Temporal Saliency Query Network for Efficient Video Recognition

Boyang Xia[1,2*], Zhihao Wang[1,2*], Wenhao Wu[3,4✉], Haoran Wang[4], and Jungong Han[5]

[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] The University of Sydney, Sydney, Australia
[4] Baidu Inc., Beijing, China
[5] Computer Science Department, Aberystwyth University, SY23 3FL, UK

Abstract. Efficient video recognition is a hot-spot research topic with the explosive growth of multimedia data on the Internet and mobile devices. Most existing methods select the salient frames without awareness of the class-specific saliency scores, which neglect the implicit association between the saliency of frames and its belonging category. To alleviate this issue, we devise a novel Temporal Saliency Query (TSQ) mechanism, which introduces class-specific information to provide fine-grained cues for saliency measurement. Specifically, we model the class-specific saliency measuring process as a query-response task. For each category, the common pattern of it is employed as a query and the most salient frames are responded to it. Then, the calculated similarities are adopted as the frame saliency scores. To achieve it, we propose a **Temporal Saliency Query Network (TSQNet)** that includes two instantiations of the TSQ mechanism based on visual appearance similarities and textual event-object relations. Afterward, cross-modality interactions are imposed to promote the information exchange between them. Finally, we use the class-specific saliencies of the most confident categories generated by two modalities to perform selection of salient frames. Extensive experiments demonstrate the effectiveness of our method by achieving state-of-the-art results on ActivityNet, FCVID and Mini-Kinetics datasets. Our project page is at `https://lawrencexia2008.github.io/projects/tsqnet`.

Keywords: Video Recognition, Transformer, Temporal Sampling

1 Introduction

In the recent years, video understanding has drawn considerable attention from the community [9,15,45,54,49,41] for the inexorable increase of video content

*: Co-first authorship. ✉: Corresponding author.

on the Internet. Much progress has been achieved on the techniques to model complex video events, which can be glimpsed on promising precision on multiple benchmark datasets [18,35]. However, computational costs grow proportionally to the recognition accuracy. This hinders the deployment of video recognition systems in resource-constraint environments, *e.g.* IoT, self-driving and mobile phone applications. Hence, it is imperative to develop efficient video recognition systems to meet the rising demands of resource-efficient applications.

There are many studies that have been conducted on efficient video recognition. One set of approaches focus on designing lightweight architectures [38,10]. At the other end of the spectrum are the dynamic inference-based approaches, which typically utilize a lightweight policy network to preview the video events, and allocate computation resources depending on the saliency of frames. They implant a policy network (or sampler network) inside the reinforce learning paradigm [47,51,22], or adopt attention weight as a proxy of policy under the attention mechanism [11,13]. The sampler networks are optimized under the assumption that the most salient frames/regions contribute most to the video representation, which produces one-size-fits-all, *i.e.,* class-agnostic frame saliency measurements.

Actually, salient patterns are tightly associated with the category semantics. However, one-size-fits-all saliencies are not sensitive to fine-grained semantics. In particular, the sampler may overestimate the saliency of some frames which seem to be representative, but they actually belong to other categories rather than the real one of the current video. By contrast, a human can precisely elect the most informative frames with the aid of prior information about the probable category of the video. Because we can naturally build the logic connection between frame sequences and the common pattern of the predicted category, which can be understood as a query-response manner. For example, in Figure 1, one can easily select the 3rd, 6th and 7th frames from the video with the assumption that the video may belongs to **Tailgate Party**. By contrast, one-sizes-fit-all sampler may also be inclined to 5th frames besides those three frames for it is quite representative for another category, *e.g.* **Parking Car**.

Inspired by this observation, in this paper, we cast frame saliency measuring as a querying process, to enable discriminative class-specific saliency measurement. To this end, we present a novel **Temporal Saliency Query (TSQ) mechanism**, which can measure saliencies of all semantic categories over frame sequence in parallel, and select the saliency of highly-confident categories as final the result. Concretely, we formulate class-specific saliency measuring as a query-response task. The common patterns of the various categories are adopted as **query**, and frame representations gathered by category-frame similarities are taken as the **response**. Then, the category-frame similarities can be regarded as frame saliencies. A conceptual overview of the TSQ mechanism is shown in Figure 1. Specifically, we use cross attention in Transformer Decoder [39] to model many-to-many category-frame similarities in parallel. On one hand, we represent the common pattern of a category, namely TSQ embedding, by visual prototypes. And the query process is performed over the visual feature of

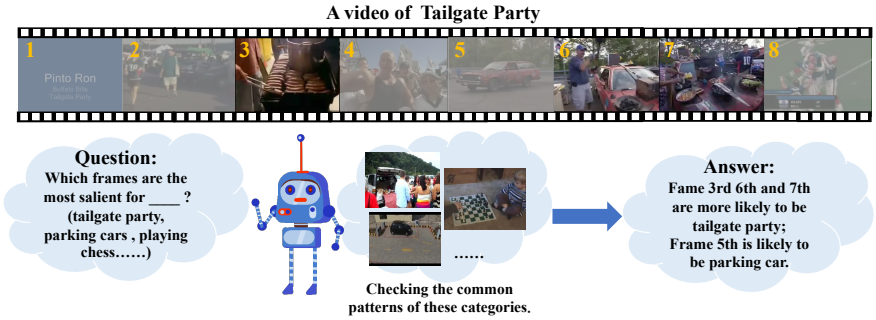


Fig. 1. A conceptual overview of the TSQ mechanism. We cast the saliency estimation task as a query-response task. We ask each category a question: Which frames are the most salient ones for it? As we can see in the above example, we get the answer that frames 3rd, 6th, 7th are most salient for *tailgate party* and the frame 5th is salient for *parking car*. No frame is salient for *playing chess*.

the frame sequence. On the other hand, to handle large intra-class variations of visual appearance, we measure saliency by textual event-object relations for complementary information. As we know, the objects in videos are closely associated with the category annotation of video. For instance, **cake, candle and balloon with birthday party**. To model the semantic relationships between object and category, we first employ BERT [7] to represent the object with word embedding of its name. Taking the product as textual embedding, we construct another textual branch in the TSQ mechanism, where the query process is executed over the embedding sequence of object names. Doing so allows us to exploit prior knowledge from off-the-shelf word representations to supply cross-modal complementary clues to saliency measurement.

Our contributions are summarized as: *First*, we propose a novel Temporal Saliency Query mechanism, to alleviate the lack of class-specific information in saliency measuring for temporal sampling frameworks. *Second*, we present an efficient multi-modal salient frame sampler **T**emporal **S**aliency **Q**uery **Net**work **(TSQNet)**, which utilize both visual appearance feature and textual feature obtained by object name embeddings to measure frame saliencies in a unified framework. *Third*, we conduct extensive experiments on three large-scale datasets, *i.e.,* ActivityNet, FCVID and Mini-Kinetics, which show TSQNet significantly outperforms the state-of-the-art approaches on accuracy-efficiency trade-off.

2 Related Work

Efficient Video Recognition. Efficient video recognition approaches can be roughly categorized into two directions. The first focus on elaborating new lightweight architectures by decomposing 3D convolution operations into 2D and 1D ones [38,53,46], channel shifting in 2D CNNs [23], *etc.* The others are based on a dynamic inference mechanism [48,52,3], which allocates computation

resources on a per-sample basis based on the saliencies of frames. Wu *et al.* [47] utilizes multi-agent reinforce learning to model parallel frame sampling and Lin *et al.* [24] make one-step decision with holistic view. Meng *et al.* [27] and Wang *et al.* [42,44] focus their attention on spatial redundancy. Panda *et al.* adaptively decide modalities for video segments. Most of the previous works are mainly based on reinforce learning or attention mechanism, which are optimized with video classification objectives. However, this paradigm makes produced adaptive sampling policy class-agnostic and lacks discrimination power in fine-grained semantics. In contrast, our temporal sampling-based framework enables discriminative class-specific frame saliency measuring and shows that class-specific mechanism combined with visual-textual multi-modal complementary measuring can push the envelope of the trade-off between accuracy and computation cost.

Transformer in Vision Tasks. Transformer [39] is initially proposed to solve the long-term dependence problem in machine translation. ViT [8], SwinTransformer [26] and DVT [43] split image to patches as words and bring Transformer Encoder to computer vision classification tasks. Query2label [25] apply Transformer Decoder to multi-label classification task. DETR [2] explore using Transformer Decoder for object detection task. Then Transformer Decoder for segmentation is also developed by MaskFormer [5]. The role of Transformer Encoder in C-Tran [21] and TransVG [6] is to model relations between different modalities.

3 Method

Given a video of T frames $X = \{x_i\}_{i=1}^{T}, x_i \in \mathbb{R}^{3\times H\times W}$, our goal is to estimate the saliency score of frames $S = \{s_i\}_{i=1}^{T}$ and sample top K frames with the highest saliency score to feed into a recognition network to obtain final video prediction P. The overview of our method is shown in Figure 2. In this section, we first introduce the Temporal Saliency Query (TSQ) mechanism in Section 3.1. Then we elaborate on the framework of our TSQNet, including two instantiations of TSQ mechanism with visual and textual modalities and cross-modality interactions of them in Section 3.2. Finally, we present the inference procedure of TSQNet in Section 3.3.

3.1 Temporal Saliency Query Mechanism

The goal of Temporal Saliency Query (TSQ) mechanism is to perform frame saliency estimation for all categories simultaneously, which is the shared building block for two branches of visual and textual modalities in TSQNet. To expand generic saliency to class-specific version, we are potentially to ask each category a question: which frames are the most similar ones to the common pattern of it? In this way, we can convert saliency estimation task to query-response task: a learnable embedding initialized with the common pattern of each category is set as the **query**, and the gathered feature from frame sequence with similarities is the **response**. Then the similarities between each category and frame sequence

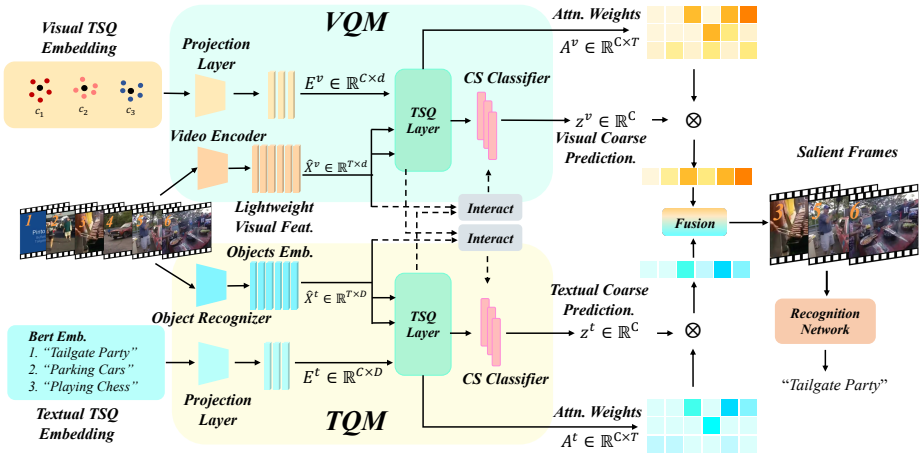


Fig. 2. The overview of the Temporal Saliency Query Networks. Frame sequence is **queried** with visual and textual TSQ embeddings of categories in VQM and TQM, then TSQNet **responded** to the queries by gathering most salient frame representations for each category. And the resultant category-frame similarities are adopted as class-specific saliency measurements for two modalities, which are post-processed and fused for final saliency scores. Top K frames with the highest saliency score are sampled and ingested to an off-the-shelf recognition network for final recognition. Cross-modality interaction ("Interaction") is considered for information exchanging during training. The projection layer is used to reduce the dimension of input features.

can be regarded as the saliency scores. We denote the learnable embedding here as **TSQ embedding**. In TSQ mechanism, a TSQ layer is proposed to enable the query-response functionality and a class-specific classifier is designed to generate coarce predictions of video category and enable discriminative learning of TSQ embedding for each category at the same time. The details of TSQ mechanism are described below.

TSQ Layer. The goal of TSQ layer is to model the many-to-many category-frame similarities simultaneously and enable learning of TSQ embeddings, denoted as $\{E_c \in \mathbb{R}^d\}_{c=1}^C$, under the video classification objective. To achieve this, TSQ layer is build on an attention structure in Transformer [39]:

$$A_c = \text{softmax}(\frac{Q_0 {K_0}^\mathsf{T}}{\sqrt{d}}), \ R_c = A_c V_0, \tag{1}$$

$Q_0 \in \mathbb{R}^d$ is a query matrix, which is obtained by projecting each TSQ embedding E_c with a parameter matrix $W_q \in \mathbb{R}^{d\times d}$: $Q_0 = E_c W_q$. $K_0 \in \mathbb{R}^{T\times d}$ and $V_0 \in \mathbb{R}^{T\times d}$ are the key and value matrix, which are generated by projecting frame feature sequence $X \in \mathbb{R}^{T\times d}$ with different parameter matrices $W_k, W_v \in \mathbb{R}^{d\times d}$: $K_0 = XW_k, V_0 = XW_v$. Then, for the TSQ embedding of the c-th category E_c, the attention weight $A_c \in \mathbb{R}^T$ is produced in querying process realized by scaled dot product operation. Then the value V_0 are gathered with attention weights

A_c and output as response vector $R_c \in \mathbb{R}^d$, which is fed to FFN of [39], *i.e.,* sequential linear layers with residual connections. The output of FFN is ingested to a class-specific classifier to generate classification predictions. In addition to functioning as gathering weights, A_c represent the frame saliency measurements of the c-th category for it characterizes the relations between the c-th category and all T frames. In TSQ mechanism, the more discriminative A_c is, the better the response vectors $\{R_c\}_{c=1}^{C}$ can represent the semantic information of the video, therefore the video classification objective can effectively optimize the this category-frame relation model.

Class-specific Classifier. We denote the output of FFN as $\hat{R} \in \mathbb{R}^{C\times d}$ here. The goal of class-specific classifier ("CS Classifier" in Figure 2) are twofold: (1) project $\hat{R} \in \mathbb{R}^{C\times d}$ to a coarse video prediction $z \in \mathbb{R}^C$, (2) enable class-specific learning of TSQ embeddings. In class-specific classifier, instead of directly using projection layer with weight matrix $W \in \mathbb{R}^{1\times d}$ as $z = W\hat{R} + b$, we apply C projection layers with different weight matrices $\{W_c \in \mathbb{R}^{1\times d}\}_{c=1}^{C}$ to each $\hat{R}_c$ separately. For the c-th category, corresponding element of p is computed as:

$$z_c = W_c \hat{R}_c^{\mathsf{T}} + b_c, \tag{2}$$

where $b_c \in \mathbb{R}^1, b \in \mathbb{R}^C$ are the bias parameters (see Appendix for illustrative examples). This class-specific design endows the response vector of each category with exclusive classifier, which effectively reserves the characteristic of each category and make model converge more easily. z is used for calculating regular cross entropy loss with video labels. Notice here the difference between the coarse video prediction z and the final video prediction P: z is used for saliency measuring while P is the final classification result of the recognition network.

3.2 Temporal Saliency Query Network

Our TSQNet mainly consists of two modules: a Visual Query module and a Textual Query module, which are instantiations of TSQ mechanism with visual and textual representations, respectively. The Visual Query module query the frame appearance sequence with the visual TSQ embedding of each category, and collect the category-frame similarities for class-specific saliency estimation. Textual Query module measures saliencies by modeling event-object (or action-object) relations on the basis of prior knowledge in off-the-shelf language models. Besides, to exchange information between two TSQ modules, cross-modality interactions are performed synchronously during training, which effective compensate scarce scene information for Textual Query module.

Visual Query Module (VQM). The goal of VQM is to generate class-specific saliency measurement from pure visual perspective, which mainly consists of a video encoder, a TSQ layer and a class-specific classifier. The video encoder is a lightweight CNN or transformer backbone, *e.g.,* MobileNetv2 [33] and Mobileformer [4], which extract features from RGB frame sequence $\{x_i\}_{i=1}^{T}$ to feature sequence $\{\hat{x}_i^v \in \mathbb{R}^d\}_{i=1}^{T}$. We further use a 1D convolutional layer to reduce the feature dimension from d to d', which we still denote as $\hat{X}^v = \{\hat{x}_i^v\}_{i=1}^{T}$ for brevity.

TSQ layer takes visual TSQ embedding as query, and frame sequence as key and value, to generate saliency measurements $A^v \in \mathbb{R}^{C\times T}$ from visual features. Class-specific classifier produce visual video coarse predictions z^v, which is further used in the post-processing procedure of saliencies. Next we describe how we obtain visual TSQ embedding.

Following the definition in Section 3.1, visual TSQ embedding $\{E_c^v \in \mathbb{R}^d\}_{c=1}^C$ here is a set of learnable embeddings initialized with common appearance patterns of categories. We propose a simple prototype based representation for common appearance patterns here. Prior works [34] find that, most of the samples belonging to the same class cluster around a prototype in feature space formed by non-linear mapping of networks. We assume that category prototypes can represent the common patterns of categories. Following definitions in [34], we use the averaged features of videos belonging to each category produced by video encoder in the training set, where a video feature is obtained by top-k pooling of frame features (see Appendix for details). A 1D convolutional layer is also used to project E_c^v to the same d'-dimension space with $\hat{x}_i^v$, which is still represented by E_c^v hereafter.

Textual Query Module (TQM). The goal of TQM is to provide knowledge-aware saliency estimation by mining generic event-object relations in videos with the help of prior knowledge in off-the-shelf language models. As observed by prior works [16,13], the event-object (or action-object) relations are generic in videos. Although this knowledge is typically represented in knowledge graph [40], we exploit it in a much more compact fashion, *i.e.,* pre-trained language models. It is proved that the semantic relationships between words can be effectively captured in pre-trained word representations, *e.g.,* Word2Vec [29] and BERT [7]. To model category-frame relations, we first build a object vocabulary $W \in \mathbb{R}^{C_o\times D}$, on a pre-defined object list, *e.g.,* ImageNet-1K category list ($C_o = 1000$) with word embeddings. Then we introduce a lightweight but precise object recognizer to extract appearing object scores from each frame $\{O_i \in \mathbb{R}^{C_o}\}_{i=1}^T$. The frame-level object embedding based feature can obtained: $\hat{X}^t = \{\hat{x}_i^t\}_{i=1}^T, \hat{x}_i^t = O_i W$. Correspondingly, the textual TSQ embedding $\{E_c^t \in \mathbb{R}^D\}_{c=1}^C$ is initialized by pre-trained word embeddings of the category name, to align with textual feature sequence in embedding space. Similar to VQM, we add a 1D convolutional layer to $\{E_c^t \in \mathbb{R}^D\}_{c=1}^C$ and $\hat{X}^t$ to reduce dimensions, which are fed into a TSQ layer and class-specific classifier for textual frame saliency measurements $A^t \in \mathbb{R}^{C\times T}$ and textual coarse video prediction $z^t \in \mathbb{R}^C$.

Cross-modality Interaction. Here we seek to enable information exchange between TSQ layers of two modalities during training and provide guidance, *e.g.,* scene knowledge, from VQM to TQM. To achieve this, we design a novel *swap-attention* structure, which gather the feature sequence with attention weights of the other modality in both VQM and TQM, to generate two additional response vectors:

$$R^{t\to v} = A^t \hat{X}^v, R^{v\to t} = A^v \hat{X}^t, \tag{3}$$

Then the two response vectors based on visual feature sequence R^v and $R^{t\to v}$ are ingested to subsequent layers and compute loss as $\mathcal{L}_v$ and $\mathcal{L}_{t\to v}$. The same

process conducted on textual features sequence renders $\mathcal{L}_t$ and $\mathcal{L}_{v\to t}$. The swap-attention structure is conducive to TQM in two ways: (1) $\mathcal{L}_{t\to v}$ help optimize scene-aware category-frame relation model (2) $\mathcal{L}_{v\to t}$ help optimize scene-aware FFN and classifier. We weighted the existing four losses to obtain the final loss function:

$$\mathcal{L} = \mathcal{L}_v + \mathcal{L}_t + \alpha\mathcal{L}_{t\to v} + \beta\mathcal{L}_{v\to t}, \tag{4}$$

3.3 Inference of TSQNet.

During inference, to yield final saliency measurements, we aggregate the generated frame saliency estimation of high-probability predicted categories for two modalities, respectively, and fuse them for final saliency results.

Saliency Aggregation. Here we only describe saliency aggregation for VQM, which is conducted for TQM with the same way. Intuitively, the higher the probability that a video belongs in c-th category, the higher the priority of the c-th row of attention weights in final saliency result. Following this intuition, we aggregate class-specific saliency measurements of VQM, $A^v \in \mathbb{R}^{C\times T}$ with the coarse video prediction $z^v \in \mathbb{R}^C$. For the i-th frame, the measured saliency of VQM is:

$$s_i^v = \sum_{c=1}^{C} z_c^v A_{c,i}^v, \tag{5}$$

In practive, to filter the noise brought about by the low-confidence categories, we only aggregate saliencies of top-5 categories with highest z^v to get final saliency measurements.

Multi-modality Saliency Fusion. We fuse the saliency measurements of VQM and TQM by taking the union of the top s_i^v frames and top s_i^t frames. The number of frames used for union in two modules are controlled by pre-defined proportion λ_v and λ_t, and the budget of selected frames K.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on three large-scale datasets: ActivityNet, FCVID and Mini-Kinetics. ActivityNet [1] contains 200 categories, it has 10024 videos for training and 4926 videos for validation, where the average duration of videos is 117 seconds. FCVID [17] includes 91,223 videos which 45,611 for training and 45,612 for validation and divided to 239 classes, where the average duration of the videos is 167 seconds. Mini-Kinetics is a small version of Kinetics [18], it consists of 121k training videos and 10k validation ones from 200 categories. Different from first two benchmarks, the videos in Mini-Kinetics are trimmed, with a average length of 10 seconds.

Evaluation metrics. For all datasets above, we apply the official train-val split to experiment our method. Following the previous work, mean Average Precision

Table 1. Example of FLOPs computation.

Module	Arch.	Res.	FLOPs/F	#F	FLOPs
Vis.Enc.	MBv2	188	0.220G	16	3.52G
Obj.Rec.	EN-B0	112	0.098G	16	1.56G
Rec.Net.	RN50	224	4.109G	5	20.55G
VQM	-	-	-	-	0.36G
TQM	-	-	-	-	0.10G
Total	-	-	-	-	26.09G

Table 2. Comparisons with simple baselines.

Method	mAP (%)	FLOPs
Uniform	70.9	195.8G
Random	70.2	195.8G
Dense	71.2	930.8G
MaxConf	74.2	930.8G
MaxConf-L	71.2	54.9G
Ours	74.3	55.3G

(mAP) is used as the main evaluation metric for ActivityNet and FCVID, and Top1 accuracy for Mini-Kinetics. We also evaluate the computation cost with giga floating point operations (GFLOPs).

Implementation details. We adopt MobileNetv2 [33] trained on target datasets as the video encoder in VQM, and Efficientnet-B0 [37] trained on ImageNet-1K as the object recognizer in TQM, respectively. For fair comparisons with previous works, we adopt three backbones in ResNet [14] series, *e.g.,* ResNet-50, 101, 152 for recognition networks. For resolution of frame processed by recognition networks, we follow previous works to scale the shorter side of frames to 256 and then center cropped them to 224×224 for all datasets. On ActivityNet and FCVID, the resolution of frames processed by VQM is 188×188 and one for TQM is 112×112 [1]. On Mini-Kinetics, the resolution is 112×112 for both VQM and TQM. Table 1 shows decomposition of computation cost of TSQNet when adopting ResNet-50 as recognition network. Please refer to Appendix for more implementation details.

4.2 Comparison with Simple Baselines

We compare our TSQNet with some simple baselines with ResNet-101 without TSN-style training as the recognizer in Table 2. There are multiple rule based baselines, "uniform" and "random" stand for uniformly and randomly selecting 10 frames from a video. "Dense" means using all frames of a video. For "MaxConf", we firstly obtain the maximum confidence among all categories for every frame by applying the model along time axis, then select K frames with highest maximum confidence. We also compare with a simple sampler based baseline, "MaxConf-L", which is a lightweight version of "MaxConf" within a uniformly pre-sampled T frames, as the same as "ours". The T in "MaxConf-L" and "ours" is 50, and K in "MaxConf", "MaxConf-L" and "ours" is 5. Our TSQNet obviously presents the best accuracy with limited FLOPs. In fact, "MaxConf-L" is an ablated baseline for our class-specific motivation, which replaces our TSQ mechanism with direct frame-level classification. Comparison with "MaxConf-L" confirms the efficacy of our TSQ mechanism.

[1] Note that the total computation cost of a 188×188 frame processed by MobileNetv2 and a 112×112 frame processed by EfficientNet-B0 equals to the cost of a 224×224 frame processed by MobileNetv2, which is the common setting of previous works [42,13].

4.3 Comparison with State-of-the-arts

Results on AcitivtyNet. We compare the proposed method with recent SOTA methods on AcitivtyNet in Table 3: SCSampler [20], AR-Net [27], AdaMML [30], VideoIQ [36], AdaFcous [42], Dynamic-STE [19] and FrameExit [12]. Experimental result shows that our method outperforms all existing methods with ResNet50 as the main recognition network. Compared with SCSampler [20] which is also a temporal sampling approach, our method surpass it by 3.7% while using 1.6× less computation overhead, which demonstrates the discrimination power of TSQ mechanism in temporal saliency estimation. Comparing to the state-of-the-art method based on early exiting, FrameExit [12], we still outperforms it by 0.5%, which shows our class-specific sampler can find more discriminative frames than this sequential early exiting framework. For a more fair comparison with above pure visual based methods, we also present the results of the visual variant of TSQNet, *i.e.,* 'VQM-only' with comparable computes. Although without text modality, it still surpass the SotA methods, which verify the superiority of our TSM mechanism.

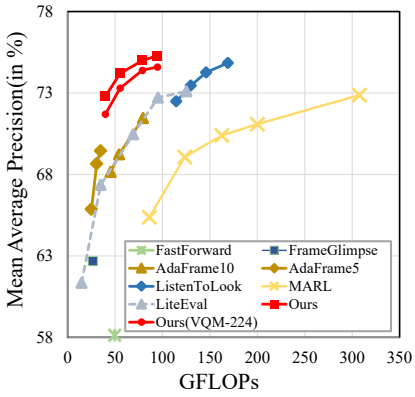


Fig. 3. Comparison of the pure VQM and the whole TSQNet with the state-of-the-art based on ResNet-101 recognition network on ActivityNet.

We further compare TSQNet with SOTA approaches in Figure 3 based on Res101 backbone. Following previous works [51,50,11,47], ResNet-101 without TSN-style training is used as the recognizer, as the same as in Section 4.2. We calculate mAP under different budget K, which varies from 3 to 10. It is shown that our method achieves clearly superior efficiency-accuracy trade-off over all methods. And the result of pure VQM illustrates the efficiency of TSQ Mechanism.

To verify that our TSQNet can collaborate with more backbones, we present experiment results with ResNet-152 and Swin-transformer [26] family as recognition networks in Table 4. It is shown that our method outperforms all method with the same ResNet-152 backbones, and achieves absolute SOTA precision (88.7 Top-1 accuracy and 93.7 mAP) with Swin-Transformer architecture.

Results on FCVID. To verify that performance promotion can be achieved on more untrimmed datasets, we also evaluate our method on FCVID in Table 5, which shows that our method outperforms competing methods in terms of accuracy while saving much computation cost. Compared with SOTA approach AdaFocus [42], which is motivated by selecting salient spatial regions, we achieve

Table 3. Comparisons with SOTA efficient video recognition methods with ResNet50 as recognition backbone on AcitivtyNet. 188 and 224 here represent resolutions.

Method	Backbone	mAP(%)	FLOPs
SCSampler [20]	ResNet50	72.9	42.0G
AR-Net [27]	ResNet18,34,50	73.8	33.5G
AdaMML [30]	ResNet50	73.9	94.0G
VideoIQ [36]	ResNet50	74.8	28.1G
AdaFocus [42]	ResNet50	75.0	26.6G
Dynamic-STE [19]	ResNet18,50	75.9	30.5G
FrameExit [12]	ResNet50	76.1	26.1G
Ours (VQM-only^{188})	ResNet50	75.7	24.3G
Ours (VQM-only^{224})	ResNet50	76.5	26.1G
Ours	ResNet50	**76.6**	26.1G

Table 4. Comparisons with SOTA video recognition methods using ResNet-152 and more advanced recognition networks on AcitivtyNet.

Method	Backbone	Pretrain	Accuracy(%)	mAP(%)
P3D [32]	ResNet-152	ImageNet	75.1	78.9
RRA [55]	ResNet-152	ImageNet	78.8	83.4
MARL [47]	ResNet-152	ImageNet	79.8	83.8
Ours	ResNet-152	ImageNet	**80.0**	**85.2**
ListenToLook [11]	R(2+1)D-152	Kinetics	-	89.9
MARL [47]	SEResNeXt152	Kinetics	-	90.1
Ours	Swin-B	Kinetics	84.7	91.2
Ours	Swin-L	Kinetics	**88.7**	**93.7**

higher mAP with less computation, which implies that our discriminative temporal sampler can capture more salient information of videos.

Results on Mini-Kinetics. We further test the capability of TSQNet on a short trimmed video dataset *i.e.,* Mini-Kinetics, which is more difficult to sample salient frames. Table 6 demonstrates that our method achieves superior Top-1 accuracy (**73.2** *v.s.* 72.9) with 2.0× less FLOPs than the state-of-the-art method [42].

Practical latency. We further conduct experiments of practical efficiency, which shows that our TSQNet significantly surpasses two state-of-the-art methods in inference latency, *i.e.,* FrameExit [12] (9.8 videos/sec *v.s.* **TSQNet 121.1** videos/sec) and AdaFocus [42] (73.8 videos/sec *v.s.* **TSQNet 121.1** videos/sec) [1]. See Appendix for more details.

4.4 Ablation Study

In this section, we inspect different aspects of our proposed TSQNet. All ablations are completed on AcitivtyNet with ResNet-101 as recognition network.

[1] Results are obtained on a NVIDIA 3090 GPU with an Intel Xeon E5-2650 v3 @ 2.30GHz CPU.

Table 5. Comparison with SOTA efficient video recognition methods on FCVID. TSQNet achieves the best mAP with significant computation savings. '188' and '224' are resolutions.

Methods	mAP(%)	FLOPs
LiteEval [50]	80.0	94.3G
AdaFrame [51]	80.2	75.1G
SCSampler [20]	81.0	42.0G
AR-Net [27]	81.3	35.1G
AdaFuse [28]	81.6	45.0G
SMART [13]	82.1	-
VideoIQ [36]	82.7	27.0G
AdaFocus [42]	83.4	26.6G
Ours (VQM-only^{188})	82.9	**24.4G**
Ours (VQM-only^{224})	83.3	26.2G
Ours	**83.5**	26.2G

Table 6. Comparison with state-of-the-art methods on Mini-Kinetics. TSQNet achieves the best Top-1 accuracy with comparable computation cost with the most efficient methods.

Methods	Top-1(%)	FLOPs
LiteEval [50]	61.0	99.0G
SCSampler [20]	70.8	42.0G
AR-Net [27]	71.7	32.0G
AdaFuse [28]	72.3	23.0G
VideoIQ [36]	72.3	20.4G
Dynamic-STE [19]	72.7	18.3G
FrameExit [12]	72.8	19.7G
AdaFocus [42]	72.9	38.6G
Ours (VQM-only)	72.9	**18.1G**
Ours	**73.2**	19.7G

Effectiveness of Class-specific Designs. We investigate the effectiveness of our class-specific designs in TSQ mechanism. Table 7 presents the results of class-specific ("CS") version and class-agnostic ("CA") version of both the attention structure and the classifier in VQM. For attention structure, the class-agnostic version refers to setting the size of visual TSQ embedding set $\{E_c^v\}_{c=1}^C$ to 1. Then generated attention weight $A^v \in \mathbb{R}^{1\times T}$ is directly used as saliency measurement. For the classifier, the class-agnostic version is to replace existing C-projection-layer classifier with a single-projection-layer one as aforementioned in Section 3.1. It is shown that "CS CS" (ours) significantly outperforms "CA CA" choice, which confirms the effectiveness of class-specific information in saliency measurements. Besides, "CS CA" choice presents an unpromising result, which demonstrates that class-specific classifier is critical for TSQ mechanism to function normally in class-specific setting. See Appendix for illustrative examples of these three settings and detailed explanation of comparison of their performance.

Effectiveness of Multi-modal and Fusion and Interactions. To verify the effectiveness of fusion of VQM and TQM and multi-modality interactions, we present experimental results on two individual modalities with different usage of $\mathcal{L}_{t\to v}$ and $\mathcal{L}_{v\to t}$ in Table 8. Without any interactions, fusion of two modules relatively impart improvements on TQM and VQM for 2.9% and 0.3% respectively, which verifies that two modules are complementary. $\mathcal{L}_{t\to v}$ clearly elevate the performance of TQM for better category-frame modelling guided by visual features from VQM. The performance of VQM is also slightly improved by introducing textual-modality attention weights. $\mathcal{L}_{v\to t}$ significantly improves the performance of TQM for better learning of textual FFN and classifier. Finally, when both losses in CIM are added, the results of both TQM and VQM branch are further promoted, and performance of overall TSQNet is obviously improved (**75.3** *v.s.* 74.9). See Appendix for detailed investigations on ratios of two losses.

Table 7. Effectiveness of Class-specific designs.

Attention	Classifier	mAP(%)
CA	CA	74.0
CS	CA	68.7
CS	CS	**74.7**

Table 8. Effectiveness of multi-modality fusion and interactions.

$\mathcal{L}_{t\to v}$	$\mathcal{L}_{v\to t}$	TQM	VQM	Ours
-	-	72.0	74.6	74.9
✓	-	72.5	74.8	75.1
-	✓	72.7	74.6	75.1
✓	✓	**73.1**	**74.8**	**75.3**

Table 9. Results of different textual feature.

Method	Usage	mAP(%)
W2V	Top10	71.2
Glove	Top10	72.0
Bert	All	71.4
Bert	Top10	**72.1**

Table 10. Impacts of initialization of TSQ embedding.

Branch	Init	mAP(%)
Vis	Random	73.8
	Prototype	**74.7**
Text	Random	71.6
	Bert Emb.	**72.1**

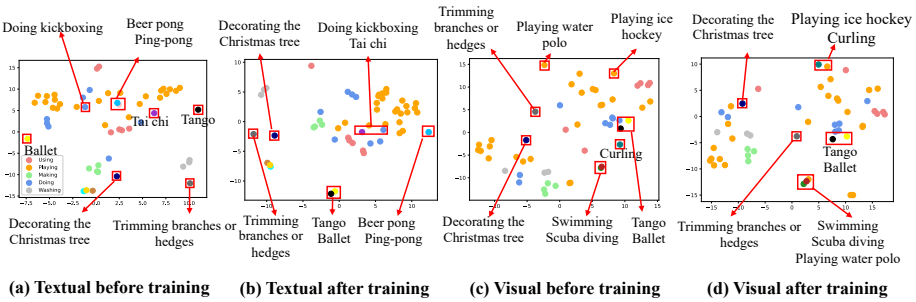
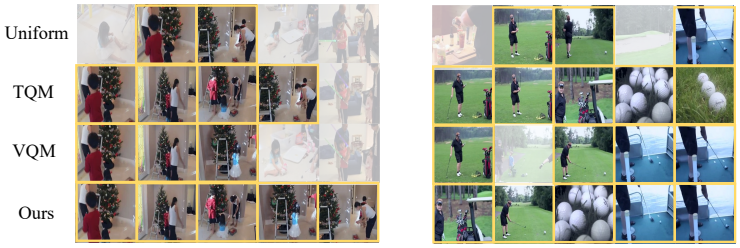


Fig. 4. The visual and textual TSQ embeddings before and after training visualized by t-SNE. The category embeddings with relevant semantics cluster together after training. See Section 4.5 for detailed explanation.

Different Textual Feature. In Table 9, we try three commonly used word embeddings, *i.e.,* Bert [7], Glove [31] and Word2Vec [29], as well as two fashions of usage of object scores O_i, *i.e.,* top-10 object categories ("Top10") and all categories ("All"). Experimental result shows that the Bert embedding with top-10 object score gain the best result, which verifies that both the quality of word embedding and noise filtering of object category count for textual instantiation of TSQ mechanism.

Impacts of Initialization of TSQ Embedding. We further explore the initialization of visual and textual TSQ embeddings in Table 10. The comparison with random initialization confirms that proposed prototype based visual TSQ embedding in VQM and word embedding based textual embedding in TQM provide meaningful and effective initialization for TSQ embeddings.



(a) Decorating Christmas tree (b) Golfing

Fig. 5. Qualitative Evaluation of Sampled frames. We visualized the most salient five frames of uniform and our proposed methods with two samples. The frames with golden border represent the identified salient frames by human intuition, and the frames with mask denote the non-salient ones.

4.5 Qualitative Analysis

We visualize visual and textual TSQ embdding by t-SNE in Figure 4, which shows that our class-specific motivation is highly interpretable in terms of relationships between categories. We also find some categories sharing similar objects are more closer in text TSQ embeddings than in visual ones. For examples, **Decorating the Christmas tree** and **Trimming branches or hedges** share tree or tree-related objects and become closer after training. This may be because TQM measure saliency based on event-object relations, which are more robust against scene variations. In Figure 5, we exhibit some qualitative examples of **Decorating Christmas tree** and **Golfing** for sampled frames by uniform baseline, TQM, VQM and TSQNet. In the case of **Decorating Christmas tree**, it is shown that TQM and VQM are clearly better than uniform baseline. After fusion, TSQNet can sample further more salient frames. Another qualitative example **Golfing** is quite interesting. VQM captures the action moments of swinging a golf club and scenes of a golf course, while TQM captures the golf balls and a golf cart. After fusion, TSQNet select the frames of these object, actions and scenes, which implies our TQM and VQM can cooperate to build a robust sampler aware of object, scene and action information.

5 Conclusions

This paper investigates efficient video recognition by proposing a novel Temporal Saliency Query mechanism and presents an efficient multi-modal salient frame sampler Temporal Saliency Query Network. Extensive experiments verify the proposed method significantly outperforms the state-of-the-art approaches on accuracy-efficiency trade-off. Our proposed method is model-agnostic and can be used with various network architectures. And since our salient score is class-specific, we can easily extend our method to multi-label efficient video recognition.

References

1. 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)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European conference on computer vision. pp. 213–229. Springer (2020)
3. Chen, X., Han, Y., Wang, X., Sun, Y., Yang, Y.: Action keypoint network for efficient video recognition. arXiv preprint arXiv:2201.06304 (2022)
4. Chen, Y., Dai, X., Chen, D., Liu, M., Dong, X., Yuan, L., Liu, Z.: Mobile-former: Bridging mobilenet and transformer. arXiv preprint arXiv:2108.05895 (2021)
5. Cheng, B., Schwing, A., Kirillov, A.: Per-pixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems **34** (2021)
6. Deng, J., Yang, Z., Chen, T., Zhou, W., Li, H.: Transvg: End-to-end visual grounding with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1769–1779 (2021)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. 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)
10. Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 203–213 (2020)
11. 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)
12. 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)
13. 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`
14. 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)
15. 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)
16. Jain, M., Van Gemert, J.C., Snoek, C.G.: What do 15,000 object categories tell us about classifying and localizing actions? In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 46–55 (2015)

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. Lanchantin, J., Wang, T., Ordonez, V., Qi, Y.: General multi-label image classification with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16478–16488 (2021)
22. 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)
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. arXiv preprint arXiv:2201.04388 (2022)
25. Liu, S., Zhang, L., Yang, X., Su, H., Zhu, J.: Query2label: A simple transformer way to multi-label classification (2021)
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. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
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. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014)
32. 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)
33. 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)

34. Snell, J., Swersky, K., Zemel, R.: Prototypical networks for few-shot learning. Advances in neural information processing systems **30** (2017)
35. 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)
36. 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)
37. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. pp. 6105–6114. PMLR (2019)
38. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: CVPR (2018)
39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
40. Wang, H., Zhang, Y., Ji, Z., Pang, Y., Ma, L.: Consensus-aware visual-semantic embedding for image-text matching. In: European Conference on Computer Vision. pp. 18–34. Springer (2020)
41. 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)
42. Wang, Y., Chen, Z., Jiang, H., Song, S., Han, Y., Huang, G.: Adaptive focus for efficient video recognition. arXiv preprint arXiv:2105.03245 (2021)
43. Wang, Y., Huang, R., Song, S., Huang, Z., Huang, G.: Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition. Advances in Neural Information Processing Systems **34**, 11960–11973 (2021)
44. 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. arXiv preprint arXiv:2112.14238 (2021)
45. 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)
46. 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)
47. 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)
48. 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)
49. Wu, W., Sun, Z., Ouyang, W.: Transferring textual knowledge for visual recognition. ArXiv **abs/2207.01297** (2022)
50. 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)
51. 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)
52. Xia, B., Wu, W., Wang, H., Su, R., He, D., Yang, H., Fan, X., Ouyang, W.: Nsnet: Non-saliency suppression sampler for efficient video recognition. ECCV (2022)

53. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: ECCV (2018)
54. 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)
55. 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)