Temporal Alignment via Event Boundary for Few-shot Action Recognition

Shuyuan Li[1]
shuyuanli@sjtu.edu.cn
Huabin Liu[1]
huabinliu@sjtu.edu.cn
Mengjuan Fei[2]
feimengjuan1@huawei.com
Xiaoyuan Yu[2]
yuxiaoyuan@huawei.com
Weiyao Lin[1]
wylin@sjtu.edu.cn

[1] Shanghai Jiao Tong University
Shanghai, China
[2] Huawei Cloud
China

Abstract

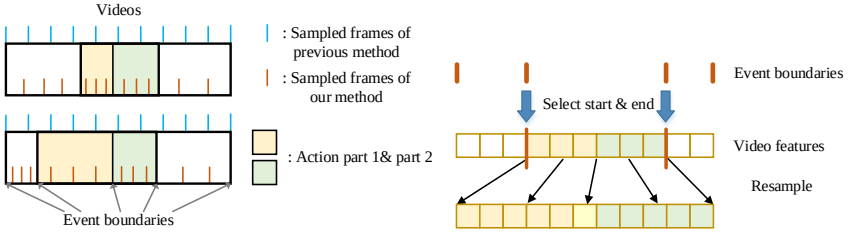
Few-shot action recognition aims to recognize novel action classes using just a few samples as knowledge. Most of the recent approaches learn to compare the similarity between videos. Recently, it has been observed that directly measuring this similarity is not ideal since the action instance cannot well aligned among videos. In this paper, we leverage the novel event boundary information to guide alignment learning in few-shot action recognition. First, a novel frame sampling strategy based on temporal boundaries is proposed to relieve the intra-class variance. Second, we propose a boundary selection module to locate the start & end time of action and further align videos to their duration. Ablation studies and visualizations demonstrate the effectiveness of the proposed methods. Extensive experiments on benchmark datasets show the potential of the proposed method in achieving state-of-the-art performance for few-shot action recognition.

1 Introduction

With the application of video analysis, action recognition has attracted much attention from researchers. Recently, data-driven deep learning has advanced the frontiers of this field. However, numerous labeled data is hard and expensive to obtain in real-world scenes. Therefore, few-shot action recognition task has received increasing interest [1, 2, 3, 18, 19], which aims to learn to recognize novel action categories using few samples for training.

The major line of existing works on few-shot action recognition follows the metric-learning paradigm [1, 2, 19], which is widely adopted in previous general few-shot learning (FSL) methods [10, 11, 15]. Specifically, they measure the distance or similarity between

[1]The paper is supported in part by the following grants: National Key Research and Development Program of China Grant (No.2018AAA0100400), National Natural Science Foundation of China (No. 61971277).



(a) Boundary-guided frame sampling. (b) Boundary-guided temporal alignment.

Figure 1: Illustration of our method compared with previous method. In both figures, we highlight the action subparts(e.g, approach & jump in high jump) with colors and different parts have different colors.

videos with learned metrics and embedding feature space. Recently, some approaches [2, 3, 18] reveal that directly measuring similarities between videos is challenging due to the *temporal misalignment* of action instances. The misalignment attributes to the fact that the temporal locations, *duration*, and *evolution* (the process of movement) of actions vary in videos. Some methods [2, 3] try to address this issue by learning an alignment among video features or metrics.

Nevertheless, two issues remain unresolved: *First*, previous methods adopt the frame sampling strategy introduced in TSN [16], which samples frames uniformly across the whole video. However, since the action duration varies from videos, this may sample a distinct number of frames for different action instances (as illustrated in Figure 1(a)). Thus, uniform sampling inevitably aggravates the problem of misalignment. Moreover, it leads to a larger intra-class variance, which is prejudicial to metric learning. *Second*, for existing methods that try to perform temporal alignment, their alignment lacks effective guidance or restriction during training.

Recently, GEBD [9] introduces the generic event boundary annotations into several video datasets. Different from the common action boundary that marks the start & end time, it divides the whole video into several semantically-coherent action *subparts* with fine-grained event boundaries. Based on such temporal boundaries, we proposed a novel method for few-shot action recognition, which learns the **T**emporal **A**lignment via **E**vent **B**oundary (TAEB). Specifically, we first devise the boundary-based sampling strategy (Figure 1(a)), which samples frame uniformly on each action subparts according to the boundaries rather than the whole video. In this way, it makes the frame-wise representation of action instances among videos more consistent and thus reduces the intra-class variation. Secondly, we propose a boundary selection module (BSM). It selects the duration boundaries (start & end time of action) from all boundary candidates. Video features are then aligned to action duration by a temporal affine transformation. Finally, an attention mechanism is introduced to further refine the alignment. The overall framework of our method is illustrated in Figure 2.

In summary, our main contributions are as follows: (1) We are the first to introduce event boundaries into few-shot action recognition task as a prior to guide temporal alignment. (2) We devise a frame sampling strategy based on temporal boundaries to reduce the intra-class variance and relieve action misalignment collaborating with proposed boundary selection module. (3) A novel boundary selection module is introduced to align video features to their action duration. (4) Extensive experiments conducted on few-shot action recognition datasets show that our proposed method achieve the state-of-the-art results.

2 Related Works

2.1 Action Recognition & Frame Sampling

Methods of action recognition in recent years mainly base on convolutional neural networks (CNNs). C3D [13] and I3D [4] are two representatives of 3D-CNN for video action recognition. C3D-like networks usually divides the whole video into clips, which are usually consecutive 16 frames. However, 3D convolutions bring expensive computational costs and memory demand. Therefore light-weight and efficient version such as P3D [8] and R(2+1)D [14] are proposed. In addition to these methods, some other methods instead of 3D convolution are proposed to capture temporal relation. TSN [16] extracts features from frames by 2D CNN and aggregates features by temporal aggregation function. TSN-based networks use sparse uniform frame sampling, which can represent a video by a few frames, *e.g.* 8 frames per video. Such uniform sampling divides the video into T parts with equal length and uniformly sample n frames in each segment, resulting in nT sampled frames.

2.2 Few-shot Action Recognition

The early study of CMN [19] represents videos by a compound memory network and store features in matrix representation. Representations in the memory can be retrieved and updated. TAEN [1] represents actions as trajectories in the learned feature space while FAN [17] encodes motion in a video into an image named dynamic image.

Due to the different distribution of action instances in videos, similarity directly measured between two videos suffers from misaligned actions and may lead to a trivial metric function. To solve the misalignment, recent works pay more attention to temporal alignment. TARN [2] proposed an attentive relation network to implicitly align actions by segment-wise temporal attention. ARN [18] reduces the temporal dimension by permutation invariant temporal attention and temporal global pooling. OTAM [3] proposes a variant of the Dynamic Time Warpping (DTW) algorithm to measure distance between videos with alignment.

Above methods only use the video frames while our method make use of event boundaries in the video to help the learning of temporal alignment. Moreover, our proposed method could further apply to them, more details could refer to Section 4.4.3.

2.3 Generic Event Boundary

Recently, the GEBD [9](Generic Event Boundary Detection) proposes a novel task and Kinetics-GEBD dataset, in which generic event boundaries segment a whole video into chunks. Conventional works in action detection and temporal segmentation focus on localizing pre-defined action categories and thus does not scale to generic videos. The GEBD aims at localizing the moments where humans naturally perceive event boundaries. And such generic event boundaries can help machine to understand videos since those boundaries segment videos into meaningful units or action sub-parts. Compared with rough action boundaries, generic event boundaries additionally segment actions into atomic parts (*e.g.* run, jump and stand up are three parts in a long jump action), which is useful for understanding and comparing actions.

3 Method

3.1 Problem Setup

We follow the standard few-shot meta-learning paradigm, dividing a dataset into three non-overlapping splits by class: meta training set $\mathcal{C}_{train}$, meta validation set $\mathcal{C}_{val}$, and meta test set $\mathcal{C}_{test}$. The meta validation set is only used to evaluate the model during training. The model is

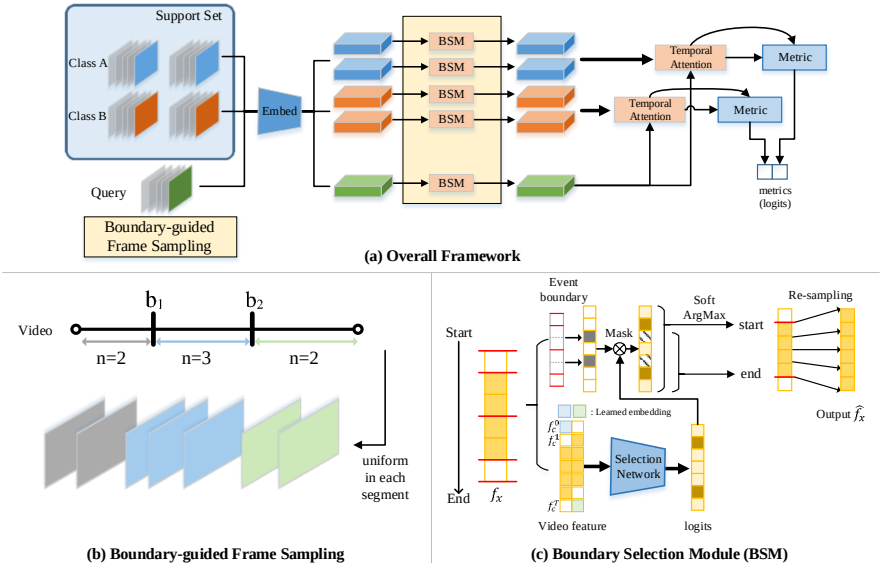


Figure 2: Overall framework illustration of our method. (a) The overall framework. We highlight our boundary-guided modules. **BSM**: Boundary Selection Module. (b) An illustration of our boundary-guided frame sampling, T=8 is shown. (c) The Boundary Selection Module (BSM). Red solid lines are event boundaries and filled blocks are where action exists.

trained and tested in a standard N-way K-shot few-shot learning setting. Under the setting, we randomly sample episodes from the dataset to train and test. An episode consists of support set $\mathcal{S}$ and query set $\mathcal{Q}$. The support set $\mathcal{S}$ contains N classes and K support examples sampled for each class while the query set $\mathcal{Q}$ contains $N \times Q$ query examples for the same N classes with $\mathcal{S}$. The support and query set has no overlap.

3.2 Model

3.2.1 Boundary-guided Frame Sampling

Most few-shot action recognition methods uniformly sample a fixed number of frames from each video to represent the whole video. However, this would aggravate the action misalignment and lead to larger intra-class variance due to the fact that the action duration varies in videos. This kind of intra-class variance is prejudicial to further metric learning. Inspired by GEBD [9], we argue that video representation will be more consistent if we sample the same number of frames for all action sub-parts. Based on this motivation, we devise the following boundary-guided frame sampling strategy with the temporal boundary provided by GEBD.

Specifically, given a video $\{I_1, I_2, \ldots, I_{T_0}\}$ and event boundaries $\{b_1, b_2, \ldots b_n\} \in [0, T_0]^n$, I_t denotes the t-th frame and b_i denotes the i-th event boundary. We first divide the whole video into $(n+1)$ segments with index ranges of $S_0 = I_{[0:b_1]}, \ldots, S_i = I_{[b_i:b_{i+1}]}, \ldots, S_n = I_{[b_n:T_0]}$. To represent a video using a fixed length frame sequence of T frames, we then uniformly sample n_i frames in the S_i segment by

$$n_i = \lfloor \frac{T}{n+1} \rfloor + \mathbb{1}(\lfloor \frac{n+1-r}{2} \rfloor < i < \lfloor \frac{n+1+r}{2} \rfloor), \tag{1}$$

$$r = T - (n+1)\lfloor \frac{T}{n+1} \rfloor, \tag{2}$$

where $\mathbb{1}$ is indicator function. By such assignment,

$$\sum_{i=0}^{n} n_i = T, \ \ \forall (i,j) : |n_i - n_j| \leq 1. \tag{3}$$

Sampled frames are concatenated in order as a fixed length frame sequence $X = I_{t_1}, I_{t_2}, \ldots, I_{t_T}$ to represent the video.

After sampling, an embedding network $f(\cdot)$ is applied on each frame and embeds sampled sequence into T frame-level features $f_x = f(X) = \{f(I_{t_1}), \ldots, f(I_{t_T})\} \in \mathbb{R}^{\mathrm{C}\times\mathrm{T}\times\mathrm{H}\times\mathrm{W}}$. In the following, we use f_s, f_q to represent the video-level feature of the support sample and query sample, respectively.

3.2.2 Boundary Selection Module

The video-level representation f_x still involves some action-irrelevant information (e.g. background frames) beyond the action duration. To address this, we proposed the Boundary Selection Module (BSM) to locate the start & end time of action, then f_x could be aligned to its action duration by a certain transformation.

The start & end boundary of actions often involves transition about video shots or scenes. Thus, we consider each paired adjacent frame feature to learn to distinguish action boundary in BSM. As illustrated in Figure 2(c), the middle time of all two adjacent frames forms the candidates of action start & end time. Naturally, the beginning and end of the whole video are also regarded as candidates. Thus, each video contains $(T+1)$ such candidates $[c_0, ..., c_T]$. For each candidate, we concatenate the features of adjacent two frames as its feature representation: $f_c^i = \texttt{Concat}(f_x^i, f_x^{i+1}) \in \mathbb{R}^{1\times 2C}$. Especially, the f_x^0 and f_x^{T+1} are two learnable embedding added by us. Then, the feature of candidate undergoes the boundary selection network ϕ to obtain the selection logit l for each candidate:

$$F_C = \texttt{Stack}(f_c^0, ..., f_c^T) \in \mathbb{R}^{(T+1)\times 2C} \tag{4}$$

$$\mathbf{l} = \texttt{Softmax}(\phi(F_c)) = [l_0, ..., l_T] \in \mathbb{R}^{T+1} \tag{5}$$

where the selection network ϕ consists of a few 1D-convolution layers, the `Stack` means stack all the features in temporal dimension, $\mathbf{l}$ is selection logits with length of $(T+1)$. Furthermore, we use the event boundaries provided by GEBD [9] to guide the selection of BSM:

$$\hat{l}_i = l_i - (1 - \mathbb{1}(c_i)) \cdot m, \tag{6}$$

where $\mathbb{1}(c_i)$ means whether c_i is labeled as event boundary and m is a hyper-parameter controls the suppression degree. By such masking, uninformative candidates are dismissed and the start & end time could be located accurately.

The candidate with maximum score is regarded as the start & end time of action, and it could be located by argmax operator. For differentiable optimization, we use soft-argmax to approximate the non-differentiable argmax operator:

$$softargmax^{\tau}(l_0, l_1, ...l_T) = \frac{\sum_{i=0}^{T} i \cdot \exp(l_i/\tau)}{\sum_{i=0}^{T} \exp(l_i/\tau)}, \tag{7}$$

where τ is the temperature parameter controlling the smoothness.

The above soft-argmax is performed twice to select the start and end boundaries of action. Noticeably, to keep the correct temporal order of the start and end boundaries, the boundary

detected the first time is regarded as the start time of action. Next, we dismiss candidates before the start boundary in logits for the end boundary localization. Given the selected start and end boundaries, we perform a temporal affine transformation $\mathbf{T}$ on the input feature f_x, which transforms the start and end boundaries to $t = 0$ and $t = T$, respectively:

$$\mathbf{T}(t \mapsto t') = \mathbf{T}(t \mapsto \alpha \cdot t + \beta),\ \mathbf{T}^{-1}(t' \mapsto t) = \mathbf{T}^{-1}(t' \mapsto \frac{1}{\alpha}(t' - \beta)) \tag{8}$$

$$\alpha = \frac{T}{end - start}, \beta = -\frac{start \cdot T}{end - start}, \tag{9}$$

$$\hat{f}_x = \mathbf{T}(f_x). \tag{10}$$

3.2.3 Temporal Attention

To further aggregate and refine the global temporal information, we further perform a self-attention mechanism on videos, which has been widely used in video tasks [2, 17]. Given a pair of features $\hat{f}_s, \hat{f}_q \in \mathbb{R}^{T \times d}$, representing a support feature and a query feature after being aligned by BSM, we perform a cross attention on temporal dimension. To calculate attention map $M \in \mathbb{R}^{T \times T}$, we first project support and query features $\hat{f}_s$ and $\hat{f}_q$ linearly by key head W_k and query head W_q respectively.

$$M = \texttt{softmax}(\frac{(W_k \cdot G(\hat{f}_s))(W_q \cdot G(\hat{f}_q))^T}{\sqrt{dim}}), \tag{11}$$

where $G(\cdot)$ is global average pooling and dim is the channel dimension of feature $G(\hat{f})$. According to the attention map M, query feature could be re-weighted. Besides, we also apply value head projection and residual sum to the support feature $\hat{f}_s$ in order to keep feature-space consistency. The re-weighting process can be formally expressed as:

$$\tilde{f}_q = \hat{f}_q + M \cdot (W_v \cdot G(\hat{f}_q)), \tag{12}$$

$$\tilde{f}_s = \hat{f}_s + W_v \cdot G(\hat{f}_s), \tag{13}$$

3.2.4 Optimization

Following previous work [3], we use the 'diagonal' distance as the metric, which is:

$$D(\tilde{f}_p, \tilde{f}_q) = \sum_{i=1}^{T} 1 - \cos \tilde{f}_{p[i]}, \tilde{f}_{q[i]}, \tag{14}$$

where f_p, f_q represents prototype and query features, $[i]$ means indexing in time dimension. Such a distance is the sum of the cosine distances between the corresponding frames of the two videos. We use the negative distances as logits and use standard cross entropy loss.

4 Experiments

4.1 Datasets and baselines

Datasets We conduct experiments on two datasets in GEBD [9], which are widely used in few-shot action recognition :

HMDB51 [6] contains 6,849 videos divided into 51 action categories. Each category contains at least 101 videos. We follow the protocol of ARN [18], which takes 31/10/10 action classes with 4280/1194/1292 videos for train/val/test.
Kinetics-GEBD [9] : Since Kinetics-GEBD releases about 38k annotations in train and val set, which have limited overlap with the widely used Kinetics-CMN [19] split, we resample a split from released Kinetics-GEBD. Following CMN, we sample 64/24/12 classes for train/val/test set and sample 100 videos for each class in the Kinetics-GEBD dataset.

For both datasets, we use ground truth event boundaries annotations. Since annotations provided in GEBD are raw annotations, we preprocess the raw annotations by selecting the annotation from the annotator with highest f1-score and deleting all non-instantaneous boundaries. Also, we add the start and the end of the video into boundaries.
Competitors Except for common baseline of ProtNet [10], we compare our method with recent FSL action recognition works related to temporal handling with state-of-the-art results, including ARN [18] and OTAM [3]. Comparisons are made on above two datasets instead of Something-V2 and Kinetics-CMN split due to no boundary annotations available in GEBD datasaets [9].

4.2 Implementation Details

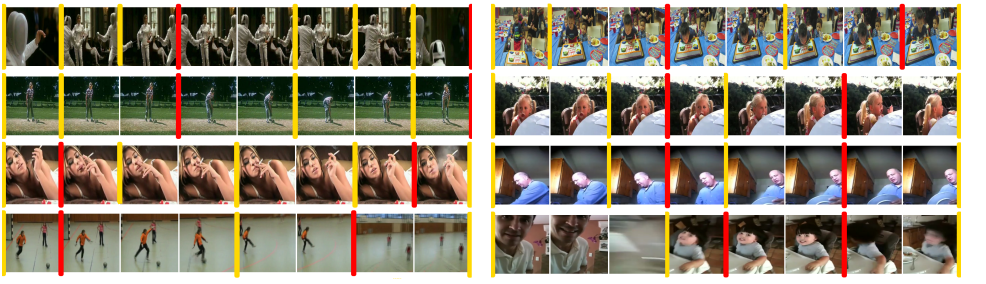
We adapt standard episode style meta-learning way to train and test models. In the N-way K-shot setting, we sample episodes in the way described in subsection 3.1. For each video, boundary-guided frame sampling is used to sample $T = 8$ frames, as described in subsubsection 3.2.1. Then we use ResNet-50 [5] to extract frame-level features so that we can make fair comparison with other works. Sampled frames are first resized to 256×256 with random horizontal flip as augment. Then random crop with the size of 224×224 is applied in training phase. For test phase, the random crop is replaced by a center crop with the same size. We use the ImageNet pre-trained checkpoint of ResNet-50 as the initial state of the feature embedding module and the whole model is trained together in the end-to-end way. For multi-shot settings, We also adapt a similar multi-head temporal attention among support features. Concretely, for K-shot setting, a temporal self attention is applied on total KT support features, which belong to the same class. To keep feature space symmetry, value head projection with residual is also applied on query feature like Eq. (13). Hyper-parameters are described in the supplementary material.

4.3 Main Results

Method	Sampling	HMDB		Kinetics-GEBD	
		1-shot	5shot	1-shot	5shot
ProtoNet [10]	Uniform	54.2	68.4	60.8	70.12
ARN [18]	Uniform	45.5	60.6	-	-
OTAM* [3]	Uniform	54.5	66.1	62.7	74.0
Ours	Bound	**58.6**	**73.8**	**67.0**	**80.9**

Table 1: Results of 5-way 1-shot and 5-shot action recognition accuracy in percent. 'Uniform' means uniform sampling proposed and 'Bound' indicates our proposed sampling methods based on event boundary. The mark * means re-implemented by us.

The Quantitative Results of ours and other competitors on HMDB51 and Kinetics-GEBD are listed in Table 1 As shown in the table, our method outperforms the powerful baseline



(a) HMDB51 (b) Kinetics-GEBD

Figure 3: Visualization results of BSM. Selected boundaries(start and end) are marked with red lines while boundaries not selected are marked with yellow lines. Note that since the selected boundaries are regressed by soft-argmax, they may not locate in annotated boundaries.

ProtoNet by a significant margin. OTAM is the most related SOTA method focusing on temporal relationship and alignment. Compared with it, our method still outperforms it by a large margin under all settings, demonstrating the effectiveness of our boundary-guided temporal alignment.

The Qualitative Results and Visualizations. For boundary selection module, in Figure 3, we visualize selected boundaries and all event boundaries by red and yellow lines respectively on sampled frame sequences. All the visualisation results come from test set. In the visualization, it can be seen that our BSM can learn to locate discriminative part in the video according to event boundaries. It is worth noting that some of the selected boundaries are not in event boundaries like the last row of HMDB51, but they may be more proper. This shows the advantage of soft-argmax and boundary-guided masking, which can trade-off between prior guidance and actual video content. This can somehow prevent the misleading affection of not detected event boundaries.

4.4 Ablation Study

4.4.1 Contribution of each module

Modules	HMDB51	Kinetics-GEBD
TA + UniformSample	57.40	66.29
TA + BoundSample	58.32	66.23
TA + BoundSample +BoundSelection w/o prior	58.05	66.85
TA + BoundSample +BoundSelection w/ prior	**58.60**	**67.01**

Table 2: Breakdown results of modules. For default settings, uniform sampling and TA (Temporal Attention) are used. 'BoundSample' means boundary-guided frame sampling and 'BoundSelection' means boundary selection module. For 'BoundSelection', 'w/o prior' means all possible positions are considered as boundaries (i.e. event boundaries are not used in both training and testing) and 'w/ prior' means only annotated event boundaries are used in training and testing. Results are reported under 5-way 1-shot setting.

We analyze the performance gain of each module by breakdown analysis. Quantitative results of boundary-guided sampling and boundary selection module on both datasets are listed in Table 2. It can be seen that boundary-guided sampling gains about 0.9% on

HMDB51 but makes little difference on Kinetics-GEBD dataset. Since HMDB51 is well-trimmed compared to Kinetics-GEBD, the boundary-guided sampling mainly addresses the variance of action parts on HMDB51, indicating the effectiveness of proposed sampling method. On HMDB51 dataset, the main performance gain comes from boundary-guided sampling while the BSM provides a little about 0.3%. This is natural because HMDB51 is trimmed and the beginning and end of the action are mostly at the beginning and end of the video. However, on Kinetics-GEBD dataset, the BSM provides about 0.7% performance gain. Longer video duration in Kinetics-GEBD makes it more valuable to locate the action instances before comparing. Also, on Kinetics-GEBD dataset, the BSM without prior (i.e. don't perform masking with event boundaries in BSM) gains about 0.5%, and the use of prior (i.e. perform masking with the event boundaries in BSM) further gains about 0.2% and BSM with prior gains about 0.3% on HMDB51 dataset. These indicate that although the BSM can learn to locate the action instance without prior, the use of prior can help the BSM learns and generalizes better.

4.4.2 Performance gain on different baselines

Baseline	HMDB		Kinetics-GEBD	
	Baseline	+BS+BSM	Baseline	+BS+BSM
ProtoNet [11]	54.2	56.5 (+2.3)	60.8	62.5 (+1.7)
OTAM [3]	54.5	56.9 (+2.4)	62.7	63.9 (+1.2)
TA	57.4	58.6 (+1.2)	66.3	67.0 (+0.7)

Table 3: Accuracy gain of BoundSample with Boundary Selection Module on three baselines (ProtoNet, OTAM and temporal attention), reported under 5-way 1-shot setting. BS meas BoundSample and BSM means Boundary Selection Module.

To further prove the effectiveness of proposed boundary-based modules, we apply BoundSample and BSM on different baselines. Results listed in Table 3 show that our proposed boundary-based method can stably improve the performance under different baselines. The gain based on TA is relatively smaller since it's already a strong baseline. Proposed boundary-based modules can bring much higher accuracy improvement on other methods.

4.4.3 Effect of boundary-guided sampling

Sampling method	HMDB51	Kinetics-GEBD
Uniform	1.146	1.414
Boundary	1.144	1.385

Table 4: Intra-class variance of uniform sampling and boundary-guided sampling on HMDB51 and Kinetics-GEBD datasets.

To verify the effectiveness of our proposed sampling strategy, we quantify the intra-class variance of videos under different sampling manner. Specifically, we feed each video to a pre-trained video classification to obtain classification probabilities of all frames, then select the probability vector $P^c = \{p_1^c, p_2^c, \ldots, p_T^c\} \in \mathbb{R}^{1\times T}$ of its ground-truth class as the representation of this video. Based on this, we quantify the intra-class variance as:

$$Var^c = \frac{1}{N_c(N_c-1)} \sum_{i\neq j, y_i=y_j=c} \|P_i^c - P_j^c\|_2, \tag{15}$$

where y_i is the class of the i-th video and N_c is the total number of videos of the c class. Variance under uniform sampling and our boundary-guided sampling are listed in Table 4. As shown in table, boundary-guided sampling reduces the intra-class variance on both datasets. Besides, such variance is reduced more on Kinetics-GEBD. Compared with HMDB51, videos in Kinetics are longer and untrimmed. Such difference makes variance reduced more on Kinetics-GEBD.

Baseline	HMDB		Kinetics-GEBD	
	Baseline	+BoundSample	Baseline	+BoundSample
OTAM [3]	54.5	56.1 (+1.6)	62.7	63.7 (+1.0)
BSM	56.3	56.5 (+0.2)	61.9	62.5 (+0.6)
TA	57.4	58.3 (+0.9)	66.3	66.2 (-0.1)

Table 5: Accuracy gain of BoundarySample on three baselines, reported under 5-way 1-shot setting.

We further perform ablation experiments to verify the accuracy gain of BoundSample as shown in Table 5. In most cases, the BoundSample brings accuracy certain gains, especially when applied on OTAM. The accuracy drop of TA on Kinetics-GEBD indicates that for more complex videos, it is hard to learn a good alignment without any prior guidance. Overall, the combination of BS and BSM obtains the largest.

4.4.4 The use of boundary prior

We change the degree of using boundary prior to explore the effect of boundary prior. If we partly use the boundary prior, then we soft mask the predicted logits by add $-m$ to non-boundary candidates, where m is the penalty value. And using no boundary prior corresponds to $m = 0$, while using strong boundary prior equals setting $m = +\infty$. It is clear that properly using the boundary prior can achieves the best accuracy, while not using prior or too dependent on the prior is harmful.

Dataset	m				
	0.1	1	5	**10**	50
HMDB51	57.11	57.72	57.84	**58.60**	56.54
Kinetics-GEBD	66.34	66.22	66.69	**67.01**	66.78

Table 6: Accuracy under 5-way 1-shot setting w.r.t dependency degree of boundary prior. Settings is the same with Table 1. $m = 10$ is the default setting used previously.

5 Conclusion

Based on recently proposed GEBD datasets, we explore utilizing the event boundary in videos to guide the temporal alignment learning for few-shot action recognition. Specifically, we devise a boundary-guided frame sampling method to generate a more consistent frame-wise video representation. Besides, we propose a boundary selection module (BSM) to locate the discriminative boundary (start and end time) and align videos to their action durations. Experiments on HMDB51 and Kinetics-GEBD datasets demonstrate that our method achieves the start of the art. More ablation experiments also verify the few-shot action recognition benefits a lot from event boundary.

References

[1] Rami Ben-Ari, Mor Shpigel, Ophir Azulai, Udi Barzelay, and Daniel Rotman. Taen: Temporal aware embedding network for few-shot action recognition. *arXiv preprint arXiv:2004.10141*, 2020.

[2] Mina Bishay, Georgios Zoumpourlis, and Ioannis Patras. Tarn: Temporal attentive relation network for few-shot and zero-shot action recognition. *arXiv preprint arXiv:1907.09021*, 2019.

[3] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10618–10627, 2020.

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011.

[7] Chih-Ting Liu, Chih-Wei Wu, Yu-Chiang Frank Wang, and Shao-Yi Chien. Spatially and temporally efficient non-local attention network for video-based person re-identification. In *British Machine Vision Conference*, 2019.

[8] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.

[9] Mike Zheng Shou, Deepti Ghadiyaram, Weiyao Wang, and Matt Feiszli. Generic event boundary detection: A benchmark for event segmentation. *CoRR*, abs/2101.10511, 2021. URL https://arxiv.org/abs/2101.10511.

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

[11] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208, 2018.

[12] Shaoqing Tan and Ruoyu Yang. Learning similarity: Feature-aligning network for few-shot action recognition. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2019.

[13] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[14] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

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

[16] 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.

[17] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.

[18] Hongguang Zhang, Li Zhang, Xiaojuan Qi, Hongdong Li, Philip HS Torr, and Piotr Koniusz. Few-shot action recognition with permutation-invariant attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020.

[19] Linchao Zhu and Yi Yang. Compound memory networks for few-shot video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 751–766, 2018.