

# Learning Temporal Action Proposals With Fewer Labels

Jingwei Ji, Kaidi Cao, Juan Carlos Niebles

Stanford University

{jingweij, kaidicao, jniebles}@cs.stanford.edu

## Abstract

*Temporal action proposals are a common module in action detection pipelines today. Most current methods for training action proposal modules rely on fully supervised approaches that require large amounts of annotated temporal action intervals in long video sequences. The large cost and effort in annotation that this entails motivate us to study the problem of training proposal modules with less supervision. In this work, we propose a semi-supervised learning algorithm specifically designed for training temporal action proposal networks. When only a small number of labels are available, our semi-supervised method generates significantly better proposals than the fully-supervised counterpart and other strong semi-supervised baselines. We validate our method on two challenging action detection video datasets, ActivityNet v1.3 and THUMOS14. We show that our semi-supervised approach consistently matches or outperforms the fully supervised state-of-the-art approaches.*

## 1. Introduction

With millions of cameras in the world, a tremendous amount of videos are generated and transmitted every day. A very important subject in these videos is humans performing activities and actions. This has motivated the computer vision community to study algorithms for understanding actions from video collections. An important task for action understanding is action detection, or temporal action localization, where the goal is to temporally localize all actions of interest within long video sequences. A common approach to tackle this problem is to first generate *temporal action proposals* to localize temporal intervals of interest, which are then fed into a classifier to obtain the corresponding action labels. In this paper, we focus on the temporal action proposal module.

To achieve high prediction accuracy, most of the existing state-of-the-art algorithms for temporal action proposals use supervised deep learning approaches [3, 14, 15, 23]. Such approaches require large amount of *labeled* videos. Different from labeling in other vision tasks like image recogni-

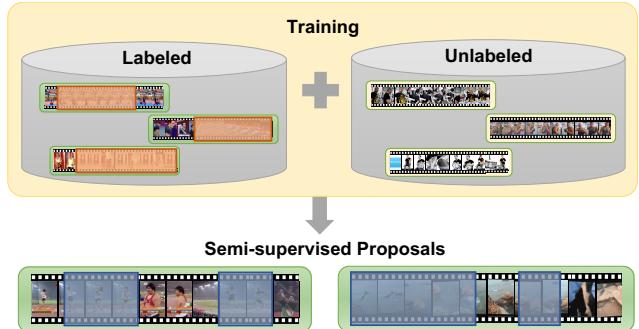


Figure 1. With only a part of training videos labeled with ground truth proposals, our semi-supervised framework can generate temporal action proposals with better quality than the state-of-the-art fully-supervised approaches.

tion, labeling temporal boundaries of actions in untrimmed videos is much more time-consuming. On the other hand are unsupervised learning approaches [34] where no label is needed for training. Although they are free from the burden of labeling, the overall performance in many tasks is usually inevitably poor than that of supervised approaches.

Semi-supervised learning is a well fit solution when large amount of data is available but only a small portion is labeled. Different from unsupervised learning, semi-supervised learning still leverages labeled data as strong supervision for high prediction accuracy. Compared to supervised learning, semi-supervised learning is less likely to overfit on the small labeled dataset because it can make use of the unlabeled data. Semi-supervised learning has been effective in image classification [21, 25, 29, 36], but has never been explored to assist generating temporal action proposals. In our problem setup (see Figure 1), we assume that during training only a part of the videos come with temporal boundary labels of actions for supervised learning. In the meanwhile, other videos with no labels or annotations are available to be leveraged by the training process. By extending the knowledge extracted from the labeled set to the unlabeled set, we can obtain a more robust model due to the regularization role that the unlabeled data can play.

One core philosophy behind semi-supervised learning methods is to train the model with smooth and consistent

classification boundaries that are robust to stochastic perturbation. To find a smooth manifold of data, Tarvainen *et al.* [36] proposes Mean Teacher which averages the “student” models at different training iterations into a “teacher” model. We embrace this architecture into our model design. To improve the robustness of the model, it is critical to introduce random perturbations on the input to the student model. In particular for the task of temporal action proposals in videos, the perturbations should be designed to benefit sequence learning. However, the prior work has not proposed appropriate perturbations for sequence data such as videos.

We propose two types of sequential perturbations: Time Warping and Time Masking. Time Warping is a resampling layer which distorts video sequences along the temporal dimension, providing perturbations for time-sensitive tasks like temporal action proposals. Time Masking randomly masks some frames of the input videos. During training, the masked student models only see parts of the videos while they are encouraged to predict the same boundaries as the unobstructed teacher model predicts. These sequential perturbations allow our optimized model to be more robust and generalize better to unseen data.

Our main contributions are as follows: **(1)** To the best of our knowledge, we are the first to incorporate semi-supervised learning in temporal action proposals to achieve label efficiency. **(2)** We have designed two essential types of sequential perturbations for this semi-supervised framework and validated them against strong semi-supervised baselines in key experiments of temporal action proposals.

## 2. Related Work

**Temporal Action Detection and Proposals.** Given a long, untrimmed video, temporal action detection aims to localize each action instance with its start and end times as well as its action class [4, 12, 14, 16, 22, 33, 40].

Traditionally, many approaches address the problem by exhaustively applying an action classifier in a sliding windows fashion [13, 19, 26, 27, 37, 39]. These methods are typically inefficient in terms of computation cost, since they need to cover temporal windows of different lengths at each location throughout the whole untrimmed video.

Inspired by recent success in proposal-plus-classification approaches of image object detection, another group of two-stage methods first propose action-agnostic temporal segment in video, then classify the action of the trimmed clips. Buch *et al.* [3] propose a network that performs single-stream temporal action proposal generation, avoiding computation cost brought by sliding window. Shou *et al.* [32] use 3D ConvNets to generate temporal proposals. There are also end-to-end frameworks that enable joint optimization of proposal generation and action classification. Buch *et al.* [2] introduce semantics constraints for curriculum training

in end-to-end temporal action localization. Chao *et al.* [8] adopt Faster R-CNN [30] for action localization task.

The proposals generated in the above methods are often dependent on pre-defined anchors, lacking flexibility and preciseness of temporal bounds. Instead, Zhao *et al.* [41] simplify the proposal generation problem into classifying the actionness of every short video snippet, post-processed by a watershed algorithm. Gao *et al.* [15] and the Boundary Sensitive Network (BSN) [23] further infer whether a video snippet is the start or end of an action to obtain more precise boundaries, in which the BSN has become the state-of-the-art on the temporal action proposal task on ActivityNet Challenge [5].

Previous research is dedicated to develop better action proposal models trained with labeled videos. In parallel, we explore how to utilize unlabeled videos to further improve proposal and detection performance. In this work, we focus on evaluating our semi-supervised framework with the BSN due to its superior performance, though our framework’s flexibility allows it to be combined with other temporal action proposal architectures as well.

**Semi-supervised Deep Learning.** Semi-supervised learning has a rich history that spans decades [9, 42]. Instead of a comprehensive review, our focus is limited to semi-supervised deep learning. A common approach is to train a neural network by jointly optimizing a supervised classification loss on labeled data and an additional unsupervised loss on both labeled and unlabeled data [21, 25, 29, 36]. Consistency regularization has been widely used for the unsupervised loss, which encourages the model to generate consistent outputs when the raw inputs or intermediate feature maps are perturbed.

Here we summarize some examples of semi-supervised deep learning using consistency regularization. Ladder Networks [29] incorporate a reconstruction branch as the unsupervised task; they enforce consistency losses between encoded and decoded activation maps at each training step. Π-Model [21] simplifies Ladder Networks and only imposes consistency loss between outputs with different perturbations on data. Next, Temporal Ensembling [21] applies a consistency loss to model outputs and a more stable target: the exponential moving average of model outputs at each epoch. Instead of averaging outputs, the more powerful Mean Teacher [36] averages the weights of models at each training step (a.k.a. “student” models) into a separate “teacher” model, whose outputs serve as the target in the consistency loss. Orthogonal to the above approaches, Virtual Adversarial Training (VAT) [25] proposes using virtual adversarial noise instead of random noise as the data perturbation. In our work, we also impose consistency regularization between outputs of student and teacher models, and propose Time Warping and Time Masking as the data perturbations specifically for video data.

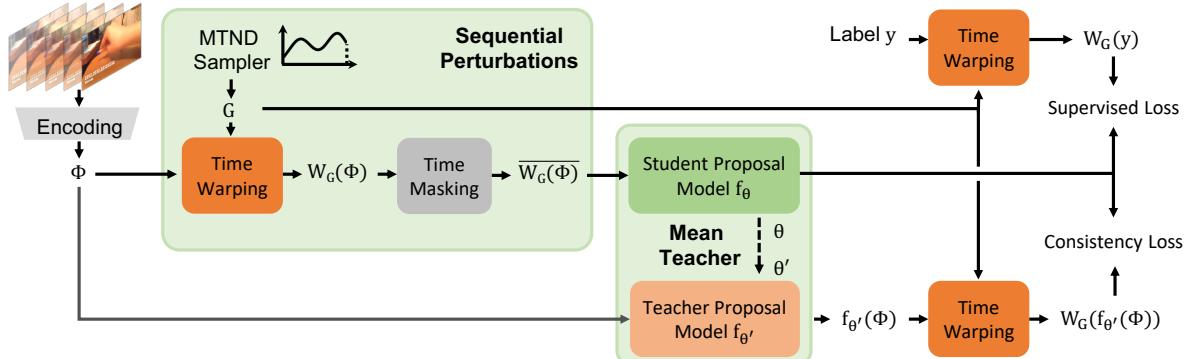


Figure 2. Overview of our method. Given an untrimmed video as input, we first encode it into a feature sequence  $\Phi$ . Next, sequential perturbations including Time Warping and Time Masking are applied to  $\Phi$  and the student proposal model takes this perturbed sequence as the input. Instead, the teacher model predicts directly on the unobstructed  $\Phi$ . In the end, the student model is jointly optimized with a supervised loss applied to labeled videos and a consistency loss to all videos.

Semi-supervised learning has been applied to sequence learning as well. Dai *et al.* [11] propose a sequence autoencoder for text classification. Prémont-Schwarz *et al.* [28] combine Ladder Networks with recurrent neural networks and evaluate their model on image classification on the Occluded Moving MNIST dataset. Clark *et al.* [10] propose cross-view training for multiple language tasks. Miyato *et al.* [24] apply VAT [25] on text classification. Although not designed for video analysis, some of the above approaches [10, 28] also embrace the idea of masking either on patches in images or words in sentences, and they inspire our Time Masking.

There is also work on *weakly supervised learning* for temporal action detection [1, 7, 17, 31], which differs from our semi-supervised setting. In the weakly supervised temporal action detection, part of the training data are fully labeled with the temporal boundaries and action classes while the rest of data are annotated with “weak” labels, either video-level classes or order lists of actions in the video. Instead, we do not assume availability of any kind of labels for the unlabeled videos used in our semi-supervised training, which entails a harder but more label-efficient task.

### 3. Technical Approach

Our main goal is to generate high-quality temporal action proposals with a relatively small amount of labels. This requires us to best utilize the labeled data with a powerful supervised proposal model while, at the same time, leveraging unlabeled data with an unsupervised auxiliary task designed for video understanding. Although our approach is agnostic to specific proposal methods, to validate the semi-supervised framework, we build our model on top of a state-of-the-art fully-supervised proposal generation network, the Boundary Sensitive Network [23]. We extend the Mean Teacher framework [36] with two types of sequential perturbations for training the proposal model: Time Warping

and Time Masking. See Figure 2 as an overview of our method.

#### 3.1. Video Encoding

The purpose of video encoding is to obtain a condensed video representation, which captures the appearance and motion patterns of a video. Given an untrimmed video with  $N$  frames as the input, we first divide it into non-overlapping short snippets which contain  $\delta$  frames each, forming a sequence of snippets  $S = \{X_1, X_2, \dots, X_T\}$ , where  $T = N/\delta$ . As illustrated in prior work [6, 38], both appearance and motion features contribute to action understanding, so we encode both the RGB frames and the optical flows of each video, then concatenate the encoded vectors. In particular, we use [38] as the video encoder  $\phi$  as in the fully-supervised baseline [23]. The encoder generates a sequence of feature vectors  $\Phi = \{\phi(X_1), \phi(X_2), \dots, \phi(X_T)\} \in \mathbb{R}^{T \times D}$ . Then we feed sequences of feature vectors into the following modules in mini-batches. Labeled and unlabeled videos share the same video encoder  $\phi$  and they co-exist in the same mini-batch.

#### 3.2. Temporal Action Proposal Model

Our semi-supervised model is sufficiently flexible that it can be built upon various fully-supervised temporal action proposal networks as long as they take sequential data as input. Specifically, we choose Boundary Sensitive Network (BSN) [23], a top performer in the temporal action proposal task in ActivityNet challenge 2018.

The same video encoding as in [23] is performed as the first step, then  $\Phi$  is directly fed into the BSN proposal model. The BSN is composed of a sequence of two trainable modules: a Temporal Evaluation Module (TEM) and a Proposal Evaluation Module (PEM). After the video encoding, TEM takes the snippet feature sequence  $\Phi$  as the input. The sequence  $\Phi$  is passed through temporal convolu-

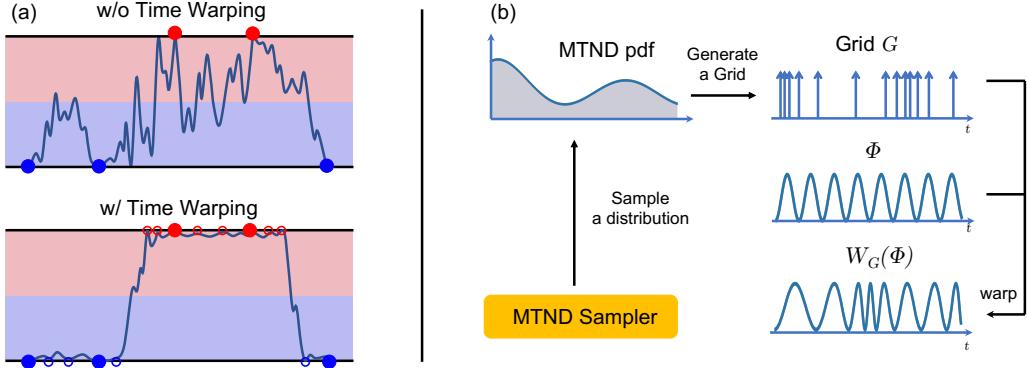


Figure 3. Time Warping. (a) With Time Warping, we can sample more snippet features in the encoded space. Here we show a simple example of binary classification for each snippet feature (dimension reduced to 1). Resampling new feature points (the empty circles) among labeled snippet features (the filled circles) encourages the student model to generate a smoother manifold for prediction.(b) To perform Time Warping, we first sample a mixed truncated normal distribution for generating the 1-D grid  $G$ . Then we apply grid sampling on the feature sequence  $\Phi$  to augment the data for training.

tional layers to generate three series of probability signals: actionness  $p^a \in \mathbb{R}^T$ , starting  $p^s \in \mathbb{R}^T$  and ending  $p^e \in \mathbb{R}^T$ . Then proposals are generated according to these three signal sequences. Finally, PEM predicts a confidence score  $p_{conf}$  for each proposal indicating how overlapped a proposal is with the closest ground truth interval, to decide if the proposal is accepted or rejected. Please refer to [23] or our supplementary materials for more details of BSN.

### 3.3. Mean Teacher Framework

Now we introduce how we construct the semi-supervised learning framework for temporal action proposals. When only a small number of labeled training samples are available, deep models like BSN tend to over-fit and not able to extract enough knowledge from the training set to generalize to unseen videos. This can be mitigated by semi-supervised learning where unlabeled videos can also be used for training. Without ground truth labels, the supervised classification loss is undefined upon unlabeled videos. Instead, we need to introduce an unsupervised auxiliary task to leverage information from unlabeled videos.

As a baseline, we can directly adapt the Mean Teacher method on the temporal action proposal model to form the semi-supervised learning framework. In this framework, there are two models: a *student* proposal model  $f_\theta$  and a *teacher* proposal model  $f_{\theta'}$ . The student learns as in fully-supervised learning, with its weights  $\theta$  optimized by the supervised classification losses applied on labeled videos. The teacher proposal model has the identical neural network architecture as the student, while its weights  $\theta'$  are generated by averaging  $\theta$  from different iterations of training:

$$\theta'_i = \alpha\theta'_{i-1} + (1 - \alpha)\theta_i \quad (1)$$

where  $\alpha$  is a smoothing coefficient parameter and  $i$  denotes the training iteration. As an ensembled model, the teacher embeds input snippet features into a smooth manifold and

outputs more consistent predictions than students. Then the unsupervised task is to impose consistency regularization between the outputs from the student and the teacher model, with both labeled and unlabeled videos as input.

### 3.4. Sequential Perturbations

Beyond the Mean Teacher framework, stochastic perturbations have been found crucial for learning robust models by many semi-supervised learning works [21, 25, 29, 36]. A typical way of perturbation is adding noise to feature maps. Mean Teacher [36] adds Gaussian noise to intermediate feature maps of both student and teacher models, whereas VAT [25] adds adversarial noise to the input. In video analysis, we further explore what other specific perturbations are necessary for sequential learning. We propose two sequential perturbations: Time Warping and Time Masking.

**Time Warping.** Time Warping is essentially a resampling layer, which resamples a sequence of feature vectors  $\Phi \in \mathbb{R}^{T \times D}$  along the time dimension guided by a randomly generated 1-D flow-field grid. Time Warping is vital for semi-supervised temporal action proposals: First, by propagating labels to unlabeled locations in the feature space, resampling leads to smoother predictions (Figure 3 (a)); second, Time Warping serves as a way of data augmentation, providing more labeled data for training, which is especially helpful in the case when we have few labels; third, stretching and compressing input signals can generate more variants to learn in certain tasks, like temporal action proposals, which require accurate starting/ending location prediction.

To perform warping on the input feature sequence  $\Phi$ , each output feature vector is computed by applying linear sampling on  $\Phi$  according to a dense 1-D grid  $G = \{g_t\}$ , where  $g_t$  is the temporal location to sample an output feature vector. Critical in performing Time Warping, the grid should include long-term distortion which slows down some parts of the video while speeds up the other parts; it

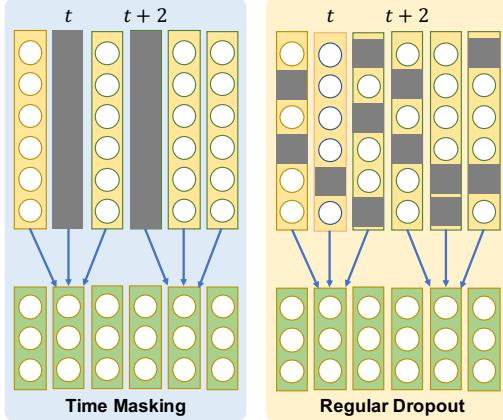


Figure 4. Time Masking. Unlike dropout which randomly zeros out some of the neurons in the input, Time Masking drops the entire feature vectors from the randomly selected time steps.

should also contain short-term stochastic noise. With these considerations, we propose a Mixed Truncated Normal Distribution (MTND) sampler (Figure 3 (b)) to generate grids.

A MTND is formed by mixing  $n$  truncated normal distributions  $\mathcal{N}_0^T(\mu_i, \sigma_i)$ ,  $i \in \{1, 2, \dots, n\}$  by different weights. Since we only want to interpolate the input sequence, the distribution is truncated at the starting (0) and ending ( $T$ ) locations. The means  $\mu_i$ 's are sampled from a uniform distribution and the standard deviations  $\sigma_i$ 's are sampled from a log-uniform distribution. Given a MTND, we sample  $T$  locations from it as the grid  $G$ , then we perform the warping and obtain  $\mathcal{W}_G(\Phi) \in \mathbb{R}^{T \times D}$ .

**Time Masking.** Besides Time Warping, we propose a Time Masking operation as another source of sequential perturbations during training. In our pipeline, Time Masking follows Time Warping and takes  $\mathcal{W}_G(\Phi)$  as input. The idea of Time Masking is simple: some snippets in the input sequence are masked out from the student model, while the teacher model can see the whole unobstructed video sequence. We denote the output of the Time Masking as  $\mathcal{W}_G(\Phi)$ . During the training, the masked student models at each iteration are encouraged to generate the same outputs as the teacher does, even though they could not access the entire information of input videos.

Time Masking can be viewed as a special Dropout layer (Figure 4). In the regular Dropout layer, the neurons in one snippet are not likely to be entirely dropped, which gives the model a chance to peek some information from every snippet in the receptive field. Instead, in Time Masking, no information of the dropped snippet will be passed to the next layers. The student model will be forced to aggregate information from temporal context to make prediction on dropped snippets. Such capability of temporal context aggregation will be learned both from supervised losses on the labeled videos and the consistency with the teacher model on all training data.

### 3.5. Training

Training our semi-supervised framework includes two parts: minimizing the supervised losses on labeled data and the consistency loss on all training data. Although we have student and teacher models, only weights in student models are optimized via back-propagation, and weights in the teacher model are the averaged weights of students.

**Supervised Losses.** Aligned with the fully-supervised proposal model, our semi-supervised framework uses the same supervised losses for training as in BSN. Please refer to [23] or our supplementary materials for details of the losses. In our semi-supervised framework, the output of the student proposal model corresponds to the sequential input distorted by Time Warping. Thus the labels  $y$  also need to be resampled according to the same grid generated by the MTND sampler. With the warped labels  $\mathcal{W}_G(y)$ , we enforce the supervised losses on the student output  $f_\theta(\mathcal{W}_G(\Phi))$ . Note that the supervised losses can only be applied on labeled videos in the training set.

**Consistency Regularization.** The consistency loss treats the outputs of the teacher model as labels and encourages the student to learn a smooth manifold like the teacher's. Unlike the supervised losses, the consistency loss can be applied to both labeled and unlabeled videos in the training set. Similar to how we handle the labels in supervised losses, we also warp the outputs of the teacher to be  $\mathcal{W}_G(f_{\theta'}(\Phi))$ . The consistency loss then measures the distance between the student outputs and the warped teacher outputs:

$$L_{cons} = D(f_\theta(\overline{\mathcal{W}_G(\Phi)}), \mathcal{W}_G(f_{\theta'}(\Phi))) \quad (2)$$

For the distance function  $D$ , we use Mean Squared Error in all experiments. Same as the supervised optimization, only weights in the student model are trained. The consistency loss and the supervised losses are summed as the total loss.

## 4. Experiments

**Datasets.** We use ActivityNet v1.3 and THUMOS14 for all experiments. **ActivityNet v1.3** [5] is a large database for temporal action proposals and detection. It contains 19,994 videos of 200 activity classes and has been used in the ActivityNet Challenge 2016 to 2019. ActivityNet v1.3 is divided into training, validation and testing sets by a ratio of 2:1:1, and temporal boundaries of action instances are annotated in all videos. **THUMOS14** [18] contains 200 and 213 temporal annotated untrimmed videos with 20 action classes in validation and testing sets, separately. The training set of THUMOS14 is the UCF-101 [35], which contains trimmed videos for action classification task. Instead of training on these trimmed videos, we train our model on the untrimmed videos in validation set, and report performance on the test set.

**Evaluation Metrics.** We evaluate our method on two tasks: temporal action proposals and temporal action localization. For proposals, we report Average Recall (AR) at various Average Number of Proposals per Video (AN). AR is defined as the average of all recall values with tIoU thresholds from 0.5 to 1 with a step size of 0.05. On ActivityNet v1.3, area under the AR vs. AN curve (AUC) is also used as a measurement, where AN varies from 0 to 100. For action localization, we calculate mean Average Precision (mAP) with different tIoU thresholds.

**Implementation Details.** We follow the same pre- and post-processing as the BSN [23], including parameters used in Soft-NMS. For feature extraction on ActivityNet v1.3, we use the two-stream network [38] pre-trained on Kinetics [20]. Different from the BSN’s setting, our features are not pre-trained on ActivityNet classification task to avoid using extra labels which will contaminate the semi-supervised setup. We use the same video features as the BSN for all THUMOS14 experiments. For the semi-supervised training, we use EMA decay  $\alpha = 0.999$ . Masking probability in Time Masking is fixed to 0.3.

#### 4.1. Temporal Action Proposals

Taking a long, untrimmed video as input, our method aims to generate temporal boundaries determining the starting and ending time of each action instances. In this section, we compare the temporal action proposals generated by our model on ActivityNet v1.3 and THUMOS14 with fully-supervised BSN and other state-of-the-art methods to verify the effectiveness of our semi-supervised framework.

**Comparison to fully-supervised methods.** We first compare the action proposal results on ActivityNet-1.3 validation set under two training setups: (1) Our semi-supervised framework, where  $x$  percent of training videos are labeled with temporal boundaries and  $100 - x$  percent of training videos are not; (2) State-of-the-art fully-supervised learning, where the same amount of labeled videos are used for training while no other data are used. With this comparison, we can see how our semi-supervised framework performs against the fully-supervised counterpart under different labeled/unlabeled ratio.

To validate the label efficiency of our method, we vary the amount of labels for training, then measure the AUC and AR@100 of proposals generated by our method and the original BSN (Figure 5). With only a part of the training set labeled, our method outperforms the fully-supervised baseline consistently under different ratio of labeled training videos. Notably, with only 60% of the videos labeled, our semi-supervised model outperforms the state-of-the-art fully-supervised BSN trained with all labels in both metrics of AUC and AR@100 (Table 1). Similarly, we examine the label efficiency on THUMOS14 (Figure 6), and observe consistent superior performance as well.

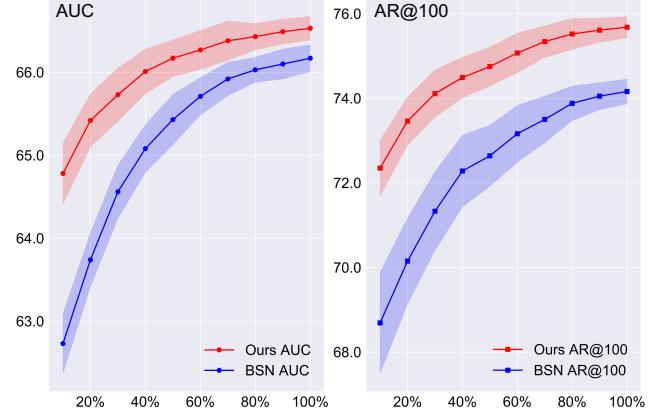


Figure 5. Label efficiency experiments on ActivityNet v1.3. Varying the percentages of labels for training, we compare the AUC and AR@100 of the proposals generated by our semi-supervised method and the fully-supervised BSN counterpart.

Method	SSN[41]	CTAP[15]	BSN[23]	Ours@60%
AR@100	63.52	73.17	74.16	<b>75.07</b>
AUC	53.02	65.72	66.17	<b>66.35</b>

Table 1. Comparison between our method and other state-of-the-art proposal generation methods on ActivityNet v1.3 in terms of AR@100 and AUC. We outperform all other methods while using only 60% of the labels.

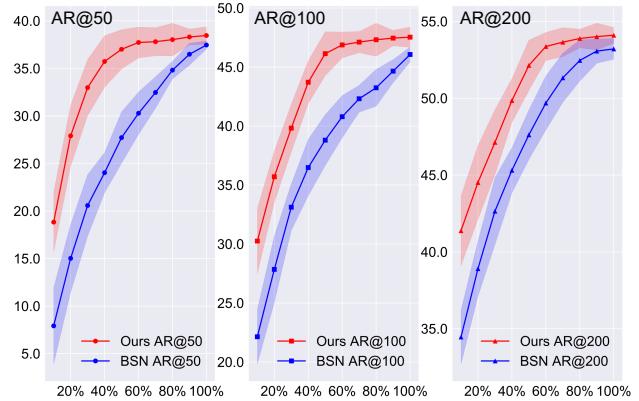


Figure 6. Label efficiency experiments on THUMOS14. We report AR@50, @100, and @200 of the proposals generated by our method and the vanilla BSN when trained with different percentages of labels in the training set.

We then compare the proposal generation on THUMOS14 with strong baseline models. Table 2 shows the comparison measured by average recall at various average number of proposals per video. Again, we outperform the BSN when trained with only 60% of labels. Moreover, when 100% of labels are available, our framework can further increase the average recalls.

**Comparison to semi-supervised baselines.** Next, we investigate the performance of our framework against multiple semi-supervised baselines on THUMOS14 proposals

Feature	Method	@50	@100	@200
C3D	DAPs [14]	13.56	23.83	33.96
C3D	SCNN-prop [32]	17.22	26.17	37.01
C3D	SST [3]	19.90	28.36	37.90
C3D	TURN [16]	19.63	27.96	38.34
C3D	BSN [23]	29.58	37.38	45.55
2-Stream	TAG [41]	18.55	29.00	39.61
Flow	TURN [16]	21.86	31.89	43.02
2-Stream	CTAP [15]	31.03	40.23	50.13
2-Stream	BSN@60% [23]	30.28	40.79	49.03
2-Stream	BSN@100% [23]	37.46	46.06	53.21
2-Stream	Ours@60%	<b>37.73</b>	<b>46.87</b>	<b>53.37</b>
2-Stream	Ours@100%	<b>38.46</b>	<b>47.53</b>	<b>54.10</b>

Table 2. Comparison between our method and other state-of-the-art proposal generation methods on THUMOS14 in terms of AR@50, AR@100 and AR@200.

AN	@50	@100	@200	@500	@1000
Vanilla BSN	30.28	40.79	49.03	<b>57.58</b>	62.35
VAT [25]	32.48	43.13	49.18	<b>57.61</b>	62.49
MT [36]	35.61	44.20	51.51	<b>58.66</b>	62.55
MT + VAT	35.63	44.21	51.49	<b>58.64</b>	62.56
MT + Dropout	35.73	44.25	51.56	<b>58.67</b>	62.58
Ours -TW	36.31	44.79	52.30	<b>58.97</b>	62.82
Ours -TM	37.24	45.37	52.65	<b>59.74</b>	63.10
Ours	<b>37.73</b>	<b>46.87</b>	<b>53.37</b>	<b>60.81</b>	<b>64.59</b>

Table 3. Comparison between fully-supervised and semi-supervised baselines trained with 60% of the labels. We report AR at various AN on THUMOS14. Abbreviations: VAT for Virtual Adversarial Training, MT for Mean Teacher, TW for Time Warping, and TM for Time Masking. Our full model outperforms strong semi-supervised baselines.

with 60% labels for training (Table 3). We first implement and evaluate VAT [25] combined with BSN. The key idea of VAT is to improve model robustness to the approximately worst case perturbations instead of random ones. Similar to the VAT application to text classification [24], we apply the adversarial noise to each video snippet embeddings, rather than directly to the raw input. VAT does not improve average recall by much, partly because that the worst case perturbations on video snippet embeddings are not significantly different to random noises.

We also investigate different variants of Mean Teacher [36]. The vanilla Mean Teacher with only random noises and no dropout layers outperforms VAT. Also, adding VAT to Mean Teacher does not help much on better proposals. Mean Teacher with regular dropout further improves the quality of proposals, but not as powerful as our approach with Time Masking. With the same dropout/masking probability, although the regular dropout zeros the same amount of neurons as Time Masking per training step, it formulates an easier task for student models to learn since the student can rely on more snippets to do inference.

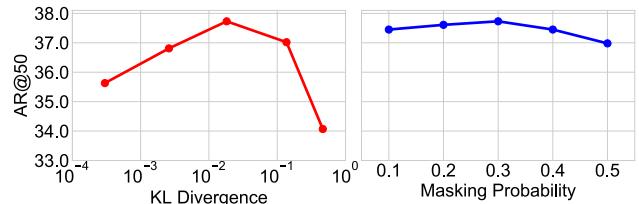


Figure 7. Ablation experiments. We assess the effects of Time Warping and Time Masking under different hyper-parameter choices to find sweet points for better performance.

Method	0.7	0.6	0.5	0.4	0.3
SST [3] + UNet	4.7	10.9	20.0	31.5	41.2
TURN [16] + UNet	6.3	14.1	24.5	35.3	46.3
BSN [23] + UNet	20.0	28.4	36.9	45.0	53.3
Ours@60% + UNet	<b>20.5</b>	<b>29.5</b>	<b>37.2</b>	<b>45.2</b>	<b>53.4</b>
Ours@100% + UNet	<b>20.7</b>	<b>29.9</b>	<b>37.9</b>	<b>46.3</b>	<b>55.1</b>

Table 4. Action detection results on the testing set of THUMOS14 in terms of mAP@tIoU. We compare with proposal + classification methods, where classification results are generated by UntrimmedNet [33].

Finally, we examine the contributions of the two proposed sequential perturbations by removing them respectively. Both of them contribute to the proposals while Time Warping appears to play a major role.

**Qualitative Results.** We visualize some temporal action proposals generated by our semi-supervised approach. Figure 8 shows that our approach is able to generate more precise temporal boundaries than the fully-supervised baseline on THUMOS14 when both are trained with 60% of labels.

## 4.2. Ablation Experiments

To assess the functionality of the two proposed sequential perturbations, we run experiments on THUMOS14 with 60% of the labels with different hyper-parameters used in Time Warping and Time Masking.

**Degree of distortion in Time Warping.** The effect of Time Warping depends on the grid sampled from MTND sampler. Varying the number of truncated normal distributions and their scales, the MTND can go from a nearly uniform distribution to a very uneven one which will greatly distort the input sequence. We examine the impact of different degree of distortion in Time Warping on generated proposals. The degree of distortion is measured by the KL-divergence  $D_{KL}(P \parallel Q)$  between the sampled MTND as  $P$  and a uniform distribution as  $Q$ . Figure 7 (a) shows a sweet spot with  $D_{KL}$  at an order of magnitude of 0.01. When  $D_{KL}$  approaches to 0, the effect of Time Warping diminishes; when the degree of distortion is too large, many parts of videos can hardly get sampled, equivalently decreasing the number of labels for training.

**Masking probability in Time Masking.** We experiment with different probabilities of zeroing feature vectors in the

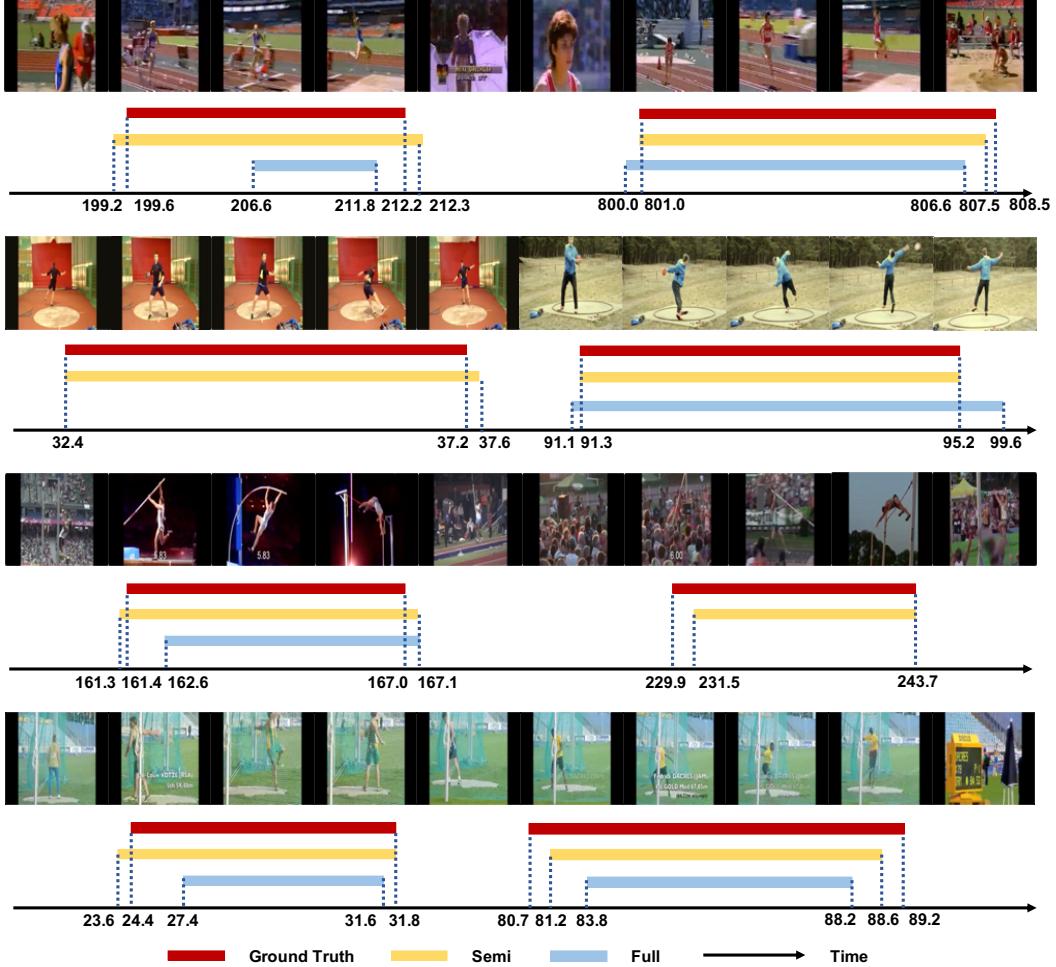


Figure 8. We compare THUMOS14 proposals generated by our semi-supervised method with the fully-supervised BSN trained using 60% of the labels. We also show ground truth intervals for reference.

sequence fed to Time Masking. As shown in Figure 7 (b),  $p = 0.3$  appears to be an optimal operating point, bringing appropriate difficulty to the students. Thus we fix this masking probability in all our experiments.

### 4.3. Temporal Action Localization

The end goal of generating temporal action proposal is temporal action localization, so we further evaluate our proposals for the localization task on THUMOS14. We follow the proposal-plus-classification two-stage approach as in [3, 16, 23]. As the BSN does, we use the top-2 video-level classes predicted by UntrimmedNet [33] on top of the proposals generated by different approaches. We report the mean Average Precision at different temporal IoU thresholds with 200 proposals per video on THUMOS14 (Table 4). The direct comparison is with the fully-supervised BSN trained with all labels, where we achieve better performance on different temporal IoU thresholds from 0.3 to 0.7. When trained with all labels, our model further improves performance on action localization.

## 5. Conclusion

We show that temporal proposal models can be trained with higher label efficiency by adopting our semi-supervised approach to learn their parameters. Our semi-supervised framework extends the Mean Teacher model with two proposed sequential perturbations for video understanding. We show empirically that our model achieves similar performance as the fully-supervised approach when trained with only 60% of the labels, outperforming other semi-supervised baselines as well. Furthermore, we show that our semi-supervised proposals can be effectively applied to the problem of temporal action localization.

**Acknowledgments.** This work has been supported by Panasonic and JD.com American Technologies Corporation (“JD”) under the SAIL-JD AI Research Initiative. This article solely reflects the opinions and conclusions of its authors and not Panasonic, JD or any entity associated with Panasonic or JD.com.

## References

- [1] Piotr Bojanowski, Rémi Lajugie, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. Weakly supervised action labeling in videos under ordering constraints. In *European Conference on Computer Vision*, pages 628–643. Springer, 2014. [3](#)
- [2] Shyamal Buch, Victor Escorcia, Bernard Ghanem, Li Fei-Fei, and Juan Carlos Niebles. End-to-end, single-stream temporal action detection in untrimmed videos. In *BMVC*, volume 2, page 7, 2017. [2](#)
- [3] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. Sst: Single-stream temporal action proposals. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6373–6382. IEEE, 2017. [1, 2, 7, 8](#)
- [4] Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. 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*, pages 1914–1923, 2016. [2](#)
- [5] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015. [2, 5](#)
- [6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 4724–4733. IEEE, 2017. [3](#)
- [7] Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, and Juan Carlos Niebles. D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3546–3555, 2019. [3](#)
- [8] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A Ross, Jia Deng, and Rahul Sukthankar. Rethinking the faster r-cnn architecture for temporal action localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1130–1139, 2018. [2](#)
- [9] Olivier Chapelle, Bernhard Scholkopf, and Alexander Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009. [2](#)
- [10] Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc V. Le. Semi-supervised sequence modeling with cross-view training. In *EMNLP*, 2018. [3](#)
- [11] Andrew M Dai and Quoc V Le. Semi-supervised sequence learning. In *Advances in neural information processing systems*, pages 3079–3087, 2015. [3](#)
- [12] Xuhuan Duan, Le Wang, Changbo Zhai, Nanning Zheng, Qilin Zhang, Zhenxing Niu, and Gang Hua. Joint spatio-temporal action localization in untrimmed videos with per-frame segmentation. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 918–922. IEEE, 2018. [2](#)
- [13] Olivier Duchenne, Ivan Laptev, Josef Sivic, Francis Bach, and Jean Ponce. Automatic annotation of human actions in video. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1491–1498. IEEE, 2009. [2](#)
- [14] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. Daps: Deep action proposals for action understanding. In *European Conference on Computer Vision*, pages 768–784. Springer, 2016. [1, 2, 7](#)
- [15] Jiyang Gao\*, Kan Chen\*, and Ram Nevatia. Ctap: Complementary temporal action proposal generation. In *ECCV*, 2018. [1, 2, 6, 7](#)
- [16] Jiyang Gao, Zhenheng Yang, Chen Sun, Kan Chen, and Ram Nevatia. Turn tap: Temporal unit regression network for temporal action proposals, 2017. [2, 7, 8](#)
- [17] De-An Huang, Li Fei-Fei, and Juan Carlos Niebles. Connectionist temporal modeling for weakly supervised action labeling. In *European Conference on Computer Vision*, pages 137–153. Springer, 2016. [3](#)
- [18] Yu-Gang Jiang, Jingen Liu, A Roshan Zamir, George Toderici, Ivan Laptev, Mubarak Shah, and Rahul Sukthankar. Thumos challenge: Action recognition with a large number of classes, 2014. [5](#)
- [19] Svebor Karaman, Lorenzo Seidenari, and Alberto Del Bimbo. Fast saliency based pooling of fisher encoded dense trajectories. In *ECCV THUMOS Workshop*, volume 1, page 7, 2014. [2](#)
- [20] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. [6](#)
- [21] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. [1, 2, 4](#)
- [22] Tianwei Lin, Xu Zhao, and Zheng Shou. Single shot temporal action detection. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 988–996. ACM, 2017. [2](#)
- [23] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *European Conference on Computer Vision*, 2018. [1, 2, 3, 4, 5, 6, 7, 8](#)
- [24] Takeru Miyato, Andrew M Dai, and Ian Goodfellow. Adversarial training methods for semi-supervised text classification. In *ICLR*, 2017. [3, 7](#)
- [25] Takeru Miyato, Shin-ichi Maeda, Shin Ishii, and Masanori Koyama. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE transactions on pattern analysis and machine intelligence*, 2018. [1, 2, 3, 4, 7](#)
- [26] Bingbing Ni, Xiaokang Yang, and Shenghua Gao. Progressively parsing interactional objects for fine grained action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1020–1028, 2016. [2](#)
- [27] Dan Ooneata, Jakob Verbeek, and Cordelia Schmid. The learner submission at thumos 2014. 2014. [2](#)
- [28] Isabeau Prémont-Schwarz, Alexander Ilin, Tele Hao, Antti Rasmus, Rinu Boney, and Harri Valpola. Recurrent ladder

- networks. In *Advances in Neural Information Processing Systems*, pages 6009–6019, 2017. 3
- [29] Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems*, pages 3546–3554, 2015. 1, 2, 4
- [30] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 2
- [31] Zheng Shou, Hang Gao, Lei Zhang, Kazuyuki Miyazawa, and Shih-Fu Chang. Autoloc: Weaklysupervised temporal action localization in untrimmed videos. In *ECCV*, pages 162–179, 2018. 3
- [32] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1049–1058, 2016. 2, 7
- [33] Gurkirt Singh and Fabio Cuzzolin. Untrimmed video classification for activity detection: submission to activitynet challenge. *arXiv preprint arXiv:1607.01979*, 2016. 2, 7, 8
- [34] Khurram Soomro and Mubarak Shah. Unsupervised action discovery and localization in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 696–705, 2017. 1
- [35] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 5
- [36] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017. 1, 2, 3, 4, 7
- [37] Limin Wang, Yu Qiao, and Xiaoou Tang. Action recognition and detection by combining motion and appearance features. *THUMOS14 Action Recognition Challenge*, 1(2):2, 2014. 2
- [38] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36. Springer, 2016. 3, 6
- [39] Jun Yuan, Bingbing Ni, Xiaokang Yang, and Ashraf A Kasim. Temporal action localization with pyramid of score distribution features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2016. 2
- [40] Ze-Huan Yuan, Jonathan C Stroud, Tong Lu, and Jia Deng. Temporal action localization by structured maximal sums. In *CVPR*, volume 2, page 7, 2017. 2
- [41] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. *ICCV, Oct*, 2, 2017. 2, 6, 7
- [42] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003. 2