北京大學 PEKING UNIVERSITY

CVPR JUNE 17-21, 2024 SEATTLE, WA

Active Object Detection With Knowledge Aggregation and Distillation from Large Models

Dejie Yang, Yang Liu*

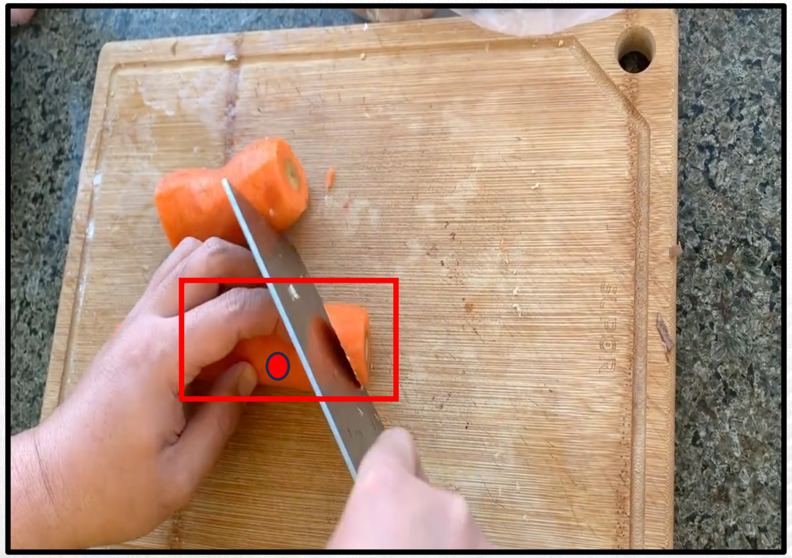
Wangxuan Institute of Computer Technology, Peking University

https://github.com/idejie/KAD

Overview

Active Object Detection(AOD)

- Detect the bounding box of the active object which is **undergoing state-change**
- For example: “**carrot** undergoing cutting”, “**pot** undergoing cleaning”



Motivation

Main Challenges of AOD

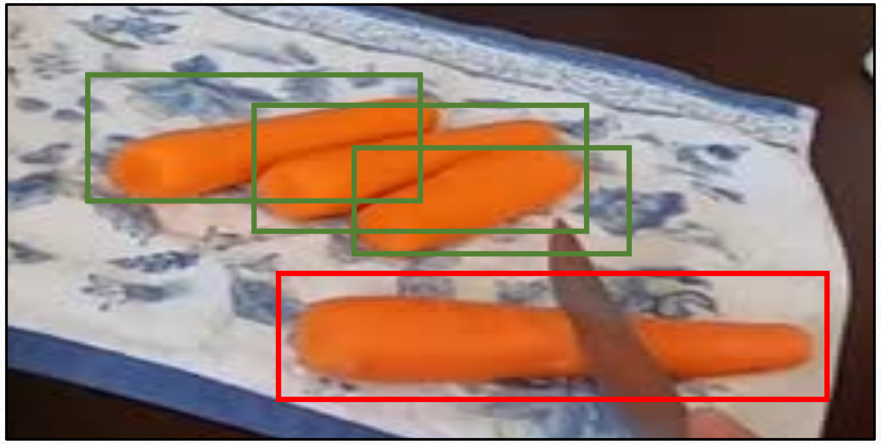
(1) **The large intra-class visual appearance variance** for the same object under state changes
e.g., carrot can be 'cutting using a knife', 'breaking by hands' or 'making into juice'

(2) **The subtle visual changes** between the instance undergoing state-change or not
e.g., multiple distracting no-change instances of the same category

1. *cutting using a knife*
2. *breaking by hands*
3. *making into juice*

…

(1)Diverse interactions and large intra-class variance



(2)Subtle visual difference and multiple distractors

Motivation

北京大学 PEKING UNIVERSITY

Contributions

(1) Introduce a **Knowledge Aggregator** that incorporates three-fold commonsense:
plausible semantic interactions, **fine-grained visual** and **spatial priors**

(2) To **avoid the extra input** at inference, propose a **Teacher-Student Knowledge Distillation** strategy

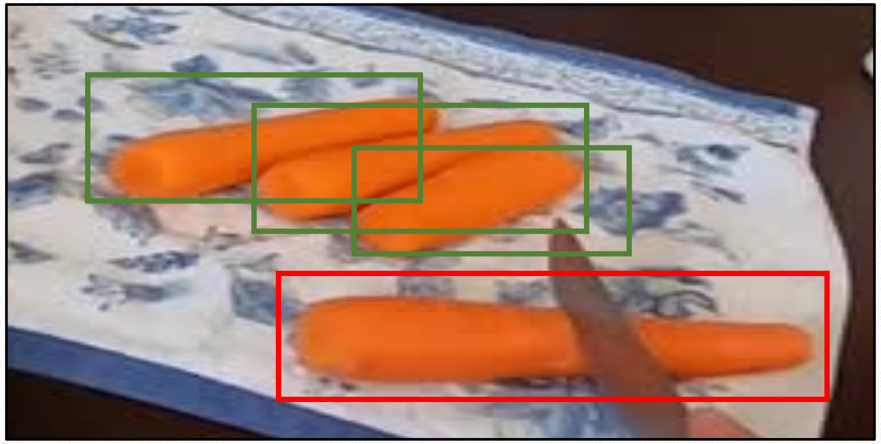
(3) Our proposed framework achieves **state-of-the-art performance on four datasets**

1. cutting using a knife

2. breaking by hands

3. making into juice

…



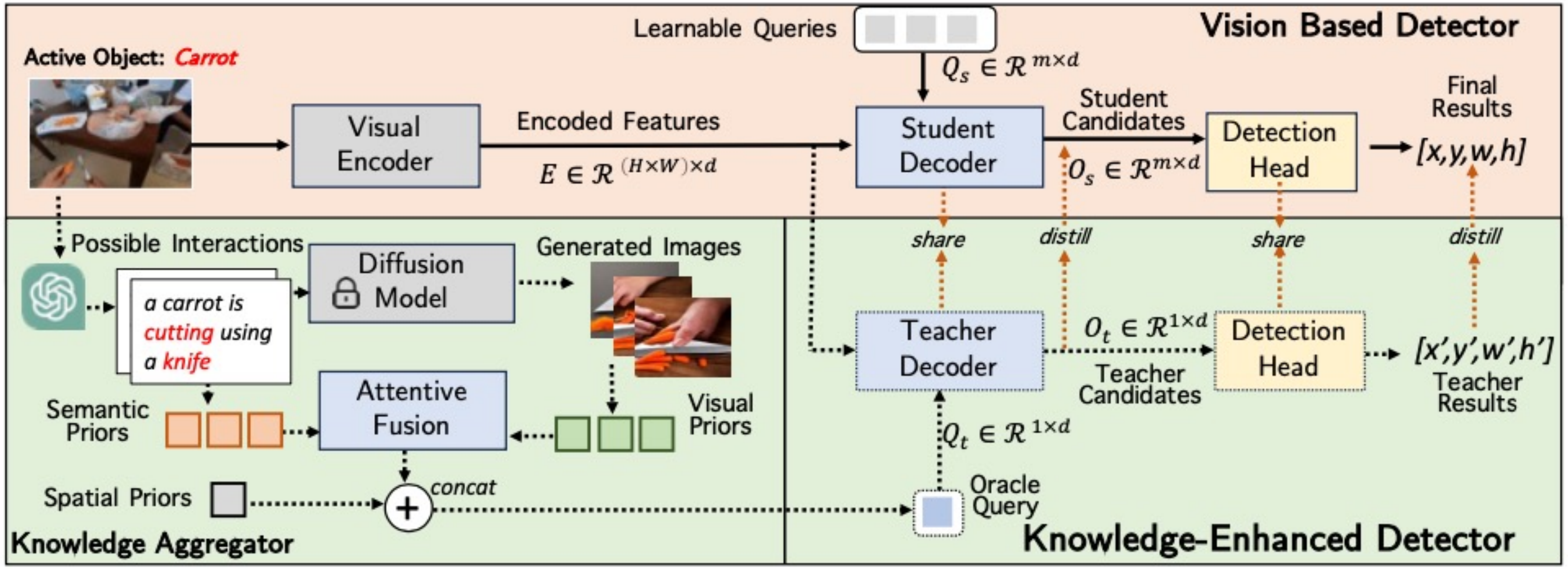
(1)Diverse interactions and large intra-class variance

(2)Subtle visual difference and multiple distractors

Proposed Method: KAD

Framework

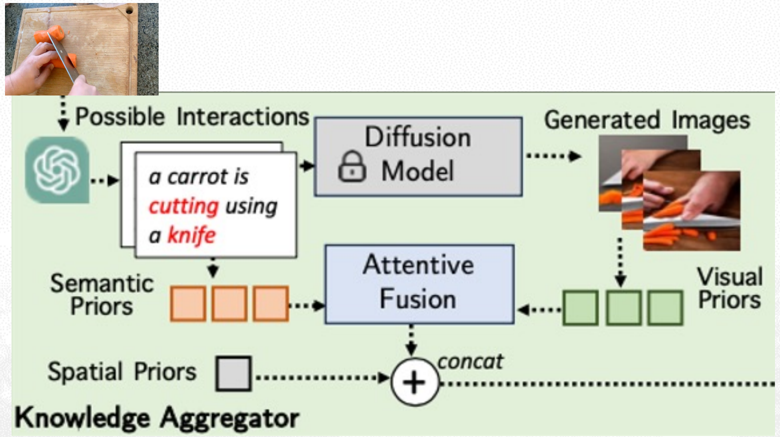
- **Vision Based Detector (Student):** to detect **active object without extra inputs**, introduce a Transformer Detector
- **Knowledge Aggregator:** to collect the **semantic-aware, visual-assisted and spatial-sensitive knowledge**, large models(GPT and Diffusion Models) and Attentive Fusion module
- **Knowledge-Enhance Detector(Teacher):** to enrich the detection process with **pertinent priors linked to active objects**, a Transformer Decoder and a Detection Head



Proposed Method: KAD

KAD: Knowledge Aggregator

- Approach:
  - *Generate*: Semantic(Interactions by GPT), Vision(Images by Diffusion Models) ,Spatial(gt bbox)
  - *Fuse*: fuse the triple priors with attention layer
- construct triple complementary priors to guided the model to distinguish where to pay more attention.



Prompt: describe 10 interaction descriptions of *[object]* undergoing state change (including tools)

1. Carrot is being washed using a faucet.
2. Carrot is being peeled using a peeler.
3. Carrot is being sliced using a knife.
4. Carrot is being grated using a grater.
5. Carrot is being boiled using a pot.
6. Carrot is being steamed using a steamer.
7. Carrot is being roasted using an oven.
8. Carrot is being pureed using a blender.
9. Carrot is being juiced using a juicer.
10. Carrot is being fermented using a fermentation jar and salt.

Generated Interactions

Generated Images

Attentive Fusion:

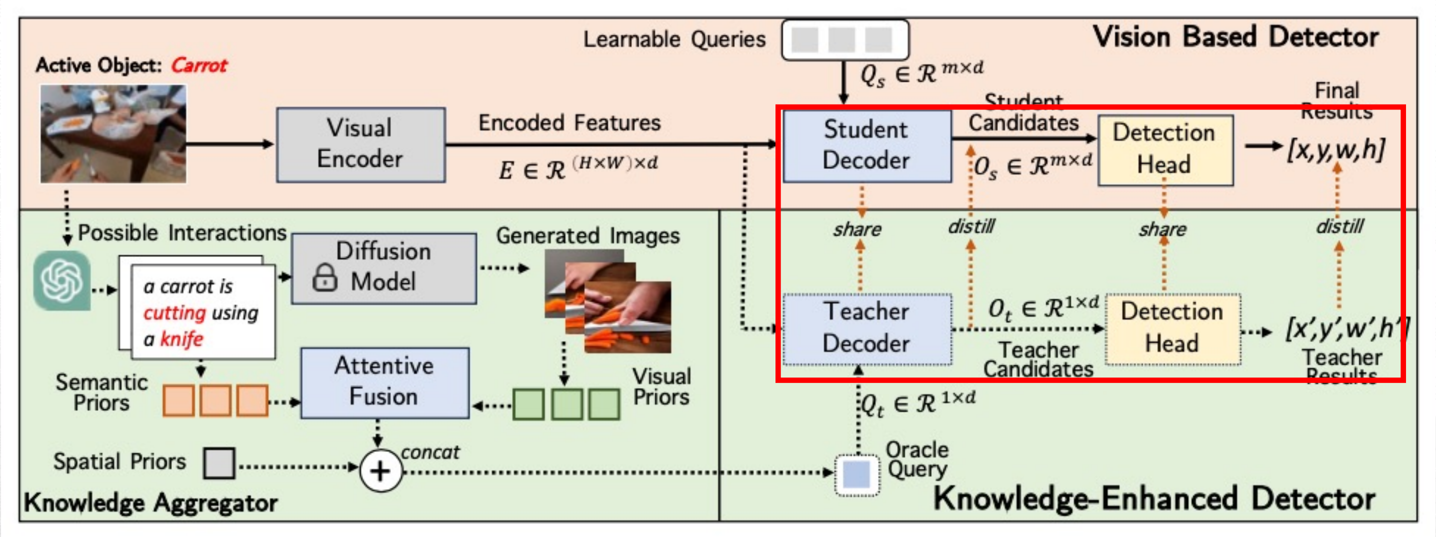
$$\mathbf{T} = pool(selfattn(\boxed{(\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_p)})) \in \mathbb{R}^{1\times d_t}$$

文本（图像）向量

Proposed Method: MRT

MRT: Knowledge Distillation

- Approach:
  - *Parameters Sharing*: decoder and detection head
  - *Knowledge Distillation*: Feature and Attention Distillation
- Leverage the oracle query to facilitate accurate representation learning
- Allow the student to emulate the ability of teacher to navigate dynamic distractors



Attention
$$L_{attn} = \sum \mathrm{KL}(A_t^l, A_{s_i}^l),$$

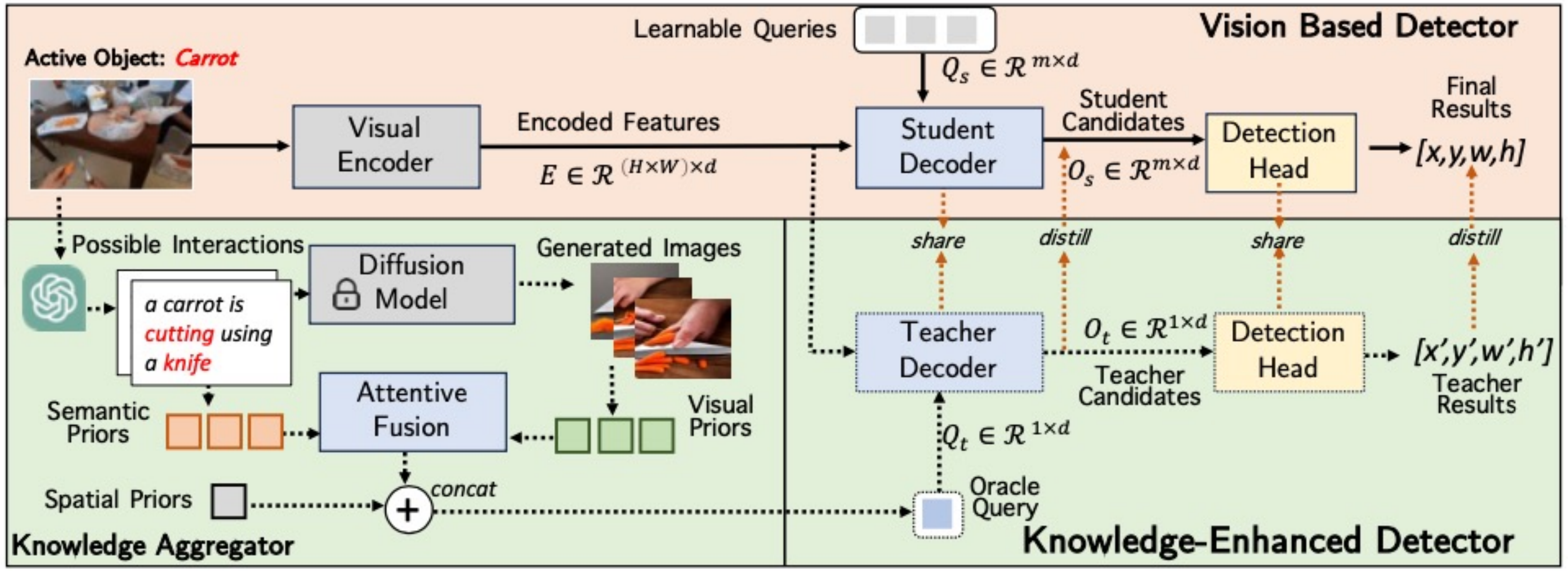
Feature
$$L_{emb} = \sum^{l} \left(1 - \frac{{O_t^l}^T O_{s_i}^l}{\|O_t^l\|_2 \|O_{s_i}^l\|_2}\right).$$

$$L_{distill} = L_{emb} + \eta L_{attn}$$

Proposed Method: KAD

The Overall Objective

- Student model : $$\mathcal{L}_v = BCE\underbrace{\boxed{(s, \hat{s}_i)}}_{\text{类别}} + \lambda(\mathcal{L}_{giou}\underbrace{\boxed{(b, \hat{b}_i)}}_{\text{边界框}} + ||b - \hat{b}_i||_1)$$
- Teacher model(only train): $$\mathcal{L}_k = BCE(s, \hat{s}_t) + \lambda(\mathcal{L}_{giou}(b, \hat{b}_t) + ||b - \hat{b}_t||_1)$$
- Overall: $$L = L_v + L_k + \alpha L_{distill}$$



Experiments

State-of-the-art performance on 4 benchmarks

Ego4D
(Daily Activity)

Epic-Kitchens
(Activity in Kitchens)

MECCANO
(Toy Assembly)

100DOH
(Daily Activity)

Experiments

State-of-the-art performance on 4 benchmarks

Ego4D

Method	Backbone	Val-Set AP	AP50	AP75
CenterNet [37]	DLA-34	6.4	11.70	6.10
FasterRCNN [23]	ResNet-101	13.4	25.6	12.5
100DOH-model [25]	ResNet-101	10.7	20.6	10.1
DETR [1]	ResNet-50	15.5	32.8	13.0
KAD(ours)	ResNet-50	**31.4**	**34.6**	**28.9**
InternVideo[31]	Uniformer-L	24.8	44.2	24.0
	Swin-L	36.4	56.5	37.6
KAD(ours)	Swin-L	**40.5**	**60.6**	**41.9**

Epic-Kitchens

Method	Backbone	Val-Set AP	AP50	AP75
DETR [1]	ResNet-50	10.4	15.7	10.1
KAD(ours)	ResNet-50	**30.2**	**30.1**	**22.5**
InternVideo[31]	Uniformer-L	19.4	38.7	17.0
	Swin-L	28.3	39.8	27.2
KAD(ours)	Swin-L	**35.2**	**44.1**	**32.5**

100DOH

Method	Backbone	AP75	AP50	AP25
100DOH-model [25]	ResNet-101	28.5	47.0	51.8
PPDM[17]	DLA-34	26.9	45.8	53.0
HOTR[15]	ResNet-50	29.3	49.3	57.8
Seq-Voting[9]	ResNet-101	29.9	53.0	57.2
KAD(ours)	ResNet-101	**31.2**	**53.9**	**58.9**

MECCANO

Method	Backbone	AP75	AP50	AP25
100DOH-model [25]	ResNet-101	-	20.2	-
Seq-Voting[9]	ResNet-101	13.0	26.3	34.9
KAD(ours)	ResNet-101	**14.4**	**28.8**	**36.2**

Experiments

Ablation Study of priors

- Each type of prior shows a performance gain.
- Combining the priors achieves the best results.

No.	Knowledge	AP	AP50	AP75
1	VBD(baseline)	35.9	55.8	36.9
2	VBD+visual	36.0	56.6	37.2
3	VBD+semantic	36.5	57.1	37.1
4	VBD+spatial	36.1	56.8	37.0
5	VBD+spatial+semantic	37.9	58.1	38.3
6	VBD+visual+semantic	39.8	59.3	40.0
7	VBD+visual+spatial	38.5	58.0	38.5
8	VBD+spatial+semantic+visual	**40.5**	**60.6**	**41.9**

Experiments

Ablation Study of distillations

- potential of leveraging feature distillation to foster the acquisition of detection capabilities by the student model (Vision-Based Detector) from the teacher model (Knowledge-Enhanced Detector).
- synergistic potential of comprehensive distillation strategies that not only align features but also bridge the gap between attentions.

No.	Distillation	AP	AP50	AP75
1	VBD	35.9	55.8	36.9
2	VBD *w emb*	38.3	59.3	41.2
3	VBD *w emb&attn*	**40.5**	**60.6**	**41.9**

Ablation Study of number of generated priors

- the scene of a state change of an object may be diverse, so diverse descriptions and images are necessary.
- When the number of generated priors is large, it can bring more improvement

No.	Number of descriptions	AP	AP50	AP75
1	No-description	37.3	57.8	37.7
2	1-description	37.5	57.9	37.7
3	10-descriptions	**40.5**	**60.6**	**41.9**

No.	Number of generated images	AP	AP50	AP75
1	No-image	37.9	58.1	38.3
2	1-image	38.1	58.2	38.4
3	10-images	39.5	58.7	39.1
4	100-images	**40.5**	**60.6**	**41.9**

北京大学 PEKING UNIVERSITY

Ablation Study of aggregation approaches

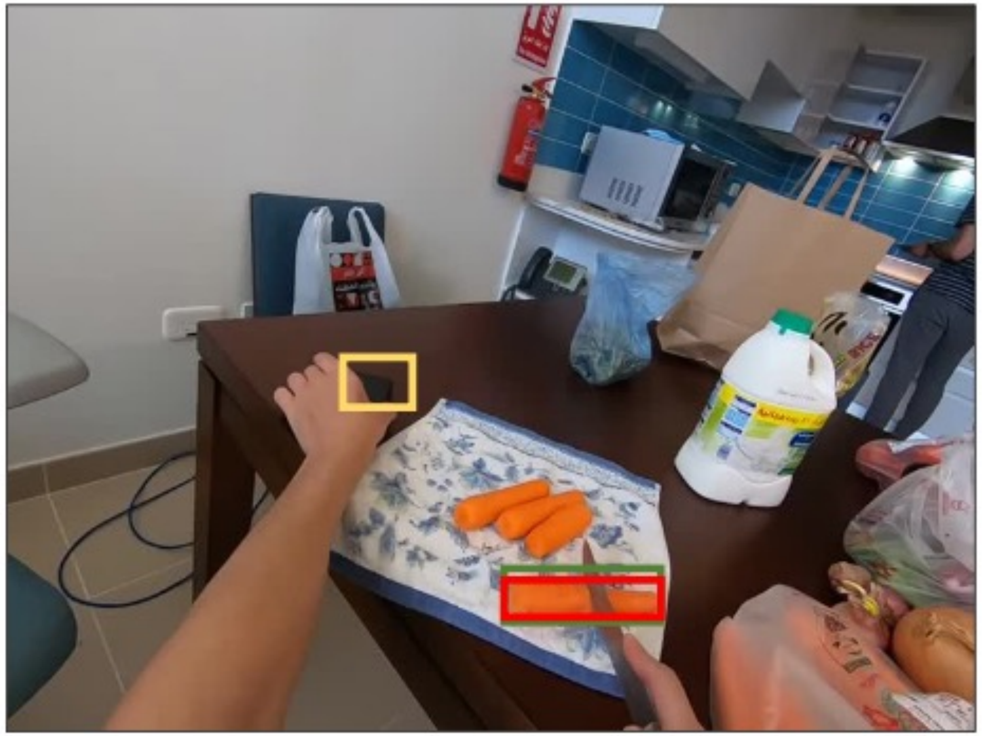
- Attentive operation has brought about 1.3% improvement on AP, which shows adaptive selection contributes to AOD.

No.	method	AP	AP50	AP75
1	max	39.2	59.5	39.6
2	avg	39.1	59.2	39.7
3	attentive	**40.5**	**60.6**	**41.9**

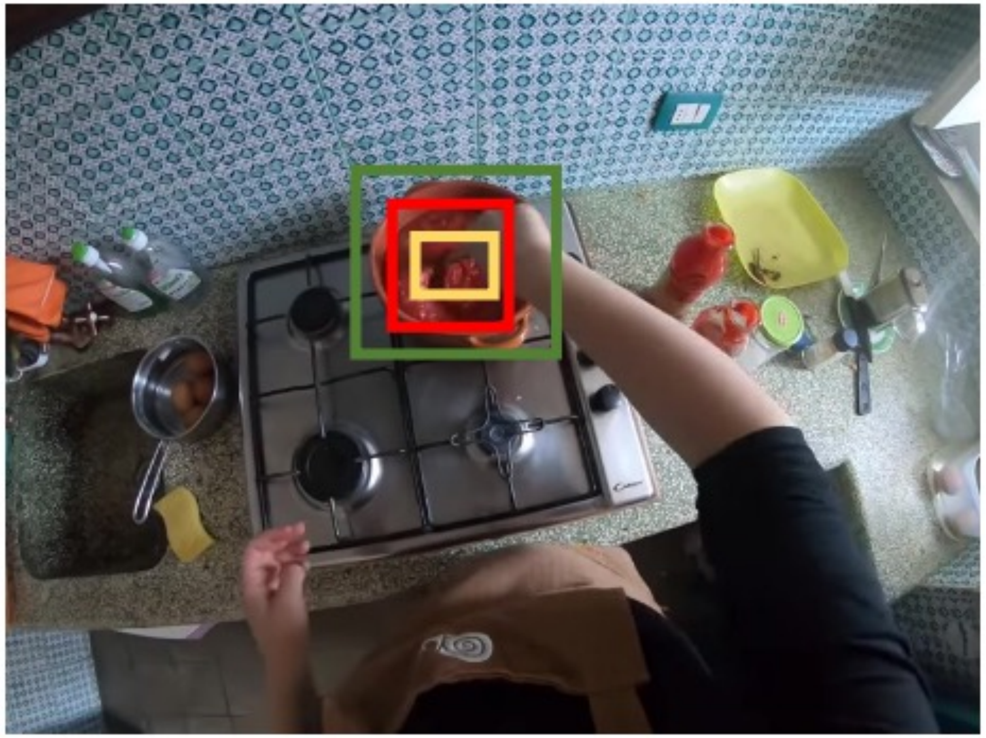
Experiments

Visual analysis

- The incorporation of related priors to active objects, effectively guides the detection process towards active objects.



(a) active object: carrot.



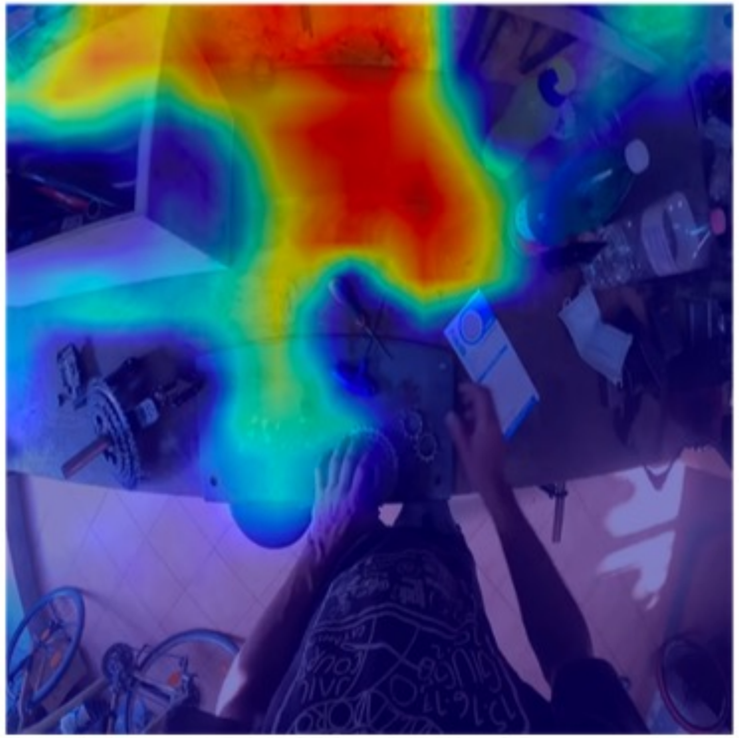
(b) active object: food.

Yellow: previous best methods (InternVideo), **Red**: ours, **Green**: ground-truth

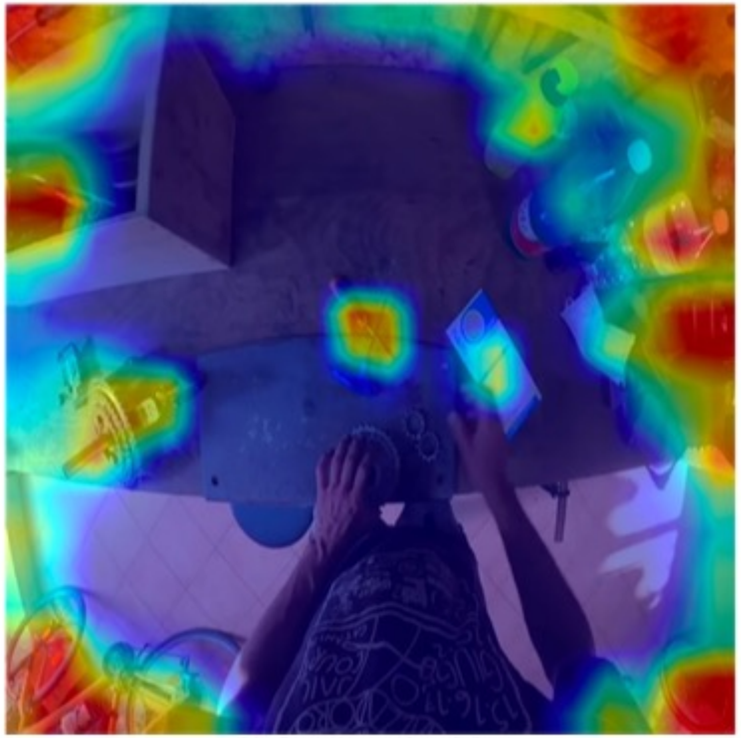
Experiments

Attention Map Visualization

- introduce prior knowledge of the active object to guide the model in inferring and locating the active object by analyzing potential interactions



(a) InternVideo Attention Map



(b) Our Attention Map

(c) Detection Results.

Attention Map: colors from blue to red means the attention from less to more

THANKS FOR YOUR ATTENTION

SCAN THE QR CODE FOR PROJECT DETAILS

Code Available

https://github.com/idejie/KAD