Neurips 2022

Rethinking Resolution in the Context of Efficient Video Recognition

Chuofan Ma*
The University of Hong Kong
b20mcf@connect.hku.hk

Qiushan Guo
The University of Hong Kong
qsguo@cs.hku.hk

Yi Jiang†
ByteDance Inc.
jiangyi0425@gmail.com

Zehuan Yuan
ByteDance Inc.
yuanzehuan@bytedance.com

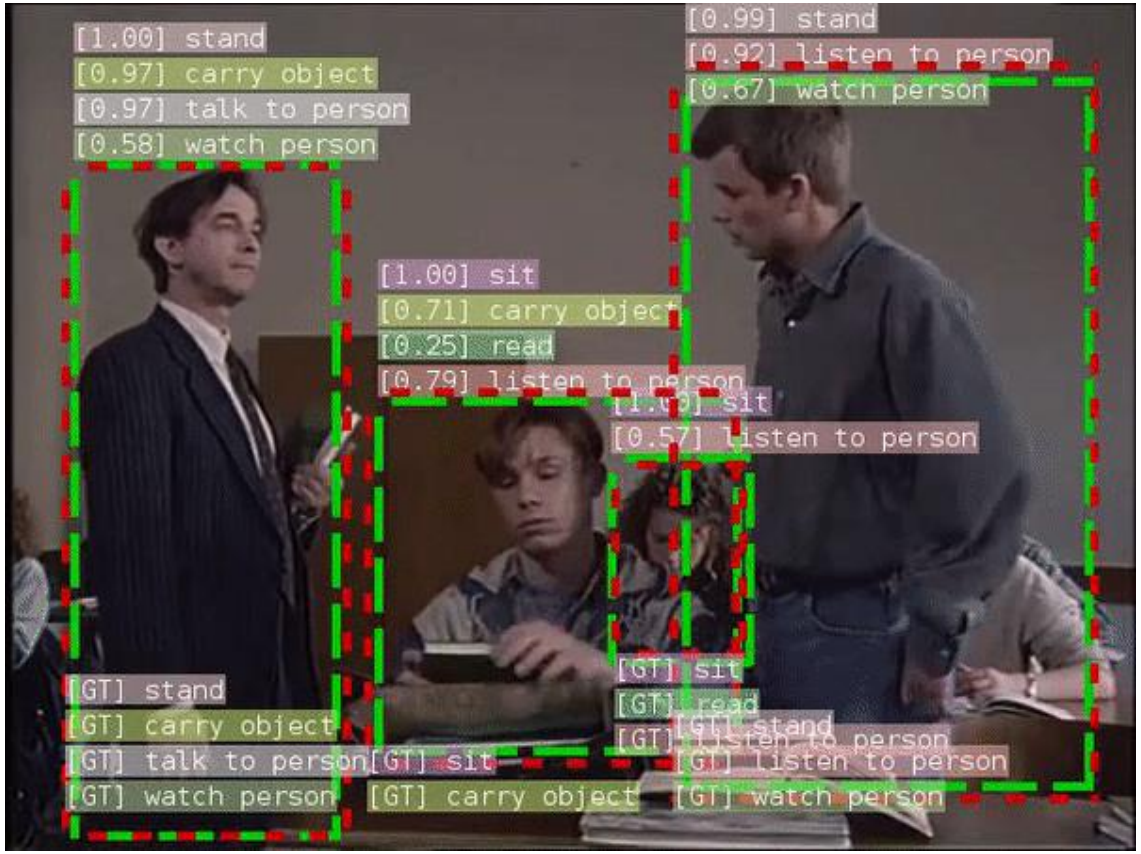
Ping Luo
The University of Hong Kong
pluo@cs.hku.hk

Xiaojuan Qi†
The University of Hong Kong
xjqi@eee.hku.hk

Presenter: Feng Zhu

Efficient Video Recognition

Video Classification is the task of producing a label that is relevant to the video, which describes the entire video.



Efficient Video Recognition

Video Dataset:

1. Kinetics: Kinetics-400, Kinetics-500, Kinetics-600 . The Kinetics-600 is a large-scale action recognition dataset which consists of around 480K videos from 600 action categories.

2. Something-Something: Something-Something V1, V2. Something-Something V2 is a temporal-related dataset which contains 168,913 training videos and 24,777 validation videos over 174 classes.

3. ActivityNet: ActivityNetv1.3 contains 10,024 training videos and 4,926 validation videos from 200 action classes, with an average duration of 117 seconds.

4. FCVIS: FCVID includes 45,611 training videos and 45,612 validation videos labeled into 239 classes, with an average length of 167 seconds.

Efficient Video Recognition

Motivation

Video task requires large amount of computation → developing compact networks or alleviating temporal redundancy of video inputs

Previous works:

1. Temporal redundancy -> Ignore some frames, or low resolution in some frames, e.g. SlowFast.
2. Spatial redundancy -> Focus on specific area, e.g. AdaFocus. AdaFocus V2.

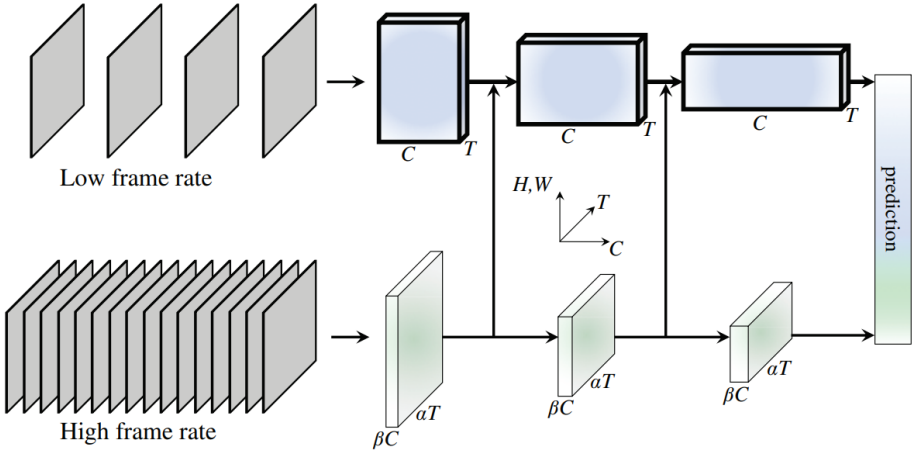
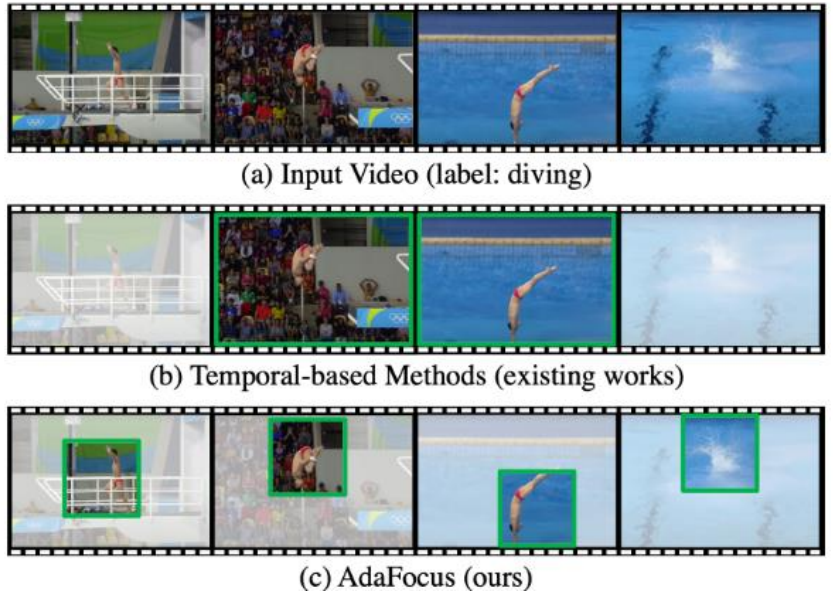


Figure 1. **A SlowFast network** has a low frame rate, low temporal



(a) Input Video (label: diving)

(b) Temporal-based Methods (existing works)

(c) AdaFocus (ours)

Efficient Video Recognition

Motivation

Low resolution → Bad performance ?

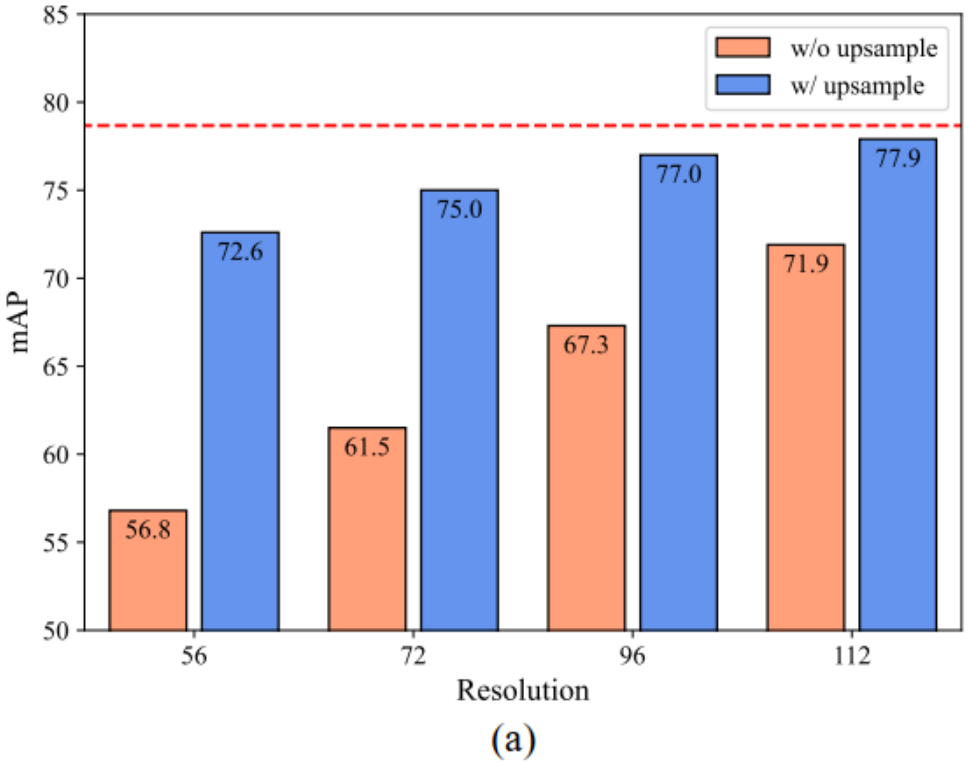
New observations:

1. Low-resolution frames are not necessarily low-quality frames.
2. Mismatch between network and input scale leads to sub-optimal performance at low resolutions.

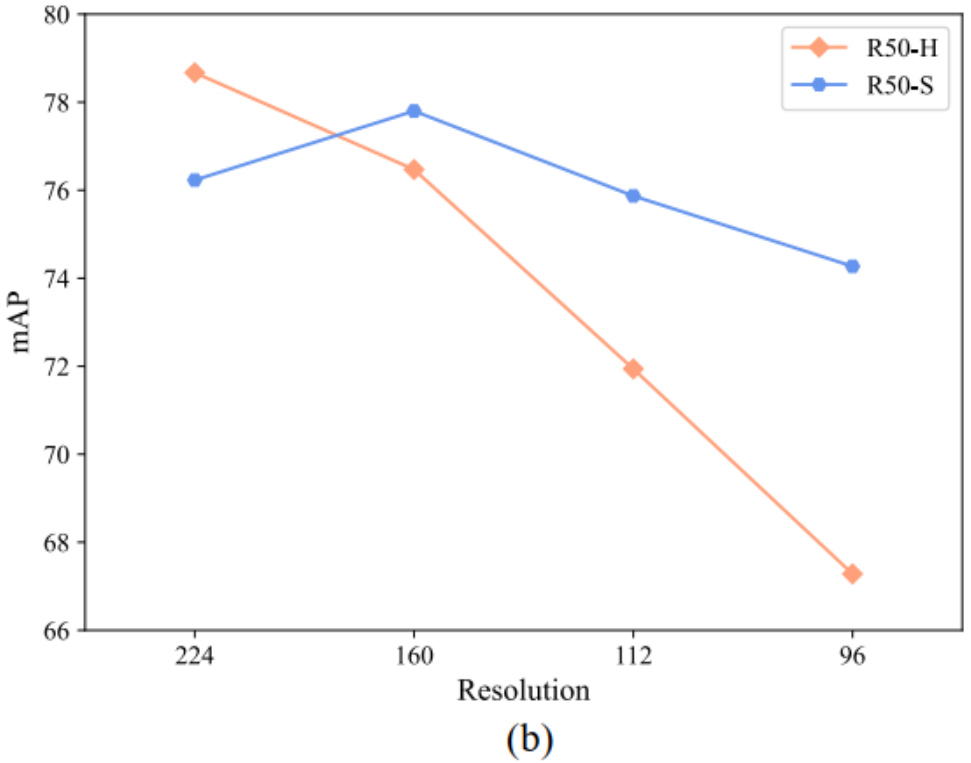
Efficient Video Recognition

A Look into Performance Degradation at Low Resolution

Exp 1. Quality of low-resolution frames



Exp 2. The mismatch between network and input scale.



Efficient Video Recognition

Cross-Resolution Knowledge Distillation

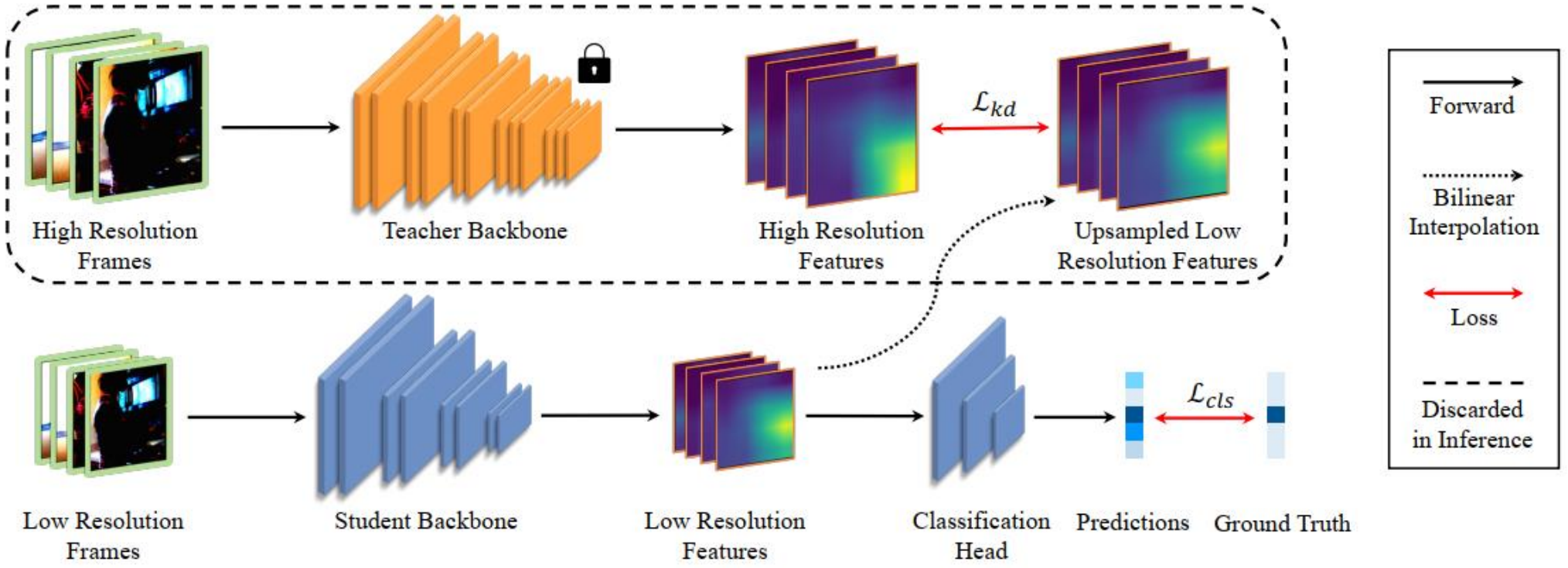


Figure 2: Overview of ResKD: In the training phase, a pre-trained teacher network taking high-resolution frames as input is leveraged to guide the learning of a student network on low-resolution frames. While for evaluation, only the student is deployed to make predictions.

Efficient Video Recognition

Experiment

Table 1: **Comparison with state-of-the-art efficient video recognition methods on ActivityNet-v1.3 and Mini-Kinetics**. MN, MN-T, and RN stand for MobileNetV2, MobileNetV2-TSM, and ResNet, respectively. The best two results are bold-faced and underlined, respectively.

Methods	Backbones	ActivityNet		Mini-Kinetics		FCVID	
		mAP	GFLOPs	Top-1 Acc.	GFLOPs	mAP	GFLOPs
LiteEval [49]	MN+RN101	72.7%	95.1	61.0%	99.0	80.0%	94.3
SCSampler [24]	MN+RN50	72.9%	42.0	70.8%	42.0	81.0%	42.0
ListenToLook [11]	MN+RN101	72.3%	81.4	-	-	-	-
AdaFrame [50]	MN+RN101	71.5%	79.0	-	-	80.2%	75.1
AR-Net [32]	MN+RN18,34,50	73.8%	33.5	71.7%	32.0	81.3%	35.1
AdaFuse [33]	RN50	73.1%	61.4	72.3%	23.0	81.6%	45.0
Dynamic-STE [20]	RN18,50	75.9%	30.5	72.7%	<u>18.3</u>	-	-
FrameExit [13]	RN50	76.1%	26.1	72.8%	19.7	-	-
VideoIQ [39]	MN+RN50	74.8%	28.1	72.3%	20.4	82.7%	27.0
AdaFocus [45]	MN+RN50	75.0%	26.6	72.2%	26.6	83.4%	<u>26.6</u>
AdaFocus V2 [46]	RN50	<u>78.9%</u>	34.1	<u>74.0%</u>	34.1	**84.5%**	34.1
OCSampler [28]	MN-T+RN50	77.2%	<u>25.8</u>	73.7%	21.6	82.7%	26.8
ResKD	RN50	**80.0%**	**17.4**	**75.4%**	**8.7**	<u>84.4%</u>	**17.4**

Efficient Video Recognition

Experiment

Does ResKD work well for SOTA models with dense sampling?

Table 3: **Effectiveness of ResKD on SlowOnly and Video Swin.** ResKD-SlowOnly uses SlowOnly-ResNet50 as the teacher and student networks. ResKD-Swin_S uses Swin_B as the teacher and Swin_S as the student. Since Swin_B and Swin_S have different numbers of output channels, some adjustments are made to ResKD, which is discussed in detail in Appendix. Numbers in brackets denote the input resolution.

Models	Top-1 Acc.	GFLOPs
SlowOnly (224)	74.5%	1260
SlowOnly (112)	69.3%	332
ResKD-SlowOnly	73.1%	332

Models	Top-1 Acc.	GFLOPs
Swin_S (224)	80.1%	1639
Swin_S (112)	76.3%	421
ResKD-Swin_S	80.0%	421

Efficient Video Recognition

Experiment

Is ResKD scalable with varying resolutions?

Table 4: **Scalability of inference speed regarding input resolutions.** Throughput (number of videos processed per second) is measured on a single Tesla V100 SXM2 GPU with the batch size of 64.

Resolution	224p	144p	112p	96p	72p	56p
mAP	81.7%	81.3%	80.0%	76.2%	73.4%	70.5%
GFLOPs	65.9	28.3 (2.3↓)	17.4 (3.8↓)	12.1 (5.4↓)	8.3 (7.9↓)	4.8 (13.7↓)
Throughput	78.7	163.2 (2.1↑)	262.8 (3.3↑)	333.9 (4.2↑)	434.7 (5.5↑)	631.5 (8.0↑)

Efficient Video Recognition

Experiment

How does ResKD help low-resolution video recognition?

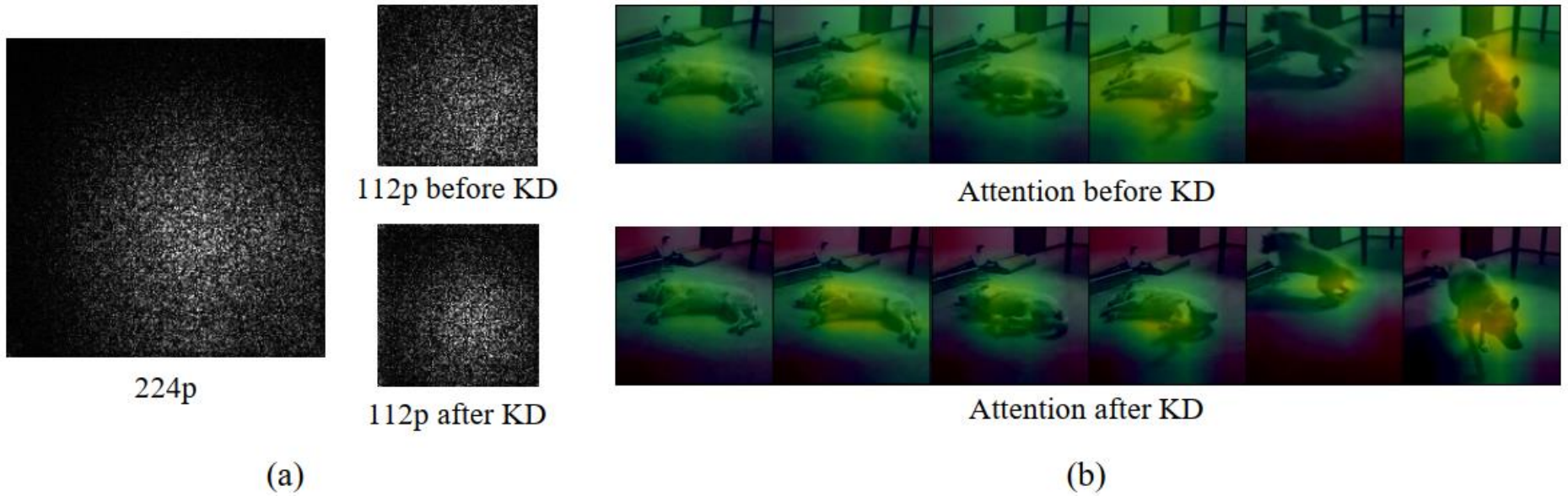


Figure 4: **Visualization before and after ResKD.** In figure (a), we choose a pixel close to the bottom right from the last-layer feature map for ERF visualization, as there is no central pixel in the feature map of 112p input. In figure (b), the focused region is highlighted in bright yellow.

Efficient Video Recognition

Experiment

Table 5: **Ablation on design choices of ResKD.** KL Div. and MSE stand for Kullback–Leibler Divergence and Mean Squared Error, respectively. We report the results on ActivityNet, with ResNet-50 as student backbone and ResNet-152 as teacher backbone.

KD method	Supervision signal	Loss type	Temporal dim.	Spatial dim.	mAP
Baseline	N/A	N/A	✗	✗	71.8%
Clip-level KD	cls. score	KL Div.	✗	✗	73.4%
Frame-level KD	cls. score	KL Div.	✓	✗	76.6%
Pixel-level KD	feature map	MSE	✓	✓	78.5%

Efficient Video Recognition

Experiment

Table 6: **Ablation on model distillation and resolution distillation.** R152 and R50 stand for ResNet-152 and ResNet-50, respectively.

Settings	Teacher	High resolution	mAP
Baseline	N/A	✗	71.8%
Model distill.	R152	✗	73.3%
Resolution distill.	R50	✓	76.2%
ResKD	R152	✓	78.5%