1-2023

Causal interventional training for image recognition

Wei QIN
Hefei University of Technology

Hanwang ZHANG
Nanyang Technological University

Richang HONG
Hefei University of Technology

Ee-Peng LIM
Singapore Management University, eplim@smu.edu.sg

Qianru SUN
Singapore Management University, qianrusun@smu.edu.sg

Causal Interventional Training for Image Recognition

Wei Qin, Hanwang Zhang, Richang Hong*, Ee-Peng Lim and Qianru Sun

***Abstract*—Deep learning models often fit undesired dataset bias in training. In this paper, we formulate the bias using *causal inference*, which helps us uncover the ever-elusive causalities among the key factors in training, and thus pursue the desired causal effect without the bias. We start from revisiting the process of building a visual recognition system, and then propose a structural causal model (SCM) for the key variables involved in dataset collection and recognition model: object, common sense, bias, context, and label prediction. Based on the SCM, one can observe that there are "good" and "bad" biases. Intuitively, in the image where a car is driving on a high way in a desert, the "good" bias denoting the common-sense context is the highway, and the "bad" bias accounting for the noisy context factor is the desert. We tackle this problem with a novel causal interventional training (`CIT`) approach, where we control the *observed* context in each object class. We offer theoretical justifications for `CIT` and validate it with extensive classification experiments on CIFAR-10, CIFAR-100 and ImageNet, *e.g.*, surpassing the standard deep neural networks ResNet-34 and ResNet-50, respectively, by 0.95% and 0.70% accuracies on the ImageNet. Our code is open-sourced on the GitHub https://github.com/qinwei-hfut/CIT.**

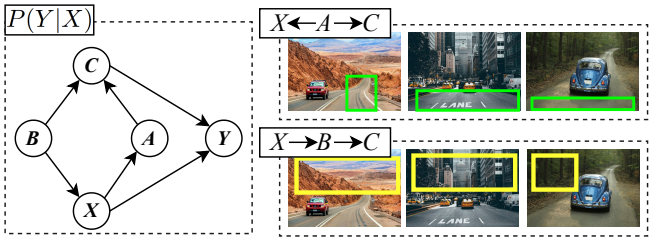
***Index Terms*—Image recognition, causality, causal intervention, deep learning, ImageNet**

I. Introduction

Deep neural networks (DNNs) achieve the state-of-the-art performance in many tasks [1]–[5]. Since deep neural networks are driven by data, biased data inevitably cause biased models, resulting in poor generalization for test domains [6]. To confront with the bias, unbiased training is proposed to directly compensate the bias effect, *e.g.*, jitter or flip the images for data augmentation [7], [8], batch normalization for stable mean and variance [9], neuron dropout for robust features [10], and re-weighting for balanced sample loss [11]–[13], just to name a few. Meanwhile, we do find that some of the bias types, such as visual contexts, are essentially good for different tasks [14], [15], *e.g.*, an image with a highway definitely increases the probability of *car* or *truck* and decreases that of *lion* or *fish*. In fact, there are evidences showing that removing such "good" bias indeed hurts the model performance [16], as the "good" bias has a high probability to appear in test cases. However, how to distinguish the "good" from the "bad" at the training stage still remains open.

In this paper, we aim to pursue the desired model trained with the "good" and without the "bad". We start from revisiting the fundamental process of building a visual recognition system.

Wei Qin and **Richang Hong** are with Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Ministry of Education, China and Institute of Artificial Intelligence, Hefei Comprehensive National Science Center. E-mail: {qinwei.hfut,hongrc.hfut}@gmail.com. Richang Hong is the corresponding author. **Hanwang Zhang** is with School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798. E-mail:hanwangzhang@ntu.edu.sg. **Ee-Peng Lim** and **Qianru Sun** are with the School of Information Systems, Singapore Management University, Singapore 178902. E-mail: {eplim, qianrusun}@smu.edu.sg



(a) our proposed SCM (b) *car* samples

X: object content ***A***: common sense ***B***: noisy context factor
C: object context ***Y***: prediction (label)

Fig. 1: Our modeling and interpretation. (a) shows our causal assumption in the format of structural causal model (SCM) specifically proposed for the task of image recognition. (b) shows some examples corresponding to two important SCM paths from object X to context C: one through "bad" bias B and the other one through "good" bias A. In specific, (b) shows that *on a highway* is a common-sense background factor for the object class *car* and *the desert* or *the CBD* are noisy background factors.

Specifically, we build a structural causal model (SCM) [17] for the key variables involved in dataset collection and its corresponding recognition model: object, common sense, bias, context, and label prediction. SCM represents the causal assumption for any image recognition models driven by large-scale training datasets. As illustrated in Figure 1(a), our assumption can be detailed as: (i) object content X directly yields the prediction Y. This is the essential causality we want to learn in image recognition; (ii) X has its context C, which is mediated by class-specific common sense denoted as A. This is the "good" bias; and (iii) dataset bias B confounds the association between X and its context C. This results in the "bad" bias effect. We justify that (i) and (ii) indicate the "good" correlation among X, C and Y, *e.g.*, Y = *car*; X = visual content of *car*; and C = visual context of *car* such as *on a highway in a desert* (*SUV car*) and *on a highway in a CBD* (*sports car*), see the Figure 1(b). We highlight that (iii) is due to the reporting bias (selection bias [8]) or the natural noise distribution, which are inevitable in dataset collection. As [18] claimed, objects features have causal relations with some context features but have no causal relations with others. Examples on the first row of Figure 1(b) show the common-sense context factors that the car has a causal relation with the highway. Examples on the second row of Figure 1(b) reveal the noisy context factors that the car has no causal relation with the desert or the CBD. The model should infer there may be a car based on cars' visual features or highways' features but not the deserts' features. We will elaborate the SCM with

formal definitions in Section *Structure Causal Model*.

Instead of training $P(Y|X)$, we leverage causal inference that is to learn $P(Y|do(X))$ where the *do*-calculus means the pursuit of the true causality between X and Y without the confounding effect [17]. The ideal way to $P(Y|do(X))$ is by intervening X physically (*a.k.a.* random controlled trial [19]), *e.g.*, if we could curate the data of any *car* in any context, we can train a unbiased classifier $P(Y = fish|do(X))$. As this is not applicable in given datasets, we thus propose an approximate approach called causal interventional training (CIT). The key idea of CIT is interventional data sampling. More specifically, we achieve a "virtual" intervention for X by fixing it as x and sample contexts c given x, to cut off the backdoor path involving C. In particular, the sample is conditional on the object class, *e.g.*, for any x of *car*, we sample from the context subset appearing in the *car* class only. Finally, we use sampled data to train $P(Y|do(X))$. We provide a theoretical justification for CIT (in Section *Causal Intervention by Backdoor Adjustment*) based on causal intervention theories [17]. We also devise two implementations in order to plug the CIT in deep neural networks (DNNs) and on large-scale training datasets, e.g., the ImageNet including millions of samples. We will elaborate with the details in Section *Experiments on the ImageNet Dataset*.

In summary, our main contributions are three-fold. (i) We propose a novel structural causal model for object recognition, demonstrating that the evil of dataset bias is the effect it causes by confounding content and context. (ii) We propose a novel CIT approach, and theoretically justify it can achieve the same performance of mitigating confounding effect as if it were able to physically control the bias. (iii) We introduce two implementation designs to plug CIT in state-of-the-art DNNs and train unbiased image classifiers on large-scale datasets, *i.e.*, CIFAR-10, CIFAR-100 [20], and ImageNet [21].

II. Related Work

Causality in Multimedia. Causal inference has been successfully applied in psychology, politics and epidemiology [22], [23]. In recent years, researchers try to introduce causal inference into multimedia tasks and computer vision tasks. To estimate temporary movement in video images, [24] introduced a novel flow extraction approach called causal flow, which can estimate the dominant causal relationships among nearby pixels. [25] used counterfactual examples to explain the learning behaviors of image classifiers. [26] introduced Invariant Risk Minimization to train robust models against spurious correlations. [16] proposed a counterfactual training framework to learn the unbiased scene graph from biased training data. [27] introduced causal intervention into visual representation learning, its backdoor adjustment is a soft-attention approximation under a quite strong assumption, i.e., the confounder inventory is well-established and fully observable. **In this paper, we relax it** to partially observable (visible), and we prove that it can be recovered by sampling in SGD.

To address the problem of domain shift, [28] use a single causal-based underlying system to model the difference of the distributions for the source (or training) domain(s) and target (or test) domain(s). [29] built a synthetic Visual Question Answering (VQA) dataset to quantify spurious correlations learned in VQA models. [30] regarded the goal of providing explanations for the decisions of machine-learning models as a causal learning task, and exploited the causal explanation (CXPlain) models that learn to evaluate to what degree certain inputs have a causal effect on outputs in another machine-learning model. [31] theoretically analyzed self-supervised representation learning using a causal framework and proposed a novel self-supervised objective– Representation Learning via Invariant Causal Mechanisms. More recently, [32], [33] proposed to tackle weakly-supervised semantic segmentation, respectively, by building SCMs for them. [34] used a causal perspective to reformulate the compositional zero-shot recognition. In specific, they formalized inference as a problem of finding the most likely intervention. [35] investigated what role do the regularizing terms play in standard regression tasks from the perspective of causal. [36] provided a causal perspective on representation learning which covers disentanglement and domain shift robustness as special cases. [37], [38] proposed a causality-inspired framework that builds structural causal model to capture the true effect of query and video content on the prediction. [39] argued that causal concepts can be used to explain the success of data augmentation by describing how they can weaken the spurious correlation between the observed domains and the task labels. [40] mitigated the noisy factors in feature space rather than example space. Different from our belief that the background contains both common-sense factors and noisy factors, [41] argued that context features should not be relied upon in classification tasks. Mitigating all information from context features may be more robust, but at the expense of performance. Although both [18] and our paper have isolated content and context of images, the investigating purposes are totally different. They try to discover the causal relation between objects in a single image. In specific, they propose a new task, distinguishing the object features from the context features, to empirically support their hypothesises. Therefore, they employ an object segmentation algorithm to isolate all possible objects in the image while our *CIT* uses a saliency detection model to isolate the single main object in the image. Compared with our work, [18] neglected what role the class and the predicted class play in the causal graph and did not involve how to recognize the object in the image. However, [18] supported the hypothesis of our work that objects features have causal relations with some context features but have no causal relations with others **Our work is the first one** to make the causal assumption and propose a new training solution for the most generic task in computer vision — image recognition.

Data Augmentation. The implementation of our CIT approach can be regarded as data augmentation. Typical data augmentation methods are random cropping, color augmentation [42] and resizing images [43]. Recently, there are some advanced data augmentation methods. For example, [44] trained a neural network on convex combinations of pairs of examples and their labels. [45] cut and pasted patches among training images whose ground truth labels are from the labels mixed proportionally to the area of the patches. [46] presented a

novel multiview-interpolation framework for wide-baseline camera arrays. To use the high-level semantics in videos, [47] augmented each frame representation with its context information. To solve the domain shift between training data and test data, [48] generated heterogeneous training images by mutually transforming the cross-modality differences and incorporating synthesized images into the learning process. [49] introduced a framework based on the deep convolutional generative adversarial networks for generating training images to augment the training set in order to improve the performance. **Our `CIT` is different from them** as it is motivated by causal intervention which is fully explainable, while the related methods are more heuristic. Besides, our work focuses on the relationship between object content and context (common-sense context and biased context), while the related works produce "new classes" using the mixup of existing classes to augment the data. More recent data augmentation methods [29], [50]–[53] are based on the counterfactual-level causal analysis. **We highlight** that according to Judea Pearl's causal ladder [17], our interventional level underpins the counterfactual level, i.e., the latter requires a well-trained SCM model to infer counterfactuals. Therefore, our work is fundamentally orthogonal to them and essentially supports them. Besides, generating counterfactual examples strongly relies on the performance of image generation models which are notoriously hard to train. We guess this is the reason why [50] and [51] showed only simple results on low-dimensional data space. [52] and [53] generated counterfactual examples using manually-designed methods.

De-bias in Image Data. Previous works [14], [54], [55] mentioned that image contexts are important visual cues for object recognition but sometimes undesirable. [56] and [57] removed context bias through adversarial learning. [58] regarded co-occurring objects in the image as context bias in the multi-label classification tasks. They thus proposed to reduce the co-occurrence. **In this paper, we define the "good" and "bad" of the context in a more general causal perspective.** We apply causality techniques (i.e., interventional training) to enable the data themselves to de-bias. Our approach can be potentially applied in a wider range of image recognition tasks.

III. Causal Interventional Training

The dataset bias B misleads the correlation between input image X and output label Y, leading to unrobust classifiers $P(Y|X)$. To model this causality, we introduce our causal assumption represented by SCM in Section *Structural Causal Model*. Based on the SCM, we point out that the key is to mitigate the confounded effect of X on Y by B. To achieve that, we propose a novel approach `CIT`, based on the general causal intervention tool — backdoor adjustment [17], in Section *Causal Intervention by Backdoor Adjustment*. Finally, in Section *Causal Interventional Training*, we detail the implementation steps of applying `CIT` to learn unbiased object classifiers $P(Y|do(X))$.

A. Structural Causal Model

The SCM in Figure 1(a) represents the causalities among the key variables in a general object recognition system: image X, label Y, context C, common sense A and bias B. Each arrow denotes the cause-effect relationship between two nodes. In the follows, we detail the underline rationale behind SCM.

$\boldsymbol{X \rightarrow A \rightarrow C}$. The context C pictures the surrounding (and/or object attributes) of the content X. The association between X and C is mediated by the latent variable of knowledge, we call it common sense A. This affects the result of data collection. For example, dataset creators tend to take images of *marine fish* particularly in the scenes of *ocean*, based on a common sense that *marine fish live in ocean*.

$\boldsymbol{X \leftarrow B \rightarrow C}$. We use a latent variable B to denote the bias which leads to negative effects. B could be of different kinds, *e.g.*, selection bias introduced by data annotators or collection tools [8], as shown in Figure 2. In most cases, B is not visible or not easy to be disentangled from the data. B confounds another correlation between X and C, which is different from the above mediation (via A). The former is "bad" as it hurts the generalization ability of the model (*i.e.*, makes the model fit to biased factors [8]), while the latter is "good" to the model. Therefore, de-biasing is to mitigate the confounding correlation caused by B. As illustrated in Figure 2, this is equivalent to $do(X)$ which cuts the path from B to X [17]. A plausible realization of this "cut" is given in Section *Causal Interventional Training*.

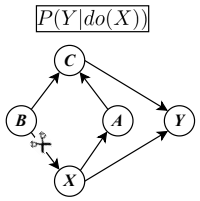


Fig. 2: The intervened SCM where we only cut the path involving "bad" bias B.

$\boldsymbol{C \rightarrow Y \leftarrow X}$. We use Y to denote the label space or the prediction of trained models. On the intervened SCM shown in Figure 2, Y is determined by X via two causal paths: (i) the direct $X \rightarrow Y$ and (ii) the mediated $X \rightarrow A \rightarrow C \rightarrow Y$ (as "cut" is applied to block $X \leftarrow B \rightarrow C \rightarrow Y$). The first path is essential. The second path (through C) is inevitable in object recognition, as object itself is contextualized, *e.g.*, the appearance of the fish *on plates* is *cooked*.

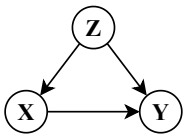


Fig. 3: A classic three-variable SCM that contains a confounder (Z).

So far, we have pinpointed the roles of bias B and common sense A played in object recognition. We understand in the conventional $P(Y|X)$, the prediction of Y given X is not only due to "X causes Y" via $X \rightarrow Y$ and $X \rightarrow A \rightarrow C \rightarrow Y$, but also the undesirable $X \leftarrow B \rightarrow C \rightarrow Y$. In the language of causal inference, the undesirable path is called backdoor [59]. To "cut" it, the general solution is backdoor adjustment, *i.e.*, using $P(Y|do(X))$ instead of $P(Y|X)$ as the training objective. Next, we elaborate how we apply this solution to image recognition in our approach.

B. The Backdoor Criterion and The Backdoor Adjustment

In Figure 3, we illustrate the most fundamental confounding case using the classical three-variable SCM, where Z denotes the confounder [60], [61]. Based on the law of total probability, it is easy to get

$$P(Y|X) = \sum_{Z} P(Y|X,Z)P(Z|X), \tag{1}$$

which includes both the direct causal effect of X on Y and the correlation confounded by Z. Taking this three-variable SCM as an example, we then introduce how to implement the causal intervention $P(Y|do(X))$ by using the conventional backdoor adjustment [62], [63].

Definition 1: (**The Backdoor Criterion**) Given a pair of variables (X, Y) in a directed acyclic graph G, a variable Z satisfies the backdoor criterion with respect to (X, Y), if (i) no node in Z is a descendant of X, and (ii) Z blocks every path between X and Y which contains an arrow into X.

If Z satisfies **The Backdoor Criterion** for (X, Y), backdoor adjustment is to replace $P(Z|X)$ with $P(Z)$, which yields the true causal effect of X on Y, *i.e.*, mitigates the confounded effect of Z. The formulation is thus as follows,

$$P(Y|do(X)) = \sum_{Z} P(Y|X,Z)P(Z). \tag{2}$$

Please note that in the original definition, Z could denote a set of variables. Here, we use a single variable (in Z) as an example.

C. Causal Intervention by Backdoor Adjustment

Preliminary. Based on the SCM in Figure 1(a), the correlation between X and Y in the conventional classifier can be formulated as:

$$P(Y|X) = \sum_{C}\sum_{A}\sum_{B} P(Y|X,C,A,B)P(C|A,B,X) \\ P(A|X)P(B|X), \tag{3}$$

which accords to the law of total probability. $P(Y|X)$ is inferior due to the effect of confounder B.

Our Approach. We propose the causal intervention by using $P(Y|do(X))$ as the new classifier, which explicitly "removes" B to achieve the true causality between X and Y. Specifically, we apply backdoor adjustment. As shown in Figure 2, *do*-Calculus "cuts" the pathway from B to X. To formulate it, we replace $P(B|X)$ with $P(B)$ in Eq. 3, and obtain:

$$P(Y|do(X)) = \sum_{C}\sum_{A}\sum_{B} P(Y|X,C,A,B) \\ P(C|A,B,X)P(A|X)P(B) \tag{4}$$

$$= \sum_{C}\sum_{A}\sum_{B} P(Y|X,C)P(C|A,B) \\ P(A|X)P(B) \tag{5}$$

$$= \sum_{C}\sum_{B} P(Y|X,C)P(C|A=a,B)P(B) \tag{6}$$

Thanks to the rule 1 of *do*-Calculus [64], A, B do not affect Y directly and X does not affect C, so $P(Y|X,C,A,B)$ and $P(C|A,B,X)$ in Eq. 4 can be replaced with $P(Y|X,C)$ and $P(C|A,B)$, respectively, yielding Eq. 5. As A is determined by the common sense knowledge of X where X is fixed, A can be specified and fixed as a, yielding Eq. 6. Then, we elaborate the case (rare) when physical intervention is applicable, followed by our approximate solution for the other case (most) when physical intervention is not applicable.

Stratify B. If assume B is visible, the physical intervention is to stratify B that produces an integrated set of contexts $\mathcal{C}_x$ (for any given content x) using every value of B. For example, if B denotes data annotators' preference, the annotator labels can be used as the values of B [65]. Using these labels, we can search the context set $\mathcal{C}_x$ introduced by all annotators and use them to contextualize x. We use the resulted images $X_{\mathcal{C}}$ to train $P(Y|do(X))$ as in Eq. 6.

However in many cases, B is not visible or not easy to disentangle from the real dataset, *e.g.*, ImageNet [21]. In other words, **Stratify** B is not always applicable. Fortunately, by leveraging causal inference, we are able to propose an equivalent solution called **Stratify** C **conditional on** A. In specific, if accumulate B in Eq. 6, we get:

$$P(Y|do(X)) = \sum_{C} P(Y|X,C)P(C|A=a). \tag{7}$$

It is worth highlighting that this "accumulate B" requires the independence between A and B, which can be guaranteed by applying D-Separation [66] on our proposed SCM (shown in Figure 1(a)): if $B \to X \to A$ exists, A and B are independent conditional on X.

Stratify C **conditional on** A is our solution for Eq. 7, by which we can learn the classifier $P(Y|do(X))$. This solution is based on the assumption that $A = a$ is visible given x. We approximate it by sampling context images $\mathcal{C}_x$ from the class of x. Our intuition is obvious. For example, if x is an example of *marine fish*, a is the common sense that *marine fish lives in ocean water*. So, the contexts in the class of *marine fish* (mostly *in ocean water*) are the most suitable samples for $\mathcal{C}_x$ to contextualize x. Similar to **Stratify** B, we can use the contextualized images to learn the classifier $P(Y|do(X))$.

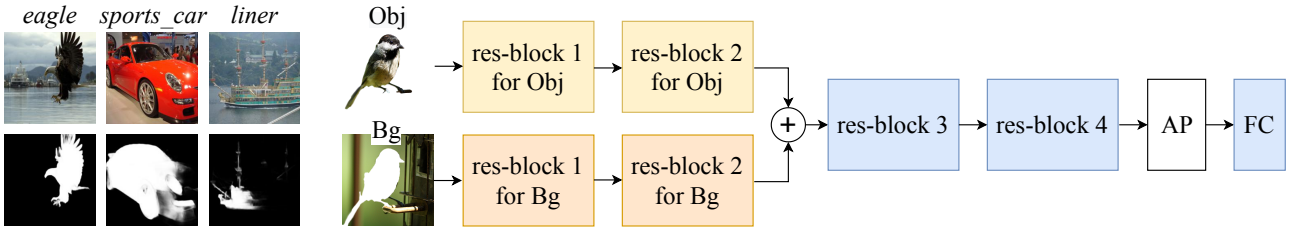
D. Causal Interventional Training

In this section, we introduce our causal interventional training (`CIT`) on real large-scale image datasets. In the general sense, there are three steps. Step 1 is to disentangle the original image I into content x_I and context c_I. Step 2 is to generate the context set $\mathcal{C}_x$ given every specific content x_I. Step 3 is to contextualize x_I using $\mathcal{C}_x$ to produce new data to train $P(Y|do(X))$. As aforementioned, we need to detail Step 2 in two cases regarding B is visible or not. We provide the pseudo code of these 3 steps in Algorithm 1.

Step 1: Given an image I, we use a function g to isolates the content and context as:

$$x_I, c_I = g(I), \tag{8}$$

where g can be (i) a pre-trained model, *e.g.*, of saliency detection [67], and (ii) a pre-defined pattern such as using colorized the backgrounds for specific classes, where the color bias is easy to recognize by the classifier itself without needing additional models. In experiments, we apply these two kinds of



(a) image samples & saliency maps

(b) double-branch neural networks (ResNet-50 or ResNet-34 as backbone)

Fig. 4: We show three ImageNet samples with their saliency detection results in (a). We place the best-quality one on the leftmost and the worse ones on the rightmost. We show the double-branch network architecture (used for ImageNet) in (b). Input layer is included in res-block 1. "res-" stands for the residual architecture, "AP" denotes the average pooling, and "FC" is the fully-connected layer.

g for ImageNet and CIFAR datasets, respectively. **Step 2:** We need to formulate two cases separately. First, if B is visible, we use **Stratify** B to generate $\mathcal{C}_{x_I}$. Let h denote the generating process of $B \rightarrow C \leftarrow A$. It uses A as a specific value a_{x_I} conditional on x_I, and then stratify all values of B (denoted as $\{b_i\}$) using the prior probability of b_i (denoted as $P(b_i)$). Therefore, we have:

$$\mathcal{C}_{x_I} = \{h(a_{x_I}, b_i) \cdot P(b_i)\}, \tag{9}$$

where $i \in \{1, 2, ..., N\}$, and $P(\cdot)$ denotes the prior distribution. Function $h(a_{x_I}, b_i)$ means using common-sense background factor a_{x_I} and noisy background factor b_i to create a new background image. Second, if B is not visible, **Stratify** C **conditional on** A is used to sample $\mathcal{C}_x$ as follows:

$$\mathcal{C}_{x_I} = \{c_{x_i}\}, \tag{10}$$

where $a_{x_i} = a_{x_I}$ meaning that x_i and x_I are in the same class (i.e., have the same common sense knowledge). **Step 3:** Contextualizing content x_I is to pair it with each specific sample c_j in $\mathcal{C}_{x_I}$. Using contextualized data to train the classifier is to minimize the following loss function:

$$\mathcal{L} = \sum_{j=1}^{M} L_{ce}(f(x_I, c_j), y_I), \tag{11}$$

where $c_j \in \mathcal{C}_{x_I}$, y_I is the ground truth label, M is the size of $\mathcal{C}_{x_I}$, L_{ce} denotes the cross-entropy loss function, and f represents the classifier. In this way, we have to learn M times of data and require M times of computation costs, which is implausible on the real large-scale datasets. We tackle this problem by introducing a new implementation method called *epoch-wise data augmentation*. In each epoch, we learn the model using the original data as well as the same number of new contextualized data. Both data contain the same set of object content. The objective is thus to minimize:

$$\mathcal{L}_j = L_{ce}(f(x_I, c_j), y_I) + L_{ce}(f(x_I, c_I), y_I), \tag{12}$$

where $\mathcal{L}_j$ is the loss at the j-th epoch, c_I is the original context of x_I and $c_j \in \mathcal{C}_{x_I}$. Empirically, we observe the model can converge at the T-th epoch where $T \ll M$.

IV. Experiments on the ImageNet Dataset

In this section, we evaluate one of our `CIT` approaches — **Stratify** C **conditional on** A on ImageNet where the bias is not visible. Below we introduce the dataset, implementation details, and comparison methods, followed by the result analysis and the interpretation of our approach.

Datasets. ImageNet 2012 [21] is a large-scale image dataset consisting of $1,000$ classes. It has $1.28M$ samples for training and $50K$ for validation. On this dataset, we employ a pre-trained saliency detection model [67] to separate the content and context for each image. The separation results are noisy (as shown in Figure 4(a)), but are still supportive to our evaluation process.

Implementation details. To realize `CIT`, i.e., by pairing different contexts to the content, we deploy two double-branch neural networks (based on ResNet-50 and ResNet-34). One branch is fed by content while the other for context. We show the architecture in Figure 4(b). We pre-train the standard ResNet and then use the learned weights to initialize the corresponding residual blocks in our double-branch networks. When training $P(Y|do(X))$, we fine-tune only two high-level blocks and the FC layer (blue blocks in Figure 4(b)). We use the SGD optimizer with Nesterov momentum [68]. Following the standard settings of ResNet [1], we set the mini-batch size to 128, the momentum to 0.9, the weight decay to $1e^{-4}$ and the initial learning rate to $1e^{-3}$. We drop the learning rate by 0.1 after every 16 epochs. We train all models for 32 epochs. Our code is based on the official PyTorch Hub. All our reporting numbers are obtained on the same validation set, and are averaged over three runs of experiments.

Comparison methods. Referring to Table I, we introduce the comparison methods. `Baseline` is the standard ResNet in the official PyTorch code. `Obj+Bg` is our implementation of a similar double-branch neural network proposed in [69]. The reason why we use a double-branch neural network is because our input contains two parts(context and content). One branch is fed with object region and the other is for background (context). `Sal IMG` is our implementation of another similar double-branch network proposed in [70]. Different from `Obj+Bg`, it inputs the whole image into one branch and the corresponding saliency map to the other branch. For both methods, we deploy

Algorithm 1 Causal Interventional Training (CIT)

STEP ONE: isolate images into content and context.
INPUT: training data $(\mathcal{I}, \mathcal{Y})$, function g that isolates the content and context of an image.
OUTPUT: isolated training data S_{iso}.

$S_{iso} = \{\}$
for I_i, y_i in $(\mathcal{I}, \mathcal{Y})$ **do**
  $x_i, c_i = g(I_i)$
  S_{iso}.ADD(x_i, c_i, y_i)
end for

STEP TWO: get the context set given each specific content.
INPUT: isolated training data S_{iso}, the function $a = f(x)$ that gets the mediator a from content x, the function $c = h(b, a)$ that simulates the process from a and b to c.
OUTPUT: intervened training data S_{itv} (images and labels).
If B is observable, we use **Stratify** B:

$S_{itv} = \{\}$
for x_i, c_i, y_i in S_{iso} **do**
  $a_{x_i} = f(x_i)$
  $\mathcal{C}_{x_i} = \{\}$
  for b_j in B **do**
    $\mathcal{C}_{x_i}$.ADD$(h(b_j, a_{x_i}))$
  end for
  S_{itv}.ADD$(x_i, \mathcal{C}_{x_i}, y_i)$
end for

If B is unobservable, we use **Stratify** C **conditional on** A:

$S_{itv} = \{\}$
for x_i, c_i, y_i in S_{iso} **do**
  $\mathcal{C}_{x_i} = \{\}$
  for x_j, c_j, y_j in S_{iso} **do**
    if $f(x_j) == f(x_i)$ **then** {On the ImageNet, this line is realized as "if $y_j == y_i$ then"}
      $\mathcal{C}_{x_i}$.ADD(c_j)
    end if
  end for
  S_{itv}.ADD$(x_i, \mathcal{C}_{x_i}, y_i)$
end for

STEP THREE: epoch-wise causal interventional training.
INPUT: intervened training data S_{itv}, model f_θ with parameters θ, learning rate λ.
OUTPUT: trained model f_θ with parameters θ.

for epc in Num_{epoch} **do**
  $S_{epc} = \{\}$
  for $(x_i, \mathcal{C}_{x_i}, y_i)$ in S_{itv} **do**
    c_{x_i} is randomly sampled from $\mathcal{C}_{x_i}$
    S_{epc}.ADD$((x_i, c_{x_i}, y_i))$
  end for
  for $X_{bat}, C_{bat}, Y_{bat}$ in S_{epc} **do**
    $\mathcal{L} = L_{ce}(f_\theta(X_{bat}, C_{bat}), Y_{bat})$
    $\theta = \theta - \lambda \cdot \partial\mathcal{L}/\partial\theta$
  end for
end for

TABLE I: Image classification accuracies (%) on the ImageNet. Top three lines are related works. Bottom shows the ablation study. In this table, $do(X)$ is general and represents any intervention conducted on the training data. * indicates using the standard ResNet (single branch).

Methods	ResNet-50	ResNet-34
`Baseline`* [1]	76.15	73.30
`Obj+Bg` [69]	76.43	73.97
`Sal+IMG` [70]	76.44	73.46
`2xBaseline`*	76.37	73.69
`Only Obj`*	73.45	70.20
`Only Bg`*	63.52	62.70
`Obj+all Bg`	71.79	69.02
`CIT`(ours)	**76.85**$_{+0.7}$	**74.25**$_{+0.95}$
`Obj+Bg`+Mixup [44]	77.65	75.31
`CIT`(ours)+Mixup	77.86$_{+1.71}$	75.60$_{+2.30}$
`Obj+Bg`+CutMix [45]	78.40	75.79
`CIT`(ours)+CutMix	78.64$_{+2.49}$	76.01$_{+2.71}$

the same architecture as in Figure 4(b). For ablation study, we have 4 settings: `2xBaseline` learns a baseline model by double times of training iterations. `Only Obj` learns a baseline model using the images with only object regions. `Only Bg` learns a baseline model using the images with only background regions. `Obj+all Bg` learns a double-branch model using the content paired with different contexts randomly sampled from the dataset. Last, we plug-in Mixup and CutMix (augmentation methods using the data of different classes) in our `CIT`. For fair comparison, we implement above methods use the same hyperparameters. Our *CIT* and the baseline methods have the same resource consumption.

A. Results and Analyses

In Table I, we demonstrate the overall results for the ablation study and the comparison to related works [1], [44], [45], [69], [70]. We detail our observations in the follows.

Context is important to object recognition. As we explained above (in Figure 1(a)), the context C is affected by both the common sense A of the class and unfortunately also by the dataset bias B. The latter B can be caused by a variety of invisible reasons. In the language of SCM, C not only bridges the "true" correlation between X and Y (through A), but also indirectly confounds their "spurious" correlation (through B). As the results shown in Table I, if we simply remove the value of context C from the input data as in `Only Obj`, the performance of the resulting model is significantly reduced (by about 3% accuracy) compared to `Baseline`. The reason is that the image of any object is intrinsically contextualized, and its representation will be degraded if the context is broken. This can be indirectly proved by using only context for object recognition. For example, the ResNet-50 model of `Only Bg` achieves the accuracy of 63.52%, which is significantly higher than the chance rate (0.1% on the 1,000-class ImageNet). Note that for this 63.52%, another inevitable cause is: `Only Bg` is trained on the context images containing the object shapes (due to saliency detection).

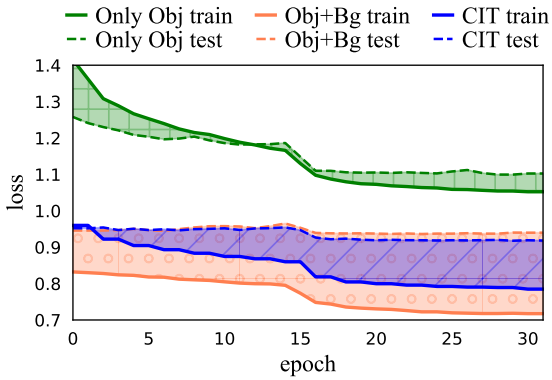


Fig. 5: Training and test loss curves on the ImageNet.

TABLE II: Test accuracies (%) with background images or object images as input. Background and object images (from one image) are separated by a pre-trained saliency detection model [67].

Backbone	Method	Test on bg images	Test on obj images
ResNet-34	Obj+Bg	20.82	52.04
	CIT	34.60$_{+13.78}$	61.91$_{+9.87}$
ResNet-50	Obj+Bg	22.11	56.41
	CIT	40.43$_{+18.32}$	65.94$_{+9.53}$

CIT performs the best for mitigating "bad" effects. The context C is important but it also indirectly causes the confounding effect (through B) causing harm to the model. Our CIT aims to eliminate such bad effect without needing to adjust B (not visible on the ImageNet). From Table I, we can see that Baseline achieves the accuracy of 76.15% on ResNet-50. Our CIT intervenes the input images and achieves the accuracy of 76.85%, the best performance over related methods (the top one achieves 76.44%) as well as the ablative models (the top one achieves 76.37%). This result is consistent for the models of ResNet-34, *e.g.*, our improvement over Baseline is more significant as 0.95%.

Other observations. From Table I, we have three additional observations. (1) 2xBaseline trained with double iterations[1] gains improvements, *e.g.*, 0.4% accuracy over Baseline ResNet-34. (2) The double-branch network architecture *per se* contributes to boosting the accuracy of image classifiers. Specifically, Obj+Bg [69] achieves the accuracy of 76.43% on ResNet-50 which is 0.3% higher than the result of Baseline [1]. The difference is that Obj+Bg is learned on the double-branch neural network. (3) Without our proposed "conditional on the common sense", the performance of $do(X)$ suffers from a considerable drop. Specifically, Obj+all Bg on both ResNet-34 and ResNet-50 are degraded by about 5% accuracies, compared to our results of using CIT. (4) The results on bottom two blocks show that our CIT can contribute additional improvements to the performance, on the shoulder of class mixup-based data augmentation methods.

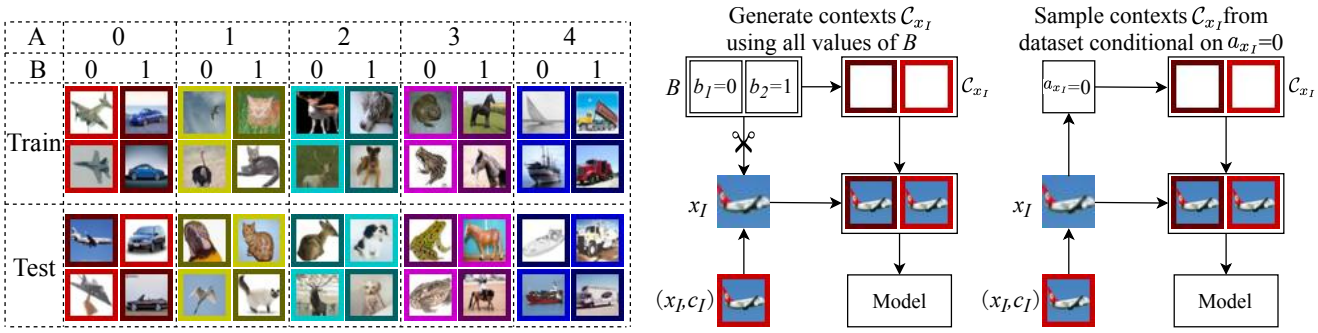
B. Interpretation

On the SCM of conventional $P(Y|X)$ (e.g., Baseline) there are three cause-effect relationships between X and Y: (i) X directly causes Y; (ii) X indirectly affects Y through mediators A and C where A is not visible and the mediating effect is reflected via the value of C; (iii) the unrobust indirect effect confounded by B (not visible) which is also reflected via the value of C. Our solution to mitigating (iii) is to train the new classifier $P(Y|do(X))$. In the following, we refer to two real models of $P(Y|do(X))$ (Only Obj and CIT) and compare their performance to that of $P(Y|X)$ (Baseline) while interpreting the realized causal effects in image recognition models. We particularly demonstrate their training and testing curves in Figure 5. **Direct causal effect generalizes the best.** We model the direct effect from X to Y by removing the context pixels from training images (identical test to ours). This corresponds to the model of Only Obj. Please note that this model can only approximate to the direct effect due to two reasons: (i) the saliency detector can not perfectly separate the content, and (ii) the content *per se* contains contextual information, *e.g.*, the lightness of the same object is different in different scenes. We have employed different saliency detection models. In a nutshell, better saliency detection technique will help us achieve better performance. From Figure 5, we can see that using Only Obj the resulted gap between training and test curves is the smallest.

This shows that Only Obj has the best generalization ability, which is definitely true as the direct effect from content X to prediction Y is the most essential causality learned by any image classifiers. If recall the results in Table I, we can see that the performance of Only Obj is poor, *i.e.*, 3.4% and 4% lower than our CIT on ResNet-50 and ResNet-34, respectively. This is also easy to explain. The contexts are caused not only by data bias but also by common sense knowledge which is of high representation (as the results of Only Bg are much higher than the chance rate). Therefore, Only Obj removing all context pixels can no longer learn any representation from the common sense related contexts.

CIT improves generalization and reduces overfitting. The input data of Obj+Bg is normal, so its resulting model learns all the correlations shown in the SCM of $P(Y|X)$ (Figure 1(a)). Obj+Bg uses the double-branch neural network to separate the input of content and context. Figure 5 shows that its performance gap between training and test (red curves) is the largest. After the 15-th epoch, the gap becomes more obvious. Therefore, its model generalization ability is the worst. Besides, its training loss is the lowest which reveals the serious problem of overfitting to training data. Compared to Obj+Bg, our CIT (using the same neural network architecture) achieves better generalization ability and suffers from less overfitting problem.

[1]Due to the constrains of lab GPU machines, we have to leave the time-consuming search of hyperparameters (*e.g.*, iterations) in the future work. We do the best to ensure the reporting results are as fair as possible, by following the standard setting of ResNets [1] and using the official Pytorch code.



(a) Illustration of our synthetic CIFAR-10 (b) Stratify B (c) Stratify C conditional on A

Fig. 6: We show the mapping among A, B, X, Y and C (using CIFAR-10 examples) in (a). Two classes with the same value of A are assigned with $B=0$ or $B=1$ *arbitrarily* in the training set but *randomly* in the test set. In (b) and (c), we demonstrate the computing flows of our `CIT` approaches.

C. Additional Experiments on ImageNet

As we mentioned, `CIT` can help the model to learn the "good" and avoid the "bad" bias. The trained `CIT` models are thus expected to have better understanding and representation abilities for the objects (content) as well as the backgrounds (context) in images. We compare such abilities to those of `Obj+Bg`, where `Obj+Bg` is a fair architecture based on baseline ResNets using exactly the same architecture and same hyperparameters to our `CIT`. These experiments are conducted on the ImageNet. We present the recognition accuracies of background images ("bg images") and object images ("obj images") in Table II.

Settings. We deploy the trained models using the comparable baseline `Obj+Bg` and our `CIT`, but set their testing images to contain only image content, i.e. the object pixels denoted as obj, or only image context, i.e. the background pixels denoted as bg. We note that using bg (or obj) means simply feeding the separated bg (or obj) images on the context (or content) branch of the model.

Observation and Conclusions. From Table II, we can see that our `CIT` achieves consistently and greatly better recognition performance than `Obj+Bg` with respect to the object classes (i.e., the labels of the input images), e.g. it gets improvement margins of over 18% and 9% with ResNet-50 for the testing of obj images and bg images, respectively. We can conclude that `CIT` enhances the machine models with a better understanding of both the content and the context of images.

Visualization of the intervened context images. We show some examples of content images and the corresponding intervened context images in Figure 7. We can observe that most intervened context images contain the "good" bias. For simplicity, we take the car in the first row as an example. As we discussed in Section I, road is the common-sense background factor in all elements of cars' background while the changing environments are noisy background factors. In our *CIT*, the car image in the first row is augmented with the intervened context images. The model will easily capture the correlation between the car and road while alleviating the overfitting to the desert or the river.

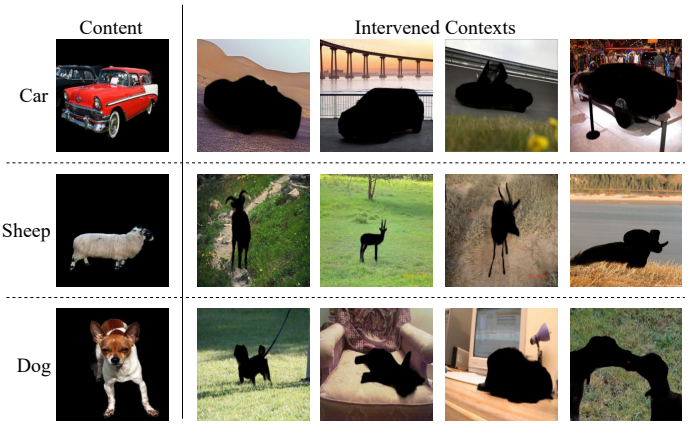


Fig. 7: We show some content images and their corresponding intervened context images.

V. Experiments on the CIFAR Datasets

Experiments on ImageNet validates the effectiveness of **Stratify** C **conditional on** A when "bad" bias B is not visible. In this section, we demonstrate the results of two `CIT` approaches, *i.e.*, both **Stratify** B and **Stratify** C **conditional on** A, by using the synthetic B (visible and controllable). Our synthesizing is based on the data of CIFAR datasets [20].

Synthetic datasets. CIFAR-10 (100) datasets [20] contains 10 (100) object classes and each with $5,000$ (500) samples for training and $1,000$ for test. The image size is 32×32. On each image, we add 4-pixel padding and change the value of the padded pixels to synthesize different contexts. Our contexts are changed according to two manual reasons: one is "biased" meaning the contexts are different between training and test sets; and the other one is "from common sense" meaning that the contexts are the same in training and test sets. We manually assign values to the nodes of SCM, which can be referred to illustration in Figure 6(a). We elaborate the details as follows. $\boldsymbol{X} \rightarrow \boldsymbol{A} \rightarrow \boldsymbol{C}$. X represents the original image on CIFAR. C denotes the synthetic context. A is the common sense integer determined by the object label of X. As shown in Figure 6(a), the value of A determines (the hue of) the color on synthetic C pixels, and it can be shared by more than one class, *e.g.*,

TABLE III: Four versions of settings designed for CIFAR datasets. "brig.(2)" on column 2 denotes the brightness on CIFAR-10 has 2 values corresponding to $|A| = 2$. Similarly, "hue(50)" on column 5 denotes the hue on CIFAR-100 has 50 values corresponding to $|B| = 50$.

Version	CIFAR-10		CIFAR-100	
	A determines	B determines	A determines	B determines
V1	brig.(2)&hue(5)	-	brig.(2)&hue(50)	-
V2	hue(5)	brig.(2)	hue(50)	brig.(2)
V3	brig.(2)	hue(5)	brig.(2)	hue(50)
V4	-	brig.(2)&hue(5)	-	brig.(2)&hue(50)

TABLE IV: Image classification accuracies (%) on the CIFAR datasets. "`on Ori.`" indicates training the models on original datasets. "`on Syn.`" indicates training the models on synthetic datasets. The last row are the differences between the model performances of **Stratify** B (ideal performance) and **Stratify** C **conditional on** A (which was proposed to handle the case when B is not visible in real-world datasets such as ImageNet). *V1-V4* are defined in Table III.

Training Methods	CIFAR-10				CIFAR-100			
	V1	*V2*	*V3*	*V4*	*V1*	*V2*	*V3*	*V4*
`Baseline` [1] `on Ori.`	91.92	91.92	91.92	91.92	68.01	68.01	68.01	68.01
`Baseline` [1] `on Syn.`	100	51.2	20.2	0.7	98.28	41.05	2.67	0.1
`CIT` - **Stratify** B	99.87	97.47	94.02	91.70	98.18	95.01	74.85	67.50
`CIT` - **Stratify** C **conditional on** A	99.81	97.49	94.03	91.76	98.15	95.03	74.83	67.52
|**Stratify** B - **Stratify** C **conditional on** A|	0.06	0.02	0.01	0.06	0.03	0.02	0.02	0.02

by two classes on the CIFAR-10. Similarly, we can generate the synthetic CIFAR-100. The only difference is that A on CIFAR-100 varies from 0 to 49, *i.e.*, generating contexts with a larger range of colors. $\boldsymbol{X} \leftarrow \boldsymbol{B} \rightarrow \boldsymbol{C}$. B is a bias integer. The value of B determines the brightness of the color on C pixels. Figure 6(a) shows how we apply B to intentively confound X and Y on CIFRA-10 (two classes with the same value of A are assigned with $B=0$ and $B=1$ *arbitrarily* in the training set but *randomly* in the test set). In this way, B introduces a correlation between X and Y (through C), which is totally random but not causal. It is also applied in the test set.

Using different integer settings for A and B, we generate 4 versions of synthetic data. *V1*: common sense only (B is null). *V2*: common sense takes a bigger proportion ($|A| > |B|$). *V3*: bias takes a bigger proportion ($|A| < |B|$). *V4*: bias only (A is null).

We elaborate the proposed four versions of synthetic datasets (taking either CIFAR-10 or CIFAR-100 as the data source). We mentioned that A is mediator, B is confounder, and they may affect C at different degrees. We thus can set different degrees for A and B and derive four version of synthetic datasets to evaluate our `CIT`. On version one (*V1*), A determines the context C and B has no effect on C. In specific, A determines both the brightness and the hue of the context pixels (the value of C) padded on the CIFAR images. On *V2*, A determines the hue of C and B determines the brightness. On *V3*, A determines the brightness and B determines the hue. On *V4*, A has no effect on C but B determines both the hue and the brightness of context pixels (the value of C). Detailed bins of brightness and hue on the CIFAR-10 and CIFAR-100 are given in Table III.

Implementation details. We deploy ResNet-20 [1], and use its official code in PyTorch [71]. Following [1], we use the SGD optimizer with Nesterov momentum [68], and set mini-batch size to 128, momentum to 0.9, weight decay to $1e^{-4}$ and initial learning rate to 0.1. We drop learning rates by 0.1 at the 80-th and 120-th training epochs. We train all models for 160 epochs. We illustrate our `CIT` approaches in Figure 6 (b) and (c), and show the results in Table IV.

`CIT` approaches achieve the "ideal performance". As aforementioned, if B is visible, we can achieve the "ideal performance" using **Stratify** B, *i.e.*, backdoor adjustment on B [17]. From the bottom two rows in Table IV, we see that the same "ideal performance" (with **Stratify** B) can be achieved by **Stratify** C **conditional on** A (which was proposed to handle the case when B is not visible). We thus empirically validate these two approaches perform equally. We are not surprised, because their formulations are derived from the same backdoor adjustment (in Eq.4). Please note that we used standard ResNet (single branch) for all methods on CIFAR, so it is not plausible to compare to `Obj+Bg` or `Sal+Img` (double-branch models designed for ImageNet).

`CIT` approaches mitigate the "bad" and preserve the "good". Row 1 in Table IV presents the original results of `Baseline` (no distinction from *V1* to *V4*). While on Row 2, if `Baseline` models are trained on "manually biased" data, the accuracy drops sharply, *e.g.*, from 91.92% to 0.7% (on the *V4* of CIFAR-10). In contrast, our `CIT` approaches do not drop at all, and even improve the results on both *V2* and *V3* — the general scenarios where both "good" and "bad" bias exist. It is because `CIT` can mitigate the "bad" and preserve the "good" of image contexts in the trained model.

VI. Conclusions

This paper presents a novel study of causality for the fundamental visual task — image recognition. We build the structural causal model (SCM) to demonstrate the cause-effect relationships among 5 key factors involved in the task. We propose a novel causal interventional training (CIT) approach to mitigate the bad effects caused by biased image contexts, but preserve the good parts to train better image classifiers. We achieve superior classification results on large-scale image benchmarks.

Acknowledgment

This research is supported by A*STAR under its AME YIRG Grant (Project No. A20E6c0101), National Research Foundation Singapore under its International Research Centres in Singapore Funding Initiative, and the National Key Research and Development Program of China under grant 2019YFA0706200, and in part by the National Natural Science Foundation of China under grant 61732007, 61932009, and Singapore MOE McRF Tier 2. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Eds., vol. 27. Curran Associates, Inc., 2014.

[3] H. Ben, Y. Pan, Y. Li, T. Yao, R. Hong, M. Wang, and T. Mei, "Unpaired image captioning with semantic-constrained self-learning," *IEEE Transactions on Multimedia*, pp. 1–1, 2021.

[4] X. Yang, P. Zhou, and M. Wang, "Person reidentification via structural deep metric learning," *IEEE transactions on neural networks and learning systems*, vol. 30, no. 10, pp. 2987–2998, 2018.

[5] X. Liu, X. Yang, M. Wang, and R. Hong, "Deep neighborhood component analysis for visual similarity modeling," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 3, pp. 1–15, 2020.

[6] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization," in *International Conference on Learning Representations*, 2017.

[7] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, p. 60, 2019.

[8] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1521–1528.

[9] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of International Conference on Machine Learning*, 2015, pp. 448–456.

[10] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, pp. 1929–1958, 2014.

[11] L. Jiang, Z. Zhou, T. Leung, L. Li, and L. Fei-Fei, "Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *Proceedings of International Conference on Machine Learning*, 2018, pp. 2309–2318.

[12] M. Ren, W. Zeng, B. Yang, and R. Urtasun, "Learning to reweight examples for robust deep learning," in *Proceedings of International Conference on Machine Learning*, 2018, pp. 4331–4340.

[13] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *Proceedings of International Conference on Machine Learning*, 2017, pp. 1885–1894.

[14] Z. Zhu, L. Xie, and A. L. Yuille, "Object recognition with and without objects," in *International Joint Conference on Artificial Intelligence*, 2017, pp. 3609–3615.

[15] Z. Liu, S. Wu, S. Jin, Q. Liu, S. Lu, R. Zimmermann, and L. Cheng, "Towards natural and accurate future motion prediction of humans and animals," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 004–10 012.

[16] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased scene graph generation from biased training," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3713–3722.

[17] J. Pearl *et al.*, "Causal inference in statistics: An overview," *Statistics surveys*, pp. 96–146, 2009.

[18] D. Lopez-Paz, R. Nishihara, S. Chintala, B. Scholkopf, and L. Bottou, "Discovering causal signals in images," in *CVPR*, 2017.

[19] J. Pearl, "Interpretation and identification of causal mediation," *Psychological methods*, vol. 19, pp. 459–481, 2014.

[20] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," *Master's thesis, Department of Computer Science, University of Toronto*, 2009.

[21] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, pp. 211–252, 2015.

[22] A. J. F. David P MacKinnon and M. S. Fritz., "Mediation analysis," *Annu. Rev. Psychol.*, 2007.

[23] R. B. Lorenzo Richiardi and D. Zugna., "Mediation analysis in epidemiology: methods, interpretation and bias." *International journal of epidemiology*, 2013.

[24] Y. Yamashita, T. Harada, and Y. Kuniyoshi, "Causal flow," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 619–629, 2012.

[25] D. Mahajan, C. Tan, and A. Sharma, "Preserving causal constraints in counterfactual explanations for machine learning classifiers," in *Workshop of Neural Information Processing Systems*, 2019.

[26] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," *arXiv:1907.02893*, 2019.

[27] T. Wang, J. Huang, H. Zhang, and Q. Sun, "Visual commonsense r-cnn," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 757–10 767.

[28] S. Magliacane, T. van Ommen, T. Claassen, S. Bongers, P. Versteeg, and J. M. Mooij, "Domain adaptation by using causal inference to predict invariant conditional distributions," in *Neural Information Processing Systems*, 2018, pp. 10 869–10 879.

[29] V. Agarwal, R. Shetty, and M. Fritz, "Towards causal VQA: revealing and reducing spurious correlations by invariant and covariant semantic editing," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9687–9695.

[30] P. Schwab and W. Karlen, "Cxplain: Causal explanations for model interpretation under uncertainty," in *Neural Information Processing Systems*, 2019, pp. 10 220–10 230.

[31] J. Mitrovic, B. McWilliams, J. Walker, L. Buesing, and C. Blundell, "Representation learning via invariant causal mechanisms," 2021.

[32] D. Zhang, H. Zhang, J. Tang, X.-S. Hua, and Q. Sun, "Causal intervention for weakly-supervised semantic segmentation," in *Neural Information Processing Systems*, 2020.

[33] F. Shao, Y. Luo, L. Zhang, L. Ye, S. Tang, Y. Yang, and J. Xiao, "Improving weakly-supervised object localization via causal intervention," *ACM Multimedia*, 2021.

[34] Y. Atzmon, F. Kreuk, U. Shalit, and G. Chechik, "A causal view of compositional zero-shot recognition," in *Neural Information Processing Systems*, 2020.

[35] D. Janzing, "Causal regularization," in *Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, 2019.

[36] R. Suter, D. Miladinovic, B. Schölkopf, and S. Bauer, "Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness," in *International Conference on Machine Learning*, 2019, pp. 6056–6065.

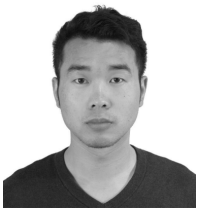
[37] X. Yang, F. Feng, W. Ji, M. Wang, and T.-S. Chua, "Deconfounded video moment retrieval with causal intervention," *SIGIR*, 2021.

[38] Y. Li, X. Yang, X. Shang, and T.-S. Chua, "Interventional video relation detection," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 4091–4099.

[39] M. Ilse, J. M. Tomczak, and P. Forré, "Designing data augmentation for simulating interventions," *ICML*, 2021.

[40] Y. Luo, P. Liu, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Category-level adversarial adaptation for semantic segmentation using purified features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[41] M. Prabhushankar and G. AlRegib, "Extracting causal visual features for limited label classification," *ICIP*, 2021.
[42] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Neural Information Processing Systems*, 2012, pp. 1106–1114.
[43] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.
[44] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *International Conference on Learning Representations*, 2018.
[45] S. Yun, D. Han, S. Chun, S. J. Oh, Y. Yoo, and J. Choe, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6022–6031.
[46] B. Ceulemans, S.-P. Lu, G. Lafruit, and A. Munteanu, "Robust multiview synthesis for wide-baseline camera arrays," *IEEE Transactions on Multimedia*, vol. 20, no. 9, p. 2235–2248, 2018.
[47] W. Zhang, S. Tang, Y. Cao, S. Pu, F. Wu, and Y. Zhuang, "Frame augmented alternating attention network for video question answering," *IEEE Transactions on Multimedia*, vol. 22, no. 4, p. 1032–1041, 2020.
[48] B. Cao, N. Wang, J. Li, and X. Gao, "Data augmentation-based joint learning for heterogeneous face recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 6, pp. 1731–1743, 2019.
[49] F. Fahimi, S. Dosen, K. K. Ang, N. Mrachacz-Kersting, and C. Guan, "Generative adversarial networks-based data augmentation for brain-computer interface," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2020.
[50] J. Yoon, J. Jordon, and M. van der Schaar, "GANITE: estimation of individualized treatment effects using generative adversarial nets," in *International Conference on Learning Representations*, 2018.
[51] L. Neal, M. L. Olson, X. Z. Fern, W. Wong, and F. Li, "Open set learning with counterfactual images," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 620–635.
[52] D. Kaushik, E. H. Hovy, and Z. C. Lipton, "Learning the difference that makes A difference with counterfactually-augmented data," in *International Conference on Learning Representations*, 2020.
[53] R. Zmigrod, S. J. Mielke, H. M. Wallach, and R. Cotterell, "Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology," in *Annual Meetings of the Association for Computational Linguistics*, 2019, pp. 1651–1661.
[54] E. Barnea and O. Ben-Shahar, "Exploring the bounds of the utility of context for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7412–7420.
[55] S. K. Divvala, D. Hoiem, J. Hays, A. A. Efros, and M. Hebert, "An empirical study of context in object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1271–1278.
[56] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim, "Learning not to learn: Training deep neural networks with biased data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9012–9020.
[57] M. S. Alvi, A. Zisserman, and C. Nellåker, "Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings," in *Workshop of Proceedings of the European Conference on Computer Vision*, 2018, pp. 556–572.
[58] K. G. Y. J. L. M. F. D. G. Krishna Kumar Singh, Dhruv Mahajan, "Don't judge an object by its context: Learning to overcome contextual bias," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 067–11 075.
[59] J. Pearl, *Causality: Models, Reasoning and Inference*. USA: Cambridge University Press, 2009.
[60] M. R., "Confounding and confounders," *Occupational and environmental medicine*, pp. 227–234, 2003.
[61] S. I. VanderWeele T J, "On the definition of a confounder," *Annals of statistics*, p. 196, 2013.
[62] J. Pearl, "Comment: graphical models, causality and intervention," *Statistical Science*, vol. 8, pp. 266–269, 1993.
[63] S. Greenland, J. Pearl, and J. M. Robins, "Causal diagrams for epidemiologic research," *Epidemiology*, pp. 37–48, 1999.
[64] J. Pearl, "The do-calculus revisited," in *Conference on Uncertainty in Artificial Intelligence*, 2012, pp. 3–11.
[65] M. Y. Guan, V. Gulshan, A. M. Dai, and G. E. Hinton, "Who said what: Modeling individual labelers improves classification," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018, pp. 3109–3118.
[66] D. Geiger, T. Verma, and J. Pearl, "d-separation: From theorems to algorithms," in *Conference on Uncertainty in Artificial Intelligence*, 1989, pp. 139–148.
[67] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3917–3926.
[68] I. Sutskever, J. Martens, G. E. Dahl, and G. E. Hinton, "On the importance of initialization and momentum in deep learning," in *Proceedings of International Conference on Machine Learning*, 2013, pp. 1139–1147.
[69] J. Liu, C. Gao, D. Meng, and W. Zuo, "Two-stream contextualized CNN for fine-grained image classification," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 4232–4233.
[70] C. F. Flores, A. Gonzalez-Garcia, J. van de Weijer, and B. Raducanu, "Saliency for fine-grained object recognition in domains with scarce training data," *Pattern Recognition*, vol. 94, pp. 62–73, 2019.
[71] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Neural Information Processing Systems*, 2019, pp. 8024–8035.



Wei Qin Wei Qin received the B.E. degree and Master degree from the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, China, where he is currently pursuing the Ph.D degree. His research interests include computer vision and machine learning.

Hanwang Zhang received the BEng (Hons.) degree in computer science from Zhejiang University, Hangzhou, China, in 2009, and the PhD degree in computer science from the National University of Singapore, in 2014. He is currently an assistant professor with Nanyang Technological University, Singapore. He was a research scientist with the Department of Computer Science, Columbia University. His research interests include computer vision, multimedia, and social media. He is the recipient of the Best Demo runner-up award in ACM MM 2012, the Best Student Paper award in ACM MM 2013, and the Best Paper Honorable Mention in ACM SIGIR 2016, and TOMM best paper award 2018. He is also the winner of Best PhD Thesis Award of School of Computing, National University of Singapore, 2014.

Richang Hong (Member, IEEE) received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2008. He was a Research Fellow of the School of Computing with the National University of Singapore, from 2008 to 2010. He is currently a Professor with the Hefei University of Technology, Hefei. He has coauthored over 70 publications in the areas of his research interests, which include multimedia content analysis and social media. He is a member of the ACM and the Executive Committee Member of the ACM SIGMM China Chapter. He was a recipient of the Best Paper Award from the ACM Multimedia 2010, the Best Paper Award from the ACM ICMR 2015, and the Honorable Mention of the IEEE Transactions on Multimedia Best Paper Award. He has served as the Technical Program Chair of the MMM 2016. He has served as an Associate Editor of Information Sciences (Elsevier) and Signal Processing (Elsevier).

Ee-Peng Lim is a professor of information systems at Singapore Management University. His research interests include social network and web mining, information integration, and digital libraries. He has published more than 300 refereed journal and conference papers in these areas. He is currently the faculty director of the Living Analytics Research Center which focuses on urban and social analytics. He is a member of the Singapore's Social Science Research Council and also serves as the Steering Committee Chair of Pacific Asia Conference on Knowledge Discovery and Data Mining (PAKDD).

Qianru Sun has been an Assistant Professor in the School of Information Systems, Singapore Management University, since 2019. From 2018 to 2019, she was a Joint Research Fellow at the National University of Singapore and the MPI for Informatics. From 2016 to 2018, she held the Lise Meitner Award Fellowship and worked at the MPI for Informatics. In 2016, she obtained her Ph.D. degree from Peking University. In 2014, she was a visiting student at the University of Tokyo. Her research interests are computer vision and machine learning that aim to develop efficient algorithms and systems for visual understanding.