Contextual Debiasing for Visual Recognition with Causal Mechanisms

Ruyang Liu*[1] Hao Liu*[2] Ge Li ✉[1] Haodi Hou[2] TingHao YU[2] Tao Yang[2]
[1]School of Electronic and Computer Engineering, Peking University [2]Tencent AI Department
{ruyang@stu,geli@ece}.pku.edu.cn {paulhliu,haodihou,maxwellyu,rigorosyang}@tencent.com

Abstract

As a common problem in the visual world, contextual bias means the recognition may depend on the co-occurrence context rather than the objects themselves, which is even more severe in multi-label tasks due to multiple targets and the absence of location. Although some studies have focused on tackling the problem, removing the negative effect of context is still challenging because it is difficult to obtain the representation of contextual bias. In this paper, we propose a simple but effective framework employing causal inference to mitigate contextual bias. We first present a Structural Causal Model (SCM) clarifying the causal relation among object representations, context, and predictions. Then, we develop a novel Causal Context Debiasing (CCD) Module to pursue the direct effect of an instance. Specifically, we adopt causal intervention to eliminate the effect of confounder and counterfactual reasoning to obtain a Total Direct Effect (TDE) free from the contextual bias. Note that our CCD framework is orthogonal to existing statistical models and thus can be migrated to any other backbones. Extensive experiments on several multi-label classification datasets demonstrate the superiority of our model over other state-of-the-art baselines.

1. Introduction

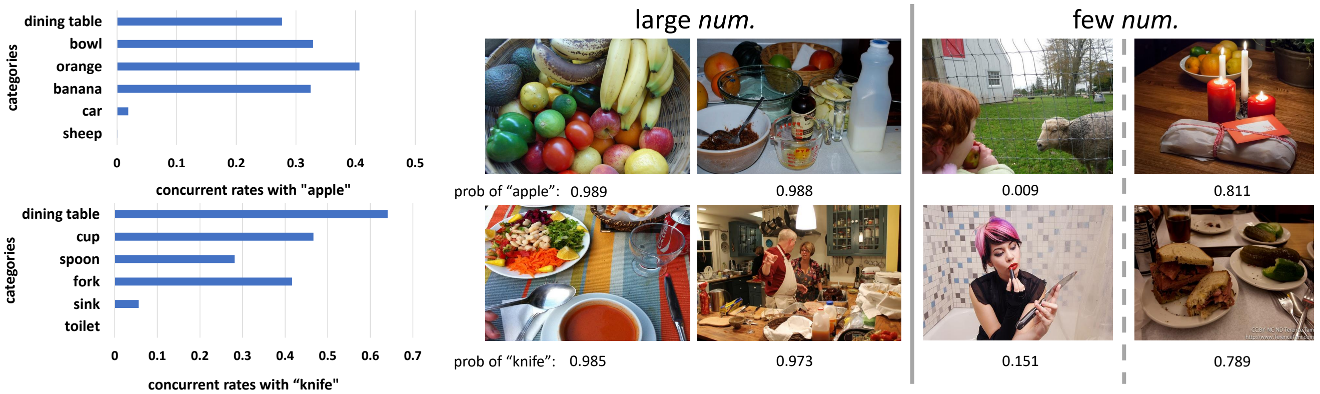
Context is a very common element in the visual world. For a single instance in an image, its context consists of other co-occurrence instances together with the background. In multi-target tasks like multi-label classification and detection, context (or, instance relation) modeling seems to have considerable potential to improve the performance. For example, the cutlery in the soup is probably a spoon rather than a fork or knife. In fact, from recurrent neural networks [40, 46], to graph convolutional networks [5,10,47], until the popular transformer-based frameworks [19, 53], recent years have witnessed numerous attempts to model the label relations in multi-instance images.

Despite the remarkable progress these models have made, they may overlook a basic question: is modeling context always beneficial in visual recognition? As is illustrated in Fig. 1, we uncovered an ever-overlooked phenomenon in multi-label classification: context may mislead the classifier, either giving the nonexistent object a high score in the scene where it usually arises, or ignoring the object appearing in a rare background. The occurrence can be partly blamed on the biased data or weak backbones; but too much attention on the label relationships will ultimately aggravate the bias. Although some works [30,47] have questioned the necessity of label modeling, they do not address the problem of contextual bias and lack the fundamental theory.

Contextual bias is not a fresh topic in the academic world. In fact, it is widely reported in many fields [11, 32]. Recent works [13] also give insight into the reason for contextual bias in computer vision: neural networks are statistics-based and ”lazy”. When the networks find the context is enough to recognize most of the objects, they often do not focus on the representations of instances. When the training data is limited (*e.g.* few-shot learning) or deficient (*e.g.* long-tail distribution), the problem is more severe and obvious. In fact, there have been some works [36, 49, 52] attending to the bias in these situation, nevertheless, they do not work on balanced datasets like MS-COCO [23] or other common tasks.

Paradoxically, context is not always bad. On one hand, bias towards context would mislead the prediction in some cases. On the other hand, it is somewhat reasonable. The appearance of contextual bias means the network indeed captured the inter-dependencies of classes: *fork* or *knife* is indeed more rational to appear on the dining table as opposed to on the street or grasslands, and rough intervention will damage the learning of the feature [18]. Actually, in the tasks like scene graph generation [45] or human object interactions [45] where the datasets are seriously biased, context priori has proved to be beneficial to the results. However, it does not mean contextual bias can be neglected. A more robust and sensible prediction should come from the object itself rather than the context, and classification by context is more likely the expediency. For tasks like multi-label classification, whose datasets are balanced and large, contextual bias will damage the final prediction.

Mitigating the contextual bias in common multi-label



(a)Co-occurrence of the two classes

(b)Qualitative results of the two classes from Resnet101 baseline

Figure 1. Examples that contextual information can do evil. We show two labels in MS-COCO: *apple* and *knife*. (a)Label dependencies $P(X|Y)$ of the two categories, which are computed through dividing the number of co-occurrence by the number of *apple*/*knife*. (b) Examples from the ResNet101 baseline. The left column depicts the objects in their common context, which is in large number and correctly predicted. And the right depicts the opposite: the objects in the unusual context or the absence of objects in their common context, which is in few number and incorrectly predicted.

tasks is challenging. In a relatively unbiased dataset, the contextual bias is not very obvious, hence it is difficult to obtain the representation of the context. Thanks to the theory of causality, we can revisit the context in a causal view: the image-specific contextual message is indeed a *mediator* preventing results from being generated directly by the representations of instances. Consequently, the final prediction is a mixed effect of the object and the context. Besides, the prior context knowledge (*e.g.*, biased datasets or pretrain model) acts as a *confounder* giving rise to spurious correlation among labels. As is illustrated in the last column of Fig.1(b), even though the scene of *dining table* is not the direct cause of *apple* or *knife*, the biased prior knowledge still fools the classifier to learn a correction between them.

In this paper, we build a Structural Causal Model [29] in Section 3 clarifying the causalities among elements mentioned above. With the assistance of causal inference, we propose a novel debiasing paradigm: Causal Context Debiasing (CCD) Module, to conquer the effect of contextual bias. We first implement the backdoor adjustment [28], which is essentially a causal intervention turning off the confounding switch and treating every contextual content equally in the prediction. Then, by counterfactual inference, we elegantly eliminate the effect of contextual bias and obtain the direct causal effect from the objects themselves, without hurting the feature representation learning. It is worth noting that our method is model-agnostic, hence, our approach can be used on a variety of backbones and achieve performance improvements. Moreover, different from many recent classification models which introduce complicated architecture (*e.g.* GCN and transformer) or additional external information (*e.g.*, word embeddings), the parameter increasement from our method is very limited.

The main contributions of our paper are summarized as:

- We establish a Structural Causal Model (SCM) to uncover the causal relevance among contextual priori, object feature, contextual bias, and final prediction in multi-target visual tasks. We find the image-specific context is indeed a mediator and contextual priori is a confounder, which sheds some light on how prediction is influenced by the context.
- We propose a simple, effective and model-agnostic framework for contextual debiasing based on causal inference. By the combination of backdoor intervention and counterfactual reasoning, we remove the obstacle of the confounder as well as contextual bias, obtaining the direct effect caused by target instances.
- We conduct extensive experiments in multi-label classification to verify the effectiveness of our methods. Results on three widely-used datasets MS-COCO [23], PASCAL-VOC [12] and NUS-WIDE [6] show that our approach can significantly improve both CNN and transformer backbones, outperforming the state-of-the-art on these datasets.

2. Related Work

Context Modeling in Mult-Label Classification. Modeling the contextual information is a common strategy in multi-target tasks. Since our experiments are mainly implemented in the multi-label classification, we take the model in this field for instance to unfold how the context is modeled. Grouped by methods, they can be summarized as attention-based, GCN-based and transformer-based, and grouped by motivations, they strive to model class dependencies and regions of interest.

Capturing class corrections has been a hot research topic in multi-label recognition. Earlier works model the label dependencies by graph model [21] and CNN-RNN frameworks [40, 46], and these methods tend to import external information like word vectors. Then, here comes an explosion of graph convolutional networks [3, 5, 10, 47]. Chen *et al.* [5] learned the classifier by propagating the semantic representations of categories through iterative GCN. Ye *et al.* [47] developed a dynamic graph along with a static graph to catch the dynamic dependencies among labels. However, as is demonstrated above, label dependencies may be harmful to the predictions, especially for the strong backbones that are able to capture features of objects themselves.

Finding regions of interest is another hotspot in contextual modeling. Attention mechanisms used to be the most common approaches [17, 43, 55], until the appearance of transformers in visual tasks. The transformer has the potential for both modeling label corrections and finding local regions, thus, a series of transformers for multi-label classification has been proposed [19, 53] recently. Lanchantin *et al.* [19] exploited dependencies among labels and features by virtue of label mask mechanisms in a transformer. Zhao *et al.* [53] employed a transformer-based dual learning framework to capture the structural relation and semantic relation. Nevertheless, whether transformer-based models or GCN-based models face the same problems: extensive architecture modification and external information dependence. For the former, complicated architecture makes it hard to evaluate if the performance is indeed increased by the growth of parameters. For the latter, robustness and flexibility are limited when extra data is unavailable. By comparison, our method is much more cost-friendly and can be easily implemented on different backbones.

Causalities in Computer Vision. Causal inference [28, 29] is aimed at pursuing the causal effect, which has been seen in medical, political and psychological research for years. Recently, causalities have also drawn growing attention in computer vision [35, 36, 42, 49, 52]. Significantly, many of them focus on various bias in different tasks (*e.g.* contextual bias and long-tail bias). Yue *et al.* [49] blamed contextual bias on the pretrain knowledge in the few-shot learning and then eliminated the bias by causal intervention. Zhang *et al.* [52] had even noticed the contextual bias in multi-label classification whose results are used as weak supervision in semantic segmentation, and improved the classification by the results of down-stream segmentation.

Most similar to our work, Tang *et al.* first introduced the combination of de-confounded training and the Total Direct Effect (TDE) [35, 36]. However, our CCD is distinctive in three aspects: 1)Intervention: the intervention is not an auxiliary for the TDE, *i.e.*, the two parts can be independently applied in our method. 2)TDE: the implementation of causal elements and bias representations are totally different, which are the core of the two models. 3)Training: TDE in [35, 36] is adopted in inference, in that it cannot converge in the training, but our Eq.7 is trainable.

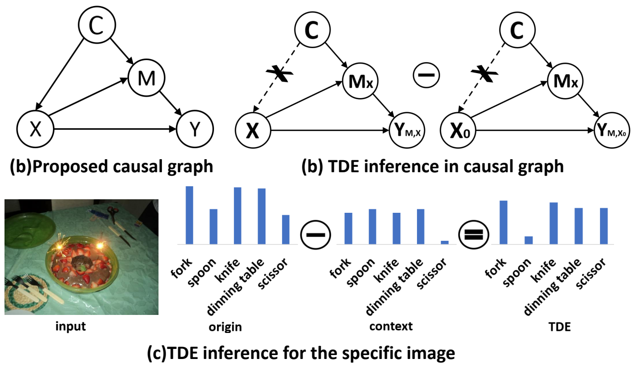
3. Structural Causal Model

As is discussed in Section 1, context as a mediator can confuse the classification and result in a suboptimal solution. To further elaborate on how classifications are misled by the bias, we build a Structural Causal Model (SCM) shown in Fig. 2 (a), which indicates the causalities among context information (C, M), object representations (X), and predictions(Y). The nodes in the directed acyclic graph denote causal entities, and the edges denote causalities between two nodes: $X \rightarrow Y$ means effect Y is caused by X. It is worth noting that the graph is applicative in many visual tasks. Next, we give a thorough introduction into the rationale behind the SCM and the causal solution is detailed in Section 4.

$\boldsymbol{C} \rightarrow \boldsymbol{X}$**.** We denote C as the prior context knowledge. There are many potential elements responsible for the contextual bias, they may be from biased datasets and pre-trained knowledge or from training process (e.g. momentum and batch normalization) or from both. Imagine that there is a biased dataset where every *fork* arises on the *dining table*, the model would be confounded to build spurious corrections between the two classes, and the learning strategy above will aggravate the existing bias. Here, we generalize those elements as a confounder set C. The link indicates the prior knowledge is highly involved in the extraction of X.

$(\boldsymbol{C}, \boldsymbol{X}) \rightarrow \boldsymbol{M}$**.** We term M as the image-specific context, which is directly from X but essentially inherited from C. For one thing, the context is a visual combination of various other objects, for example, a streetscape can be depicted as the label *car* with its context containing *person* and *building*, and it is the same when the lead is *person* or *building*. For another, theoretically, the contextual bias as representations from the image can be formed by semantic manifolds of different contextual templates [2]. In particular, how to obtain context representations of present samples is crucial for the counterfactual inference, and we will introduce the implementation in more detail in the next section.

$(\boldsymbol{X}, \boldsymbol{M}) \rightarrow \boldsymbol{Y}$**.** The links denote that the final prediction effect can be disentangled into two ways: the direct effect $X \rightarrow Y$ and the mediate effect $X \rightarrow M \rightarrow Y$. The causality in $M \rightarrow Y$ is easy to understand: the contextual information has considerable impact on the prediction for labels. The object itself and its context together affect the recognition of it. In some cases when these instances are imperceptible, context is likely to become the main effect determining the predictions. However, what determines the occurrence of an object is that the object is actually present in the image, not that ”it should occur”, which motivates us to mitigate the effect of context. Free from the negative effect of context, the predictions are more robust and reliable,



(b)Proposed causal graph

(b) TDE inference in causal graph

(c)TDE inference for the specific image

Figure 2. (a)The proposed causal graph for the causalities in visual recognition. (b)The TDE inference to remove contextual bias in the proposed causal graph. (c)An example for causal reasoning given specific input. By capturing and removing contextual bias, we filter out the hard negative and emphasize the hard positive.

and the "lazy" models are consequently forced to learn the feature of objects themselves.

4. The Proposed Solution

Contextual bias originates from many potential factors. However, these confounders on the other hand are indispensable for the training. Thanks to the causal theory, we can realize the de-confounded training through backdoor adjustment [27] without changing these useful components:

$$\mathrm{P}(Y|\mathrm{do}(X))=\sum_{c}\mathrm{P}(Y|X,C=c)\mathrm{P}(C=c), \quad (1)$$

where $\mathrm{do}(X)$ is the causal intervention cutting off the edge $C \to X$ as illustrated in Fig. 2(b). Through the *do*-operator, we can fairly evaluate the effect of the context, *e.g.*, cut off the correction that *fork* relies on *dining table*.

Now we can clarify our final destination: obtaining the classification from the objects themselves, *i.e.*, the direct effect $X \to Y$. Compared with intervention, the counterfactual is about hindsight: if the object did not occur in the image, what predictions would we make? As is conveyed in Fig. 2 (b), we first let the context become the main effect, and then, through a simple minus, we naturally remove the influence of contextual bias. In causal inference, this process is named Total Direct Effect (TDE) [26, 38, 39]:

$$\mathrm{TDE}(Y)=\mathrm{P}(Y|\mathrm{do}(X=x))-\mathrm{P}(Y|\mathrm{do}(X=x_0)), \quad (2)$$

where x_0 means the link $X \to Y$ is turned off, hence the classification is totally determined by context. Note that the link $X \to M$ is preserved because context M is affected by both C and X, if X is invisible for the M, it would be impossible to obtain the context. As is shown in Fig. 2 (c), removing the contextual bias by causal inference not only clears away the contextual hard-negative objects (*spoon*), but also highlights the hard-positive objects (*scissor*).

4.1. Causal Intervention

The key for performing backdoor adjustment is the implementation of $P(Y|X, C = c)$ and $P(C = c)$ in Eq. (1). However, C is unobservable during the training. Although an approximation will be given in Section 4, the sampling of C is still laborious. Now, we might as well look at Eq. (1) in another perspective:

$$\mathrm{P}(Y|\mathrm{do}(X=x))=\sum_{c}\frac{\mathrm{P}(Y,X=x|C=c)\mathrm{P}(C=c)}{\mathrm{P}(X=x|C=c)}. \quad (3)$$

This is a form of Inverse Probability Weighting [27] (IPW), where $1/\mathrm{P}(X = x|C = c)$ is the weight for each $(X = x|C = c)$. The equation gives us some inspiration on how to attain the desired effect: despite the infinite of C, there is correspondingly one x given one c in Eq. (3). In other words, the values of (y, x) and c are one-to-one mapped. Consequently, we can skip the $P(c)$ when conditioning on X and sample the origin confounded logits to approximate the intervened effect as below:

$$\mathrm{P}(Y|\mathrm{do}(X=x))\approx\frac{1}{N}\sum_{n=1}^{N}\frac{\mathrm{P}(Y,X=x_n)}{\mathrm{P}(X=x_n|C=c)}. \quad (4)$$

where the $C = c$ in the numerator is omitted following the common form in IPW, and we should remember X relies on C. The N means that our method divides the representation X into n parts, due to our belief that multiple sampling leads to more fine-grained and precise approximation.

Finally, Eq. (4) can be naturally modeled as an energy-based model [20], and the numerator is the common logits $w \cdot x$. Meanwhile, the denominator, *i.e.*, inverse probability weight, becomes the propensity score [1] under the energy-based model, where the effect is divided into the controlled group $(\|w\|_2 \cdot \|x\|_2)$ and uncontrolled group $(\alpha \cdot \|x\|_2)$. Therefore, we compute the final logits by assembling them as follows:

$$\mathrm{P}(Y|\mathrm{do}(X=x))=\frac{\tau}{N}\sum_{n=1}^{N}\frac{w_n x_n}{(\|w_n\|_2+\alpha)\,\|x_n\|_2}, \quad (5)$$

which is a combination of multi-head and normalization trick, and τ is a scaling factor the same with other normalization classifiers (*e.g.* cosine classifier [14]). Surprisingly, it sheds some light on why the models equipped with multi-head (*e.g.*, transformer) or normalization classifiers are more robust to defective and noisy datasets.

4.2. Total Direct Effect

After the causal intervention, the model is prepared for the direct effect. However, before conducting the counterfactual inference, we need to obtain context features given current samples. As is discussed in Section 3, the context of the certain object can be viewed as another form of "object", *e.g.*, the context of most *fork* can be summarized as

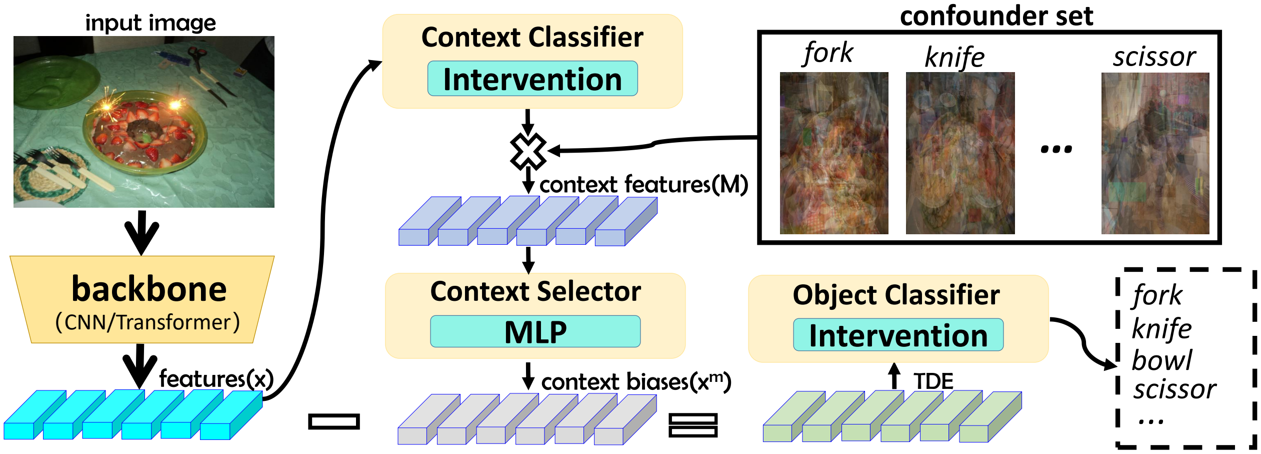


Figure 3. Overview of our proposed model. The model is composed of two modules: the intervention (interventional classifier) and the counterfactual (elimination of context bias), where we compute context bias from prior context confounder and probability for the context.

the dining or kitchen scene with other tablewares. Consequently, just like foreground objects, we define the context of each class as representations in low-dimensional manifolds [31]. Specifically, we assume C as a confounder set $\{\mathbf{c}_i\}_{i=1}^{K}$, where K is the number of categories in dataset and $\mathbf{c}_i$ is the prototype for the context of class i in feature space.

Object features can be linearly or non-linearly represented by the manifolds [2, 37], and so are the context features. Therefore, we model the image-specific context features M of current samples as follows:

$$M = f(x, C) = \sum_{i=1}^{K} \mathrm{P}(\mathbf{c}_i|x)\mathbf{c}_i, \tag{6}$$

where $P(\mathbf{c}_i|x)$ is the probability of classification that feature x belongs to the context of class i. Imagining there is a model that is able to classify the context, our implementation is originated from the viewpoint that the classifier can be viewed as the distilled knowledge [16].

Now, the last remaining difficulty is the implementation of the contextual confounder set C. Visual contexts are reported to emerge in higher layers of CNN during training [51,54], which is indeed a feature map from backbones. In tasks that rely highly on the contexts, the feature map can even approximate representations of contexts. However, the approach is somewhat unreasonable in common multi-label tasks where the context is just one of the causes for results.

Consequently, we implement C as the mean features from the model in early training, *i.e.*, training after the first several epochs. Our idea is inspired from two aspects. Theoretically, as is proposed by Zeiler *et al.* [50], lower layers of CNN encoding similar textures converge within a few epochs, while upper layers carrying semantic messages need much more time. In other words, the models in early stage are enabled to capture contextual information but have no idea about advanced semantics; hence, the classification in this period mainly depends on the contexts. Experimentally, we will show in Fig. 4 that ”hard negative” from the model sees the abnormal trend in the first several epochs. As is demonstrated in Section 1, context bias may confuse the model when facing hard negative samples, thereby, we assume the feature representations during this period are strongly affected by the context.

Given the modeling of C and M, we are prepared for the representations of context bias. Considering the context for different inputs may be different, there are several options to constitute the ultimate context features, and we model them as $x^m = f_s(x, M)$. The implementation of f_s includes identical mapping, multi-layer perceptron (MLP) and self-attention and we will explore the options in ablation study. Finally, we define the total direct effect inference as:

$$\mathrm{TDE}(Y) = \frac{\tau}{N}\sum_{n=1}^{N} \frac{w_n}{\|w_n\|_2 + \alpha}\left(\frac{x_n}{\|x_n\|_2} - \frac{x_n^m}{\|x_n^m\|_2}\right). \tag{7}$$

Different from previous works, Eq.7 can train and test in the same form. Through the simple minus, we can remove context biases in image representations and force the model to concentrate on the objects.

5. Experiments

We choose the multi-label classification as the main task in our experiment, which is free from bounding box or linguistic information and can intuitively evaluate the superiority of our proposed model. We conduct extensive experiments on several common datasets: MS-COCO [23], Pascal-VOC [12], and NUS-WIDE [6]. Following classic works [5,10,47], we employ mean average precision (mAP) as the main evaluation metric, and many other metrics as supplements, including the overall precision (OP), overall recall (OR), overall F1-measure (OF1), per-category precision (CP), per-category recall (CR), and per-category F1-measure (CF1).

Table 1. Comparison of the mAP between our method and the state-of-the-arts on the MS-COCO. Two types of backbones are presented: CNN-based (upper) and transformer-based (lower). 22k means that the backbone is pre-trained on the ImageNet-22k.

Methods	Backbones	Resolution	mAP	All						Top3					
				CP	CR	CF1	OP	OR	OF1	CP	CR	CF1	OP	OR	OF1
ResNet-101 [15]	ResNet101	448 × 448	81.5	80.4	72.9	76.3	83.6	76.7	80.0	87.1	63.6	73.5	89.4	66.0	76.0
CADM [4]	ResNet101	448 × 448	82.3	82.5	72.2	77.0	84.0	75.6	79.6	87.1	63.6	73.5	89.4	66.0	76.0
ML-GCN [5]	ResNet101	448 × 448	83.0	85.1	72.0	78.0	85.8	75.4	80.3	87.2	64.6	74.2	89.1	66.7	76.3
MS-CMA [48]	ResNet101	448 × 448	83.8	82.9	74.4	78.4	84.4	77.9	81.0	88.2	65.0	74.9	90.2	67.4	77.1
KSSNet [24]	ResNet101	448 × 448	83.7	84.6	73.2	77.2	87.8	76.2	81.5	-	-	-	-	-	-
SSGRL [3]	ResNet101	576 × 576	83.8	**89.9**	68.5	76.8	91.3	70.8	79.7	**91.9**	62.5	72.7	93.8	64.1	76.2
C-Trans [19]	ResNet101	576 × 576	85.1	86.3	74.3	79.9	87.7	76.5	81.7	90.1	65.7	76.0	92.1	**71.4**	77.6
ADD-GCN [47]	ResNet101	576 × 576	85.2	84.7	**75.9**	80.1	84.9	**79.4**	82.0	88.8	**66.2**	75.8	90.3	68.5	**77.9**
CCD-R101(Ours)	ResNet101	448 × 448	84.0	87.2	70.9	77.3	**88.8**	74.6	81.1	89.7	63.9	72.9	92.0	66.5	77.2
CCD-R101(Ours)	ResNet101	576 × 576	**85.3**	88.3	73.1	**80.2**	**88.8**	76.3	**82.1**	91.0	65.2	**76.0**	**92.3**	67.3	**77.9**
VIT [9]	VIT-L16(22k)	224 × 224	80.9	84.6	67.2	74.1	87.5	71.1	78.5	87.6	60.9	70.0	91.1	64.3	75.4
Swin-transformer [25]	Swin-B(22k)	384 × 384	88.4	83.0	82.4	82.2	83.1	84.9	84.0	89.8	70.0	77.2	90.6	71.4	79.9
Swin-transformer [25]	Swin-L(22k)	384 × 384	89.5	**87.5**	81.7	83.6	**86.8**	84.4	85.6	**92.1**	70.2	78.1	**92.4**	71.7	80.8
CCD-VIT(Ours)	VIT-L16(22k)	224 × 224	85.2	85.1	76.0	79.3	84.7	78.8	81.7	89.8	66.5	74.7	90.5	68.6	78.1
CCD-Swin(Ours)	Swin-B(22k)	384 × 384	89.1	84.7	82.7	83.2	84.1	85.5	84.8	90.8	70.0	77.6	91.2	71.8	80.4
CCD-Swin(Ours)	Swin-L(22k)	384 × 384	**90.3**	85.9	**84.0**	**84.6**	85.1	**86.4**	**85.7**	91.9	**70.9**	**78.8**	92.0	**72.4**	**81.0**

Table 2. Comparison between our method and the state-of-the-arts on the Pascal VOC2007. Both mAP and AP of each class are presented. All of the inputs are resized into 448 × 448 except ADD-GCN and SSGRL.

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
RCP [41]	98.6	97.1	98.0	95.6	75.3	94.7	95.8	97.3	73.1	90.2	80.0	97.3	96.1	94.9	96.3	78.3	94.7	76.2	97.9	91.5	90.9
ML-GCN [5]	99.5	**98.5**	98.6	98.1	80.8	94.6	97.2	98.2	82.3	95.7	86.4	98.2	98.4	96.7	99.0	84.7	96.7	84.3	98.9	93.7	94.0
ASL [30]	**99.9**	98.4	**98.9**	98.7	86.8	**98.2**	**98.7**	98.5	83.1	98.3	89.5	98.8	**99.2**	**98.6**	**99.3**	**89.5**	99.4	86.8	99.6	95.2	95.8
SSGRL(576) [3]	99.7	98.4	98.0	97.6	85.7	96.2	98.2	98.8	82.0	98.1	89.7	98.8	98.7	97.0	99.0	86.9	98.1	85.8	99.0	93.7	95.0
ADD-GCN(576) [47]	99.8	99.0	98.4	**99.0**	**86.7**	98.1	98.5	98.3	85.8	98.3	88.9	98.8	99.0	97.4	99.2	88.3	98.7	**90.7**	99.5	**97.0**	**96.0**
CCD-R101(Ours)	**99.9**	98.2	98.4	98.9	84.9	97.7	97.8	**99.0**	**86.4**	**98.8**	**90.2**	**99.2**	98.9	97.8	98.8	87.3	**99.4**	88.8	**99.7**	96.6	95.8

5.1. Implementations Details

We adopt ResNet101 [15] pre-trained on ImageNet 1k [7] as the backbone. For data preprocessing, we apply the standard data augmentation [3,5,30] with the resolution resized into 448 × 448. For x^m, we adopt the simplest implementation $x^m = M$ with no parameter increasement, and use the model trained for 5 epochs as the confounder set. The model is trained by minimizing a focal loss [22] due to the imbalance between positive and negative samples. Followinging [30,47], we apply the model pre-trained on COCO for Pascal VOC to accelerate the converge. We use Adam as optimizer with a weight decay of $2e-4$ and $(\beta_1, \beta_2) = (0.9, 0.9999)$, and the learning rate is $2e-4$ with a 1-cycle policy. All our codes were implemented in Pytorch and ran on 4 V100s, with batch size of 128 on each GPU and training for 40 epochs in total.

5.2. Comparisons with the State-of-the-Art

MS-COCO. MS-COCO [23] is first built for detection, segmentation, and caption, and then becomes the most popular benchmark for multi-label image recognition. It contains 82,081 images for training and 40,137 for validation, which covers 80 categories and 2.9 for each image on average. Given the result of COCO is highly related to image resolution and backbone, we design different backbones and different input resolutions in experiments to confirm the effectiveness of our model.

For CNN backbone, we use ResNet101 pre-trained on ImageNet 1k, and the inputs are resized into 448 × 448 and 576 × 576 respectively for fair comparisons with previous works. For transformer backbone, we adopt vision transformer [8] and swin-transformer [25] pre-trained on ImageNet 22k, and they both use the backbones trained for 2 epochs as confounder set. All quantitative results are conveyed in Table 1. The upper block shows the performances of ResNet101 backbone, in which our methods outperform all other state-of-the-arts. Considering Graph-Based methods(*e.g.* ML-GCN [5] and KSSNet [3]) and Transformer-Based methods (*e.g.* C-Tran [19]) have much more complicated and time-consuming pipelines, the superiority of our models is more convincing. The lower block shows the results on transformer backbone, which has a more elaborate architecture. Based on such high performance, common methods like asymmetric loss [30] lose efficiency, but our approach still brings about some progress.

Table 3. Comparison between our method and the state-of-the-arts on the Pascal VOC2012. Both mAP and AP of each class are presented. All of the inputs are resized into 448 × 448 except ADD-GCN and SSGRL. Results are evaluated by official evaluation server.

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
VGG+SVM [34]	96.7	83.1	94.2	92.8	61.2	82.1	89.1	94.2	64.2	83.6	70.0	92.4	91.7	84.2	93.7	59.8	93.2	75.3	99.7	78.6	84.0
HCP [44]	99.1	92.8	97.4	94.4	79.9	93.6	89.8	98.2	78.2	94.9	79.8	97.8	97.0	93.8	96.4	74.3	94.7	71.9	96.7	88.6	90.5
RCP [41]	99.3	92.2	97.5	94.9	82.3	94.1	92.4	98.5	83.8	93.5	83.1	98.1	97.3	96.0	98.8	77.7	95.1	79.4	97.7	92.4	92.2
SSGRL(576) [3]	99.7	96.1	97.7	96.5	86.9	95.8	95.0	98.9	88.3	97.6	87.4	99.1	99.2	97.3	99.0	84.8	98.3	85.8	99.2	94.1	94.8
ADD-GCN(576) [47]	**99.8**	97.1	**98.6**	96.8	**89.4**	97.1	96.5	**99.3**	89.0	97.7	87.5	**99.2**	99.1	97.7	**99.1**	86.3	**98.8**	87.0	99.3	95.4	95.5
CCD-R101(Ours)	**99.8**	**98.2**	98.3	**98.0**	88.6	**97.4**	**96.9**	99.1	**90.8**	**98.9**	**90.2**	**99.2**	**99.6**	**98.4**	99.0	**87.7**	98.4	**88.8**	**99.7**	**96.4**	**96.1**

Table 4. Comparison between our method and the state-of-the-arts on the NUS-WIDE. Both mAP, CF1 and OF1 are presented. All results have the resolution of 448 × 448.

Methods	mAP	CF1	OF1
S-CLs [24]	60.1	58.7	73.7
MS-CMA [48]	61.4	60.5	73.8
SRN [55]	62.0	58.5	73.4
ICME [5]	62.8	60.7	74.1
ASL [30]	63.9	**62.7**	74.6
CCD-R101(Ours)	**65.1**	61.3	**75.0**

Table 5. Ablation studies on two modules of our method. Experiments are based on two backbones: ResNet101 and Vit-L16.

Methods	mAP	CF1	OF1
ResNet-101	81.5	76.3	80.0
Intervention-R101	82.5	76.2	79.7
Counterfactual-R101	82.8	76.9	80.6
CCD-R101	**84.0**	**77.3**	**81.1**
VIT	80.9	74.1	78.5
Intervention-VIT	84.6	78.6	81.2
Counterfactual-VIT	84.2	78.5	80.7
CCD-VIT	**85.2**	**79.3**	**81.7**

Table 6. Ablation studies on the implementation of $f_s(x, M)$. W denotes the learnable projection matrices and $\otimes$ denotes the element-wise multiplication. All experiments are implemented on MS-COCO following the default setting.

$f_s(x, M)$	mAP	Extra Weights(M)
M	84.0	1
$W \cdot M$	83.9	17
$W \cdot \text{concat}(x, M)$	**84.4**	33
$\tanh(W \cdot x) \otimes M$	84.3	17

PASCAL-VOC. PASCAL-VOC 2007 and 2012 [12] are also widely used in multi-label classification, and the model for them obeys default settings. Following the previous [5, 10, 47], we train the model on the *train-val* set and test it on the *test* set. VOC 2007 contains 5,011 images in *train-val* set and 4,952 images in *test* set with 20 categories. The results, posed in table 2, report the Average Precise (AP) of each class and the mAP. We can see our CCD has comparable performances with previous methods, even with the weaker backbone and resoltion. VOC 2012 has 11,540 images in the *train-val* set and 10,991 test images. Different from VOC 2007, all results must be evaluated on an official evaluation server, giving a fairer comparison than local tests. As is illustrated in table 3, our method outperforms other state-of-the-art methods with a larger margin.

NUS-WIDE. NUS-WIDE [6] is another common multi-label classification dataset consisting of real-world web images. It has 269,648 Flickr images with 81 classes. We follow the default settings and the steps of evaluation in [30]. The comparison between ours and the previous best model is provided in table 4. Once again, our model obtains a new state-of-the-art result, indicating that our model still works well on a more challenging dataset with lower resolutions and more noise.

5.3. Ablation Study

Effects of the two modules in CCD. There are two main modules in CCD: the intervention part and the counterfactual part. To further study the effects of two components, we conduct controlled experiments on different backbones, each engaging one of the parts on the CNN-based or the transformer-based baseline, and the results are listed in table 5. It can be seen that both parts have an obvious advantage over baselines of different backbones. Specifically, the intervention part makes stable contribution to the result on various backbones, while the counterfactual part works better on weaker backbones or defective data that are easier to bring about the context bias.

The choice of confounder set. As is mentioned above, we have chosen the model training for the first several epochs as the confounder set, and now we provide some experimental validation. To get rid of the distractions, we close the intervention part and all data augmentation, set models in different training period as the confounder set and record the final mAP. The blue line in Fig. 4 conveys the mAP keeps on growing in the first several epochs, reaches the highest in the 5th epoch, and then continues to drop. The results prove that models in early training (1-10 epochs in this case) as confounder sets can indeed improve the performance.

Intuitively, the more the model depends on the context to recognize the instances, the more false positives it will generate. Hence, we check the amount of ”hard negative” in the training loop, where the ”hard negative” is defined as ”the score of a negative sample is larger than the mean score of positive samples in this class”. As is illustrated by

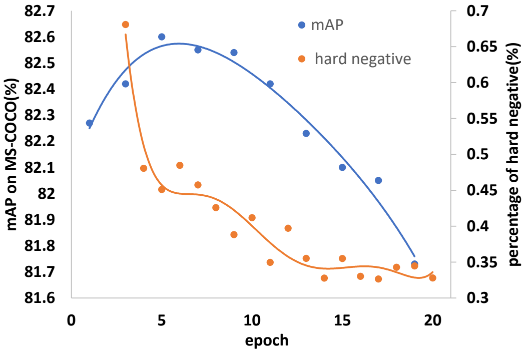


Figure 4. Ablation study on the choice of confounder set. The two curves are generated from the same model training for different epochs. The orange one depicts the percentage of hard negative, and the blue one is the mAP from counterfactual-R101 which uses the model in according epoch as a confounder set.

the orange line in Fig. 4, the percentage of hard negative drops rapidly when the training first starts due to the random initialization of classifier. Then, the decline meets a sudden slowdown, and at this stage, the model employed as a confounder set achieves the best performance. After a short pause, the percentage begins to drop again, and if using the model in this period, the performance will continue to drop as well. By the contrast between the two curves, we shed some light on why we choose the incompletely trained model as the confounder set.

The implementation details of $f_s(x, M)$. When we obtain the image-specific context feature M, whether our method should compute the final effect by $x - M$ is open to question. Context is not always bad, and different samples may need different degrees of context information. Therefore, it seems better to add more networks to learn how much we need from the context. The ablation studies are shown in Table 6, where we try some more complicated structures. Note that the increased 1 M is from the context classifier. The results indicate that although a simple minus is good enough, extra structures do bring about some improvement. Whether it merits exchanging extra computation against potential enhancement of performance depends on self-selection.

5.4. Qualitative Results

To further illustrate the advantage and disadvantage of our method, we visualize the activation maps via Grad-CAM [33]. As is shown in Fig. 5, for objects appearing in an unusual scene, the baseline fails to locate the target exactly. And for a nonexistent object, the baseline which activates more on the contextual regions, tends to make mistakes in a familiar context. Contrastively, our model focuses on the direct effect, *i.e.*, the instances themselves. Hence, given the ambiguous context, our model is qualified to find the most discriminative regions.

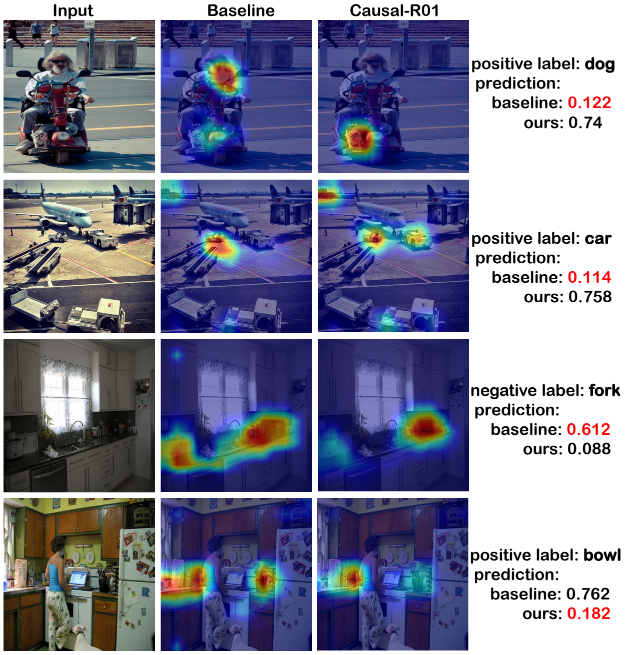


Figure 5. Visualization for the feature map of baseline and our Causal-R101 using the CAM [33].

In the last row, we also state the weakness of our method. When the instance is too obscure to be found, the model has no choice but to seek evidence from the context. In such case, our method fails to make correct prediction. Potential solutions include stronger backbones and more fine-grained locations of objects (*e.g.*, visual attention). Moreover, as is mentioned in Section 1, context plays an important role in some visual tasks, therefore, more works are desired to discriminate when the context is harmful to the recognition, and more experiments are expected for generalizing our approach to wider domains of visual learning and recognition.

6. Conclusion

In this paper, we present a simple, adaptive and powerful causal context debiasing recognition framework. We first uncover the damage caused by context bias and propose a structural causal model depicting the causalities in multi-label tasks. Then, by the combination of the causal intervention and the counterfactual training, we elegantly remove the effect of contextual bias through a simple minus without any increasement of learning parameters. Extensive experiments on four widely used multi-label classification datasets convey that our method has apparent advantage over the state-of-the-art on different datasets and different backbones with better performance and less computation.

Acknowledgements. This work was supported by National Natural Science Foundation of China (No. 62172021).

*: equal contribution

References

[1] Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3):399–424, 2011. 4

[2] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):1–37, 2011. 3, 5

[3] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 522–531, 2019. 3, 6, 7

[4] Zhao-Min Chen, Xiu-Shen Wei, Xin Jin, and Yanwen Guo. Multi-label image recognition with joint class-aware map disentangling and label correlation embedding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 622–627. IEEE, 2019. 6

[5] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2019. 1, 3, 5, 6, 7

[6] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009. 2, 5, 7

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 6

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021. 6

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 6

[10] Thibaut Durand, Nazanin Mehrasa, and Greg Mori. Learning a deep convnet for multi-label classification with partial labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 647–657, 2019. 1, 3, 5, 7

[11] Thomas KP Egglin and Alvan R Feinstein. Context bias: a problem in diagnostic radiology. *Jama*, 276(21):1752–1755, 1996. 1

[12] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 5, 7

[13] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2018. 1

[14] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4367–4375, 2018. 4

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6

[16] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *stat*, 1050:9, 2015. 5

[17] Dat Huynh and Ehsan Elhamifar. A shared multi-attention framework for multi-label zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8776–8786, 2020. 3

[18] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2019. 1

[19] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 16478–16488, 2021. 1, 3, 6

[20] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006. 4

[21] Qiang Li, Maoying Qiao, Wei Bian, and Dacheng Tao. Conditional graphical lasso for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2977–2986, 2016. 3

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 6

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European conference on computer vision (ECCV)*, pages 740–755. Springer, 2014. 1, 2, 5, 6

[24] Yongcheng Liu, Lu Sheng, Jing Shao, Junjie Yan, Shiming Xiang, and Chunhong Pan. Multi-label image classification via knowledge distillation from weakly-supervised detection. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 700–708, 2018. 6, 7

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. pages 10012–10022, 2021. 6

[26] J PEARL. Direct and indirect effects. In *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence, 2001*, pages 411–420. Morgan Kaufman, 2001. 4

[27] Judea Pearl. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146, 2009. 4

[28] Judea Pearl. Interpretation and identification of causal mediation. *Psychological methods*, 19(4):459, 2014. 2, 3

[29] Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19, 2000. 2, 3

[30] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 82–91, 2021. 1, 6, 7

[31] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000. 5

[32] Constantine Sedikides, W Keith Campbell, Glenn D Reeder, and Andrew J Elliot. The self-serving bias in relational context. *Journal of Personality and Social Psychology*, 74(2):378, 1998. 1

[33] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 8

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015. 7

[35] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in Neural Information Processing Systems*, 33, 2020. 3

[36] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3716–3725, 2020. 1, 3

[37] Matthew A Turk and Alex P Pentland. Face recognition using eigenfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–587, 1991. 5

[38] Tyler VanderWeele. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015. 4

[39] Tyler J VanderWeele. A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology (Cambridge, Mass.)*, 24(2):224, 2013. 4

[40] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2285–2294, 2016. 1, 3

[41] Meng Wang, Changzhi Luo, Richang Hong, Jinhui Tang, and Jiashi Feng. Beyond object proposals: Random crop pooling for multi-label image recognition. *IEEE Transactions on Image Processing*, 25(12):5678–5688, 2016. 6, 7

[42] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10760–10770, 2020. 3

[43] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 464–472, 2017. 3

[44] Yunchao Wei, Wei Xia, Min Lin, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. Hcp: A flexible cnn framework for multi-label image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1901–1907, 2015. 7

[45] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5419, 2017. 1

[46] Vacit Oguz Yazici, Abel Gonzalez-Garcia, Arnau Ramisa, Bartlomiej Twardowski, and Joost van de Weijer. Orderless recurrent models for multi-label classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 13440–13449, 2020. 1, 3

[47] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *Proceedings of the European conference on computer vision (ECCV)*, pages 649–665. Springer, 2020. 1, 3, 5, 6, 7

[48] Renchun You, Zhiyao Guo, Lei Cui, Xiang Long, Yingze Bao, and Shilei Wen. Cross-modality attention with semantic graph embedding for multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12709–12716, 2020. 6, 7

[49] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. 33, 2020. 1, 3

[50] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 818–833. Springer, 2014. 5

[51] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5831–5840, 2018. 5

[52] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. 33, 2020. 1, 3

[53] Jiawei Zhao, Ke Yan, Yifan Zhao, Xiaowei Guo, Feiyue Huang, and Jia Li. Transformer-based dual relation graph for multi-label image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 163–172, 2021. 1, 3

[54] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 5

[55] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5513–5522, 2017. 3, 7