Improving Weakly-supervised Object Localization via Causal Intervention

Feifei Shao[a], Yawei Luo *[a], Li Zhang[b], Lu Ye[c], Siliang Tang[a], Yi Yang[a], and Jun Xiao[a]

[a]*Zhejiang University*
[b]*Zhejiang Insigma digital technology CO.LTD*
[c]*Zhejiang University of Science and Technology*

Abstract

The recently emerged weakly-supervised object localization (WSOL) methods can learn to localize an object in the image only using image-level labels. Previous works endeavor to perceive the interval objects from the small and sparse discriminative attention map, yet ignoring the co-occurrence confounder (e.g., duck and water), which makes the model inspection (e.g., CAM) hard to distinguish between the object and context. In this paper, we make an early attempt to tackle this challenge via causal intervention (CI). Our proposed method, dubbed CI-CAM, explores the causalities among image features, contexts, and categories to eliminate the biased object-context entanglement in the class activation maps thus improving the accuracy of object localization. Extensive experiments on several benchmarks demonstrate the effectiveness of CI-CAM in learning the clear object boundary from confounding contexts. Particularly, on the CUB-200-2011 which severely suffers from the co-occurrence confounder, CI-CAM significantly outperforms the traditional CAM-based baseline (58.39% vs 52.4% in Top-1 localization accuracy). While in more general scenarios such as ILSVRC 2016, CI-CAM can also perform on par with the state of the arts.

1. Introduction

Object localization [37, 5] aims to indicate the category, the spatial location and scope of an object in a given image, in forms of bounding box [8, 30]. This task has been studied extensively in the computer vision community [37] due to its broad applications, such as scene understanding and autonomous driving. Recently, the techniques based on deep convolutional neural networks (DCNNs) [32, 34, 12, 17, 19] promote the localization performance to a new level. How-
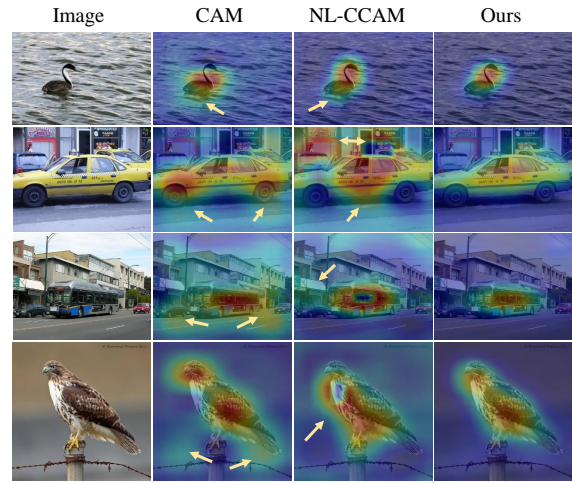
---

*Corresponding author: yaweiluo329@gmail.com

Figure 1. Visualization comparison between CAM, NL-CCAM and CI-CAM (ours). The yellow arrows indicate the regions suffer from entangled context.

ever, this performance promotion is at the price of huge amounts of fine-grained human annotations[18, 16]. To alleviate such a heavy burden, weakly-supervised object localization (WSOL) has been proposed by only resorting to image-level labels.

To capitalize on the image-level labels, the existing studies [31, 6, 14, 40] follow the Class Activation Mapping (CAM) approach [50] to generate class activation maps first and segment the highest activation area for a coarse localization. Albeit, CAM is initially designed for the classification task and tends to focus only on the most discriminative feature to increase its classification accuracy. Targeting on this issue, recent prevailing works [6, 14, 40, 49, 10, 21, 44] endeavor to perceive the interval objects from the small and sparse "discriminative regions". On one hand, they adjust the network structure to make the detector more tailored for object localization in a weak supervision setting. For exam-

ple, some methods [6, 40] use a three stages network structure to continuously optimize the prediction results by training the current stage using the output of the previous stages as supervision. Besides, some methods [14, 49, 10, 21] use two parallel branches that its first branch is responsible for digging out the most discriminative small regions, and the second branch is responsible for detecting the second discriminative large regions. On the other hand, they also make full use of the image information to improve the prediction results. For example, TS^2C [40] detects bounding boxes by comparing the confidence values between its pixels and surrounding pixels, and NL-CCAM [44] combines activation maps of low probability class to alleviate the impact of the bright background area on the object localization.

In contrast to the prevailing efforts focusing on the most discriminative feature, in this work we target another key issue which we called "entangled context". The reason behind this issue is that the objects usually co-occur with a certain background. For example, if most "duck" appears concurrently with "water" in the images, these two concepts would be inevitably entangled and a classification model would wrongly generate ambiguous boundary between "duck" and "water" with only image-level supervision. In contrast to the vanilla CAM which yields a relatively small bounding box on the small discriminative region, we notice that our targeted problem would cause a biased bounding box that includes the wrongly entangled background, which impairs the localization accuracy in terms of the object range. Consequently, we argue that resolving the "entangled context" issue is vital for WSOL but remains unnoticed and unexplored.

In this paper, we propose a principled solution, dubbed CI-CAM, tailored for WSOL based on Causal Inference [25] (CI). CI-CAM ascribes the "entangled context" issue to the frequently co-occurred background that misleads the image-level classification model to learn spurious correlations between pixels and labels. To find those intrinsic pixels which truly cause the image-level label, CI-CAM first establishes a structural causal model (SCM) [27]. Based on SCM, CI-CAM further regards the object context as a confounder and explores the causalities among image features, contexts, and labels to eliminate the biased co-occurrence in the class activation maps. More specifically, the CI-CAM causal context pool accumulates the contextual information of each image for each category and employs it as attention in convolutional layers to enhance the feature representation for making feature boundary clearer. To our knowledge, we are making an early attempt to apprehend and approach the "entangled context" issue for WSOL. To sum up, the contributions of this paper are as follows.

- We are among the pioneers to concern and reveal the "entangled context" issue of WSOL that remains un-

explored by prevailing efforts.

- We propose a principled solution for the "entangled context" issue based on causal inference, in which we pinpoint the context as a confounder and eliminate its negative effects on image features via backdoor adjustment [27].

- We design a new network structure, dubbed CI-CAM, to embed the causal inference into the WSOL pipeline with an end-to-end scheme.

- The proposed CI-CAM achieves state-of-the-art performance on the CUB-200-2011 dataset [38], which significantly outperforms the baseline method by 2.16% and 5.99% in terms of classification and localization accuracy, respectively. While in more general scenarios such as ILSVRC 2016 [30] which suffer less from the "entangled context" due to the huge amount of images and various backgrounds, CI-CAM can also perform on par with the state of the arts.

2. Related Work

2.1. Weakly-supervised Object Localization

Since CAM [50] is prone to bias to the most discriminative part of the object rather than the integral object, the research attention of most of the current methods is how to improve the accuracy of object localization. These methods can be broadly categorized into two groups: enlarging the proposal region and discriminative region removal.

1) **Enlarging proposal region**: enlarging the box size appropriately of the initial prediction box [6, 40]. WCCN [6] introduces three cascaded networks trained in an end-to-end pipeline. The latter stage continuously enlarges and refines the output proposals of its previous stage. TS^2C [40] selects the box by comparing the mean pixel confidence values of the initial prediction region and its surrounding region. If the gap of the mean values of two regions is large, the initial prediction region is the final prediction box; otherwise, the surrounding region.

2) **Discriminative region removal**: detecting the bigger region after removing the most discriminative region [14, 49, 4, 21]. TP-WSL [14] first detects the most discriminative region in the first network, Then, it erases this region of the conv5-3 feature maps in the second network (*e.g.*, zero). ACoL [49] uses the masked feature maps by erasing the most discriminative region discovered in the first classifier as the input feature maps of the second classifier. ADL [4] uses the self-attention map to produce an erased mask using thresholding and an importance map using sigmoid activation, respectively. Then, it stochastically selects the erased mask or the importance map at each iteration as a final attention map projected in the feature maps of images.

EIL [21] is an adversarial erasing method that simultaneously computes the erased branch and the unerased branch by sharing one classifier.

The above methods basically focus on how to solve the problem of poor localization performance caused by the most discriminative part of the object. However, they ignore the problem of the fuzzy boundary between the objects and the co-occur certain context background. For example, if most "duck" appears concurrently with "water" in the images, these two concepts would be inevitably entangled and a classification model would wrongly generate ambiguous boundary between "duck" and "water" with only image-level supervision.

2.2. Causal Inference

Causal inference [33, 46, 26] is a critical research topic across many domains, such as statistics [27], politics [13], psychology, and epidemiology [20, 29]. The purpose of causal inference is to give the model the ability to pursue causal effects: we can eliminate false bias [2], clarify the expected model effects [3], and modularize reusable features to make them well generalized [24]. Nowadays, causal inference is used repeatedly in computer vision tasks [23, 28, 35, 36, 39, 45, 47, 48]. Specifically, Zhang *et al*. [48] utilize a SCM [7, 27] to deeply analyze the causalities among image features, contexts, and class labels and propose a new network: Context Adjustment (CONTA) that achieves the new state-of-the-art performance in weakly-supervised semantic segmentation task. Yue *et al*. [47] use the causal intervention in few-shot learning. They uncover the pre-trained knowledge is indeed a confounder that limits the performance. And they propose a novel FSL paradigm: Interventional Few-Shot Learning (IFSL), which is implemented via the backdoor adjustment [27]. Tang *et al*. [35] show that the SGD momentum is essentially a confounder in long-tailed classification by using a structural causal model.

In our work, we also leverage a structural causal model [27] - every node in the graph can be located and implemented in the WSOL pipeline. By analyzing the causalities among image features, contexts, and class labels, we find that context is a confounder factor. In §3.4, we will introduce a causal context pool that is used for eliminating the negative effects of the context and keeping its positive effects.

3. Methodology

In this section, we describe the details of the CI-CAM method as shown in Figure 3. We first introduce the preliminaries of CI-CAM including problem settings, causal inference, and baseline method in §3.1. Second, we formulate the causalities among pixels, context, and labels with a Structural Causal Model (SCM) in §3.2. Based on SCM, we
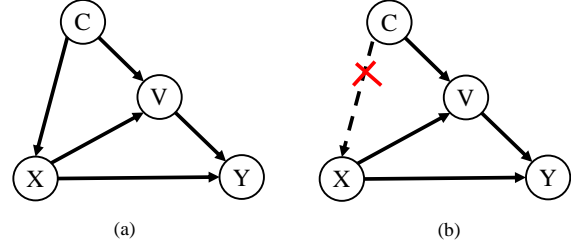

Figure 2. (a) The structural causal model (SCM) for causality of classifier in WSOL, (b) The intervened SCM for the causality of classifier in WSOL.

approach the "entangled context" issue in a principled way via causal inference in §3.3. We design the network structure of CI-CAM to embed causal inference in the WSOL pipeline detailed in §3.4, at the core of which is the causal context pool. Finally, we give the training objective of CI-CAM in §3.5.

3.1. Preliminaries

Problem Settings. Before presenting our method, we first introduce the problem settings of WSOL formally. Given an image I, WSOL targeting at classifying and locating one object in terms of the class label and the bounding box. However, only image-level labels Y can be accessed during the training phase.

Causal Inference. Causal inference enables to equip the model with the ability to pursue causal effects: it can eliminate false bias [2] as well as clarify the expected model effects [3]. SCM [27] is a directed graph in which each node represents each participant of the model, and each link denotes the causalities between the two nodes. Nowadays, SCM is widely used in causal inference scenes [47, 48, 35]. Backdoor adjustment [27] is responsible for finding the wrong impact between two nodes and eliminating this issue by leveraging the three do-calculus rules [22].

Baseline Method. Class activation maps (CAMs) are widely employed for locating the object boxes in the WSOL task. Yang *et al*. [44] argue that using only one activation map of the highest probability class for segmenting object boxes is problematic since it often biases into limited regions or sometimes even highlights background regions. Based on such observation, they propose the NL-CCAM [44] method to combine all activation maps from the highest to the lowest probability class to a localization map using a specific combinational function and achieves good localization performance.

Based on vanilla fully convolutional network (FCN)-based backbone, *e.g*., VGG16 [32], NL-CCAM [44] inserts four non-local blocks before every bottleneck layer excluding the first bottleneck layer simultaneously to produce a non-local fully convolutional network (NL-FCN). Given an image I, it is fed into the NL-FCN to produce its feature

maps $X \in \mathbb{R}^{c \times h \times w}$, where c is the number of channels and $h \times w$ is the spatial size. Then, the feature maps X are forwarded to a global average pooling (GAP) layer followed by a classifier with a fully connected layer. The prediction scores $S = \{s_1, s_2, \ldots, s_n\}$ are computed by using a softmax layer on the top of the classifier for classification. The weight matrix of the classifier is denoted as $W \in \mathbb{R}^{n \times c}$, where n is the number of image classes. Therefore, the activation maps M_i of class i among class activation maps (CAMs) $M \in \mathbb{R}^{n \times h \times w}$ proposed in [50] are given as follows.

$$M_i = \sum_k^c W_{i,k} \cdot X_k, \tag{1}$$

where $i \in \{1, 2, \ldots, n\}$.

NL-CCAM [44] produces a localization map by using a combinational function in CAMs instead of using the activation map of the highest probability class among CAMs. Firstly, it ranks the activation maps from the highest probability class to the lowest and uses M_{t_k} to denote the activation map of the k highest probability class. The class label with the highest probability t_1 is computed as follows.

$$t_1 = argmax_k(\{S_k\}), \tag{2}$$

where $k \in \{1, 2, \ldots, n\}$. Then it combines the class activation maps M to a localization map $H \in \mathbb{R}^{h \times w}$ as follows.

$$H = \sum_k^n \gamma(k) M_k, \tag{3}$$

where $\gamma(\cdot)$ is a combinational function. Finally, it segments the localization map H using a thresholding technique proposed in [50] to generate a bounding box for object localization.

Our method is based on NL-CCAM [44] but introduces substantial improvements. We equip the baseline network with the ability of causal inference to tackle the "entangled context" issue, which will be detailed in the following.

3.2. Structural Causal Model

In this section, we will reveal the reason why the context hurts the object localization quality. We formulate the causalities among image features X, context confounder C, and image-level labels Y, with a structural causal model (SCM) [27] shown in Figure 2 (a). The direct links denote the causalities between the two nodes: cause → effect.

$C \rightarrow X$: This link indicates that the backbone generates feature maps X under the effect of context C. Although the confounder context C is helpful for a better association between the image features X and labels Y via a model $P(Y \mid X)$, e.g., it is likely a "duck" when seeing a "water" region, $P(Y \mid X)$ mistakenly associates non-causal but

positively correlated pixels to labels, e.g., the "water" region wrongly belongs to "duck". That is one reason for the inaccurate localization in WSOL. Fortunately, as we will introduce later in §3.4, we can avoid it by using a causal context pool in causal intervention.

$C \rightarrow V \leftarrow X$: V is an image specific-representation using the contextual templates from C. For example, V tells us that the shape and location of a "duck" (foreground) in a scene (background). Note that this assumption is not ad-hoc in our model, it underpins almost every concept learning method from the classic Deformable Part Models [9] to modern CNNs [11].

$X \rightarrow Y \leftarrow V$: These links indicate that X and V together affect the label Y of an image. $V \rightarrow Y$ denotes that the contextual templates directly affect the image label. It is worth noting that even if we do not explicitly take V as an input for the WSOL model, $V \rightarrow Y$ still holds. On the contrary, if $V \nrightarrow Y$ in Figure 2 (a), the only path left from C to Y: $C \rightarrow X \rightarrow Y$. Considering the effect of context C on image features X, we will cut off the link of C between X. Then, no context C are allowed to contribute to the image label Y by training $P(Y \mid X)$, which results in never uncovering the context. So, WSOL would be impossible.

So far, we have pinpointed the role of context C played in the causal graph of image-level classification in Figure 2 (a). Next, we will introduce a causal intervention method to remove the confounding effect.

3.3. Causal Intervention for WSOL

We propose to use $P(Y \mid do(X))$ based on the backdoor adjustment [27] as the new image-level classifier, which removes the confounder C and pursues the true causality from X to Y shown in Figure 2 (b). By this way, we can achieve better classification and localization in WSOL. The key idea is to 1) cut off the link $C \rightarrow X$, and 2) stratify C into pieces $C = \{c_1, c_2, \ldots, c_n\}$, and c_i denotes the ith class context. Formally, we have

$$P(Y \mid do(X)) = \sum_i^n P(Y \mid X = x, V = f(x, c_i)) P(c_i), \tag{4}$$

where $f(X, c)$ abstractly represents that V is formed by the combination of X and c, and n is the number of image class. As C does not affect X, it guarantees X to have a fair opportunity to incorporate every context c into Y's prediction, subject to a prior $P(c)$. To simplify the forward propagation of the network, we adopt the Normalized Weighted Geometric Mean [42] to optimize Eq. (4) by moving the outer sum $\sum_i^n P(c_i)$ into the feature level

$$P(Y \mid do(X)) = P(Y \mid X = x, V = \sum_i^n f(x, c_i) P(c_i)). \tag{5}$$
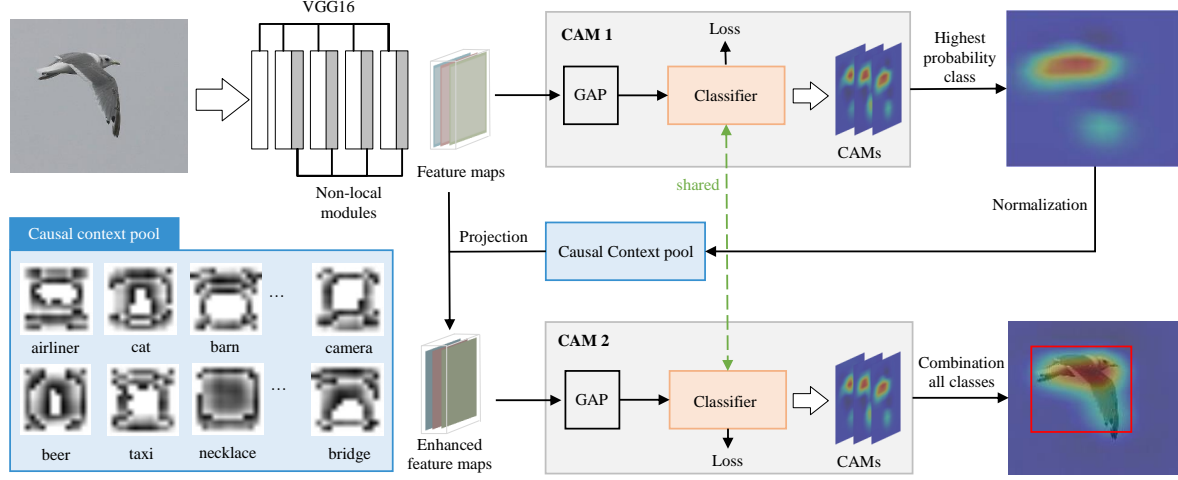
Figure 3. Overview of the proposed CI-CAM approach. CI-CAM consists of four parts: a backbone to extract the feature maps, the share-weighted CAM modules to generate the class activation maps, a causal context pool (which is the core of the CI-CAM method) to enhance the feature maps by eliminating the negative effect of confounder, and a combinational module to generate the final bounding box.

Therefore, we only need to feed-forward the network once instead of n times. Since the number of samples for each class in the dataset is roughly the same, we set $P(c)$ to uniform $1/n$. After further optimizing Eq. (5), we have

$$P(Y \mid do(X)) = P(Y \mid x \oplus \frac{1}{n} \sum_i^n f(x, c_i)), \quad (6)$$

where $\oplus$ denotes projection. So far, the "entangled context" issue has been transferred into calculating $\sum_i^n f(x, c_i)$. We will introduce a causal context pool Q to represent $\sum_i^n f(x, c_i)$ in §3.4.

3.4. Network Structure

In this section, we implement causal inference for WSOL with a tailored network structure, at the core of which is a causal context pool. The main idea of the causal context pool is to accumulate all contexts of each class, and then re-project the contexts to the feature maps of convolutional layers shown in Eq. (6) to pursue the pure causality between the cause X and the effect Y. Figure 3 illustrates the overview of CI-CAM that includes four parts: backbone, CAM module, causal context pool, and combinational part.

Backbone. Inherited from the baseline method, we design our backbone by inserting multiple non-local blocks at both low- and high-level layers of a FCN-based network simultaneously. It acts as a feature extractor that takes the RGB images as input and producing high-level position-aware feature maps.

CAM module. It includes a global average pooling (GAP) layer and a classifier with a fully connected layer [50]. Image feature maps X generated by the backbone are fed into GAP and classifier to produce prediction

scores $S = \{s_1, s_2, \ldots, s_n\}$. Then, the CAM network multiplies the weight W of the classifier to X to produce class activation maps $M \in \mathbb{R}^{n \times h \times w}$ shown in Eq. (1). In our model, we use two CAM modules with shared weights. The first CAM module is designed to produce initial prediction scores S and class activation maps M, and the second CAM network is responsible for producing more accurate prediction scores $S^e = \{s_1^e, s_2^e, \ldots, s_n^e\}$ and class activation maps $M^e \in \mathbb{R}^{n \times h \times w}$ using the feature maps $X^e \in \mathbb{R}^{c \times h \times w}$ enhanced by the causal context pool.

Causal context pool. We maintain a causal context pool $Q \in \mathbb{R}^{n \times h \times w}$ during the network training phase, where Q_i denotes the context of all i^{th} class images. Q ceaselessly stores all contextual information maps of each class by accumulating the activation map of the highest probability class. Then, it projects all contexts of each class as attentions onto the feature maps of the last convolutional layer to produce enhanced feature maps. The idea behind using a causal context pool is not only to cut off the negative effect of entangled context on image feature maps but also to spotlight the positive region of the image feature maps for boosting localization performance.

Combinational part. The input of the combinational part is class activation maps M generated from the CAM module, and the corresponding output is a localization map $H \in \mathbb{R}^{h \times w}$ calculated by Eq. (3): First, the combinational part ranks the activation maps from the highest probability class to the lowest. Second, it combines these sorted activation maps by a combinational function as Eq. (3).

With all the key modules presented above, we would give a brief illustration of the data flow in our network. Given an image I, we first forward I to the backbone to produce feature maps X. X is then fed into the following two parallel

CAM branches. The first CAM branch produces initial prediction scores S and class activation maps M. Then, the causal context pool Q would be updated by fusing the activation map of the highest probability class in M as follows:

$$Q_\pi = BN(Q_\pi + \lambda \times BN(M_\pi))), \quad (7)$$

where $\pi = argmax(\{s_1, s_2, \ldots, s_n\})$, λ denotes the update rate, and BN denotes the batch normalization operation. The second branch is responsible for producing more accurate prediction scores S^e and class activation maps M^e. The input of the second branch is enhanced feature maps X^e projected by the context among causal context pool Q of the highest probability class generated from the first branch. More concretely, the feature enhancement can be calculated as

$$X^e = X + X \odot Conv_{1 \times 1}(Q_\pi), \quad (8)$$

where $\odot$ denotes matrix dot product. In the combinational part, we first build a localization map $H \in \mathbb{R}^{h \times w}$ by aggregating all activation maps from the highest to the lowest probability class using a specific combinational function [44] in Eq. (3). Then, we use the simple thresholding technique proposed by [50] to generate a bounding box B from the localization map. Finally, the bounding box B and prediction scores S^e as the final prediction of CI-CAM.

3.5. Training Objective

During the phase of training, our proposed network learns to minimize image classification losses for both classification branches. Given an image I, we can obtain initial prediction scores $S = \{s_1, s_2, \ldots, s_n\}$ and more accurate prediction scores $S^e = \{s_1^e, s_2^e, \ldots, s_n^e\}$ of the two classifiers shown in Figure 3. We follow a naive scheme to train the two classifiers together in an end-to-end pipeline using the following loss function L.

$$L = \left(-\sum_{i=1}^{n} s_i^* log(s_i) \right) + \left(-\sum_{i=1}^{n} s_i^* log(s_i^e) \right), \quad (9)$$

where s^* is the ground-truth label of an image.

4. Experiments

4.1. Datasets and Evaluation Metrics

Datasets. The proposed CI-CAM was evaluated on two public datasets, *i.e.*, **CUB-200-2011** [38] and **ILSVRC 2016** [30]. 1) **CUB-200-2011** is an extended version of Caltech-UCSD Birds 200 (CUB-200) [41] containing 200 bird species which focuses on the study of subordinate categorization. Based on the CUB-200, CUB-200-2011 adds

more images for each category and labels new part localization annotations. CUB-200-2011 contains $5,994$ images in the training set and $5,794$ images in the test set. Each image of CUB-200-2011 is annotated by the bounding boxes, part locations, and attribute labels. 2) **ILSVRC 2016** is the dataset originally prepared for the ImageNet Large Scale Visual Recognition Challenge (ILSVRC). It contains 1.2 million images of $1,000$ categories in the training set, $50,000$ in the validation set and $100,000$ images in the test set. For both datasets, we only utilize the image-level classification labels for training, as constrained by the problem setting in WSOL.

Evaluation Metrics. We leverage the **classification accuracy (cls) and localization accuracy (loc)** as the evaluation metrics for WSOL. The former includes Top-1 and Top-5 classification accuracy, while the latter includes Top-1, Top-5, and GT-know localization accuracy. Top-1 classification accuracy denotes the accuracy of the highest prediction score (likewise for localization accuracy). Top-5 classification accuracy denotes that if one of the five predictions with the highest score is correct, it counts as correct (likewise for localization accuracy). GT-know localization accuracy is the accuracy that only considers localization regardless of classification result compared to Top-1 localization accuracy [21].

4.2. Implementation Details

We use PyTorch for implementation. We adopt the VGG16 [32] pre-trained on ImageNet [30] as the backbone. We insert four non-local blocks to the backbone before every bottleneck layer excluding the first one. The newly added blocks are randomly initialized except for the batch normalization layers in the non-local blocks, which are initialized as zero. We use Adam [15] to optimize CI-CAM with $\beta_1 = 0.9$, $\beta_2 = 0.99$. We finetune our network with the learning rate 0.0005, batch size 6, update rate λ 0.01, and epoch 100 on the CUB-200-2011 (and learning rate 0.0001, batch size 40, update rate λ 0.01, and epoch 20 on the ILSVRC 2016). At test stage, we resize images to 224×224. For object localization, we produce the localization map with the threshold $\theta = 0.0$, followed by segmenting it to generate a bounding box. The source code will be made public available.

4.3. Comparison with State-of-the-Art Methods

We compare CI-CAM with other state-of-the-art methods on the CUB-200-2011 and ILSVRC 2016 as shown in Table 1 and Table 2, respectively.

on the CUB-200-2011, we observe that our CI-CAM significantly outperforms the baseline and can be on par with the existing methods under all the evaluation metrics: CI-CAM yields Top-1 classification accuracy of 75.56%,

Table 1. Performance on the CUB-200-2011 test set. ∗ indicates our re-implemented results.

Methods	Top-1 Cls(%)	Top-1 Loc(%)
VGG-CAM [50]	71.24	44.15
VGG-ACoL [49]	71.90	45.92
VGG-ADL [4]	65.27	52.36
VGG-DANet [43]	75.40	52.52
VGG-NL-CCAM [44]	73.4	52.4
VGG-MEIL [21]	74.77	57.46
VGG-Rethinking-CAM [1]	74.91	**61.3**
VGG-Baseline∗ [44]	74.45	52.54
VGG-CI-CAM(ours)	**75.56** (1.11 ↑)	58.39 (5.85 ↑)

Table 2. Performance on the ILSVRC 2016 validation set. ∗ indicates the our re-implemented results.

Methods	Top-1 Cls(%)	Top-1 Loc(%)
VGG-CAM [50]	66.6	42.8
VGG-ACoL [49]	67.5	45.83
VGG-ADL [4]	69.48	44.92
VGG-NL-CCAM [44]	72.3	**50.17**
VGG-MEIL [21]	70.27	46.81
VGG-Rethinking-CAM [1]	67.22	45.4
VGG-Baseline∗ [44]	72.15	48.55
VGG-CI-CAM(ours)	**72.62** (0.47 ↑)	48.71 (0.16 ↑)

which is 1.11% higher than the baseline and brings about 5.85% improvement to the baseline in terms of Top-1 localization accuracy. Compared with the current state-of-the-art method in terms of classification accuracy, *i.e.*, DANet [43], CI-CAM outperforms it by 0.16%. While under the localization metric, CI-CAM brings a significant performance gain of 5.87% over DANet. Compared with the current state-of-the-art method in terms of Top-1 localization accuracy, *i.e.*, Rethinking-CAM, our model yields a slightly lower localization accuracy but outperforms it on classification. In conclusion, the introduction of causal inference to WSOL task is effective for on both object localization and classification.

For more general scenarios as in ILSVRC 2016 which suffer less from the "entangled context" due to the huge amount of images and various backgrounds, CI-CAM can also perform on par with the state of the arts. Compared with the NL-CCAM model (Baseline∗ re-implemented by ourselves), which enjoys state-of-the-art classification and localization accuracy simultaneously, CI-CAM yields a slightly higher result under both metrics. Compared with MEIL [43], CI-CAM brings a significant performance gain of 2.35% and 1.9% in classification and localization accuracy over it, respectively.

4.4. Ablation Study

To better understand the effectiveness of the causal context pool module, we conducted several ablation studies on the CUB-200-2011 and ILSVRC 2016 using VGG16. The
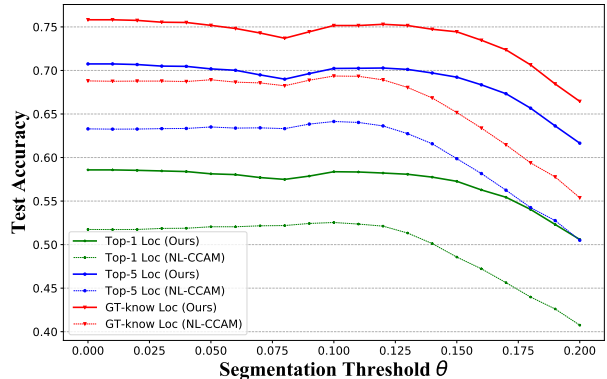

Figure 4. Ablation studies of threshold θ on the CUB-200-2011 test set (casual context pool update rate $\lambda = 0.01$).

results of our ablation studies are illustrated in Table 3.

Comparing the first-row and the second-row experimental results on the CUB-200-2011 dataset, employing a causal context pool can comprehensively improve the accuracy of classification and localization. Especially it has increased by 1.13% and 1.23% in the Top-1 classification accuracy and the Top-1 localization accuracy, respectively. Meanwhile, it also improves by 0.18%, 0.42%, and 0.31% in the Top-5 classification accuracy, Top-5 localization accuracy, and GT-know localization accuracy, respectively. In addition, comparing the third-row and the fourth-row experimental results on the ILSVRC 2016 dataset, we are surprised to find that the causal context pool also performs well on the ILSVRC 2016 dataset which suffers less from the "entangled context" due to the huge amount of images and various backgrounds. More specifically, using a causal context pool has improved by 0.46%, 0.56%, and 0.54% in the Top-1, Top-5, and GT-know localization accuracy. At the same time, the Top-1 and Top-5 classification accuracy has been increased by 0.02% and 0.03% by using a causal context pool, respectively.

In conclusion, employing a causal context pool can improve classification and localization together on the CUB-200-2011 dataset. And the main improvement of employing the causal context pool on the ILSVRC 2016 dataset is localization accuracy. The experimental results from Table 3 verify that the introduced causal context pool module can boost the accuracy in the WSOL task.

4.5. Analysis

As shown in the Eq. (7), we introduce a hyperparameter λ for updating the causal context pool. Besides, we use a necessary hyperparameter segmentation threshold θ for segmenting bounding box. Therefore, we will discuss their effects on detection performance when λ and θ take different values in this section on the CUB-200-2011 dataset.

1) *Update rate λ*. To inspect the effect of the update rate

Table 3. Ablation study on the CUB-200-2011 and ILSVRC 2016 datasets. 1) TwoC: Two classifiers, 2)ConPool: Causal context pool.

Dataset	TwoC	ConPool	Top-1 Cls(%)	Top-5 Cls(%)	Top-1 Loc(%)	Top-5 Loc(%)	GT-know Loc(%)
CUB-200-2011	√		74.43	91.85	57.16	70.12	75.37
test set	√	√	**75.56**	**92.03**	**58.39**	**70.54**	**75.68**
ILSVRC 2016	√		72.60	90.90	48.25	58.20	61.82
val set	√	√	**72.62**	**90.93**	**48.71**	**58.76**	**62.36**
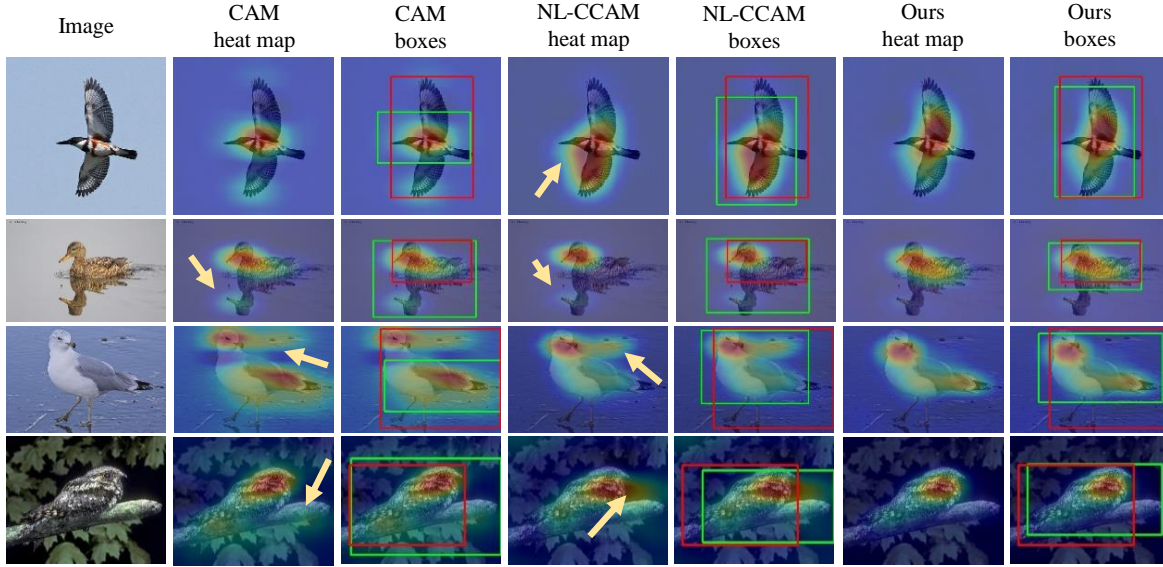
(a) CUB-200-2011
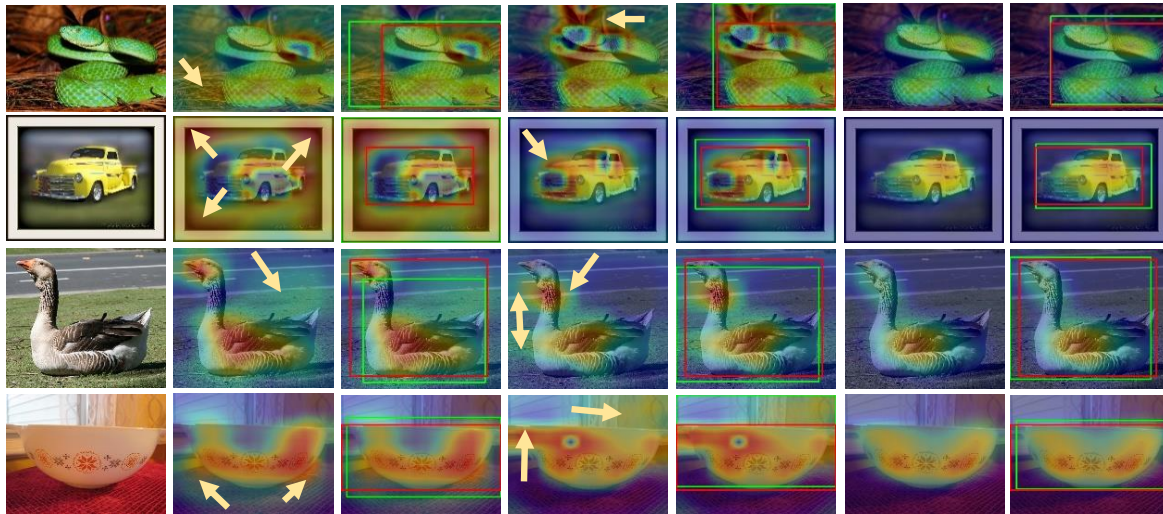
(b) ILSVRC 2016

Figure 5. Qualitative object localization results compared with the CAM and NL-CCAM methods. The predicted bounding boxes are in green, and the ground-truth boxes are in red. The yellow arrows indicate the regions suffer from entangled context.

λ on classification and localization accuracy, we report the results of using different values of λ shown in Table 4. By comparing the results we can observe that the update rate λ has a great impact on the classification and localization

accuracy of the model, especially in GT-know localization (75.68% *vs* 77.80%). The highest Top-1 classification accuracy outperforms the lowest Top-1 classification accuracy by 1.36%, and the highest Top-1 localization accuracy out-

Table 4. Ablation studies of casual context pool update rate λ on the CUB-200-2011 test set (threshold $\theta = 0.0$).

Update rate λ	Top-1 Cls(%)	Top-1 Loc(%)	GT-know Loc
0.001	74.28	58.28	76.37
0.002	74.77	**59.23**	**77.80**
0.005	74.49	58.27	76.94
0.01	**75.56**	58.39	75.68
0.02	74.58	57.90	76.54
0.04	74.53	59.03	77.63
0.08	74.20	58.70	77.77

performs the lowest Top-1 localization accuracy by 1.33%. However, in these experiments, there is no λ that performs best in both classification and localization. Therefore, when we choose the update rate λ, we should determine it according to the specific needs of the task.

2) *Segmentation threshold θ.* Although θ does not participate in the training of the model, it still plays a very important role in object localization. If the value of θ is low, the detector tends to introduce some highlighted background area around the object. So previous methods [50, 49] used segmentation threshold $\theta = 0.2$, and NL-CCAM used segmentation threshold $\theta = 0.12$. Based on a large segmentation threshold, they tend to filter out low-light object regions and focus on the most discriminative part of the object rather than the whole object. Fortunately, we leverage a causal context pool to resolve the above problem by making the boundary of object and co-occurrence context (such as duck and water) clearer.

To inspect the effect of causal context pool on classification and localization accuracy, we test the different segmentation threshold θ shown in Figure 4. Firstly, we find that the best localization of NL-CCAM is $\theta = 0.1$. When θ becomes smaller or larger, it will reduce the localization accuracy. Secondly, we can observe that the localization accuracy of CI-CAM is higher than that of NL-CCAM. Especially we obtain the highest Top-1 localization, Top-5 localization, and GT-know localization accuracy when $\theta = 0.0$, which means that CI-CAM can locate a larger part of the object without including the background. Therefore, we can indirectly conclude that CI-CAM is better in dealing with the boundary between the instance and the co-occurrence context background. To illustrate the effect of CI-CAM more vividly, we present the localization maps of the CAM, NL-CCAM, and our model on the CUB-200-2011 as well as the ILSVRC 2016 datasets in Figure 5. Our visualization in Figure 5 indicates our method is effective in dealing with the co-occurrence context.

5. Conclusions

In this paper, we targeted the "entangled context" problem in the WSOL task, which remains unnoticed and un-

explored by existing efforts. Through analyzing the causal relationship between image features, context, and image labels using a structural causal model, we pinpointed the context as a confounder and tried to utilize probability formula transformation to cut off the link between context and image features. Based on the causal analysis, we proposed an end-to-end CI-CAM model, which uses a causal context pool to accumulate all contexts of each class, and then re-project the fused contexts to the feature maps of convolutional layers to make the feature boundary clearer. To our knowledge, we have made a very early attempt to apprehend and approach the "entangled context" issue for WSOL. Extensive experiments have demonstrated that the "entangled context" is a practical issue within the WSOL task and our proposed method is effective towards it: CI-CAM achieved the new state-of-the-art performance on the CUB-200-2011 which severely suffers from the co-occurrence background, which significantly outperforms the traditional CAM-based baseline (58.39% *vs* 52.4% in Top-1 localization accuracy). While in more general scenarios such as in ILSVRC, CI-CAM can also perform on par with the state of the arts.

References

[1] Wonho Bae, Junhyug Noh, and Gunhee Kim. Rethinking class activation mapping for weakly supervised object localization. In *European Conference on Computer Vision*, pages 618–634. Springer, 2020.

[2] Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In *Artificial Intelligence and Statistics*, pages 100–108. PMLR, 2012.

[3] Michel Besserve, Arash Mehrjou, Rémy Sun, and Bernhard Schölkopf. Counterfactuals uncover the modular structure of deep generative models. *arXiv preprint arXiv:1812.03253*, 2018.

[4] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2219–2228, 2019.

[5] Sandipan Choudhuri, Nibaran Das, Ritesh Sarkhel, and Mita Nasipuri. Object localization on natural scenes: A survey. *International Journal of Pattern Recognition and Artificial Intelligence*, 32(02):1855001, 2018.

[6] Ali Diba, Vivek Sharma, Ali Pazandeh, Hamed Pirsiavash, and Luc Van Gool. Weakly supervised cascaded convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 914–922, 2017.

[7] Vanessa Didelez and Iris Pigeot. Judea pearl: Causality: Models, reasoning, and inference. *Politische Vierteljahresschrift*, 42(2):313–315, 2001.

[8] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[9] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively

trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2009.

[10] Yan Gao, Boxiao Liu, Nan Guo, Xiaochun Ye, Fang Wan, Haihang You, and Dongrui Fan. C-midn: Coupled multiple instance detection network with segmentation guidance for weakly supervised object detection. In *ICCV*, 2019.

[11] Ross Girshick, Forrest Iandola, Trevor Darrell, and Jitendra Malik. Deformable part models are convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 437–446, 2015.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[13] Luke Keele. The statistics of causal inference: A view from political methodology. *Political Analysis*, pages 313–335, 2015.

[14] Dahun Kim, Donghyeon Cho, Donggeun Yoo, and In So Kweon. Two-phase learning for weakly supervised object localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3534–3543, 2017.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Adversarial style mining for one-shot unsupervised domain adaptation. In *Advances in Neural Information Processing Systems*, pages 20612–20623, 2020.

[17] Yawei Luo, Ping Liu, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Category-level adversarial adaptation for semantic segmentation using purified features. *IEEE Transactions on Pattern Analysis & Machine Intelligence (TPAMI)*, 2021.

[18] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2507–2516, 2019.

[19] Yawei Luo, Zhedong Zheng, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Macro-micro adversarial network for human parsing. In *Proceedings of the European conference on computer vision (ECCV)*, pages 418–434, 2018.

[20] David P MacKinnon, Amanda J Fairchild, and Matthew S Fritz. Mediation analysis. *Annu. Rev. Psychol.*, 58:593–614, 2007.

[21] Jinjie Mai, Meng Yang, and Wenfeng Luo. Erasing integrated learning: A simple yet effective approach for weakly supervised object localization. In *CVPR*, 2020.

[22] Leland Gerson Neuberg. Causality: Models, reasoning, and inference, 2003.

[23] 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.

[24] Giambattista Parascandolo, Niki Kilbertus, Mateo Rojas-Carulla, and Bernhard Schölkopf. Learning independent causal mechanisms. In *International Conference on Machine Learning*, pages 4036–4044. PMLR, 2018.

[25] Judea Pearl. Interpretation and identification of causal mediation. *Psychological methods*, 19(4):459, 2014.

[26] Judea Pearl et al. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009.

[27] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

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

[29] Lorenzo Richiardi, Rino Bellocco, and Daniela Zugna. Mediation analysis in epidemiology: methods, interpretation and bias. *International journal of epidemiology*, 42(5):1511–1519, 2013.

[30] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[31] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *arXiv*, 2014.

[33] Michael E Sobel. An introduction to causal inference. *Sociological Methods & Research*, 24(3):353–379, 1996.

[34] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[35] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *arXiv preprint arXiv:2009.12991*, 2020.

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

[37] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 648–656, 2015.

[38] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[39] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10760–10770, 2020.

[40] Yunchao Wei, Zhiqiang Shen, Bowen Cheng, Honghui Shi, Jinjun Xiong, Jiashi Feng, and Thomas Huang. Ts2c: Tight box mining with surrounding segmentation context for weakly supervised object detection. In *Proceedings of the*

European Conference on Computer Vision (ECCV), pages 434–450, 2018.

[41] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.

[42] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[43] Haolan Xue, Chang Liu, Fang Wan, Jianbin Jiao, Xiangyang Ji, and Qixiang Ye. Danet: Divergent activation for weakly supervised object localization. In *ICCV*, 2019.

[44] Seunghan Yang, Yoonhyung Kim, Youngeun Kim, and Changick Kim. Combinational class activation maps for weakly supervised object localization. In *WACV*, 2020.

[45] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *arXiv preprint arXiv:2003.03923*, 2020.

[46] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *arXiv preprint arXiv:2002.02770*, 2020.

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

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

[49] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1325–1334, 2018.

[50] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.