

What If We Could Not See?

Counterfactual Analysis for Egocentric Action Anticipation

Tianyu Zhang^{1,2}, Weiqing Min^{1,2}, Jiahao Yang^{1,2}, Tao Liu^{3,1}, Shuqiang Jiang^{1,2}, Yong Rui⁴

¹Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, CAS, Beijing, 100190, China

²University of Chinese Academy of Sciences, Beijing, 100049, China

³Shandong University, Jinan, 250101, China

⁴Lenovo Group, Beijing, 100085, China
tianyu.zhang@vipl.ict.ac.cn

Abstract

Egocentric action anticipation aims at predicting the near future based on past observation in first-person vision. While future actions may be wrongly predicted due to the dataset bias, we present a counterfactual analysis framework for egocentric action anticipation (CA-EAA) to enhance the capacity. In the factual case, we can predict the upcoming action based on visual features and semantic labels from past observation. Imagining one counterfactual situation where no visual representation had been observed, we would obtain a counterfactual predicted action only using past semantic labels. In this way, we can reduce the side-effect caused by semantic labels via a comparison between factual and counterfactual outcomes, which moves a step towards unbiased prediction for egocentric action anticipation. We conduct experiments on two large-scale egocentric video datasets. Qualitative and quantitative results validate the effectiveness of our proposed CA-EAA.

1 Introduction

Forecasting the near future is crucial for both humans and intelligent systems. With the development of wearable cameras, egocentric (first-person) vision offers an interesting scenario to study the action anticipation problem [Furnari and Farinella, 2019], which is essential for real-world applications. For example, the ability to predict what action the camera wearer is going to perform based on observation from the past is critical for intelligent wearable systems to understand the user’s goal and provide seamless assistance [Kanade and Hebert, 2012]. In general, this is a challenging task which requires understanding observed actions and making assumptions about unobserved but upcoming actions.

For the problem of action anticipation, what can be exploited from observed information is essential for the anticipation result of target action. In addition to visual features which capture spatial and temporal representations, semantic labels act as high-level abstraction about what has happened [Miech *et al.*, 2019; Sener *et al.*, 2020]. While the distribution

of egocentric actions is imbalanced, future actions may be wrongly predicted due to the dataset bias. A typical example is illustrated in Figure 1 (a). After observing the action “take onion”, models tend to wrongly predict the future action as “cut onion” instead of the ground-truth action “peel onion”. Considering two consecutive actions as a pair, the incorrect prediction can be attributed to the fact that pair (“take onion”, “cut onion”) has a higher frequency than (“take onion”, “peel onion”). Similar examples are widespread, which indicate predictions of future actions could be misled by dataset bias.

To tackle this issue, we present a causal view of the relationship between observed past action and predicted future action. Since egocentric action anticipation is a vision-based task, we consider that visual representation of past observation has a main causal effect on predicting the future action. On the other hand, the semantic label of observed action may bring side-effect to the prediction result because models tend to learn from the statistical shortcut linking two consecutive actions. Therefore, semantic labels should be properly used to prevent models from neglecting visual representation.

In this paper, we propose a Counterfactual Analysis framework for Egocentric Action Anticipation (CA-EAA). As shown in Figure 1 (b), the pipeline of CA-EAA consists of three stages. In the first stage, we still predict the upcoming action based on visual features with semantic labels as auxiliary information. We consider the predicted future action as a factual outcome where “cut onion” is the first candidate and the probability of ground-truth “peel onion” is smaller than “cut onion”. In the second stage of CA-EAA, we can imagine a counterfactual situation: “*what action would be predicted if we had not observed any visual representation?*” If we had not observed the action “take onion”, we would predict the upcoming action only using the semantic label of past observation, which is considered as a counterfactual outcome. The counterfactual outcome is obtained by purely exploiting statistical correlations. In the counterfactual outcome, “cut onion” still ranks first while the gap between “cut onion” and “peel onion” is larger than that in the factual outcome, which reflects the side-effect that past semantic labels have on predicting future actions.

According to the effect analysis in causal inference [VanderWeele, 2013; Pearl and Mackenzie, 2018], we can mitigate

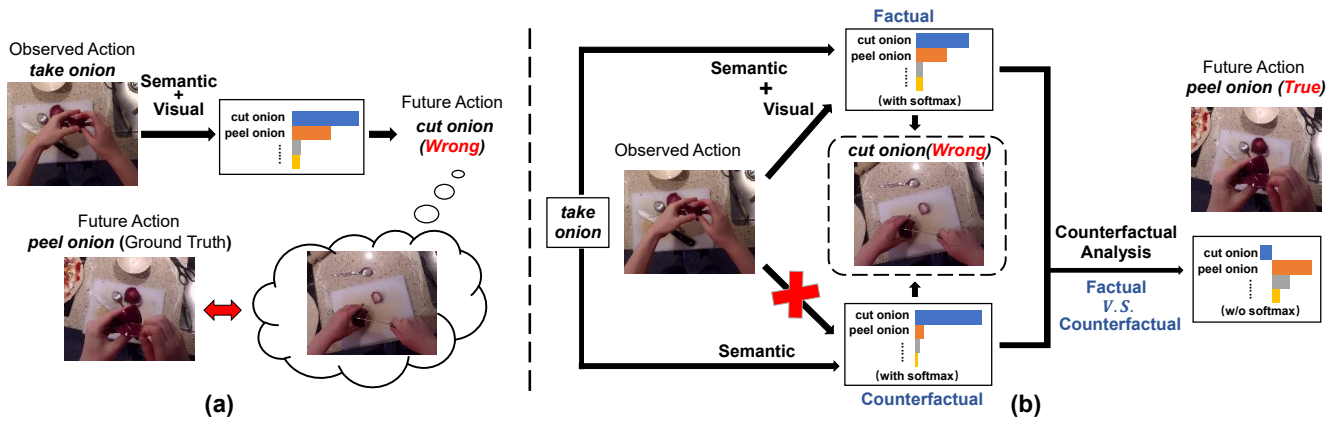


Figure 1: An example showing the comparison of (a) biased predictions and (b) our proposed counterfactual analysis for egocentric action anticipation (CA-EAA). (a) The future action may be wrongly predicted due to the dataset bias. (b) With the help of counterfactual analysis, we can mitigate biased predictions by reducing the side-effect of past semantic labels on predicting future actions (indicated by the red cross).

the side-effect caused by semantic labels in the third stage of CA-EAA. Specifically, the comparison between factual and counterfactual outcomes will decrease the probability of predicting the future action as “cut onion” and increase the chance of anticipating the ground-truth action “peel onion”. This strategy captures the main causal effect of visual features while reducing the side-effect of semantic labels on predicting future actions, which moves a step towards unbiased prediction. To verify the effectiveness of our approach, we conduct experiments on two large-scale egocentric datasets EPIC-Kitchens [Damen *et al.*, 2018] and EGTEA Gaze+[Li *et al.*, 2018]. Qualitative and quantitative results demonstrate the effectiveness and superiority of our proposed CA-EAA.

To summarize, our main contributions are as follows:

- Predicting the future and counterfactual thinking are both innate abilities of humans. To our best knowledge, we are the first to integrate them, which moves closer to human intelligence.
- We propose a counterfactual analysis framework for egocentric action anticipation which captures the main causal effect of visual features and reduces the side-effect of semantic labels on predicting future actions.
- Experimental results on two large-scale egocentric datasets verify the effectiveness of our approach.

2 Related Work

2.1 Egocentric Action Anticipation

Action anticipation aims at predicting an action before it happens. Compared to the widely studied action recognition problem [Simonyan and Zisserman, 2014; Wang *et al.*, 2016; Carreira *et al.*, 2018], action anticipation needs not only to understand what has happened, but also to predict what will happen next. Despite its challenging nature, egocentric action anticipation is critical for real-world intelligent systems such as wearable assistants, which has attracted increasing attention in recent years [Damen *et al.*, 2018; Furnari *et al.*, 2018; Furnari and Farinella, 2019; Liu *et al.*, 2020; Zhang *et al.*,

2020; Wu *et al.*, 2020]. What can be exploited from past observation includes both visual features and semantic labels [Miech *et al.*, 2019; Sener *et al.*, 2020]. In this paper, we consider that past semantic labels have a side-effect on predicting future actions by enforcing models to learn from the statistical shortcut linking two consecutive actions, which is neglected by previous works. To our best knowledge, we first provide a causal view of egocentric action anticipation and introduce a counterfactual analysis framework to capture the main causal effect of visual features on predicting future actions.

2.2 Counterfactual Analysis

Counterfactual thinking is derived from psychology, which describes the human capacity to reason the outcome of an alternative operation that could have been performed [Pearl and Mackenzie, 2018]. It has been widely studied in economics, politics and epidemiology [Chernozhukov *et al.*, 2013; King, 2008; Richiardi *et al.*, 2013] as the tool to study the effect of certain treatments or policies. Recently, counterfactual thinking has gained popularity in the computer vision community to pursue unbiased outcomes in several applications, including general long-tailed visual recognition [Tang *et al.*, 2020a], scene graph generation [Tang *et al.*, 2020b] and visual question answering [Niu *et al.*, 2020]. In our work, we take a cause-effect look at the biased prediction and propose a counterfactual analysis framework for egocentric action anticipation. The goal of proposed CA-EAA is to capture the main causal effect of visual features and mitigate the side-effect of semantic labels on predicting future actions, which makes a step towards unbiased prediction in this field.

3 Method

3.1 Preliminaries

Problem Description. According to previous work [Damen *et al.*, 2018], the egocentric action anticipation problem can be described as follows. For an action occurring at the time step $t + 1$, the task is to categorize the future action by observing a video segment before time step t . Since

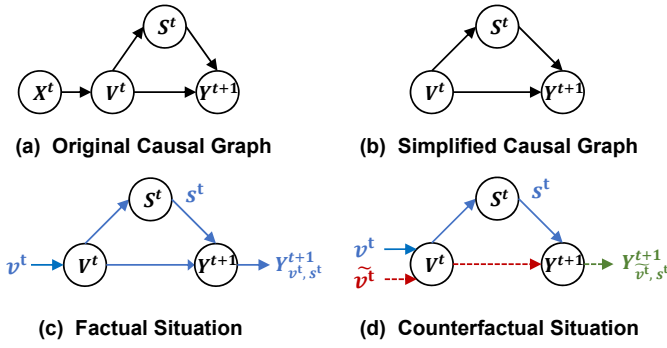


Figure 2: The causal graph for egocentric action anticipation with four elements: past observation X^t , semantic labels S^t , visual features V and future action Y^{t+1} . We show its (a) original version and (b) simplified version. To reduce the side-effect of S^t on predicting Y^{t+1} , we compare the outcomes obtained under (c) factual situation and (d) counterfactual situation.

it is a vision-based task, visual features can be extracted from past observation to directly predict the future action. In addition to visual features, semantic labels can also be used as auxiliary information to help with anticipation. Specifically, semantic labels can be obtained by applying action recognition to visual features of past observation. Similar to visual features, semantic labels can be used to predict the future action.

Causal Graph. The causal graph is expressed as a directed acyclic graph [Pearl *et al.*, 2016; Pearl and Mackenzie, 2018]. It consists of nodes and directed edges where the nodes denote variables, and the directed edges (*i.e.*, arrows) represent the cause-effects between two nodes. For example, $A \rightarrow B$ represents that there exists a direct path from A to B , which means A has direct effect on B . As a highly general roadmap, causal graph is used to analyze the causal effects among variables. Combined with causal graph, we will detail the methodology of CA-EAA.

3.2 CA-EAA

We establish the causal graph and analyze the causal effects for egocentric action anticipation. As shown in Figure 2 (a), we formulate the causalities among past observation X^t , past visual features V^t , past semantic labels S^t and future action Y^{t+1} . Specifically, visual features V^t are extracted from past observation X^t while semantic labels S^t are further obtained from V^t by action recognition. The future action Y^{t+1} is predicted by merging information from both V^t and S^t . The overall impact that X^t has on Y^{t+1} consists of two parts. One can be represented as $X^t \rightarrow V^t \rightarrow Y^{t+1}$ while the other is denoted as $X^t \rightarrow V^t \rightarrow S^t \rightarrow Y^{t+1}$. Since the single path from X^t to V^t is shared by these two parts, the simplified causal graph can be found in Figure 2 (b).

We denote a random variable as an uppercase letter (*e.g.*, V^t) and its value as a lowercase letter (*e.g.*, v^t). As illustrated in Figure 2 (c), under the factual situation, the prediction result is obtained when V^t is set to v^t and S^t takes the value s^t based on $V^t = v^t$, which can be represented as:

$$Y_{v^t, s^t}^{t+1} = Y^{t+1} (V^t = v^t, S_{V^t}^t = S^t(V^t = v^t)) \quad (1)$$

In the factual scenario, the total effect is composed of direct effect $V^t \rightarrow Y^{t+1}$ and indirect effect $V^t \rightarrow S^t \rightarrow Y^{t+1}$. Recall that for egocentric action anticipation, $V^t \rightarrow Y^{t+1}$ is the main causal path because observed visual representation contains concrete information that indicates what will happen next. However, we cannot estimate the main causal effect in the factual case because S^t acts as a mediator which brings side-effect simultaneously. Thanks to counterfactual analysis, we can capture the pure indirect effect $V^t \rightarrow S^t \rightarrow Y^{t+1}$ via imagining the counterfactual scenario, which can be found in Figure 2 (d). Under this situation, V^t is set to the empty value \tilde{v}^t while S^t still takes the value s^t as if it had seen the real v^t . The prediction result can be denoted as:

$$Y_{\tilde{v}^t, s^t}^{t+1} = Y^{t+1} (V^t = \tilde{v}^t, S_{V^t}^t = S^t(V^t = v^t)) \quad (2)$$

It is worth noting that V^t can only be simultaneously set to different values v and \tilde{v}^t under the counterfactual situation. The counterfactual outcome captures the pure indirect effect by blocking the direct path at the same time. Following the decomposition way in the literature of causal inference [Van-derWeele, 2013], we can disentangle the main causal effect as the total direct effect (TDE) by subtracting the pure indirect effect from the total effect:

$$TDE = Y_{v^t, s^t}^{t+1} - Y_{\tilde{v}^t, s^t}^{t+1} \quad (3)$$

where we keep S^t as the same value s^t . By comparing factual and counterfactual outcomes, this strategy can reduce the side-effect of S^t on Y^{t+1} and capture the main causal effect of V^t on Y^{t+1} from the overall impact that V^t has on Y^{t+1} .

3.3 Implementation

CA-EAA is a unified framework for egocentric action anticipation. Without loss of generality, we implement CA-EAA based on a temporal aggregation architecture which contains both action recognition module F_R and anticipation module F_A [Sener *et al.*, 2020]. As shown in Figure 3, there are two paths leading to the factual outcome. On one hand, we can obtain predictions directly from visual features as $Y_{v^t}^{t+1} = F_A(v^t)$ where different visual modalities are merged by late fusion. A cross-entropy loss based on ground-truth labels Y^{t+1} is applied to the anticipation scores $Y_{v^t}^{t+1}$:

$$\mathcal{L}_{A_v} = - \sum_{n=1}^N (Y^{t+1})_n (\log(Y_{v^t}^{t+1}))_n \quad (4)$$

where N is the number of training examples. On the other hand, semantic labels s^t are obtained as action recognition scores from v^t , *i.e.*, $s^t = F_R(v^t)$. A cross-entropy loss based on ground-truth of past observation Y^t is applied to s^t as the recognition loss to regulate the generation process of semantic labels:

$$\mathcal{L}_R = - \sum_{n=1}^N (Y^t)_n (\log(s^t))_n \quad (5)$$

The anticipation scores with respect to semantic labels are then obtained as $Y_{s^t}^{t+1} = F_A(s^t)$. Similar to Equation 4, a

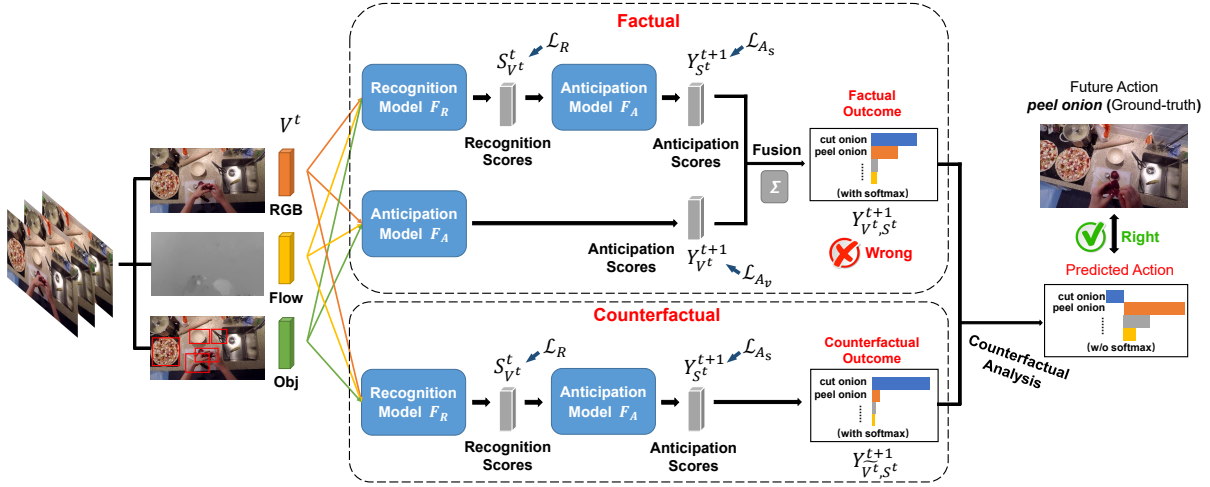


Figure 3: Overview of CA-EAA implementation. In addition to directly obtained from visual features, anticipation scores can also be obtained by applying the anticipation model to recognition scores (semantic labels of past observation). The factual outcome is the fusion of these two parts while the counterfactual outcome is obtained relying only on semantic labels. We obtain the anticipation result of CA-EAA by comparing the factual and counterfactual outcomes.

cross-entropy loss is also computed as the anticipation loss:

$$\mathcal{L}_{A_s} = - \sum_{n=1}^N (Y^{t+1})_n (\log(Y_{s^t}^{t+1}))_n \quad (6)$$

The factual outcome Y_{v^t, s^t}^{t+1} is obtained as the late fusion of $Y_{v^t}^{t+1}$ and $Y_{s^t}^{t+1}$. The overall loss is computed as:

$$\mathcal{L} = \mathcal{L}_{A_v} + \mathcal{L}_{A_s} + \mathcal{L}_R \quad (7)$$

Considering the side-effect caused by semantic labels, we introduce counterfactual analysis instead of treating Y_{v^t, s^t}^{t+1} as the final result. Under the counterfactual situation where visual features are not given, the counterfactual outcome is predicted as $Y_{v^t, s^t}^{t+1} = F_A(s^t)$ while s^t is still obtained as $s^t = F_R(v^t)$. The final result of CA-EAA is obtained as $Y_{v^t, s^t}^{t+1} - Y_{v^t, s^t}^{t+1}$, which is the comparison of factual and counterfactual outcomes. In this way, we capture the main causal effect of visual representation on predicting future actions.

4 Experiment

4.1 Datasets

EPIC-Kitchens [Damen *et al.*, 2018] is a large-scale video dataset in first-person vision. It contains 55 hours of non-scripted daily activities captured by head-mounted GoPro cameras in 32 wearers' native kitchen environments, making the dataset close to real-world data. The dataset consists of 2,513 unique action classes. Similar to [Furnari and Farinella, 2019], we split the public training set into training (23,493 action instances) and validation (4,979 action instances) sets. **EGTEA Gaze+** [Li *et al.*, 2018] is an egocentric dataset which is also recorded in a naturalistic kitchen environment with SMI wearable eye-trackers. The dataset contains 10,321 action instances annotated with 106 classes. Similar to [Liu *et al.*, 2020], we use the first split of the dataset where 8,299 action instances are used for training and 2,022 for testing.

4.2 Evaluation Metrics

We adopt both class-agnostic and class-aware metrics to evaluate our method. As for the class-agnostic metric, the standard classification accuracy is used in most works on action anticipation as the key performance measurement, which is computed over all examples. As for class-aware metrics, Average Class Precision (AvgCP) refers to averaging precision values over classes. Similarly, Average Class Recall (AvgCR) refers to averaging recall values over classes. Different from previous work where AvgCP and AvgCR are reported on many-shot classes [Damen *et al.*, 2018], we compute them over all classes to obtain a comprehensive evaluation.

4.3 Implementation Details

The time distance between X^t and Y^{t+1} is set to one second for fair comparisons. Similar to [Sener *et al.*, 2020], the time length of X^t is set to 6 seconds with the same spanning scale, recent scale and recent starting points. As for visual features, we use the RGB frame features, optical flow frame features and object-based features provided by [Furnari and Farinella, 2019] whose dimensions are 1,024, 1,024 and 352 respectively. We apply late fusion to merge information on the score level, not only for the three visual modalities but also for the fused visual features and semantic labels. The word embedding of semantic labels is set to 1024-dimension. We use the Adam optimizer to train the framework with batch size of 16. The learning rate is set to 0.0001 initially and divided by 10 every 10 epochs. We train 25 epochs and apply early stopping. To regularize the training and avoid overfitting, dropout with retain probability 0.3 is used.

4.4 Quantitative Results

Experiments on EPIC-Kitchens. We evaluate the performance on the unseen test set of EPIC-Kitchens. The unseen

Method	Accuracy	AvgCP	AvgCR
2SCNN [Damen <i>et al.</i> , 2018]	2.29	0.85	1.14
ATSN [Damen <i>et al.</i> , 2018]	2.39	0.80	1.07
MCE [Furnari <i>et al.</i> , 2018]	5.57	1.99	2.39
Trans [Miech <i>et al.</i> , 2019]	7.24	2.20	3.36
RU [Furnari and Farinella, 2019]	8.16	3.64	4.83
IAI [Zhang <i>et al.</i> , 2020]	8.57	3.33	4.56
ImagineRNN [Wu <i>et al.</i> , 2020]	9.25	3.47	5.21
MotorJoint [Liu <i>et al.</i> , 2020]	9.94	4.40	5.18
TempAgg [Sener <i>et al.</i> , 2020]	10.04	4.92	6.26
CA-EAA (Ours)	10.14	5.71	6.23

Table 1: Action anticipation results on the unseen test set of EPIC-Kitchens (%).

means kitchen scenes in the test set do not appear in the training set, for videos from 4 subjects are held out [Damen *et al.*, 2018]. The generalization capacity can be apparently reflected because the test and train distributions are distinct. As we can see from Table 1, CA-EAA outperforms the previous methods under both class-agnostic and class-aware metrics. Our method generally achieves improvement on top of TempAgg [Sener *et al.*, 2020], especially with respect to accuracy and AvgCP. Considering the challenging nature of the EPIC-Kitchens dataset and the action anticipation problem, this performance improvement is significant, which demonstrates that CA-EAA is better at anticipating future actions. The probable reason for the comparable performance under the AvgCR metric is that class-aware metrics are computed by giving equal weight to all classes, which is sensitive to the large number of complicated action categories in the dataset.

We conduct the ablation study on the validation set of EPIC-Kitchens. Table 2 shows the effectiveness of counterfactual analysis based on four visual modalities: RGB frame features (RGB), optical flow frame features (Flow), object-based features (Obj) and the late fusion of them (Fusion). The baseline for each visual modality (w/o CA-EAA) refers to factual outcomes obtained by fusing anticipation scores from both the visual modality itself and semantic labels generated from it. These results show a steady improvement by introducing counterfactual analysis with respect to all modalities. Specially, when compared to other modalities, CA-EAA significantly improves the performance based on object-based features. This is mainly because egocentric actions involve numerous object manipulations. Overall, these comparison results prove the effectiveness of counterfactual analysis.

We also evaluate the effectiveness of the recognition loss \mathcal{L}_R used in CA-EAA. Table 3 shows the comparison of results with and without recognition loss based on the Fusion modality, which is the last row in Table 2. It can be seen that the recognition loss can bring improvement of 1.08%, 0.78% and 1.15% under three metrics respectively. These performance gains show that it is necessary to regulate the recognition process from past observation, which is the generation process of semantic labels.

Experiments on EGTEA Gaze+. We also conduct experiments on the EGTEA Gaze+ dataset. Table 4 reports the results of CA-EAA and other methods: RU [Furnari and

Modality	Method	Accuracy	AvgCP	AvgCR
RGB	w/o CA-EAA	13.48	4.59	5.23
	with CA-EAA	15.07	5.93	6.07
Flow	w/o CA-EAA	9.16	2.02	2.21
	with CA-EAA	10.78	3.55	3.56
Obj	w/o CA-EAA	11.24	3.56	4.09
	with CA-EAA	13.98	5.83	5.95
Fusion	w/o CA-EAA	15.21	5.28	5.66
	with CA-EAA	17.89	6.70	7.53

Table 2: The effectiveness of counterfactual analysis based on different visual modalities on the validation set of EPIC-Kitchens (%).

Method	Accuracy	AvgCP	AvgCR
CA-EAA (w/o \mathcal{L}_R)	16.81	5.92	6.38
CA-EAA	17.89	6.70	7.53

Table 3: The effectiveness of recognition loss used in CA-EAA on the validation set of EPIC-Kitchens (%).

Farinella, 2019], IAI [Zhang *et al.*, 2020], MotorJoint [Liu *et al.*, 2020] and TempAgg [Sener *et al.*, 2020]. As we can see from Table 4, the performance of our CA-EAA in general favors that of other methods under all metrics. It is also worth noting that the values of metrics obtained on EGTEA are much higher than on EPIC-Kitchens. This is probably due to the smaller-scale of EGTEA Gaze+ containing 106 action categories while the number of action category is 2,513 in EPIC-kitchens, which means anticipation on the EPIC-Kitchens dataset is more challenging.

4.5 Qualitative Analysis and Visualization

Figure 5 illustrates some qualitative examples from EPIC-Kitchens. We show prediction results in the second and third columns to demonstrate the improvement brought by counterfactual analysis. Take the first case for example, given the observed action as “take bowl”, the future action is wrongly predicted as “put bowl” instead of the ground-truth “put down meat” due to the dataset bias. Thanks to counterfactual analysis, we can correct the prediction result by reducing the side-effect of the past semantic label on predicting the future action. It should be noted that in the last case, limited by the challenge of distinguishing between “spoon” and “fork” with similar visual features, the prediction result is still incorrect after we introduce counterfactual analysis. However, the ground-truth “take fork” ranks higher, which proves the effectiveness of CA-EAA.

We further visualize the performance difference on each action class of EPIC-Kitchens in Figure 4. It can be seen

Method	Accuracy	AvgCP	AvgCR
RU [Furnari and Farinella, 2019]	32.74	20.12	22.35
IAI [Zhang <i>et al.</i> , 2020]	33.65	21.23	21.80
MotorJoint [Liu <i>et al.</i> , 2020]	36.89	25.34	26.68
TempAgg [Sener <i>et al.</i> , 2020]	36.36	25.75	27.11
CA-EAA (Ours)	37.67	27.10	27.64

Table 4: Action anticipation results on EGTEA Gaze+ (%).

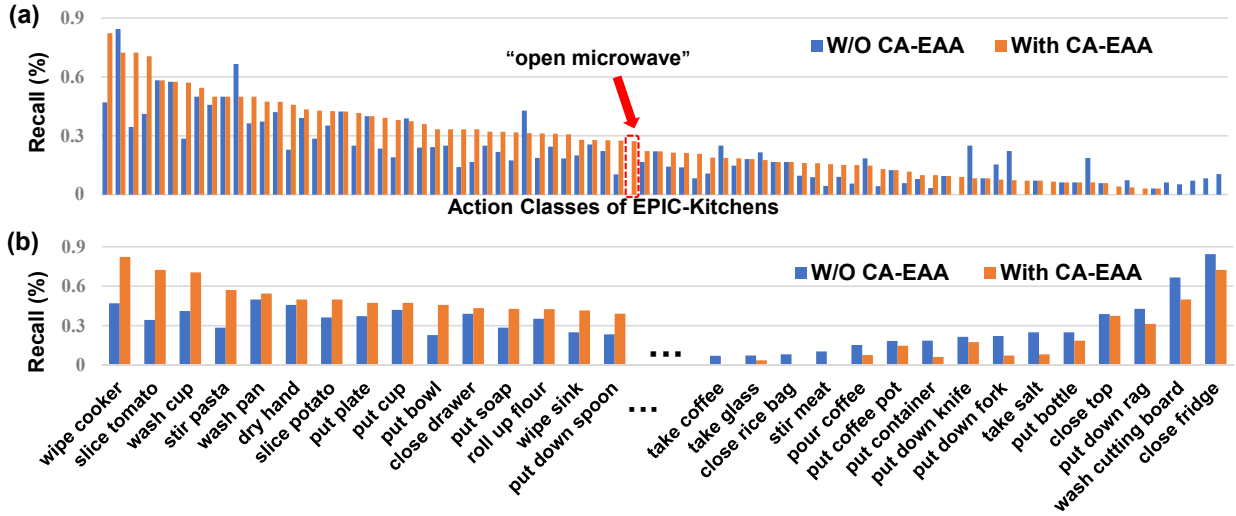


Figure 4: Visualization of per-class recall values changed by CA-EAA on EPIC-Kitchens. (a) Overview of changes that CA-EAA brings to each action class where most classes achieve performance gain. (b) Selected classes corresponding to both performance gain and decay.









Past Observation	Anticipation W/O CA-EAA	Anticipation With CA-EAA	Future Action								
	<table border="1"> <tr><td>put bowl</td></tr> <tr><td>put down meat</td></tr> <tr><td>open yoghurt bowl</td></tr> <tr><td>put cherry</td></tr> </table>	put bowl	put down meat	open yoghurt bowl	put cherry	<table border="1"> <tr><td>put down meat</td></tr> <tr><td>open yoghurt bowl</td></tr> <tr><td>put cherry</td></tr> <tr><td>get meat</td></tr> </table>	put down meat	open yoghurt bowl	put cherry	get meat	
put bowl											
put down meat											
open yoghurt bowl											
put cherry											
put down meat											
open yoghurt bowl											
put cherry											
get meat											
	<table border="1"> <tr><td>put down knife</td></tr> <tr><td>take cutting board</td></tr> <tr><td>take knife</td></tr> <tr><td>take lid</td></tr> </table>	put down knife	take cutting board	take knife	take lid	<table border="1"> <tr><td>take cutting board</td></tr> <tr><td>put down knife</td></tr> <tr><td>take knife</td></tr> <tr><td>take lid</td></tr> </table>	take cutting board	put down knife	take knife	take lid	
put down knife											
take cutting board											
take knife											
take lid											
take cutting board											
put down knife											
take knife											
take lid											
	<table border="1"> <tr><td>put in dishwasher</td></tr> <tr><td>take plate</td></tr> <tr><td>put bowl</td></tr> <tr><td>close dishwasher</td></tr> </table>	put in dishwasher	take plate	put bowl	close dishwasher	<table border="1"> <tr><td>take plate</td></tr> <tr><td>put bowl</td></tr> <tr><td>close drawer</td></tr> <tr><td>put in dishwasher</td></tr> </table>	take plate	put bowl	close drawer	put in dishwasher	
put in dishwasher											
take plate											
put bowl											
close dishwasher											
take plate											
put bowl											
close drawer											
put in dishwasher											
	<table border="1"> <tr><td>close drawer</td></tr> <tr><td>take spoon</td></tr> <tr><td>take spoon</td></tr> <tr><td>take fork</td></tr> </table>	close drawer	take spoon	take spoon	take fork	<table border="1"> <tr><td>take spoon</td></tr> <tr><td>take fork</td></tr> <tr><td>put down spoon</td></tr> <tr><td>wash fork</td></tr> </table>	take spoon	take fork	put down spoon	wash fork	
close drawer											
take spoon											
take spoon											
take fork											
take spoon											
take fork											
put down spoon											
wash fork											

Figure 5: Qualitative examples from EPIC-Kitchens (The ground-truth labels of future actions are highlighted in red).

from Figure 4 (a) that CA-EAA generally has a positive effect on most classes. In particular, the action class “open microwave” pointed by a red arrow has never been correctly predicted before we introduce CA-EAA. Considering (*, “open microwave”) as any consecutive action pair where “open microwave” is the future action, the poor performance of “open microwave” is mainly because (*, “open microwave”) occupies a small proportion. For example, we find that (“take fork”, “open microwave”) occupies only 0.68% of all (“take fork”, *) pairs. Given the observed action as “take fork”, it

is almost impossible to predict the upcoming action as “open microwave” without CA-EAA. Fortunately, CA-EAA makes it possible for action classes like “open microwave” to stand out. Figure 4 (b) lists some classes corresponding to both performance gain and decay. It is worth noting that CA-EAA also lowers the performance of several action classes. As we can observe, these classes share similar peculiarities. For example, “close fridge” can be found in action pair (“take turkey”, “close fridge”) that occupies the vast majority of (“take turkey”, *). Given the observed action “take turkey”, the probability distributions of both factual and counterfactual outcomes are highly consistent. After the introduction of CA-EAA, the probability of each class becomes almost equal and the ground-truth “close fridge” may be confused with other classes, which leads to incorrect predictions. We will further investigate how to adaptively use counterfactual analysis to avoid this kind of “overcorrection” in the future.

5 Conclusion

In this paper, we propose CA-EAA, a counterfactual analysis framework for egocentric action anticipation. It is the first time that causal inference is explored to address this task. We can enhance anticipative ability by leveraging counterfactual analysis to reduce the side-effect of past semantic labels on predicting future actions. Experimental results on two large-scale egocentric video datasets demonstrate the effectiveness of our method. In the future, we consider deeper investigation into the problem by exploring causal inference from more aspects including adaptively using counterfactual analysis and analyzing effects among different visual modalities.

Acknowledgements

This work was supported by National Key Research and Development Project of New Generation Artificial Intelligence of China, under Grant 2018AAA0102500.

References

- [Carreira *et al.*, 2018] João Carreira, Andrew Zisserman, and Quo Vadis. Action recognition? a new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4724–4733, 2018.
- [Chernozhukov *et al.*, 2013] Victor Chernozhukov, Iván Fernández-Val, and Blaise Melly. Inference on counterfactual distributions. *Econometrica*, 81(6):2205–2268, 2013.
- [Damen *et al.*, 2018] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision*, pages 720–736, 2018.
- [Furnari and Farinella, 2019] Antonino Furnari and Giovanni Maria Farinella. What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6252–6261, 2019.
- [Furnari *et al.*, 2018] Antonino Furnari, Sebastiano Battiato, and Giovanni Maria Farinella. Leveraging uncertainty to rethink loss functions and evaluation measures for egocentric action anticipation. In *Proceedings of the European Conference on Computer Vision Workshops*, 2018.
- [Kanade and Hebert, 2012] Takeo Kanade and Martial Hebert. First-person vision. *Proceedings of the IEEE*, 100(8):2442–2453, 2012.
- [King, 2008] Brayden G King. A political mediation model of corporate response to social movement activism. *Administrative Science Quarterly*, 53(3):395–421, 2008.
- [Li *et al.*, 2018] Yin Li, Miao Liu, and James M Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Proceedings of the European Conference on Computer Vision*, pages 619–635, 2018.
- [Liu *et al.*, 2020] Miao Liu, Siyu Tang, Yin Li, and James M Rehg. Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In *Proceedings of the European Conference on Computer Vision*, pages 704–721, 2020.
- [Miech *et al.*, 2019] Antoine Miech, Ivan Laptev, Josef Sivic, Heng Wang, Lorenzo Torresani, and Du Tran. Leveraging the present to anticipate the future in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019.
- [Niu *et al.*, 2020] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. *arXiv preprint arXiv:2006.04315*, 2020.
- [Pearl and Mackenzie, 2018] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic Books, 2018.
- [Pearl *et al.*, 2016] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.
- [Richiardi *et al.*, 2013] Lorenzo Richiardi, Rino Bellocco, and Daniela Zugna. Mediation analysis in epidemiology: methods, interpretation and bias. *International journal of epidemiology*, 42(5):1511–1519, 2013.
- [Sener *et al.*, 2020] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding. In *Proceedings of the European Conference on Computer Vision*, pages 154–171, 2020.
- [Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in Neural Information Processing Systems*, 27:568–576, 2014.
- [Tang *et al.*, 2020a] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in Neural Information Processing Systems*, 33, 2020.
- [Tang *et al.*, 2020b] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3716–3725, 2020.
- [VanderWeele, 2013] Tyler J VanderWeele. A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology (Cambridge, Mass.)*, 24(2):224, 2013.
- [Wang *et al.*, 2016] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Proceedings of the European Conference on Computer Vision*, pages 20–36, 2016.
- [Wu *et al.*, 2020] Yu Wu, Linchao Zhu, Xiaohan Wang, Yi Yang, and Fei Wu. Learning to anticipate egocentric actions by imagination. *IEEE Transactions on Image Processing*, 30:1143–1152, 2020.
- [Zhang *et al.*, 2020] Tianyu Zhang, Weiqing Min, Ying Zhu, Yong Rui, and Shuqiang Jiang. An egocentric action anticipation framework via fusing intuition and analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 402–410, 2020.