NSNet: Non-saliency Suppression Sampler for Efficient Video Recognition

Boyang Xia[1,2*], Wenhao Wu[3,4*✉], Haoran Wang[4], Rui Su[5], Dongliang He[4], Haosen Yang[6], Xiaoran Fan[2], and Wanli Ouyang[5,3]

[1] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, China
[2] University of Chinese Academy of Sciences, Beijing, China
[3] SenseTime Computer Vision Group, The University of Sydney, Sydney, Australia
[4] Baidu Inc., Beijing, China
[5] Shanghai AI Laboratory, Shanghai China
[6] Harbin Institute of Technology, Harbin, China

Abstract. It is challenging for artificial intelligence systems to achieve accurate video recognition under the scenario of low computation costs. Adaptive inference based efficient video recognition methods typically preview videos and focus on salient parts to reduce computation costs. Most existing works focus on complex networks learning with video classification based objectives. Taking all frames as positive samples, few of them pay attention to the discrimination between positive samples (salient frames) and negative samples (non-salient frames) in supervisions. To fill this gap, in this paper, we propose a novel **Non-saliency Suppression Network (NSNet)**, which effectively suppresses the responses of non-salient frames. Specifically, on the frame level, effective pseudo labels that can distinguish between salient and non-salient frames are generated to guide the frame saliency learning. On the video level, a temporal attention module is learned under dual video-level supervisions on both the salient and the non-salient representations. Saliency measurements from both two levels are combined for exploitation of multi-granularity complementary information. Extensive experiments conducted on four well-known benchmarks verify our NSNet not only achieves the state-of-the-art accuracy-efficiency trade-off but also present a significantly faster (2.4∼4.3×) practical inference speed than state-of-the-art methods. Our project page is at `https://lawrencexia2008.github.io/projects/nsnet`.

Keywords: Video Recognition, Adaptive Inference, Temporal Sampling

1 Introduction

The prevalence of digital devices exponentially increases the data amount of video content. Meanwhile, the proliferation of videos poses great challenges for

*: Co-first authorship. This work was done when Boyang was an intern at Baidu.
✉: Corresponding author.

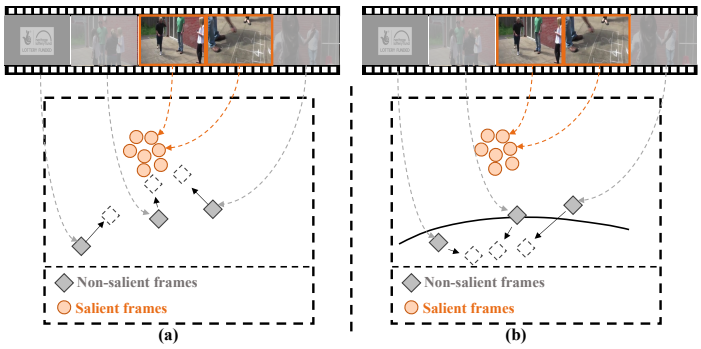

Fig. 1. A conceptual comparison between our proposed non-saliency suppression based sampler and existing approaches. (a) Feature space learned by sampler networks optimized by vanilla classification objectives with video labels. (b) Feature space learned by proposed NSNet. A video of *Hopscotch* is used for illustration. The arrows indicate the moving directions of frame features during training process. In (a), the features of non-salient frames, the 1^{st}, 2^{nd}, 5^{th} frames are forced to cluster near the centroid of the category. In contrast, our approach introduce a *non-salient* category and those low saliency frames are labeled as *non-salient* and function as negative samples against the video category during training. In this way, features of the 1^{st}, 2^{nd}, 5^{th} frames in (b) are pushed away to form another cluster of *non-salient* category.

existing video analysis systems, and consequently draws more and more attention from the research community. Thanks to the renaissance of deep neural networks [12,16,45], a surge of progress has been made to promote the development of advanced video understanding techniques [35,36,59,49,4,43,25,15]. Although achieving promising performance on some benchmarks [18,2] with supervised learning or unsupervised learning [14,6], many of them apply computationally heavy networks, which hinders their deployment to the practical applications, such as autonomous driving and personalized recommendations. Accordingly, building an efficient video understanding system is a crucial step towards widespread deployment in the real world.

To achieve efficient video recognition, a rich line of studies have been proposed, which roughly fall into two paradigms: i) **lightweight architecture based methods** and ii) **adaptive inference based methods**. The first category of approaches [23,50] devote to reducing the computational cost via designing lightweight networks. By contrast, another series of works propose to achieve efficient recognition by leveraging adaptive inference strategy to flexibly allocate resources according to the the saliency of frames. Specifically, the adaptive inference mechanism has been applied on multiple dimensions of video, including temporal sampling [20], spatial sampling [44,27], *etc.* Compared to the former, the adaptive inference based methods is easier to be incorporated into the existing advanced recognition backbones. For example, in the pipeline of a adaptive temporal sampling method, a sampler network is first trained based on

lightweight feature to sample key frames, and then an off-the-shell computation-consuming recognizer is evoked on sampled frames for final recognition.

Semantic saliency of each frame is the fundamental basis for the adaptive inference based methods [20,30]. Nonetheless, it is difficult to obtain explicit supervision of frame-level saliency in the general setting of video recognition. Therefore, existing methods are mainly based on either reinforce learning (RL) or attention mechanism [56,51,27,10], where the agent (*resp.*, attention module) is optimized to take actions (*resp.*, attend) to those salient frames by classification objective based rewards (*resp.*, loss). In this way, the sampler is trained to determine the salient frames by using all frames as the positive samples of the corresponding video category, as shown in Figure 1. Due to the lack of negative samples in training, it is hard for the sampler to accurately determine the non-salient frames from the salient ones within one video, which may easily overestimate the saliency of the non-salient frames. As a result, it is reasonable to introduce the negative samples for the frame-level video category classification objective based learning, which helps suppress the response of the non-salient frames to the video category during the sampling process.

To this end, we propose a novel **Non-saliency Suppression (NS) mechanism**, to provide negative-sample-aware supervision for saliency measurement, which can effectivly suppress the response of non-salient frames in the adaptive inference based framework. Specifically, the key principle is that the salient frames should belong to the corresponding video category while the non-salient frames should fall into a special category, which is distinguishable from all video categories. We term this special category as *non-salient* category, as shown in Figure 1. As the video categories are the only annotation we can use during training, in order to guarantee high-quality negative samples (*i.e.,* the non-salient frames) in the frame-level saliency learning, we propose a Frame Scrutinize Module (FSM) to generate frame-level pseudo labels for supervision. By doing so, the salient frames can be effectively distinguished from the non-salient frames. In addition to the frame-level supervision, we then propose a temporal attention module named Video Glimpse Module (VGM) to compensate for high-level information of video events by using video-level supervision. In order to introduce NS mechanism on video level, we first formulate a video representation as a linear combination of two components: the representation of the salient parts and the representation of the non-salient parts of a video. Following the aforementioned principle, we then assign the label of that salient representation as current video label, and the label of non-salient representation as the special *non-salient* category.

Overall, our contributions are three-folds:

- We introduce the **Non-saliency Suppression mechanism** for suppressing the responses of non-salient frames, which considerably improve the discrimination power of the temporally sampled video representations without increment of computation overhead.
- We propose an discriminative and flexible multi-granularity frame sampling framework **N**on-saliency **S**uppression **Net**work (**NSNet**), which leverages

supervisions from both video level and frame level to measure frame saliency. We design two specific schemes for realizing Non-saliency Suppression mechanism on the two granularities.
– Extensive experiments are conducted with multiple backbone architectures on four well-known benchmarks, *i.e.*, ActivityNet, FCVID, Mini-Kinetics and UCF101, which show that our NSNet achieves superior performance over existing state-of-the-art methods with limited computational costs.

2 Related Work

Video Recognition. Video Recognition has made significant progress in past decade for successful application of neural networks including 2D CNNs [42], 3D CNNs [3] and Transformers [1,53]. Although decent results are achieved by powerful spatiotemporal networks, it is still challenging for applying video recognition in resource-constraint scenarios for its superfluous computational complexity. TSM [23], TEA [22], MVFNet [50], *etc.*, try to realize temporal modelling with pure 2D CNNs to improve efficiency by shifting operations, motion based channel selection and multi views fusion. While P3D [31], S3D [58], R(2+1)D [40], SlowFast [7], Ada3D [21], DSANet [54] are proposed to improve the efficiency by decomposing 3D convolution or designing hybrid 2D-3D frameworks. Different from these approaches, we seek to achieve efficient video recognition by adaptively sampling salient frames and recognize selectively on a per-sample basis.

Adaptive Inference. The core idea of adaptive inference is to dynamically allocate computational resources (network layers, parameters, *etc.*) conditioned on the input to improve the trade-off between performance and cost [11,52]. For video analysis, adaptive inference are realized in several perspectives including temporal sampling, resolution, sub-networks and modality. FastForward [5], FrameGlimpse [60], AdaFrame [56], MARL [51] and OCSampler [24] model temporal sampling as a decision-making process, which is optimized by policy gradients or differentiable alternatives. ListenToLook [8], SMART [10], TSQNet [57] design temporal frame samplers based on attention mechanism. Besides temporal sampling, AdaFocus series [44,46] samples salient patches for each frame to reduce spatial redundancy. AR-Net [27], LiteEval [55], AdaMML [30] strategically allocate higher resolution, more powerful sub-networks, or more expensive modality to more informative frames, respectively.

The most closely related work to ours is an adaptive temporal sampling method, SCSampler [20], which proposes a frame-level classification task with video label and measure saliency based on classification confidence. However, there exist substantial differences between SCSampler and the proposed NSNet. SCSampler assigns each frame with the video label, while we argue that only the salient ones belong to video category, other ones should be labeled as a special category distinguishable from all semantic categories. Besides, we also consider to measure frame saliency with video level supervisions to enable context-aware saliency measurements, which is overlooked by SCSampler. Compared with SC-

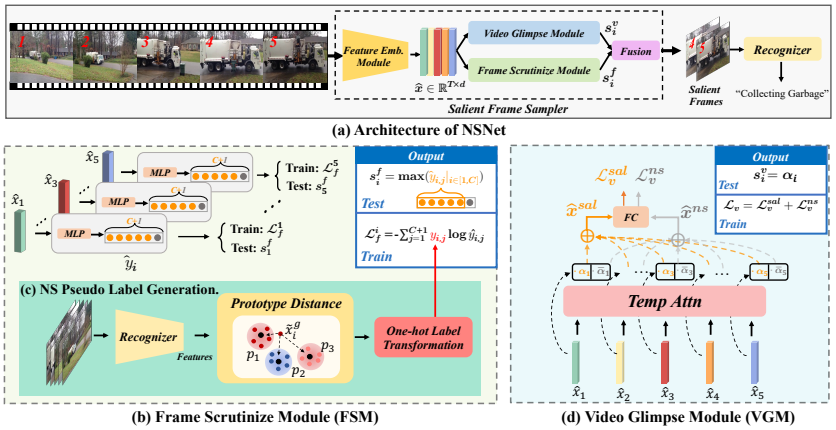

Fig. 2. An overview of architecture of NSNet. (a) shows the whole architecture. (b) shows the Frame Scrutinize Module (FSM), which estimates the saliency of each frame by the prediction confidence in frame-level classification. (c) shows the proposed Non-saliency Suppression (NS) frame-level pseudo label generation strategy based on the distance between each frame and video category prototypes. (d) shows the Video Glimpse module (VGM), which measures the saliency of each frame using temporal attentions in video-level classification.

Sampler, our NSNet achieves significantly superior performance with much less computational overhead.

3 Approach

3.1 Problem Definition

Let $V = \{x_i\}_{i=1}^{T}$ denote a video of T frames. Firstly, our NSNet is designed to select K most salient frames from all T frames. Then, the K salient frames are fed into a recognizer model, the predictions of which are aggregated to yield the final video-level prediction. The workflow of our system is illustrated in Figure 2a. Note that we use an off-the-shell model as our recognizer, and therefore, the problem can be formulated as how to effectively sample the most salient frames from the input frames. In the following sections, we introduce the process of salient frames selection for our NSNet.

3.2 The Overview of NSNet

In this section, we elaborate on the proposed **N**on-saliency **S**uppression **Net**work (NSNet), which mainly consists of three components: a **Feature Embedding module (FEM)**, a **Frame Scrutinize module (FSM)**, and a **Video Glimpse module (VGM)**. The feature embedding module generates feature embedding from the input video frames. The FSM (see Figure 2b) measures saliency of

each frame by predicting the saliency confidence scores in frame-level classification, and the VGM (see Figure 2d) models saliency of frames from the temporal attention weights used to aggregate attention-based feature for video-level classification. To alleviate the lack of negative samples, we further apply NS mechanism in the FSM and VGM in different ways. For each of these two modules, a *non-salient* category is attached onto the frame-level and the video-level supervisions, respectively, resulting in a total of $C + 1$ categories for each supervision, where C is the total number of the original video categories. In the FSM, a Non-saliency Suppression frame-level pseudo label generation strategy is proposed to separate the negative samples from the truly salient frames for frame-level saliency learning. In the VGM, a Non-saliency Suppression loss is proposed to impose an extra constraint of non-salient representations of videos besides original classification objectives.

3.3 Feature Embedding Module (FEM)

Here we encode the input video frames $\{x_i\}_{i=1}^{T}$ into the robust feature sequence $\{\hat{x}_i\}_{i=1}^{T}$, which is then used by our FSM and VGM. Specifically, we first use the off-the-shelf lightweight feature extractor, *e.g.,* MobileNet, EfficientNet, *etc*, to take the video frames as input and extract features for all frames. In order to allow message passing among features for all frames, we then apply a transformer encoder [41] on top of these features and output the feature sequence $\{\hat{x}_i\}_{i=1}^{T}$.

3.4 Frame Scrutinize Module (FSM)

In our FSM, we first generate frame-level Non-saliency Supression (NS) pseudo labels, and then use them as supervision to train our FSM to perform frame-level saliency classification and produce saliency scores. Details of our FSM are provided as follows.

Non-saliency Suppression Pseudo Label Generation. Here we denote the label of a video of the c-th category as a C-dimension one-hot vector $y_v \in \mathbb{R}^C$, where $y_{v,c} = 1$ and $y_{v,m} = 0|_{m\in[1,C],m\neq c}$. To distinguish between the salient frames and non-salient frames in frame labels, we then introduce a **guiding saliency score** g_i, which is obtained from the recognizer, *i.e.*, the one we used for final recognition as described in Section 3.1 (Figure 2a). Although the classification response produced by the pre-trained model (*i.e.,* recognizer) is widely used for pseudo labeling in weakly-supervised learning [39,29,47], we propose a prototype-based strategy for more robust frame level pseudo label generation. According to [33,61], a sample could be more representative when it is closer to the centroid in feature space. As a result, we use distances of the feature for the i-th frame from the prototype features of all categories to obtain g_i. Specifically, we first use the recognizer to extract features and confidence scores on ground-truth category for all frames for each video in training set. Then the prototype feature of each category is then calculated by averaging all video features in that category. Here each video feature is obtained by applying average pooling on the features of the top-K frames based on the predicted confidence scores (see our

Appendix for more details). The guiding saliency score g_i for the i-th frame is as follows:

$$g_i = \frac{e^{\phi(\tilde{x}_i^g, p_c)}}{\sum_{j=1}^{C} e^{\phi(\tilde{x}_i^g, p_j)}}, \tag{1}$$

where ϕ is a distance function measuring the similarity of two feature vectors, *e.g.,* Euclidean Distance, $\tilde{x}_i^g$ is the feature for the i-th frame extracted by the recognizer, p_j and p_c are the prototype features of the j-th category and the ground truth category, respectively. Finally, we use g_i to generate the NS pseudo label $y_i^{ns} = [\ g_i y_{v,1},\ g_i y_{v,2},\ \ldots\ g_i y_{v,C}, 1 - g_i] \in \mathbb{R}^{C+1}$.

Frame-level Saliency Classification. After generating the NS pseudo labels, we then use them to train our FSM to perform frame-level saliency classification over the feature sequence $\{\hat{x}_i\}_{i=1}^{T}$. Mathematically, the frame-level classification objective is defined as follows:

$$\mathcal{L}_f = -\sum_{i=1}^{T}\sum_{j=1}^{C+1} y_{i,j}^{ns} \log(\hat{y}_{i,j}^{ns}), \tag{2}$$

where $y_{i,j}^{ns}$ is the element of the j-th category in y_i^{ns}, and $\hat{y}_{i,j}^{ns}$ is the classification prediction. It is noteworthy that y_i^{ns} is a soft one-hot target, the cross entropy loss of which is similar to label smooth [38]. During inference, the frames with very high response to any one of C categories are identified as salient frames. To this end, the maximum confidence across C semantic categories (except the $C+1$-th category) of classification score after softmax normalization is used for saliency measurement. We then apply additional softmax normalization along the time axis to obtain final saliency score s_i^f.

3.5 Video Glimpse Module (VGM)

In our VGM, we first generate attention weights $\alpha_i = \text{TempAttn}(\hat{x}_i)$ for the features of all observed frames, where $\text{TempAttn}(\cdot)$ is implemented by a fully-connected layer followed by a L1 normalization layer, which is used to rescale attention weights to $[0, 1]$ range. The features of all observed frames are then aggregated with the attention weights to generate the video salient representation $\hat{x}_v^{sal} = \sum_{i=1}^{T} \alpha_i \hat{x}_i$. To perform video-level classification, the salient representation of the video $\hat{x}_v^{sal}$ is fed to a fully-connected layer to compute the cross-entropy loss with video label. In order to guide our VGM to separate negative samples (non-salient frames) from positive samples (salient frames) of current video category, we propose a Non-saliency Suppression (NS) loss to impose a constraint other than the regular classification objective. During inference, the attention weights are used as saliency scores $\{s_i^v\}_{i=1}^{T}$. The details of the NS loss are described next.

Non-saliency Suppression Loss. It is obvious that all videos contains both salient and non-salient frames for a specific video category. Therefore, it is natural that a holistic video representation $\hat{x}_v$ can be formulated as a linear combination of the salient representation $\hat{x}_v^{sal}$ and the non-salient representation

$\hat{x}_v^{ns}$ [29], *i.e.,* $\hat{x}_v = \hat{x}_v^{sal} + \gamma\hat{x}_v^{ns}$. The non-salient representation $\hat{x}_v^{ns}$ can be obtained as follows. We first compute the complementary part of attention weights $\overline{\alpha}_i = \frac{1}{T}(1-\alpha_i)$, and then aggregate the feature sequence with $\overline{\alpha}_i$ and produce the non-salient representation $\hat{x}_v^{ns} = \sum_{i=1}^{T} \overline{\alpha}_i \hat{x}_i$. In this way, a video can be regarded as a positive sample of both its ground truth category and *non-salient* category to different proportions, at the same time. Both $\hat{x}_v^{sal}$ and $\hat{x}_v^{ns}$ will be fed into the classification fully-connected layer to get different predictions $\hat{y}_v^{sal}$ and $\hat{y}_v^{ns}$. Then we defines labels for both $\hat{y}_v^{sal}$ and $\hat{y}_v^{ns}$: $y_v^{ns} = [0, 0, ..., 0, 1] \in \mathbb{R}^{C+1}$, $y_v^{sal} = [y_{v,1}, y_{v,2}, ...y_{v,C}, 0] \in \mathbb{R}^{C+1}$, where y_v is the original video label. The cross-entropy loss between $\hat{y}_v^{sal}$ and y_v^{sal} is the original classification loss $\mathcal{L}_{cls}$, and the one between $\hat{y}_v^{ns}$ and y_v^{ns} is the NS loss $\mathcal{L}_{ns}$. Consequently, the objective function of this module is defined as follows, where γ is the weight of $\mathcal{L}_{ns}$.

$$\mathcal{L}_v = \mathcal{L}_{cls} + \gamma\mathcal{L}_{ns}, \tag{3}$$

3.6 Learning Objectives

The overall objective function of our NSNet is formulated as follows:

$$\mathcal{L} = \mathcal{L}_v + \mathcal{L}_f, \tag{4}$$

where $\mathcal{L}_v$ and $\mathcal{L}_f$ denote the loss function of the VGM and the FSM, respectively. This objective not only drives model to conduct discriminative saliency measuring according to video semantics and frame discrepancy, but also facilitates information exchange between the video context and parts in shared feature encoding.

4 Experiments

4.1 Datasets and Evaluation Metrics

We evaluate our method on four large-scale video recognition benchmarks, *i.e.,* ActivityNet, FCVID, Mini-Kinetics and UCF101. ActivityNet [2] contains 19994 videos of 200 categories of most popular actions in daily life. FCVID [17] contains 91,223 videos collected from YouTube and divided into 239 classes covering most common events, objects, and scenes in our daily lives. Mini-Kinetics [27] is a subset of Kinetics [18] presented by [27], including 200 categories of videos of Kinetics, with 121k videos for training and 10k videos for validation. UCF101 [34] has 101 classes of actions and 13K videos with short duration (7.2sec). Mean Average Precision (mAP) is used as the main evaluation metric for ActivityNet and FCVID, while Top-1 accuracy is used for Mini-Kinetics and UCf-101 following previous works. We also report the computational cost (in FLOPs) to evaluate the efficiency of the proposed method. FLOPs of our method are composed of following parts: $P_{total} = P_{rec} \times K + P_{fem} + P_{vgm} + P_{fsm}$, where

$P_{rec}, P_{fem}, P_{vgm}, P_{fsm}$ represent the FLOPs of the recognizer, FEM, VGM and FSM, respectively. An example of FLOPs computation of our model with the setting in Table 2 is: 4.109×5 (ResNet-50 with 5 frames) + (0.320×16+0.315) (MobileNetv2 with 16 frames+transformer) + 0.004 + 0.002 = 25.99(G).

4.2 Implementation Details

Training. Following previous works [10,56], we mainly use MobileNetv2 [32] as the lightweight feature extractor in our FEM. Different high-capacity networks trained on target datasets are used as recognizers at the same time: ResNet family [12] and Swin-Transformer family [26], *etc.* For the transformer encoder in our FEM, 2 encoder layers with 8 heads and learnable positional embedding are used. The distance function ϕ in our FSM used for guiding saliency score is Euclidean Distance. The non-saliency suppression loss weight γ is set to 0.2. See Appendix for more details of training.
Inference. We fuse the results of FSM and VGM to obtain the final saliency measurements. Score *sum, max, mul* and index *union, intersect, join* are considered for fusion. See Appendix for details.

4.3 Main Results

Comparison with Simple Baselines. As shown in Table 1, we compare our approach with multiple hand-crafted sampling methods on ActivityNet and UCF101 with ResNet-101 and ResNet-50 as recognizers (without TSN training strategy [42]), respectively. The simple baselines include Uniform, Random, Dense, and Top-K sampling. For uniform and random, we uniformly and randomly sample 10 frames from all frames, respectively, while for Top-K, we sample top 10 frames with highest predicted confidence scores (*i.e.*, the maximum confidence among all categories), from all frames. For our method, we first uniformly sample an observation number (100 for ActivityNet and 50 for UCF-101) of frames as the observation frames from the input videos, and then use our method to sample 5 frames from the observation frames. Dense sampling makes use of all frames for recognition. We observe that our NSNet outperforms all simple baselines by a large margin on both two datasets. In ActivityNet, our method relatively outperforms the competitive but heavy Top-K baseline by 2.4% in terms of mAP with 11.7× less GFLOPs, which verifies the effectiveness of our sampler. In UCF-101, the videos are much shorter than those in ActivityNet (7sec *v.s.* 119sec on average), which constructs a much more difficult setting for sampler. However, the top-1 accuracy of our NSNet still relatively exceeds that of the most competitive Dense baseline by 2.1% with much less GFLOPs, which demonstrates NSNet can improve the video classification performance on trimmed videos.
Comparison with SOTA on ActivityNet. We make a comprehensive comparison with recent state-of-the-art efficient video recognition methods on the ActivityNet dataset in Table 2-3, and Figure 3. As shown in Table 2, we compare our NSNet with other state-of-the-art efficient approaches using ResNet-50

Table 1. Comparison with several hand-crafted sampling strategies. ResNet-101 and ResNet-50 are adopted as the recognizers for ActivityNet and UCF-101, respectively.

	ActivityNet		UCF101	
	mAP(%)	FLOPs	Top-1(%)	FLOPs
Uniform	68.6	195.8G	75.9	61.7G
Random	68.1	195.8G	75.7	61.7G
Dense	69.0	930.8G	76.1	753.4G
Top-K	72.5	930.8G	74.5	753.4G
Ours	**74.9**	**73.2G**	**77.6**	**37.6G**

Table 2. Comparisons with SOTA efficient video recognition methods with ResNet50 as the main recognizer on the ActivityNet dataset. The backbones used for sampler and recognizer are reported. MBv2 denotes MobileNetv2.

Method	Backbone	mAP(%)	FLOPs
SCSampler[20]	MBv2+Res50	72.9	42.0G
AR-Net [27]	MBv2+ResNets	73.8	33.5G
AdaMML [30]	MBv2+Res50	73.9	94.0G
VideoIQ [37]	MBv2+Res50	74.8	28.1G
AdaFocus [44]	MBv2+Res50	75.0	26.6G
Dynamic-STE [19]	Res18+Res50	75.9	30.5G
FrameExit [9]	ResNet-50	76.1	26.1G
Ours	MBv2+Res50	**76.8**	**26.0G**

as the recognizer. NSNet consistently outperforms all existing methods including sampler-based and sampler-free approaches. When compared with AR-Net [27], an adaptive resolution method, the mAP of our method improve by 3% with much less computational cost (**26.0G** *v.s.* 33.5G). Our NSNet also outperforms AdaMML [30], an adaptive modality approach, by 2.9% in terms of mAP while having 3.6× less FLOPs. In addition, sampler-free approaches often have relatively low FLOPs, because they do not need sampling process. However, although our NSNet has extra computational on the sampling process, it can achieve higher accuracy than FrameExit [9](**76.8%** *v.s.* 76.1%), a competing sampler-free framework, with comparable FLOPs, which demonstrates our sampler greatly improve the discrimination power of the video representation. In Table 3, we show the results of the SOTAs using ResNet-152 and more advanced backbones on the ActivityNet dataset. We can see that our NSNet outperforms the sampler-free methods P3D [31], RRA [63] by at least 1.7% in terms of mAP. When compared with competing sampler-based methods (MARL [51], Listen-ToLook [8], SMART [10]), our method still shows superiority over them. Besides, when using a more advanced transformer-based network Swin-Transformer [26] as the recognizer in our NSNet, the mAP can be further promoted to 94.3%, which is the highest performance on ActivityNet to our best knowledge.

Figure 3 illustrates the GFLOPs-mAP curve on the ActivityNet dataset. On this curve, the observation number T is set to 50, while the number of sampled frames K is tuned up as FLOPs budget increases. Following previous works

Table 3. Comparisons with SOTA video recognition methods with ResNet152 and more advanced networks as the recognizers on the ActivityNet dataset.

Method	Recognizer	Pretrain	Top-1(%)	mAP(%)
ResNet-152 w/ ImageNet				
P3D [31]	ResNet-152	ImageNet	75.1	78.9
RRA [63]	ResNet-152	ImageNet	78.8	83.4
MARL [51]	ResNet-152	ImageNet	79.8	83.8
ListenToLook [8]	ResNet-152	ImageNet	80.3	84.2
Ours	ResNet-152	ImageNet	**80.7**	**85.1**
ResNet-152 w/ Kinetics				
SMART [10]	ResNet-152	Kinetics	-	84.4
Ours	ResNet-152	Kinetics	**84.5**	**88.7**
More Advanced Networks w/ Kinetics				
DSN [62]	R(2+1)D-34	Kinetics	82.6	87.8
Ada3D [21]	SlowOnly-50	Kinetics	-	84.0
ListenToLook [8]	R(2+1)D-152	Kinetics	-	89.9
MARL [51]	SEResNeXt-152	Kinetics	85.7	90.1
Ours	Swin-B	Kinetics	86.7	91.6
Ours	Swin-L	Kinetics	**90.2**	**94.3**

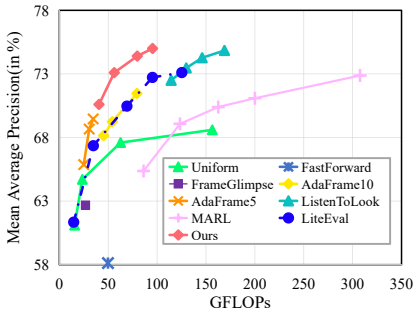

Fig. 3. Comparison with state-of-the-art sampling methods on ActivityNet dataset. Our proposed NSNet achieves better mAP with much fewer GLOPs (per video) than other methods. It is worth noting that we compare these methods with the same recognizer ResNet-101, under different computation budgets. The results are quoted from the published works [8,27].

[56,55,8,51], we use ResNet-101 without TSN-style training as the recognizer. Our NSNet presents significant accuracy improvement with much lower GFLOPs than other methods.

Comparison with SOTA on FCVID and Mini-Kinetics. We further evaluate the performance of our method on two large-scale video recognition benchmarks, *i.e.,* FCVID and Mini-Kinetics, with ResNet-50 as the recognizer in Table 4. We have a similar observation that our NSNet can achieve superior mAP with the much lower computational cost, which demonstrate the efficacy of non-saliency suppression (NS) mechanism in both untrimmed video and trimmed video scenarios.

Practical Efficiency. We present the comparison results on inference speed between our NSNet and two SotA methods, FrameExit [9] and AdaFocus [44], in Table 5. Latency and throughput with batch size of 1 and 32 are reported[1]. It can be observed that our method achieves significantly superior latency (42.0 ms)

[1] The latency and throughput results of two SotA methods are obtained by running their official code [9,44] on the same hardware (a NVIDIA 3090 GPU with a Intel Xeon E5-2650 v3 @ 2.30GHz CPU) as ours.

Table 4. Comparison with previous methods on FCVID and Mini-Kinetics. Our NSNet consistently outperforms state-of-the-art in terms of accuracy and efficiency using ResNet-50 as the recognizer.

Methods	FCVID		Mini-Kinetics	
	mAP(%)	FLOPs	Top-1(%)	FLOPs
LiteEval [55]	80.0	94.3G	61.0	99.0G
AdaFrame [56]	80.2	75.1G	-	-
SCSampler [20]	81.0	42.0G	70.8	42.0G
AR-Net [27]	81.3	35.1G	71.7	32.0G
AdaFuse [28]	81.6	45.0G	72.3	23.0G
SMART [10]	82.1	-	-	-
VideoIQ [37]	82.7	27.0G	72.3	20.4G
Dynamic-STE [19]	-	-	72.7	18.3G
FrameExit [9]	-	-	72.8	19.7G
AdaFocus [44]	83.4	26.6G	72.9	38.6G
Ours	**83.9**	**26.0G**	**73.6**	**18.1G**

Table 5. Comparison of practical efficiency between SotA methods.

Method	mAP(%)	GFLOPs	Latency (bs=1)	Throughput (bs=32)
AdaFocus [44]	75.0	26.6	181.8ms	73.8 vid/s
FrameExit [9]	76.1	26.1	102.0ms	-
Ours	**76.8**	**26.1**	**42.0ms**	**132.5 vid/s**

and than two methods (2.4× than FrameExit [9] and 4.3× than AdaFocus [44]), which demonstrate the superiority of our parallel temporal sampling framework over existing methods on practical efficiency.

4.4 Ablation Studies

To comprehensively evaluate our NSNet, we provide extensive ablation studies on ActivityNet in Table 6. Accordingly, the effectiveness of each component in our framework is analyzed as follows. We use ResNet-101 without TSN style training as the recognizer, as the same as in Table 1 and Figure 3 in Section 4.3.

Effectiveness of non-saliency suppression. We explore the effectiveness of *non-saliency suppression* (NS) mechanism for two modules. For VGM, "baseline" denotes the variant without $\mathcal{L}_{ns}$. For FSM, "baseline" denotes the variant replacing our NS frame label with video label. As shown in Table 6a, for the VGM, simply adding a *non-salient* class (from C classes to $C+1$ classes) without according supervisions can not elevate performance. In contrast, by applying NS mechanism, the mAP significantly improves by 0.4%. In the FSM, we observe that "baseline" present very low performance, similar to UNIFORM baseline(68.4 *v.s.* 68.6), which is because it imposes many label noises when assigning the label of video to irrelevant or low-quality frames. We also observe that simply adding the *non-salient* class cannot address the issue (68.6 *v.s.* 68.4). NS mechanism elevates the performance significantly with effective and reasonable supervisions (74.7 *v.s.* 68.4). More ablations in supervision signals of FSM are in Table 6d.

Table 6. Ablation studies on ActivityNet with mAP (%) as the evaluation metric. Unless otherwise specified, MobileNetv2 and ResNet-101 are used as the backbone for observation network and recognizer respectively.

(a) Evaluation of the effectiveness of NS mechanism.

Method	VGM	FSM
baseline	73.4	68.4
+ *non-salient* class	73.4	68.6
+ NS	**73.8**	**74.7**

(b) Performance of different number of sampled frames.

#F	VGM	FSM	NSNet
5	72.6	73.9	**74.9**
10	73.8	74.7	**75.5**

(c) Ablation of transformer encoder in feature embedding module.

Network	mAP(%)
1D Conv	73.6
LSTM	74.0
MLP	74.3
Transformer	**75.5**

(d) Results of FS module with different learning objectives.

FS Objective	mAP(%)
Baseline	68.4
Regression	72.0
Ranking	72.2
Baseline+	73.7
Ours	**74.7**

Ablations of transformer encoder in feature embedding module. Table 6c presents the results of different choice in FEM to passing message among the features from the input frames, including Long short-term memory networks (LSTM) [13], 1-D convolutional networks (1D Conv), multi-layer perceptron (MLP) [48], Transformer Encoder (Transformer) [41]. Among all these choices, the transformer encoder achieves the highest performance.

Different learning objectives of the FSM. In Table 6d we compare various objectives of FSM mentioned in Section 1 and Section 2, including frame classification with video labels ("baseline", as the same one as in Table 6a), ranking [20], and regression [10] with guiding saliency scores. With guiding saliency scores as supervisions, "regression" and "ranking" model saliency sampling as saliency score regression and ranking tasks respectively. They can achieve higher performance than "baseline" (72.0% & 72.2%), whereas they overlook the exploitation of class-specific information, which limits their performance. We modify "baseline" by transforming hard one-hot video label to soft one-hot label using the guiding saliency score, which is denoted as "baseline+". This modification improves the result significantly (**73.7%** *v.s.* 68.4%) by taking into account both the discrimination between salient frames and non-salient frames and the use of category-specific information. With the same setting of guiding saliency score, our NS mechanism based objective outperforms 'baseline+' in a large margin (**74.7%** *v.s.* 73.7%), which verifies the proposed supervisions offer more robust saliency for supplying negative samples.

Different numbers of sampled frames. As shown in Table 6b, we report the performance of different numbers of sampled frames for NSNet. We can see that the fusion of two modules always improves the performance by 0.8% and 1.0% when sampling 10 and 5 frames, respectively. It demonstrates the effectiveness of our fusion strategies, especially when fewer frames are sampled. Besides, when

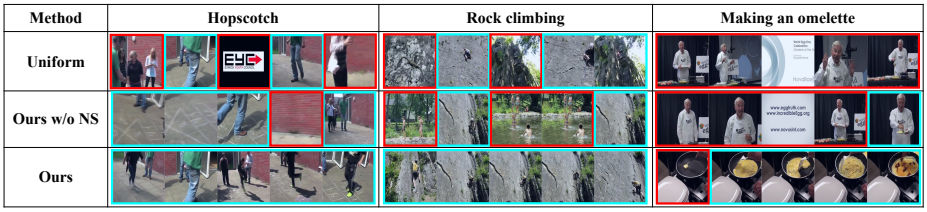

Fig. 4. Visualization of selected frames with different variants of our approach. 1st row: Uniform, 2nd row: Our approach without NS mechanism (Ours w/o NS), 3rd row: Our approach (Ours). Intuitively salient frames are are outlined in aqua while non-salient ones are outlined in red. Please zoom in for best view.

fewer frames are sampled, the performance of FSM degrades more slowly than that of VGM (0.8% *v.s.* 1.2%), which is because FSM can distinguish between salient frames and non-salient frames in finer granularity with the help of frame-level supervisions.

4.5 Qualitative Analysis

Figure 4 shows frames sampled by different methods. Our NSNet can sample more discriminative salient frames than uniform baseline and the variant without non-saliency suppression. For example, in the 4th column, the 3rd row of this column shows that "ours w/o NS" is mainly attracted by frames with scenes of a cook, which is not discriminative for frequently appearing in other cooking events. In contrast, 4th row shows NSNet can sample more indicative frames.

5 Conclusions

In this paper, we present the Non-saliency Suppression Network (NSNet) to measure the saliency of frames by leveraging both video-level and frame-level supervisions. In Frame Scrutinize module, we propose a pseudo label generation strategy to enable negative sample aware frame-level saliency learning. Meanwhile, in Video Glimpse module, an attention module constrained by dual classification objectives is presented to compensate high-level information. Experiments show that our NSNet outperforms the state-of-the-arts on accuracy-efficiency trade-off. In the future, we plan to explore non-saliency suppression on both spatial and temporal dimensions to save more redundancy.

Acknowledgment. Wanli Ouyang was supported by the Australian Research Council Grant DP200103223, Australian Medical Research Future Fund MRFAI000085, CRC-P Smart Material Recovery Facility (SMRF) – Curby Soft Plastics, and CRC-P ARIA - Bionic Visual-Spatial Prosthesis for the Blind.

References

1. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? arXiv preprint arXiv:2102.05095 (2021)
2. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the ieee conference on computer vision and pattern recognition. pp. 961–970 (2015)
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
4. Chen, X., Han, Y., Wang, X., Sun, Y., Yang, Y.: Action keypoint network for efficient video recognition. arXiv preprint arXiv:2201.06304 (2022)
5. Fan, H., Xu, Z., Zhu, L., Yan, C., Ge, J., Yang, Y.: Watching a small portion could be as good as watching all: Towards efficient video classification. In: IJCAI International Joint Conference on Artificial Intelligence (2018)
6. Fang, B., Wu, W., Liu, C., Zhou, Y., He, D., Wang, W.: Mamico: Macro-to-micro semantic correspondence for self-supervised video representation learning. In Proc. ACMMM (2022)
7. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)
8. Gao, R., Oh, T.H., Grauman, K., Torresani, L.: Listen to look: Action recognition by previewing audio. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10457–10467 (2020)
9. Ghodrati, A., Bejnordi, B.E., Habibian, A.: Frameexit: Conditional early exiting for efficient video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15608–15618 (2021)
10. Gowda, S.N., Rohrbach, M., Sevilla-Lara, L.: SMART frame selection for action recognition **35**(2), 1451–1459 (2021), `https://ojs.aaai.org/index.php/AAAI/article/view/16235`
11. Han, Y., Huang, G., Song, S., Yang, L., Wang, H., Wang, Y.: Dynamic neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
14. Huang, D., Wu, W., Hu, W., Liu, X., He, D., Wu, Z., Wu, X., Tan, M., Ding, E.: Ascnet: Self-supervised video representation learning with appearance-speed consistency. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8096–8105 (2021)
15. Huang, W., Fan, L., Harandi, M., Ma, L., Liu, H., Liu, W., Gan, C.: Toward efficient action recognition: Principal backpropagation for training two-stream networks. IEEE Transactions on Image Processing **28**(4), 1773–1782 (2018)
16. Ji, Z., Chen, K., Wang, H.: Step-wise hierarchical alignment network for image-text matching. In: Zhou, Z. (ed.) Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021. pp. 765–771. ijcai.org (2021). https://doi.org/10.24963/ijcai.2021/106, `https://doi.org/10.24963/ijcai.2021/106`

17. Jiang, Y.G., Wu, Z., Wang, J., Xue, X., Chang, S.F.: Exploiting feature and class relationships in video categorization with regularized deep neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(2), 352–364 (2018). https://doi.org/10.1109/TPAMI.2017.2670560
18. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017)
19. Kim, H., Jain, M., Lee, J.T., Yun, S., Porikli, F.: Efficient action recognition via dynamic knowledge propagation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13719–13728 (2021)
20. Korbar, B., Tran, D., Torresani, L.: Scsampler: Sampling salient clips from video for efficient action recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
21. Li, H., Wu, Z., Shrivastava, A., Davis, L.S.: 2d or not 2d? adaptive 3d convolution selection for efficient video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6155–6164 (2021)
22. Li, Y., Ji, B., Shi, X., Zhang, J., Kang, B., Wang, L.: Tea: Temporal excitation and aggregation for action recognition. In: CVPR. pp. 909–918 (2020)
23. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7083–7093 (2019)
24. Lin, J., Duan, H., Chen, K., Lin, D., Wang, L.: Ocsampler: Compressing videos to one clip with single-step sampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13894–13903 (2022)
25. Liu, Y., Ma, L., Zhang, Y., Liu, W., Chang, S.F.: Multi-granularity generator for temporal action proposal. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3604–3613 (2019)
26. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)
27. Meng, Y., Lin, C.C., Panda, R., Sattigeri, P., Karlinsky, L., Oliva, A., Saenko, K., Feris, R.: Ar-net: Adaptive frame resolution for efficient action recognition. In: European Conference on Computer Vision. pp. 86–104. Springer (2020)
28. Meng, Y., Panda, R., Lin, C.C., Sattigeri, P., Karlinsky, L., Saenko, K., Oliva, A., Feris, R.: Adafuse: Adaptive temporal fusion network for efficient action recognition. arXiv preprint arXiv:2102.05775 (2021)
29. Nguyen, P.X., Ramanan, D., Fowlkes, C.C.: Weakly-supervised action localization with background modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5502–5511 (2019)
30. Panda, R., Chen, C.F., Fan, Q., Sun, X., Saenko, K., Oliva, A., Feris, R.: Adamml: Adaptive multi-modal learning for efficient video recognition. arXiv preprint arXiv:2105.05165 (2021)
31. Qiu, Z., Yao, T., Mei, T.: Learning spatio-temporal representation with pseudo-3d residual networks. In: proceedings of the IEEE International Conference on Computer Vision. pp. 5533–5541 (2017)
32. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)
33. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Advances in neural information processing systems **30** (2017)

34. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
35. Su, R., Ouyang, W., Zhou, L., Xu, D.: Improving action localization by progressive cross-stream cooperation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
36. Su, R., Yu, Q., Xu, D.: Stvgbert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1533–1542 (October 2021)
37. Sun, X., Panda, R., Chen, C.F.R., Oliva, A., Feris, R., Saenko, K.: Dynamic network quantization for efficient video inference. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7375–7385 (2021)
38. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
39. Tang, P., Wang, X., Bai, X., Liu, W.: Multiple instance detection network with online instance classifier refinement. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2843–2851 (2017)
40. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: CVPR (2018)
41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
42. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: European conference on computer vision. pp. 20–36. Springer (2016)
43. Wang, X., Zhu, L., Wu, Y., Yang, Y.: Symbiotic attention for egocentric action recognition with object-centric alignment. IEEE transactions on pattern analysis and machine intelligence (2020)
44. Wang, Y., Chen, Z., Jiang, H., Song, S., Han, Y., Huang, G.: Adaptive focus for efficient video recognition. arXiv preprint arXiv:2105.03245 (2021)
45. Wang, Y., Lv, K., Huang, R., Song, S., Yang, L., Huang, G.: Glance and focus: a dynamic approach to reducing spatial redundancy in image classification. Advances in Neural Information Processing Systems **33**, 2432–2444 (2020)
46. Wang, Y., Yue, Y., Lin, Y., Jiang, H., Lai, Z., Kulikov, V., Orlov, N., Shi, H., Huang, G.: Adafocus v2: End-to-end training of spatial dynamic networks for video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20062–20072 (June 2022)
47. Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1568–1576 (2017)
48. Werbos, P.J.: Applications of advances in nonlinear sensitivity analysis. In: System modeling and optimization, pp. 762–770. Springer (1982)
49. Wu, J., Zhang, W., Li, G., Wu, W., Tan, X., Li, Y., Ding, E., Lin, L.: Weakly-supervised spatio-temporal anomaly detection in surveillance video. IJCAI (2021)
50. Wu, W., He, D., Lin, T., Li, F., Gan, C., Ding, E.: Mvfnet: Multi-view fusion network for efficient video recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 2943–2951 (2021)
51. Wu, W., He, D., Tan, X., Chen, S., Wen, S.: Multi-agent reinforcement learning based frame sampling for effective untrimmed video recognition. In: Proceedings

of the IEEE/CVF International Conference on Computer Vision. pp. 6222–6231 (2019)
52. Wu, W., He, D., Tan, X., Chen, S., Yang, Y., Wen, S.: Dynamic inference: A new approach toward efficient video action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. pp. 676–677 (2020)
53. Wu, W., Sun, Z., Ouyang, W.: Transferring textual knowledge for visual recognition. arXiv e-prints pp. arXiv–2207 (2022)
54. Wu, W., Zhao, Y., Xu, Y., Tan, X., He, D., Zou, Z., Ye, J., Li, Y., Yao, M., Dong, Z., et al.: Dsanet: Dynamic segment aggregation network for video-level representation learning. In Proc. ACMMM (2021)
55. Wu, Z., Xiong, C., Jiang, Y.G., Davis, L.S.: Liteeval: A coarse-to-fine framework for resource efficient video recognition. arXiv preprint arXiv:1912.01601 (2019)
56. Wu, Z., Xiong, C., Ma, C.Y., Socher, R., Davis, L.S.: Adaframe: Adaptive frame selection for fast video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1278–1287 (2019)
57. Xia, B., Wang, Z., Wu, W., Wang, H., Han, J.: Temporal saliency query network for efficient video recognition. ECCV (2022)
58. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: ECCV (2018)
59. Yang, H., Wu, W., Wang, L., Jin, S., Xia, B., Yao, H., Huang, H.: Temporal action proposal generation with background constraint. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 3054–3062 (2022)
60. Yeung, S., Russakovsky, O., Mori, G., Fei-Fei, L.: End-to-end learning of action detection from frame glimpses in videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2678–2687 (2016)
61. Zhang, M., Song, G., Zhou, H., Liu, Y.: Discriminability distillation in group representation learning. In: European Conference on Computer Vision. pp. 1–19. Springer (2020)
62. Zheng, Y.D., Liu, Z., Lu, T., Wang, L.: Dynamic sampling networks for efficient action recognition in videos. IEEE Transactions on Image Processing **29**, 7970–7983 (2020)
63. Zhu, C., Tan, X., Zhou, F., Liu, X., Yue, K., Ding, E., Ma, Y.: Fine-grained video categorization with redundancy reduction attention. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 136–152 (2018)