Counterfactual Debiasing Inference for Compositional Action Recognition

Pengzhan Sun*
University of Electronic Science and Technology of China
pzsun@std.uestc.edu.cn

Bo Wu*
MIT-IBM Watson AI Lab
bo.wu@ibm.com

Xunsong Li
University of Electronic Science and Technology of China
lixunsong@std.uestc.edu.cn

Wen Li†
University of Electronic Science and Technology of China
liwen@uestc.edu.cn

Lixin Duan
University of Electronic Science and Technology of China
lxduan@uestc.edu.cn

Chuang Gan
MIT-IBM Watson AI Lab
ganchuang1990@gmail.com

ABSTRACT

Compositional action recognition is a novel challenge in the computer vision community and focuses on revealing the different combinations of verbs and nouns instead of treating subject-object interactions in videos as individual instances only. Existing methods tackle this challenging task by simply ignoring appearance information or fusing object appearances with dynamic instance tracklets. However, those strategies usually do not perform well for unseen action instances. For that, in this work we propose a novel learning framework called Counterfactual Debiasing Network (CDN) to improve the model generalization ability by removing the interference introduced by visual appearances of objects/subjects. It explicitly learns the appearance information in action representations and later removes the effect of such information in a causal inference manner. Specifically, we use tracklets and video content to model the factual inference by considering both appearance information and structure information. In contrast, only video content with appearance information is leveraged in the counterfactual inference. With the two inferences, we conduct a causal graph which captures and removes the bias introduced by the appearance information by subtracting the result of the counterfactual inference from that of the factual inference. By doing that, our proposed CDN method can better recognize unseen action instances by debiasing the effect of appearances. Extensive experiments on the Something-Else dataset clearly show the effectiveness of our proposed CDN over existing state-of-the-art methods.

CCS CONCEPTS

• **Computing methodologies → Activity recognition and understanding**; *Causal reasoning and diagnostics.*

*Both authors contributed equally to this research.
†Corresponding author.

MM '21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475472

KEYWORDS

compositional action recognition; action recognition; causal reasoning; counterfactual inference

ACM Reference Format:
Pengzhan Sun, Bo Wu, Xunsong Li, Wen Li, Lixin Duan, and Chuang Gan. 2021. Counterfactual Debiasing Inference for Compositional Action Recognition. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475472

1 INTRODUCTION

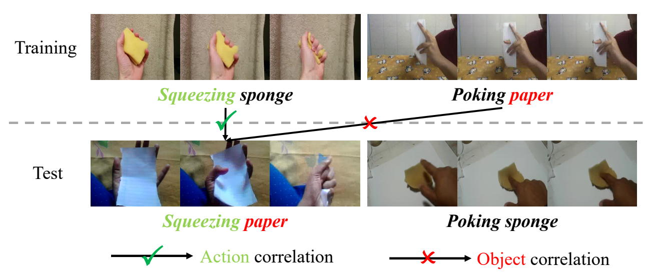


Figure 1: Examples of non-overlapping object-action compositions. The action model never sees [*squeezing paper*] during training, but sees [*paper*] occurred in action [*poking*]. Thus it gives prediction [*poking*] according to the object correlation instead of [*squeezing*] according to the action correlation when being tested with sample [*squeezing paper*].

Action recognition [3, 12, 23, 38] has been receiving much attention in computer vision area for many years. Benefited from the distribution learning power of deep networks, mainstream action recognition models [3, 8, 10, 13, 14, 25, 34, 37, 38] attempt to learn effective representations of observed dynamic actions from videos. However, it's still difficult to recognize a seen action when facing to never seen objects. Therefore, a recent research [27] proposes a novel challenge: compositional action recognition. In the setting of this task, combinations of an action and instances are not overlapped in the training set and the test set as shown in Figure 1. For existing action recognition methods, compositional action recognition is still an open-issue. Because they rely heavily

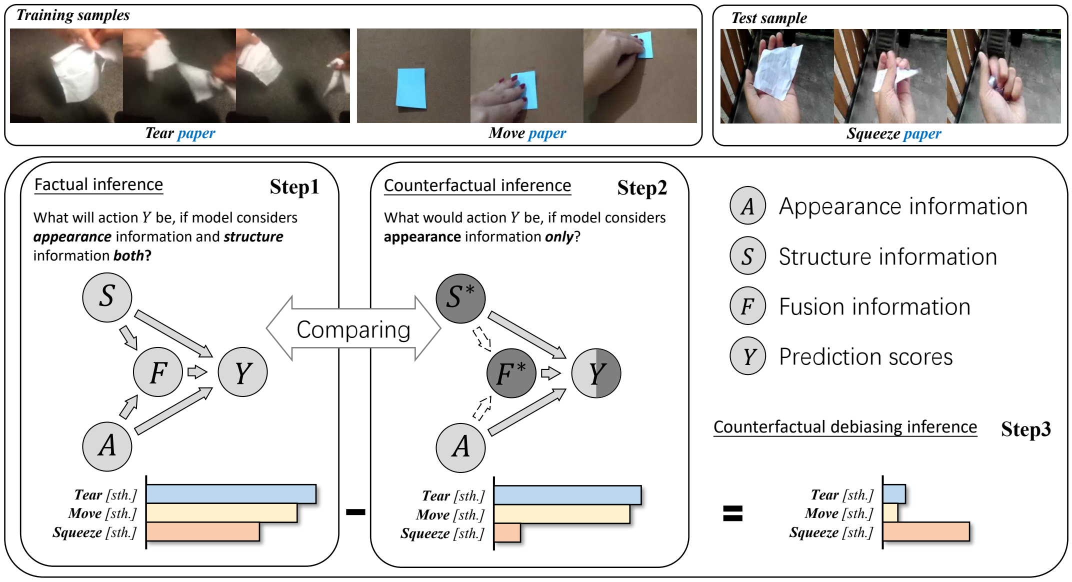


Figure 2: The illustrated example shows the counterfactual debiasing inference for compositional action recognition. Factual inference depicts the actual situation where the model considers appearance information, structure information and their fusion information together to give a prediction. Counterfactual inference depicts the virtual scenario where the model considers appearance information only. Total indirect effect used as the criterion is obtained by subtracting natural direct effect from total effect. Detailed explanation of causal graph refers to Figure 3.

on the correlation between the visual features and the prediction results [7, 22] which learned by data-driven methods. When instances and actions in given test samples combine in a way that the model has not seen before, the model will tend to give the wrong prediction results based on the prior distribution of the seen visual clues.

An intuitive solution to tackle the challenge is to break the object appearance dependency when learning a dynamic interaction, which means to inhibit the co-occurrence bias in the same action with distinct objects. By capturing the instance tracklet of an action (a continuous set of bounding boxes coordinates), Spatial-Temporal Interaction Network [27] achieves comparable performance against I3D [3]. But the strategy excusably fails when actions are associated more with the changes in terms of the intrinsic property of an object, such as "poking" and "tearing". Besides, there is another line of works [21, 44] insisting that visual information contains effective cues for compositional action recognition. Based on attention mechanism [21] or the auxiliary prediction task [44], fusing appearance information and structure information [1, 18, 21, 26, 40] brings observed improvements. However, the potential risk of appearance interference has not been solved positively in these fusion methods.

To address the aforementioned problems, in this paper, we propose a novel framework called Counterfactual Debiasing Network (CDN) by explicitly control the effect of instance appearance for compositional action recognition. Our motivation comes from the fact that the instance appearance contains both beneficial and harmful cues for compositional action recognition. As a result, the traditional appearance dependency decreasing methods or the appearance fusing methods cannot well handle this issue. We think counterfactual debiasing inference [28, 30, 31] offers a rational way to address such situation. Based on the counterfactual debiasing inference, we consider that action knowledge learned from instance appearance can be divided into two components in the causal graph. One is the bias which can be represented by the direct effect of appearance information, and the other is an effective cue that can be captured by the indirect effect through fusion information on final prediction results. With this perspective, we then propose a counterfactual debiasing inference framework to perform unbiased action prediction for compositional action recognition. By conducting counterfactual debiasing inference on the causal graph, we remove natural direct effect from total effect. More specifically, in the training stage, the classification result of the model comes from the joint contribution of appearance information A, structure information S and their fusion information F. While in the test phase, as illustrated in Figure 2, we empower CDN the ability of counterfactual analysis so that a more accurate classification result can be gained by comparing factual inference outcome and counterfactual inference outcome:

> **Factual Inference**: *What will action be, if model observes appearance information, structure information and fusion information of the above two?*
>
> **Counterfactual Inference**: *What would action be, if model observes appearance information, but* ***had not*** *observed structure information and fusion information?*

To be specific, as shown in Figure 2, given a test video with the ground truth label [*squeeze something*], CDN **first** makes factual inference to predict classification scores based on the observed appearance and tracklets of [*paper*], which is denoted as total effect. As for Total Effect (TE), scores of [*tear something*] and [*move something*] are higher than the correct answer [*squeeze something*] influenced by the appearance model activation. This is because instance [*paper*] involved in video samples with labels [*tear something*] and [*move something*] for the most samples in training set. For the **second** step, CDN conducts counterfactual inference to output classification scores only based on the appearance of [*paper*], which can be denoted as Natural Direct Effect (NDE) on classification results. The score of the wrong answer [*tear something*] dominates in NDE, for the model is cheated by the unreliable correlation learned only from appearance information. At the **last** step, by subtracting NDE from TE, the model gives its debiased final prediction [*squeeze something*] by thinking twice and comparing the answers obtained from factual inference and counterfactual inference. We verify the effectiveness of our approach on the challenging Something-Else task from the Something-Something V2 dataset [15]. CDN using Total Indirect Effect (TIE) as criterion achieves 4.0% top-1 accuracy and 3.9% top-5 accuracy improvement over state-of-the-art performance.

Our contributions can be summarized as follows:

- We observe that prior knowledge learned from appearance information is mixed with the spurious correlation between action and instance appearance, which badly inhibits the model's ability of action learning.
- We remove the pure appearance effect from total effect by counterfactual debiasing inference on our novel framework CDN proposed for compositional action recognition.
- We achieve state-of-the-art performance for compositional action recognition on the Something-Else dataset.

2 RELATED WORK

2.1 Compositional Action Recognition

Compositional action recognition [27] makes the combination of objects and actions disjoint between training and testing. This non-overlapping splitting leads to appearance bias becoming a major problem when learning actions. To tackle this issue, [27] proposed Spatial-Temporal Interaction Network (STIN) to represent actions by leveraging instance bounding boxes only to model the transformation of object geometric relations in both spatial and temporal domain. STIN generalizes well over some actions associated with object movements but fails to recognize actions about the intrinsic state changes of objects. To model such more complex actions, RGB information is introduced and fused with the spatio-temporal geometric information obtained from instance bounding boxes [21, 44]. [21] designs an attention mechanism to fuse this structure information from instance bounding boxes and visual information from RGB frames. [44] fuses these information in object-level and designs an auxiliary prediction task to guide the fusion process. In this paper, we focus on mitigating the appearance bias by conducting counterfactual debiasing inference based on the proposed causal graph.

2.2 Causal Inference in Computer Vision

Causal inference has recently inspired a wide range of works in computer vision community, which includes scene graph generation [5, 35], image recognition [35], video analysis [6, 9, 45], few-shot learning [47], zero-shot learning [46], semantic segmentation [49], and vision-language tasks [4, 32, 36, 42]. Among them, the idea of counterfactual reasoning has achieved promising results and make a step towards unbiased prediction in many tasks, especially in Visual Question Answering [29]. We need to mention that the types of bias between VQA and Compositional Action Recognition are different. For Compositional Action Recognition, the bias in the task comes from the combination distributions of verbs and nouns. Such bias from the composition is widespread in the real world and can hardly be avoided during dataset construction. In contrast, the bias in VQA comes from the imbalanced sample distribution of the dataset. In this work, we provide a new comprehension with the counterfactual debiasing inference perspective for the compositional action recognition task, for the spurious correlation exists from visual appearance when recognizing actions.

3 METHODOLOGY

Based on the analysis on the Something-Else dataset, we first observe that the prior knowledge learned from spurious visual correlation seriously inhibits the model ability of action learning. To solve this problem, we propose a causal graph for the compositional action recognition from the causal inference view. Then we introduce how to get unbiased prediction classification results using counterfactual debiasing inference on this causal graph. Finally, a novel counterfactual debiasing inference framework for compositional action recognition is given to verify our approach.

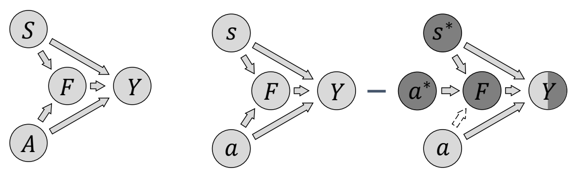
3.1 Graphical Causal Model

3.1.1 Appearance Bias in Compositional Action Recognition. Let us first take a closer look at the role of the prior action knowledge learned from appearance information. We break the correlation between object appearance and action categories by leveraing CutMix [48] and mixup [50] operations on the level of instances to explore the effect of object appearance on action predictions. For instance-level CutMix, given a video sample, each object in it is cut out according to its bounding box coordinates. Another object is sampled from the training set randomly, then resized and pasted to this given video. Similarly, we leverage mixup at the instance level. Different from [50], we fix the mixup weight as 0.5. We observe significant improvements in the performance of appearance model I3D as illustrated in Table 1. It shows that the prior action knowledge provided by objects involved in videos is mixed with the spurious correlation, which badly inhibits action learning and misleads the model to converge in this unreliable shortcut between instance appearance and action categories.

3.1.2 Causal Graph. A causal graph is constructed with four variables which includes instance appearance information A, action structure information S, fusion information F and model prediction Y, which is illustrated in Figure 3(a). It is a directed acyclic graph $\mathcal{G} = \{\mathcal{N}, \varepsilon\}$, showing how a set of variables interact with each other through causal effect links.

Table 1: Performance of I3D with instance-level CutMix and mixup on the Something-Else dataset. A noticeable improvement is profited from breaking the combinations of actions and instances.

Method	original	I3D [3] with CutMix [48]	mixup [50]
Image			
Top-1 (%)	50.5	55.4	**55.9**
Top-5 (%)	76.9	80.8	**81.4**



(a) (b)

Figure 3: (a) Causal graph for compositional action recognition. S: structure information. A: appearance information. F: fusion information. Y: prediction scores. (b) Counterfactual analysis between factual inference outcome and counterfactual inference outcome given a video sample and corresponding observed values a and s. Light node denotes real value input while dark node denotes dummy value input.

Our causal graph designed for the compositional action recognition task is highly general, which imposes no constraints on the implementation details. Now we give a detailed description of each node and link.

Node $\mathcal{A}$ (Appearance Backbone & Instance Appearance Information): A video appearance feature extractor (we use I3D in our implementation) is fixed into this node. Given a video sample V, this node outputs video-level appearance representation A:

$$Input : \{V\} \Rightarrow Output : \{A\},$$

where A is aggregated from multiple instances' appearance feature. The appearance information of instances contains useful contextual information and bias that misleads the model. However, existing compositional action recognition methods can only choose to accept or reject appearance information as a whole. We will describe how to make unbiased action predictions based on the biased appearance information.

Node $\mathcal{S}$ (Structure Backbone & Action Structure Information): Tracklets of instances in the video are available through the object detector [16, 33] and tracker [2, 19]. The action structure module takes tracklets of instances as input and outputs action structure information [21, 40] S:

$$Input : \{V\} \Rightarrow Output : \{S\}.$$

Tracklets of each instance depict how it moves and interacts with others, which are abstract and essential representation of actions and provide critical cues for correct prediction. Also, they denote unbiased information for action learning since object categories and visual information are not involved. [27] has shown that this representation of an action will achieve superior results than other state-of-the-art convolution-based video models.

Links $\{\mathcal{A}, \mathcal{S}\} \rightarrow \mathcal{F}$ (Appearance and Structure Information Input for Fusion Module): Appearance information and structure information are transposed to a fusion module to generate a better video-level representation. The effectiveness of fusion between appearance information and structure information is verified in [21], where a particular attention module guides the fusion process and leads to a better generalization ability for compositional action recognition.

Node $\mathcal{F}$ (Fusion Module & Video Fusion Information): Given the appearance information A and the structure information S of a video, the fusion module aggregates them into the video fusion information F, which is more comprehensive than either.

$$Input : \{A, S\} \Rightarrow Output : \{F\}.$$

Different modules of fusing instance appearance information and action structure information can be applied in this node, such as bilinear pooling [24, 41], attention mechanisms [21, 39, 43], and other approaches [11, 34]. For simplicity, we use a concatenation operation following with fully connected layers as the fusion module.

Link $\{\mathcal{A},\mathcal{S},\mathcal{F}\} \rightarrow \mathcal{Y}$ (Classifiers): This procedure can be formalized as:

$$Input : \{A\} \Rightarrow Output : \{Z_a\},$$
$$Input : \{S\} \Rightarrow Output : \{Z_s\},$$
$$Input : \{F\} \Rightarrow Output : \{Z_f\},$$

where Z_a, Z_s and Z_f are classification scores corresponding to A, S and F mentioned above. It is worth mentioning that Z_a is a biased classification result, and we will reduce this effect caused by bias in the subsequent counterfactual debiasing inference part.

Node $\mathcal{Y}$ (Fusion Function & Action Classification Result): The final classification prediction score $Z_{a,s,f}$ is generated by fusing all activation $\{Z_a, Z_s, Z_f\}$ using a score fusion function.

$$Input : \{Z_a, Z_s, Z_f\} \Rightarrow Output : \{Z_{a,s,f}\}.$$

We try two fusion functions in our implementation: 1) Naive Sum: $Z_{a,s,f} = Z_a + Z_s + Z_f$ 2) Log-sigmoid Sum [29]: $Z_{a,s,f} = log(\sigma(Z_a + Z_s + Z_f))$, where $\sigma(\cdot)$ is the sigmoid function.

3.2 Counterfactual Debiasing Inference

We present counterfactual debiasing inference to exclude the pure instance appearance effect through $A \rightarrow Y$ to reduce appearance bias. We denote the appearance model, the structure model and the fusion module as M_A, M_S and M_F respectively. Then we have formulations below, where a is RGB frames input and s is tracklets input:

$$\begin{aligned} M_A(a) &= \{Z_a, f_a\}, \\ M_S(s) &= \{Z_s, f_s\}, \\ M_F(f_a, f_s) &= Z_f, \end{aligned} \tag{1}$$

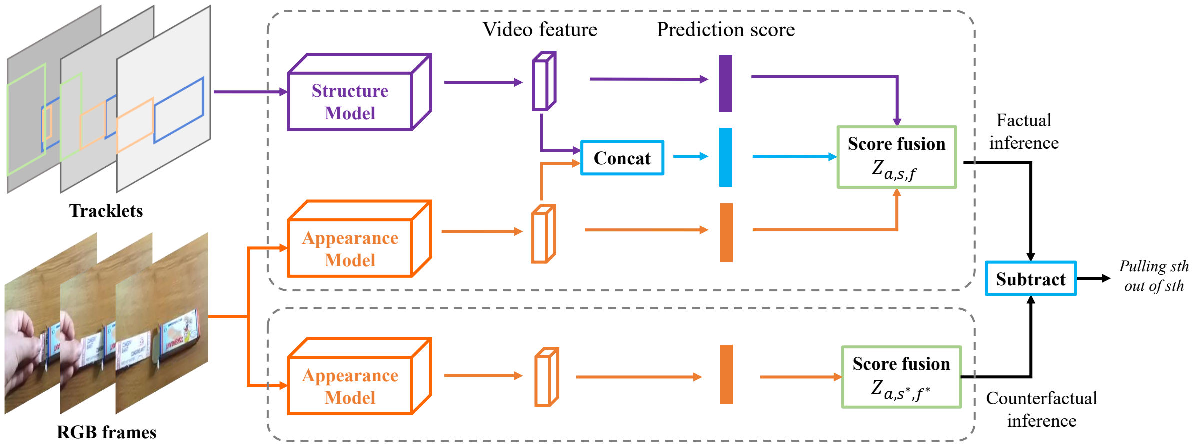


Figure 4: An overview of CDN implementation. There are no strict requirements in the specific implementation of the structure model and appearance model. The factual outcome is score fusion function's activation based on three branches. The counterfactual outcome is score fusion function's activation based on appearance branch and two zero value as placeholders.

where f_a and f_s represent features extracted from the appearance and structure backbone respectively. The final score

$$Z_{a,s,f} = h(Z_a, Z_s, Z_f), \tag{2}$$

is gained by aggregating three paths activation directly connected to Y using a fusion function h.

We denote a random variable as a capital letter and represent the corresponding observed value as a lowercase letter. The lowercase letter with the superscript $*$ represents under no-treatment control condition. For example, to recognize an action, $A = a$ represents having observed instance appearance in this action video, then $A = a^*$ represents having not observed instance appearance.

To capture the appearance bias, we need to observe the causal effect of direct path $A \longrightarrow Y$ when blocking the activation from other pathways. However, neural networks cannot make an inference when fed with variables of the dummy value. Therefore, we manually set the output to be a zero score for brevity instead of a learnable score like [29] when the model input is a dummy value. Our setting can be formalized as:

$$Z_a = \begin{cases} z_a = M_A(a) & A = a \\ z_a^* = 0 & A = a^* \end{cases}, \tag{3}$$

$$Z_s = \begin{cases} z_s = M_S(s) & S = s \\ z_s^* = 0 & S = s^* \end{cases}, \tag{4}$$

$$Z_f = \begin{cases} z_f = M_F(f_a, f_s) & A = a \text{ and } S = s \\ z_f^* = 0 & A = a^* \text{ or } S = s^* \end{cases}. \tag{5}$$

Total effect [30] denotes the effect of individual and mediator together on the outcome, which can be decomposed as the sum of direct effect and indirect effect. Total effect of $A = a$ and $S = s$ on the classification result Y can be represented as:

$$TE = Z_{a,s,f} - Z_{a^*,s^*,f^*}, \tag{6}$$

where $Z_{a,s,f}$ is the inference outcome based on $A = a$ and $S = s$, and Z_{a^*,s^*,f^*} is the inference outcome based on $A = a^*$ and $S = s^*$. According to our causal graph, the effect of appearance information A on classification result Y can be divided into direct effect $A \longrightarrow Y$ and indirect effect $A \longrightarrow F \longrightarrow Y$. Counterfactual debiasing inference aims for blocking the direct effect $A \longrightarrow Y$ while retaining the indirect effect $A \longrightarrow F \longrightarrow Y$. In this way, we achieve removing the bias while keeping the good context cue in appearance information. Natural direct effect [30] denotes the effect of an individual on the outcome with the blocked mediator. The direct effect of appearance information can be captured using natural direct effect (NDE):

$$NDE = Z_{a,s^*,f^*} - Z_{a^*,s^*,f^*}. \tag{7}$$

Finally, by doing a simple minus calculation as shown in Figure 3(b), we subtract counterfactual inference outcome NDE from factual inference outcome TE to eliminate visual bias and obtain a more reasonable and accurate result, total indirect effect [30] (TIE):

$$TIE = TE - NDE = Z_{a,s,f} - Z_{a,s^*,f^*}. \tag{8}$$

In our implementation, a hyperparameter α controls the proportion of NDE we want to remove from TE.

We formalize the implementation of TIE as follow:

$$TIE = Z_{a,s,f} - \alpha \cdot Z_{a,s^*,f^*}. \tag{9}$$

We choose the classification result with the highest TIE, which is different from the traditional method based on the posterior possibility.

3.3 Framework Implementation

We propose a framework CDN with implementations based on the causal graph built before. Thanks to our causal graph and model framework, other modules can be embedded into our CDN so long as the corresponding output has the same semantic information.

Note that without loss of generality, we implement our model as simply as possible. For the appearance model, we choose I3D

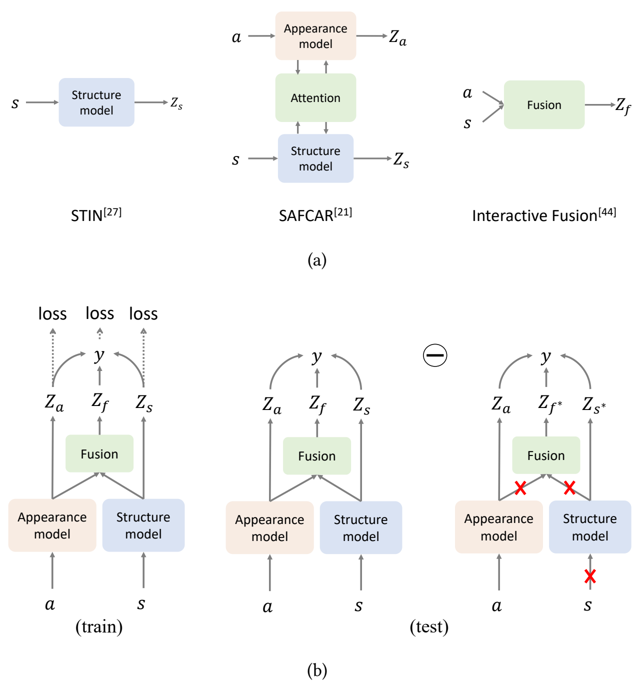


Figure 5: (a) Existing frameworks for compositional action recognition. (b) Different from existing frameworks, a three-branch framework is designed corresponding to our causal graph. Note that the debiased effect is set as the criterion, which is different from traditional posterior probability.

as our feature extractor backbone because of its generality and simplicity. With the guidance of instance bounding boxes annotated [18, 27] or detected [17, 40] in each video frame, instance-level appearance features can be gained by using RoI-pooling [16, 33]. The video-level action appearance feature is generated by average pooling all instance appearance features in both spatial and temporal space. Our action structure model adopts the similar way in [27] that takes instance bounding box coordinates and its identity embedding as input and feeds them into fully connected layers to obtain instance-centric representation. We get the frame-level representation through performing pair-wise reasoning between instances at each frame and aggregate these frame descriptors in the temporal domain to get the video-level action structure feature. We use a concatenation operation followed with an MLP as our fusion module.

In **training** stage, inspired by Counterfactual VQA [29], two auxiliary loss items are added into our model to stabilize the causal influence of each independent branch. Without these auxiliary loss items, the model tends to converge to a single branch which converges fastest. That would lead other branch activation to output meaningless perturbations. The whole loss function can be formalized as follow

$$\mathcal{L} = \mathcal{L}_F(a, s, f) + \mathcal{L}_A(a) + \mathcal{L}_S(s), \tag{10}$$

where $\mathcal{L}_F(a, s, f)$, $\mathcal{L}_A(a)$ and $\mathcal{L}_S(s)$ are cross-entropy losses over $Z_{a,s,f}$, Z_a and Z_s.

During **inference** stage, we use the outcome of counterfactual debiasing inference, total indirect effect, as the criterion, which is implemented as Eq. (9).

4 EXPERIMENTS

4.1 Dataset

We validate our approach on the Something-Else [27] task, which is the extension of the Something-Something V2 [15] dataset but follows the compositional data split setting. The Something-Else task defines a subset of frequent object categories (appearing in more than 100 videos in the dataset) and splits it into two disjoint groups, $\mathcal{A}$ and $\mathcal{B}$. The total 174 action categories are divided into two groups (1 and 2) as well. According to the splits of groups, each video in the Something-Else dataset will be assigned as one of $1\mathcal{A}$, $1\mathcal{B}$, $2\mathcal{A}$, $2\mathcal{B}$. Then the training set is a collection of $1\mathcal{A} + 2\mathcal{B}$ and the validation set is $1\mathcal{B} + 2\mathcal{A}$. As a result, there are 112,795 videos (54,919 for training and 57,876 for validation) with the compositional setting.

4.2 Implementation Details

We sample 16 frames for RGB input and 8 frames for bounding box tracklets input (follow the parameter settings in [27]). We use the ground-truth bounding boxes annotations released in [27]. I3D [3] is selected as the backbone of our appearance model and initialized with Kinetics-400 [20] pre-trained weights. The dimension of both video appearance feature and structure feature is $d = 512$. The structure model of our CDN is trained for 30 epochs with a learning rate 0.01 using SGD with 0.0001 weight decay and 0.9 momentum, the learning rate is decayed by the factor of 10 at epochs 24. The learning rate of the appearance model in CDN is set to 0.6 times that of the structure model. We set a batch size of 16 and implement our method using PyTorch on 4 Nvidia GeForce RTX 2080Ti GPUs.

4.3 Methods and Baselines

To validate the effectiveness of our CDN, we compare CDN with the recent methods in the follows:

- **I3D** [3]: Applying 3D convolution over RGB frames to obtain action representations.
- **STIN** [27]: Leveraging instance bounding boxes and category information to represent instances and performing spatial-temporal interaction to model the geometric relation transformation of actions.
- **SAFCAR** [21]: A two-branch model takes RGB frames and instance tracklets as input and fuses the two branch information with an attention module.
- **Interactive Fusion** [44]: Fusing information from appearance and tracklets information in object-level and designing an auxiliary prediction task to guide the fusion process to represent actions.
- **CDN w/o CF**: A basic version of our approach with the Log-sigmoid Sum fusion function using the traditional posterior probability as criterion. Note that counterfactual debiasing inference is not used in this basic version.

- **CDN**: The complete version our of approach with the Log-sigmoid Sum fusion function using our total indirect effect observed from the difference between factual inference results and counterfactual inference results as criterion.

Figure 5(a) shows the input information and overall architectures of existing compositional action recognition models. Figure 5(b) shows a brief training and test pipeline of our approach CDN.

4.4 Results

As shown in Table 2, methods that use the appearance and structure information both within an action outperform than those processing only the single one, which means that instance appearance information brings prior knowledge for compositional action recognition. Based on the causal graph, our designed model CDN achieves slightly higher performance than baseline methods by using traditional posterior probability as the criterion. After applying counterfactual debiasing inference, CDN can easily improve its prediction accuracy on Top-1 (1.7%) and Top-5 (0.9%) by using total indirect effect as the criterion. This shows that our counterfactual debiasing inference could mitigate the bias and keep effective cues in appearance information by only adopting a minor modification during the test stage. Overall, the complete result of our **CDN** outperforms state-of-the-art performance [21, 27, 44] with a noticeable margin.

Table 2: Recognition accuracy comparison against state-of-the-art methods on the Something-Else dataset.

Method	Input		Something-Else	
	RGB	Track	Top-1 (%)	Top-5 (%)
I3D [3]	o		50.5	76.9
STIN [27]		o	51.4	79.3
STIN+I3D [27]	o	o	54.6	79.4
Interactive Fusion [44]	o	o	59.6	85.8
SAFCAR [21]	o	o	60.5	84.3
Our CDN w/o CF	o	o	62.8	87.3
Our CDN	o	o	**64.5**	**88.2**

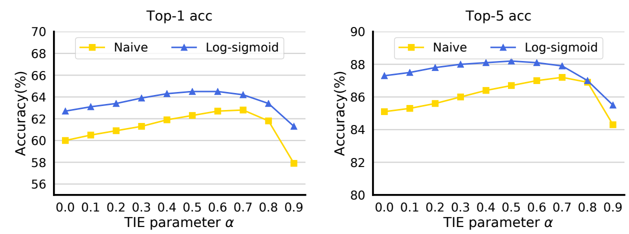
4.5 Ablation Study

Fusion function: Note that the score fusion function is an indispensable part instead of an ensemble trick for CDN. Both the factual and counterfactual outcomes are calculated by the fusion function. Therefore, the model cannot give any output when using CF without the fusion function. For reference, we provide the performance of each single model without fusion function and CF as shown in Table 3. We try Naive Sum and Log-sigmoid Sum respectively to generate the final prediction results. We only substitute Naive Sum function with Log-sigmoid Sum function, leading to a performance improvement. This suggests that the selection of score fusion function has a great impact on the final prediction results.

Effect of different TIE parameter α: The hyperparameter α used in our implementation controls the trade-off between total indirect effect and total effect. The higher value of α, the less dependent on appearance information of model prediction results.

Table 3: Ablation of fusion function effectiveness on CDN.

Method	Something-Else	
	Top-1 (%)	Top-5 (%)
Single Appearance Model	58.9	84.1
Single Structure Model	53.8	80.5
Single Fusion Module	34.0	63.6
CDN w/o CF (Naive Sum)	60.1	85.0
CDN w/o CF (Log-sigmoid Sum)	62.8	87.3
CDN (Naive Sum)	62.8	87.2
CDN (Log-sigmoid Sum)	64.5	88.2



(a) (b)

Figure 6: Naive Sum and Log-sigmoid Sum used in accuracy with different TIE weight.

When α equals 0, total indirect effect on classification results Z degenerates into total effect, which is equivalent to results gained from traditional inference strategies based on posterior probability. As α increases from 0 to 1, the performance of CDN first increases and drops down around α=0.7 as shown in Figure 6. Here we select α=0.5 for Log-sigmoid Sum score fusion function and α=0.7 for Naive Sum.

By searching for a proper value of α, CDN successes in mitigating the bias while keeping the good context in appearance information. This further illustrates that a compromise between learning action knowledge from visual information and totally discarding visual cues is the most reasonable solution for compositional action recognition.

Category Analysis: We compare the accuracy improvement on individual action categories when applying counterfactual debiasing inference on our CDN. As illustrated in Figure 7, actions that are more associated with instance appearance information benefit a lot from our counterfactual analysis. For example, [*pulling two ends of something so that it separates into two pieces*] depicts a situation where objects appearance changing much, from a whole instance into two pieces. [*pouring something into sth. until it overflows*] describes a scenario where liquid such as water and milk flows out of a container.

Example Analysis: Figure 8 visualizes examples of how our CDN performs when applying counterfactual debiasing inference or not. For example, [*paper*] is shared by action [*squeezing something*] in test and action [*poking a hole into something soft*] in training. Three objects occurring in [*poking a hole into something soft*] most

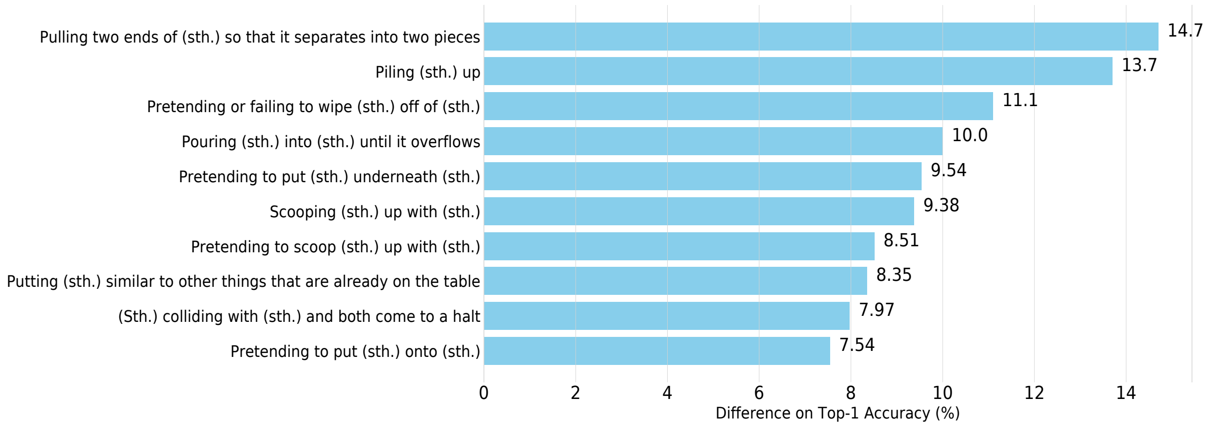


Figure 7: Top 10 action categories on which counterfactual debiasing inference exceeds traditional inference.

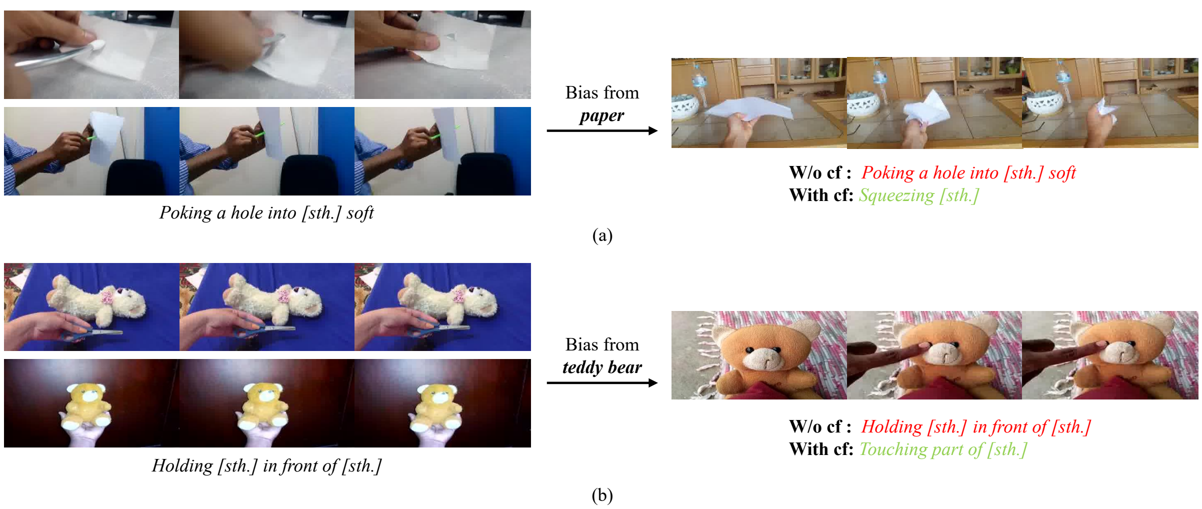


Figure 8: Visualization on representative samples. With cf represents applying counterfactual inference while W/o cf represents not applying counterfactual inference. The correct and false predictions are highlighted in green and red respectively.

frequently are [*paper*], [*pillow*] and [*bread*], accounting for 29.7%, 26.6%, and 9.4% respectively. This verifies the action [*poking a hole into something soft*] is biased due to the high object correlation with [*paper*] and [*pillow*]. Therefore, the correlation between [*paper*] appearance and action [*poking a hole into something soft*] learned from the training set misleads the model to give a wrong prediction classification result if we use posterior probability as the criterion. However, CDN can overcome its biased prior distribution learned from the dataset with the help of counterfactual debiasing inference. A correct answer can be given since it does not rely on the shortcut provided by spurious appearance correlation through subtracting the biased classification results from the total effect.

5 CONCLUSION

In this paper, we first observed that a spurious correlation between instance appearance and action category exists, which badly inhibits the model's ability of action learning. To solve this problem, we presented a novel counterfactual framework for compositional action recognition to provide an elegant solution for blocking the shortcut that the model learned from pure vision bias. With the help of counterfactual thinking, we captured the pure appearance direct effect on classification scores, which would be subtracted from total effect on the predictions. We validate our approach on the Something-Else dataset, and a new state-of-the-art performance is established by unbiased inference on our model framework.

ACKNOWLEDGEMENT

This research was supported by the National Key Research and Development Program of China under Grant No. 2018AAA0100400.

REFERENCES

[1] Fabien Baradel, Natalia Neverova, Christian Wolf, Julien Mille, and Greg Mori. 2018. Object Level Visual Reasoning in Videos. In *Computer Vision - ECCV 2018 - 15th European Conference*. 106–122.

[2] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. 2016. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*. IEEE, 3464–3468.

[3] João Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*. 4724–4733.

[4] Long Chen, Xin Yan, Jun Xiao, Hanwang Zhang, Shiliang Pu, and Yueting Zhuang. 2020. Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10800–10809.

[5] Long Chen, Hanwang Zhang, Jun Xiao, Xiangnan He, Shiliang Pu, and Shih-Fu Chang. 2019. Counterfactual critic multi-agent training for scene graph generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4613–4623.

[6] Zhenfang Chen, Jiayuan Mao, Jiajun Wu, Kwan-Yee Kenneth Wong, Joshua B Tenenbaum, and Chuang Gan. 2021. Grounding physical concepts of objects and events through dynamic visual reasoning. *arXiv preprint arXiv:2103.16564* (2021).

[7] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. 2019. Why Can't I Dance in the Mall? Learning to Mitigate Scene Bias in Action Recognition. *arXiv preprint arXiv:1912.05534* (2019).

[8] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2625–2634.

[9] Zhiyuan Fang, Shu Kong, Charless Fowlkes, and Yezhou Yang. 2019. Modularized textual grounding for counterfactual resilience. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6378–6388.

[10] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6202–6211.

[11] Yuqian Fu, Li Zhang, Junke Wang, Yanwei Fu, and Yu-Gang Jiang. 2020. Depth Guided Adaptive Meta-Fusion Network for Few-shot Video Recognition. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*. 1142–1151.

[12] Chuang Gan, Chen Sun, Lixin Duan, and Boqing Gong. 2016. Webly-supervised video recognition by mutually voting for relevant web images and web video frames. In *ECCV*. Springer, 849–866.

[13] Chuang Gan, Naiyan Wang, Yi Yang, Dit-Yan Yeung, and Alex G Hauptmann. 2015. Devnet: A deep event network for multimedia event detection and evidence recounting. In *CVPR*. 2568–2577.

[14] Chuang Gan, Ting Yao, Kuiyuan Yang, Yi Yang, and Tao Mei. 2016. You lead, we exceed: Labor-free video concept learning by jointly exploiting web videos and images. In *CVPR*. 923–932.

[15] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. 2017. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*. 5842–5850.

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[17] Roei Herzig, Elad Levi, Huijuan Xu, Hang Gao, Eli Brosh, Xiaolong Wang, Amir Globerson, and Trevor Darrell. 2019. Spatio-temporal action graph networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*. 0–0.

[18] Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. 2020. Action genome: Actions as compositions of spatio-temporal scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10236–10247.

[19] Rudolph Emil Kalman. 1960. A new approach to linear filtering and prediction problems. (1960).

[20] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[21] Tae Soo Kim and Gregory D. Hager. 2020. SAFCAR: Structured Attention Fusion for Compositional Action Recognition. abs/2012.02109 (2020). arXiv:2012.02109

[22] Yingwei Li, Yi Li, and Nuno Vasconcelos. 2018. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 513–528.

[23] Ji Lin, Chuang Gan, and Song Han. 2019. TSM: Temporal Shift Module for Efficient Video Understanding. In *2019 IEEE/CVF International Conference on Computer Vision*. 7082–7092.

[24] Jinlai Liu, Zehuan Yuan, and Changhu Wang. 2018. Towards good practices for multi-modal fusion in large-scale video classification. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 0–0.

[25] Xiang Long, Chuang Gan, Gerard De Melo, Jiajun Wu, Xiao Liu, and Shilei Wen. 2018. Attention clusters: Purely attention based local feature integration for video classification. In *CVPR*. 7834–7843.

[26] Chih-Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, and Hans Peter Graf. 2018. Attend and interact: Higher-order object interactions for video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6790–6800.

[27] Joanna Materzynska, Tete Xiao, Roei Herzig, Huijuan Xu, Xiaolong Wang, and Trevor Darrell. 2020. Something-Else: Compositional Action Recognition With Spatial-Temporal Interaction Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1046–1056.

[28] Leland Gerson Neuberg. 2003. Causality: Models, Reasoning, and Inference.

[29] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12700–12710.

[30] Judea Pearl. 2001. Direct and indirect effects. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*. 411–420.

[31] Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

[32] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. 2020. Two causal principles for improving visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10860–10869.

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497* (2015).

[34] Karen Simonyan and Andrew Zisserman. 2014. Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199* (2014).

[35] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. 2020. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3716–3725.

[36] Damien Teney, Ehsan Abbasnedjad, and Anton van den Hengel. 2020. Learning what makes a difference from counterfactual examples and gradient supervision. *arXiv preprint arXiv:2004.09034* (2020).

[37] Du Tran, Lubomir D. Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features with 3D Convolutional Networks. In *2015 IEEE International Conference on Computer Vision*. 4489–4497.

[38] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*. Springer, 20–36.

[39] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7794–7803.

[40] Xiaolong Wang and Abhinav Gupta. 2018. Videos as space-time region graphs. In *Proceedings of the European conference on computer vision (ECCV)*. 399–417.

[41] Yunbo Wang, Mingsheng Long, Jianmin Wang, and Philip S Yu. 2017. Spatiotemporal pyramid network for video action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1529–1538.

[42] Bo Wu, Haoyu Qin, Alireza Zareian, Carl Vondrick, and Shih-Fu Chang. 2020. Analogical reasoning for visually grounded language acquisition. *arXiv preprint arXiv:2007.11668* (2020).

[43] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. 2019. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 284–293.

[44] Rui Yan, Lingxi Xie, Xiangbo Shu, and Jinhui Tang. 2020. Interactive Fusion of Multi-level Features for Compositional Activity Recognition. arXiv:2012.05689

[45] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. 2019. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442* (2019).

[46] Zhongqi Yue, Tan Wang, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. 2021. Counterfactual Zero-Shot and Open-Set Visual Recognition. *arXiv preprint arXiv:2103.00887* (2021).

[47] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. 2020. Interventional few-shot learning. *arXiv preprint arXiv:2009.13000* (2020).

[48] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. 2019. CutMix: Regularization Strategy to Train Strong Classifiers With Localizable Features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[49] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xiansheng Hua, and Qianru Sun. 2020. Causal intervention for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2009.12547* (2020).

[50] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* (2017).