as $r_S = [t_s - d_g/10, t_s + d_g/10]$ and $r_E = [t_e - d_g/10, t_e + d_g/10]$ separately. Then, for a temporal location t_n within F_w , we denote its local region as $r_{t_n} = [t_n - d_f/2, t_n + d_f/2]$, where $d_f = t_n - t_{n-1}$ is the temporal interval between two locations. Then we calculate overlap ratio IoR of r_{t_n} with r_S and r_E separately, and denote maximum IoR as $g_{t_n}^S$ and $g_{t_n}^E$ separately, where IoR is defined as the overlap ratio with ground truth proportional to the duration of this region. Thus we can generate $G_{S,w} = \{g_{t_n}^S\}_{n=1}^{L_w}$ and $G_{E,w} = \{g_{t_n}^E\}_{n=1}^{L_w}$ as label of Boundary Transformer. For Proposal Transformer, we need to generate **Sparse Proposal sequence** label $G_C \in R^L$, where L is number of sparse proposals. For a proposal $\varphi_i \in P$, we calculate its Intersection-over-Union (IoU) with all φ_g in Ψ_w , and denote the maximum IoU as g_i^C . Thus we can generate $G_C = \{g_i^C\}_{i=1}^L$ as label of Proposal Transformer.

2 MORE VISUALIZATION EXAMPLES

Figure 1 illustrates two visualization examples and comparison with classical method BMN [?]. The example on the left shows that the BTR module can significantly reduce the adverse effects of boundary noise. The other example indicates that BTR can enhance

the boundary awareness. These two remarkable advantages benefit from the ability of BTR to capture the long-term frame-level dependencies.

Figure 2 illustrates more visualization examples on ActivityNet1.3. The generated proposals of our method with the highest k scores are visualized in each video, where k is the number of ground truth. The scenarios of all examples are relatively complicated. In the first and second videos which contain complex action scenes, our method predicts the various scenes correctly as a complete action. For instance, the action of wrapping presents contains various scenes including putting the present in the wrapping paper, packing the present and so on. The correct predictions show that our method understands the inter-frame relation and inter-proposal relation properly. In the third and fourth videos, where the background frames and action frames are of the similar scene, our top-2 proposals are successfully aligned with the positions of ground truth action instances. Our method is able to identify the noisy snippets even when they have very similar appearance to the action. In the bottom video, which contains complex background scenes, our model correctly understands the semantic information and identify the true action. For instance, the background of the platform diving action contains various scenes including audience cheering, the diving platform scene without competition going on and so on.

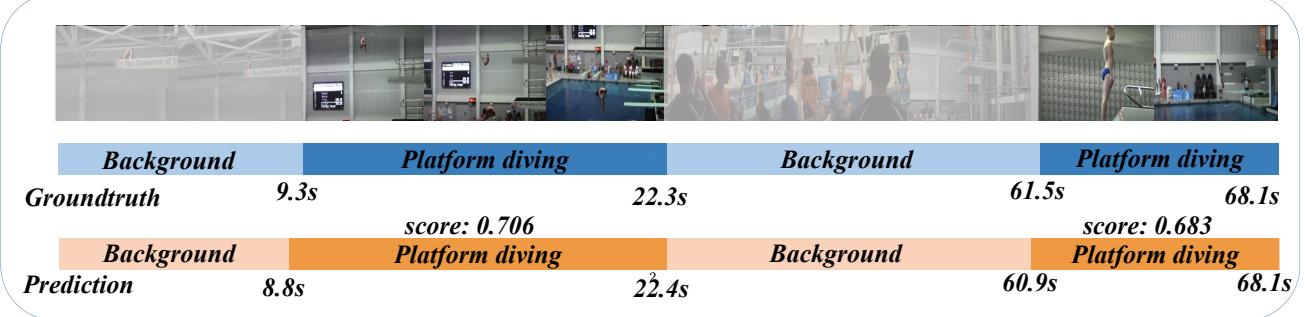
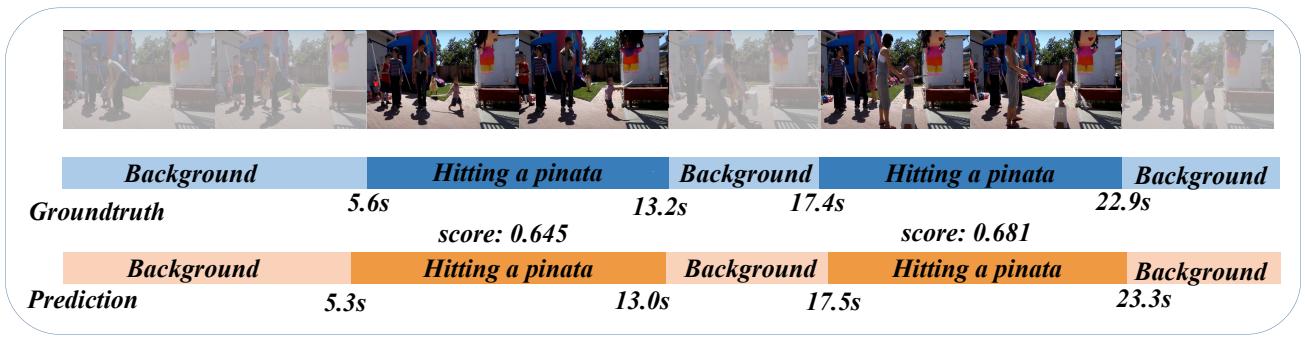
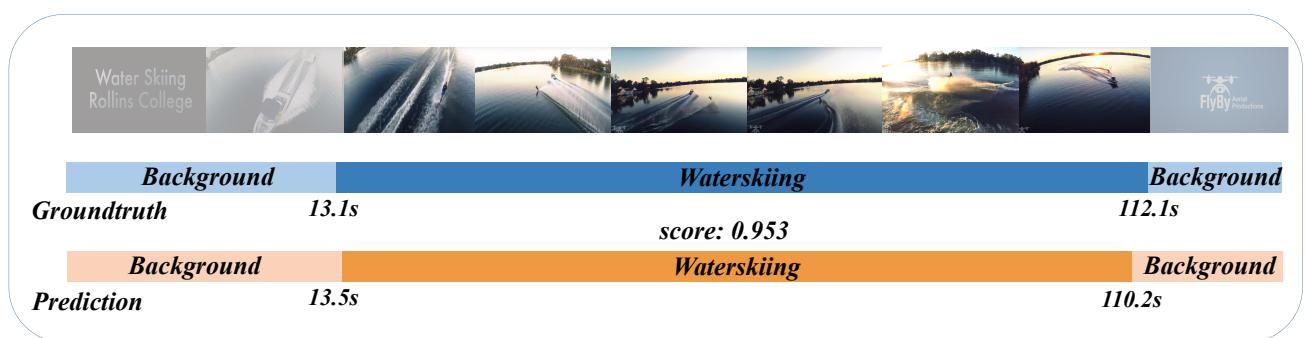
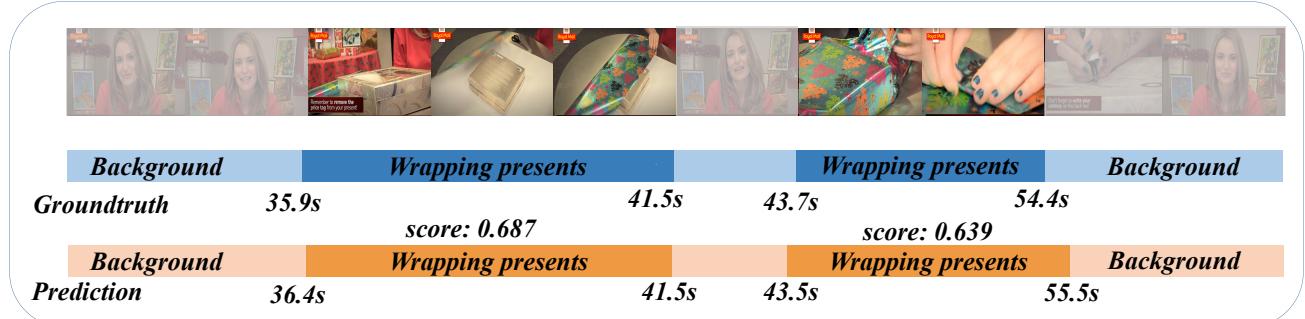


Figure 2: Visualization examples of generated proposals on ActivityNet1.3.