Active Object Detection with Knowledge Aggregation and Distillation from Large Models

Dejie Yang Yang Liu*
Wangxuan Institute of Computer Technology, Peking University
ydj@stu.pku.edu.cn yangliu@pku.edu.cn

Abstract

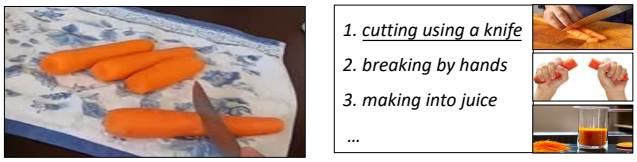
Accurately detecting active objects undergoing state changes is essential for comprehending human interactions and facilitating decision-making. The existing methods for active object detection (AOD) primarily rely on visual appearance of the objects within input, such as changes in size, shape and relationship with hands. However, these visual changes can be subtle, posing challenges, particularly in scenarios with multiple distracting no-change instances of the same category. We observe that the state changes are often the result of an interaction being performed upon the object, thus propose to use informed priors about object related plausible interactions (including semantics and visual appearance) to provide more reliable cues for AOD. Specifically, we propose a knowledge aggregation procedure to integrate the aforementioned informed priors into oracle queries within the teacher decoder, offering more object affordance commonsense to locate the active object. To streamline the inference process and reduce extra knowledge inputs, we propose a knowledge distillation approach that encourages the student decoder to mimic the detection capabilities of the teacher decoder using the oracle query by replicating its predictions and attention. Our proposed framework achieves state-of-the-art performance on four datasets, namely Ego4D, Epic-Kitchens, MECCANO, and 100DOH, which demonstrates the effectiveness of our approach in improving AOD. The code and models are available at https://github.com/idejie/KAD.git.

1. Introduction

Active object detection (AOD) focuses on localizing key objects that are undergoing state changes as a result of a sequence of human actions, interactions, or manipulations, which has broad potential applications [8, 12, 13].

Existing methods explore conventional object detectors, i.e., Faster RCNN [23], DETR [1], CenterNet [6], or specialized hand-object interaction models [9, 26] for AOD, which primarily rely on visual appearance of the objects *within input*, such as changes in size, shape and its relationship with hands. However, only using visual cues is often inadequate for AOD due to the following reasons: (1) *The visual changes can be subtle between the instance undergoing state-change or not*, posing challenges, particularly in scenarios with multiple distracting no-change instances of the same category in Figure1a, or the target active object suffers from hands occlusion. (2) *The intra-class visual appearance variance for the same object under state changes is large*. As shown in Figure1b, a state-change of the carrot can be caused by many interactions, such as 'cutting using a knife', 'breaking by hands' or 'making into juice'.

In practice, we notice that alterations in the object's state frequently stem from interactions carried out on the object. This underscores the significance of common-sense understanding of object affordance in the context of AOD. If we can establish well-informed priors that model this commonsense knowledge concerning plausible object interactions beforehand (out of the scope of the current input), we can uncover the aforementioned 'intra-class variance'. In principle,



(a) subtle difference and multiple distractors

(b) diverse interaction and large intra-class variance

Figure 1. **An example of state-change carrot.** Active objects detection (state change carrots) is difficult, as there are (1) visual changes can be subtle between the carrot undergoing state-change or not, and multiple distractors, (2) intra-class visual appearance variance for the carrot under state changes is large. To achieve accurate detection, we propose to construct triple priors to provide hints for the model, including semantic interaction priors, fine-grained visual priors, and spatial priors of active objects.

*Corresponding Author

this revelation can be leveraged to assist AOD, analogous to leveraging hints during exams to facilitate question-solving.

Motivated by this, in this paper, we propose a new framework, namely 'knowledge aggregation and distillation'(KAD), aiming at aggregating the object affordance common sense knowledge and incorporating it into the AOD process as shown in Figure2. Firstly, we propose to model such common sense in three ways :(1) we employ a language model to provide *semantic interaction priors* of "how the active object can be interacted," capturing multiple possibilities that may induce changes in the object's state; (2) we use an image generation model to equip the above semantic by providing vivid images depicting state changes during corresponding interactions, offering *fine-grained visual priors*; (3)we utilize the ground truth position of the active object as *spatial priors*, indicating regions that require heightened attention to enhance the model's spatial sensitivity, particularly in the presence of multiple distractors. Secondly, we introduce a Knowledge Aggregator designed to reconcile conflicts and harness complementary cues from the three aforementioned common-sense modalities, serving as an oracle query for the decoder. Thirdly, in practical scenarios, the category of active objects remains unknown during inference, rendering the acquisition of the aforementioned oracle query challenging. As a solution, we propose a distillation strategy that uses a plain detector as student (using learnable queries) to mimic the attention and intermediate outputs in a teacher detector using an oracle query. Throughout the training process, we facilitate the transfer of knowledge from the 'cheated' teacher to the student through distillation, thereby endowing the student with a level of proficiency in active object detection ability akin to that of the teacher. During inference, we leverage the student detector exclusively, effectively circumventing the need for unnecessary inputs.

Our contributions can be summarised as follows:

- We introduce a Knowledge Aggregator that incorporates three-fold commonsense pertaining to active objects, encompassing plausible semantic interactions, fine-grained visual and spatial priors, serving as a 'cheated' teacher to facilitate more accurate AOD localization.
- To avoid the extra commonsense input at inference, we propose a Teacher-student Knowledge Distillation strategy, enabling the training of a simple student detector that possesses robust AOD capabilities by mimicing the attention and intermediate outputs from the teacher.
- Comprehensive experiments conducted on extensive egocentric datasets, Ego4D, Epic-Kitchens, MECCANO and 100DOH, demonstrate the efficacy of our proposed approach.

2. Related Work

2.1. Active Object Detection

Active object detection (AOD) involves identifying manipulated objects entwined with human actions[7, 9, 29, 31]. InternVideo[31] proposes a series of models for Ego4D tasks. For state-change object detection, InternVideo[31] explores the transfer learning from general object detection to egocentric state-change object detection, which takes the Swin-Transformer[18] as the backbone and employs DINO[35] as the detection head. HOTR[15] proposes an end-to-end human or hand object interaction detection transformer, which detects hand(or human) and object independently with a detector and identifies the hand-object interaction to match hands and objects detected in the detector. Seq-Voting[9] takes the hand detection results as cues and proposes a voting function with a box field to leverage each pixel of the input image as evidence to predict the bounding box of the active object. While the investigation of hand-object interactions[9, 15, 26] stands out as a crucial facet of object manipulation, it is crucial to recognize that objects undergoing state changes might not always be directly engaged with hands. Only using visual cues is often inadequate because of the subtle difference between state-change or not and the large intra-class visual appearance of the same active object. In this paper, we aim to provide three-fold priors: semantic interaction priors, fine-grained visual priors, and spatial-sensitive priors to enhance active object detection.

2.2. Object-State Change

Human actions often induce changes of the state of an object. Previous works have studied detecting object states in images [10, 19, 20, 24] or learning actions and their modifiers in videos[27, 28]. Gouidis et al.[10] propose an approach for the task of zero-shot state classification, which combines the knowledge graph on object states and visual information that relates the appearance of certain objects to their states. ChopNLearn[24] aims to learn compositional generalization, the ability to recognize and generate unseen compositions of objects in different states. [32] aims to grounding the active objects, and proposes a prompting pipeline to extract knowledge for objects undergoing state change. Unlike these methods for object state grounding[32], recognition or generation[10, 19, 20, 24], the AOD task aims to detect/locate objects undergoing state changes. Locating active objects not only requires semantic and visual priors but also relies on spatial prior to provide clear hints to distinguish between active objects and other distractors. In this paper, we aim to construct an oracle query for the detector and the query contains semantic, visual, and spatial priors.

2.3. Knowledge Distillation in Object Detection

Traditional object detection entails the task of identifying and localizing all objects within an image or video frame, encompassing the simultaneous duties of classification and spatial localization. Recently, the successful application of knowledge distillation in traditional object detection has garnered attention[2, 16, 30, 33, 36]. In the pursuit of compact and efficient object detection networks, [3] have seamlessly integrated knowledge distillation, achieving heightened efficiency with minimal accuracy trade-offs. Additionally, [14] proposes an ingenious approach that incorporates instance annotations into an attention mechanism, effectively pinpointing significant regions. FGD[34] proposes a focal and global knowledge distillation to separate the objects and background and rebuild the relation between different pixels from teachers to students. While knowledge distillation is effective in traditional object detection, its use in active object detection is novel. Attention or output-based distillation methods [14, 34] in traditional object detection can provide certain spatial knowledge to distinguish between objects and backgrounds. However, AOD is more difficult, such as the subtle difference between state change or not. It is inadequate to only provide spatial prior knowledge and distillation. Therefore, our method also integrates semantic interaction and fine-grained visual priors of active objects.

3. Method

3.1. Overview of the pipeline

As depicted in Figure 2, our newly devised Knowledge Aggregation and Distillation (KAD) framework encompasses two distinct detectors: the **Vision Based Detector** (emphasized in orange) and the **Knowledge-Enhanced Detector** (shaded in green). Moreover, we use **Knowledge Transfer** to enhance the detection capability of Vision Based Detector.

Vision Based Detector: The detector is based on Transformer architecture, following [1, 35]. To extract the features of the frame/image, we adopt a visual backbone and a Transformer-based encoder to extract the feature map and encode the map. Then, a Transformer-based decoder is introduced to predict a set of candidates of active objects and model their relation. Finally, we use an active object detection head based on feed-forward networks to predict the normalized box of the active object and the confidence score of predictions.

Knowledge-Enhanced Detector: To enrich the detection process with pertinent priors linked to active objects, we establish the Knowledge Aggregator, a component responsible for collecting the semantic-aware, visual-assisted and spatial-sensitive knowledge aligned with the active object's category (detailed in section 3.3). And we use the encoded feature from vision based detector as input, and introduce a Transformer decoder and an active object detection head (both the components shared with vision based detector) to predict the normalized box of active object and the confidence score of prediction.

Knowledge Transfer: Due to the unknown active object in the input image during testing, it is also difficult to obtain the relevant prior knowledge. To enhance the ability of active object detection for vision-based detector, we share the parameters of the decoders and detection heads of the above two detectors, and propose a knowledge distillation strategy to transfer the knowledge from knowledge-enhanced detector to vision-based detector by aligning the attention and immediate outputs.

3.2. Vision Based Detector

In accordance with the conventional Transformer-based Detectors paradigm [1, 35], we use a visual backbone in tandem with a Transformer-based encoder to encode region features $\mathbf{E} \in \mathbb{R}^{H\times W\times d}$ from the frame $\mathbf{I}$. In these expressions, $H \times W$ shows the spatial resolutions of the encoded features. The parameters d represents the dimension of the encoded features. Subsequently, a decoder furnished with a collection of learnable queries $\mathbf{Q}_s \in \mathbb{R}^{m\times d}$ is introduced to forecast potential active object candidates, denoted as student candidates $\mathbf{O}_s \in \mathbb{R}^{m\times d}$. In this context, m corresponds to the number of learnable queries, which is equal to the number of student candidates. Concluding this process, an active object detection head (depicted in yellow) outputs both the class-score $\hat{s} \in [0, 1]$ (representing the confidence of active object detection) and the box location $\hat{b} \in \mathbb{R}^{m\times 4}$ for the student candidates. To optimize the final predictions $(\hat{s}, \hat{b})$, we use the bipartite matching following [1] to find the lowest-cost pair between ground-truth (s, b) (where $s = 1$ due to only taking one active object annotation without inactive object annotations) and m predictions. Assume the i-th prediction $(\hat{s}_i, \hat{b}_i)$ is corresponding to lowest-cost $\mathcal{L}_{match}(i) = -\hat{s}_i + \lambda(\mathcal{L}_{giou}(b, \hat{b}_i) + ||b - \hat{b}_i||_1)$, where $\mathcal{L}_{giou}$ is the Generalized IOU Loss and $||\cdot, \cdot||_1$ is the L1 distance, and λ is the balanced parameter. Then we try to optimize the i-th confidence score and box with the objective:

$$\mathcal{L}_v = BCE(s, \hat{s}_i) + \lambda(\mathcal{L}_{giou}(b, \hat{b}_i) + ||b - \hat{b}_i||_1), \quad (1)$$

where BCE is the binary cross-entropy loss.

3.3. Knowledge-Enhanced Detector

As discussed in sections 1 and 3.1, active object detection should consider the subtle appearance changes (*vision*), diverse possible interactions(*semantic*), and distractions by other objects(*spatial*). However, conventional vision-based detectors[1, 9, 25, 31] overlook the significance of such common sense. Consequently, they lack the robustness required for accurate active object detection. Recognizing the potential of aggregating knowledge connected to active objects to

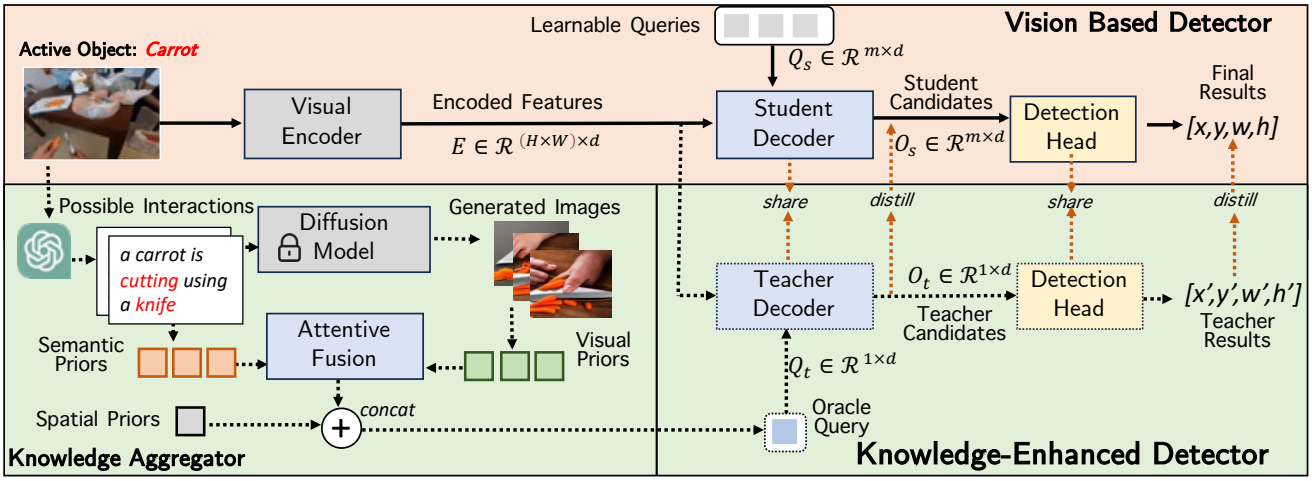


Figure 2. **Proposed Architecture: Knowledge Aggregation and Distillation (KAD)**. Our KAD architecture comprises two distinct detectors: the Vision-Based Detector (highlighted in orange, detailed in Section 3.1) and the Knowledge-Enhanced Detector (emphasized in green, elaborated in Section 3.3). Knowledge and concepts related to active object categories are systematically gathered and consolidated within the Knowledge Aggregator (shown in gray and positioned at the lower left, discussed in Section 3.3.2). Best view in color.

furnish essential priors and cues, we introduce the **Knowledge Aggregator** in this section. Its primary role is to gather and fuse visual-assited, semantic-aware and spatial-sensitive knowledge related to active objects, thereby enriching the training phase with active-relevant priors.

Conversely, during the reasoning phase, predicting the category of active objects becomes inherently challenging, as they cannot always be effectively represented by the aforementioned related concepts. To address this, we present a **Knowledge Distillation** strategy. This approach compels vision-based detectors to align their intermediate outputs and attention mechanisms with those of knowledge-enhanced counterparts. By distillation, the student imitates the teacher's ability to detect AOD with priors and avoids the extra commonsense input at inference.

3.3.1 Knowledge Aggregator

In order to overcome the challenge of AOD, the subtle appearance difference and large intra-class variance, we aim to aggregate the object affordance common sense knowledge and incorporating it into the AOD. Specifically, we construct triple complementary priors: semantic interactions priors to capture multiple possibilities that may cause changes in object state, fine-grained visual priors to provide vivid images depicting state changes, and spatial priors to guided the model to distinguish where to pay more attention.

Semantic Interaction Priors. To build an oracle query that includes active object related interactions, we use GPT to generate multiple descriptions of the scene where an object may be undergoing state change, for example: ***carrot** is cutting using a knife*. These describe concepts such as related objects and actions involved in the state changes of objects, and include relevant knowledge of active objects. We use language encoder to extract features of these descriptions $[\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_p]$ to provide possible semantic priors to build oracle query, where p is the number of descriptions.

Fine-grained Visual Priors. Semantic description of interaction is still abstract for vision-based detector. More directly, images provide more vivid visual information of related objects and interaction to better identify active objects. Therefore, we use the interaction description of the active object as prompt, and then use the Diffusion Model to generate the corresponding image. Similarly, we also extract the corresponding features from these images as a candidate set of visual concepts for the oracle query: $[\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_q]$, where q is the number of images.

In order to select more important information from these text concepts and visual concepts, we propose an Attentive Fusion module to selectively aggregate these concepts. Specifically, taking text concepts as an example, we use the self-attention layer and max-pooling to select relatively important information:

$$\mathbf{T} = pool(selfattn([(\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_p)])) \in \mathbb{R}^{1\times d_t}. \quad (2)$$

Similarly, we can also obtain selectively aggregated visual priors: $\mathbf{V} \in \mathbb{R}^{1\times d_v}$. d_t and d_v are dimensions of the fused semantic prior and visual prior, respectively.

Spatial Priors. AOD is also limited by the influence of other no change objects in the image, especially the objects with the same category of active object, which can easily distract attention in spatial. Active object is usually unique in an image. It is also necessary to provide accurate spatial location for AOD to heighten attention to enhance the model's

spatial sensitivity. We utilize the ground truth bounding box of the active object as spatial priors.

Finally, we merge text concepts, visual concepts and active object locations in the input image as the final oracle query, which has provided rich active object clues: $Q_t = [\mathbf{T}; \mathbf{V}; b] \in \mathbb{R}^{1\times d}$, where $b \in \mathbb{R}^{1\times 4}$ is the bounding box(spatial priors).

3.3.2 Knowledge-Enhanced Detector

We strive to formulate an oracle query for active objects that encapsulates not only the associated knowledge but also the category and the normalized ground-truth bounding box of the active object. By harmonizing these three types of embeddings, we craft a comprehensive query denoted as $\mathbf{Q}_t \in \mathbb{R}^{1\times d}$. Subsequently, the teacher decoder interfaces with the encoded features from the Vision-Based Detector, utilizing the amalgamated oracle query to provide critical semantic priors and precise positional data that serve as key indicators for detection. The resulting outputs from this teacher decoder, termed the Teacher Candidates $\mathbf{O}_t \in \mathbb{R}^{1\times d}$, are depicted in Figure 2. Ultimately, this enhanced detector also employs detection heads to forecast the active object's presence, leveraging the insights gleaned from the oracle query $\mathbf{Q}_t$ The optimization of the detector is similar to L_v(Eq.1):

$$\mathcal{L}_k = BCE(s, \hat{s}_t) + \lambda(\mathcal{L}_{giou}(b, \hat{b}_t) + ||b - \hat{b}_t||_1), \quad (3)$$

where $(\hat{s}_t \in \mathbb{R}^1, \hat{b}_t \in \mathbb{R}^4)$ are the final active object result (confidence score and box corresponding to oracle query).

3.4. Knowledge Distillation between Detectors

The oracle query $\mathbf{Q}_t$ integrates the three-fold informed priors for AOD: semantic interaction priors, fine-grained visual priors and spatial priors. Compared to Vision-Based Detector(student) using learnable queries, Knowledge-Enhanced Detector(teacher) using the oracle query can use the above priors to more accurately locate active objects. Due to the unknown active object in the input image during testing, it is also difficult to obtain the relevant prior knowledge. We need to transfer the knowledge of the teacher to the student to avoid extra commonsense input and make student work well in alone at inference.

Parameter Sharing. In order to achieve knowledge transfer from teacher to student, we share the parameters of decoders and detection heads between teacher and student.

Knowledge Distillation. Furthermore, we adopt a distillation strategy from teacher to student, allowing the student to mimic the output and attention of the teacher. Specifically, Knowledge-Enhanced Detector takes encoded features E as inputs and incorporates the oracle query to predict and learn the representation O_t associated with an active object. To augment the aptitude of the Vision-Based Detector for active object detection, we introduce a knowledge distillation mechanism between the predictions $\mathbf{O}_t$ of the Knowledge-Enhanced Detector (teacher) and $\mathbf{O}_s$ of the Vision-Based Detector (student).

Specifically, we align the cross attention $A_{s,i}$ and decoder embedding $O_{s,i}$ for i-th prediction of student network with those (A_t and O_t) of teacher via distillation:

$$\begin{aligned} L_{attn} &= \sum^{l} \mathrm{KL}(A_t^l, A_{s_i}^l), \\ L_{emb} &= \sum^{l} \left(1 - \frac{O_t^{l\,T} O_{s_i}^l}{\|O_t^l\|_2 \|O_{s_i}^l\|_2}\right). \end{aligned} \quad (4)$$

Here, i corresponds to the index associated with the lowest-cost in the bipartite matching of the Vision-Based Detector. The parameter l indicates the l-th layer in the decoders, which are shared between the two detectors. For a given l-th decoder layer, our approach involves aligning the cross attention pertaining to the i-th prediction in the student decoder, denoted as $A_{s_i}^l$, with its counterpart from the teacher, denoted as A_t^l, through a Kullback-Leibler divergence loss (KL). Additionally, we align the intermediate embeddings of the two networks using a cosine similarity loss. Aligning the embeddings forces the student to mimic the teacher's ability to express active objects. And attention can enable student to learn the teacher's ability where to pay attention to active objects in spatial.

The overall distillation loss is a combination of these two components, modulated by a hyper-parameter η, to achieve balance: $L_{distill} = L_{emb} + \eta L_{attn}$.

In essence, the strategy we employ begins by leveraging the oracle query to facilitate accurate representation learning in the teacher network, which helps address the challenges posed by dynamic distractors. Subsequently, we synchronize the intermediate outputs of the student network with those of the teacher network via distillation. This approach allows the student network to emulate the teacher network's ability to navigate dynamic distractors adeptly and to acquire the robust representation skills it possesses.

3.5. Training and Inference

Objective Functions The final objective function is as follows:

$$L = L_v + L_k + \alpha L_{distill}\,, \quad (5)$$

Training and Inference During the training phase, the student detector's representation and attention mechanisms are structured to mimic those garnered from the teacher detector, thereby steering the student network via a distillation loss. This approach facilitates the transfer of valuable insights and knowledge from the teacher to the student, enhancing the student detector's aptitude.

Upon transitioning to the inference stage, the teacher detector is no longer in play. This strategic abandonment of the

Table 1. Comparisons with other methods on Ego4D. We bold the best results and underline the second best ones.

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	<u>15.5</u>	<u>32.8</u>	<u>13.0</u>
KAD(ours)	ResNet-50	**31.4**	**34.6**	**28.9**
InternVideo[31]	Uniformer-L	24.8	44.2	24.0
	Swin-L	<u>36.4</u>	<u>56.5</u>	<u>37.6</u>
KAD(ours)	Swin-L	**40.5**	**60.6**	**41.9**

Table 2. Comparisons with other methods on Epic-Kitchens. We bold the best results and underline the second best ones.

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

teacher model ensures that no supplementary computational overhead is incurred.

4. Experiment

4.1. Dataset

Ego4D [11] stands as one of the latest expansive egocentric video datasets. We focus on subsets of this dataset for our state-change object detection (SCOD) tasks. The original train and validation sets encompass 19,070 and 12,800 annotated frames, respectively, marking the point of no return, or the initiation of a state change.

Epic-Kitchens [4] is a large-scale dataset in the domain of egocentric vision. We convert the segmentation annotations of action-related objects within the VISOR subset [5] into bounding boxes, specifically tailored for the active object detection task. Notably, we employ a total of 67,217 and 9,668 annotated frames for our train and validation splits.

MECCANO [22] is an egocentric dataset for human-object interaction understanding in industrial-like settings. And it has been acquired in an industrial-like scenario in which subjects built a toy model of a motorbike. It contains 64,349 frames which are annotated with active object boxes. Following prior splits[8], the training set, validation set, and test set contain 21686, 4270 and 12111 images, respectively.

100DOH [25] is a large-scale benchmark for hand-object interaction. It has 99,899 frames (79,921 for training, 9,995 for validation and 9,983 for testing). The focus of the dataset is hand contact, and it includes both first-person and third-person perspectives.

4.2. Implementation Details

The embeddings of semantic and visual features are extracted through CLIP[21]. During training, we utilize AdamW optimization, setting the initial learning rate of the Transformer to 10^{-4}, and the learning rate for the backbone to 10^{-5}. The hyperparameters α, λ, and η are configured to 0.2, 5.0, and 1.0, respectively, to govern the optimization process. Our models are trained over 50 epochs using a cosine annealing strategy with warm restarts. The training process is executed across NVIDIA A100, and employs a batch size of 4.

4.3. Comparisons Results

Comparisons on Ego4D and Epic-Kitchen We present the COCO-style Average Precision (AP), AP50, and AP75 results achieved by our method on the Ego4D validation and test sets. The comparative outcomes for Ego4D are meticulously tabulated in Table 1. (1) Our KAD method attains state-of-the-art performance across all metrics under a fair comparison protocol (with two backbones). This indicates the generalization of our model on different backbones. (2) From the Table 1, it can be seen that the transformer based method can achieve higher performance. In particular, KAD outperforms the best existing method[31] with the same transformer-based backbone by 4.1%, 4.1%, and 4.3% on AP, AP50, and AP75, respectively. This proves the effectiveness of priors for active object detection. Notably, our method dose not introduce extra priors during testing(traditional detectors[1, 31]), but instead forced traditional detectors to learn the AOD capabilities brought by priors through parameter sharing and knowledge distillation.

Table 2 presents a comprehensive comparison between

Table 3. Comparisons with other methods on MECCANO. We bold the best results and underline the second best ones.

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

Table 4. Comparisons with other methods on 100DOH. We bold the best results and underline the second best ones.

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	<u>57.8</u>
Seq-Voting[9]	ResNet-101	<u>29.9</u>	<u>53.0</u>	57.2
KAD(ours)	ResNet-101	**31.2**	**53.9**	**58.9**

our KAD approach and other methods on the Epic-Kitchens dataset[4]. To gauge the performance, we employed the DETR[1] and InternVideo[31] methods on the Epic-Kitchens dataset, retraining and evaluating their outcomes using the corresponding configurations (by 6.9%, 4.3%, and 5.3% on AP, AP50, and AP75 compared to best baseline method[31]). Importantly, our KAD method outperforms other approaches across all metrics. This improvement of performance underscores the effectiveness of our proposed KAD framework and its generalization ability on different datasets.

Comparisons on MECCANO and 100DOH Due to the lack of object categories provided in 100DOH and fine-grained toy components in MECCANO[1], it is difficult to generate semantic aware interaction descriptions and visual assigned images. Therefore, we only use the spatial location of the active object as the oracle query. (1)From Tables 3 and 4, it can be seen that although our method only includes spatial cues, it still improves the AOD detection performance without leveraging any external knowledge, which indicates the effectiveness of our network design. On the MECCANO dataset [22], our method outperforms the previous best method[8] in terms of AP75, AP50 and AP25 improved by 1.4%,2.5% and 1.3% respectively. On 100DOH [25], our method improves the detection performance by 1.3%, 0.9%, and 1.7%, respectively.It can be seen that spatial cues help the model focus on active objects and improve detection performance. (2) These two tables contain comparisons with more AOD specific methods which considering the interactiveness. And compared to them, our improvement indicates that only hand or visual information is not sufficient, as well as the effectiveness on spatial priors.

4.4. Ablation Study

In this part, we evaluate the effectiveness of different modules or variants of our KAD on the Ego4D validation set.

Different knowledge aggregation variants. Our exploration into the influence of various knowledge aggregation approaches on active object detection has shed light on nuanced improvements, as shown in Table5. Beginning with the baseline, represented in the first row, which exclusively employs the Vision-Based Detector. The introduction of the Knowledge-Enhanced Detector combined with an oracle query containing solely the ground-truth normalized box of the active object showcased in the second row (VBD+spatial), demonstrates an initial boost in performance (on AP +0.2% improvement). This indicates that leveraging positional information of active object is indeed beneficial and provides valuable cues for detection. We delve deeper into enriching the knowledge aggregation process. In the third row(VBD+spatial+semantic), we take a significant step forward by incorporating semantic features of possible interactions. The outcome is a further improvement in detection performance(on AP +1.8% improvement). This noteworthy progress underscores the pivotal role of semantic information tied to active objects. Furthermore, we provide more direct image information for these interactions as part of the oracle query (results in the last row, VBD+spatial+semantic+visual). The visual features provide the best performance(on AP +2.6% improvement). The comparisons show the necessity of triple knowledge: visual-assited,semantic-aware and spatial-sensitive.

Table 5. Different knowledge aggregations on Ego4D.

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**

Ablation of different knowledge distillation strategy. Additionally, our investigation extended to the realm of different distillation techniques and their influence on detection

[1]i.e., *gray angled perforated bar*.

Table 6. Ablation of knowledge distillation on Ego4D.

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**

outcomes in Table 6. Similar to our earlier explorations, the baseline model(VBD, Vision-Based Detector) is initiated without the incorporation of knowledge aggregation and distillation. Subsequently, we introduced the Knowledge-Enhanced Detector as an extension of the baseline, with distillation exclusively applied to features. The outcomes reveal that feature-level distillation yields a discernible performance boost with an average enhancement of 2.4% on AP. This underscores the 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). Building upon this foundation, the introduction of distillation across attentions imparts a substantial augmentation to the model's proficiency, underscoring the synergistic benefits of comprehensive distillation techniques with an average enhancement of 2.2% on AP. We introduce distillation across attention mechanisms in addition to feature distillation. The outcomes of this approach yielded significant advancements in the model's proficiency. This underscores the synergistic potential of comprehensive distillation strategies that not only align features but also bridge the gap between attentions. By orchestrating the transfer of intermediate outputs and attention maps from the teacher to the student, our distillation scheme enables the Vision-Based Detector to harness the enhanced knowledge of the Knowledge-Enhanced Detector.

Ablation of the number of generated descriptions. We validate the performance of the model using different numbers of semantic descriptions. We generated 10 interaction descriptions of a state-change object using GPT, with the prompt "describe 10 interaction descriptions of *[object]* undergoing state change (including tools)". From Table 7, it can be seen that diverse descriptions bring significant improvements to model performance(when using 10 text descriptions, detection performance improved by 3.2% on AP). This may be due to the fact that the scene of a state change of an object may be diverse, so diverse descriptions are necessary.

Ablation of the number of generated images . We also validate the impact of using different numbers of generated images on model performance, as shown in Table 8. It can be seen that as the number of generated images increases, the performance of the model also increases. Compared to not using images, when each object uses 1 generated image (randomly selecting a text description to generate one image), the model improves by 0.2% on AP. When the number increased to 10 (10 descriptions for each object and each text description generated 1 image), the model improved by 1.6% on AP. At 100 images (each text description generated 10 images), the model performance provided 2.6% on AP. This indicates that diverse visual features can provide more performance improvements to the model.

Table 7. Different number of generated interaction descriptions.

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**

Table 8. Different number of generated images.

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**

Table 9. Different aggregation approaches.

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**

Different aggregation approaches. In Table 9, we validate the impact of aggregation methods(attentive indicates use the way described in Eq.2) on model performance. In the first two rows, we directly perform max- or average-pooling on semantic features or visual features without any attention operation. It can be seen that the performance of maximum pooling is relatively high(+0.1% on AP). Furthermore, we first perform a self attention operation on the features and then max-pooling (the third line, attentive). Attentive operation has brought about 1.3% improvement on AP, which shows adaptive selection contributes to AOD.

5. Conclusion

We aim to address the challenges inherent in active object detection by knowledge aggregation and distillation. We propose a framework that significantly improves the accuracy and efficiency of active object detection. Our proposed Knowledge Aggregator aggregates three-fold commonsense pertaining to active objects, encompassing plausible semantic interactions, fine-grained visual and spatial priors. Furthermore, our Knowledge Distillation strategy empowers the traditional detector with the capability for localizing active objects without extra prior inputs. The results of comprehensive experiments on Ego4D, Epic-Kitchens, 100DOH and MECCANO, demonstrate the efficacy of our method.

Acknowledgements. This work was supported by the grants from the National Natural Science Foundation of China 62372014.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1, 3, 6, 7

[2] Jiahao Chang, Shuo Wang, Hai-Ming Xu, Zehui Chen, Chenhongyi Yang, and Feng Zhao. Detrdistill: A universal knowledge distillation framework for detr-families—supplementary material—. 3

[3] Guobin Chen, Wongun Choi, Xiang Yu, Tony Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. *Advances in neural information processing systems*, 30, 2017. 3

[4] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *IJCV*, 130:33–55, 2022. 6, 7

[5] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*, 2022. 6

[6] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6569–6578, 2019. 1

[7] Matteo Dunnhofer, Antonino Furnari, Giovanni Maria Farinella, and Christian Micheloni. Visual object tracking in first person vision. *International Journal of Computer Vision*, 131(1):259–283, 2023. 2

[8] Qichen Fu, Xingyu Liu, and Kris Kitani. Sequential voting with relational box fields for active object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2374–2383, 2022. 1, 6, 7

[9] Qichen Fu, Xingyu Liu, and Kris M Kitani. Sequential decision-making for active object detection from hand. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 2, 3, 7

[10] Filipos Gouidis, Theodore Patkos, Antonis Argyros, and Dimitris Plexousakis. Leveraging knowledge graphs for zero-shot object-agnostic state classification. *arXiv preprint arXiv:2307.12179*, 2023. 2

[11] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, pages 18995–19012, 2022. 6

[12] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1

[13] Xiaoning Han, Huaping Liu, Fuchun Sun, and Xinyu Zhang. Active object detection with multistep action prediction using deep q-network. *IEEE Transactions on Industrial Informatics*, 15(6):3723–3731, 2019. 1

[14] Zijian Kang, Peizhen Zhang, Xiangyu Zhang, Jian Sun, and Nanning Zheng. Instance-conditional knowledge distillation for object detection. *Advances in Neural Information Processing Systems*, 34:16468–16480, 2021. 3

[15] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim. Hotr: End-to-end human-object interaction detection with transformers. In *CVPR*. IEEE, 2021. 2, 7

[16] Zhihui Li, Pengfei Xu, Xiaojun Chang, Luyao Yang, Yuanyuan Zhang, Lina Yao, and Xiaojiang Chen. When object detection meets knowledge distillation: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 3

[17] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. Ppdm: Parallel point detection and matching for real-time human-object interaction detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 482–490, 2020. 7

[18] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[19] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1792–1801, 2017. 2

[20] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 169–185, 2018. 2

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 6

[22] Francesco Ragusa, Antonino Furnari, and Giovanni Maria Farinella. Meccano: A multimodal egocentric dataset for humans behavior understanding in the industrial-like domain. *Computer Vision and Image Understanding (CVIU)*, 2023. 6, 7

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1, 6

[24] Nirat Saini, Hanyu Wang, Archana Swaminathan, Vinoj Jayasundara, Bo He, Kamal Gupta, and Abhinav Shrivastava. Chop & learn: Recognizing and generating object-state compositions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20247–20258, 2023. 2

[25] Dandan Shan, Jiaqi Geng, Michelle Shu, and David Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, 2020. 3, 6, 7

[26] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9869–9878, 2020. 1, 2

[27] Tomáš Souček, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic. Look for the change: Learning object states and state-modifying actions from untrimmed web videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13956–13966, 2022. 2

[28] Tomáš Souček, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic. Multi-task learning of object state changes from uncurated videos. *arXiv preprint arXiv:2211.13500*, 2022. 2

[29] Sanket Thakur, Cigdem Beyan, Pietro Morerio, Vittorio Murino, and Alessio Del Bue. Anticipating next active objects for egocentric videos. *arXiv preprint arXiv:2302.06358*, 2023. 2

[30] Tao Wang, Li Yuan, Xiaopeng Zhang, and Jiashi Feng. Distilling object detectors with fine-grained feature imitation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4933–4942, 2019. 3

[31] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 2, 3, 6, 7

[32] Te-Lin Wu, Yu Zhou, and Nanyun Peng. Localizing active objects from egocentric vision with symbolic world knowledge. *arXiv preprint arXiv:2310.15066*, 2023. 2

[33] Chenhongyi Yang, Mateusz Ochal, Amos Storkey, and Elliot J Crowley. Prediction-guided distillation for dense object detection. In *European Conference on Computer Vision*, pages 123–138. Springer, 2022. 3

[34] Zhendong Yang, Zhe Li, Xiaohu Jiang, Yuan Gong, Zehuan Yuan, Danpei Zhao, and Chun Yuan. Focal and global knowledge distillation for detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4643–4652, 2022. 3

[35] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 2, 3

[36] Linfeng Zhang and Kaisheng Ma. Improve object detection with feature-based knowledge distillation: Towards accurate and efficient detectors. In *International Conference on Learning Representations*, 2020. 3

[37] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019. 6