TSI: Temporal Scale Invariant Network for Action Proposal Generation

Shuming Liu[1], Xu Zhao[1(✉)], Haisheng Su[1], and Zhilan Hu[2]

[1] Department of Automation, Shanghai Jiao Tong University, Shanghai, China
{shumingliu,zhaoxu}@sjtu.edu.cn
[2] The Central Media Technology Institute of Huawei Co., Ltd., Shenzhen, China

Abstract. Despite the great progress in temporal action proposal generation, most state-of-the-art methods ignore the impact of action scales and the performance of short actions is still far from satisfaction. In this paper, we first analyze the sample imbalance issue in action proposal generation, and correspondingly devise a novel scale-invariant loss function to alleviate the insufficient learning of short actions. To further achieve proposal generation task, we adopt the pipeline of boundary evaluation and proposal completeness regression, and propose the **Temporal Scale Invariant network**. To better leverage the temporal context, boundary evaluation module generates action boundaries with high-precision-assured global branch and high-recall-assured local branch. Simultaneously, the proposal evaluation module is supervised with introduced scale-invariant loss, predicting accurate proposal completeness for different scales of actions. Comprehensive experiments are conducted on ActivityNet-1.3 and THUMOS14 benchmarks, where TSI achieves state-of-the-art performance. Especially, AUC performance of short actions is boosted from 36.53% to 39.63% compared with baseline.

1 Introduction

As an important and fundamental video understanding task, temporal action detection has attracted extensive attention recently. Akin to object detection, detecting action clips in a given untrimmed video can be divided into two stages: temporal action proposal generation and proposal classification. For action proposal generation task, the start and end time of real action instances in the video need to be temporally localized. Action proposal generation is extremely useful for many advanced video understanding tasks, such as action recognition, video captioning, spatio-temporal action localization, and so forth.

Previous proposal generation methods can be mainly categorized as three frameworks. The first one follows the ***top-down*** pathway. By utilizing sliding window or anchor mechanism, a large number of default proposals are generated

This work has been supported in part by the funding from NSFC (61673269, 61273285), Huawei cooperative project and the project funding of the Institute of Medical Robotics at Shanghai Jiao Tong University.

H. Ishikawa et al. (Eds.): ACCV 2020, LNCS 12626, pp. 530–546, 2021.
https://doi.org/10.1007/978-3-030-69541-5_32

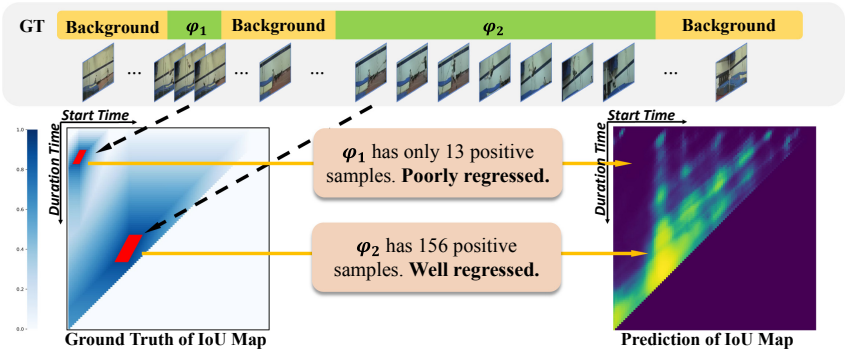


Fig. 1. What's the impact of action's temporal scale in proposal generation? For an untrimmed video with two actions φ_1 and φ_2, current proposal confidence prediction module would regress a promising score for long action φ_2 yet miss the short action φ_1. This problem is caused by the imbalance of positive samples for different actions.

densely, which are designed to cover different duration ground truth. Then these redundant proposals are revised by offset prediction and confidence regression, such as [1–6]. The second framework takes ***bottom-up*** methodology, where the temporal feature sequence is firstly used for boundary detection and actionness evaluation, and the proposals are explicitly formed by pairing the start and end points. Then, proposals are also refined by confidence regression, such as BSN [7] and TSA [8]. In the third framework, to combine the advantage of both bottom-up and top-down methods, boundary detection and dense confidence regression are performed simultaneously by using ROI align. This complementary framework obtains impressive results in BMN [9] and DBG [10].

Despite the remarkable progress achieved in action proposal generation, there are still many issues remain unsolved. Among them, how to deal with the ***scale*** change in temporal dimension is a long-standing problem. As shown in Fig. 1, in an untrimmed video with two ground truth actions, the shorter action is prone to be missed in completeness prediction, which is reflected as the extreme low recall compared to long actions in Table 5. We delve deep into this phenomenon, and conclude that the ignorance of short actions can be caused by the unbalanced positive sample distribution. Another bottleneck that limits performance gains is the module of boundary detection. Current methods mainly focus on local information and low-level features, however the critical global context is missed when determining the action boundaries. Local-global combination is an intuitive and promising direction to widen this bottleneck.

To address the aforementioned issues, we first analyze the sample-imbalance problem in action proposal generation, and correspondingly propose a general scale-invariant loss function for confidence regression, which can evidently enhance the detection ability for short actions. Furthermore, in order to achieve complete action proposal generation, we combine the bottom-up and top-down pathways, and introduce our **Temporal Scale Invariant network (TSI)**.

To be specific, TSI novelly adopts a multi-branch temporal boundary detector to capture action boundaries with both high recall and high precision. Simultaneously, IoU map regressor, supervised by the proposed scale-invariant loss function, is able to regress accurate confidence score especially for short actions. The main contributions of this work are summarized as:

1. Centered on the temporal scale issue, we analyze the sample-imbalance phenomena behind it, and accordingly devise a scale-invariant loss function to improve the detection performance on short actions.
2. To achieve the complete action proposal generation, besides handling the scale issue, TSI also takes advantage of temporal context for boundary detection with local-global-complementary structure to enhance the performance.
3. Comprehensive experiments are conducted on THUMOS14 and ActivityNet benchmarks. Results show that TSI outperforms other state-of-the-art action proposal generation methods and achieves AUC of 68.35% on ActivityNet.

2 Related Work

Temporal action detection can be grouped into two types of methods: the first type is "one-stage" method that intends to localize the actions and predict its category simultaneously. The other type is "two-stage" method, which follows the pipeline of "detection by classifying proposals".

Temporal Action Detection. The advantage of one-stage method is to naturally avoid sub-optimization for action localization and classification. For example, akin to SSD in object detection, SSAD [3] defines multi-scale anchors and uses temporal convolution to extract corresponding contextual features for offset regression and category prediction. What's more, GTAN [11] uses Gaussian kernels to model the temporal structure, which can dynamically optimize the temporal scale of each proposal. Besides, P-GCN [12] and G-TAD [13] exploits proposal-proposal relations and temporal-temporal relations by graph convolution networks and achieves significant performance gains.

Temporal Action Proposal Generation. The motivation of two-stage method is the success of video classification task for a given trimmed video [14–18]. Therefore, how to localize possible action instance with precise boundary and high overlap in long untrimmed video becomes the key issue in action detection. The mainstream of top-down action proposal generation methods would first initiate a default proposal set, which is often predefined by clustering ground truth actions, and then revise them with confidence regression [5,10,19–24]. For example, RapNet [6] introduces a relation-aware module to exploit long-range temporal relations and follows a two-stage adjustment scheme to refine the proposal boundaries and measure their confidence. As for bottom-up methods [8], TAG [2] designs temporal watershed algorithm to generate proposals, yet missing the regression for proposal confidence. Considering the boundary information, BSN [7] firstly utilizes temporal evaluation module to predict the starting

and ending probabilities, and uses proposal evaluation module to regress the confidence score. To take advantage of both bottom-up and top-down method, MGG [25] first attempts to embed position information and generate proposals from different granularities. Improved from BSN, BMN [9] develops boundary matching mechanism to regress the confidence of all potential proposals. To further regress densely distributed proposals, DBG [10] propose an unified framework to achieve boundary classification and action completeness regression.

Although the great progress in action detection, the long-standing problem of temporal scale variation still has not been pertinently studied, which is the main motivation of this paper.

3 Our Approach

3.1 Problem Definition and Video Representation

Given an untrimmed video X, the temporal proposal annotation is denoted as $\Psi_g = \{\varphi_i = [t_{s,i}, t_{e,i}]\}_{i=1}^{N_g}$, where N_g is the number of ground truth, and $[t_{s,i}, t_{e,i}]$ is the start and end time of action instance φ_i. The aim of temporal action proposal generation is to predict candidate proposal set $\Psi_p = \{\varphi_i = [t_{s,i}, t_{e,i}, s_i]\}_{i=1}^{N_p}$ to cover Ψ_g with high recall and high overlap, where s_i is the confidence score of predicted φ_i and will be used for proposal ranking.

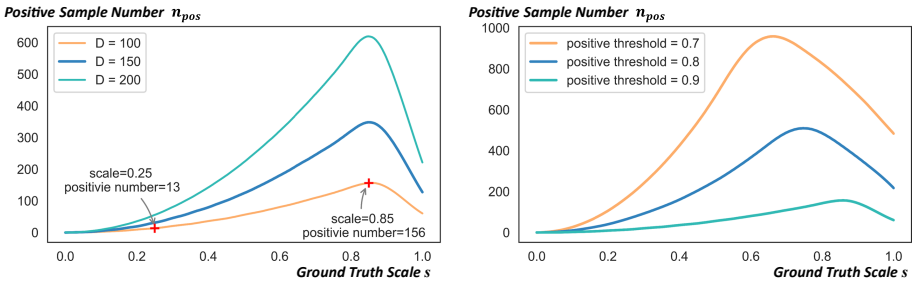
Following previous work [7,9,22,26], we adopt two-stream network [14] to encode the raw RGB frames and optical flow of video X into representative video feature sequence $F_0 \in \mathbb{R}^{C \times T}$, where C is the fixed feature channel and T is the video feature length. Then we rescale the feature sequence to length D by linear interpolation and eventually obtain the feature $F \in \mathbb{R}^{C \times D}$, as the input of action proposal generation network.

It is worth noticing that, in BMN, DBG and other methods, a proposal is considered as the composition of a start point and an end point, which are both selected from D quantified time index. Therefore, the upper limit number of candidate proposals can be calculated as $N = \binom{D}{2} = \frac{D(D-1)}{2}$, meaning the algorithm need to retrieve real actions from these N candidates.

3.2 Scale-Imbalance Analysis in Proposal Generation

As introduced before, short actions are prone to be missed in confidence regression. By contrast, AUC can decrease significantly from 94.48% of long actions to 36.53% of short actions in state-of-the-art baseline BMN (see Table 5). In fact, the inferior detection ability of short actions can be interpreted as many folds, such as the insufficient feature representation with limited granularity, the stringent boundary overlap requirements due to the IoU evaluation metrics. In addition to above reasons, **the unbalanced training** also leads to the overwhelming learning of large actions but severe weak learning for short actions.

To discuss this issue, we need to clarify the definition of action scale first. The scale of an action s is regarded as *the ratio of action length to video length*, thus,



(a) set positive threshold = 0.9 (b) set rescale length D = 100

Fig. 2. The distribution of positive sample numbers with action scale. Take D equals to 100 in (a) for example, a long action with $s = 0.85$ will have 156 positive samples while a short action with $s = 0.25$ has only 13 positive samples. This sample imbalance causes severe weak learning for short actions but excessive learning for long actions.

s should belong to $(0, 1)$. Now we inspect a video with two ground truth actions. By computing the IoU between GT actions with aforementioned N proposals, the IoU map is obtained as shown in Fig. 1 left. In this map, point(i, j) represents the maximum IoU between GTs with proposal(i, j) (following the definition in BMN, proposal(i, j) indicates a proposal with duration time i and start time j). Therefore IoU values around GT are closer to 1 and should be considered as *high quality proposals*. However, as visualized in Fig. 1, the area of high quality proposals of long action φ_2 is much larger than short action φ_1, which reminds us: **Is the short action overlooked in such dense regression mechanism?**

The answer is Yes. No matter what loss function we choose in IoU regression, for example binary logistic loss used in BMN and L2 loss in DBG, positive samples and negative samples need to be defined first. Normally, a proposal with its $IoU > \varepsilon$ is regarded as positive, where ε is a predefined threshold. Thus, we can use sampling or reweight methods to balance the positive/negative samples. However, inside the positive samples, with the change of action scale s, the number of positive samples n_{pos} of each ground truth would vary significantly, as shown in Fig. 2. Take Fig. 2(a) for instance, when $D = 100$, an action with scale 0.85 has 10x positive samples than the action with scale 0.25. Consequently, the short action with less positive samples can not be learned adequately.

To address above problem, the loss function of confidence regression must satisfy two conditions (1) the contribution of each ground truth should be equal considering the n_{pos} (2) the positive/negative samples should be balanced appropriately. To this end, we propose the **scale-invariant loss** (SI-Loss) as Eq. 1.

$$L_{SI} = \sum\nolimits_{i,j} b_{i,j} w_{i,j} \log\left(p_{i,j}\right) + \left(1 - b_{i,j}\right) w_{i,j} \log\left(1 - p_{i,j}\right) \tag{1}$$

$$w_{i,j} = \begin{cases} (1-\alpha)/n_{pos,c} & \text{if } b_{i,j} = 1 \\ \alpha/(N - \sum_c n_{pos,c}) & \text{if } b_{i,j} = 0 \end{cases} \tag{2}$$

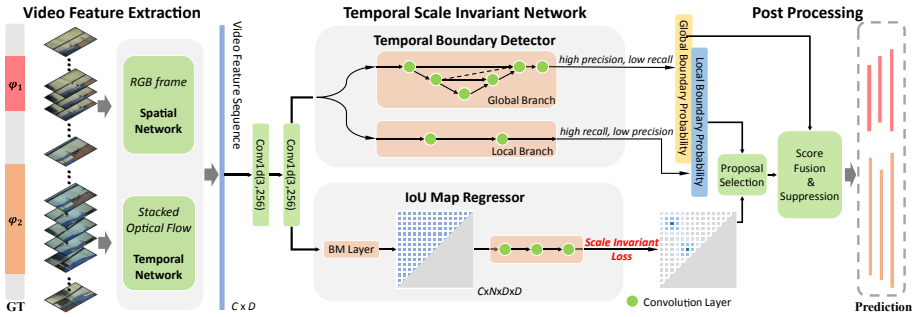


Fig. 3. The framework of our method. TSI contains two modules: temporal boundary detector (TBD) and IoU map regressor (IMR). In TBD, local branch focuses on local information and generates high-recall-assured boundaries, while U-shaped global branch distills contextual features and provides high precision-assured boundaries. Meanwhile, IMR densely regresses the completeness of potential proposals supervised with scale-invariant loss, which can greatly improve the detection ability for short actions.

SI-Loss essentially is a scale-weighted binary logistic loss. In Eq. 1, $w_{i,j}$ is the weight coefficient for proposal(i, j). $b_{i,j}$ stands for the positive mask whether proposal(i, j) is a positive sample given threshold ε. To balance the change of n_{pos} in loss contribution, we define $w_{i,j}$ as following: if a proposal(i, j) is a positive sample and it belongs to annotation φ_c, we divide its loss with φ_c's total positive sample number $n_{pos,c}$, which can guarantee the aggregate positive loss of each GT the same. In this way, taking positive sample number distribution into consideration, each action in a video can be learned equally in the training loss, which achieves the scale-invariant purpose. What's more, to control the balance of positive and negative samples, hyper-parameter α is adopted in SI-Loss.

When video only contains one annotation and α takes 0.5, scale-invariant loss would degenerate into normal binary logistic loss. What's more, when α is bigger than 0.5, SI-Loss would have a higher weight on negative samples, which can reduce the false positive response. Supervised with SI-Loss, in proposal completeness regression module, the ability to retrieve small targets is greatly enhanced and its effectiveness has been proved as shown in Table 5.

3.3 Temporal Scale Invariant Network

With the scale-invariant loss, to achieve the complete action proposal generation process, we combine the bottom-up and top-down pathways and propose Temporal Scale Invariant Network. The framework of TSI can be demonstrated as Fig. 3, which contains two modules: **Temporal Boundary Detector (TBD)** and **IoU Map Regressor (IMR)**.

Temporal Boundary Detector. It is acknowledged that one of the necessary conditions for a well-performed action proposal generation method is the precise

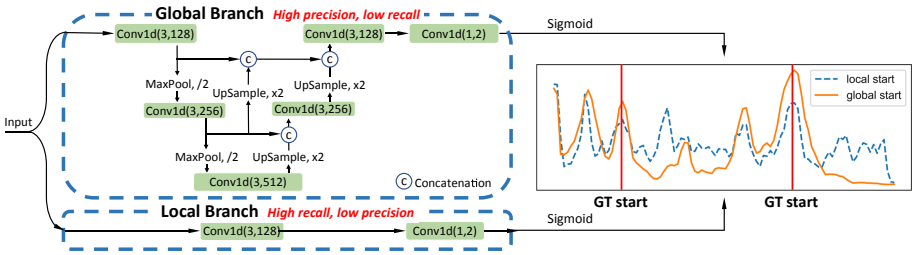


Fig. 4. TBD architecture. TBD contains local branch and global branch to detect boundaries with high precision and high recall. *c* stands for the concatenation operation.

prediction for action boundary. Conventional approaches [7,9] hold that boundary is a local information which does not require much attention on temporal context or deep semantic features, thus they both share a limited receptive field.

Such viewpoints, however, are biased as revealed in [6,8]. Actions with different scales should require corresponding receptive field, thus the boundary detection module need to be able to leverage local apparent information and global contextual semantic information in a unified framework. Taking into account of such requirements, we design a local-global complementary network named TBD to detect accurate temporal boundaries, as shown in Fig. 4.

In TBD, the local branch observes a small receptive field with only two temporal convolution layers. Therefore this branch focuses on the local abrupt change and generates a rough boundary with high recall to cover all actual start/end points, yet bringing extreme low precision. To make up this shortcoming, global branch enlarges the receptive field and presents boundaries with contextual U-shaped network, which is inspired by UNet [27]. The global branch uses multiple temporal convolution layers followed by down-sampling steps to distill semantic information of different granularity. To restore the resolution of temporal feature sequence, several up-sampling operations are repeated and features in the same resolutions are concatenated.

In Fig. 4, ***conv1d(3,128)*** represents the temporal convolution layer with kernel size 3 and output channel 128. If not stated specifically, ReLU is used for activation function. At last, 1×1 convolution with 2 channels and sigmoid function is used to generate starting and ending boundaries for both branches. To sum up, this combination of local and global structure will best leverage the low-level fine-grained features with contextual features and extract accurate boundaries with high recall and high precision.

IoU Map Regressor. Besides the bottom-up pathway of boundary evaluation, proposal confidence regression is also vital for action proposal generation. To densely regress potential proposal confidence, we adopt the Boundary-Matching mechanism in BMN [9], which can transfer temporal feature sequence $F \in \mathbb{R}^{C \times D}$ to proposal feature matrix $M_F \in \mathbb{R}^{C \times M \times D \times D}$ through BM layer.

Boundary-Matching mechanism essentially is a ROI align layer implemented in matrix product. By using such module, the completeness of all proposals can be regressed simultaneously.

For fair comparison, we follow the exact network structure of proposal evaluation module in BMN. After IoU Map Regressor, each proposal will be predicted with two confidence score, which is supervised with IoU classification loss and IoU regression loss. However, the classification loss in BMN ignores the impact of action scales that leads to the low recall of short actions. Therefore, we use the aforementioned scale-invariant loss as the IoU classification loss to enforce the network to focus on different scale actions equally.

4 Training and Inference

4.1 Training of TSI

Label Assignment. For a ground truth action $\varphi_g = [t_s, t_e]$, action starting region is defined as $r_s = [t_s - d/10, t_s + d/10]$, where $d = t_e - t_s$. Then by computing the maximum overlap ratio of each temporal interval with r_s, we can obtain $G_s = \{g_i^s\}$ as the starting label of TBD. The same label assignment process is adopted for ending label G_e. As for IMR, the label of IoU map is denoted as $G_{iou} = \{g_{i,j}\}$, which follows the definition in BMN.

Loss of TBD. The output of TBD are the starting and ending probability sequence from local and global branch, denoted as $P_{s,l}$, $P_{e,l}$, $P_{s,g}$, and $P_{e,g}$ respectively. We follow [7] to adopt binary logistic loss L_{bl} to supervise the boundary prediction with G_s, G_e, denoted as

$$L_{TBD} = \frac{1}{2}\left(L_{bl}\left(P_{s,l}, G_s\right) + L_{bl}\left(P_{e,l}, G_e\right) + L_{bl}\left(P_{s,g}, G_s\right) + L_{bl}\left(P_{e,g}, G_e\right)\right) \quad (3)$$

Loss of IMR. The output of IMR is a probability map P_{iou} with two channels. Following BMN, we construct the classification loss and regression loss as the IMR loss, where we use proposed SI-Loss as classification loss L_C and L2 loss as regression loss L_R. Especially, positive threshold ε is set as 0.9 in SI-Loss.

$$L_{IMR} = L_C(P_{iou,c}, G_{iou}) + L_R(P_{iou,r}, G_{iou}) \quad (4)$$

The training objective of TSI is the multi-task learning in the unified framework. The overall loss function contains TBD loss, IMR loss, and L2 regularization term, where λ is the weight term set to 10^{-4}:

$$L = L_{TBD} + L_{IMR} + \lambda \cdot L_2\left(\Theta\right) \quad (5)$$

4.2 Inference of TSI

Proposal Selection. To ensure the diversity of proposals and guarantee a high recall, only local branch of TBD is used for proposal selection. Following [7,9], all temporal locations satisfying (1) local peak in boundary probabilities and (2) probabilities higher than $0.5 \cdot max(P)$ are regarded as the starting and ending locations. Then we match all starting and ending locations to generate redundant candidate proposals, denoted as Ψ_p.

Score Fusion and Proposal Suppression. For each proposal(i, j) in Ψ_p, whose duration time is i, start time is j and end time is $i+j$, its IoU completeness is denoted as fusion of classification score and regression score $p_{iou} = p_{i,j,c} \cdot p_{i,j,r}$. Its starting probability is denoted as $p_{start} = \sqrt{p_{s,l}(i) \cdot p_{s,g}(i)}$, which is the same as p_{end} for ending probability. Therefore the proposal confidence score is defined as $p_f = p_{start} \cdot p_{end} \cdot p_{iou}$. Then we adopt Soft-NMS [28] to remove redundant proposals to retrieve final high quality proposals.

5 Experiments

5.1 Datasets and Settings

ActivityNet-1.3 [29] is a large-scale video understanding dataset, consisting of 19,994 videos annotated for action proposal task. The dataset is divided into training, validation and testing set with the ratio of 2:1:1.

THUMOS14 dataset contains 200 annotated untrimmed videos in validation set and 213 annotated untrimmed videos in testing set. We use the validation set to train TSI and evaluate our model on testing set.

Implementation Details. On ActivityNet dataset, rescaling length D is set to 100. On THUMOS dataset, we slide the temporal window with length 128 and overlap ratio 0.5 by following [7]. On both datasets, we use batch size of 16 and Adam optimizer to train TSI. The learning rate is set to 10^{-3} and decay it to 10^{-4} after 7 epochs. Besides, α is set to 0.8 as experimented on ablation study.

5.2 Temporal Action Proposal Generation

For action proposal generation task, Average Recall (AR) under Average Number of proposals (AN) with different IoU thresholds is the widely used evaluation metrics. Besides, the area under AR vs AN curve (AUC) is also used for comparison on ActivityNet-1.3 in our experiments.

Comparison with State-of-the-Art Methods. Table 1 illustrates the performance of our proposal generation method compared with other state-of-the-art methods on ActivityNet-1.3 dataset. The result shows that TSI outperform other methods and improves the AUC from 67.10% to 68.35% on validation set.

Table 1. Comparison between TSI and other state-of-the-art temporal action proposal generation methods on ActivityNet-1.3 in terms of AR@100 and AUC.

Method	CTAP [5]	BSN [7]	MGG [25]	BMN [9]	DBG [10]	**TSI**
AR@100(val)	73.17	74.16	74.56	75.01	**76.65**	76.31
AUC(val)	65.72	66.17	66.54	67.10	68.23	**68.35**
AUC(test)	–	66.26	66.47	67.19	68.57	**68.85**

Table 2. Comparison between TSI and other state-of-the-art temporal action proposal generation methods on test set of THUMOS14 dataset in terms of AR@AN.

Method	Feature	@50	@100	@200	@500	Feature	@50	@100	@200	@500
TURN [22]	C3D	19.63	27.96	38.34	53.52	Flow	21.86	31.89	43.02	57.63
MGG [25]	C3D	29.11	36.31	44.32	54.95	2Stream	39.93	47.75	54.65	61.36
BSN [7]	C3D	29.58	37.38	45.55	54.67	2Stream	37.46	46.06	53.21	60.64
BMN [9]	C3D	32.73	40.68	47.86	56.42	2Stream	39.36	47.72	54.70	62.07
DBG [10]	C3D	32.55	41.07	48.83	57.58	2Stream	40.89	49.24	55.76	62.21
TSI	C3D	**33.46**	**41.64**	**49.97**	**57.73**	2Stream	**42.30**	**50.51**	**57.24**	**63.43**

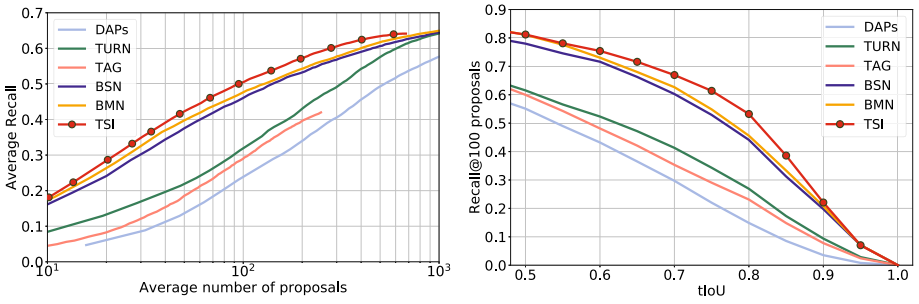


Fig. 5. Comparison between TSI and other state-of-the-art methods on test set of THUMOS14 in terms of **(left)** AR@AN **(right)** Recall@100 with different tIoU.

Especially, the AR@100 is improved from 75.01% to 76.31%, suggesting TSI can generate rich and accurate proposals.

We also implement our method on THUMOS14, as shown in Table 2. C3D feature [15] and two stream feature [14] used in BMN are adopted for fair comparison. Experiment shows that TSI outperforms other methods in all AN sets and achieves state-of-the-art performance. Figure 5 further illustrates that TSI can guarantee higher recall with fewer proposals and in terms of different tIoU.

Ablation Study. To fully confirm the effectiveness of TSI, we conduct extensive ablation experiments on our proposed method.

Table 3. Ablation study of different boundary detection modules on ActivityNet-1.3.

	BSN-TEM	BMN-TEM	TSI-TBD
AUC(val)	64.80	65.17	**66.31**
AR@100	73.57	73.72	**74.13**

Table 4. Ablation study of hyper parameter α in Scale-Invariant loss.

α	0.5	0.6	0.7	0.8	0.9
AUC(val)	67.98	68.08	68.13	**68.35**	68.33

Table 5. Ablation study of Scale-Invariant Loss with AUC performance of different action scales on ActivityNet-1.3 validation set. s stands for the scale of ground truth

Method	AUC	$0.0 \leq s < 0.06$	$0.06 \leq s < 0.65$	$0.65 \leq s \leq 1.0$
BMN	67.10	36.53	70.43	94.48
BMN+SI-Loss	**67.98**	40.24	70.32	94.41
DBG	67.90	39.07	72.18	93.08
DBG+SI-Loss	**68.23**	40.57	70.25	94.73
TSI(TBD)	66.31	36.65	68.55	94.59
TSI(TBD+IMR)	67.47	36.87	71.11	**95.20**
TSI(TBD+IMR+SI-loss)	**68.35**	**39.63**	**71.40**	94.79

Effectiveness of Temporal Boundary Detector. First, we evaluate our temporal boundary detector with other boundary-based methods. As shown in Table 3, we only use TBD without IMR to generate action proposals, which can already achieve higher AUC and recall of 66.31%, compared with other temporal evaluation module in BSN and BMN. This result proves that TBD with local-global branches can better leverage the temporal context to detect precise boundaries and well balance the recall and precision of retrieved boundary location. Note that in all comparisons, Soft-NMS is used for redundant proposal suppression.

Ablation Study of α In Scale-Invariant Loss. The hyper parameter α is the coefficient to balance the positive/negative samples. As shown in Table 4, with the increase of α, AUC is correspondingly boosted from 67.98% to 68.35%, indicating (1) network supervised with scale-invariant loss can achieve high AUC regardless of α (2) the larger α would reduce the false positive response of IoU prediction, which can improve the detection ability.

Effectiveness of Scale-Invariant Loss. To further verify the effectiveness proposed scale-invariant loss function, we conducted several ablation experiments, which is shown in Table 5. (Note: we use the same video feature of BMN on DBG and TSI, thus the result of DBG is lower than reported in their paper.)

Table 6. Generalization ability of TSI on validation set of ActivityNet-1.3 in terms of AR@100 and AUC.

BMN/TSI	*Seen*		*Unseen*	
	AR@100	AUC	AR@100	AUC
Training with *Seen+Unseen*	72.96/74.69	65.02/66.54	72.68/74.31	65.06/66.14
Training with *Seen*	72.47/73.59	64.37/65.60	72.46/**73.07**	64.47/**65.05**

First, to verify the detection ability on short actions, we compare the AUC performance on different scales of actions on ActivityNet-1.3 validation set. According to the value of s from small to large, we artificially divide the dataset into three groups: small scale actions that $0 \leq s < 0.06$, middle scale actions that $0.06 \leq s < 0.65$, and large scale actions $0.65 \leq s \leq 1.0$. Each subset has almost the same amount of ground truth, which guarantees the fairness of comparison. Then we evaluate methods on each sub dataset.

What's more, we transfer our scale-invariant loss to our methods to prove its generality. The results demonstrate:

1. Both BMN, DBG and TSI behave worse on the subset of short actions compared with long actions. This phenomenon is intuitive because small actions naturally don't have sufficient feature representation against the background, and the IoU evaluation metrics are sensitive especially on small action length, not surprisingly, bringing the extreme low recall.
2. Without bells and whistles, we transfer the scale-invariant loss into BMN and DBG, and achieve steady improvements. In BMN, AUC of short actions has been boosted from 36.53% to 40.24%. Specifically, because the imbalance issue in DBG is not severer as BMN, AUC gains for DBG is not as much as BMN, which is acceptable.
3. Except for the significant improvement on short actions, to middle actions and long actions, TSI also provide performance gains than baseline BMN.
4. If we only adopt boundary detection module in TSI, AUC can achieve 66.31%. When we integrate TBD with IMR, performance is already better than the BMN baseline. Overall, our TSI achieves the 68.35% in validation set and 68.85% on test set of ActivityNet-1.3.

Generalization Ability. To evaluate the generalization ability of action proposal methods for unseen videos, following [9], we choose two un-overlapped subsets "Sports, Exercise, and Recreation" and "Socialing, Relaxing, and Leisure" classes of ActivityNet-1.3, as seen and unseen subset respectively. C3D [15] pretrained on Sports-1M dataset [30] is adopted for feature extraction. We train TSI with seen and seen+unseen videos separately and evaluate on both sub-datasets. As shown in Table 6, TSI can localize the actions of unseen data with high AUC. Compared with BMN, TSI also achieves better generalization ability.

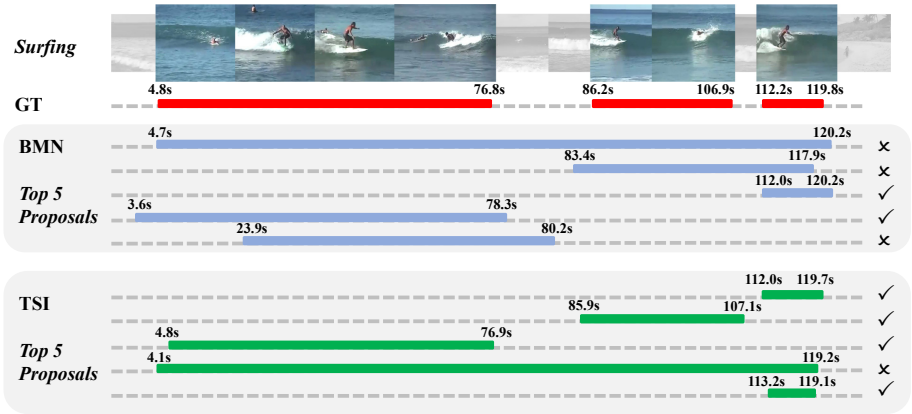


Fig. 6. Qualitative results of top-5 proposals generated by BMN and TSI on ActivityNet-1.3.

Visualization of Qualitative Results. As illustrated in Fig. 6, we visualize the top-5 proposal prediction of BMN and TSI on ActivityNet dataset. The demonstrated surfing video has three ground truth actions. However, due to the excessive learning for long actions, BMN may regard two individual actions as only one and predict more proposals with long duration. Besides, the temporal boundary of BMN is also not accurate enough. Compared with BMN, our proposed method can retrieve three actions independently with higher overlap and more accurate boundaries, because of the introduced modules.

5.3 Temporal Action Proposal Detection

With retrieved high quality action proposals, many video understanding tasks will be benefited, such as temporal action detection. In the detection task, Mean Average Precision (mAP) is used as the evaluation metrics. For a fair comparison, we combine our TSI proposals with state-of-the-art action classifier to achieve "detection by classifying proposals" framework.

On THUMOS14, we select top-200 TSI proposals with UntrimmedNet classifier [17] following [9]. The results on THUMOS14 datasets are shown in Table 7. Experiments prove that our generated proposals can satisfy the demand for detection task and outperform other state-of-the-art methods on THUMOS14 benchmarks, indicating that TSI can retrieve high quality action proposals.

On ActivityNet-1.3, we adopt top-100 TSI proposals with top-2 video level classification results provided by CUHK [31] as detection results. More specific, to enhance the detection performance on ActivityNet, we first adopt the proposal selection introduced in Sect. 4.2. Then, instead of using $p_{start} \cdot p_{end} \cdot p_{iou}$ as proposal confidence, we directly use the p_{iou} as final proposal confidence and utilize NMS with threshold 0.5 to reduce the redundant proposals, which is the same setting in BMN for fair comparison. The results are reported in Table 8.

Table 7. Action detection results on testing set of THUMOS14, where video-level classifier UntrimmedNet [17] is combined with our proposals.

Method	0.7	0.6	0.5	0.4	0.3
TURN [22]	6.3	14.1	25.6	35.3	46.3
BSN [7]	20.0	28.4	36.9	45.0	53.5
MGG [25]	21.3	29.5	37.4	46.8	53.9
BMN [9]	20.5	29.7	38.8	47.4	56.0
DBG [10]	21.7	30.2	39.8	49.4	57.8
G-TAD [13]	**23.4**	30.8	40.2	47.6	54.5
TSI	22.4	**33.2**	**42.6**	**52.1**	**61.0**

Table 8. Action detection results on validation set of ActivityNet-1.3, where video-level classification results generated by [31] are combined with our proposals.

Method	Validation			
	0.5	0.75	0.95	Average
SSN [2]	39.12	23.48	5.49	23.98
BSN [7]	46.45	29.96	8.02	30.03
DBG [10]	42.59	26.24	6.56	29.72
BMN [9]	50.07	34.78	8.29	33.85
G-TAD [13]	50.36	34.60	**9.02**	34.09
TSI	50.86	33.89	7.28	33.71
TSI(reweight)	**51.18**	**35.02**	6.59	**34.15**

To further improve the detection performance, we reweight the iou classification score and iou regression score, which can achieve the mAP of 34.15%.

It is worth discussing the differences and connections between temporal action proposal generation task and temporal action detection task. Although the proposal generation results with proposal classification results can be combined for the detection task, however, the proposal confidence used for ranking must be carefully designed. For example, DBG has achieved state-of-the-art action proposal generation performance with AUC of 68.23%, while the detection performance is unexpected low with only 29.72% mAP, which is far below current baseline methods. The reason of this phenomenon is the different evaluation metrics of each task. The action proposal generation focuses on the diversity of retrieved proposals and judges the performance by the recall of top N proposals. However, the action detection task focuses on the precision of top proposals, such as top-5. Therefore, some action proposal generation method, such as DBG, may retrieve the actions with well diversity, yet sacrificing the precision of top 1 proposal. In fact, the top-1 precision of DBG is much lower than TSI, leading to the low detection performance. This insight also reminds us that one possible trick for improving detection performance, which is, using two-stage methods to

learn the proposal confidence again with proposal generation results, and re-rank proposals with proposal-relation-aware model, such as P-GCN.

6 Conclusion

In this paper, we introduced the Temporal Scale Invariant Network (TSI) for action proposal generation, which can predict precise action boundaries with temporal contextual information and regress accurate proposal confidence. Especially, we analyze the positive sample imbalance problem in temporal action proposal generation and correspondingly devise a scale-invariant loss function to make up the insufficient learning of short actions and reduce the impact of the action scale change. Extensive experiments prove the effectiveness of proposed TSI and the state-of-the-art performance on ActivityNet-1.3 and THUMOS14 benchmarks is reported.

References

1. Buch, S., Escorcia, V., Shen, C., Ghanem, B., Carlos Niebles, J.: SST: single-stream temporal action proposals. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2911–2920 (2017)
2. Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D.: Temporal action detection with structured segment networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2914–2923 (2017)
3. Lin, T., Zhao, X., Shou, Z.: Single shot temporal action detection. In: Proceedings of the 25th ACM international conference on Multimedia, pp. 988–996. ACM (2017)
4. Chao, Y.W., Vijayanarasimhan, S., Seybold, B., Ross, D.A., Deng, J., Sukthankar, R.: Rethinking the faster R-CNN architecture for temporal action localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1130–1139 (2018)
5. Gao, J., Chen, K., Nevatia, R.: CTAP: complementary temporal action proposal generation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11206, pp. 70–85. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01216-8_5
6. Gao, J., et al.: Accurate temporal action proposal generation with relation-aware pyramid network. In: AAAI Conference on Artificial Intelligence, pp. 10810–10817 (2020)
7. Lin, T., Zhao, X., Su, H., Wang, C., Yang, M.: BSN: boundary sensitive network for temporal action proposal generation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11208, pp. 3–21. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01225-0_1
8. Gong, G., Zheng, L., Mu, Y.: Scale matters: temporal scale aggregation network for precise action localization in untrimmed videos. In: IEEE International Conference on Multimedia and Expo, pp. 1–6 (2020)
9. Lin, T., Liu, X., Li, X., Ding, E., Wen, S.: BMN: boundary-matching network for temporal action proposal generation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3889–3898 (2019)

10. Lin, C., et al.: Fast learning of temporal action proposal via dense boundary generator. In: AAAI Conference on Artificial Intelligence, pp. 11499–11506 (2020)
11. Long, F., Yao, T., Qiu, Z., Tian, X., Luo, J., Mei, T.: Gaussian temporal awareness networks for action localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 344–353 (2019)
12. Zeng, R., et al.: Graph convolutional networks for temporal action localization. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 7094–7103 (2019)
13. Xu, M., Zhao, C., Rojas, D.S., Thabet, A., Ghanem, B.: G-TAD: sub-graph localization for temporal action detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10156–10165 (2020)
14. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in Neural Information Processing Systems, pp. 568–576 (2014)
15. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497 (2015)
16. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3D residual networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5533–5541 (2017)
17. Wang, L., Xiong, Y., Lin, D., Van Gool, L.: UntrimmedNets for weakly supervised action recognition and detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4325–4334 (2017)
18. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6299–6308 (2017)
19. Caba Heilbron, F., Carlos Niebles, J., Ghanem, B.: Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1914–1923 (2016)
20. Shou, Z., Chan, J., Zareian, A., Miyazawa, K., Chang, S.F.: CDC: convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5734–5743 (2017)
21. Buch, S., Escorcia, V., Ghanem, B., Fei-Fei, L., Niebles, J.C.: End-to-end, single-stream temporal action detection in untrimmed videos. In: British Machine Vision Conference, vol. 2, p. 7 (2017)
22. Gao, J., Yang, Z., Chen, K., Sun, C., Nevatia, R.: Turn tap: temporal unit regression network for temporal action proposals. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3628–3636 (2017)
23. Ji, J., Cao, K., Niebles, J.C.: Learning temporal action proposals with fewer labels. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 7073–7082 (2019)
24. Schlosser, P., Munch, D., Arens, M.: Investigation on combining 3d convolution of image data and optical flow to generate temporal action proposals. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops (2019)
25. Liu, Y., Ma, L., Zhang, Y., Liu, W., Chang, S.F.: Multi-granularity generator for temporal action proposal. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3604–3613 (2019)

26. Escorcia, V., Caba Heilbron, F., Niebles, J.C., Ghanem, B.: DAPs: deep action proposals for action understanding. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 768–784. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_47
27. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
28. Bodla, N., Singh, B., Chellappa, R., Davis, L.S.: Soft-NMS-improving object detection with one line of code. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 5561–5569 (2017)
29. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: ActivityNet: a large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 961–970 (2015)
30. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1725–1732 (2014)
31. Zhao, Y., et al.: CUHK & ETHZ & SIAT submission to ActivityNet challenge 2017. In: CVPR ActivityNet Workshop (2017)