Few-Shot Temporal Action Localization with Query Adaptive Transformer

Sauradip Nag[1,2]
s.nag@surrey.ac.uk
Xiatian Zhu [1]
xiatian.zhu@surrey.ac.uk
Tao Xiang[1,2]
t.xiang@surrey.ac.uk

[1] Centre for Vision Speech and Signal Processing (CVSSP)
University of Surrey, UK

[2] iFlyTek-Surrey Joint Research Centre on Artificial Intelligence

Abstract

Existing temporal action localization (TAL) works rely on a large number of training videos with exhaustive segment-level annotation, preventing them from scaling to new classes. As a solution to this problem, few-shot TAL (FS-TAL) aims to adapt a model to a new class represented by as few as a single video. Exiting FS-TAL methods assume trimmed training videos for new classes. However, this setting is not only unnatural – actions are typically captured in untrimmed videos, but also ignores background video segments containing vital contextual cues for foreground action segmentation. In this work, we first propose a new FS-TAL setting by proposing to use untrimmed training videos. Further, a novel FS-TAL model is proposed which maximizes the knowledge transfer from training classes whilst enabling the model to be dynamically adapted to both the new class and each video of that class simultaneously. This is achieved by introducing a query adaptive Transformer in the model. Extensive experiments on two action localization benchmarks demonstrate that our method can outperform all the state-of-the-art alternatives significantly in both single-domain and cross-domain scenarios. The source code can be found in https://github.com/sauradip/fewshotQAT

1 Introduction

Temporal action localization (TAL) aims to identify the temporal duration (*i.e.*, the start and end points) and class label of action instances in naturally untrimmed videos [3, 14]. Existing TAL methods [2, 19, 33, 36, 43] use training datasets composed of a large number of videos (e.g., hundreds) per class with exhaustive segment-level annotation. The annotation is tedious and costly. Further, for some rare classes collecting sufficient video instances may not even be feasible. These have severely limited the scalability and general usability of existing TAL methods. Inspired by the success of few-shot image classification [5, 9, 22, 25, 27], few-shot learning (FSL) has been recently introduced to TAL [37, 38, 41]. A few-shot learning model is designed to eliminate the annotation of large training data. This is achieved by meta-learning which enables a model to adapt to any new class with as few as a single video. One of the key challenges in FS-TAL is how to capture the intra-class variation

© 2021. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

using only a handful (*e.g.*, 1-5) training instances of a new class. One of the key objectives of meta-learning is to thus transfer such intra-class variation information from a large set of seen training classes to the new class to compensate for lack of training data.

Nonetheless, existing few-shot TAL (FS-TAL) methods [37, 38, 41] all adopt a setting under which trimmed videos are used to represent the new classes for model adaptation. This setting seems to problematic: (1) As mentioned earlier, the TAL problem exists because most action instances are first captured in untrimmed videos sandwiched by background segments. An analogy is that objects always co-exist with background (*e.g.*, tree/road/wall) in an image. So to obtain the trimmed new class video, one needs to first manually annotate (trim) the untrimmed videos. This begs the question: *why not use the untrimmed video together with the annotation for model adaptation?* (2) Each new action class occurs in its own specific context (background), which carries important cues on how to segment it. Using trimmed videos means that a FS-TAL model is unable to exploit the contextual information for both knowledge transfer from seen classes and new unseen class adaptation.

In this work, we first introduce a new and more practical few-shot TAL (FS-TAL) problem setting. During both the training (meta-learning) and inference (model adaptation) stages, each class is represented by a support set comprising untrimmed videos with temporal annotation. A segmentation model is then built using the support set and applied to a query set containing untrimmed videos of the same class to locate the foreground action instances. This change of setting means that instead of meta-learning a model to temporally align the support set instances with the foreground segments of the query video as in [37, 38, 41], we aim to meta-learn a foreground/background classifier that can be quickly adapted to new classes. To this end, we propose a novel FS-TAL model which meta-learns a query adaptive Transformer (QAT) for fast adaption of foreground/background classifier to a new class. In particular, this leverages the attention mechanism across the query video and few-shot classifier in order to better capture the intra-class invariance. As shown in Figure 1, our model has two key components, a snippet classifier that labels each video snippet into foreground or background, and a query adaptation module designed for query video adaptation. The former is a simple binary classifier constructed using the annotated untrimmed support set videos. The latter is formulated as a Transformer that takes both the classifier weight vector and query video features as input and outputs an updated classifier adapted to each query video. This QAT module is meta-learned and fixed during inference; therefore the whole model is inductive. Importantly, our model is flexible in that it can work in both the new setting proposed in this paper and the existing setting with trimmed support set. We make the following contributions: **(1)** We introduce a new and more practical FS-TAL problem setting. **(2)** We propose a novel FS-TAL model with a query adaptive Transformer for model adaptation to both a given new class and each query video. **(3)** Extensive experiments show that the proposed method yields new state-of-the-art performance on two TAL datasets (ActivityNet-v1.3 and Thumos'14). Under a more challenging and more realistic cross-domain setting, the advantage of our method remains.

2 Related Works

Temporal Action Localization An intuitive approach to temporal action localization (TAL) is based on sliding window – first generating multi-scale segments and then classifying them [6]. A key limitation with this pipeline is that a large number (thousands) of possible segments are necessary for achieving reasonable accuracy, which is computationally expensive. To overcome this issue, foreground/background modeling is introduced to generate action

proposals [10, 19, 20, 24, 24, 43]. When proposal generation is poor, incorporating sliding windows could be helpful [11]. For improving local segment-level feature representation, [36, 39] exploit graph convolutional networks to capture long-range contextual information. Nonetheless, assuming a pre-collected dataset of all action class during training, all these methods have poor scalability to large number of classes, due to the high annotation cost.

Few-shot Learning For fast adaptation of a model to any given new class with few training samples, few-shot learning (FSL) provides a solution [25, 27, 31]. It is often realized by meta-learning which simulates the behaviour of new tasks with novel classes represented by only a handful of labeled samples. This eliminates the requirement of labeling a large dataset for a new class. Representative approaches include hallucination (data augmentation) [12, 34], initialization optimization [9, 21, 22], metric learning [17, 27]. Beyond image classification, FSL has also been introduced to object detection [7, 13, 15] and semantic segmentation [23, 32, 40, 42]. In contrast to these image analysis problems, here we focus on the more challenging TAL problem. Note that the model in [13], though developed for object detection in images, can also work in the FS-TAL setting with trimmed support set. More specifically, unlike our query adaptive Transformer for classifier adaptation at the sample level, it leverages self-attention to contrast the regional features exhaustively across the query and support samples. We will compare with [13] in our experiments (Table 1).

Few-shot Temporal Action Localization FSL has been introduced to temporal action location recently [37, 38, 41]. Yang et al. [37] propose the first FS-TAL setting with trimmed support set. It incorporates the sliding window idea in the matching network [31] to localize action instances in untrimmed query videos. Later on, Zhang et al. [41] consider weak video-level annotation of untrimmed training videos. Similar to our proposed setting, the latest work [37] also focuses on a singe new class at one time. However, a common limitation with these existing FS-TAL problem settings stems from the assumption of trimmed support set. As explained earlier, trimmed videos do not exist naturally and need to be obtained with the same amount of manual annotation as our setting. Importantly, the ignorance of background content in the original untrimmed video leads to the failure to exploit useful context information. We will compare the two FS-TAL settings in our experiments (Table 1).

3 Proposed Methodology

Problem Formulation Given only a few videos from any unseen action class, we aim to learn a TAL model for that class. For FS-TAL, we assume a base category set C_{base} for training, and a novel category set C_{novel} for testing. For testing cross-class generalization, we ensure that the two class sets are disjoint: $C_{base} \bigcap C_{novel} = \varnothing$. Accordingly, the base and novel datasets are denoted as $D_{base} = \{(V_i, Y_i), Y_i \in C_{base}\}$ and $D_{novel} = \{(V_i, Y_i), Y_i \in C_{novel}\}$ respectively. Under the proposed new setting, each training video V_i is associated with segment-level annotation $Y_i = \{(s_t, e_t, c), t \in \{1, .., M\}, c \in C\}$ including M segment labels each with the start and end time locations and action class. In evaluation, for each task, we randomly sample a class $L \sim C_{novel}$ from which K and one labeled videos are randomly sampled to construct the support set S and the query set Q respectively. The labels of S are accessible for model few-shot learning whilst that of Q are used for performance evaluation.

3.1 Model Architecture

Our FS-TAL model is illustrated in Figure 1. It consists of a task-generic video embedding module (Sec. 3.2), and a task-specific snippet classification module (Sec. 3.3). We aim to

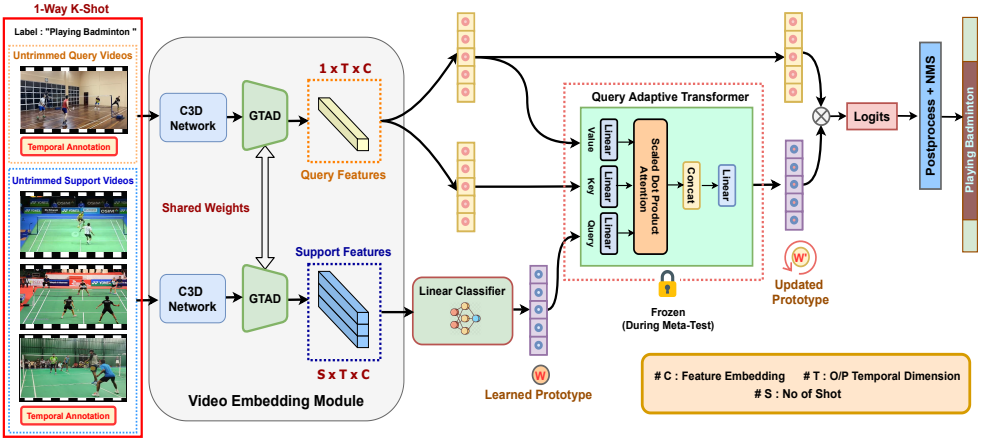


Figure 1: **Overview of the proposed FS-TAL deep learning architecture.** There are two main modules: (1) Video embedding for feature representation: It is pre-trained on the whole training set, and shared by all different tasks for more effective knowledge transfer from training classes to test classes. (2) Snippet classification for foreground prediction: It is learned specifically for every individual task in two steps. Initialized with the average of foreground snippet features, the first step learns the classifier on the support videos in a supervised manner. The second step further adapts the classifier weights to every query video with a query adaptive Transformer. The Transformer is meta-trained. The final localization result is obtained by thresholding snippet-level classification scores and temporal non-maximum suppression.

achieve optimal video embedding and classification for any new task with only a few (1 or 5) labeled support videos. To that end, we share video embedding component across all tasks, and exploit the classification component for tackling the task specificity. With the output of task adapted classification on every snippet of a test video, we apply a non-parametric localization process to obtain the segment predictions (Sec. 3.4).

3.2 Task-Generic Video Embedding

To capture action location information of a video, we construct a video embedding module with two components including feature backbone and snippet embedding.

Feature backbone In general, any video action models can be used such as C3D [29], I3D [4] and TSM [18]. For fair comparison with [33], we adopt the same backbone C3D as our default choice. It is characterized by conducting 3D convolution and pooling operations in 2D spatial and 1D temporal dimensions simultaneously, capturing both appearance and motion information. Given an input video V, we extract RGB $X_r \in \mathbb{R}^{T\times d_1}$ and optical flow $X_o \in \mathbb{R}^{T\times d_1}$ features at the snippet level, where T denotes the number of snippets and d_1 denotes the feature dimension. Each snippet is a short sequence of (*e.g.*, 16 in our case) consecutive frames. We denote the concatenated features as $X = [X_r; X_o] \in \mathbb{R}^{T\times 2d_1}$. As in most TAL methods [20, 35, 36], the feature backbone is pre-trained on a large video classification dataset (*e.g.*, Kinetics [16]) and then frozen to serve as a feature extractor.

Snippet embedding Whilst C3D features have already encoded local motion information due to using 3D convolution and optical flow, long-term structural information is lacking but critical for action localization. To address this issue, we adopt an off-the-shelf tempo-

ral proposal model called GTAD [36]. Other proposal models can be similarly integrated [19, 20, 36]. In particular, GTAD exploits a temporal graph and a semantic graph for modeling long-term temporal and contextual information concurrently. In our context, we utilize GTAD as a means for refining the C3D snippet features in a way that they become more sensitive to foreground (action content) and background. We use the output of some intermediate layer of GTAD as the snippet embedding. The layer selection will be evaluated in our experiments (Sec. 4).

Formally, taking C3D features $X \in \mathbb{R}^{T\times 2d_1}$ of a video as input, GTAD can output the snippet embedding as $X_{se} \in \mathbb{R}^{T\times C}$ where C is the embedding dimension. Support and query videos share the same GTAD model. We denote X^s_{se} and X^q_{se} as the embedding of the support and query videos. Consider that snippet embedding would be largely shareable among different tasks, we train the GTAD model on the base dataset in a standard supervised learning way. The objective function includes a classification loss and a localization loss with respect to the ground-truth foreground mask [36]. The trained GTAD and C3D form the video embedding module which provides generic video representations for the subsequent few-shot learning stage.

3.3 Task-Specific Snippet Classification

In our architecture, few-shot learning is focused on the snippet classification component for capturing each task's specificity. We aim to build a binary classifier h_ϕ (with ϕ the parameters) that can distinguish foreground action from background content in a video. Formally, the classifier model for predicting the foreground likelihood is a simple linear classifier as

$$p(t) = h_\phi(X_{se}(t)) = \sigma(\tau \cdot [cos(X_{se}(t), \phi)]), \tag{1}$$

where σ specifies the sigmoid function, τ is a temperature hyper-parameter, and cos is the cosine similarity. The snippet is indexed by $t \in \{1, \cdots, T\}$.

To make a classifier discriminative for each specific task, we introduce a two-step learning-and-adapting strategy. In the first step, we learn the classifier weights on the support set in a supervised way. In the second step, we further adapt the support-set trained classifier weight to every query video with a query adaptive Transformer model. This aims to solve the intra-class variation problem.

New class adaptation As the support set is composed of untrimmed videos with segment-level annotation, we can adapt the classifier to a new class with standard supervised learning. Given the ground-truth annotation, we label each snippet with foreground or background. To train the classifier, we use the cross entropy loss as the objective function:

$$L_{ce} = -\frac{1}{2K}\sum_{k=1}^{K}[L_{fg}(X^s_k) + L_{bg}(X^s_k)], \tag{2}$$

$$L_{fg}(X^s_k) = \frac{l_{fg}+l_{bg}}{\varepsilon + l_{fg}} \sum_{t\in\{1,\cdots,T\}} \hat{y}^s_k(t)\log[p^s_k(t)], \tag{3}$$

$$L_{bg}(X^s_k) = \frac{l_{fg}+l_{bg}}{\varepsilon + l_{bg}} \sum_{t\in\{1,\cdots,T\}} (1-\hat{y}^s_k(t))\log[1-p^s_k(t)], \tag{4}$$

where $p^s_k(t)$ is the prediction of the t-th snippet $X^s_k(t)$ from the k-th support video. ε is used to tackle extreme cases such as zero background/foreground. To balance the effect of

foreground and background snippets in training, we introduce a balancing policy based on their sizes (l_{fg} and l_{bg}). The idea is intuitive – less is more important.

The classifier can be trained by a small number of (*e.g.*, 50∼100) iterations. We denote ϕ^* as the support-set trained classifier's weights. Given only a handful of labeled samples, how to initialize the classifier weights becomes more critical. Instead of random initialization, we found that the mean of foreground snippet's embedding serves as a stronger choice.

Query video adaptation Under the few-shot setting, a key challenge to overcome is the insufficient training samples in the support set for capturing the intra-class invariance of the new class. As a result, training the classifier only on the support videos often fails to capture the discriminative informative generalizable to individual query videos. To address this limitation, we propose a query adaptive Transformer model (with the parameters ψ) which is based on self-attention [30].

Taking an input in a triplet of $(query, key, value)$, our Transformer outputs an undated *query* with attentive association the *value*. As our objective is to associate the classifier weights ϕ^* with the query video X^q_{se}, we set $(query, key, value) = (\phi^*, X^q_{se}, X^q_{se})$. The attentive learning is then formulated as

$$A_i(\phi^*) = \phi^* + softmax(\frac{\phi^* W_Q (X^q_{se} W_K)^T}{\sqrt{d}})(X^q_{se} W_V), \tag{5}$$

where $W_Q/W_K/W_V$ are learnable parameters (each is realized by a fully-connected layer) that projects the respective input to a d-dimension latent space. In a multi-head attention (MA) design, we combine a set of independent A_i to form a richer learning process:

$$\phi^{**} = \underbrace{[A_1(\phi^*)..A_m(\phi^*)]}_{MA} + MLP(\phi^*) \in \mathbb{R}^{L\times 256}. \tag{6}$$

The MLP block has one fully-connected layer with residual skip connection. Layer norm is applied before both the MA and MLP block.

Learning objective After the classifier has been learned on both support and query videos, it can be applied to the query video. We classify each snippet with Eq. (1) with the foreground probability as:

$$p'(t) := h_{\phi^{**}}(X^q_{se}(t)). \tag{7}$$

For training our Transformer (ψ), this prediction is then used to compute a cross-entropy loss (Eq. (2)) as objective. In meta-training, we conduct loss gradient back-propagation only once for each episode. We denote ψ^* as the optimized Transformer's parameters.

3.4 Model Inference

During testing, each time we are given a new task with one random unseen action class sampled from the novel dataset D_{novel}. With the frozen video embedding module, we need to obtain a task-specific snippet classifier in two steps: supervised learning with K shots of support videos (Eq. (2)) and classifier weight adaptation on a query video by applying the meta-trained Transformer (note that our Transformer itself is frozen here). The classifier is then applied to predict the foreground probability of every snippet of a query/test video.

Action instance generation After we obtain the snippet-level classification results, we threshold on their foreground probabilities and take those consecutive snippets as action instance candidates. To indicate the prediction confidence of each candidate, we use the highest snippet foreground probability. We then adjust the confidence scores using temporal soft Non-Maximal Suppression (NMS) [1, 20]. Finally, we select top N candidates as the localization result.

4 Experiments

Datasets We evaluate on two large-scale temporal action localization datasets. **ActivityNet-v1.3** [3] is a popular TAL benchmark. It contains 19,994 temporally annotated untrimmed videos in 200 action categories. **THUMOS'14** [14] is another widely used benchmark for action recognition and localization. There are 413 untrimmed videos from 20 different categories. The 20 classes are a subset of the 101 classes in UCF101 [26].

Few-shot learning setting To facilitate performance comparison, we use the same class split as introduced in [38]. For both datasets, we split the videos into single instance and multi-instance according to the number of action instances per video. For the single instance case, we divide the videos with multiple action instances into independent single-instance videos. Every newly generated video is no longer than 768 frames. For each of the two cases, we divide all the classes into three non-overlapping subsets for training (80%), validation (10%) and testing (10%), respectively. The validation set is used for model parameter tuning and best model selection. We consider 1-shot and 5-shot. In our setting we adopt untrimmed support set, as opposed to [38] using trimmed videos. For each test, we use 5000 random tasks and report their average result.

Implementation Details For each untrimmed video, we extract its RGB frames at 16 FPS and at the resolution of 256 × 256. We averagely divide each video into 100 (256 for THUMOS) non-overlapping snippets and sample 8 frames for each snippet (*i.e.*, $T = 100$). As [38], we filter out videos having less than 768 frames. We consider single-instance and multi-instance test videos, separately. The dimension of C3D feature is 500 (*i.e.*, $d_1 = 500$). We take the penultimate layer's output (Layer-5) of GTAD's localization module as video embedding (256-D). The latent feature dimension d (Eq. (5)) of our query adaptive Transformer is 256. Dropout is used in our Transformer to alleviate model overfitting. We set the NMS threshold of 0.7/0.6 for ActivityNet/THUMOS. As the final TAL result, we take top 100/200 for ActivityNet/THUMOS. We adopt the Adam optimizer [28] with learning rate 0.004. We train the model for 50 epochs each with 200 episodes.

	Single instance videos												**Multi-instance videos**											
	ActivityNet-v1.3						THUMOS'14						ActivityNet-v1.3						THUMOS'14					
map@	0.5	0.6	0.7	0.8	0.9	mean	0.5	0.6	0.7	0.8	0.9	mean	0.5	0.6	0.7	0.8	0.9	mean	0.5	0.6	0.7	0.8	0.9	mean
1 Shot													*1 shot*											
Hu et al. [13]	41.0	33.0	27.1	15.9	6.8	24.8	-	-	-	-	-	-	29.6	23.2	12.7	7.4	3.1	15.2	-	-	-	-	-	-
Feng et al. [8]	43.5	35.1	27.3	16.2	6.5	25.7	-	-	-	-	-	-	31.4	25.5	16.1	8.9	3.2	17.0	-	-	-	-	-	-
Yang et al. [38]	53.1	40.9	29.8	18.2	8.4	29.5	48.7	-	-	-	-	-	42.1	36.0	18.5	11.1	7.0	22.9	-	-	-	-	-	-
Ours	55.1	**45.2**	35.5	**25.3**	**13.2**	**32.5**	49.2	36.9	**24.3**	**16.5**	**10.1**	**27.2**	44.1	37.8	**29.5**	21.4	**11.5**	25.8	7.3	4.2	3.1	2.0	1.5	3.7
Ours†	**55.6**	44.6	**35.7**	24.6	12.7	31.8	**51.2**	**38.1**	22.7	14.8	9.2	27.0	**44.9**	**38.0**	29.2	**21.4**	11.2	**25.9**	**9.1**	**6.8**	**4.9**	**3.5**	**2.3**	**5.3**
5 Shot													*5 shot*											
Buch et al. [1]	39.7	33.6	27.0	14.0	4.6	23.3	35.7	29.4	20.8	11.7	3.4	20.2	30.4	25.1	19.6	12.9	6.6	18.9	2.7	1.9	1.4	0.9	0.4	1.5
Hu et al. [13]	45.4	35.0	29.9	17.6	5.2	27.0	42.2	32.6	20.3	13.7	5.2	22.8	38.9	27.2	18.3	12.7	7.3	20.9	6.8	3.1	2.2	1.8	1.3	3.1
Yang et al. [38]	56.5	47.0	37.4	21.5	11.9	34.9	51.9	42.7	24.4	17.7	10.1	29.3	43.9	37.4	20.2	13.4	7.7	24.5	8.6	5.6	3.8	2.5	1.7	4.4
Ours	63.0	**54.5**	**44.2**	30.9	15.8	38.4	54.3	43.6	**35.8**	**24.5**	12.2	31.6	48.2	39.1	29.7	22.5	**12.8**	28.2	10.4	7.1	5.7	4.8	2.9	5.4
Ours†	**63.8**	54.2	43.9	**31.4**	**16.4**	**38.5**	**56.1**	**47.2**	32.4	24.3	**13.7**	**32.7**	**51.8**	**42.7**	**32.6**	**23.4**	11.9	**30.2**	**13.8**	**11.3**	**8.4**	**6.3**	**4.2**	**7.1**

Table 1: FS-TAL results (%). †: Using untrimmed support set (i.e., the new setting).

4.1 Comparison with state-of-the-art

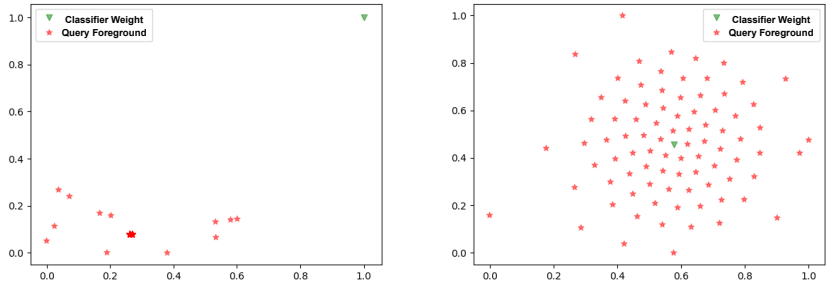
Competitors For comparative evaluation, we consider a few-shot object detection model [13], a one-shot video re-localization model [8], and the latest FS-TAL model [38]. Because [8] cannot tackle multiple support videos, we compare with a modified version of temporal action proposal model SST [2] for 5-shot case. As in [38], a fusion layer is added on top of SST's GRU layer to incorporate the support video features, and the proposal with the largest confidence score is taken as the prediction. All the methods use the same C3D video feature backbone. For all the competitors, we use trimmed support set to keep their original designs. We evaluate the proposed model under both the previous setting (trimmed support set) and our new setting (untrimmed support set). This allows for absolute fair model comparison as well as setting comparison. When feeding trimmed support videos into our model, the background loss term L_{bg} in Eq. (2) will become zero; without any other formulation change, our model can be applied to the old setting. The difference is that now the Transformer is used to adapt a foreground template/prototype to each query video, instead of a foreground/background classifier. Note that none of the existing methods can be easily extended to operate under our new setting.

Results The results are compared in Table 1. It is evident that our method achieves the best performance in all test settings when using the same trimmed support set. This suggests the superiority of our model over all alternative designs, verifying the proposed few-shot learning architecture. The margin is even larger in more strict metrics. Importantly, we see that the margin further increases in 5-shot case, indicating the superior capability of our method in leveraging multiple training videos. This is mainly due to the proposed query adaptive Transformer that can amplify the benefit of larger support-set via attentive query video adaptation, which is lacking in all existing methods. In the multi-instance setting on THUMOS'14, all the methods do not work well due to longer videos and short action instances. However, it is still encouraging that our model can double or triple the performance of alternatives at mAP@0.6-0.9 in such challenging test.

We further examine the two FS-TAL problem settings with the proposed method. We make the following observations. In the single-instance setting, the model performance is marginally better in previous setting with trimmed videos in most cases. Our observation suggests that this is potentially due to lack of background diversity. However, when it comes to the more practical and challenging multi-instance setting, the opposite is true especially in the 5-shot case. This indicates that background helps model learning with useful context cues co-existing with action instances. Given these observations, we consider that the proposed setting is not only more practical but also provides more information for better modeling, as compared to the previous settings.

4.2 Effect of Query Video Adaptation

In Section 3.3 we introduce a query adaptive Transformer for fast adapting the support-set trained classifier's weights to each query video. This aims to solve intra-class variation with FS-TAL as there is no sufficient training samples in support set to capture this variation. There may exist big appearance difference between the support and query video action instances (see Figure 2 in Supplementary). Query video adaptation is thus critical. From Table 2 we can see that without the proposed query video the performance will drop significantly ($3 \sim 8\%$) in 1/5-shot settings of both datasets. This verifies the importance of learning the intra-class invariance problem and the ability of our Transfer model in adapting the classifier's weight to each query video, *i.e.*, video-specific adaptation. In Figure 2 we visually



(a) Before Query Adaptation (b) After Query Adaptation

Figure 2: **The effect of Query Video Adaptation with t-SNE visualization.** It is shown that with the proposed query video adaptation, the classifier weight can be effectively pushed to be aligned with the foreground content of the query video sample. This improves learning the intra-class invariance of the new class.

show that our query adaptive transformer is effective in adapting the classifier's weight to capture the specificity of the query video's foreground content.

Dataset	ActivityNet		Thumos	
mAP	0.5	mean	0.5	mean
Without QVA				
Ours @ 1-shot	37.3	21.7	3.6	2.3
Ours @ 5-shot	43.8	25.3	7.9	4.0
With QVA				
Ours @ 1-shot	44.9 (↑ **7.6**)	25.9 (↑ **4.2**)	9.1 (↑ **5.5**)	5.3 (↑ **3.0**)
Ours @ 5-shot	51.8 (↑ **8.0**)	30.2 (↑ **4.9**)	13.8 (↑ **5.9**)	7.1 (↑ **3.1**)

Table 2: **Effect of query video adaptation (QVA)** in the multi-instance setting.

4.3 Cross-Domain Localization

Following the above single domain (dataset) FS-TAL evaluation, we further introduce a more challenging and more realistic cross-domain setting. As THUMOS'14 and ActivityNet-v1.3 present large differences in action instance length and background characteristics, they are suitable for cross-domain evaluation. We consider the single-instance setting. We compare our method with the state-of-the-art model [38].

THUMOS → ActivityNet In the first cross-domain experiment, we train a model on the base classes of THUMOS'14 (source domain) and test the model on the novel classes of ActivityNet-v1.3 (target domain). The results are reported in Table 3. It is shown that the performance advantage of our method remains compared to the single-domain setting. For example, the mAP@0.5 margin of our model over [38] is 4.8%/6.2% in the 1/5-shot cases. Comparing the single-domain results in Table 1, we can see that domain shift indeed negatively affects the performance of both models.

ActivityNet → THUMOS The second experiment considers the opposite transfer direction. At large we have similar observations with our model again outperforming [38] in both 1/5-shot setting. This suggests that our model can generalize to different transfer setups with consistent performance advantages.

Cross Domain	**Thumos → ActivityNet**		**ActivityNet → Thumos**	
mAP	0.5	mean	0.5	mean
Yang et al. [38] @ 1-shot	41.1	25.2	36.2	21.4
Ours @ 1-shot	**45.9**	**26.6**	**38.1**	**22.5**
Yang et al. [38] @ 5-shot	48.2	27.8	37.5	23.6
Ours @ 5-shot	**54.4**	**31.6**	**43.8**	**27.2**

Table 3: **Cross-domain FS-TAL.**

4.4 Effect of Video Embedding Module

We evaluate the generality of our FS-TAL architecture in different video embedding designs. In this test we select BMN [20]. Table 4 shows that BMN is slightly inferior to GTAD for video embedding, which is consistent with the previous finding [36].

4.5 Inference Efficiency

In inference, our model runs a small number of iterations for learning the linear classifier's weights on the support set, which increases slightly the computational overhead. We conduct a quantitative cost analysis in 5-shot multi-instance setting on ActivityNet-v1.3. We compared to the state-of-the-art model [38]. For both methods, we track the speed of 100 FS-TAL tasks on a machine with one RTX2080Ti GPU. Table 5 shows that our method has very similar inference speed as [38], without efficiency disadvantage.

Dataset	ActivityNet-v1.3	
mAP	0.5	mean
BMN [20]	61.6	37.5
GTAD [36]	63.8	38.5

Table 4: **Effect of video embedding** in the 5-shot multi-instance setting.

Dataset	ActivityNet-v1.3
Metrics	Speed (seconds / task)
Yang et al. [38]	0.81
Ours	0.83

Table 5: **Inference efficiency test** in the 5-shot multi-instance setting with a RTX2080 GPU.

5 Conclusion

We have presented a new and more practical few-shot temporal action localization (FS-TAL) problem. Unlike all existing settings, in our setting a new action class is represented by untrimmed support set with useful background segments to provide contextual information. We introduce a novel FS-TAL architecture that effectively transfers class-generic representation knowledge from training classes to any unseen test classes whilst adapting the model to any new class. To solve the large intra-class variation problem, we introduce a query adaptive Transformer that further dynamically adapts the support-set trained classifier's weights to each query video. Experiments on two popular TAL datasets verify the superiority of our method over existing alternatives in both the newly proposed setting with untrimmed labeled support set and previous settings with trimmed counterpart. Moreover, our method remains to be advantageous under a more realistic and challenging cross-domain setting.

References

[1] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *ICCV*, 2017.

[2] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. Sst: Single-stream temporal action proposals. In *CVPR*, 2017.

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015.

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017.

[5] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019.

[6] Xiyang Dai, Bharat Singh, Guyue Zhang, Larry S Davis, and Yan Qiu Chen. Temporal context network for activity localization in videos. In *ICCV*, 2017.

[7] Xuanyi Dong, Liang Zheng, Fan Ma, Yi Yang, and Deyu Meng. Few-example object detection with model communication. *TPAMI*, 41(7), 2018.

[8] Yang Feng, Lin Ma, Wei Liu, Tong Zhang, and Jiebo Luo. Video re-localization. In *ECCV*, 2018.

[9] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *ICML*. PMLR, 2017.

[10] Jiyang Gao, Zhenheng Yang, Kan Chen, Chen Sun, and Ram Nevatia. Turn tap: Temporal unit regression network for temporal action proposals. In *ICCV*, 2017.

[11] Jiyang Gao, Kan Chen, and Ram Nevatia. Ctap: Complementary temporal action proposal generation. In *ECCV*, 2018.

[12] Bharath Hariharan and Ross Girshick. Low-shot visual recognition by shrinking and hallucinating features. In *ICCV*, 2017.

[13] Tao Hu, Pascal Mettes, Jia-Hong Huang, and Cees GM Snoek. Silco: Show a few images, localize the common object. In *CVPR*, 2019.

[14] 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.

[15] Bingyi Kang, Zhuang Liu, Xin Wang, Fisher Yu, Jiashi Feng, and Trevor Darrell. Few-shot object detection via feature reweighting. In *ICCV*, 2019.

[16] 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.

[17] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015.

[18] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *CVPR*, 2019.

[19] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. Bsn: Boundary sensitive network for temporal action proposal generation. In *ECCV*, 2018.

[20] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *CVPR*, 2019.

[21] Sachin Ravi and Hugo Larochelle. Optimization as a model for few-shot learning. 2016.

[22] Andrei A Rusu, Dushyant Rao, Jakub Sygnowski, Oriol Vinyals, Razvan Pascanu, Simon Osindero, and Raia Hadsell. Meta-learning with latent embedding optimization. *arXiv preprint arXiv:1807.05960*, 2018.

[23] Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. In *BMVC*, 2017.

[24] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *CVPR*, 2016.

[25] Jake Snell, Kevin Swersky, and Richard S Zemel. Prototypical networks for few-shot learning. *arXiv preprint arXiv:1703.05175*, 2017.

[26] 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.

[27] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, 2018.

[28] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *ICML*. PMLR, 2013.

[29] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[31] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra. Matching networks for one shot learning. *arXiv preprint arXiv:1606.04080*, 2016.

[32] Haochen Wang, Xudong Zhang, Yutao Hu, Yandan Yang, Xianbin Cao, and Xiantong Zhen. Few-shot semantic segmentation with democratic attention networks. In *ECCV*, 2020.

[33] Limin Wang, Yuanjun Xiong, Dahua Lin, and Luc Van Gool. Untrimmednets for weakly supervised action recognition and detection. In *CVPR*, 2017.

[34] Yu-Xiong Wang, Ross Girshick, Martial Hebert, and Bharath Hariharan. Low-shot learning from imaginary data. In *CVPR*, 2018.

[35] Huijuan Xu, Abir Das, and Kate Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *ICCV*, 2017.

[36] Mengmeng Xu, Chen Zhao, David S. Rojas, Ali Thabet, and Bernard Ghanem. G-tad: Sub-graph localization for temporal action detection. In *CVPR*, June 2020.

[37] Hongtao Yang, Xuming He, and Fatih Porikli. One-shot action localization by learning sequence matching network. In *CVPR*, 2018.

[38] Pengwan Yang, Vincent Tao Hu, Pascal Mettes, and Cees GM Snoek. Localizing the common action among a few videos. In *ECCV*. Springer, 2020.

[39] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. Graph convolutional networks for temporal action localization. In *ICCV*, 2019.

[40] Chi Zhang, Guosheng Lin, Fayao Liu, Rui Yao, and Chunhua Shen. Canet: Class-agnostic segmentation networks with iterative refinement and attentive few-shot learning. In *CVPR*, 2019.

[41] Da Zhang, Xiyang Dai, and Yuan-Fang Wang. Metal: Minimum effort temporal activity localization in untrimmed videos. In *CVPR*, 2020.

[42] Xiaolin Zhang, Yunchao Wei, Yi Yang, and Thomas S Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *IEEE Transactions on Cybernetics*, 2020.

[43] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *ICCV*, 2017.