Towards robust inference against distribution shifts in computer vision

Tang, Kaihua

2021

Tang, K. (2021). Towards robust inference against distribution shifts in computer vision. Doctoral thesis, Nanyang Technological University, Singapore.
https://hdl.handle.net/10356/154119

Towards Robust Inference Against Distribution Shifts in Computer Vision

Kaihua Tang

School of Computer Science and Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

2021

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

13/07/2021

.....

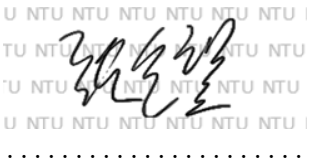
Date

.....

Kaihua Tang

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.



13/07/2021

.....

Date

.....

Ass. Prof. Hanwang Zhang

Authorship Attribution Statement

This thesis contains material from 3 papers published in the following peer-reviewed conferences and 1 paper that is currently submitted to a conference (also released to arXiv preprint) in which I am listed as the first author for all 4 papers.

Chapter 3 is published as Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, Wei Liu. "Learning to Compose Dynamic Tree Structures for Visual Contexts." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR, Oral and Best Paper Finalists(45/5160)**). Long Beach, United States. 2019.

The contributions of the co-authors are as follows:

- Prof. Hanwang Zhang pointed out the research direction and participated in the paper writing.
- Baoyuan Wu, Wenhan Luo, Wei Liu joined the early discussion and polished the paper before submission.
- I designed the algorithm, took charge of the entire code implementation and wrote the draft paper for other co-authors to refine.

Chapter 4 is published as Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, Hanwang Zhang. "Unbiased Scene Graph Generation from Biased Training." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR, Oral**). Seattle, United States. 2020.

The contributions of the co-authors are as follows:

- Prof. Hanwang Zhang pointed out the research direction and participated in the paper writing.
- Yulei Niu joined the early discussion of algorithms. Jiaxin Shi contributed part of non-core codes in the proposed framework. Jianqiang Huang helped with polishing the paper before submission.
- I designed the core algorithm, took charge of most code implementation and wrote the draft paper for other co-authors to refine.

Chapter 5 is published as Kaihua Tang, Jianqiang Huang, Hanwang Zhang. "Long-Tailed Classification by Keeping the Good and Removing the Bad Momentum Causal Effect." Advances in Neural Information Processing Systems (**NeurIPS, Poster**). Virtual. 2020.

The contributions of the co-authors are as follows:

- Prof. Hanwang Zhang pointed out the research direction and participated in the paper writing.
- Jianqiang Huang joined the early discussion and helped with polishing the paper before submission.
- I designed the algorithm, took charge of the entire code implementation and wrote the draft paper for other co-authors to refine.

Chapter 6 is submitted as Kaihua Tang, Mingyuan Tao, Xian-Sheng Hua, Hanwang Zhang. "Adversarial Visual Robustness by Causal Intervention." **arXiv preprint (Under Review)**. 2021.

The contributions of the co-authors are as follows:

- Prof. Hanwang Zhang pointed out the research direction and participated in the paper writing.
- Mingyuan Tao and Xian-Sheng Hua helped with polishing the paper before submission.
- I designed the algorithm, took charge of the entire code implementation and wrote the draft paper for other co-authors to refine.

13/07/2021

TANG Kaihua

Date

Kaihua Tang

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor Prof. Hanwang Zhang. Without his patient guidance and continuous help, I would never have accomplished these research projects and published corresponding papers. He opened the door of academic research for me with his invaluable expertise. Before I met him, I was just a green hand that had a pretty bad taste of research and zero experience of publishing top-tier papers. However, he reinvented me with his immense knowledge and insightful support, which guided me to find my research topics, sharpen my critical thinking, build my academic skills and bring my work to a higher level.

I would like to acknowledge my collaborators, Dr. Xu Yang, Ph.D. Xinting Hu, Ph.D. Jianqiang Huang, Dr. Yulei Niu, Dr. Jiaxin Shi, etc., for their support and insightful comments. Besides, I also want to thank Ph.D. Xin Tan, Dr. Min Wang, Dr. Long Chen, Ph.D. Chi Zhang, Ph.D. Dong Zhang, Ph.D. Daqing Liu, Ph.D. Jiaxin Qi, Ph.D. Beier Zhu, Ph.D. Zhongqi Yue, Ph.D. Huaizheng Zhang, whose precious advice helped me a lot in both academic research and life decisions.

I would like to thank the School of Computer Science Nanyang Technological University for their great facilities, knowledgeable faculty members, kind administration staff, and the NTU Research Scholarship that helped me to complete my Ph.D. studies. I also want to thank Alibaba DAMO Academy, especially the City Brain Group, for their technical help and generous sharing of computational resources.

Last but not least, I am extremely grateful to my father Jun Tang and my mother Dingxian Wang for their unconditional love. Without their support, I wouldn't be able to pursue my Ph.D. degree. I also want to express my gratitude to all the friends in my life, Xuchao Lu, Min Wen, Zhiming Ding, Honglin Chen, Yuxi Zhang, Yushan Liu, Wencan Zhang, and those who are not listed here. Their presence and companionship made me a better person.

Kaihua Tang, July 2021

Abstract

After a decade of prosperity, the development of machine learning based on deep neural networks (DNNs) seems to reach a new turning point. A variety of tasks and fields have proved that recklessly feeding a massive volume of data and increasing the model capacity would no longer bring us a panacea for all the problems. The ubiquitous bias in the model structures, long-tailed distributions, and optimization strategies stops the DNN from learning the underlying causal mechanisms, resulting in the catastrophic drop of performances when facing distribution shift problems like rare spatial layouts, misalignment between source domains and targeted domains, or adversarial perturbations .

To tackle these challenges and increase the robustness of DNNs for better generalization abilities, a line of research, including dynamic network with attention architectures, long-tailed recognition, and adversarial robustness, have attracted significant attention in recent years. In this thesis, we systematically study the threats of model robustness against distribution shifts from three aspects: 1) network architectures, 2) long-tailed distributions, 3) adversarial perturbations. The latter two can also be interpreted as the explicit and implicit distribution shifts on patterns, respectively. To address these threats, we propose several algorithms that successfully increase the robustness of deep neural networks in a wide range of computer vision tasks, including image classification, object detection, instance segmentation, scene graph generation, and visual question answering.

In summary, the major contributions of this thesis are as follows:

- **We propose to compose dynamic tree structures**, called VCTree, that put objects in an image into robust visual contexts. Such a robust architecture can help visual reasoning tasks like scene graph generation and visual question answering, which are known to suffer from biased datasets. Visual contexts, which play a critical role in the modern convolutional neural networks, can

be captured by either pixel-level large receptive fields or object-level message-passing mechanisms. Since robust structures of the former have been well studied by dilated convolutions or feature pyramids, we mainly focus on the object-level contexts in this thesis. Compared with existing structured object representations that utilize fully-connected graphs or chains to connect the objects, the proposed VCTree have two key advantages: 1) the effective and efficient binary tree structure that encodes both parallel and hierarchical relationships among objects, e.g., "hands" and "legs" are usually co-occurred and belong to "person"; 2) the dynamic constructions, which allows the VCTree varying from image to image and question to question, making the message passing among objects more content-specific and task-specific. These advantages prevent the proposed VCTree from blindly favoring the statistical preference by memorizing a fixed layout of objects from training data, *i.e.*, overfitting the biased dataset. We also develop a hybrid learning framework, integrating supervised learning for end tasks and reinforcement learning for tree structures, which further increases the robustness against memorizing the biased distributions.

- **We introduce a multi-modal causal inference framework** for scene graph generation to obtain unbiased prediction through the Total Direct Effect (TDE). Due to the different components of a multi-modal network usually having diverse difficulties to converge, the easy path tends to be overfitted by the model, hurting the robustness of DNNs. Taking scene graph generation as an example, recent methods are far from practical for the severe training bias. To be specific, they collapse diverse predictions of `people sit on/ walk on/lay on beach` into `people on beach`, because the language embedding and statistical prior converge faster than visual features, making visual appearances less important. Those generated scene graphs can hardly benefit downstream tasks like Visual Question Answering as they are merely a bag of objects. However, achieving robustness in Scene Graph Generate (SGG) is not trivial because traditional debiasing methods cannot distinguish between the good and bad bias, *i.e.*, hurting the learning of rich contexts in feature extraction backbones. Therefore, we present a novel multi-modal TDE framework based on counterfactual causal inference rather than conventional likelihood inference, which maintains the good knowledge during training while removes the bad prejudice from the counterfactual indirect effect. Note that the proposed framework is agnostic to any SGG

model and thus can be widely applied in the community that seeks unbiased predictions.

- **We firstly design a novel De-confound TDE algorithm** to solve the general long-tailed problems in pure (single-modal) computer vision tasks without relying on the accessibility of prior statistics of the class distribution. As we all know, maintaining a balanced dataset across many classes is challenging. Because the data are long-tailed in nature, so the cost of balancing a dataset will increase exponentially with the growing vocabulary. It is even impossible when the sample-of-interest co-exists with each other in one collectible unit, *e.g.*, multiple visual instances in one image. Therefore, obtaining long-tailed robustness is the key to conduct deep learning at a large scale. However, existing solutions are mainly based on heuristic re-weighting or re-sampling strategies that lack a fundamental theory. In this thesis, we establish a novel De-confound TDE framework based on causal inference for the general long-tailed robustness. It not only unravels the underlying mechanisms of previous methods but also derives a new principled solution. Different from the multi-modal TDE, it can be directly applied to most of the single-modal computer vision tasks by incorporating the de-confound training with counterfactual TDE inference.

- **We present a Causal intervention by instrumental Variable (CiiV) regularization** that not only offers a proactive defender avoiding the endless adversarial training, but also opens a novel yet fundamental viewpoint of adversarial robustness research. Despite the remarkable progress achieved by DNNs, adversarial vulnerability keeps haunting the computer vision community. Among a variety of potential solutions, adversarial training is the de facto most promising defense against adversarial examples. Yet, its passive nature inevitably prevents it from being immune to unknown attackers. To achieve a proactive defense, we need a more fundamental understanding of adversarial examples beyond the popular bounded threat model. In this thesis, we provide a causal viewpoint of adversarial vulnerability: the cause is the confounder ubiquitously existing in learning, where attackers are precisely exploiting the confounding effect. Therefore, a fundamental solution for adversarial robustness is by causal intervention. As the confounder is unobserved in general, we propose to use the instrumental variable that achieves intervention without the need for confounder observation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Robust Deep Neural Network

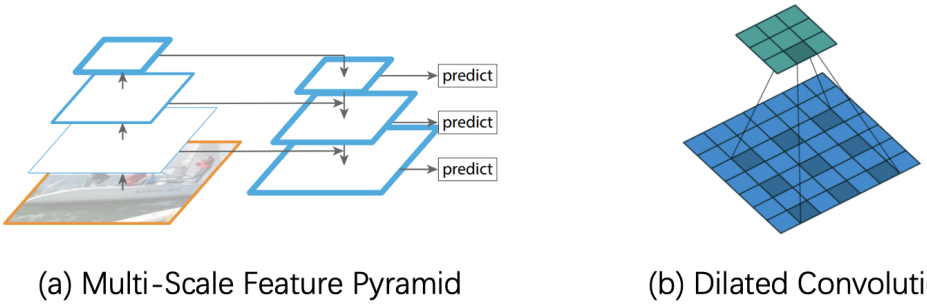
Due to the rise of deep neural networks (DNNs), a variety of computer vision fields have witnessed significant progress in recent years, *e.g.*, Convolutional Neural Network (CNN) based image classification models [24–26] have outperformed human performances, detection or segmentation algorithms [27–30] are able to recognize over 9,000 object categories [31], and some of them are even faster enough to be run on mobile devices [32, 33], etc. The remaining obstacle of widely deploying DNN-based computer vision systems is that they are still not as robust as our human beings [34], especially under the distribution shift. Without achieving the robustness of machine learning systems, most of the well-studied fields and problems in the experimental environments and few limited datasets may face the drastic performance drop in real-world scenarios [35], which could eventually lead to the coming of another period of disappointment to AI systems, *i.e.*, "AI Winter" [36].

Due to the fact that DNNs have an inherent tendency to learn spurious correlations as shortcuts [35], achieving robust deep neural networks is not trivial. It has long been noticed that DNN systems are brittle when facing the rare situations [37, 38], misalignment between unbalanced source domains and balanced target domains [10, 39], or adversarial perturbations [40, 41]. It's because statistics-based DNNs are learning from associations [42], *i.e.*, memorizing the data as a look-up table without understanding the underlying causality. Therefore, memorized pattern distributions could take the place of causal features as they are

more efficiently interpolating the data points. For example, only predicting the banana from "yellow ellipse" patterns (neglecting the green banana) due to selection bias [43] or simply memorizing human always "sitting on" (rather than "standing on") chairs for the long-tailed data distribution [10, 44]. Those biased models could create a false sense of security when the test sets that are drawn from the same distribution, because the robustness against distribution shifts is not considered in the evaluation. For example, if a self-driving system is only trained and tested under the daylight, we would overestimate its reliability until it crashes under the night scene at one day.

Therefore, there is a series of research, including robust network structure [8, 45, 46], long-tailed recognition [10, 39], and adversarial robustness [40, 41] that attempts to increase the robustness of DNNs. To tackle the biased prediction problem, Chen *et al.* [15] introduce a routing mechanism based on a structured knowledge graph to alleviate the bias problem. In robust network structure problems, Liu *et al.* [45] adopt a tree-structured neural module network based on the dependency parsing tree of description sentences to increase the interpretability and robustness of the visual grounding task. Yang *et al.* [46] also use dynamic neural modules to imitate sentence patterns, which alleviates the over-fitting by regularizing the diverse training with different modules. In long-tailed recognition, Liu *et al.* [10] use the dynamic meta-embedding to transfer the knowledge from head categories to tail categories. Kang *et al.* [11] improve the long-tailed robustness by decoupling the instance-balanced training of backbone feature extraction with the class-balanced learning of classifier. Tan *et al.* [39] solve the long-tailed object detection by introducing an equalization re-weighting loss. As to the adversarial robustness, adversarial training and its variants [47, 48] are proposed to intuitively alleviate the adversarial perturbations. De-noising methods [49–53] treat adversarial perturbations as noises then conducting de-noising modules to remove their effects.

However, due to various forms of bias that can be either explicit or implicit, and broad potential solutions ranging from network structure to the inference strategy or training pipeline, most of the studies [10, 41, 46] mainly focus on a narrow definition of robustness against the various distribution shifts. To obtain the general robust deep neural network, in this thesis, we will systematically investigate



(a) Multi-Scale Feature Pyramid (b) Dilated Convolutions

FIGURE 1.1: The robust architectures to capture the pixel-level contexts.

the distributional robustness in three aspects: 1) architectural robustness, 2) long-tailed robustness and 3) adversarial robustness. In architectural robustness, we are going to demonstrate how vulnerable the fully-connected/chained network structures are to the shortcuts or over-fitting the training distributions, which may shed some light on how important the robustness is to DNNs. After that, we firstly introduce causal inference [54] to the computer vision fields and design causality-oriented frameworks to increase the distributional robustness of DNNs by pursuing the causal effects rather than associations. We categorize the data bias into explicit distribution bias and implicit pattern bias. The representative of the former is the long-tailed problem while the latter is revealed by adversarial attacks. Note that the latter is the hardest problem in distribution shifts due to its dynamic and learning nature. The details of these three aspects will be introduced in the following paragraphs.

1.2 Architectural Robustness

In modern computer vision fields, CNN architecture is indispensable for many state-of-the-art visual systems [24, 25, 27], which has a variety of advantages, *e.g.*, encoding the visual contexts, equivariant to translation, etc. Among them, visual contexts are both important and tricky. The bright side is that it can provide additional information to better infer attributes of an object, especially under occlusions, stains, or limited resolutions. However, inappropriately extracting the visual contexts could drastically hurt the robustness, because the neural network model could ignore the hard causal features by taking the statistical shortcut of certain context combinations, *e.g.*, recognizing the furniture store as the bedroom

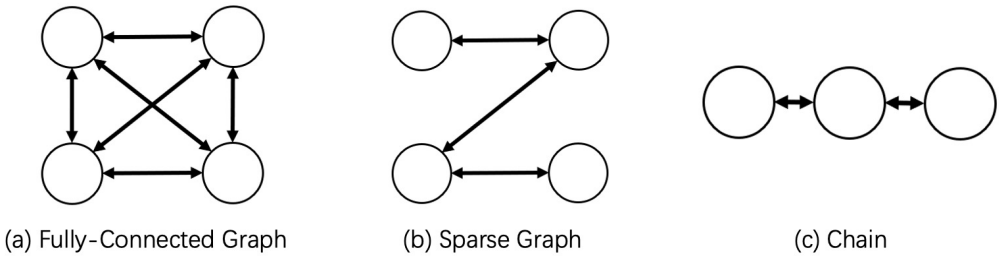


FIGURE 1.2: The previous architectures to capture the object-level scene contexts, where each node is an object.

by the appearance of beds. The modern CNN architectures usually extract the visual contexts by enlarging the receptive field of convolutions. Yet, it also gradually degenerates the CNN into fully connected layers, which are prone to overfit the visual data. Therefore, as illustrated in Figure 1.1, the robust variant of architectures like multi-scale feature map pyramids and dilated convolutions [18, 55, 56] are proposed to merge various receptive fields rather than enlarging one. Such pixel-level local contexts [57] play one of the significant roles in closing the performance gap of the "mid-level" vision between humans and machines, such as R-CNN based object detection [18, 27, 28], instance segmentation [29, 58], and FCN based semantic segmentation [55, 59, 60].

Despite the above feature pyramids achieving robust pixel-level contexts, achieving the robustness of object-level scene contexts is still challenging. In this thesis, we mainly focus on the "higher-level" of scene contexts that are explicitly extracted on object features. Most of the existing methods try to capture the contexts by arranging the objects into chains [1] and graphs [2, 61–64] (Figure 1.2), where sequential models such as bidirectional LSTM [65] and CRF-RNN [66] or graphical models such as GCN [67] are used to encode the context information for chains and graphs, respectively. However, they are sub-optimal prior structures. First, the chains over-simplify the 3-D spatial relationships and thus may only capture simple spatial information or co-occurrence bias; though graphs are better replacements, they fail to discriminate the hierarchical relations and parallel relations; in addition, the commonly used dense fully connections could also lead to message passing saturation in the subsequent context encoding [2], which causes the overfitting to the statistical preference in training data. Second, visual contexts should be content-/task-driven, *e.g.*, the object layouts should vary from content to content, task to task. Therefore, fixed structures like chains and graphs inevitably memorize

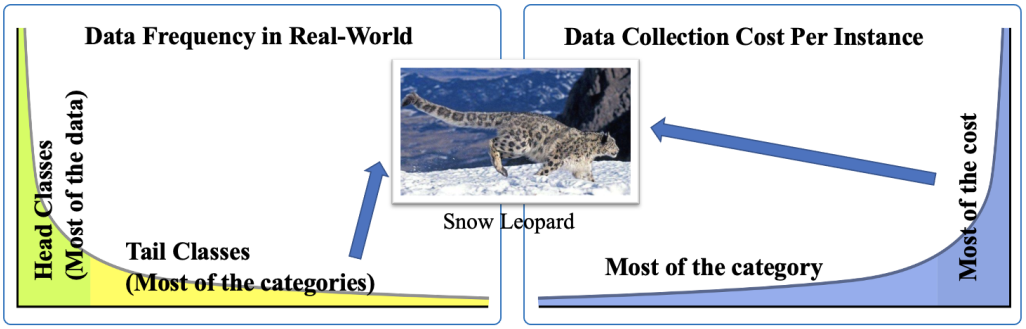


FIGURE 1.3: Both the number of instances for each category and the cost to collect them follow the long-tailed distribution.

the dominant situations and hurt the robustness against distribution shifts as they are incompatible with the dynamic scenes [68].

Driven by the above drawbacks, we are going to introduce a novel VCTREE algorithm that dynamically composes tree structures for object-level visual contexts, which significantly improves the architectural robustness of visual contexts by preventing the overfitting to fixed structures. The hybrid learning strategy that fine-tunes the tree-construction network through reinforcement learning further stops the model memorizing the training distribution. Through comprehensive evaluations, we observe that the proposed VCTREE consistently outperforms the other algorithms on high-level visual reasoning tasks like Scene Graph Generation or Visual Question Answering.

1.3 Long-Tailed Robustness

In the past decades, we have witnessed large and balanced datasets like ImageNet [69] and MS-COCO [70] leading the fast evolution of computer vision algorithms [24, 26, 27]. The recent blooming era of Transformer-based models further increases the demand for huge data by several orders of magnitude. However, along with the rapid growth of the digital data created by human activities, the crux of making the larger datasets is no longer about how to collect or where to collect, but how to balance.

However, the cost of expanding balanced datasets to a larger set of categories is not linear, but exponential, because the distribution of natural data is inevitably

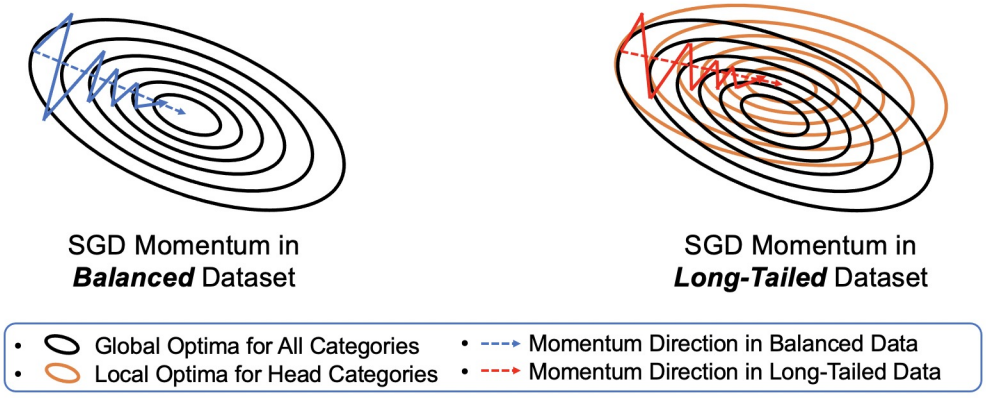


FIGURE 1.4: An intuitive illustration of how the momentum amplifies the model bias in long-tailed dataset.

governed by Zipf's law [71], *i.e.*, following the long-tailed distribution. To be specific, every single data point increased for tail categories will come along with more samples from head categories. Meanwhile, as shown in Figure 1.3, the cost of collecting them also follows a reversed long-tailed distribution. For example, the cost of capturing a photo of the snow leopard is much higher than most of the common animals, making the balanced dataset extremely expensive. What's worse, re-balancing the class is impossible in certain tasks. For example, in instance segmentation [19], if we increase the images of rare instances like "remote controller", we have to bring in more common instances like "sofa" and "TV" simultaneously in every new images [72].

Therefore, the development of robust algorithms under long-tailed datasets is crucial and indispensable for training deep models at a large scale. Most of recent methods [16, 39] adopt intuitive re-weighting and re-sampling strategies, or their decoupled variants [11] to solve this problem. However, the fundamental theory that explains the whys and wherefores of the long-tailed effect is still missing. We humans also live in a long-tailed world, but why can we overcome the head preference and successfully recognize the rare situations? In this thesis, we first point out that the accumulative momentum effect could be the main cause of the long-tailed problem in DNNs, which implicitly encodes the data distribution, causing the optimization direction to deviate from the global optima, as explained in Figure 1.4. Yet, simply removing the momentum from the optimizer won't solve the problem, because it will result in unstable gradients, trapping the model in local optimums. Besides, SGD [73, 74] itself is accumulatively updating the parameters.

To tackle the paradox of the momentum effect, we embrace the causal inference [42, 54] framework and design the Total Direct Effect(TDE) strategy that keeps the momentum in training and only removes its bad causal effect through a counterfactual inference. There are two different applications of TDE strategy: 1) the multi-modal TDE and 2) the general de-confound TDE. The former can be directly applied to existing Scene Graph Generation frameworks or other multi-modal tasks, eliminating the long-tailed bias from the overfitted single modality. Experiments on Visual Genome dataset demonstrate that it can significantly improve the Zero-Shot Recall@K and the proposed balanced metric mean Recall@K compared with conventional methods, resulting better performances of downstream tasks like image retrieval. The latter combines the de-confound training and TDE inference, obtaining the robustness on a variety of pure computer vision tasks. We evaluate it on ImageNet-LT, Long-Tailed CIFAR-10/-100 for image classification and LVIS V0.5/V1.0 for object detection and instance segmentation. The experimental results significant outperform previous methods. Besides, the improvements are achieved without introducing any additional modules or training stages.

1.4 Adversarial Robustness

Apart from the explicit long-tailed bias in the category level, the implicit pattern-level bias is an even stronger threat to intelligent computer vision systems. Despite the remarkable progress achieved by DNNs, adversarial vulnerability [41] keeps haunting the computer vision community since it has been spotted by Szegedy *et al.* [47]. Adversarial attackers exploit the trivial pattern-level bias to drastically alter the model predictions by modifying non-robust patterns that are barely noticed by humans [75], as illustrated in Figure 1.5.

However, achieving the adversarial robustness is much more difficult as the physical meaning of mid-level features in DNNs is implicit, making the de-biasing of adversarial perturbations intractable. Such a distribution shift is also harder than its category-level counterpart, because the attacker that generates the distribution shift is dynamic and learnable. Therefore, the most promising defender still remains to be the intuitive Adversarial Training and its variants [76, 77], yet their performances are largely dependent on whether the training set contains sufficient

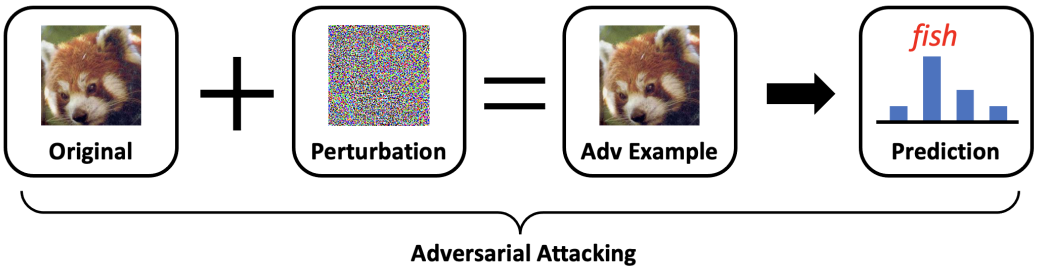


FIGURE 1.5: An example of adversarial attacking and how an attacker modify the model prediction, where adv example is the abbreviation of adversarial example.

adversarial samples from various attackers as many as possible [78]. Besides, adversarial training may easily overfit to known attackers [79]. Particularly, in few-/zero-shot scenarios, it is even impossible to collect enough adversarial training samples based on the out-of-distribution/unseen original samples [80].

To overcome the passive immunity of adversarial training methods and proactively defend against all the known and unknown attackers, we have to de-mystify the cause of adversarial examples. Recent studies [75, 81] show that they are predictive features that can only be exploited by machines, *i.e.*, non-robust features. Yet, Ilyas *et al.* [75] use the definition of adversarial samples to define the robust and non-robust features, which is unfortunately the circular reasoning, because it only allows us to recognize the non-robust features as "adversarial samples" again. In this thesis, we postulate that adversarial attacks are confounding effects, which are spurious correlations established by non-causal features. For example, if dark spots appear on bananas in most samples, a model trained by associating samples with labels will recklessly use the spot pattern—the confounding effect—to recognize the banana category, which would thus be exploited by attackers to fool the model.

In the causality field, removing the unobserved and unknown confounders is also challenging in practice [82], because their effects are entangled with causal effects and thus hard to be eliminated. However, human visions can successfully ignore the confounded features and reveal surprisingly adversarial robustness. Given the fact that biological visions are more complex than machines in terms of both neuron amount [83] and diversity [84, 85], there is no reason for human vision to extract "less feature" than machines. Inspired by the retinotopic sampling [86] of human visions, we conjecture that such sampling is the answer, which can be viewed as causal intervention by using instrumental variable [87] that deliberately

ignores non-robust features. Therefore, we imitate such a mechanism and introduce a CiiV framework for training robust DNN against adversarial attacks, which does not only offer a proactive defender, but also opens a novel yet fundamental viewpoint of adversary research. To verify our hypothesis, we also construct a Confounded Toy (CToy) dataset to demonstrate the relationship between its adversarial perturbations and pre-defined confounders. The experimental results and visualizations also prove that the proposed CiiV can significantly increase the adversarial robustness in various settings.

1.5 Outline of the Thesis

The thesis is organized as follows.

- In Chapter 1, we firstly introduce the background of robust DNNs and the reason of why it's so important for the computer vision community. After briefly review the recent progress of robust DNN, we categorize the challenges in three aspects: 1) architectural robustness, 2) long-tailed robustness and 3) adversarial robustness. Afterwards, we discuss each case of robustness from their descriptions and limitations of current approaches to the motivation and basic ideas of our proposed solutions.

- In Chapter 2, we systematically review all the related tasks, preliminary concepts, especially those in the causal inference, as causality is alien to most of the audiences. At last, we introduce some background knowledge of previous solutions and recent state-of-the-art methods.

- In Chapter 3, we propose to compose dynamic tree structures that place the objects in an image into a robust visual context, helping visual reasoning tasks such as scene graph generation and visual question answering to overcome the data bias. To construct the proposed visual context tree model, dubbed VCTree, we design a score function that calculates the task-dependent validity between each object pair, and the tree is the binary version of the maximum spanning tree from the score matrix. Then, visual contexts are encoded by bidirectional TreeLSTM and decoded by task-specific models. We develop a hybrid learning procedure which integrates

end-task supervised learning and the tree structure reinforcement learning, where the former's evaluation result serves as a self-critic for the latter's structure exploration. Experimental results on two benchmarks, which require reasoning over contexts: Visual Genome for scene graph generation and VQA2.0 for visual Q&A, show that VCTree outperforms state-of-the-art results while discovering interpretable visual context structures. The codes are publicly available on Github: `https://github.com/KaihuaTang/VCTree-Scene-Graph-Generation` for scene graph generation task and `https://github.com/KaihuaTang/VCTree-Visual-Question-Answering` for visual question answering task.

- In Chapter 4, we present a novel SGG framework (multi-modal TDE) based on causal inference but not the conventional likelihood. We first build a causal graph for SGG, and perform traditional biased training with the graph. Then, we propose to draw the counterfactual causality from the trained graph to infer the effect from the bad bias, which should be removed. In particular, we use Total Direct Effect as the proposed final predicate score for unbiased SGG. Note that our framework is agnostic to any SGG model and thus can be widely applied in the community who seeks unbiased predictions. By using the proposed Scene Graph Diagnosis toolkit, which is publicly available on GitHub: `https://github.com/KaihuaTang/Scene-Graph-Benchmark.pytorch`, on the SGG benchmark Visual Genome and several prevailing models, we observed significant improvements over the previous state-of-the-art methods.

- In Chapter 5, we establish a causal inference framework (de-confound TDE) for the general long-tailed recognition problem, which not only unravels the whys of previous intuitive re-balancing methods and decoupling strategies, but also derives a new principled solution. Specifically, our theory shows that the SGD momentum is essentially a confounder in long-tailed classification. On one hand, it has a harmful causal effect that misleads the tail prediction biased towards the head. On the other hand, its induced mediation also benefits the representation learning and head prediction. Our framework elegantly disentangles the paradoxical effects of the momentum, by pursuing the direct causal effect caused by an input sample. In particular, we use causal intervention in training, and counterfactual reasoning in inference, to remove the "bad" while keep the "good". We

achieve new state-of-the-arts on three long-tailed visual recognition benchmarks: Long-tailed CIFAR-10/-100, ImageNet-LT for image classification and LVIS for instance segmentation. Our code is also available on `https://github.com/KaihuaTang/Long-Tailed-Recognition.pytorch`.

- In Chapter 6, we provide a causal viewpoint of adversarial vulnerability: the cause is the confounder ubiquitously existing in learning, where attackers are precisely exploiting the confounding effect. Therefore, a fundamental solution for adversarial robustness is by causal intervention. As the confounder is unobserved in general, we propose to use instrumental variable that achieves intervention without the need for confounder observation. We term our robust training method as Causal intervention by instrumental Variable (CiiV). It's a regularization loss using the retinotopic sampling augmentation, which is stable and guaranteed not to suffer from gradient obfuscation. Extensive experiments on a wide spectrum of attackers and settings applied in CIFAR-10, CIFAR-100 and mini-ImageNet datasets empirically demonstrate that CiiV is robust to adaptive attacks.

- In Chapter 7, we summarize the contributions of the proposed algorithms from the above mentioned three types: 1) architectural robustness, 2) long-tailed robustness, and 3) adversarial robustness. Then we conclude the importance of causal inference in achieving robust neural networks against distribution shifts, and how to use it for the guidance of future designs of robust DNNs. Afterwards, we point out the new directions in those three kinds of robustness.

Chapter 2

Literature Review

2.1 Related Tasks

In this section, we will introduce the tasks that are used to evaluate the robustness of the proposed methods in this thesis, including Scene Graph Generation, Visual Question Answering, Long-Tailed Recognition, and Adversarial Examples.

2.1.1 Scene Graph Generation

Scene Graph Generation(SGG) [2, 6] is a mid-level task that jointly detects all the objects and their relationships in an image, which aims to bridge the gap between low-level recognition and high-level visual intelligence and reasoning. It has received increasing attention in the computer vision community, because it has the potential to bring revolutions to the downstream visual reasoning tasks, like visual question answering [88], image captioning [89], visual grounding [90], image generation [91], etc. Most of the existing methods [2, 8, 62, 64, 92–97] struggle for better feature extraction networks. For example, Xu *et al.* [2] design an iterative message passing module to simultaneously update the object feature and relationship feature. Yang *et al.* [94] adopt an attention-based GCN to fine-tune the representations. Newell *et al.* [98] train a stacked hourglass architecture [99] to directly predict relationships from the feature map. Li *et al.* [93] try to improve the features by multi-task learning.

Recently, Zellers *et al.* [6] firstly bring the bias problem of SGG into attention. Without using any visual features, the frequency baseline based on pure statistical preferences easily outperforms plenty of previous heavy models, revealing the systemic bias towards frequent object combinations. To obtain the unbiased scene graph generation, we firstly propose the unbiased metric, mean Recall, to fairly evaluate models on both common and rare situations. Other related work [100] attempts to balance the dataset by pruning those dominant and easy-to-predict relationships in the training set, yet, it could significantly increase the cost of building an SGG system in real-world applications due to the tremendous cost in fair data collection.

2.1.2 Visual Question Answering

Visual Question Answering (VQA) [101] is a high-level task that bridges the gap between computer vision and natural language processing. It provides a comprehensive QA system that is able to answer the questions based on the visual knowledge of given images. It also serves as the bedrock for other more advanced tasks like visual dialogue. Most of the previous state-of-the-art VQA models [14, 102, 103] rely on various multi-modal visual attention mechanisms between the language embedding and bag of objects.

However, due to the complexity of combining the vision and language modalities, most of existing vision-and-language models are easily taking the shortcut of language bias, for example, gender bias, object bias, and action bias [104–106]. In VQA, there are two types of dominant biases: 1) the correlation between questions and answers [107, 108], *i.e.*, the question type bias; 2) the preference of asking existing objects in the image [101, 107, 109], *i.e.*, the selection bias. Both of them can be alleviated by the question-guided VCTree proposed by this thesis due to its dynamic nature cutting the shortcuts of learning biases.

2.1.3 Long-Tailed Recognition

Long-Tailed Recognition [10, 11, 39, 110] including a variety of tasks like image classification, object detection, instance segmentation, multi-label classification, etc., that seek to directly train deep neural networks on the highly skewed dataset,

i.e., under long-tailed distribution, while still performs well on the balanced test set. Due to the exponential cost of balancing the large dataset, long-tailed recognition is the key to deep learning at scale.

Similar to the popular large and balanced datasets such as ImageNet [69] and MS-COCO [70] for image classification [24] and object detection [18, 27], there are also corresponding long-tailed version of datasets for these tasks, *e.g.*, ImageNet-LT [10], Long-tailed CIFAR-10/-100 [16, 111] and LVIS [19]. The ImageNet-LT and Long-tailed CIFAR-10/-100 are directly sampled from the original ImageNet and CIFAR-10/-100 datasets by manually introducing a long-tailed distribution. LVIS re-annotates the MS-COCO 2017 with more detailed bounding boxes and segmentation masks, and a more fine-grained larger vocabulary. Thanks to these benchmarks, recent work [10, 11, 16] is able to fill in the performance gap between class-balanced and long-tailed train data, making the direct training on naturally collected data possible.

2.1.4 Adversarial Examples

Adversarial examples [40, 47] are specifically manipulated inputs that are able to drastically change the model predictions with insignificant perturbations, which can barely be observed by humans. It's a thorn in the side of the modern machine learning community that undermines the reliability of DNN models in various domains [112–119] and settings [120–125], causing serious security issues in computer vision systems, like face recognition [126] and autonomous vehicle [127]. Therefore, a line of related research [47, 48, 51, 77, 85, 128–131] has attracted significant attention in the machine learning community.

However, due to the lack of interpretability for the adversarial perturbations, an arms race has been started between the attackers and defenders in recent years. What's worse, the development of defenders [51, 77] is relatively slower than the progress on attackers [120]. Generally, the defenders fall into the following five categories: adversarial training [47, 48], data augmentation [132], generative classifier [128, 129], de-noising [51, 77], and certified defense [130]. Among them, some defenders [85, 131] are inspired by biological vision systems. Yet, there is still no theoretical framework to unify the various defenders, making the practical

adversarial robustness intractable. What's worse, the emerging attacking techniques [120, 133] keep challenging the existing defenders, making the adversarial robustness hard to be obtained [134] once for all.

2.2 Preliminary Concepts

In this section, we will introduce some preliminary concepts that would be mentioned in the thesis, including visual contexts, unbiased training and a variety of terminologies in the causal inference.

2.2.1 Visual Context

The visual context is a coherent object configuration based on the fact that some kinds of objects would co-vary with each other in certain scenes. Studies in cognitive science [135–137] further prove that human brains inherently exploit visual contexts to comprehensively understand cluttered objects.

To extract the visual context in computer vision systems, despite the consensus on its value, existing context models are diversified into a variety of implicit or explicit approaches. Implicit models directly encode surrounding pixels into multi-scale feature maps, *e.g.*, dilated convolution [60] presents an efficient way to increase receptive field, applicable in various dense prediction tasks [55, 59]; feature pyramid structure [18] combines low-resolution contextual features with high-resolution detailed features, facilitating object detection with rich semantics. Explicit models group objects into fixed layouts, *i.e.*, chains or graphs, then incorporate contextual cues through object connections [1, 2, 63]. The former methods are commonly used in low-level vision tasks, like image classification or object detection, while the latter ones are usually adopted by higher-level tasks, like scene graph generation or visual question answering, as they are directly built on top of the detected objects.

2.2.2 Unbiased Training

The bias problem has long been investigated in machine learning [138], because the real-world data *per se* and the way we describe it are biased: there are more

situations of "human sitting on chair" than "human standing on chair"; there are more "children" than "adults" playing the seesaw. However, those biased data could hurt the fairness and robustness of models, causing systematic bias in machine learning systems.

Existing debiasing methods can be roughly categorized into three types: 1) data augmentation or re-balancing(re-sampling and re-weighting) [139–144], 2) transfering the knowledge from sample-rich categories to the sample-scarce categories [10, 145, 146], 3) multi-stage/multi-branch decoupled training [11, 16]. Most of them rely on intuitively adjusting the contribution of images from different categories according to the data distribution. Meanwhile, the proposed TDE algorithm does not require to train additional layers like [147, 148] to model the bias, or require the accessibility to the data distribution, making it a general solution for various settings and tasks.

2.2.3 Causal Inference

In the past decades, causal inference [42, 149] has been widely adopted in psychology, politics, and epidemiology for years [150–152]. It does not just serve as an interpretation framework, but also provides solutions to achieve the desired objectives by pursuing causal effect.

Recently, causal inference has been shown effective in mitigating data bias in computer vision tasks [153–157]. They use confounder adjustment [106] or counterfactual inference [156] to alleviate the biased causal effect measured in domain-specific tasks. Recent surveys [158, 159] have also mentioned the potential of causality in computer vision fields, which not only provides an extensive framework to analyze the ubiquitous bias and shortcuts in a system but also helps to guide the development of solutions for these problems.

To better understand the causal inference framework used in the thesis, we will briefly introduce several common concepts in this field.

2.2.3.1 Causal Graph

Causal graph, which is also called structural causal model (SCM) [42, 54, 149] is a directed acyclic graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, indicating how a set of variables $\mathcal{N}$ interacts with each other through the causal links $\mathcal{E}$, where $\mathcal{N}$ is the set of nodes in the graph and the $\mathcal{E}$ is the set of edges. It provides a sketch of the causal relations behind the data and how variables obtain their values, *e.g.*, the price of the house is determined by its size and location, so the corresponding causal graph is illustrated in Figure 2.1.

2.2.3.2 Confounder

A confounder is a variable that influences both correlated and independent variables, creating a spurious statistical correlation. Considering the following causal graph, $exercise \leftarrow age \rightarrow cancer$, the elder people spend more time on physical exercise after retirement and they are also easier to get cancer due to the elder age, so the confounder age creates a spurious correlation that more physical exercise will increase the chance of getting cancer.

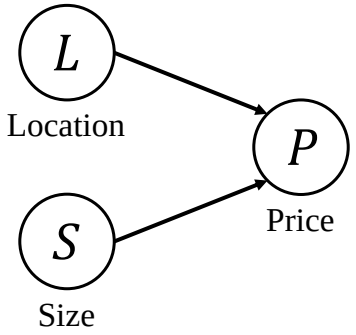


FIGURE 2.1: An example of causal graph with $\mathcal{N} = \{L, S, P\}$ and $\mathcal{E} = \{L \rightarrow P, S \rightarrow P\}$.

2.2.3.3 Mediator

The mediator is the side effect of the independent causation variable. The example of a mediator would be $drug \rightarrow placebo \rightarrow cure$, where mediator placebo is the side effect of taking a drug that prevents us from getting the direct effect of $drug \rightarrow cure$.

2.2.3.4 Causal Intervention

The causal intervention is the action of giving a variable a certain value without following the original causation relationships of this node, which can be regarded as removing all the incoming links of a node. The ultimate goal of causal intervention is to identify the causal effect of $X \rightarrow Y$ by removing all spurious correlations [149], denoted as $P(Y|do(X = x))$. However, the physical intervention may not always

applicable in the experiments, so there are three major ways to conduct virtual interventions: backdoor adjustment, front-door adjustment and instrumental variable. Given the underlying causal graph of the model, these adjustments can approximate the real causal intervention results without taking physical actions. For example, in the situation of containing a confounder C, causal intervention is well-defined as d-separation [149], by which observing the confounder can block the spurious path, *e.g.*, given $C = c$, the path $X \leftarrow C \rightarrow Y$ is blocked.

2.3 Previous Methods

In this section, we will briefly introduce the previous methods that are related to our work, including the former state-of-the-arts, the common strategies and the popular solutions.

2.3.1 Learning to Compose Structures

In order to prevent the heuristic embeddings overfitting the dataset, and increase the transparency and robustness of intermediate processings, learning to compose dynamic structures is becoming popular in NLP for sentence representation, *e.g.*, Cho *et al.* [160] apply a gated recursive convolutional neural network (grConv) to control the bottom-up feature flow for a dynamic structure; Choi *et al.* [161] combine TreeLSTM with Gumbel-Softmax, allowing task-specific tree structures automatically learned from plain text. Yet, very few work composes visual structures for images. Conventional approaches construct a statistical dependency graph/tree for the entire dataset based on object categories [162] or exemplars [163]. Those statistical methods cannot put per-image objects in a context as a whole to reason over content-/task-specific fashion. Socher *et al.* [164] construct a bottom-up tree structure to parse images; however, their tree structure learning is supervised while ours is reinforced, which does not require tree ground truth.

2.3.2 Re-Balanced Training

The most widely-used strategy for long-tailed classification is arguably to re-balance the contribution of each class in the training phase. It can be either achieved by re-sampling [165–169] or re-weighting [39, 111, 170, 171]. Want *et al.* [145] propose a dynamic curriculum learning framework that gradually changes the sampling strategy from instance-balanced to class-balanced. Cui *et al.* [170] introduce an effective way to design the re-weighting loss based on the effective number of samples. Cao *et al.* [111] design a novel LDAM loss and manually adopt two-stage training, learning from easy to hard categories. However, they inevitably cause the under-fitting/over-fitting problem to head/tail classes. Besides, relying on the accessibility of data distribution also limits their application scope, *e.g.*, not applicable in online and streaming data.

2.3.3 Hard Example Mining

The hard example mining [144, 172, 173] is also a practical solution to increase the robustness of models. Instead of hacking the prior distribution of classes, focusing on the hard samples also alleviates the long-tailed issue. Jamal *et al.* [174] adopt meta-learning to dynamically learn the weights for each instance by iteratively optimize an inner loop. Li *et al.* [175] enhance the hard samples of by group softmax. Lin *et al.* [144] dynamically adjust the weight for each instance based on the current likelihood, giving higher weights to the hard samples.

2.3.4 Transfer Learning/Two-Stage Learning

Recent work shows a new trend of addressing the long-tailed problem by transferring the knowledge from data-rich head categories to data-scarce tail categories. Zhou *et al.* [16] design a weight-sharing bilateral-branch network that learning the instance-balanced sampler for one branch and class-balanced sampler for another branch. Kang *et al.* [11] propose a two-stage training strategy, which firstly learns from the biased data for better representations then freezes the backbone and fine-tune a balanced classifier. Liu *et al.* [10] use a memory module, called dynamic meta-embedding, to transfer the knowledge from head to tail. Liu *et al.* [146]

utilize the feature cloud to represent each category and transfer the distribution of head categories to enrich the tail. Yet, those methods either significantly increase the parameters or require a complicated training strategy.

2.3.5 Adversarial Defense

To increase the adversarial robustness, a variety of methods are proposed as defenders. Generally, the defenders fall into the following five categories: 1) adversarial training [47, 48] that intuitively generates adversarial examples to train the model, 2) data augmentation [132] that regards the adversarial vulnerability as the lack of training data, 3) generative classifier [128, 129] that attempts to find which category better generate the input image rather than directly outputs the prediction based on likelihood, 4) de-noising methods [51, 77] that consider adversarial perturbations as noises, which can be eliminated by either pre-network or in-network feature purification, and 5) certified defense [130], which tries to provide a safe bound of each image so any perturbation within the bound will not change the prediction.

Chapter 3

Dynamic Tree Structures for Robust Visual Contexts[1]

3.1 Introduction

Objects are not alone. They are placed in the visual context: a coherent object configuration attributed to the fact that they co-vary with each other. Extensive studies in cognitive science show that our brains can inherently exploit visual contexts to understand cluttered visual scenes comprehensively [135–137]. For example, even the girl's leg and the horse are not fully observed in Figure 3.1, we can still infer "girl riding horse". Inspired by this, modeling visual contexts is also indispensable in many modern computer vision systems. Yet, its implementation is tricky. For example, directly increasing the receptive field to the same size of images will degrade the CNN layers to fully connected layers, causing serious overfitting issues. To tackle this problem, state-of-the-art CNN architectures capture the context by convolutions of various receptive fields and encode it into multi-scale feature map pyramid [18, 55, 56]. Such robust pixel-level visual context (or local context [57]) arguably plays one of the key roles in closing the performance gap of

[1]The work in this chapter has been published in the paper : Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, Wei Liu. "Learning to Compose Dynamic Tree Structures for Visual Contexts." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR, Oral and Best Paper Finalists(45/5160)**). Long Beach, United States. 2019.

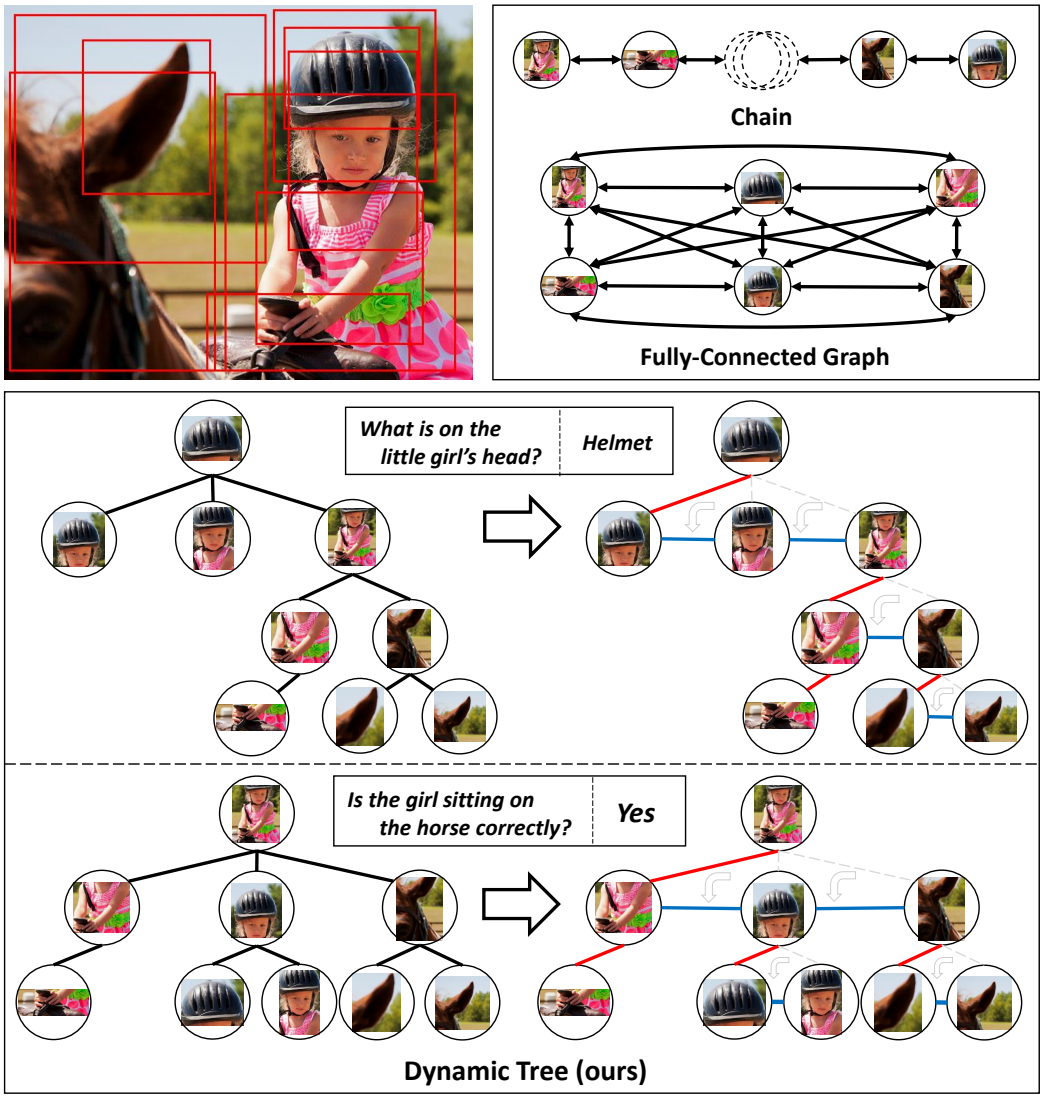


FIGURE 3.1: Illustrations of different object-level visual context structures: chains [1], fully-connected graphs [2], and dynamic tree structures constructed by the proposed VCTREE. For the purpose of efficiently and robustly encoding visual contexts by using TreeLSTM [3], we transform the multi-branch trees (left) to the equivalent left-child right-sibling binary trees [4], where the left branches (red) indicate the hierarchical relations and right branches (blue) indicate the parallel relations. The key advantages of VCTREE over chains and graphs are hierarchical, dynamic, robust, and efficient.

the "mid-level" vision between humans and machines, such as R-CNN based object detection [18, 27, 28], instance segmentation [29, 58], and FCN based semantic segmentation [55, 59, 60].

Except for the above implicit visual contexts at pixel levels, modeling visual contexts *explicitly* on the object level has also been shown effective in "high-level" vision tasks such as image captioning [176] and visual question answering [177]. In

fact, the visual context serves as a powerful inductive bias that connects objects in a particular layout for high-level reasoning [93, 176–178]. For example, the spatial layout of "person" on "horse" is useful for determining the relationship "ride", which is in turn informative to localize the "person" if we want to answer "who is riding on the horse?". However, those works assume that the context is a scene graph, whose detection *per se* is a high-level task and not yet reliable. Without high-quality scene graphs, we have to use a prior layout structure. As shown in Figure 3.1, two popular structures are chains [1] and fully-connected graphs [2, 61–64], where the context is encoded by sequential models such as bidirectional LSTM [65] for chains and CRF-RNN [66] for graphs.

However, unlike the robustness of pixel-level visual contexts that are well captured by dynamic receptive fields like feature pyramids or dilated convolutions, these two prior structures are sub-optimal and easily fooled by biased dataset. First, chains are oversimplified and may only capture simple spatial information or co-occurrence bias; though fully-connected graphs are complete, they lack the discrimination between hierarchical relations, *e.g.*, "helmet affiliated to head", and parallel relations, *e.g.*, "girl on horse"; in addition, dense connections could also lead to message passing saturation in the subsequent context encoding [2]. Second, visual contexts are inherently content-/task-driven, *e.g.*, the object layouts should vary from content to content, question to question. Therefore, fixed chains and graphs are incompatible with the dynamic nature of visual contexts [68].

In this chapter, we propose a robust context model dubbed VCTree, pioneering to compose dynamic tree structures for encoding object-level visual context for high-level visual reasoning tasks, such as scene graph generation (SGG) and visual question answering (VQA). Given a set of object proposals in an image (*e.g.*, obtained from Faster-RCNN [27]), we maintain a trainable task-specific score matrix of the objects, where each entry indicates the contextual validity of the pairwise objects. Then, a maximum spanning tree can be trimmed from the score matrix, *e.g.*, the multi-branch trees shown in Figure 3.1. This dynamic structure represents a "hard" hierarchical layout bias of what objects should gain more contextual information from others, *e.g.*, objects on the person's head are most informative given the question "what on the little girl's head?"; while the whole person's body is more important given the question "Is the girl sitting on the horse correctly?". To avoid memorizing the data bias caused by the densely connected arbitrary number

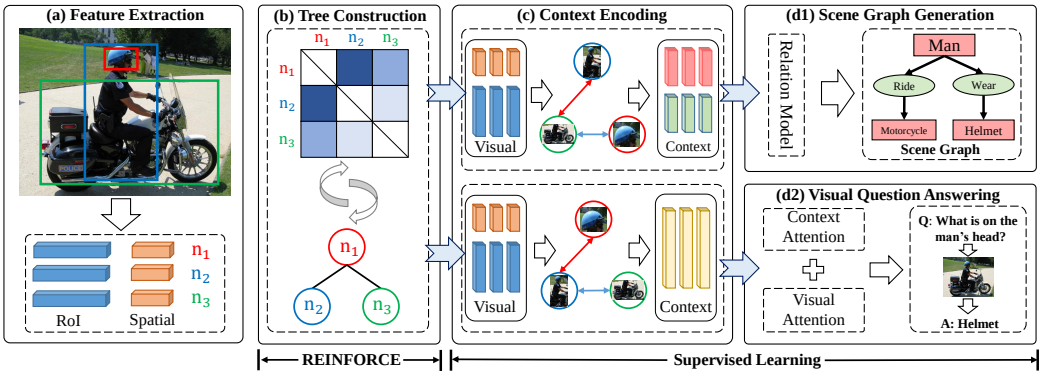


FIGURE 3.2: The framework of the proposed VCTREE model. We extract visual features from proposals and construct a dynamic VCTREE using the learnable score matrix. The tree structure is used to encode the object-level visual context, which will be decoded for each specific end-task. Parameters in stages (c)&(d) are trained by supervised learning, while those in stage (b) are using REINFORCE with a self-critic baseline.

of children, we further morph the multi-branch trees to the equivalent left-child right-sibling binary trees [4], where the left branches (red) indicate the hierarchical relations and right branches (blue) indicate the parallel relations, then use TreeLSTM [3] to encode the context.

As the above robust VCTREE construction is in a discrete and non-differentiable nature, we develop a hybrid learning strategy using REINFORCE [179–181] for tree structure exploration and supervised learning for context encoding and its subsequent tasks, which further increases the architectural robustness of the VCTree by reducing the effect of shortcuts in dataset. In particular, the evaluation result (Recall for SGG and Accuracy for VQA) from supervised task can be exploited as a "critic" function that guide the "action" of tree construction. We evaluate VCTREE on two benchmarks: Visual Genome [5] for SGG and VQA2.0 [107] for VQA. For SGG, we achieve a new state-of-the-art on all three standard tasks, *i.e.*, Scene Graph Generation, Scene Graph Classification, and Predicate Classification; for VQA, we achieve competitive results on single model performances. In particular, VCTREE helps high-level vision models fight against the dataset bias. For example, we achieve 4.1% absolute gain in proposed Mean Recall@100 metric of Predicate Classification than MOTIFS [1], and observe higher improvement in VQA2.0 balanced pair subset [103] than normal validation set. Qualitative results also show that VCTREE composes interpretable structures.

3.2 Approach

As illustrated in Figure 3.2, our robust VCTree model can be summarized into the following four steps. (a) We adopt Faster-RCNN to detect object proposals [27]. The visual feature of each proposal i is presented as $\boldsymbol{x}_i$, concatenating a RoIAlign feature [29] $\boldsymbol{v}_i \in \mathbb{R}^{2048}$ and spatial feature $\boldsymbol{b}_i \in \mathbb{R}^8$, where 8 elements indicate the bounding box coordinates (x_1, y_1, x_2, y_2), center $(\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2})$, and size $(x_2 - x_1, y_2 - y_1)$, respectively. Note that the visual feature $\boldsymbol{x}_i$ is not limited to bounding box; segment feature from instance segmentations [29] or panoptic segmentations [30] could also be alternatives. (b) In Section 3.2.1, a learnable matrix will be introduced to construct VCTree. Moreover, since the VCTree construction is discrete in nature and the score matrix is non-differentiable from the loss of end-task, we develop a hybrid learning strategy to explore the optimal structure in Section 3.2.5. (c) In Section 3.2.2, we employ Bidirectional Tree LSTM (BiTreeLSTM) to encode the contextual cues using the constructed VCTree. (d) The encoded contexts will be decoded for each specific end-task detailed in Section 3.2.3 and Section 3.2.4.

3.2.1 VCTree Construction

VCTree construction aims to learn a score matrix $\boldsymbol{S}$, which approximates the task-dependent validity between each object pair. Two principles guide the formulation of this matrix: 1) inherent object correlations should be maintained, *e.g.*, "man wears helmet" in Figure 3.2; (2) task related object pair has higher score than irrelevant ones, *e.g.*, given question "what is on the man's head?", "man-helmet" pair should be more important than "man-motorcycle" and "helmet-motorcycle" pairs. Therefore, we define each element of $\boldsymbol{S}$ as the product of the object correlation score function $f(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and the pairwise task-dependency score function $g(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{q})$:

$$\begin{cases} \boldsymbol{S}_{ij} = f(\boldsymbol{x}_i, \boldsymbol{x}_j) \cdot g(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{q}), \\ f(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma\left(\mathrm{MLP}(\boldsymbol{x}_i, \boldsymbol{x}_j)\right), \\ g(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{q}) = \sigma(h(\boldsymbol{x}_i, \boldsymbol{q})) \cdot \sigma(h(\boldsymbol{x}_j, \boldsymbol{q})), \end{cases} \tag{3.1}$$

where $\sigma(\cdot)$ is the sigmoid function; $\boldsymbol{q}$ is the task feature, *e.g.*, the question feature encoded by GRU in VQA; MLP is a multi-layer perceptron; $h(\boldsymbol{x}_i, \boldsymbol{q})$ is the object-task correlation in VQA, which will be introduced later in Section 3.2.4. In SGG,

the entire $g(\boldsymbol{x}_i, \boldsymbol{x}_j, \boldsymbol{q})$ is set to 1, as we assume that each object pair contributes equally without the question prior. We pretrain $f(\boldsymbol{x}_i, \boldsymbol{x}_j)$ on Visual Genome [5] for a reasonable binary prior if two objects are related. Yet, such a pretrained model is not perfect due to the lack of coherent graph-level constraint or question prior, so it will be further fine-tuned in Section 3.2.5.

Considering $\boldsymbol{S}$ as a symmetric adjacency matrix, we can obtain a maximum spanning tree using the Prim's algorithm [182], with a root, *i.e.*, the source node i, satisfying $\arg\max_i \sum_{j \neq i} \boldsymbol{S}_{ij}$. In a nutshell, as illustrated in the Figure 3.3, we construct the tree recursively by connecting the node from the pool to the tree node if it has the most validity. Note that during the tree structure exploration in Section 3.2.5, each of the i-th step $t^{(i)}$ in the above tree construction is sampled from all possible choices in a multinomial distribution with the probability $p(t^{(i)}|t^{(1)}, ..., t^{(i-1)}, \boldsymbol{S})$ in proportion to the validity score.

However, the resultant tree structure is multi-branch and merely a sparse graph with only one type of connection, which is still unable to discriminate the hierarchical and parallel relations in the subsequent context encoding. To this end, we convert the multi-branch tree into an equivalent binary tree, *i.e.*, VCTree by changing non-leftmost edges into right branches as in Figure 3.1. In this fashion, the right branches (blue) indicate parallel contexts, and left ones (red) indicate hierarchical contexts. Such a binary tree structure achieves significant improvements on the architectural robustness as its deeper connects in TreeLSTM allow automatically neglecting trivial objects, which reduces the chance of taking shortcuts from irrelevant objects in the given tasks. Consequently, the proposed VCTree achieves better performances in the SGG and VQA experiments compared to its multi-branch alternative.

3.2.2 TreeLSTM Context Encoding

Given the above constructed VCTree, we adopt Bidirectional TreeLSTM (BiTreeLSTM) as our context encoder:

$$D = \text{BiTreeLSTM}(\{\boldsymbol{z}_i\}_{i=1,2,...,n}), \tag{3.2}$$

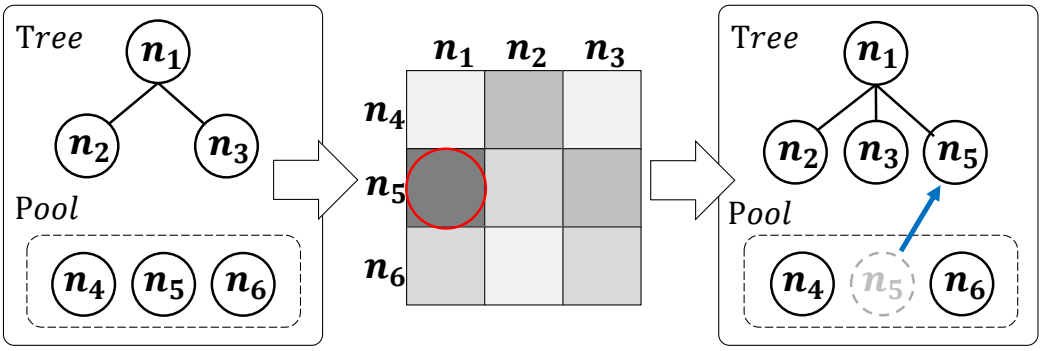


FIGURE 3.3: The maximum spanning tree from $\boldsymbol{S}$. In each step, a node in the remaining pool is connected to the current tree, if it has the highest validity score.

where $\boldsymbol{z}_i$ is the input node feature, which will be specified in each task, and $D = [\boldsymbol{d}_1, \boldsymbol{d}_2, ..., \boldsymbol{d}_n]$ is the encoded object-level visual context. Each $\boldsymbol{d}_i = [\vec{\boldsymbol{h}}_i; \overleftarrow{\boldsymbol{h}}_i]$ is the concatenated hidden states from both TreeLSTM [3] directions:

$$\vec{\boldsymbol{h}}_i = \text{TreeLSTM}(\boldsymbol{z}_i, \vec{\boldsymbol{h}}_p), \tag{3.3}$$

$$\overleftarrow{\boldsymbol{h}}_i = \text{TreeLSTM}(\boldsymbol{z}_i, [\overleftarrow{\boldsymbol{h}}_l; \overleftarrow{\boldsymbol{h}}_r]), \tag{3.4}$$

where $\rightarrow$and$\leftarrow$ denote the top-down and bottom-up directions, respectively; we slightly abuse the subscripts p, l, r to denote the parent, left child, and right child of node i. The order of the concatenation $[\overleftarrow{\boldsymbol{h}}_l; \overleftarrow{\boldsymbol{h}}_r]$ in Eq. (3.4) indicates the explicit discrimination between the left and right branches in context encoding. We use zero vectors to pad all the missing branches.

3.2.3 Scene Graph Generation Model

Now we detail the implementation of Eq. (3.2) and how to decode them for the SGG task as illustrated in Figure 3.4.

Object Context Encoding. We employ BiTreeLSTM from Eq. (3.2) to encode object context representation into $D^o = [\boldsymbol{d}_1^o, \boldsymbol{d}_2^o, ..., \boldsymbol{d}_n^o], \boldsymbol{d}_i^o \in \mathbb{R}^{512}$. We set inputs $\boldsymbol{z}_i$ of Eq. (3.2) to $[\boldsymbol{x}_i; \boldsymbol{W}_1 \hat{\boldsymbol{c}}_i]$, *i.e.*, concatenation of object visual features and embedded N-way original Faster-RCNN class probabilities, where $\boldsymbol{W}_1$ is the embedding matrix that maps each original label distribution $\hat{\boldsymbol{c}}_i$ into $\mathbb{R}^{200}$.

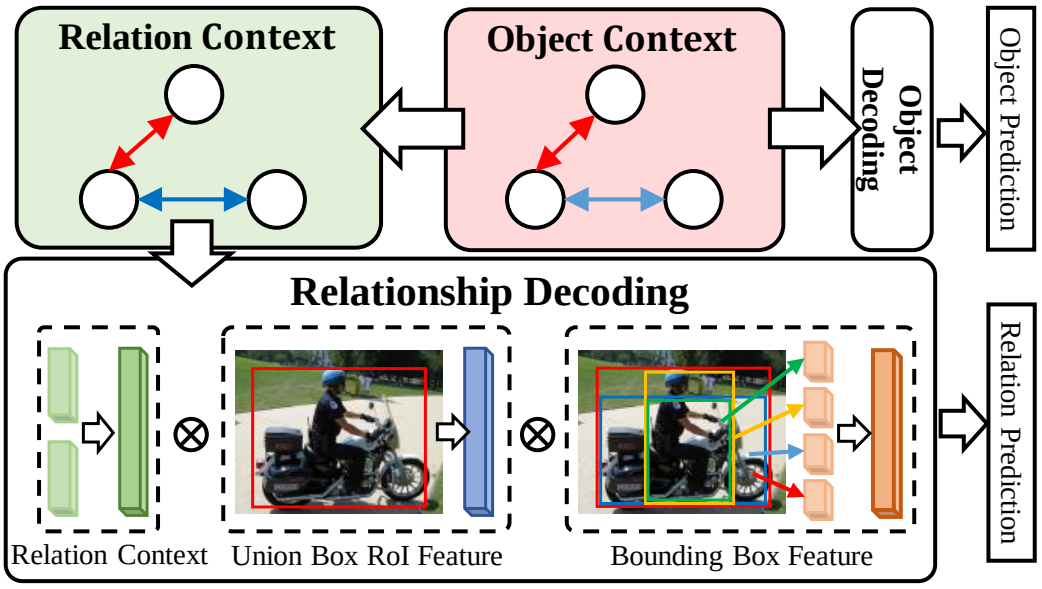


FIGURE 3.4: The overview of our SGG Model. The object context feature will be used to decode object categories, and the pairwise relationship decoding jointly fuses the relation context feature, RoIAlign feature of union box, and bounding box feature, before prediction.

Relation Context Encoding. We apply an additional BiTreeLSTM using the above $\boldsymbol{d}_i^o$ as input $\boldsymbol{z}_i$ to further encode the relation context $D^r = [\boldsymbol{d}_1^r, \boldsymbol{d}_2^r, ..., \boldsymbol{d}_n^r], \boldsymbol{d}_i^r \in \mathbb{R}^{512}$.

Context Decoding. The goal of SGG is to detect objects and then predict their relationship. Similar to the framework of [1], we adopt a dynamic object prediction which can be viewed as a decoding process in a top-down direction using Eq. (3.3), that is, the object class of a child is dependent on its parent. Specifically, we set the input $\boldsymbol{z}_i$ of Eq. (3.3) to be $[\boldsymbol{d}_i^o; \boldsymbol{W}_2\boldsymbol{c}_p]$, where $\boldsymbol{c}_p$ is the predicted label distribution of the i's parent, and $\boldsymbol{W}_2$ embeds it into $\mathbb{R}^{200}$, then the output hidden is passed to a softmax classifier to achieve object label distribution $\boldsymbol{c}_i$.

The relationship prediction is in a pairwise fashion. First, we collect three pairwise features for each object pair: (1) $\boldsymbol{d}_{ij} = \text{MLP}([\boldsymbol{d}_i^r; \boldsymbol{d}_j^r])$ as the context feature, (2) $\boldsymbol{b}_{ij} = \text{MLP}([\boldsymbol{b}_i; \boldsymbol{b}_j; \boldsymbol{b}_{i\cup j}; \boldsymbol{b}_{i\cap j}])$ as the bounding box pair feature, with $i\cup j, i\cap j$ being union box and intersection box, (3) $\boldsymbol{v}_{ij}$ as the RoIAlign feature [29] from the union bounding box of the object pair. All $\boldsymbol{d}_{ij}, \boldsymbol{v}_{ij}, \boldsymbol{b}_{ij}$ are under the same dimension $\mathbb{R}^{2048}$. Then, we fuse them into a final pairwise feature: $\boldsymbol{g}_{ij} = \boldsymbol{d}_{ij} \cdot \boldsymbol{v}_{ij} \cdot \boldsymbol{b}_{ij}$, before feed it into the softmax predicate classifier, where $\cdot$ is element-wise product.

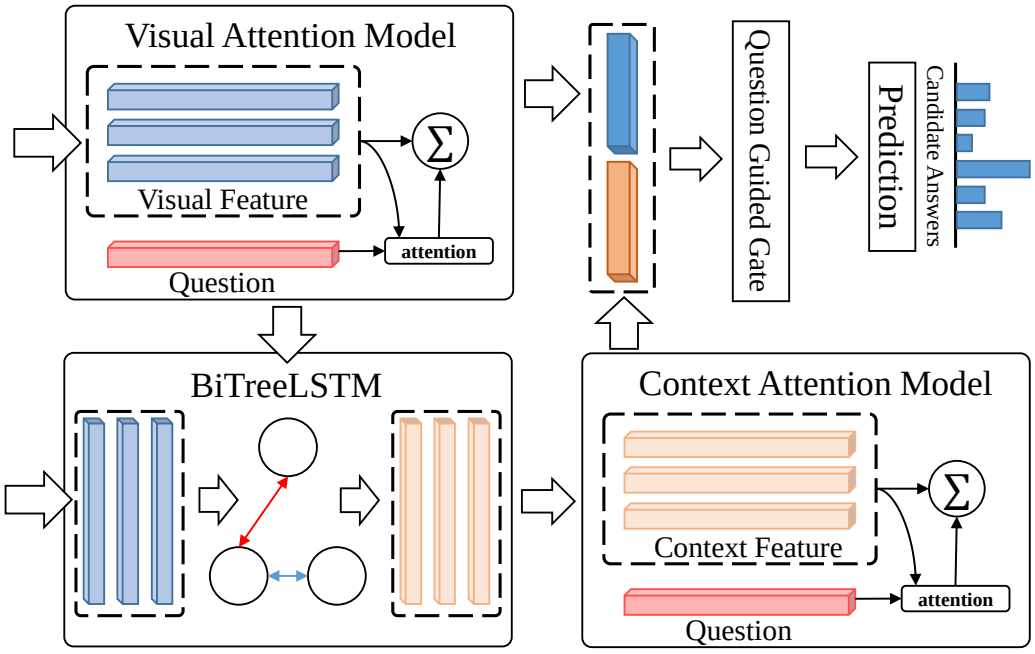


FIGURE 3.5: The overview of our VQA framework. It contains two multimodal attention models for visual feature and context feature. Outputs from both models will be concatenated and passed to a question-guided gate before answer prediction.

3.2.4 Visual Question Answering Model

Now we detail the implementation of Eq. (3.2) for VQA, and illustrate our VQA model in Figure 3.5.

Context Encoding. The context feature in VQA: $D^q = [\boldsymbol{d}_1^q, \boldsymbol{d}_2^q, ..., \boldsymbol{d}_n^q]$, $\boldsymbol{d}_i^q \in \mathbb{R}^{1024}$ is directly encoded from the bounding box visual feature $\boldsymbol{x}_i$ by Eq. (3.2).

Multimodal Attention Feature. We adopt a popular attention model from previous work [102, 103] to calculate the multimodal joint feature $\boldsymbol{m} \in \mathbb{R}^{1024}$ for each question and image pair:

$$\boldsymbol{m} = f_d(\hat{\boldsymbol{z}}, \boldsymbol{q}), \tag{3.5}$$

where $\boldsymbol{q} \in \mathbb{R}^{1024}$ is the question feature from a one-layer GRU encoding the sentence; $\hat{\boldsymbol{z}} = \sum_{i=1}^{N} \alpha_i \boldsymbol{z}_i$ is the attentive image feature calculated from the input feature set $\{\boldsymbol{z}_i\}$, $\alpha_i = \exp(u_i) / \sum_k \exp(u_k)$ is the attention weight from object-task correlation $u_i = h(\boldsymbol{z}_i, \boldsymbol{q}) = \text{MLP}\big(f_d(\boldsymbol{z}_i, \boldsymbol{q})\big)$, with the output of MLP being a scalar; f_d can be any multi-modal feature fusion function, in particular, we adopt

$f_d(\boldsymbol{x}, \boldsymbol{y}) = \text{ReLU}(\boldsymbol{W}_3\boldsymbol{x} + \boldsymbol{W}_4\boldsymbol{y}) - (\boldsymbol{W}_3\boldsymbol{x} - \boldsymbol{W}_4\boldsymbol{y})^2$ as in [183], with $\boldsymbol{W}_3$ and $\boldsymbol{W}_4$ projecting $\boldsymbol{x}, \boldsymbol{y}$ into the same dimension. Therefore, we can use Eq. (3.5) to obtain both the multimodal visual attention feature $\boldsymbol{m}_x$ by setting input $\boldsymbol{z}_i$ to $\boldsymbol{x}_i$ and multimodal contextual attention feature $\boldsymbol{m}_d$ by setting $\boldsymbol{z}_i$ to $\boldsymbol{d}_i^q$.

Question Guided Gate Decoding. However, the importance of $\boldsymbol{m}_x$ and $\boldsymbol{m}_d$ could vary from question to question. For example, "is there a dog?" only requires visual features for detection, while "is the man dressed formally?" is highly context dependent. Inspired by [184], we adopt a question guided gate to select the most related channels from $[\boldsymbol{m}_x; \boldsymbol{m}_d]$. The gate vector $\boldsymbol{g} \in \mathbb{R}^{2048}$ is defined as:

$$\boldsymbol{g} = \sigma\big(\text{MLP}([\boldsymbol{q}; \boldsymbol{W}_5\boldsymbol{l}_q])\big), \tag{3.6}$$

where $\boldsymbol{l}_q \in \mathbb{R}^{65}$ is a one-hot question type vector defined by prefixed words of questions, which is embedded into $\mathbb{R}^{256}$ by matrix $\boldsymbol{W}_5$, and $\sigma(\cdot)$ denotes the sigmoid function.

Finally, we fuse $\boldsymbol{g} \cdot [\boldsymbol{m}_x; \boldsymbol{m}_d]$ as the final VQA feature and feed it into the softmax classifier.

3.2.5 Hybrid Learning

Due to the discrete nature of VCTree construction, the score matrix $\boldsymbol{S}$ is not fully differentiable from the loss back-propagated from the end-task loss. Inspired by [179], we use a hybrid learning strategy that combines reinforcement learning, *i.e.*, policy gradient [181] for the parameters θ of $\boldsymbol{S}$ in the tree construction and supervised learning for the rest parameters. It also allows VCTree to explore more structures without taking shortcuts of data bias, increasing the architectural robustness of the model.

Suppose a layout l, *i.e.*, a constructed VCTree, is sampled from $\pi(l|I, q; \theta)$, *i.e.*, the construction procedure explained in Section 3.2.1, where I is the given image, q is the task, *e.g.*, questions in VQA. To avoid clutter, we drop I and q. Then, we define the reinforcement learning loss $L_r(\theta)$ as:

$$L_r(\theta) = -E_{l\sim\pi(l|\theta)}[r(l)], \tag{3.7}$$

where $L_r(\theta)$ aims to minimize the negative expected reward $r(l)$, which can be the end-task evaluation metrics such as Recall@100 for SGG and Accuracy for VQA. Then, the above gradient will be $\nabla_\theta L_r(\theta) = -E_{l\sim\pi(l|\theta)}[r(l)\nabla_\theta log\pi(l|\theta)]$. Since it is impractical to estimate all possible layouts, we use the Monte-Carlo sampling to estimate the gradient:

$$\nabla_\theta L_r(\theta) \approx -\frac{1}{M}\sum_{m=1}^{M}\Big(r(l_m)\nabla_\theta log\pi(l_m|\theta)\Big), \tag{3.8}$$

where we set M to 1 in our implementation.

To reduce the gradient variance, we apply a self-critic baseline [180] $b = r(\hat{l})$, where $\hat{l}$ is the greedy constructed tree without sampling. So the original reward $r(l_m)$ can be replaced by $r(l_m) - b$ in Eq. (3.8). We observe faster convergence than using the traditional moving baseline [185].

The overall hybrid learning will be alternatively conducted between supervised learning and reinforcement learning, where we first train the supervised end-task on pretrained $\pi(l|\theta)$, then fix the end-task as reward function to learn our reinforcement policy network, after that, we update the supervised end-task by new $\pi(l|\theta)$. The latter two stages are running alternatively 2 times in our model.

3.3 Bidirectional TreeLSTM

In this section, we will introduce the details of the bidirectional TreeLSTM applied to encode the object-level visual contexts. For the bottom-up direction, we employ N-ary TreeLSTM [3] for binary trees, *i.e.*, VCTrees and Overlap Trees, and the normalized Child-Sum [3] TreeLSTM for Multi-Branch Trees used in ablation studies. For the top-down direction, since each node only has one parent, TreeLSTM is similar to the traditional LSTM [65].

3.3.1 N-ary TreeLSTM for Binary Trees

According to the definition of N-ary TreeLSTM [3], it can be applied to the tree structures with at most N ordered branches for each node. In our work, we adopt

binary TreeLSTM as our bottom-up TreeLSTM for the proposed binary tree structures, *i.e.*, VCTrees and Overlap Trees. It can be formulated as follows:

$$\overleftarrow{\boldsymbol{h}}_t = \text{TreeLSTM}\big(\boldsymbol{z}_t, [\overleftarrow{\boldsymbol{h}}_l; \overleftarrow{\boldsymbol{h}}_r]\big), \tag{3.9}$$

$$\boldsymbol{i}_t = \sigma\big(\boldsymbol{W}^{(i)}\boldsymbol{z}_t + \boldsymbol{U}^{(i)}[\overleftarrow{\boldsymbol{h}}_l; \overleftarrow{\boldsymbol{h}}_r] + \boldsymbol{b}^{(i)}\big), \tag{3.10}$$

$$\boldsymbol{f}_l = \sigma\big(\boldsymbol{W}_l^{(f)}\boldsymbol{z}_t + \boldsymbol{U}_l^{(f)}[\overleftarrow{\boldsymbol{h}}_l; \overleftarrow{\boldsymbol{h}}_r] + \boldsymbol{b}_l^{(f)}\big), \tag{3.11}$$

$$\boldsymbol{f}_r = \sigma\big(\boldsymbol{W}_r^{(f)}\boldsymbol{z}_t + \boldsymbol{U}_r^{(f)}[\overleftarrow{\boldsymbol{h}}_l; \overleftarrow{\boldsymbol{h}}_r] + \boldsymbol{b}_r^{(f)}\big), \tag{3.12}$$

$$\boldsymbol{o}_t = \sigma\big(\boldsymbol{W}^{(o)}\boldsymbol{z}_t + \boldsymbol{U}^{(o)}[\overleftarrow{\boldsymbol{h}}_l; \overleftarrow{\boldsymbol{h}}_r] + \boldsymbol{b}^{(o)}\big), \tag{3.13}$$

$$\boldsymbol{u}_t = \tanh\big(\boldsymbol{W}^{(u)}\boldsymbol{z}_t + \boldsymbol{U}^{(u)}[\overleftarrow{\boldsymbol{h}}_l; \overleftarrow{\boldsymbol{h}}_r] + \boldsymbol{b}^{(u)}\big), \tag{3.14}$$

$$\overleftarrow{\boldsymbol{c}}_t = \boldsymbol{i}_t \odot \boldsymbol{u}_t + \boldsymbol{f}_l \odot \overleftarrow{\boldsymbol{c}}_l + \boldsymbol{f}_r \odot \overleftarrow{\boldsymbol{c}}_r, \tag{3.15}$$

$$\overleftarrow{\boldsymbol{h}}_t = \boldsymbol{o}_t \odot \tanh(\overleftarrow{\boldsymbol{c}}_t), \tag{3.16}$$

where $\boldsymbol{z}_t \in \mathbb{R}^d$ is the input feature for node t; $\overleftarrow{\boldsymbol{h}}_t, \overleftarrow{\boldsymbol{h}}_l, \overleftarrow{\boldsymbol{h}}_r \in \mathbb{R}^h$ are the hidden states; $\overleftarrow{\boldsymbol{c}}_t, \overleftarrow{\boldsymbol{c}}_l, \overleftarrow{\boldsymbol{c}}_r \in \mathbb{R}^h$ are memory cells; $\boldsymbol{W}^{(i)}, \boldsymbol{W}_l^{(f)}, \boldsymbol{W}_r^{(f)}, \boldsymbol{W}^{(o)}, \boldsymbol{W}^{(u)} \in \mathbb{R}^{h\times d}$ and $\boldsymbol{U}^{(i)}, \boldsymbol{U}_l^{(f)}, \boldsymbol{U}_r^{(f)}, \boldsymbol{U}^{(o)}, \boldsymbol{U}^{(u)} \in \mathbb{R}^{h\times 2h}$ are learnable matrices; $\boldsymbol{b}^{(i)}, \boldsymbol{b}_l^{(f)}, \boldsymbol{b}_r^{(f)}, \boldsymbol{b}^{(o)}, \boldsymbol{b}^{(u)} \in \mathbb{R}^h$ are vectors; σ denotes sigmoid function; tanh denotes tanh activation function; $\odot$ means element-wise product. Note that we slightly abuse the subscripts l, r of $\overleftarrow{\boldsymbol{c}}_l, \overleftarrow{\boldsymbol{c}}_r, \overleftarrow{\boldsymbol{h}}_l, \overleftarrow{\boldsymbol{h}}_r$ to denote hidden states and memory cells from the left-child and right-child of node t. The hidden states and memory cells of the missing branches will be filled with zero vectors.

3.3.2 Child-Sum TreeLSTM for Multi-Branch Trees

The Child-Sum TreeLSTM [3] is able to deal with the tree structure where each node has arbitrary number of children. Therefore, we adopt it as the bottom-up TreeLSTM of the context encoder for the Multi-Branch Trees in the ablation studies. For each node t of a Multi-Branch Tree, we define $C(t)$ as the set of its children. Compared with the original paper [3], we replace the Child-Sum with the Child-Mean in our implementation for better normalization, then it is formulated

as:

$$\overleftarrow{\boldsymbol{h}}_t = \text{TreeLSTM}\big(\boldsymbol{z}_t, \{\overleftarrow{\boldsymbol{h}}_k\}\big), k \in C(t), \tag{3.17}$$

$$\overleftarrow{\boldsymbol{h}}_{mean} = \frac{\sum_{k \in C(t)} \overleftarrow{\boldsymbol{h}}_k}{|C(t)|}, \tag{3.18}$$

$$\boldsymbol{i}_t = \sigma\big(\boldsymbol{W}^{(i)}\boldsymbol{z}_t + \boldsymbol{U}^{(i)}\overleftarrow{\boldsymbol{h}}_{mean} + \boldsymbol{b}^{(i)}\big), \tag{3.19}$$

$$\boldsymbol{f}_k = \sigma\big(\boldsymbol{W}^{(f)}\boldsymbol{z}_t + \boldsymbol{U}^{(f)}\overleftarrow{\boldsymbol{h}}_k + \boldsymbol{b}^{(f)}\big), \tag{3.20}$$

$$\boldsymbol{o}_t = \sigma\big(\boldsymbol{W}^{(o)}\boldsymbol{z}_t + \boldsymbol{U}^{(o)}\overleftarrow{\boldsymbol{h}}_{mean} + \boldsymbol{b}^{(o)}\big), \tag{3.21}$$

$$\boldsymbol{u}_t = \tanh\big(\boldsymbol{W}^{(u)}\boldsymbol{z}_t + \boldsymbol{U}^{(u)}\overleftarrow{\boldsymbol{h}}_{mean} + \boldsymbol{b}^{(u)}\big), \tag{3.22}$$

$$\overleftarrow{\boldsymbol{c}}_t = \boldsymbol{i}_t \odot \boldsymbol{u}_t + \frac{\sum_{k \in C(t)} \boldsymbol{f}_k \odot \overleftarrow{\boldsymbol{c}}_k}{|C(t)|}, \tag{3.23}$$

$$\overleftarrow{\boldsymbol{h}}_t = \boldsymbol{o}_t \odot \tanh(\overleftarrow{\boldsymbol{c}}_t), \tag{3.24}$$

where $\overleftarrow{\boldsymbol{h}}_t, \overleftarrow{\boldsymbol{h}}_k, \in \mathbb{R}^h$ are the hidden states; $\overleftarrow{\boldsymbol{c}}_t, \overleftarrow{\boldsymbol{c}}_k \in \mathbb{R}^h$ are memory cells; $\boldsymbol{W}^{(i)}$, $\boldsymbol{W}^{(f)}$, $\boldsymbol{W}^{(o)}$, $\boldsymbol{W}^{(u)} \in \mathbb{R}^{h \times d}$ and $\boldsymbol{U}^{(i)}, \boldsymbol{U}^{(f)}, \boldsymbol{U}^{(o)}, \boldsymbol{U}^{(u)} \in \mathbb{R}^{h \times h}$ are learnable matrices; $\boldsymbol{b}^{(i)}, \boldsymbol{b}^{(f)}, \boldsymbol{b}^{(o)}, \boldsymbol{b}^{(u)} \in \mathbb{R}^h$ are vectors; $|C(t)|$ is the number of children for node t; $\overleftarrow{\boldsymbol{h}}_{mean}$ denotes the mean hidden state of all the children of node t.

3.3.3 Top-Down TreeLSTM

We use the traditional LSTM [65] as the top-down TreeLSTM for all the VCTrees, Overlap Trees, and Multi-Branch Trees, because each node only has at most one parent. The only difference with the traditional LSTM is that our structures are trees rather than chains, the previous hidden state is from the parent of node t.

For the proposed VCTree, we assigned different learnable matrices for the hidden states from the left-branch parents and right-branch parents. However, the result didn't show significant improvements in the end-tasks, so we employ traditional LSTM as our top-down LSTM for efficiency.

	Scene Graph Generation			Scene Graph Classification			Predicate Classification		
Model	R@20	R@50	R@100	R@20	R@50	R@100	R@20	R@50	R@100
VRD [186]	-	0.3	0.5	-	11.8	14.1	-	27.9	35.0
AsscEmbed [98]	6.5	8.1	8.2	18.2	21.8	22.6	47.9	54.1	55.4
IMP$^\diamond$ [2]	14.6	20.7	24.5	31.7	34.6	35.4	52.7	59.3	61.3
TFR [187]	3.4	4.8	6.0	19.6	24.3	26.6	40.1	51.9	58.3
FREQ$^\diamond$ [1]	20.1	26.2	30.1	29.3	32.3	32.9	53.6	60.6	62.2
MOTIFS$^\diamond$ [1]	21.4	27.2	30.3	32.9	35.8	36.5	58.5	65.2	67.1
Graph-RCNN [188]	-	11.4	13.7	-	29.6	31.6	-	54.2	59.1
Chain	21.2	27.1	30.3	33.3	36.1	36.8	59.4	66.0	67.7
Overlap	21.4	27.3	30.4	33.7	36.5	37.1	59.5	66.0	67.8
Multi-Branch	21.5	27.3	30.6	34.3	37.1	37.8	59.5	66.1	67.8
VCTree-SL	21.7	27.7	31.1	35.0	37.9	38.6	59.8	66.2	67.9
VCTree-HL	**22.0**	**27.9**	**31.3**	**35.2**	**38.1**	**38.8**	**60.1**	**66.4**	**68.1**

TABLE 3.1: SGG performances (%) of various methods. $^\diamond$ denotes the methods using the same Faster-RCNN detector as ours. IMP$^\diamond$ is reported from the re-implemented version [1].

3.4 Experiments on Scene Graph Generation

3.4.1 Dataset

Visual Genome (VG) [5] is a popular benchmark for SGG. It contains 108,077 images with tens of thousands of unique object and predicate relation categories, yet most of categories have very limited instances. Therefore, previous works [2, 93, 189] proposed various VG splits that remove rare categories. We adopted the most popular one from [2], which selects top-150 object categories and top-50 predicate categories by frequency. The entire dataset is divided into the training set and test set by 70%, 30%, respectively. We further picked 5,000 images from training set as the validation set for hyper-parameter tuning.

3.4.2 Protocols

We followed three conventional protocols to evaluate our SGG model: (1) **Scene Graph Generation (SGGen)**: given an image, detect object bounding boxes and their categories, and predict their relationships; (2) **Scene Graph Classification (SGCls)**: given ground-truth object bounding boxes in an image, predict the object categories and their relationships; (3) **Predicate Classification (PredCls)**: given the object categories and their bounding boxes in the image, predict their relationships.

	Scene Graph Generation			Scene Graph Classification			Predicate Classification		
Model	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100	mR@20	mR@50	mR@100
MOTIFS [1]	4.2	5.7	6.6	6.3	7.7	8.2	10.8	14.0	15.3
FREQ [1]	4.5	6.1	7.1	5.1	7.2	8.5	8.3	13.0	16.0
Chain	4.6	6.3	7.2	6.3	7.9	8.8	11.0	14.4	16.6
Overlap	4.8	6.5	7.5	7.2	9.0	9.3	12.5	16.1	17.4
Multi-Branch	4.7	6.5	7.4	6.9	8.6	9.2	11.9	15.5	16.9
VCTree-SL	5.0	6.7	7.7	8.0	9.8	10.5	13.4	17.0	18.5
VCTree-HL	**5.2**	**6.9**	**8.0**	**8.2**	**10.1**	**10.8**	**14.0**	**17.9**	**19.4**

TABLE 3.2: Mean recall (%) of various methods across all the 50 predicate categories. MOTIFS [1] and FREQ [1] are using the same Faster-RCNN detector as ours.

3.4.3 Metrics

Since the annotation in VG is incomplete and biased, we followed the conventional Recall@K (R@K = 20,50,100) as the evaluation metrics [1, 2, 186]. However, it is well-known that SGG models trained on biased datasets such as VG have low performances for less frequent categories. To this end, we introduced a balanced metric called **Mean Recall (mR@K)** for better evaluation of architectural robustness of the proposed VCTree. It calculates the recall on each predicate category independently, and then averages the results. So, each category contributes equally. Such a metric reduces the influence of some common yet meaningless predicates, *e.g.*, "on", "of", and gives equal attention to those infrequent predicates, *e.g.*, "riding", "carrying", which are more valuable to high-level reasoning.

3.4.4 Implementation Details

We adopted Faster-RCNN [27] with VGG backbone to detect object bounding boxes and extract RoI features. Since the performance of SGG highly depends on the underlying detector, we used the same set of parameters as [1] for fair comparison. Object correlations $f(\boldsymbol{x}_i, \boldsymbol{x}_j)$ in Eq. (3.1) will be pretrained on ground-truth bounding boxes with class-agnostic relationships (*i.e.*, foreground/background relationships), using all possible symmetric pairs without sampling. In SGGen, top-64 object proposals were selected after non-maximal suppression (NMS) with 0.3 IoU. We set background/foreground ratio for predicate classification to 3, and capped the number of training samples at 64 (retained all foreground pairs if possible). Our model is optimized by SGD with momentum, using learning rate $lr = 6 \cdot 10^{-3}$ and

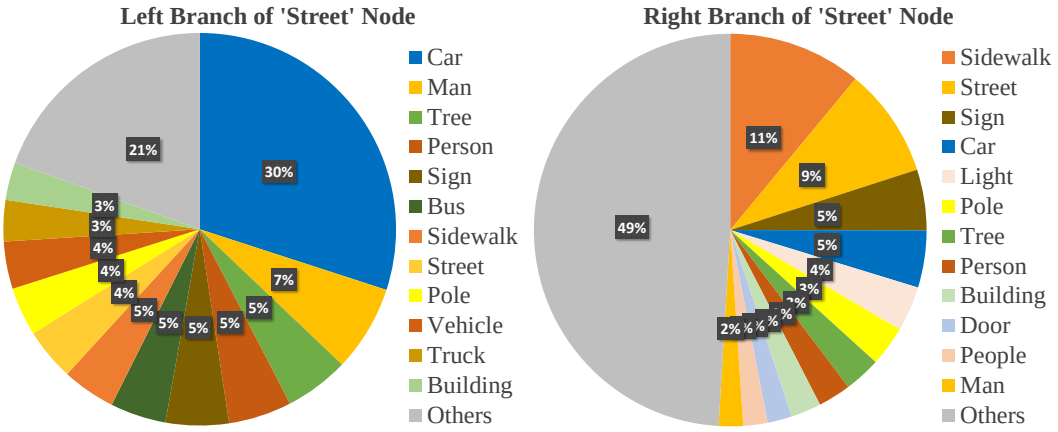


FIGURE 3.6: The statistics of left-branch (hierarchical) nodes and right-branch (parallel) nodes of the "street" category.

batch size $b = 5$ for supervised learning, and $lr = 6 \cdot 10^{-4}, b = 1$ for reinforcement learning.

3.4.5 Ablation Studies

We investigated the influence of different structure construction policies. They are reported on the bottom half of Table 3.1. The ablative methods are (1) **Chain**: sorting all the objects by $\sum_{j:j\neq i} \boldsymbol{S}_{ij}$, then constructing a chain, which is different from the left-to-right ordered chain in MOTIFS [1]; (2) **Overlap**: iteratively constructing a binary tree by selecting the node with largest number of overlapped objects as parent, and dividing the rest nodes into left/right sub-trees by relatively positions of their bounding boxes; (3) **Multi-Branch**: the maximum spanning tree generated from score matrix $\boldsymbol{S}$, using Child-Sum TreeLSTM [3] to incorporate context; (4) **VCTree-SL**: the proposed VCTREE trained by supervised learning; (5) **VCTree-HL**: the complete version of VCTREE, trained by hybrid learning for structure exploration in Section 3.2.5. As we will show that Multi-Branch is significantly worse than VCTREE, so there is no need to conduct hybrid learning experiment on Multi-Branch. We observe that VCTREE performs better than other structures, and it is further improved by hybrid learning for structure exploration.

VQA2.0 Validation Accuracy					
Model	Yes/No	Number	Other	All	Balanced Pairs
Graph	81.8	44.9	56.6	64.5	36.3
Chain	81.8	44.5	56.9	64.6	36.3
Overlap	81.8	44.8	57.0	64.7	36.4
Multi-Branch	82.1	44.3	56.9	64.7	36.6
VCTree-SL	82.3	45.0	57.0	64.9	36.9
VCTree-HL	**82.6**	**45.1**	**57.1**	**65.1**	**37.2**

TABLE 3.3: Accuracies (%) of various context structures on the VQA2.0 validation set.

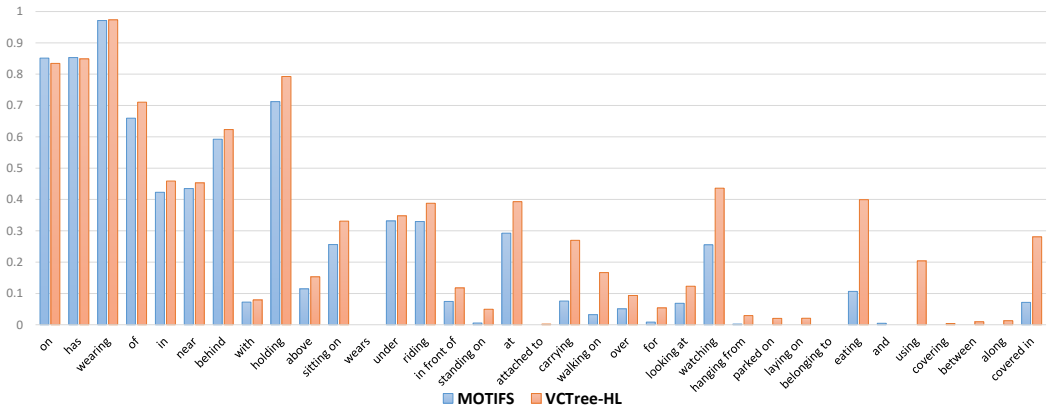


FIGURE 3.7: Recall@100 of MOTIFS [1] and the proposed VCTREE-HL under PredCls for each Top-35 category ranking by frequency.

3.4.6 Comparisons with State-of-the-Arts

3.4.6.1 Comparing Methods

We compared VCTREE with state-of-the-art methods in Table 3.1: (1) **VRD** [186], **FREQ** [1] are methods without using visual contexts. (2) **AssocEmbed** [98] assembles implicit contextual features by stacked hourglass backbone [99]. (3) **IMP** [2], **TFR** [187], **MOTIFS** [1], **Graph-RCNN** [188] are explicit context models with a variety of structures.

3.4.6.2 Quantitative Analysis

From Table 3.1, compared with the previous state-of-the-art MOTIFS [1], the proposed VCTREE has the best performances. Interestingly, Overlap tree and Multi-Branch tree are better than other non-tree context models. We also report detailed

results of the proposed **Mean Recall (mR@K)** in Table 3.2. The proposed VCTree-HL shows best performance among all the ablative structures. Note that MOTIFS [1] has lower mR@100 than FREQ [1] baseline in SGCls and PredCls, which means that MOTIFS is even worse at predicting infrequent predicate categories. However, its mR@20 and mR@50 are higher than FREQ in SGCls and PredCls, which indicates that MOTIFS better separates the foreground relationships from the background ones than FREQ.

To better visualize the improvement of the proposed VCTree-HL on infrequent predicate categories, we rank all the predicate categories by frequency, and show the PredCls Recall@100 of MOTIFS [1] and VCTree-HL for each top-35 category independently in Figure 3.7. We can observe significant improvements on those less frequent but more semantically meaningful predicates.

3.4.6.3 Qualitative Analysis

To better understand what context is learned by VCTree, we visualized a statistics of left-/right-branch nodes for nodes classified as "street" in Figure 3.6. From the left pie, the hierarchical relations, we can see the node categories are long-tailed, *i.e.*, top-10 categories cover the 73% of the instances; while the right pie, the parallel relations, are more uniformly distributed. This demonstrates that VCTree captures the two types of context successfully. Qualitative examples of VCTrees and their generated scene graph can be viewed in Figure 3.8. The common errors are generally synonymous labels, *e.g.*, "jeans" vs. "pants", "man" vs. "person", and over-interpretation, *e.g.*, the "tail" of bottom left "dog" is considered as "leg", as it appears at the place where "leg" should be.

We further investigated more misclassified results of the proposed VCTree-HL. The corresponding tree structures and the generated scene graphs are reported in Figure 3.9. We observed 3 types of interesting misclassifications: 1) In the image (a) of Figure 3.9, the proposed VCTree-HL predicts more appropriate predicates "in front of" and "behind" than original "near". 2) In the image (b) and (d), the ground truth "man in snow" and "window near building" are improper, while our method shows more appropriate predicates. 3) In the image (c) and (d), the objects isolated from the Scene Graph (only considering R@20 predicates) are easier to be misclassified.

VQA2.0 test-dev				
Model	Yes/No	Number	Other	All
Teney [103]	81.82	44.21	56.05	65.32
MUTAN [190]	82.88	44.54	56.50	66.01
MLB [191]	83.58	44.92	56.34	66.27
DA-NTN [14]	**84.29**	47.14	57.92	67.56
Count [183]	83.14	**51.62**	58.97	68.09
Chain	82.74	47.31	58.93	67.42
Graph	83.53	47.09	58.6	67.56
VCTree-HL	84.28	47.78	**59.11**	**68.19**

Table 3.4: Single-model accuracies (%) on VQA2.0 test-dev, where MUTAN and MLB are re-implemented versions from [14].

VQA2.0 test-standard				
Model	Yes/No	Number	Other	All
Teney [103]	82.20	43.90	56.26	65.67
MUTAN [190]	83.06	44.28	56.91	66.38
MLB [191]	83.96	44.77	56.52	66.62
DA-NTN [14]	**84.60**	47.13	58.20	67.94
Count [183]	83.56	**51.39**	59.11	68.41
Chain	83.06	47.38	58.95	67.68
Graph	84.03	47.08	58.82	68.0
VCTree-HL	84.55	47.36	**59.34**	**68.49**

Table 3.5: Single-model accuracies (%) on VQA2.0 test-standard, where MUTAN and MLB are re-implemented versions from [14].

3.5 Experiments on Visual Q&A

3.5.1 Datasets

We evaluated the proposed VQA model on VQA2.0 [107]. Compared with VQA1.0 [101], VQA2.0 has more question-image pairs for training (443,757) and validation (214,354), and all the question-answer pairs are balanced by making sure the same question can have different answers. In VQA2.0, the ground-truth accuracy of a candidate answer is considered as the average of $\min(\frac{\#\text{Humans votes}}{3}, 1)$ over all 10 select 9 sets. Question-answer pairs are organized in three answer types: *i.e.* "Yes/No", "Number", "Other". There are also 65 question types determined by prefixed words, which we used to generate question-guided gates. We also tested our models on a balanced subset of validation set, called Balanced Pairs [103], which requires the

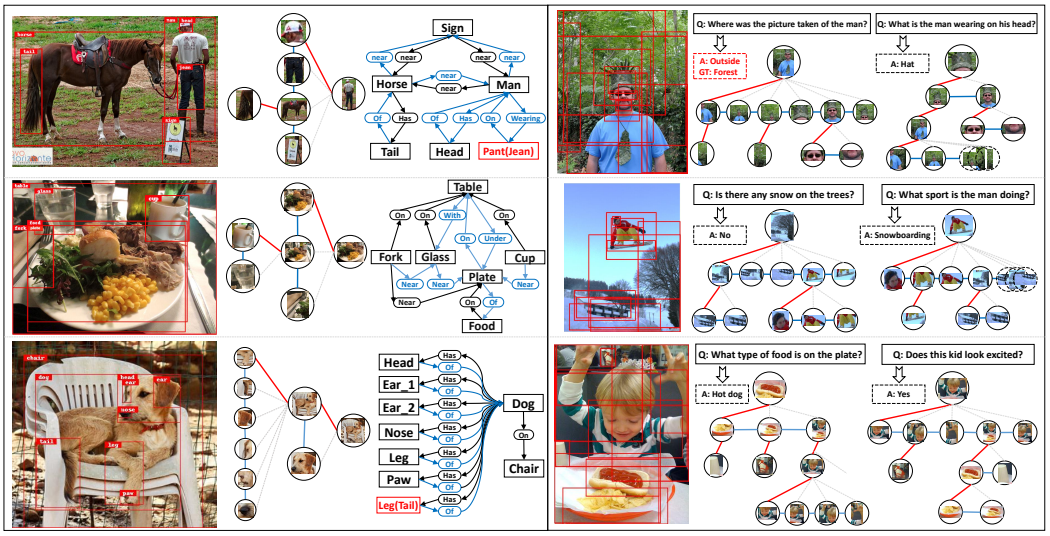


FIGURE 3.8: **Left:** the learned tree structure and generated scene graphs in VG. Black color indicates correctly detected objects or predicates; red indicates the misclassified ones; blue indicates correct predictions that not labeled as ground-truth. **Right:** interpretable and dynamic trees subject to different questions in VQA2.0.

same question on different images with two different yet perfect (with 1.0 ground-truth score) answers. Since Balanced Pairs strictly removes question-related bias, it reflects the ability of a context model to distinguish subtle differences between images.

3.5.2 Implementation Details

We employed a simple text preprocessing that is commonly used by other methods for questions and answers, which changes all characters into lower-case and removes special characters. Questions were encoded into a vocabulary of the size 13,758 without trimming. Answers used a 3,000 vocabulary selected by frequency. For fair comparison, we used the same bottom-up feature [102] as previous methods [14, 102, 103, 183], which contains 10 to 100 object proposals per image extracted by Faster-RCNN [27]. We used the same Faster-RCNN detector to pretrain the $f(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Since candidate answers were represented by probabilities rather than one-hot vectors in VQA2.0, we allowed the cross-entropy loss calculating soft categories, *i.e.*, probabilities of ground-truth candidate answers. We used Adam optimizer with learning rate $lr = 0.0015$ and batch size $b = 256$, lr decayed at ratio of 0.5 every 20 epochs.

3.5.3 Ablation Studies

In addition to the 5 structure construction policies introduced in Section 3.4.5, we also implemented a fully-connected graph structure using the message passing mechanism [2]. From Table 3.3, the proposed VCTree-HL outperforms all the context models on three answer types.

We further evaluated the above context models on VQA2.0 balanced pair subset [103]: the last column of Table 3.3, and found that the absolute gains between VCTree-HL and other structures are even larger than those on the original validation set. Meanwhile, as reported in [103], different architectures or hyper-parameters in non-contextual VQA model normally gain less improvements on the balanced pair subset than overall validation set. Thus, it suggests that VCTree indeed use better context structures to alleviate the question-answer bias in VQA.

3.5.4 Comparisons with State-of-the-Arts

3.5.4.1 Comparing Methods

Table 3.4 & 3.5 reports the single-model performances of various state-of-the-art methods [14, 103, 183, 190, 191] on both test-dev and test-standard sets. For fair comparison, the reported methods are all using the same Faster-RCNN features [102] as ours.

3.5.4.2 Quantitative Analysis

The proposed VCTree-HL shows the best overall performance in both test-dev and test-standard. Note that though Count [183] has close overall performance to our VCTree, it mainly improves the "Number" task by the elaborately designed model, while the proposed VCTree is a more general solution.

3.5.4.3 Qualitative Analysis

We visualized several examples of VCTree-HL on the validation set. They illustrate that the proposed VCTree is able to learn dynamic structures with interpretability, *e.g.*, in Figure 3.8, given the right middle image with the question "Is there any snow on the trees?", the generated VCTree locates the "tree" then searching for the "snow", while with question "What sport is the man doing?", the "man" appears to be the root.

More constructed VCTrees for VQA2.0 are visualized in Figure 3.10. The dynamic tree structures are subject to different questions, which allow the objects in an image to incorporate the different contextual cues according to each question. The proposed VCTree also helps us understand how the model predicts the answer of the question given the image, *e.g.*, in image (a) of Figure 3.10, given the question "does this dog have a collar?", we find that our model first focuses on the collar-like object rather than the dog; in image (b) of Figure 3.10, given the question "what sport is being played?", we find that our model focuses on the sportsman rather than playground to answer this question.

3.6 Conclusions

In this chapter, we proposed a dynamic tree structure called VCTree, which significantly increase the architectural robustness in capturing the task-specific visual contexts. It can be used to support two high-level vision tasks: SGG and VQA. By exploiting VCTree, we observed consistent performance gains in SGG on Visual Genome and in VQA on VQA2.0, compared to models with or without visual contexts, especially in those fair metrics, *e.g.*, the proposed mean Recall@K in SGG and the balanced pair accuracy in VQA. Besides, to justify that VCTree indeed increases the architectural robustness to learn non-trivial contexts, we conducted additional experiments against the category bias in SGG and the question-answer bias in VQA, respectively.

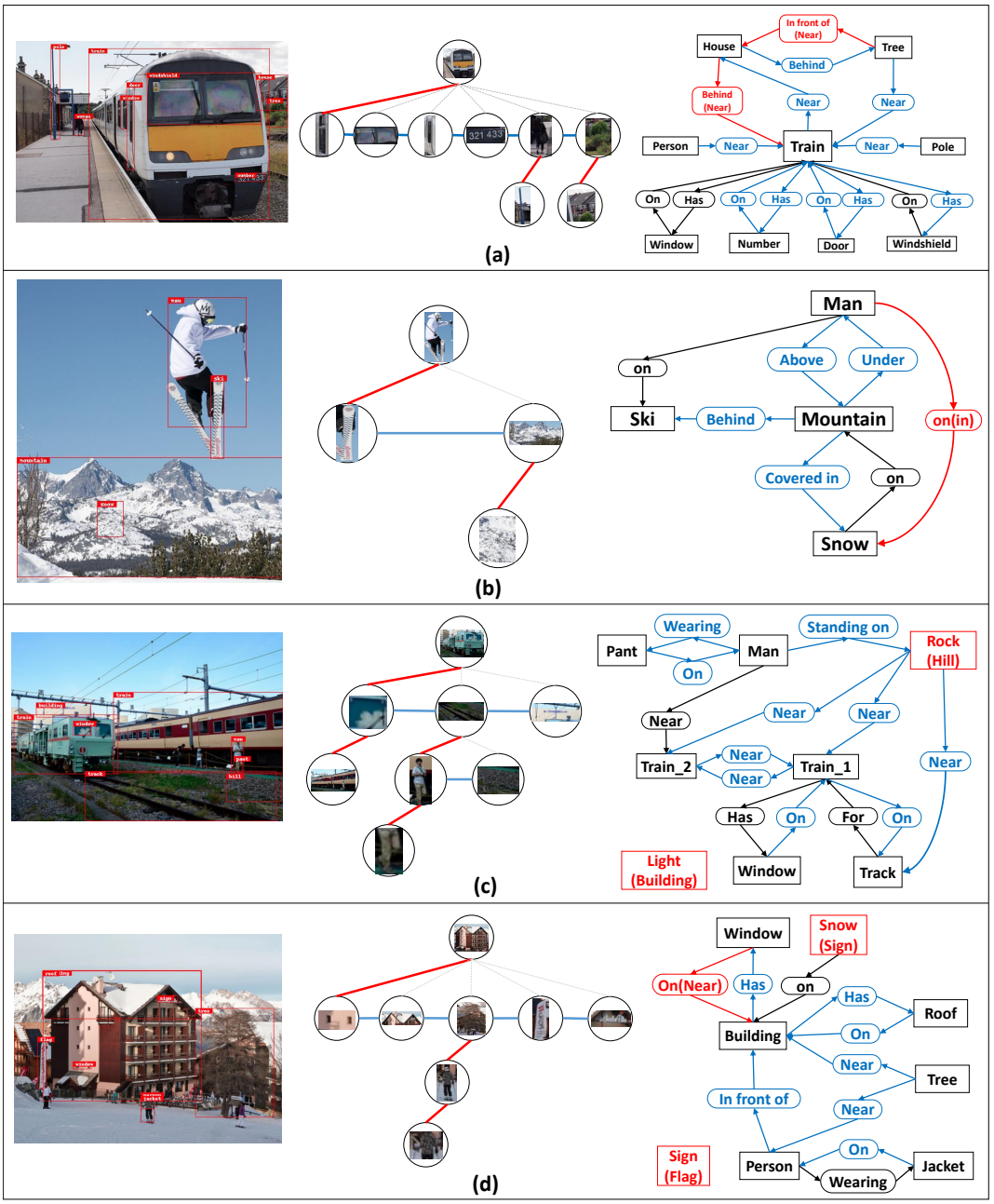


FIGURE 3.9: The learned tree structures and generated scene graphs in VG. We selectively report the predicates from R@20 and all the ground-truth predicates. Black color indicates correctly detected objects or predicates; red indicates the misclassified ones; blue indicates correct predictions that not labeled as ground-truth.

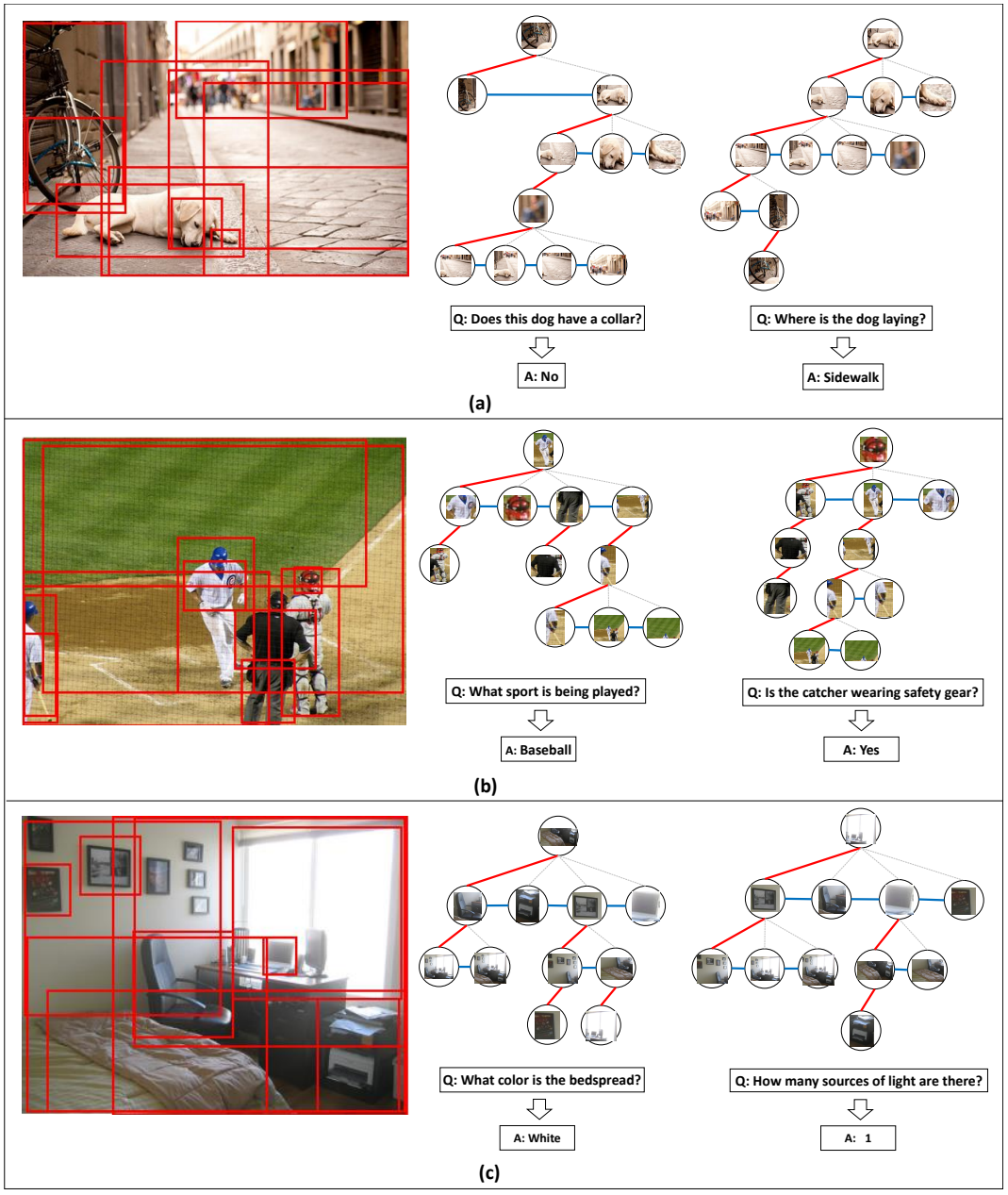


FIGURE 3.10: The dynamic and interpretable tree structures that subject to different questions, which allow the objects in an image incorporate different contextual cues according to each question.

Chapter 4

Total Direct Effect for Unbiased Scene Graph Generation[1]

4.1 Introduction

Scene graph generation (SGG) [2] — a visual detection task of objects and their relationships in an image — seems to have never fulfilled its promise: a comprehensive visual scene representation that supports *graph reasoning* for high-level tasks such as visual captioning [89, 192] and VQA [193, 194]. Once equipped with SGG, these high-level tasks have to abandon the ambiguous visual relationships — yet on which are our core efforts made [6, 8, 15], then pretend that there is a graph — nothing but a sparse object layout with binary links, and finally shroud it into graph neural networks [195] for merely more contextual object representations [89, 91, 193]. Although this is partly due to the research gap in graph reasoning [88, 196, 197], the crux lies in the *biased* relationship prediction as we mentioned in the previous chapter.

Figure 4.1 visualizes an example of SGG results from a state-of-the-art model [6]. We can see a frustrating scene: among almost perfectly detected objects, most of their visual relationships are trivial and less informative. For example in Figure 4.1(c), except the trivial 2D spatial layouts, we know little about the image

[1]The work in this chapter has been published in the paper : Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, Hanwang Zhang. "Unbiased Scene Graph Generation from Biased Training." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (**CVPR, Oral**). Seattle, United States. 2020.

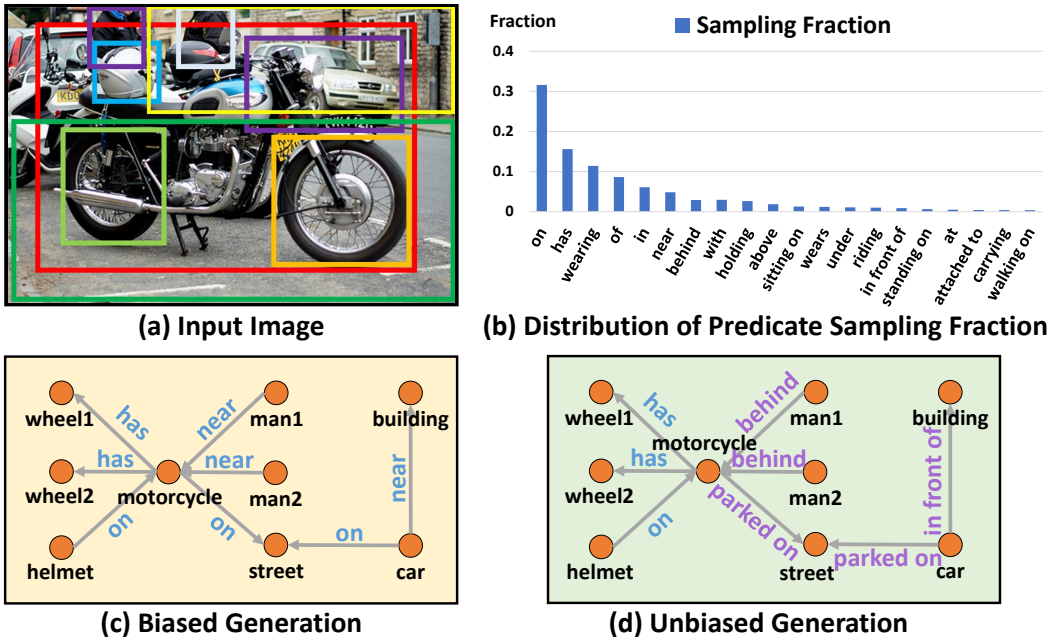


FIGURE 4.1: An example of scene graph generation (SGG). (a) An input image with bounding boxes. (b) The distribution of sample fraction for the most frequent 20 predicates in Visual Genome [5]. (c) An example of SGG results from re-implemented MOTIFS [6]. (d) An example of SGG results by the proposed unbiased prediction from the same model.

from `near`, `on`, and `has`. Such heavily biased generation comes from the *biased training data*, more specifically, as shown in Figure 4.1(b), the highly-skewed long-tailed relationship annotations. For example, if a model is trained for predicting `on` 1,000 times more than `standing on`, then, during test, the former is more likely to prevail over the latter, which is unfortunately inevitable in large-scale dataset for the cost of balancing the dataset increasing exponentially with the size of vocabulary. Therefore, to perform a sensible graph reasoning, we need to distinguish more fine-grained relationships from the ostensibly probable but trivial ones, such as replacing `near` with `behind/in front of`, and `on` with `parking on/driving on` in Figure 4.1(d).

However, we should not blame the biased training because both our visual world *per se* and the way we describe it are biased: there are indeed more `person carry bag` than `dog carry bag` (*i.e.*, the long-tail theory); it is easier for us to label `person beside table` rather than `eating on` (*i.e.*, bounded rationality [7]); and we prefer to say `person on bike` rather than `person ride on bike` (*i.e.*, language or reporting bias [147]). In fact, most of the biased annotations can help the model learn good contextual prior [6, 186] to filter out the unnecessary search

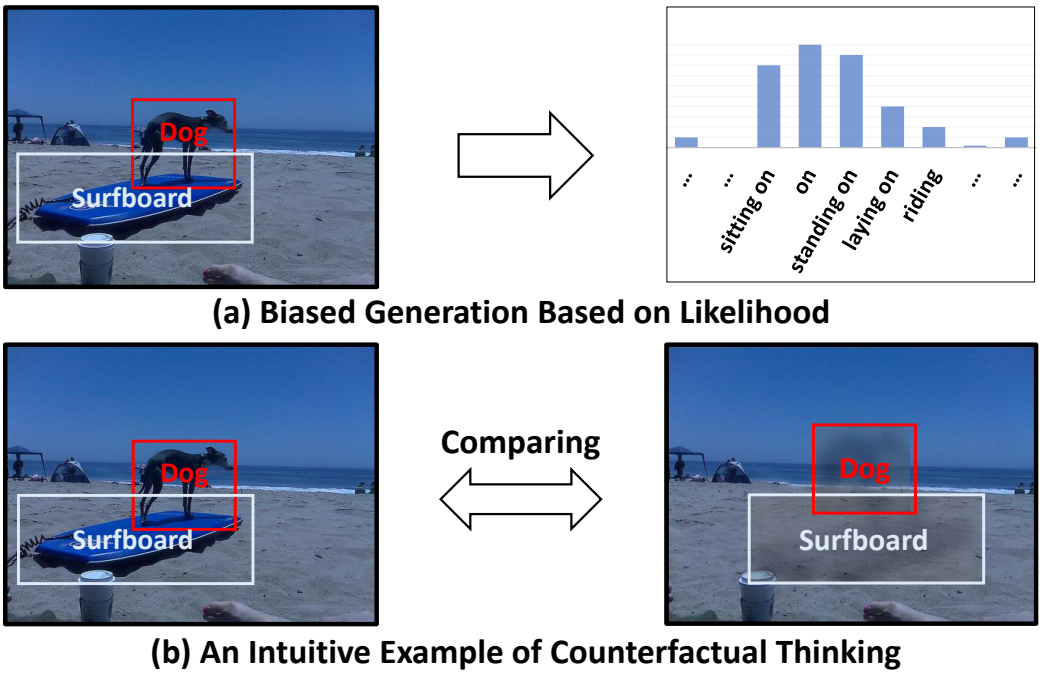


FIGURE 4.2: (a) The biased generation that directly predicts labels from likelihood. (b) An intuitive example of the proposed total direct effect, which calculates the difference between the real scene and the counterfactual one. Note that the "wipe-out" is only for the illustrative purpose but not considered as visual processing.

candidates such as `apple park on table` and `apple wear hat`. A promising but embarrassing finding [6] is that: by only using the statistical prior of detected object class in the Visual Genome benchmark [5], we can already achieved 30.1% on Recall@100 for Scene Graph Detection — rendering all the much more complex SGG models almost useless — that is only 1.1-1.5% lower than the state-of-the-art [8, 198, 199]. Not surprisingly, as we will show in Section 4.5, conventional debiasing methods who do not respect the "good bias" during training, *e.g.*, re-sampling [142] and re-weighting [144], fail to generalize to unseen relationships, *i.e.*, zero-shot SGG [186].

For both machines and humans, decision making is a collaboration of *content* (endogenous reasons) and *context* (exogenous reasons) [200]. Take SGG as an example, in most SGG models [6, 198, 199], the content is the visual features of the subject and object, and the context is the visual features of the subject-object union regions and the pairwise object classes. We humans — born and raised in the biased nature — are ambidextrous in embracing the good while avoiding the bad context,

and making unbiased decisions together with the content. The underlying mechanism is *causality-based*: the decision is made by pursuing the main causal effect caused by the content but not the side-effect by context. However, on the other hand, machines are usually *likelihood-based*: the prediction is analogous to look-up the content and its context in a huge likelihood table, interpolated by population training. We believe that the key is to teach machines how to distinguish between the "main effect" and "side-effect".

In this chapter, we propose to empower machines the ability of *counterfactual causality* [42] to pursue the "main effect" in unbiased prediction:

If I ***had not*** *seen the content, would I still make the same prediction?*

The counterfactual lies between the fact that "I see" and the imagination "I had not", and the comparison between the factual and counterfactual will naturally *remove* the effect from the context bias, because the context is the only thing unchanged between the two alternatives. Thanks to the causal graph, it can effectively disentangle the contributions from different modalities, *e.g.*, visual feature vs. language embedding vs. statistical prior, extracting and eliminating biased components of predictions by the counterfactual inference stage.

To better illustrate the profound yet subtle difference between likelihood and counterfactual causality, we present a `dog standing on surfboard` example in Figure 4.2(a). Due to the biased training, the model will eventually predict the `on`, as the modality of language embedding and statistical prior taking the shortcuts of memorizing data, making the visual feature of standing less significant in the final prediction. Note that even though the rest choices are not all exactly correct, thanks to the bias, they still help to filter out a large amount of unreasonable ones. To take a closer look at what relationship it is in the context bias, we are essentially comparing the original scene with a counterfactual scene (Figure 4.2(b)): only the visual features of the `dog` and `surfboard` are wiped out, while keeping the rest — the scene and the object classes — untouched, as if the visual features had ever existed. By doing this, we can focus on the main visual effects of the relationship without losing the context.

We propose a novel unbiased SGG method based on the Total Direct Effect (TDE) analysis framework in causal inference [201–203]. Figure 4.3(a) shows the underlying causal graphs [42, 149] of the two alternate scenes: factual and counterfactual.

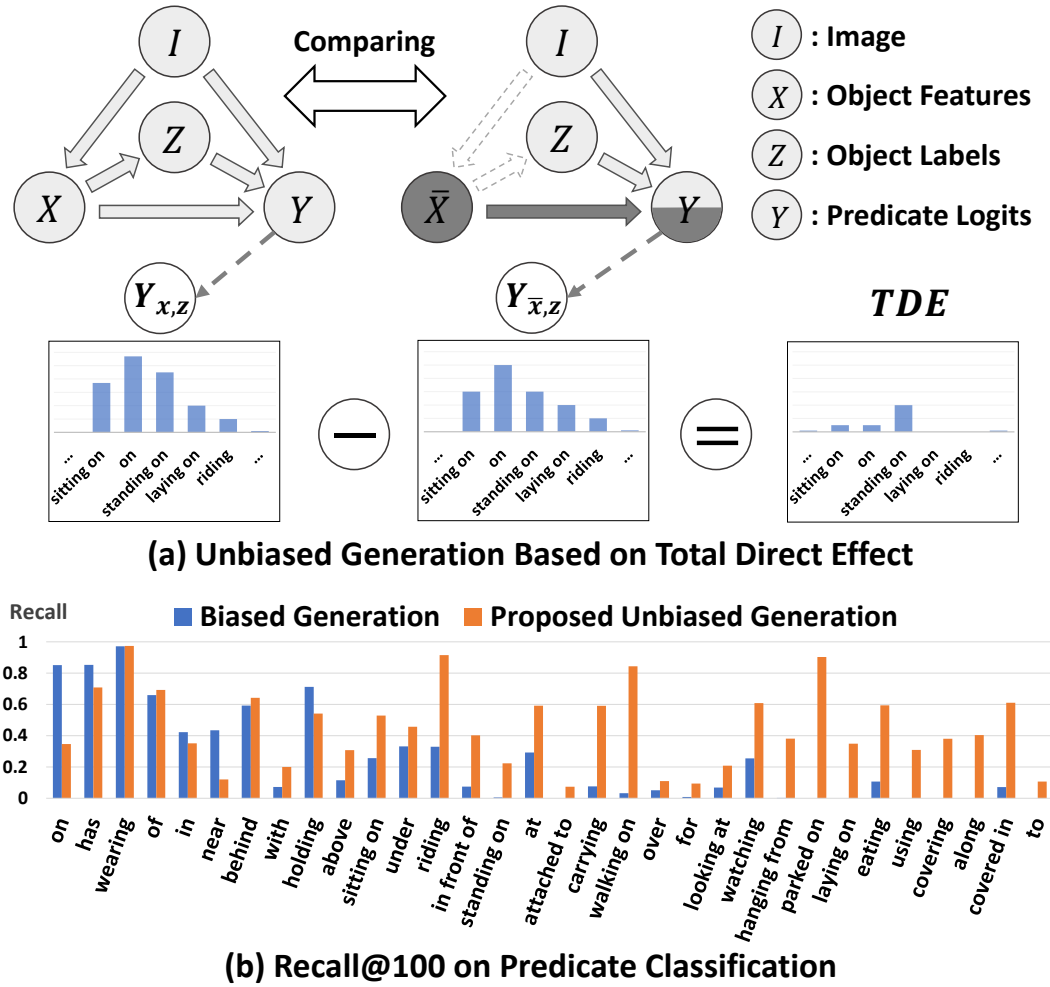


FIGURE 4.3: (a) The example of total direct effect calculation and corresponding operations on the causal graph, where $\bar{X}$ represents wiped-out X. (b) Recall@100 of Predicate Classification for selected predicates ranking by sampling fraction. The biased generation refers to re-implemented MOTIFS [6] and the proposed unbiased generation is the result from the same model using TDE.

Although a formal introduction of them is given in Section 4.2-4.3, now you can simply understand the nodes as data features and the directed links as (parametric) data flows. For example, $X \rightarrow Y$, $Z \rightarrow Y$, and $I \rightarrow Y$ indicate that the relationship Y is a combined effect caused by *content*: the pair of object visual features X, *context*: their object classes Z, and *scene*: the image I; the faded links denote that the wiped-out $\bar{X}$ is no longer caused by I or affects Z. These graphs offer an algorithmic formulation to calculate TDE, which exactly realizes the counterfactual thinking in Figure 4.2. As shown in Figure 4.3(b), the proposed TDE significantly improves most of the predicates, and impressively, the distribution of the improved performances is no longer long-tailed, indicating the fact that our improvement is indeed from the proposed method, but NOT from the better

exploitation of the context bias. A closer analysis in Figure 4.9 further shows that the worse predictions like `on` — though very few — are due to turning to more fine-grained results such as `stand on` and `park on`. We highlight that TDE is a model-agnostic prediction strategy and thus applicable for a variety of models and fusion tricks [6, 8, 9].

Last but not least, we propose a new standard of SGG diagnosis toolkit[2] (cf. Section 4.5.2) for more comprehensive SGG evaluations. Besides traditional evaluation tasks, it consists of the bias-sensitive metric: mean Recall [8, 15] and a new Sentence-to-Graph Retrieval for a more comprehensive graph-level metric. By using this toolkit on SGG benchmark Visual Genome [5] and several prevailing baselines, we verify the severe bias in existing models and demonstrate the effectiveness of the proposed unbiased prediction over other debiasing strategies.

4.2 Biased Training Models in Causal Graph

As illustrated in Figure 4.4, we summarize the multimodal SGG framework in the form of *Causal Graph* (*a.k.a.*, structural causal model) [42, 54, 149]. As we mentioned in Chapter 2, it is a directed acyclic graph $\mathcal{G} = \{\mathcal{N}, \mathcal{E}\}$, indicating how a set of variables $\mathcal{N}$ interact with each other through the causal links $\mathcal{E}$. It provides a sketch of the causal relations behind the data and how variables obtain their values, *e.g.*, $(I, X, Z) \rightarrow Y$. Before we conduct counterfactual analysis that deliberately manipulates the values of nodes and prunes the causal graph, we first revisit the conventional biased SGG model training in the graphical view.

The causal graph in Figure 4.4(b) is applicable to a variety of SGG models, since it is highly general, imposing no constraints on the detailed implementations. We case-study three representative model formulations: the classic VTransE [9], the state-of-the-art MOTIFS [6] and VCTree [8], using the language of nodes and links.

Node *I* (**Input Image&Backbone**). A Faster R-CNN [27] is pre-trained and frozen in this node to detect the objects, It outputs a set of bounding boxes $B = \{b_i | i = 1...n\}$ and the feature map $\mathcal{M}$ from image I.

[2]Codes are publicly available on https://github.com/KaihuaTang/Scene-Graph-Benchmark.pytorch

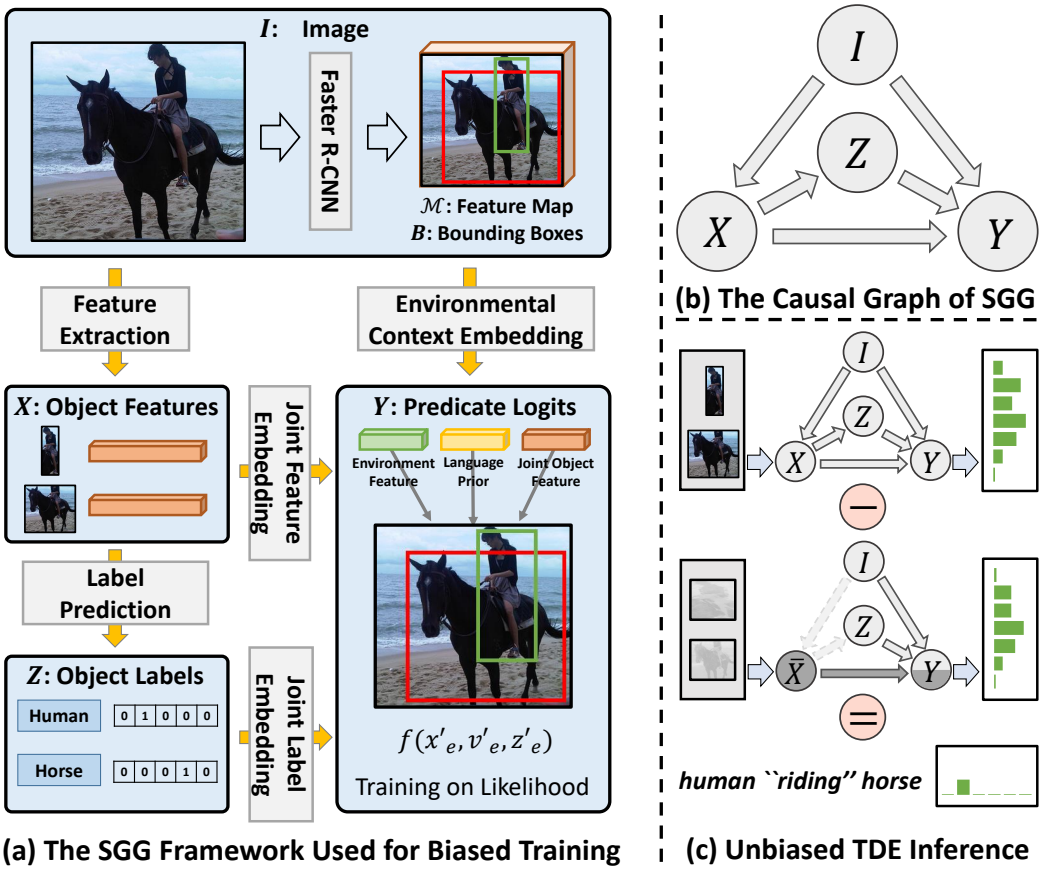


FIGURE 4.4: (a) The framework used in our biased training. (b) The causal graph of the SGG framework. (c) An illustration of the proposed TDE inference.

Link $I \rightarrow X$ (Object Feature Extractor). It firstly extracts RoIAlign features [29] $R = \{r_i\}$ and tentative object labels $L = \{l_i\}$ by the object classifier on Faster R-CNN. Then, like MOTIFS [6] or VCTree [8], we can use the following module to encode visual contexts for each object:

$$Input: \{(r_i, b_i, l_i)\} \Longrightarrow Output: \{x_i\}, \tag{4.1}$$

where MOTIFS implements it as bidirectional LSTMs (Bi-LSTMs) and VCTree [8] adopts bidirectional TreeLSTMs (Bi-TreeLSTMs) [3], early works like VTransE [9] simply use fully connected layers.

Node X (Object Feature). The pairwise object feature X takes value from $\{(x_i, x_j)|i \neq j; i, j = 1...n\}$. We slightly abuse the notation hereinafter, denoting the combination of representations from i and j as subscript e: $x_e = (x_i, x_j)$.

Link $X \rightarrow Z$ (Object Classification). The fine-tuned label of each object is decoded from the corresponding feature vector x_i by:

$$Input : \{x_i\} \Longrightarrow Output : \{z_i\}, \tag{4.2}$$

where MOTIFS [6] and VCTree [8] utilizes LSTM and TreeLSTM as decoders to capture the co-occurrence among object labels, respectively. The input of each LSTM/ TreeLSTM cell is the concatenation of feature and the previous label $[x_i; z_{i-1}]$. VTransE [9] uses the conventional fully connected layer as the classifier.

Node Z (Object Class). It contains a pair of one-hot vectors for object labels $z_e = (z_i, z_j)$.

Link $X \rightarrow Y$ (Object Feature Input for SGG). For relationship classification, pairwise feature X are merged into a joint representation by the module:

$$Input : \{x_e\} \Longrightarrow Output : \{x'_e\}, \tag{4.3}$$

where another Bi-LSTMs and Bi-TreeLSTMs layers are applied in MOTIFS [6] and VCTree [8], respectively, before concatenating the pair of object features. VTransE [9] uses fully connected layers and element-wise subtraction for feature merging.

Link $Z \rightarrow Y$ (Object Class Input for SGG). The language prior is calculated in this link through a joint embedding layer $z'_e = W_z[z_i \otimes z_j]$, where $\otimes$ generates the one-hot unique vector $\mathbb{R}^{N \times N}$ for the pair of N-way object labels.

Link $I \rightarrow Y$ (Visual Context Input for SGG). This link extracts the contextual union region features $v'_e = \text{Convs}(\text{RoIAlign}(\mathcal{M}, b_i \cup b_j))$ where $b_i \cup b_j$ indicates the union box of two RoIs.

Node Y (Predicate Classification). The final predicate logits Y that takes inputs from the three branches is then generated by using a fusion function. In Section 4.5, we test two general fusion functions: 1) SUM: $y_e = W_x x'_e + W_v v'_e + z'_e$, 2) GATE: $y_e = W_r x'_e \cdot \sigma(W_x x'_e + W_v v'_e + z'_e)$, where $\cdot$ is element-wise product, $\sigma(\cdot)$ is a sigmoid function.

Training Loss. All models are trained by using the conventional cross-entropy losses of object labels and predicate labels. To avoid any single link spontaneously dominating the generation of logits y_e, especially $Z \rightarrow Y$, we further add auxiliary cross-entropy losses that individually predict y_e from each branch, which also help to disentangle the effects from different modalities for further counterfactual inference.

4.3 Unbiased Prediction by Causal Effects

Once the above training has been done, the causal dependencies among the variables are learned, in terms of the model parameters. The conventional biased prediction can only see the output of the entire graph given an image $I = u$ without any idea about how a specific pair of objects affect their predicate. However, causal inference [42] encourages us to think out of the black box. From the graphical point of view, we are no longer required to run the entire graph as a whole. We can directly manipulate the values of several nodes and see what would be going on. For example, we can cut off the link $I \rightarrow X$ and assign a dummy value to X, then investigate what the predicate would be. The above operation is termed *intervention* in causal inference [149]. Next, we will make unbiased predictions by intervention and its induced counterfactuals.

4.3.1 Notations

Intervention. It can be denoted as $\boldsymbol{do(\cdot)}$. It wipes out all the in-coming links of a variable, demanding the variable to take a certain value, *e.g.*$do(X = \bar{x})$ in Figure 4.5(b), which means X is no longer affected by its causal parents.

Counterfactual. It means *"counter to the facts"* [204], and takes one step further that assigns the "clash of worlds" combination of values to variables. Take Figure 4.5(c) as an example, if the intervention $do(X = \bar{x})$ is conducted on X, the variable Z still takes the original z as if x had existed.

Causal Effect. Throughout this section, we will use the pairwise object feature X as our control variable where the intervention is conducted, aiming to assess its effects, due to the fact that there wouldn't be any valid relationship if the pair

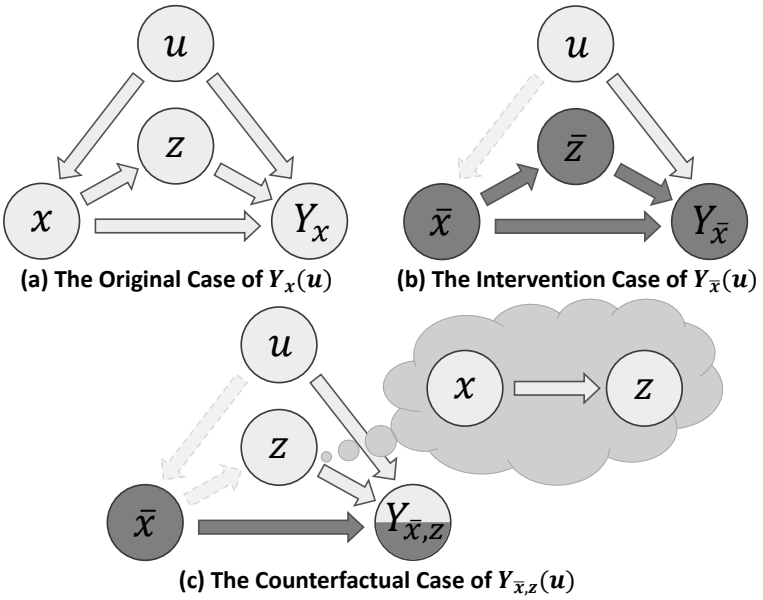


FIGURE 4.5: The original causal graph of SGG together with two interventional and counterfactual alternates.

of objects do not exist. The observed X is denoted as x while the intervened unseen value is $\bar{x}$, which is set to either the mean feature of the training set or zero vector. The object label z on Figure 4.5(c) is calculated from Eq. (4.2), taking x as input. We denote the output logits Y after the intervention $X = \bar{x}$ as follows (Figure 4.5(b)):

$$Y_{\bar{x}}(u) = Y(do(X = \bar{x})|u), \tag{4.4}$$

where u is the input image in SGG. Following the above notation, the original and counterfactual Y, *i.e.*, Figure 4.5(a,c), can be re-written as $Y_x(u)$ and $Y_{\bar{x},z}(u)$, respectively.

4.3.2 Total Direct Effect

As we discussed in Section 4.1, instead of the static likelihood that tends to be biased, the unbiased prediction lies in the difference between the observed outcome $Y_x(u)$ and its counterfactual alternate $Y_{\bar{x},z}(u)$. The later one is a context-specific bias that we want to remove from prediction. Intuitively, the unbiased prediction that we seek is the visual stimuli from blank to the observed real objects with specific attributes, states, and behaviors, but not merely from the surroundings and language priors. Those specific visual cues of objects are the key to the more

fine-grained and informative unbiased predictions, because even if the overall prediction is biased towards the relationship like `dog on surfboard`, the "straight legs" would cause more effect on `standing on` rather than `sitting on`. In causal inference [201, 203], the above prediction process can be calculated as Total Direct Effect (TDE):

$$TDE = Y_x(u) - Y_{\bar{x},z}(u), \tag{4.5}$$

where the first term is from the original graph and the second one is from the counterfactual, as illustrated in Figure 4.5.

Note that there is another type of effect [201], Total Effect (TE), which is easy to be mixed up with TDE. Instead of deriving counterfactual bias $Y_{\bar{x},z}(u)$, TE lets all the descendant nodes of X change with intervention $do(X = \bar{x})$ as shown in Figure 4.5(b). TE is therefore formulated as:

$$TE = Y_x(u) - Y_{\bar{x}}(u). \tag{4.6}$$

The main difference lies in the fact that $Y_{\bar{x}}(u)$ is not conditioned on the original object labels (those caused by x), so TE only removes the general bias in the whole dataset (similar to the b in $y = k \cdot x + b$), rather than the specific bias caused by the mediator we care about. The subtle difference between TE and TDE is further defined as Natural Indirect Effect (NIE) [201] or Pure Indirect Effect (PIE) [203]. More experimental analyses among these three types of effect are given in Section 4.5. More detailed introduction of these different types of effects will be given in the next section. The choosing of these effects depends on specific tasks and which kind of information you desire.

Overall SGG. At last, the proposed unbiased prediction $y_e^\dagger$ is obtained by replacing the conventional one-time prediction with TDE, which essentially "thinks" twice: one for observational $Y_{x_e}(u) = y_e$, the other for imaginary $Y_{\bar{x},z_e}(u) = y_e(\bar{x}, z_e)$. The unbiased logits of Y is therefore defined as follows:

$$y_e^\dagger = y_e - y_e(\bar{x}, z_e). \tag{4.7}$$

It is also worth mentioning that the proposed TDE doesn't introduce any additional parameters and is widely applicable to a variety of models.

4.4 Review of Causal Effect Analysis

In this section, a comprehensive review of causal effect analysis is given in the form of the proposed causal graph. More detailed background knowledge about causal inference can be found in [42, 149] while the extension of effect analysis (a.k.a. mediation analysis) is given in [201–203, 205].

4.4.1 Total, Direct and Indirect Effects

As we discussed in Section 4.3, without further counterfactual intervention on the mediator Z, the overall effect of X towards Y is regarded as the Total Effect (TE) of X on Y, which can be calculated as:

$$TE = Y_x(u) - Y_{\bar{x}}(u). \tag{4.8}$$

As illustrated in Figure 4.6, other than the path $I \rightarrow X$ that is cut off by the intervention $X = \bar{x}$, all the other variables will take their values through the links of causal graph. Especially, the mediator Z will get value $\bar{z}$, which is calculated from Eq. (2) given $\bar{x}$ as input.

However, by only using the TE, we are still not able to separate the mediator-specific "causal effect" from "side effect", which limits the value of causal effect analysis. Thanks to the development of causal inference, here comes the decomposition of TE [202, 203]. Generally, the TE of X is composed of the Direct Effect (DE) caused by the causal path $X \rightarrow Y$ and Indirect Effect (IE) caused by the side-effect path $X \rightarrow Z \rightarrow Y$. Depending on whose effect we want to obtain, two kinds of decomposition can be applied.

Decomposition 1: The first kind of decomposition is what we used in the Section 4.3, which separates the TE into the Total Direct Effect (TDE) and the Natural/Pure Indirect Effect (NIE/PIE). The former one has already been defined in previous section as:

$$TDE = Y_x(u) - Y_{\bar{x},z}(u), \tag{4.9}$$

which can be regarded as the effect of X in the real situation, *i.e.*, Z always takes the value z as if it had seen the real x. Meanwhile, the NIE or PIE is the effect caused by the mediator Z under a pure/natural situation, *i.e.*, X will not take the

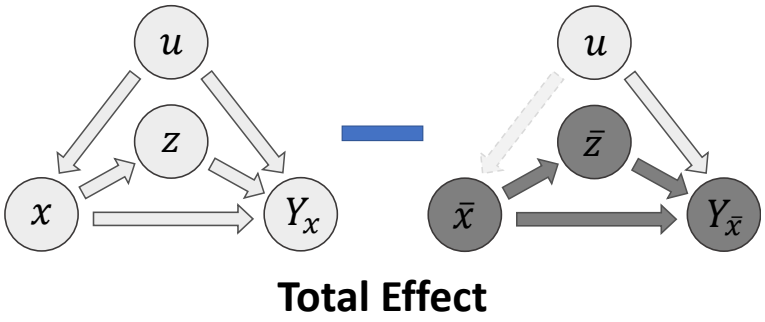


FIGURE 4.6: The illustration of Total Effect on causal graph.

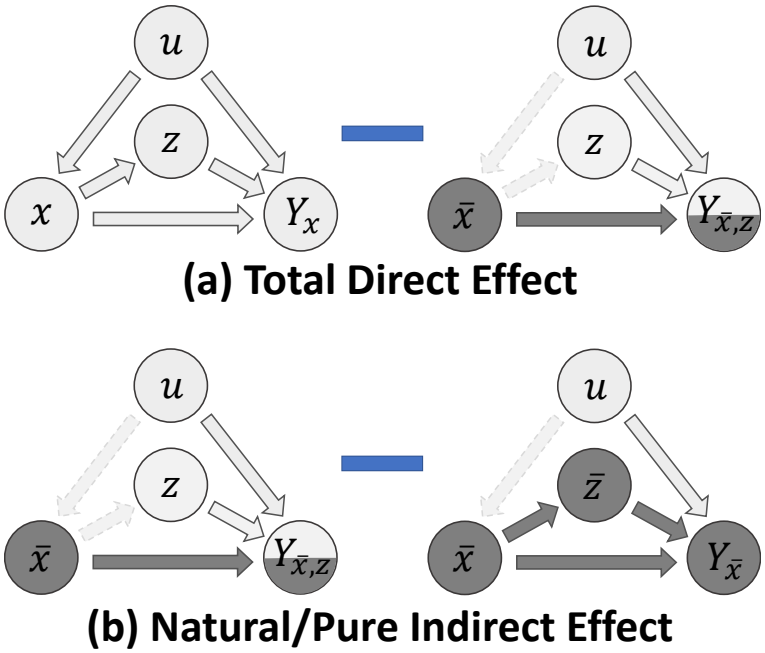


FIGURE 4.7: The illustration of Total Direct Effect and Pure/Natural Indirect Effect on causal graph.

value x under the specific case and it's only assigned to the general unactivated value $\bar{x}$. Therefore, the NIE of Z is denoted as:

$$NIE = Y_{\bar{x},z}(u) - Y_{\bar{x}}(u) \tag{4.10}$$

$$= TE - TDE, \tag{4.11}$$

where we can easily identify that NIE is the effect of Z when it changes from $\bar{z}$ to z in a pure environment, *i.e.*, $X = \bar{x}$. The illustrations of TDE and NIE are given in Figure 4.7.

Decomposition 2: The second type of decomposition is opposite to the first

one. It's mainly adopted when the indirect effect of the mediator is what we are looking for. For example, in the study of carcinogenesis by smoke (Cigarette → Nicotine → Cancer), sometimes the side effect of Nicotine is what researchers really care about. In this case, TE can be decomposed into Total Indirect Effect(TIE) and Natural/Pure Direct Effect (NDE/PDE). The definition of the former one is very similar to the NIE except for the environment being the real case $X = x$, which is therefore formulated as:

$$TIE = Y_x(u) - Y_{x,\bar{z}}(u). \tag{4.12}$$

At the same time, since direct effect is not the target, their pure/natural effect should be removed from the TE. The calculation of NDE/PDE is following:

$$NDE = Y_{x,\bar{z}}(u) - Y_{\bar{x}}(u) \tag{4.13}$$

$$= TE - TIE, \tag{4.14}$$

where NDE is the effect of X changing from $\bar{x}$ to x under the pure environment $Z = \bar{z}$. In general, we should put the effect we care under the real environment, *i.e.*TDE or TIE, so we can get the results specific to each cases.

The above two types of decomposition are both commonly used in medical, political or psychological research [134, 150–152, 206], which depends on which effect we want to obtain, main effect or side effect. Note that, if the system is a pure linear system, both two types of decomposition would be exactly the same. The difference between NXE/PXE and TXE are whether the multimodal context needs to be maintained or not. Those start with T- encode the contexts from multimodal interactions between two branches.

4.5 Experiments

4.5.1 Settings and Models

Dataset. For SGG, we used Visual Genome (VG) [5] dataset to train and evaluate our models, which is composed of 108k images across 75k object categories and 37k predicate categories. However, as 92% of the predicates have no more than

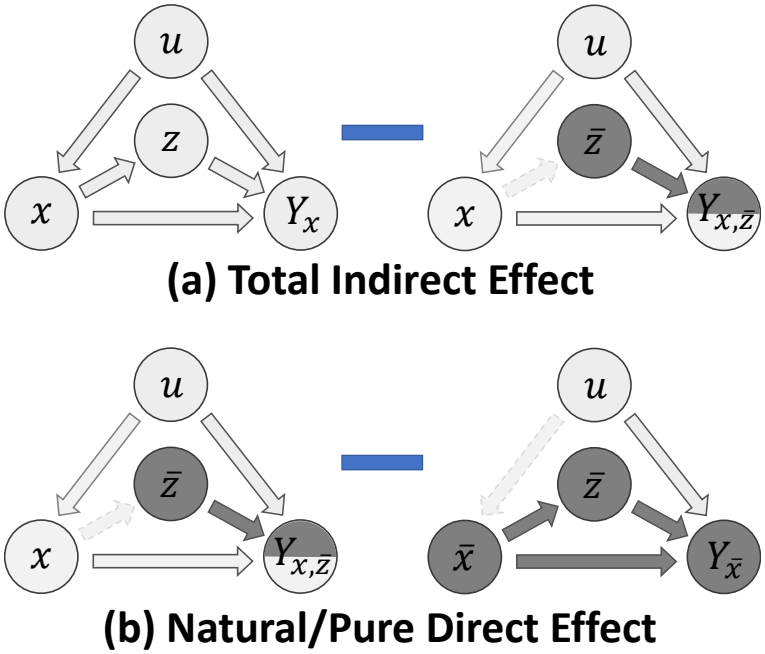


FIGURE 4.8: The illustration of Total Indirect Effect and Pure/Natural Direct Effect on causal graph.

10 instances, we followed the widely adopted VG split [2, 6, 8, 198] containing the most frequent 150 object categories and 50 predicate categories. The original split only has training set (70%) and test set (30%). We followed [6] to sample a 5k validation set from training set for parameter tuning. For Sentence-to-Graph Retrieval (cf. Section 4.5.2), we selected the overlapped 41,859 images between VG and MS-COCO Caption dataset [70] and divided them into train/test-1k/test-5k (35,859/1,000/5,000) sets. The later two only contain images from VG test set in case of exposing to grount-truth SGs. Each image has at least 5 captions serving as human queries, the same as how we use searching engines.

Model Zoo. We evaluated three models: VTransE [9], MOTIFS [6], VTree [8], and two fusion functions: SUM and GATE. They were re-implemented using the same codebase as we proposed. All models shared the same hyper-parameters and the pre-trained detector backbone.

4.5.2 Scene Graph Generation Diagnosis

Our proposed SGG diagnosis has the following three evaluations:

			Predicate Classification			Scene Graph Classification			Scene Graph Detection		
Model	Fusion	Method	mR@20	mR@50	mR100	mR@20	mR50	mR100	mR@20	mR50	mR100
IMP+ [2, 15]	-	-	-	9.8	10.5	-	5.8	6.0	-	3.8	4.8
FREQ [6, 8]	-	-	8.3	13.0	16.0	5.1	7.2	8.5	4.5	6.1	7.1
MOTIFS [6, 8]	-	-	10.8	14.0	15.3	6.3	7.7	8.2	4.2	5.7	6.6
KERN [15]	-	-	-	17.7	19.2	-	9.4	10.0	-	6.4	7.3
VCTree [8]	-	-	14.0	17.9	19.4	8.2	10.1	10.8	5.2	6.9	8.0
MOTIFS†	SUM	Baseline	11.5	14.6	15.8	6.5	8.0	8.5	4.1	5.5	6.8
		Focal	10.9	13.9	15.0	6.3	7.7	8.3	3.9	5.3	6.6
		Reweight	16.0	20.0	21.9	8.4	10.1	10.9	**6.5**	**8.4**	**9.8**
		Resample	14.7	18.5	20.0	9.1	11.0	11.8	5.9	8.2	9.7
		X2Y	13.0	16.4	17.6	6.9	8.6	9.2	5.1	6.9	8.1
		X2Y-Tr	11.6	14.9	16.0	6.5	8.4	9.1	5.0	6.9	8.1
		TE	18.2	25.3	29.0	8.1	12.0	14.0	5.7	8.0	9.6
		NIE	0.6	1.1	1.4	6.1	9.0	10.6	3.8	5.1	6.0
		TDE	**18.5**	**25.5**	**29.1**	**9.8**	**13.1**	**14.9**	5.8	8.2	**9.8**
	GATE	Baseline	12.2	15.5	16.8	7.2	9.0	9.5	5.2	7.2	8.5
		TDE	**18.5**	**24.9**	**28.3**	**11.1**	**13.9**	**15.2**	**6.6**	**8.5**	**9.9**
VTransE†	SUM	Baseline	11.6	14.7	15.8	6.7	8.2	8.7	3.7	5.0	6.0
		TDE	**17.3**	**24.6**	**28.0**	**9.3**	**12.9**	**14.8**	**6.3**	**8.6**	**10.5**
	GATE	Baseline	13.6	17.1	18.6	6.6	8.2	8.7	5.1	6.8	8.0
		TDE	**18.9**	**25.3**	**28.4**	**9.8**	**13.1**	**14.7**	**6.0**	**8.5**	**10.2**
VCTree†	SUM	Baseline	11.7	14.9	16.1	6.2	7.5	7.9	4.2	5.7	6.9
		TDE	**18.4**	**25.4**	**28.7**	**8.9**	**12.2**	**14.0**	**6.9**	**9.3**	**11.1**
	GATE	Baseline	12.4	15.4	16.6	6.3	7.5	8.0	4.9	6.6	7.7
		TDE	**17.2**	**23.3**	**26.6**	**8.9**	**11.8**	**13.4**	**6.3**	**8.6**	**10.3**

TABLE 4.1: The SGG performances of Relationship Retrieval on mean Recall@K [8, 15]. The SGG models re-implemented under our codebase are denoted by the superscript †.

4.5.2.1 Relationship Retrieval (RR)

It can be further divided into three sub-tasks: (1) Predicate Classification **(PredCls)**: taking ground truth bounding boxes and labels as inputs, (2) Scene Graph Classification **(SGCls)**: using ground truth bounding boxes without labels, (3) Scene Graph Detection **(SGDet)**: detecting SGs from scratch. The conventional metric of RR is **Recall@K (R@K)**, which was abandoned in this chapter due to the reporting bias [147]. As illustrated in Figure 4.3(b), previous methods like [6] with good performance on R@K unfairly cater to "head" predicates, *e.g.*, `on`, while neglect the "tail" ones, *e.g.*, predicates like `parked on`, `laying on` have embarrassingly 0.0 Recall@100. To speak for the valuable "tail" rather than the trivial "head", we adopted a recent replacement, **mean Recall@K (mR@K)**, proposed by Tang *et al.* [8] and Chen *et al.* [15]. mR@K retrieves each predicate separately and then averages R@K for all predicates.

			Predicate Classification		Scene Graph Classification		Scene Graph Detection	
Model	Fusion	Method	R@20 / 50 / 100	mR@20 / 50 / 100	R@20 / 50 / 100	mR@20 / 50 / 100	R@20 / 50 / 100	mR@20 / 50 / 100
IMP+ [2, 15]	-	-	52.7 / 59.3 / 61.3	- / 9.8 / 10.5	31.7 / 34.6 / 35.4	- / 5.8 / 6.0	14.6 / 20.7 / 24.5	- / 3.8 / 4.8
FREQ [6, 8]	-	-	53.6 / 60.6 / 62.2	8.3 / 13.0 / 16.0	29.3 / 32.3 / 32.9	5.1 / 7.2 / 8.5	20.1 / 26.2 / 30.1	4.5 / 6.1 / 7.1
MOTIFS [6, 8]	-	-	58.5 / 65.2 / 67.1	10.8 / 14.0 / 15.3	32.9 / 35.8 / 36.5	6.3 / 7.7 / 8.2	21.4 / 27.2 / 30.3	4.2 / 5.7 / 6.6
KERN [15]	-	-	- / 65.8 / 67.6	- / 17.7 / 19.2	- / 36.7 / 37.4	- / 9.4 / 10.0	- / 27.1 / 29.8	- / 6.4 / 7.3
VCTree [8]	-	-	60.1 / 66.4 / 68.1	14.0 / 17.9 / 19.4	35.2 / 38.1 / 38.8	8.2 / 10.1 / 10.8	22.0 / 27.9 / 31.3	5.2 / 6.9 / 8.0
MOTIFS†	SUM	Baseline	59.5 / 66.0 / 67.9	11.5 / 14.6 / 15.8	35.8 / 39.1 / 39.9	6.5 / 8.0 / 8.5	25.1 / 32.1 / 36.9	4.1 / 5.5 / 6.8
		Focal	59.2 / 65.8 / 67.7	10.9 / 13.9 / 15.0	36.0 / 39.3 / 40.1	6.3 / 7.7 / 8.3	24.7 / 31.7 / 36.7	3.9 / 5.3 / 6.6
		Reweight	45.4 / 57.0 / 61.7	16.0 / 20.0 / 21.9	24.2 / 29.5 / 31.5	8.4 / 10.1 / 10.9	18.3 / 24.4 / 29.3	6.5 / 8.4 / 9.8
		Resample	57.6 / 64.6 / 66.7	14.7 / 18.5 / 20.0	34.5 / 37.9 / 38.8	9.1 / 11.0 / 11.8	23.2 / 30.5 / 35.4	5.9 / 8.2 / 9.7
		X2Y	58.3 / 65.0 / 66.9	13.0 / 16.4 / 17.6	35.2 / 38.6 / 39.5	6.9 / 8.6 / 9.2	24.8 / 32.1 / 36.7	5.1 / 6.9 / 8.1
		X2Y-Tr	59.0 / 65.3 / 66.9	11.6 / 14.9 / 16.0	35.5 / 38.9 / 39.7	6.5 / 8.4 / 9.1	25.5 / 32.8 / 37.2	5.0 / 6.9 / 8.1
		TE	34.3 / 46.7 / 51.7	18.2 / 25.3 / 29.0	25.5 / 32.5 / 35.4	8.1 / 12.0 / 14.0	14.8 / 20.1 / 23.9	5.7 / 8.0 / 9.6
		NIE	0.6 / 1.0 / 1.3	0.6 / 1.1 / 1.4	28.6 / 35.0 / 37.4	6.1 / 9.0 / 10.6	17.3 / 22.7 / 26.8	3.8 / 5.1 / 6.0
		TDE	33.6 / 46.2 / 51.4	18.5 / 25.5 / 29.1	21.7 / 27.7 / 29.9	9.8 / 13.1 / 14.9	12.4 / 16.9 / 20.3	5.8 / 8.2 / 9.8
	GATE	Baseline	58.9 / 65.5 / 67.4	12.2 / 15.5 / 16.8	36.2 / 39.4 / 40.1	7.2 / 9.0 / 9.5	25.8 / 33.3 / 37.8	5.2 / 7.2 / 8.5
		TDE	38.7 / 50.8 / 55.8	18.5 / 24.9 / 28.3	21.8 / 27.2 / 29.5	11.1 / 13.9 / 15.2	5.9 / 7.4 / 8.4	6.6 / 8.5 / 9.9
VTransE†	SUM	Baseline	59.0 / 65.7 / 67.6	11.6 / 14.7 / 15.8	35.4 / 38.6 / 39.4	6.7 / 8.2 / 8.7	23.0 / 29.7 / 34.3	3.7 / 5.0 / 6.0
		TDE	36.9 / 48.5 / 53.1	17.3 / 24.6 / 28.0	19.7 / 25.7 / 28.5	9.3 / 12.9 / 14.8	13.5 / 18.7 / 22.6	6.3 / 8.6 / 10.5
	GATE	Baseline	58.7 / 65.3 / 67.1	13.6 / 17.1 / 18.6	34.6 / 38.1 / 38.9	6.6 / 8.2 / 8.7	24.5 / 31.3 / 35.5	5.1 / 6.8 / 8.0
		TDE	40.0 / 50.7 / 54.9	18.9 / 25.3 / 28.4	23.0 / 28.8 / 31.1	9.8 / 13.1 / 14.7	13.7 / 19.0 / 22.9	6.0 / 8.5 / 10.2
VCTree†	SUM	Baseline	59.8 / 66.2 / 68.1	11.7 / 14.9 / 16.1	37.0 / 40.5 / 41.4	6.2 / 7.5 / 7.9	24.7 / 31.5 / 36.2	4.2 / 5.7 / 6.9
		TDE	36.2 / 47.2 / 51.6	18.4 / 25.4 / 28.7	19.9 / 25.4 / 27.9	8.9 / 12.2 / 14.0	14.0 / 19.4 / 23.2	6.9 / 9.3 / 11.1
	GATE	Baseline	59.1 / 65.5 / 67.4	12.4 / 15.4 / 16.6	35.4 / 38.9 / 39.8	6.3 / 7.5 / 8.0	24.8 / 31.8 / 36.1	4.9 / 6.6 / 7.7
		TDE	39.1 / 49.9 / 54.5	17.2 / 23.3 / 26.6	22.8 / 28.8 / 31.2	8.9 / 11.8 / 13.4	14.3 / 19.6 / 23.3	6.3 / 8.6 / 10.3

TABLE 4.2: The SGG performances of Relationship Retrieval on both conventional **Recall@K** and **mean Recall@K** [8, 15]. The SGG models reimplemented under our codebase are denoted by the superscript †.

4.5.2.2 Zero-Shot Relationship Retrieval (ZSRR)

It was introduced by Lu *et al.* [186] as **Zero-Shot Recall@K** and was firstly evaluated on VG dataset in this chapter, which only reports the R@K of those subject-predicate-object triplets that have never been observed in the training set. ZSRR also has three sub-tasks as RR.

4.5.2.3 Sentence-to-Graph Retrieval (S2GR)

It uses the image caption sentence as the query to retrieve images represented as SGs. Both RR and ZSRR are triplet-level evaluations, ignoring the graph-level coherence. Therefore, we design S2GR, using human descriptions to retrieve detected SGs. We didn't use proxy vision-language tasks like captioning [89, 192] and VQA [193, 194] as the diagnosis, because their implementations have too many components unrelated to SGG and their datasets are challenged by their own biases [104, 108, 207]. In S2GR, the detected SGs (using SGDet) are regarded as the only representations of images, cut off all the dependencies on black-box visual features, so any bias on SGG would sensitively violate the coherence of SGs, resulting in worse retrieval results. For example, if `walking on` was detected as the biased alternative `on`, images would be mixed up with those have `sitting on` or `laying on`. Note that S2GR is fundamentally different from the previous image

Zero-Shot Relationship Retrieval			PredCls	SGCls	SGDet
Model	Fusion	Method	R@50/100	R@50/100	R@50/100
MOTIFS†	SUM	Baseline	10.9 / 14.5	2.2 / 3.0	0.1 / 0.2
		Focal	10.9 / 14.4	2.2 / 3.1	0.1 / 0.3
		Reweight	0.7 / 0.9	0.1 /0.1	0.0 / 0.0
		Resample	11.1 / 14.3	2.3 / 3.1	0.1 / 0.3
		X2Y	11.8 / 17.6	2.3 / 3.7	1.6 / 2.7
		X2Y-Tr	13.7 / 17.6	3.1 / 4.2	1.8 / 2.8
		TE	14.2 / 18.1	1.4 / 2.0	1.4 / 1.8
		NIE	2.4 / 3.2	0.2 / 0.4	0.3 / 0.6
		TDE	**14.4 / 18.2**	**3.4 / 4.5**	**2.3 / 2.9**
	GATE	Baseline	7.4 / 10.6	0.9 / 1.3	0.2 / 0.4
		TDE	**7.7 / 11.0**	**1.9 / 2.6**	**1.9 / 2.5**
VTransE†	SUM	Baseline	11.3 / 14.7	2.5 / 3.3	0.8 / 1.5
		TDE	**13.3 / 17.6**	**2.9 / 3.8**	**2.0 / 2.7**
	GATE	Baseline	4.2 / 5.9	1.9 / 2.6	**1.9** / 2.6
		TDE	**5.3 / 7.9**	**2.1 / 3.0**	**1.9 / 2.7**
VCTree†	SUM	Baseline	10.8 / 14.3	1.9 / 2.6	0.2 / 0.7
		TDE	**14.3 / 17.6**	**3.2 / 4.0**	**2.6 / 3.2**
	GATE	Baseline	4.4 / 6.8	2.5 / 3.3	1.8 / 2.7
		TDE	**5.9 / 8.1**	**3.0 / 3.7**	**2.2 / 2.8**

TABLE 4.3: The results of Zero-Shot Relationship Retrieval.

retrieval with scene graph [208, 209], because the latter still consider the images as visual features but not SGs. **Recall@20/100 (R@20/100)** and median ranking indexes of retrieved results **(Med)** on the gallery size of 1,000 and 5,000 were evaluated. Note that S2GR should have diverse implementations as long as its spirit: graph-level symbolic retrieval, is fulfilled. We provide our implementation in the next sub-section.

4.5.3 Implementation Details

4.5.3.1 Object Detector

Following the previous works [2, 6, 8], we pre-trained a Faster R-CNN [27] and froze it to be the underlying detector of our SGG models. We equipped the Faster R-CNN with a ResNeXt-101-FPN [18, 26] backbone and scaled the longer side of input images to be 1k pixels. The detector was trained on the training set of VG using SGD as optimizer. We set the batch size to 8 and the initial learning rate to

Sentence-to-Graph Retrieval								
Gallery Size			1000			5000		
Model	Fusion	Method	R@20	R@100	Med	R@20	R@100	Med
MOTIFS†	SUM	Baseline	11.6	39.9	155	3.1	12.1	708
		Focal	10.9	39.0	163	2.9	11.1	737
		Reweight	9.7	36.8	159	3.0	11.4	725
		Resample	13.1	43.6	124	2.5	13.4	593
		X2Y	14.3	44.8	125	3.5	14.6	556
		X2Y-Tr	14.5	45.6	114	3.9	16.8	525
		TE	15.9	49.9	100	4.4	16.9	469
		NIE	6.7	29.2	202	1.6	8.6	1050
		TDE	**17.0**	**53.6**	**91**	**5.2**	**18.9**	**425**
	GATE	Baseline	13.7	45.6	143	4.4	16.2	618
		TDE	**20.8**	**59.2**	**72**	**5.2**	**21.3**	**325**
VTransE†	SUM	Baseline	12.3	42.3	129	**3.6**	15.0	596
		TDE	**14.7**	**48.4**	**106**	**3.6**	**16.3**	**483**
	GATE	Baseline	12.9	41.8	136	3.8	14.3	634
		TDE	**18.5**	**50.4**	**110**	**4.5**	**19.1**	**486**
VCTree†	SUM	Baseline	9.9	37.4	150	3.1	11.5	745
		TDE	**19.0**	**57.0**	**82**	**5.0**	**20.0**	**385**
	GATE	Baseline	13.4	44.1	121	3.7	13.6	583
		TDE	**19.1**	**55.5**	**87**	**5.1**	**20.3**	**395**

TABLE 4.4: The results of Sentence-to-Graph Retrieval.

8×10^{-3}, which was decayed by the factor of 10 on the $30k^{th}$ and $40k^{th}$ iterations. The final detector achieved 28.14 mAP on VG test set (using 0.5 IoU threshold). 4 2080ti GPUs were used for the pre-training.

4.5.4 Scene Graph Generation

On top of the frozen detector, we trained SGG models using SGD as optimizer. Batch size and initial learning rate were set to be 12 and 12×10^{-2} for PredCls and SGCls; 8 and 8×10^{-2} for SGDet. The learning rate would be decayed by 10 two times after the validation performance plateaus. For SGDet, 80 RoIs were sampled for each image and Per-Class NMS [6, 210] with 0.5 IoU was applied in object prediction. We sampled up to 1,024 subject-object pairs containing 75% background pairs during training. Different from previous works [6, 8, 198], we didn't assume that non-overlapping subject-object pairs are invalid in SGDet, making SGG more general.

4.5.5 Sentence-to-Graph Retrieval

We handled S2GR as a graph-to-graph matching problem. The query captions of each image were stuck together and parsed to a text-SG using [209]. We set all the subject/object and predicates that appear less than 5 times to "UNKNOWN" tokens, obtaining a dictionary of size 4,459 subject/object entities and 645 predicates, respectively. The original image SG generated from SGDet contains a fixed number of RoIs and forces all valid subject-object pairs to predict foreground relationships, to serve the K number in mR@K, which is inappropriate for S2GR. Therefore, we used a threshold of 0.1 to filter RoIs by the label probabilities and removed all background predicates from the graph. Recall that the vocabulary size of the entity and predicate for image SGs are 150 and 50 as we mentioned before. To match the two heterogeneous graphs: image SG and text SG, in a unified space, we used BAN [211] to encode the two graph types into fixed-dimension vectors to facilitate the retrieval. More details can be found in supplementary material.

4.5.6 Network Details

In Section 4.2, we simplified the feature extraction module in Link $I \rightarrow X$ and the visual context module in Link $I \rightarrow Y$. Their details will be given in this subsection.

4.5.6.1 Feature Extraction Module

Since we adopted ResNeXt-101-FPN [18, 26] as the backbone, the extracted $\mathcal{M}$ contains feature maps from 4 scales: $(1/4, 1/8, 1/16, 1/32) \rightarrow (\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3)$. Each bounding box will be assigned to the corresponding $\mathcal{M}_k, (k = 0, 1, 2, 3)$ based on their areas [212]. Given a bounding box b_i with area a_i, the corresponding index k of feature map is calculated as follows:

$$k = \max(2, \min(5, \lfloor 4 + \log_2(a_i/224 + 1 \times 10^{-6}) \rfloor)) - 2. \tag{4.15}$$

Then ROIAlign [29] will be applied to the selected bounding box b_i on the corresponding $\mathcal{M}_k$ for the feature r_i as we described in Section 3.

Index	Input	Operation	Output
(1)	$(\mathcal{M}_0, b_i \cup b_j)$	ROIAlign	$(7 \times 7 \times 256)$
(2)	$(\mathcal{M}_1, b_i \cup b_j)$	ROIAlign	$(7 \times 7 \times 256)$
(3)	$(\mathcal{M}_2, b_i \cup b_j)$	ROIAlign	$(7 \times 7 \times 256)$
(4)	$(\mathcal{M}_3, b_i \cup b_j)$	ROIAlign	$(7 \times 7 \times 256)$
(5)	(1-4)	Concatenation	$(7 \times 7 \times 1024)$
(6)	(5)	Conv	$(7 \times 7 \times 256)$
(7)	b_i, b_j	dummy mask	$(27 \times 27 \times 2)$
(8)	(7)	Conv+ReLU+BatchNorm	$(14 \times 14 \times 128)$
(9)	(8)	MaxPool	$(7 \times 7 \times 128)$
(10)	(9)	Conv+Relu+BatchNorm	$(7 \times 7 \times 256)$
(11)	(6),(10)	Element-wise Addition	$(7 \times 7 \times 256)$
(12)	(11)	Flatten	12,544
(13)	(12)	FC+ReLU	4,096
(14)	(13)	FC+ReLU	4,096

TABLE 4.5: The details of Visual Context Module.

4.5.6.2 Visual Context Module

To extract the visual context feature v'_e for the union box $b_i \cup b_j$, we consider all 4 feature maps will provide complementary contextual information from different levels. Therefore, we extract ROIAlign [29] features on all 4 feature maps before we project the visual context feature into a feature space of $\mathbb{R}^{4096}$. The entire module is summarized in the Table 4.5, where the dummy mask operation in (7) generates two masks for b_i and b_j independently, assigning 1.0 to the pixels inside the bounding box and 0.0 for the rest.

4.5.6.3 The Special Treatment for PredCls

In previous sections, we skipped a special case of causal graph, *i.e.*, causal graph for Predicate Classification (PredCls), for simplification. In PredCls, the ground truth object labels are given, which means the link $X \rightarrow Z$ is blocked by assigning ground truth labels. It won't affect TDE calculation, where Z takes the real value z. However, it's involved in the ablation studies of TE and NIE, where Z could be assigned to $\bar{z}$. In this case, $\bar{z}$ will directly use to the mean vector of training set rather than be calculated from Eq.(2). We also need to notice that, for MOTIFS [6], Eq.(3) will take z_e as input too, which is simplified in previous, because z_e itself is

derived from x_e and it can be considered as the interaction between link $X \rightarrow Y$ and $Z \rightarrow Y$ in the causal graph.

4.5.6.4 Sentence-to-Graph Retrieval

As we mentioned before, we treated Sentence-to-Graph Retrieval (S2GR) as the graph-to-graph matching problem, parsing query captions to text-SGs by [209]. Both detected image-SGs and parsed text-SGs are composed of entities $E^k = \{e_i^k\}$ and relationships $R^k = \{r_{ij}^k = (s_i^k, p_{ij}^k, o_j^k)\}$, where $k \in \{text, image\}$, subject and object categories (s_i^k, o_j^k) share the same dictionary with e_i^k for each k, p_{ij}^k denotes the onehot vector of the predicate category.

The image-SGs and text-SGs are equipped with different embedding layers, because they have different dictionaries. The entities and relationships are encoded as:

$$E_{embed}^k = W_e^k E^k, \tag{4.16}$$

$$R_{embed}^k = [W_s^k S^k; W_p^k P^k; W_o^k O^k], \tag{4.17}$$

where $E_{embed}^k \in \mathbb{R}^{N_d \times N_e^k}$, $R_{embed}^k \in \mathbb{R}^{3N_d \times N_r^k}$, $N_d = 512$ is the dimension of embedded feature, N_e^k, N_r^k are numbers of entities and relationships for each image.

4.5.6.5 Bilinear Attention Scene Graph Encoding

Since entities and relationships are both important for SGs, we apply Bilinear Attention Network (BAN) [211] to encode their multimodal interactions into the same representation space. The same BAN model is used for both text-SGs and image-SGs, hence we remove k hereinafter for simplification. The original BAN involves two steps: 1) attention map generation, and 2) bilinear attended feature calculation. Because scene graph has already provides connections between entities and relationships, we skipped the first step and used normalized scene graph connection as attention map $A_{ij} = M_{ij} / \sum_j M_{ij}$, where $A, M \in \mathbb{R}^{N_e \times N_r}$, the scene graph connection M is defined as follows:

$$M_{ij} = \begin{cases} 1, \ if \ E_i \ in \ R_j, \\ 0, \ if \ E_i \ not \ in \ R_j. \end{cases} \tag{4.18}$$

Index	Input	Loop	Operation	Output
(1)	E_{embed}		Input Shape	$(N_e \times 512)$
(2)	R_{embed}		Input Shape	$(N_r \times 512)$
(3)	A		Input Shape	$(N_e \times N_r)$
(4)	(1)	start	Transpose + Unsqueeze	$(512 \times 1 \times N_e)$
(5)	(2)	$\downarrow$	Transpose + Unsqueeze	$(512 \times N_r \times 1)$
(6)	(3)	$\downarrow$	Unsqueeze	$(1 \times N_e \times N_r)$
(7)	(4),(6)	$\downarrow$	Matrix Multiplication	$(512 \times 1 \times N_r)$
(8)	(5),(7)	$\downarrow$	Matrix Multiplication	$(512 \times 1 \times 1)$
(9)	(8)	$\downarrow$	Squeeze + FC	(512)
(10)	(4),(9)	end	Unsqueeze + Element-wise Addition	$(512 \times 1 \times N_e)$
(11)	(10)		Sum Over N_e	512
(12)	(11)		FC + ReLU + FC + ReLU	1024

TABLE 4.6: The details of Bilinear Attention Scene Graph Encoding Module.

The bilinear attended scene graph encoding is calculated by Table 4.6, where steps (4-10) are calculated 2 times, and the final output $E_{graph} \in \mathbb{R}^{1024}$ is a feature vector representing the whole SG. The same BAN is used for both text-SG or image-SG, *i.e.*, the parameters of the BAN are shared.

The model was trained by the triplet loss [213] with L1 distance. The model was trained in 30 epochs by SGD optimizer and set batch size to be 12. Learning rate was set to be 12×10^{-2}, which was decayed at 10^{th} and 25^{th} epochs by the factor of 10.

4.5.7 Ablation Studies

Except for the models and fusion functions that we've discussed before, we also investigated three conventional debiasing methods, two intuitive causal graph surgeries, and other two types of causal effects: 1) **Focal**: focal loss [144] automatically penalizes well-learned samples and focuses on the hard ones. We followed the hyper-parameters ($\gamma = 2.0, \alpha = 0.25$) optimized in [144]. 2) **Reweight**: weighted cross-entropy is widely used in the industry for biased data. The inversed sample fractions were assigned to each predicate category as weights. 3) **Resample** [143]: rare categories were up-sampled by the inversed sample fraction during training. 4) **X2Y**: since we argued that the unbiased effect was under the effect of object features X, it directly generated SG by the outputs of $X \rightarrow Y$ branch after biased training. 5) **X2Y-Tr**: it even cut off other branches, using $X \rightarrow Y$ for both

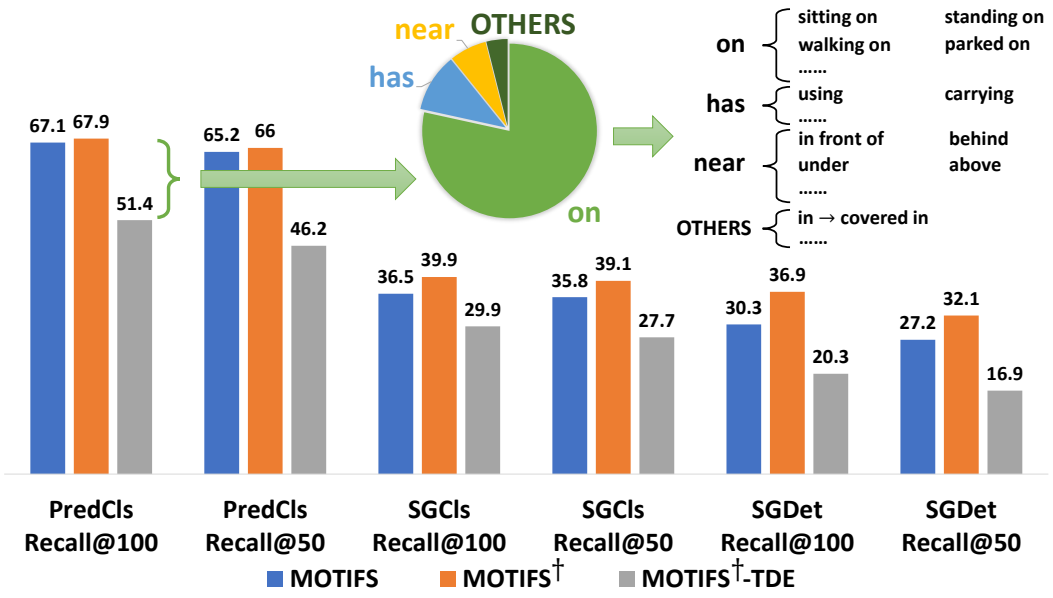


FIGURE 4.9: The pie chart summarizes all the relationships, that are correctly detected by the baseline model but considered "incorrect" by TDE. The right side of the pie chart shows the corresponding labels given by the TDE. Combining with our qualitative examples, we believe that the drop of Recall@K is caused by two reasons: 1) the annotators' preference towards simple annotations caused by bounded rationality [7], 2) TDE tends to predict more action-like relationships rather than vague prepositions.

training and testing. 6) **TE**: as we introduced in Section 4.3, TE is the debiasing method that not conditioned on the contexts. 7) **NIE**: it is the marginal difference between TDE and TE, *i.e.*, NIE = TE-TDE, which can be considered as the pure effect caused by introducing the bias $Z \rightarrow Y$. **NOTE**: although zero vector can also be used as the wiped-out input $\bar{x}$, we chose the mean feature of training set for minor improvements.

4.5.8 Quantitative Studies

4.5.8.1 RR & ZSRR

The results are listed in Table 4.1& 4.3. Despite the fact that conventional debiasing methods: Reweight and Resample, directly hack the mR@K metric, they only gained limited advantages in RR but not in ZSRR. In contrast to the high mR@K of Reweight in RR SGDet, it got embarrassingly 0.0/0.0 in ZSRR SGDet, indicating that such debiased training methods ruin the useful context prior. Focal loss [144]

barely worked for both RR and ZSRR. Causal graph surgeries, X2Y and X2Y-Tr, both improved RR and ZSRR from the baseline, yet their increases were limited. TE had a very similar performance to TDE, but as we discussed, it removed the general bias rather than the subject-object specific bias. NIE is the marginal improvements from TE to TDE, which was even worse than baseline. Although R@K is not a qualified metric for RR as we discussed, we still reported the R@50/100 performance of MOTIFS†-SUM in Figure 4.9. We can observe a performance drop from baseline to TDE, but a further analysis shows that those considered as correct in baseline and "incorrect" in TDE were mainly the "head" predicates, and they are classified by TDE into more fine-grained "tail" classes. Among all three models and two fusion functions, even the worst TDE performance outperforms previous state-of-the-art methods [8, 15] by a large margin on RR mR@K.

The full results of Relationship Retrieval, including both conventional Recall@K and the adopted mean Recall@K [8, 15], are given in Table 4.2. Although a performance drop on conventional Recall@k is observed on TDE, the detailed analysis of the "decreased" predicates in Figure 4.9 implies that it's caused by a more fine-grained predicate classification.

The detailed predicate-level Recall@100 on PredCls of all three models, two fusion functions and baseline *vs.* TDE are given in Figure 4.12 4.13 4.14. Impressively, the distribution of the improved performances is no longer long-tailed while those conventional debiasing methods illustrated in Figure 4.11 can't surpass the dataset distribution anyway. For TDE, very few decreased predicates are mainly due to the more fine-grained classification and we can observe significant improvements on their subclass predicates. Note that, unlike Reweight, which blindly hurt all frequent predicates, the proposed TDE will even improve some of the top-10 frequent predicates, like `behind` and `above`, which themselves are the subclasses of `near`. It further proves that the improvement of the proposed TDE doesn't come from hacking the distribution.

4.5.8.2 S2GR

In S2GR, Focal and Reweight are even worse than the baseline. Among all the three conventional debiasing methods, Resample was the most stable one based on our experiments. X2Y and X2Y-Tr have minor advantages over baseline. TE takes

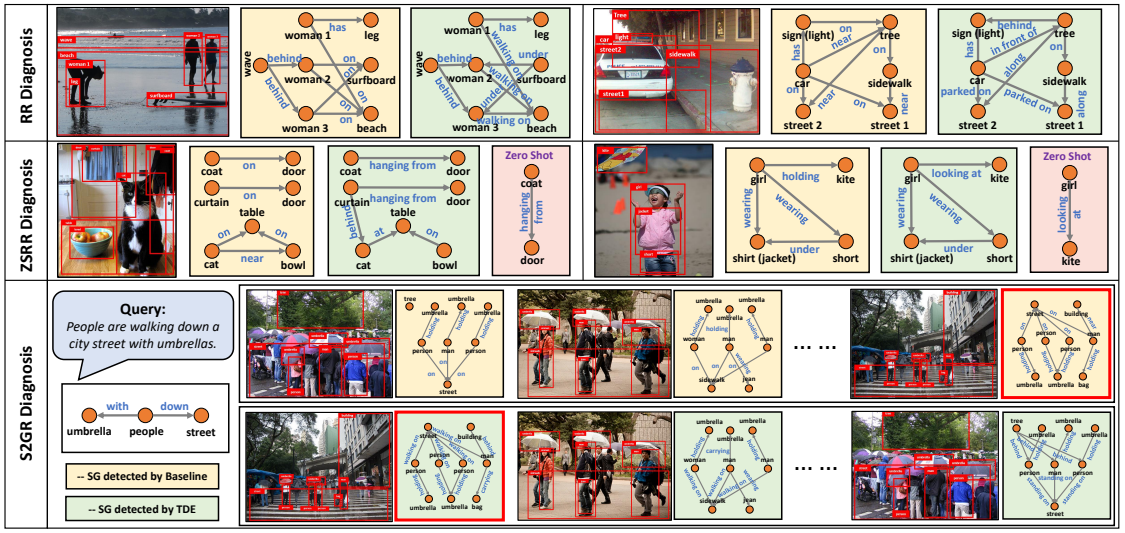


FIGURE 4.10: Results of scene graphs generated from MOTIF†-SUM baseline (yellow) and corresponding TDE (green). Top: relationship retrieval results. Mid: zero shot relationship retrieval results. Red boxes indicate the zero shot triplets. Bottom: results of S2GR. Red boxes mean the correctly retrieved SGs. Part of the trivial detected objects are removed from the graphs, due to space limitation.

the 2nd place and was only a little bit worse than TDE. NIE is the worst as we expected because it is only based on the pure context bias. It is worth highlighting that all the three models and two fusion functions had significant improvements after we applied TDE.

4.5.9 Qualitative Studies

We visualized several SGCls examples that generated from MOTIFS†-SUM baseline and TDE in the top and mid rows of Figure 4.10, scene graphs generated by TDE are much more discriminative compared to the baseline model which prefers trivial predicates like `on`. The right half of the mid row shows that the baseline model would even generate `holding` due to the long-tail bias when the girl is not touching the kite, implying that the biased predictions are easy to be "blind", while TDE successfully predicted `looking at`. The bottom of Figure 4.10 is an example of S2GR, where the SGs detected by baseline model lost the detailed actions of people, considering both `person walking on street` and `person standing on street` as `person on street`, which caused worse retrieval results. All the examples show

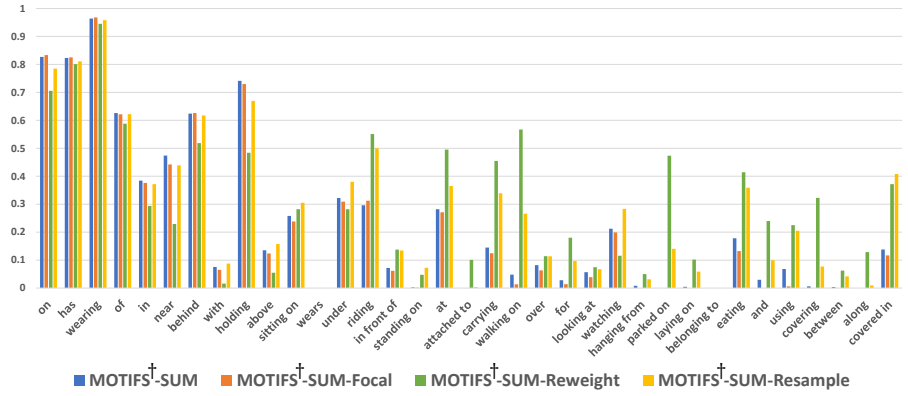


FIGURE 4.11: Conventional Debiasing Methods: Recall@100 on Predicate Classification for the most frequent 35 predicates.

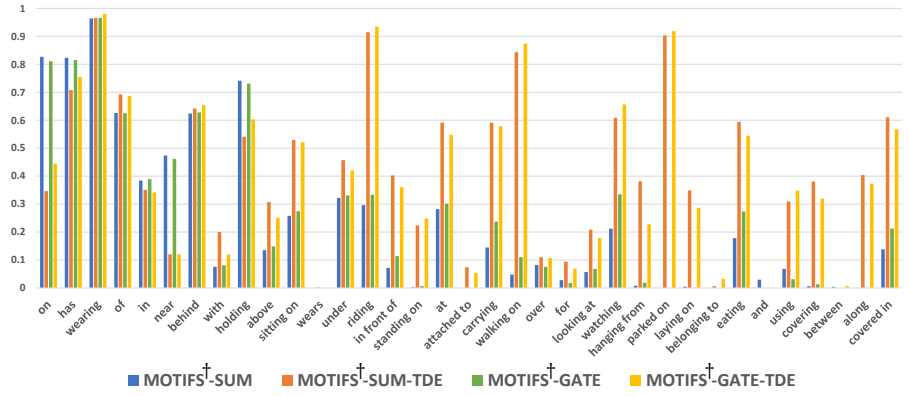


FIGURE 4.12: MOTIFS† [6]: Recall@100 on Predicate Classification for the most frequent 35 predicates.

a clear trend that TDE is much more sensitive to those semantically informative relationships instead of the trivially biased ones.

More Relationship Retrieval (RR) and Zero-Shot Relationship Retrieval (ZSRR) results are given in Figure 4.15, where top 10 relationships under SGCls are selected for each image. As we can see, other than the trivial relationship problem, conventional baseline barely distinguishes different entities. For example, in the left bottom image, the same `sign` is almost `on` every `pole` in the baseline while the TDE results are more sensitive to different entities. However, one of the problem of TDE is that it over emphasizes the action predicates. It even uses `holding` for `pole` and `sign` while the predicate `on` used by the baseline is more natural in this case.

Another example of Sentence-to-Graph Retrieval (S2GR) is illustrated in Figure 4.16. Although we only reported sub-graphs of the original SGDet results,

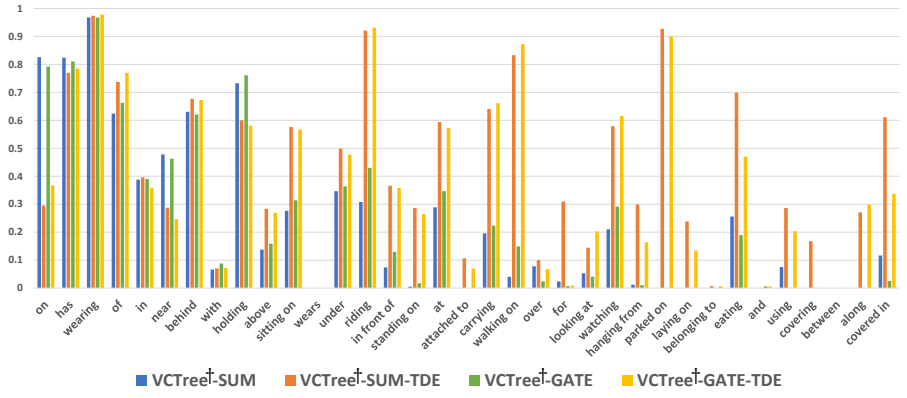


FIGURE 4.13: VCTree† [8]: Recall@100 on Predicate Classification for the most frequent 35 predicates.

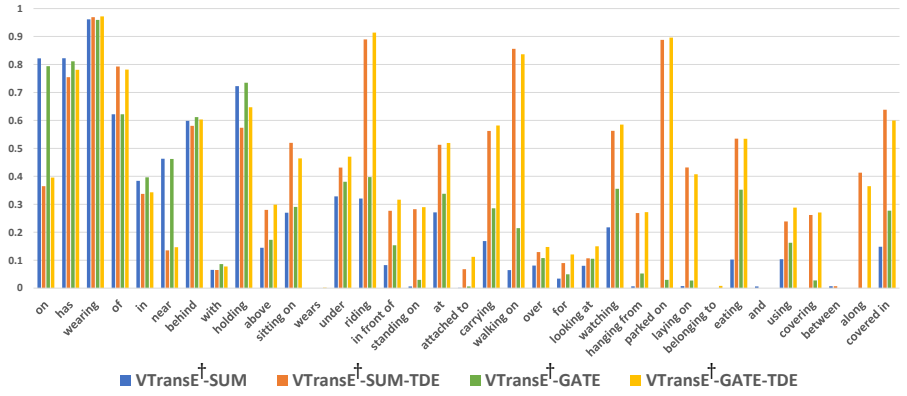


FIGURE 4.14: VTransE† [9]: Recall@100 on Predicate Classification for the most frequent 35 predicates.

due to the limited space, we can still find that the conventional baseline model is not able to detect predicate like `eating`, which causes the detected SGs only provide the spatial relationships, missing the most discriminative word `eating` in the query caption.

4.6 Conclusions

We presented a general multimodal TDE framework, which can be used to obtain the unbiased predictions from biased training in multimodal tasks. To be specific, we examined the proposed framework on Scene Graph Generation, which is the first work addressing the serious bias issue in SGG. With the power of *counterfactual causality*, we can remove the harmful bias from the good context bias, *i.e.*,

disentangle effects from different modalities, which cannot be easily identified by traditional debiasing methods such as data augmentation [139, 142] and unbiased learning [144]. We achieved the unbiasedness by calculating the Total Direct Effect (TDE) with the help of a causal graph, which is a roadmap for training any SGG model. By using the proposed Scene Graph Diagnosis toolkit, our unbiased SGG results are considerably better than their biased counterparts. What's more, it provided a novel and effective strategy to obtain the robustness in various multimodal tasks.

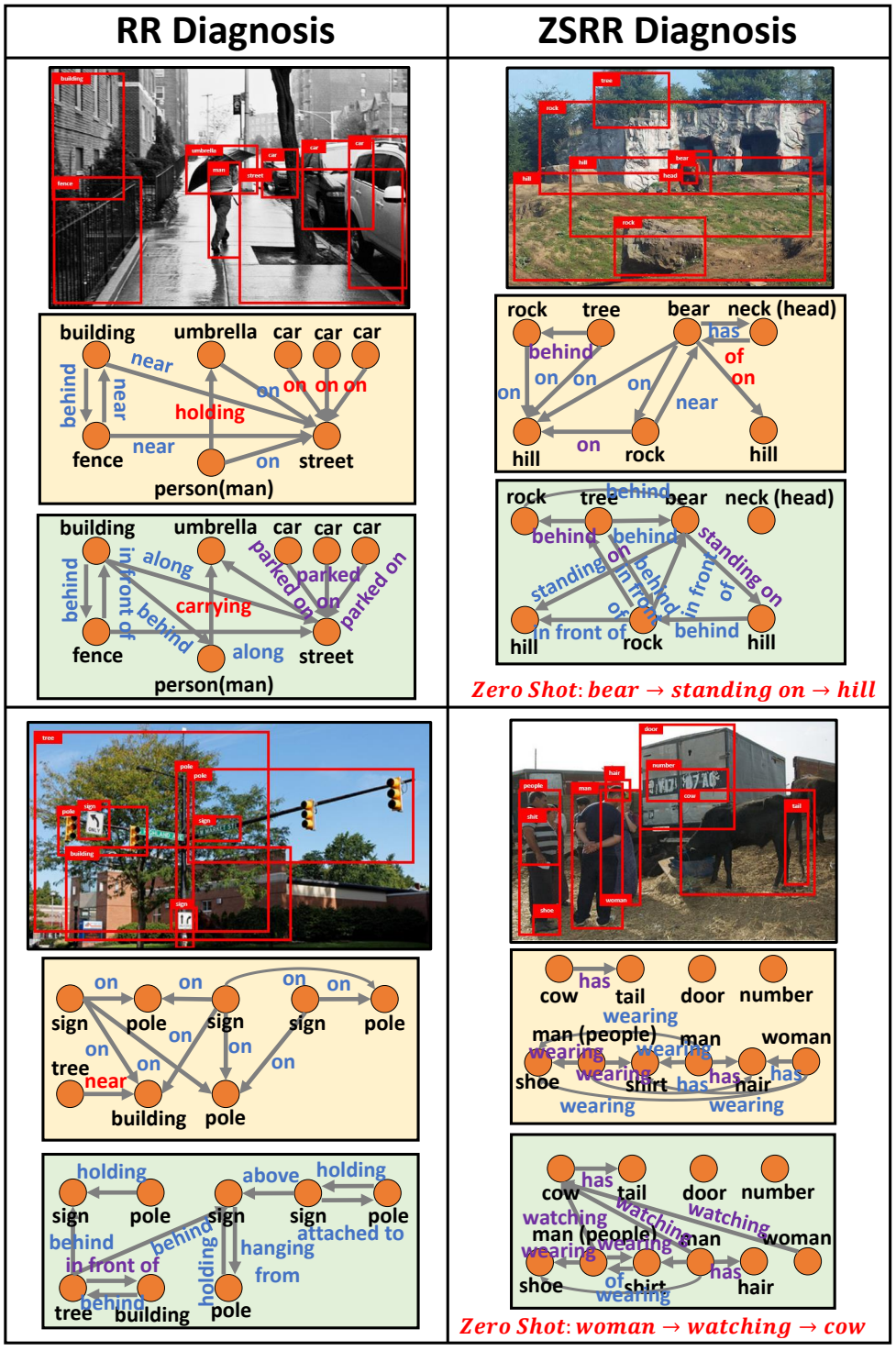


FIGURE 4.15: Top 10 Relationship Retrieval (RR) and Zero-Shot Relationship Retrieval (ZSRR) results of SGCls for MOTIFS†+SUM baseline (yellow box) and corresponding TDE (green box). The red predicates indicate misclassified relationships, the purple predicates are those correctly classified relationships (in ground truth), the blue predicates are those not labeled in ground truth.

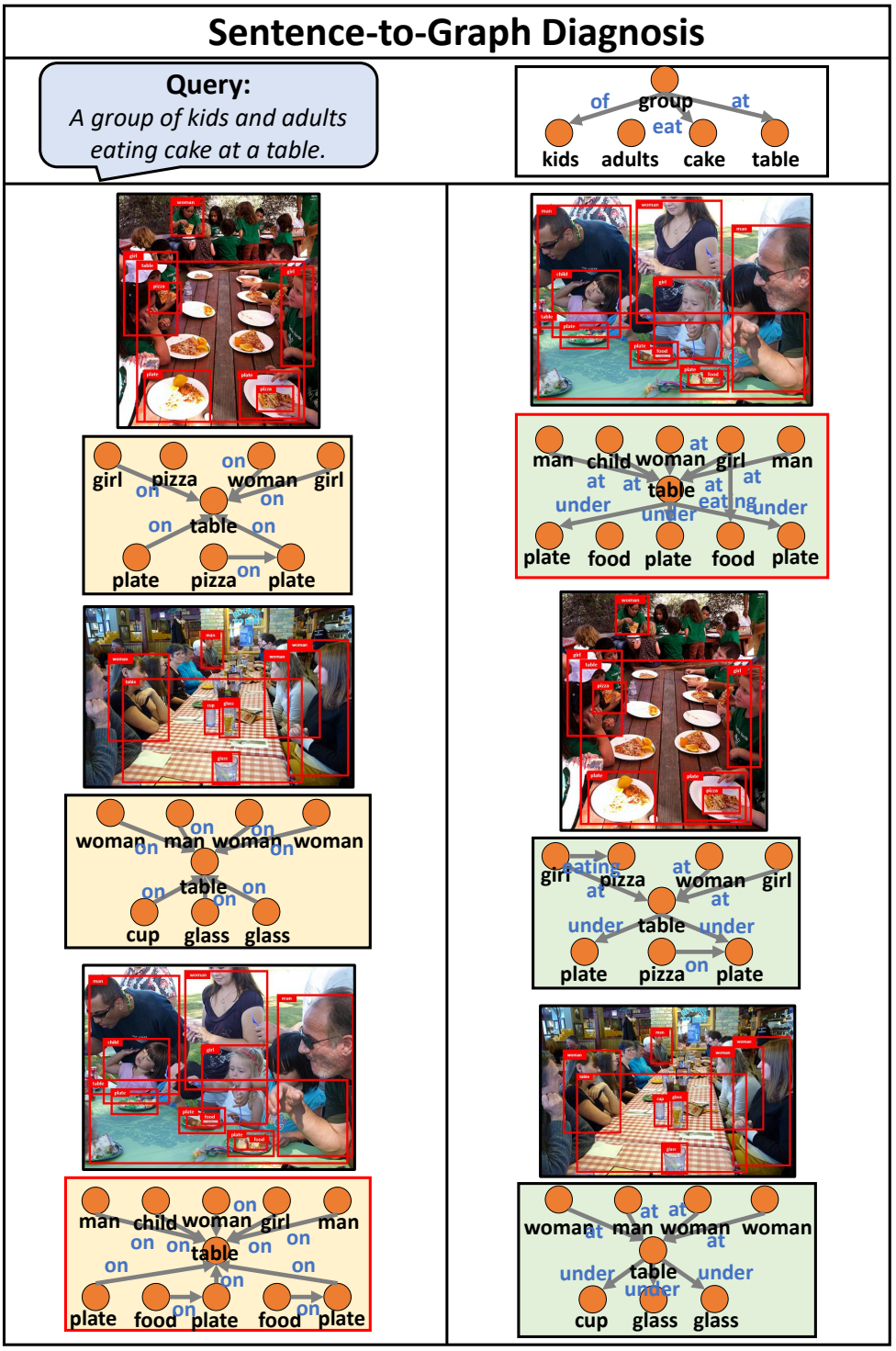


FIGURE 4.16: An example of Sentence-to-Graph Retrieval (S2GR) results for MOTIFS†+SUM baseline (yellow box) and corresponding TDE (green box). The red boxes indicate ground truth matching results. Note that we only draw subgraphs containing important objects and predicates, because the original detected scene graphs from SGDet have too many trivial objects and predicates.

Chapter 5

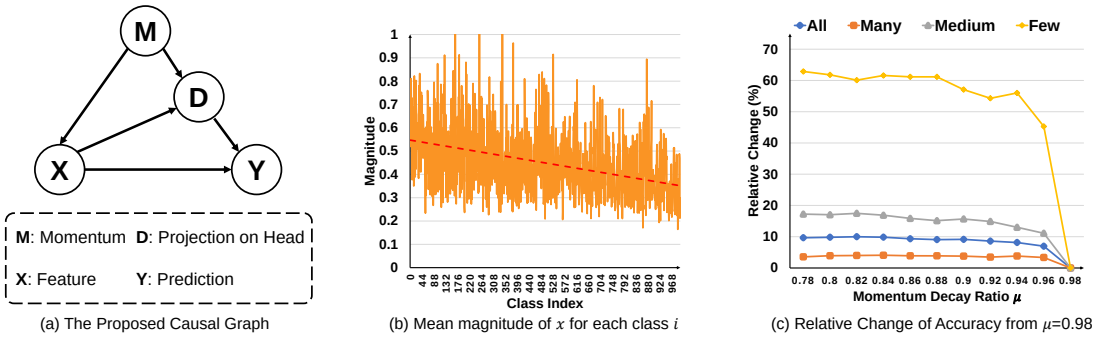
De-confound TDE for General Long-Tailed Robustness[1]

5.1 Introduction

In the previous chapter, we investigated the long-tailed bias in the "mid-level" vision task, Scene Graph Generation, and proposed a multimodal TDE to alleviate the bias under multiple modalities. However, the highly-skewed long-tailed distribution also ubiquitously exists in other low-level vision tasks, like image classification, object detection or instance segmentation, etc.

Over the years, we have witnessed the fast development of computer vision techniques [24, 26, 27], stemming from large and balanced datasets such as ImageNet [69] and MS-COCO [70]. Along with the growth of the digital data created by us, the crux of making a large-scale dataset is no longer about where to collect, but how to balance. However, the cost of expanding them to a larger class vocabulary with balanced data is not linear — but exponential — as the data will be inevitably long-tailed by Zipf's law [71]. Specifically, a single sample increased for any data-poor tail class will result in more samples from the data-rich head. What's worse, sometimes, re-balancing the class is impossible. For example, in instance segmentation [19], if we target at increasing the images of tail class instances

[1]The work in this chapter has been published in the paper : Kaihua Tang, Jianqiang Huang, Hanwang Zhang. "Long-Tailed Classification by Keeping the Good and Removing the Bad Momentum Causal Effect." Advances in Neural Information Processing Systems (**NeurIPS, Poster**). Virtual. 2020.



(a) The Proposed Causal Graph (b) Mean magnitude of x for each class i (c) Relative Change of Accuracy from μ=0.98

FIGURE 5.1: (a) The proposed causal graph explaining the causal effect of momentum. See Section 5.2 for details. (b) The mean magnitudes of feature vectors for each class i after training with momentum $\mu = 0.9$, where i is ranking from head to tail. (c) The relative change of the performance on the basis of $\mu = 0.98$ shows that the few-shot tail is more vulnerable to the momentum.

like "remote controller", we have to bring in more head instances like "sofa" and "TV" simultaneously in every newly added image [72].

Therefore, long-tailed classification is indispensable for training deep models at scale. Recent work [10, 11, 16] starts to focus on the long-tailed robustness by filling in the performance gap between class-balanced and long-tailed datasets, while new long-tailed benchmarks are springing up such as Long-tailed CIFAR-10/-100 [16, 111], ImageNet-LT [10] for image classification and LVIS [19] for object detection and instance segmentation. Despite the vigorous development of this field, we find that the fundamental theory is still missing. We conjecture that it is mainly due to the paradoxical effects of long tail. On one hand, it is bad because the classification is severely biased towards the data-rich head. On the other hand, it is good because the long-tailed distribution essentially encodes the natural inter-dependencies of classes — "TV" is indeed a good context for "controller" — any disrespect of it will hurt the feature representation learning [16], *e.g.*, re-weighting [170, 171] or re-sampling [167, 168] inevitably causes under-fitting to the head or over-fitting to the tail.

Inspired by the above paradox, latest studies [11, 16] show promising long-tailed robustness in disentangling the "good" from the "bad", by the naïve two-stage separation of *imbalanced* feature learning and *balanced* classifier training. However, such disentanglement does not explain the whys and wherefores of the paradox, leaving critical questions unanswered: given that the re-balancing causes under-fitting/over-fitting, why is the re-balanced classifier good but the re-balanced feature learning bad? The two-stage design clearly defies the end-to-end merit that

we used to believe since the deep learning era; but why does the two-stage training significantly outperform the end-to-end one in long-tailed classification?

In this chapter, we propose a causal framework that not only fundamentally explains the previous methods [10, 11, 16, 39, 167, 168], but also provides a principled solution to further improve long-tailed classification. The proposed causal graph of this framework is given in Figure 5.1 (a). We find that the momentum M in any SGD optimizer [73, 74] (also called betas in Adam optimizer [214]), which is indispensable for stabilizing gradients, is a confounder who is the common cause of the sample feature X (via $M \rightarrow X$) and the classification logits Y (via $M \rightarrow D \rightarrow Y$). In particular, D denotes the X's projection on the head feature direction that eventually deviates X. We will justify the graph later in Section 5.2. Here, Figure 5.1 (b&c) sheds some light on how the momentum affects the feature X and the prediction Y. From the causal graph, we may revisit the "bad" long-tailed bias in a causal view: the backdoor [215] path $X \leftarrow M \rightarrow D \rightarrow Y$ causes the spurious correlation even if X has nothing to do with the predicted Y, *e.g.*, misclassifying a tail sample to the head. Also, the mediation [202] path $X \rightarrow D \rightarrow Y$ mixes up the pure contribution made by $X \rightarrow Y$. For the "good" bias, $X \rightarrow D \rightarrow Y$ respects the inter-relationships of the semantic concepts in classification, that is, the head class knowledge contributes a reliable evidence to filter out wrong predictions. For example, if a rare sample is closer to the head class "TV" and "sofa", it is more likely to be a living room object (*e.g.*, "remote controller") but not an outdoor one (*e.g.*, "car").

Based on the graph that explains the paradox of the "bad" and "good", we propose a principled solution for long-tailed classification. It is a natural derivation of pursuing the direct causal effect along $X \rightarrow Y$ by removing the momentum effect. Thanks to causal inference [149], we can elegantly keep the "good" while remove the "bad". First, to learn the model parameters, we apply de-confounded training with causal intervention: while it removes the "bad" by *backdoor adjustment* [215] who cuts off the backdoor confounding path $X \leftarrow M \rightarrow D \rightarrow Y$, it keeps the "good" by retaining the mediation $X \rightarrow D \rightarrow Y$. Second, we calculate the direct causal effect of $X \rightarrow Y$ as the final prediction logits. It disentangles the "good" from the "bad" in a *counterfactual* world, where the bad effect is considered as the Y's indirect effect when X is zero but D retains the value when $X = \boldsymbol{x}$. In contrast to the prevailing two-stage design [11] that requires unbiased re-training

in the 2nd stage, our solution is one-stage and re-training free. Interestingly, as discussed in Section 5.3.4, we show that why the re-training is inevitable in their method and why ours can avoid it with even better performance.

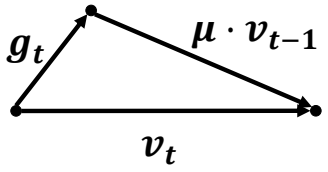
On image classification benchmarks Long-tailed CIFAR-10/-100 [16, 111] and ImageNet-LT [10], we outperform previous state-of-the-arts [11, 16] on all splits and settings, showing that the performance gain is not merely from catering to the long tail or a specific imbalanced distribution. In object detection and instance segmentation benchmark LVIS [19], our method also has a significant advantage over the former winner [39] of LVIS 2019 challenge. We achieve 3.5% and 3.1% absolute improvements on mask AP and box AP using the same Cascade Mask R-CNN with R101-FPN backbone [17]. The proposed De-confound-TDE successfully fill the gap of previous multimodal TDE, obtaining the robust DNNs on pure computer vision tasks.

5.2 A Causal View on Momentum Effect

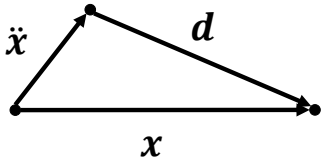
To systematically study the long-tailed classification and how momentum affects the prediction, we construct a **causal graph** [149, 202] in Figure 5.1 (a) with four variables: momentum (M), object feature (X), projection on head direction (D), and model prediction (Y). As we introduced in the Chapter 2, the causal graph is a directed acyclic graph used to indicate how variables of interest $\{M, X, D, Y\}$ interacting with each other through causal links.

The nodes M and D constitute a confounder and a mediator, respectively. A *confounder* is a variable that influences both correlated and independent variables, creating a spurious statistical correlation. Considering a causal graph **exercise** $\leftarrow$ **age** $\rightarrow$ **cancer**, the elder people spend more time on physical exercise after retirement and they are also easier to get cancer due to the elder age, so the confounder *age* creates a spurious correlation that more physical exercise will increase the chance of getting cancer. The example of a *mediator* would be **drug** $\rightarrow$ **placebo** $\rightarrow$ **cure**, where mediator *placebo* is the side effect of taking *drug* that prevents us from getting the direct effect of **drug** $\rightarrow$ **cure**.

Before we delve into the rationale of our causal graph, let's take a brief review on the SGD with momentum [74]. Without loss of generality, we adopt the Pytorch



(a) Decompose the gradient velocity



(b) Decompose the biased feature vector

FIGURE 5.2: Based on Assumption 1, the feature vector $\boldsymbol{x}$ can be decomposed into a discriminative feature $\ddot{\boldsymbol{x}}$ and a projection on head direction $\boldsymbol{d}$

implementation [216]:

$$v_t = \underbrace{\mu \cdot v_{t-1}}_{momentum} + g_t, \qquad \theta_t = \theta_{t-1} - lr \cdot v_t, \tag{5.1}$$

where the notations in the t-th iteration are: model parameters θ_t, gradient g_t, velocity v_t, momentum decay ratio μ, and learning rate lr. Other versions of SGD [73, 74] only change the position of some hyper-parameters and we can easily prove them equivalent with each other. The use of momentum in SGD optimizer considerably dampens the oscillations caused by each single sample. In our causal graph, momentum M is the overall effect of $\mu \cdot v_{T-1}$ at the convergence $t = T$, which is the exponential moving average of the gradient over all past samples with decay rate μ. Eq. (5.1) shows that, given fixed hyper-parameters μ and lr, each sample $M = \boldsymbol{m}$ is a function of the model initialization and the mini-batch sampling strategy, that is, M has infinite samples.

In a balanced dataset, the momentum is equally contributed by every class. However, when the dataset is long-tailed, it will be dominated by the head samples, emerging the following causal links:

$M \rightarrow X$. This link shows that the backbone parameters used to generate feature vectors X, are trained under the effect of M. This is obvious from Eq. (5.1) and can be illustrated in Figure 5.1 (b), where we visualize how the magnitudes of X change from head to tail.

$(M, X) \rightarrow D$. This link denotes that the momentum also causes feature vector X deviates to the head direction D, which is also determined by M. In a long-tailed dataset, few head classes possess most of the training samples, who have less variance than the data-poor but class-rich tail, so the moving averaged momentum will thus point to a stable head direction. Specifically, as shown in Figure 5.2, we

can decompose any feature vector $\boldsymbol{x}$ into $\boldsymbol{x} = \ddot{\boldsymbol{x}} + \boldsymbol{d}$, where $D = \boldsymbol{d} = \hat{\boldsymbol{d}} cos(\boldsymbol{x}, \hat{\boldsymbol{d}}) \|\boldsymbol{x}\|$. In particular, the head direction $\hat{\boldsymbol{d}}$ is given in Assumption 1, whose validity is detailed in Section 5.2.1.

Assumption 1: *The head direction $\hat{\boldsymbol{d}}$ is the unit vector of the exponential moving average features with decay rate μ like momentum, i.e., $\hat{\boldsymbol{d}} = \overline{\boldsymbol{x}}_T / \|\overline{\boldsymbol{x}}_T\|$, where $\overline{\boldsymbol{x}}_t = \mu \cdot \overline{\boldsymbol{x}}_{t-1} + \boldsymbol{x}_t$ and T is the number of the total training iterations.*

Note that the above assumption says that the head direction is exactly determined by the sample moving average in the dataset, which does not need the accessibility of the class statistics at all. In particular, as we show in Section 5.2.1, when the dataset is balanced, Assumption 1 also holds but suggests that $X \to Y$ is naturally not affected by M.

$X \to D \to Y$ & $X \to Y$. These links indicate that the effect of X can be disentangled into an indirect (mediation) and a direct effect. Thanks to the above orthogonal decomposition: $\boldsymbol{x} = \ddot{\boldsymbol{x}} + \boldsymbol{d}$, the indirect effect is affected by $\boldsymbol{d}$ while the direct effect is affected by $\ddot{\boldsymbol{x}}$, and they together determine the total effect. As shown in Figure 5.4, when we change the scale parameter α of $\boldsymbol{d}$, the performance of the tail classes monotonically increases with α, which inspires us to remove the mediation effect of D in Section 5.3.2.

5.2.1 Additional Explanations of Assumption 1

To better understand the $(M, X) \to D$ and Assumption 1, let's take a simple example. Given a learnable parameter $\theta \in \mathcal{R}^2$, and its gradients of instances for class A, B approximate to (1, 1) and (-1, 1) respectively. If each of these two classes has 50 samples, the mean gradient would be (0, 1), which is the optimal gradient direction shared by both A and B. The momentum will thus accelerate on this direction that optimizes the model to fairly discriminate two classes. However, if there are 99 samples from class A and only 1 sample from class B (long-tailed dataset), the mean gradient would be (0.98, 1). In this case, the momentum direction now approximates to the class A (head) gradients, encouraging the backbone parameters to generate head-like feature vectors, *i.e.*, creating an unfair deviation towards the head.

Since the momentum in SGD [73, 74, 216] usually dominates the gradient velocity, the effect of such a deviation is not trivial, which will eventually create the head projection D on all feature vectors generated by the backbone. It's worth noting that although there are non-linear activation layers in the backbone, due to the central limit theorem [217], the overall effect of these deviated parameters is still following the normal distribution, which means we can use the moving averaged feature to approximate this head direction, *i.e.*, the Assumption 1 in the previous section.

In addition, even in a balanced dataset, the Assumption 1 still holds. Considering the above example, the mean gradient is (0, 1) for balanced A and B, which is not biased towards either direction: (1, 1) or (-1, 1). In other word, the D still exists for the balanced dataset, but the $cos(\boldsymbol{x}, \hat{\boldsymbol{d}})$ should be almost the same for all classes. Therefore, the $M \rightarrow D \rightarrow Y$ won't cause any preference in the balanced dataset, which naturally allows $X \rightarrow Y$ free from the effect of M. It's also intuitively easy to understand, because when the dataset is balanced, the mean feature only represents the common patterns shared by all classes, *e.g.*, the D in a balanced face recognition dataset is the mean face, which would be a contour of human head that not biased towards any specific face categories.

5.3 The Proposed Solution

Based on the proposed causal graph in Figure 5.1 (a), we can delineate our goal for long-tailed classification: the pursuit of the direct causal effect along $X \rightarrow Y$. In causal inference, it is defined as Total Direct Effect (TDE) [202, 203]:

$$\arg\max_{i \in C} \quad TDE(Y_i) = [Y_{\boldsymbol{d}} = i | do(X = \boldsymbol{x})] - [Y_{\boldsymbol{d}} = i | do(X = \boldsymbol{x}_0)], \tag{5.2}$$

where $\boldsymbol{x}_0$ denotes a null input (0 in this chapter). We define the causal effect as the prediction logits Y_i for the i-th class. Subscript $\boldsymbol{d}$ denotes that the mediator D always takes the value $\boldsymbol{d}$ in the *deconfounded* causal graph model of Figure 5.1 (a) with $do(X = \boldsymbol{x})$, where the *do*-operator denotes the causal intervention [149] that modifies the graph by $M \not\rightarrow X$. Thus, Eq. (5.2) shows an important principle in long-tailed classification: before we calculate the final TDE (Section 5.3.2), we need

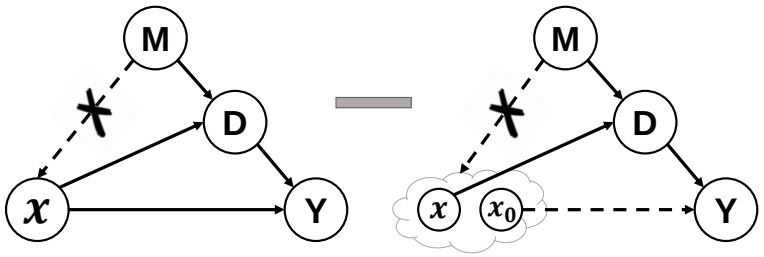


FIGURE 5.3: The TDE inference (Eq. (5.2)) for the long-tailed classification after de-confounded training. Subtracted left: $[Y_{\boldsymbol{d}} = i|do(X = \boldsymbol{x})]$, minus right: $[Y_{\boldsymbol{d}} = i|do(X = \boldsymbol{x}_0)]$.

to first perform de-confounded training (Section 5.3.1) to estimate the "modified" causal graph parameters.

We'd like to highlight that Eq. (5.2) removes the "bad" while keeps the "good" in a reconcilable way. First, in training, the *do*-operator removes the "bad" confounder bias while keeps the "good" mediator bias, because the *do*-operator retains the mediation path. Second, in inference, the mediator value $\boldsymbol{d}$ is imposed in both terms to keep the "good" of the mediator bias (towards head) in the predicted logit; it also removes its "bad" by subtracting the second term: the prediction when the input X is null ($\boldsymbol{x}_0$) but the mediator D is still the value $\boldsymbol{d}$ when X had been $\boldsymbol{x}$. Note that such a *counterfactual* minus elegantly characterizes the "bad" mediation bias, just like how we capture the tricky placebo effect: we cheat the patient to take a placebo drug, setting the direct drug effect **drug** $\rightarrow$ **cure** to zero; thus, any cure observed must be purely due to the non-zero placebo effect **drug** $\rightarrow$ **placebo** $\rightarrow$ **cure**.

5.3.1 De-confounded Training

The model for the proposed causal graph is optimized under the causal intervention $do(X = \boldsymbol{x})$, which aims to preserve the "good" feature learning from the momentum and cut off its "bad" confounding effect. We apply the backdoor adjustment [215] to derive the de-confounded model:

$$P(Y = i|do(X = \boldsymbol{x})) = \sum_{\boldsymbol{m}} P(Y = i|X = \boldsymbol{x}, M = \boldsymbol{m})P(M = \boldsymbol{m}) \tag{5.3}$$

$$= \sum_{\boldsymbol{m}} \frac{P(Y = i, X = \boldsymbol{x}|M = \boldsymbol{m})P(M = \boldsymbol{m})}{P(X = \boldsymbol{x}|M = \boldsymbol{m})}. \tag{5.4}$$

As there are infinite number of $M = \boldsymbol{m}$, it is prohibitively to achieve the above backdoor adjustment. Fortunately, the Inverse Probability Weighting [149] formulation in Eq. (5.4) provides us a new perspective in approximating the infinite sampling $(i, \boldsymbol{x})|\boldsymbol{m}$. For a finite dataset, no matter how many $\boldsymbol{m}$ there are, we can only observe one $(i, \boldsymbol{x})$ given one $\boldsymbol{m}$. In such cases, the number of $\boldsymbol{m}$ values that Eq. (5.4) would encounter is equal to the number of samples $(i, \boldsymbol{x})$ available, not to the number of possible $\boldsymbol{m}$ values, which is prohibitive. In fact, thanks to the backdoor adjustment, which connects the equivalence between the originally confounded model P and the deconfounded model P with $do(X)$, we can collect samples from the former, that act as though they were drawn from the latter. Therefore, Eq. (5.4) can be approximated as

$$P(Y = i|do(X = \boldsymbol{x})) \approx \frac{1}{K}\sum_{k=1}^{K} \widetilde{P}(Y = i, X = \boldsymbol{x}^k|M = \boldsymbol{m}), \tag{5.5}$$

where $\widetilde{P}$ is the inverse weighted probability and we will drop $M = \boldsymbol{m}$ in the rest of the paper for notation simplicity and bear in mind that $\boldsymbol{x}$ still depends on $\boldsymbol{m}$. In particular, compared to the vanilla trick, we apply a multi-head strategy [218] to equally divide the channel (or dimensions) of weights and features into K groups, which can be considered as K times more fine-grained sampling.

We model $\widetilde{P}$ in Eq. (5.5) as the softmax activated probability of the energy-based model [219]:

$$\widetilde{P}(Y = i, X = \boldsymbol{x}^k) \propto E(i, \boldsymbol{x}^k; \boldsymbol{w}_i^k) = \tau \frac{f(i, \boldsymbol{x}^k; \boldsymbol{w}_i^k)}{g(i, \boldsymbol{x}^k; \boldsymbol{w}_i^k)}, \tag{5.6}$$

where τ is a positive scaling factor akin to the inverse temperature in Gibbs distribution. Recall Assumption 1 that $\boldsymbol{x}^k = \ddot{\boldsymbol{x}}^k + \boldsymbol{d}^k$. The numerator, *i.e.*, the unnormalized effect, can be implemented as logits $f(i, \boldsymbol{x}^k; \boldsymbol{w}_i^k) = (\boldsymbol{w}_i^k)^\top(\ddot{\boldsymbol{x}}^k + \boldsymbol{d}^k) = (\boldsymbol{w}_i^k)^\top \boldsymbol{x}^k$, and the denominator is a normalization term (or propensity score [220]) that only balances the magnitude of the variables: $g(i, \boldsymbol{x}^k; \boldsymbol{w}_i^k) = \|\boldsymbol{x}^k\| \cdot \|\boldsymbol{w}_i^k\| + \gamma\|\boldsymbol{x}^k\|$, where the first term is a class-specific energy and the second term is a class-agnostic baseline energy.

Putting the above all together, the logit calculation for $P(Y = i|do(X = \boldsymbol{x}))$ can be formulated as:

$$[Y = i|do(X = \boldsymbol{x})] = \frac{\tau}{K}\sum_{k=1}^{K}\frac{(\boldsymbol{w}_i^k)^\top(\ddot{\boldsymbol{x}}^k + \boldsymbol{d}^k)}{(\|\boldsymbol{w}_i^k\| + \gamma)\|\boldsymbol{x}^k\|} = \frac{\tau}{K}\sum_{k=1}^{K}\frac{(\boldsymbol{w}_i^k)^\top\boldsymbol{x}^k}{(\|\boldsymbol{w}_i^k\| + \gamma)\|\boldsymbol{x}^k\|}. \quad (5.7)$$

Interestingly, this model also explains the effectiveness of normalized classifiers like cosine classifier [20, 21]. We will further discuss it in Section 5.3.4.

5.3.2 Total Direct Effect Inference

After the de-confounded training, the causal graph is now ready for inference. The TDE of $X \to Y$ in Eq. (5.2) can thus be depicted as in Figure 5.3. By applying the counterfactual consistency rule [221], we have $[Y_{\boldsymbol{d}} = i|do(X = \boldsymbol{x})] = [Y = i|do(X = \boldsymbol{x})]$. This indicates that we can use Eq. (5.7) to calculate the first term of Eq. (5.2). Thanks to Assumption 1, we can disentangle $\boldsymbol{x}$ by $\boldsymbol{x} = \ddot{\boldsymbol{x}} + \boldsymbol{d}$, where $\boldsymbol{d} = \|\boldsymbol{d}\| \cdot \hat{\boldsymbol{d}} = cos(\boldsymbol{x}, \hat{\boldsymbol{d}})\|\boldsymbol{x}\| \cdot \hat{\boldsymbol{d}}$. Therefore, we have $[Y_{\boldsymbol{d}} = i|do(X = \boldsymbol{x}_0)]$ that replaces the $\ddot{\boldsymbol{x}}$ in Eq. (5.7) with zero vector, just like "cheating" the model with a null input but keeping everything else unchanged. Overall, the final TDE calculation for Eq. (5.2) is

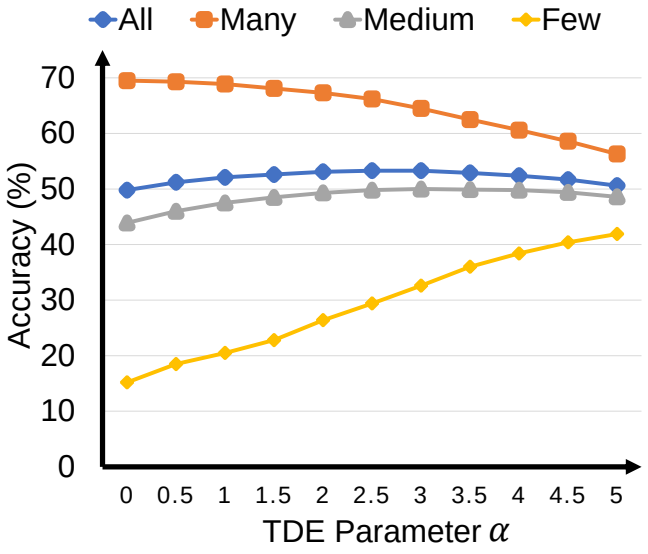
$$TDE(Y_i) = \frac{\tau}{K}\sum_{k=1}^{K}\left(\frac{(\boldsymbol{w}_i^k)^\top\boldsymbol{x}^k}{(\|\boldsymbol{w}_i^k\| + \gamma)\|\boldsymbol{x}^k\|} - \alpha \cdot \frac{cos(\boldsymbol{x}^k, \hat{\boldsymbol{d}}^k) \cdot (\boldsymbol{w}_i^k)^\top\hat{\boldsymbol{d}}^k}{\|\boldsymbol{w}_i^k\| + \gamma}\right), \quad (5.8)$$

where α controls the trade-off between the indirect and direct effect as shown in Figure 5.4.

In summary, the proposed De-confound-TDE can be intuitively interpreted as treating the long-tailed bias from the magnitudes and directions of feature vectors separately. The de-confound training approximates the backdoor adjustment by normalizing the magnitude bias. The TDE inference further adjusts the direction by removing the biased components of features. Hence, the overall De-confound-TDE can jointly increase the long-tailed robustness.

Methods	Two-stage	Re-balancing ($do(D)$)	De-confound ($do(X)$)	Direct Effect
Cosine [20, 21]	-	-	✔	-
LDAM [111]	-	✔	✔	CDE
OLTR [10]	✔	✔	-	NDE
BBN [16]	✔	✔	-	NDE
Decouple [11]	✔	✔	-	NDE
EQL [39]	-	✔	-	-
Our method	-	-	✔	TDE

TABLE 5.1: Revisiting the previous state-of-the-arts in our causal graph. CDE: Controlled Direct Effect. NDE: Natural Direct Effect. TDE: Total Direct Effect.



(a) Accuracy for different TDE parameter α

FIGURE 5.4: The influence of parameter α in Eq. (5.8) on ImageNet-LT val set [10] shows how D controls the head/tail preference.

5.3.3 Background-Exempted Inference

To make the De-confound-TDE more general, some classification tasks need a special "background" class to filter out samples belonging to none of the classes of interest, *e.g.*, object detection and instance segmentation use the background class to remove non-object regions [17, 27], and recommender systems assume that the majority of the items are irrelevant to a user [222]. In such tasks, most of the training samples are background and hence the background class is a good head class, whose effect should be kept and thus exempted from the TDE calculation. To this end, we propose a *background-exempted* inference that particular uses the original

inference (total effect) for background class. The inference can be formulated as:

$$\arg\max_{i \in C} \begin{cases} (1-p_0) \cdot \frac{q_i}{1-q_0} & i \neq 0 \\ p_0 & i = 0 \end{cases}, \tag{5.9}$$

where $i = 0$ is the background class, $p_i = P(Y = i|do(X = \boldsymbol{x}))$ is the de-confounded probability that we defined in Section 5.3.1, q_i is the softmax activated probability of the original $TDE(Y_i)$ in Eq. (5.8). Note that Eq. (5.9) adds up to 1 from $i = 0$ to C.

5.3.4 Revisiting Two-stage Training

What's more, the proposed framework also theoretically explains the previous state-of-the-art two-stage strategy as shown in Table 5.1. The detailed revisit for each methods will be discussed in the next section.

5.3.4.1 Two-stage Re-balancing

The Naïve re-balanced training fails to retain a natural mediation D that respects the inter-dependencies among classes, causing the bad feature extraction backbones that over-fit to the tail and under-fit to the head. Therefore, the two-stage training is adopted by most of the re-balancing methods: imbalanced pre-training the backbone with natural D and then balanced re-training a fair classifier with the fixed backbone for feature representation. Later, we will show that the second stage re-balancing essentially plays a counterfactual role, which reveals the reason why the stage-2 is indispensable.

5.3.4.2 De-confounded Training

Technically, the proposed de-confounded training in Eq. (5.7) is the multi-head classifier with normalization. The normalized classifier, like cosine classifier, has already been embraced by various methods [10, 11, 20, 21] based on empirical practice. However, as we will show in Table 5.2, without the guidance of our causal graph, their normalizations perform worse than the proposed de-confounded model.

Methods	Many-shot	Medium-shot	Few-shot	Overall
Focal Loss† [144]	64.3	37.1	8.2	43.7
OLTR† [10]	51.0	40.8	20.8	41.9
Decouple-OLTR† [10, 11]	59.9	45.8	27.6	48.7
Decouple-Joint [11]	65.9	37.5	7.7	44.4
Decouple-NCM [11]	56.6	45.3	28.1	47.3
Decouple-cRT [11]	61.8	46.2	27.4	49.6
Decouple-τ-norm [11]	59.1	46.9	30.7	49.4
Decouple-LWS [11]	60.2	47.2	30.3	49.9
Baseline	66.1	38.4	8.9	45.0
Cosine† [20, 21]	67.3	41.3	14.0	47.6
Capsule† [10, 22]	67.1	40.0	11.2	46.5
(Ours) De-confound	**67.9**	42.7	14.7	48.6
(Ours) Cosine-TDE	61.8	47.1	30.4	50.5
(Ours) Capsule-TDE	62.3	46.9	30.6	50.6
(Ours) De-confound-TDE	62.7	**48.8**	**31.6**	**51.8**

TABLE 5.2: The performances on ImageNet-LT test set [10]. All models were using the ResNeXt-50 backbone. The superscript † denotes being re-implemented by our framework and hyper-parameters.

For example, methods like decouple [11] only applies normalization in the 2nd stage balanced classifier training, and hence its feature learning is not de-confounded.

5.3.4.3 Direct Effect

The one-stage re-weighting/re-sampling training methods, like LDAM [111], can be interpreted as calculating Controlled Direct Effect (CDE) [149]: $CDE(Y_i) = [Y = i|do(X = \boldsymbol{x}), do(D = \boldsymbol{d}_0)] - [Y = i|do(X = \boldsymbol{x}_0), do(D = \boldsymbol{d}_0)]$, where $\boldsymbol{x}_0$ is a dummy vector and $\boldsymbol{d}_0$ is a constant vector. CDE performs a physical intervention — re-balancing — on the training data by setting the bias D to a constant. Note that the second term of CDE is a constant that does not affect the classification. However, CDE removes the "bad" at the cost of hurting the "good" during representation learning, as D is no longer a natural mediation generated by X. In other words, it results bad feature extraction backbones that over-fit to the tail and under-fit to the head.

The two-stage methods [11, 16] are essentially Natural Direct Effect (NDE), where the stage-2 re-balanced training is actually an intervention on D that forces the

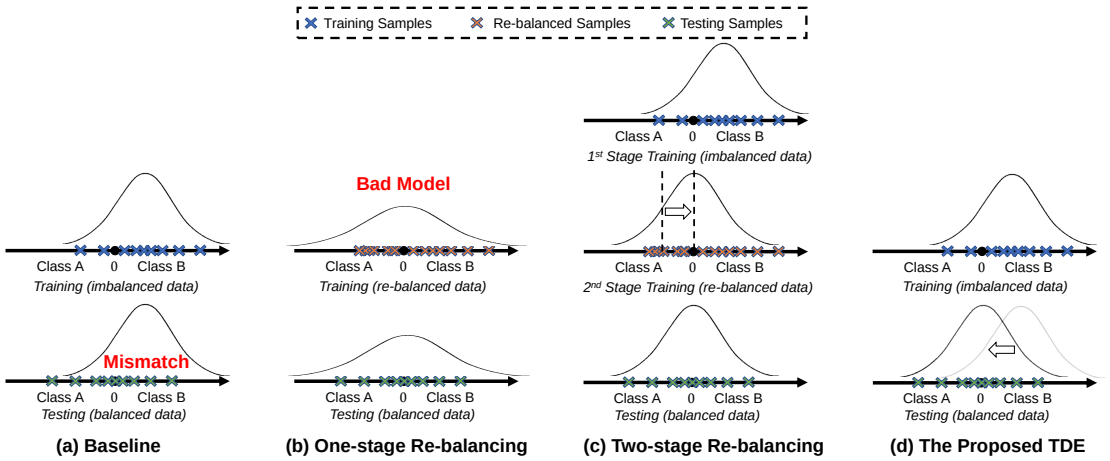


FIGURE 5.5: A simple one-dimensional binary classification example of conventional classifier, one-/two-stage re-balancing classifiers, and the proposed TDE.

direction $\hat{\boldsymbol{d}}$ do not head to any class. Therefore, when attached with the stage-1 imbalanced pre-trained features, the balanced classifier calculates the NDE: $NDE(Y_i) = [Y_{\boldsymbol{d}_0} = i|do(X = \boldsymbol{x})] - [Y_{\boldsymbol{d}_0} = i|do(X = \boldsymbol{x}_0)]$, where $\boldsymbol{x}_0$ and $\boldsymbol{d}_0$ are dummy vectors, because the stage-2 balanced classifier forces the logits to nullify any class-specific momentum direction; $do(X = \boldsymbol{x})$ as stage-1 backbone is frozen and $M \not\to X$; the second term can be omitted as it is a class-agnostic constant. Besides that their stage-1 training is still confounded, as we will show in experiments, our TDE is better than NDE because the latter completely removes the entire effect of D by setting $D = \boldsymbol{d}_0$, which is however sometimes good, *e.g.*, mis-classifying "warthog" as the head-class "pig" is better than "car"; TDE admits the effect by keeping $D = \boldsymbol{d}$ as a baseline and further compares the fine-grained difference via the direct effect, *e.g.*, by admitting that "warthog" does look like "pig", TDE finds out that the tusk is the key difference between "warthog" and "pig", and that is why our method can focus on more discriminative regions in Figure 5.7.

5.3.4.4 The Difference Between NDE and TDE

In this section, we will further discuss the relationship between two-stage re-balancing NDE and the proposed TDE. As we discussed in previous, the 2nd-stage re-balanced classifier essentially calculates the $NDE(Y_i) = [Y_{\boldsymbol{d}'} = i|do(X = \boldsymbol{x})] - [Y_{\boldsymbol{d}'} = i|do(X = \boldsymbol{x}')]$, where the second term can be omitted because $\boldsymbol{x}'$ is a dummy vector and the moving averaged $\boldsymbol{d}'$ in a balanced set won't point to any

specific classes, so it is actually a constant offset. Therefore, the crux of understanding the NDE would be why the 2nd-stage re-balanced training equals to the first term $[Y_{\boldsymbol{d}'} = i|do(X = \boldsymbol{x})]$. It is because when the backbone is frozen, it breaks the dependency between $M \rightarrow X$, which is a straightforward implementation of causal intervention $do(X = \boldsymbol{x})$. The original OLTR [10] violates this intervention by fine-tuning the backbone parameters in the 2nd stage, and it thus performs much worse than the Decouple-OLTR, which freezes the backbone parameters. Meanwhile, the balanced re-sampling also brings a fair $\boldsymbol{d}'$ as we discussed in the third paragraph of Section 5.2.1.

To better illustrate both the similarity and the difference between re-balancing NDE and the proposed TDE, we constructed a one-dimensional binary classification example for conventional classifier, one-/two-stage re-balancing classifiers, and the proposed TDE in Figure 5.5, where the gaussian distribution curve represents the feature distribution generated by the backbone, and the 0 point is the classifier's decision boundary. The conventional classifier and one-stage re-balancing are fundamentally problematic, because they either cause the mismatching in the inference or learn a bad backbone model. In the meantime, both two-stage re-balancing and the proposed TDE are able to correctly remove the bias by proper adjustments. The 2nd-stage re-balanced training (NDE) fixes the backbone parameters $do(X = \boldsymbol{x})$ learnt from 1st-stage imbalanced training, *i.e.*, the frozen curve in the image, and then re-samples an artificially balanced data distribution to create a fair $\boldsymbol{d}'$. The overall re-balancing NDE can be considered as subtracting a bias offset from original decision boundary. Meanwhile, the proposed TDE removes the bias effect (head projection) from feature vectors.

Both two types of adjustments can properly remove the head bias in this example. That's why TDE and NDE should be theoretically identical in the long-tailed classification scenario. However, the 2nd-stage re-balancing NDE has two disadvantages: 1) its adjustment requires an additional training stage to fine-tune the classifier weights, which relies on the accessibility of data distribution; 2) if non-linear modules are applied to the feature vectors, *e.g.*, a global context layer that conducts interactions among all objects $\{\boldsymbol{x}_j\}$ in an image, the NDE can only remove a linear approximation of this non-linear activated head bias, while the TDE would be able to maintain the natural interactions of features in both original logit term and the subtracted counterfactual term. It explains why the Decouple-OLTR

(a) Magnitude of w_i for each class i

FIGURE 5.6: The magnitudes of classifier weights $\|\boldsymbol{w}_i\|$ for each class after training with momentum $\mu = 0.9$, where i is ranking by the number of training samples in a descending order.

doesn't perform as good as Decouple-τ-norm or Decouple-LWS, because OLTR involves non-linear interactions between feature vectors and memory vectors, so a linear adjustment on classifier's decision boundary cannot completely remove the head bias.

5.3.5 Revisiting Other Strategies

In this section, we will discuss other previous strategies in long-tailed recognition: the normalized classifiers and the re-balancing strategies.

5.3.5.1 Normalized Classifiers

The normalized classifiers [10, 11, 20, 21] have already been widely adopted in long-tailed classification based on empirical practice. As we discussed in the Section 4, the correctly applied normalized classifiers are approximations of the proposed de-confounded training. However, without the guidance of the proposed causal framework, most of them are not utilized in a proper way. We define the general

normalized classifier as the following equation:

$$\arg\max_{i \in C} \; P(Y=i|X=\boldsymbol{x}) = \frac{e^{\boldsymbol{z}_i}}{\sum_{c=1}^{C} e^{\boldsymbol{z}_c}}, \quad \text{where} \quad z_i = \frac{\tau}{K} \sum_{k=1}^{K} \frac{(\boldsymbol{w}_i^k)^\top \boldsymbol{x}^k}{N(\boldsymbol{x}^k, \boldsymbol{w}_i^k)}. \tag{5.10}$$

Since in most of the previous methods, K is set to 1, so we slightly abuse the notation to omit the superscript k for simplicity.

The cosine classifier [20, 21] is defined based on the cosine similarity, which has $N(\boldsymbol{x}, \boldsymbol{w}_i) = \|\boldsymbol{x}\| \cdot \|\boldsymbol{w}_i\|$. It is commonly used in the tasks like few-shot learning [223]. In previous sections, we have proved its effectiveness in the long-tailed classification. The capsule classifier is proposed by Liu *et al.* [10] as the replacement of vanilla cosine classifier in OLTR. It changes the $l2$ norm of $\boldsymbol{x}$ into the squashing non-linear function proposed in Capsule Network [22], which allows the normalized $\boldsymbol{x}$ having a magnitude range from 0 to 1, representing the probability of $\boldsymbol{x}$ in its direction. The final normalization term can thus be defined as $N(\boldsymbol{x}, \boldsymbol{w}_i) = (\|\boldsymbol{x}\|+1) \cdot \|\boldsymbol{w}_i\|$. However, the OLTR [10] doesn't use it to de-confound the visual feature. Instead, its $\boldsymbol{x}$ is the joint embedding of the feature vector and an attentive memory vector. The Decouple [11] also invents two different types of normalized classifiers: τ-norm classifier and Learnable Weight Scaling (LWS) classifier. They empirically found that the $l2$ norm of $\boldsymbol{w}_i$ is not uniform in the long-tailed dataset, and has a positive correlation with the number of training samples for class i, as shown in Figure 5.6. Therefore, their normalized classifiers only normalize the $\boldsymbol{w}_i$: the τ-norm classifier is defined as $N(\boldsymbol{x}, \boldsymbol{w}_i) = \|\boldsymbol{w}_i\|^\tau, \tau \in [0, 1]$ while LWS is $N(\boldsymbol{x}, \boldsymbol{w}_i) = g_i$, where g_i is a learnable parameter. Yet, these decouple classifiers fail to de-confound the $M \rightarrow X$ for two reasons: 1) they don't considering the confounding effect on $\boldsymbol{x}$; 2) they only apply the normalized classifiers on the 2nd stage when the backbone has already been frozen.

5.3.5.2 Re-balancing Strategies

Both OLTR [10] and Decouple [11] adopt the same class-aware sampler in their 2nd stage training, which forces each class to contribute the same number of samples regardless of the size. To dynamically combine the two training stages, the BBN [16] utilizes a bilateral-branch design to smoothly transfer the sampling strategy from the imbalanced branch to the re-balancing branch, where two branches

share the same set of parameters but learn from different sampling strategies, which has the same spirit as two-stage design in OLTR [10] and Decouple [11]. As to the EQL [39], since the re-sampling is complicated in the object detection and instance segmentation tasks, where objects from different classes co-exist in one image, they choose the re-weighted loss to balance the contributions of different classes.

5.4 Experiments

The proposed method was evaluated on three long-tailed benchmarks: Long-tailed CIFAR-10/-100, ImageNet-LT for image classification and LVIS for object detection and instance segmentation. The consistent improvements across different tasks demonstrate our broad application domain.

5.4.1 Datasets and Protocols

We followed [16, 111] to collect the long-tailed versions of CIFAR-10/-100 with controllable degrees of data imbalance ratio ($\frac{N_{max}}{N_{min}}$, where N is number of samples in each category), which controls the distribution of training sets. ImageNet-LT [10] is a long-tailed subset of ImageNet dataset [69]. It consists of 1k classes over 186k images, where 116k/20k/50k for train/val/test sets, respectively. In train set, the number of images per class is ranged from 1,280 to 5, which imitates the long-tailed distribution that commonly exists in the real world. The test and val sets were balanced and reported on four splits: Many-shot containing classes with > 100 images, Medium-shot including classes with ≥ 20 $\& \leq 100$ images, Few-shot covering classes with < 20 images, and Overall for all classes. LVIS [19] is a large vocabulary instance segmentation dataset with 1,230/1,203 categories in V0.5/V1.0, respectively. It contains a 57k/100k train set (V0.5/V1.0) under a significant long-tailed distribution, and relatively balanced 5k/20k val set (V0.5/V1.0) and 20k test set.

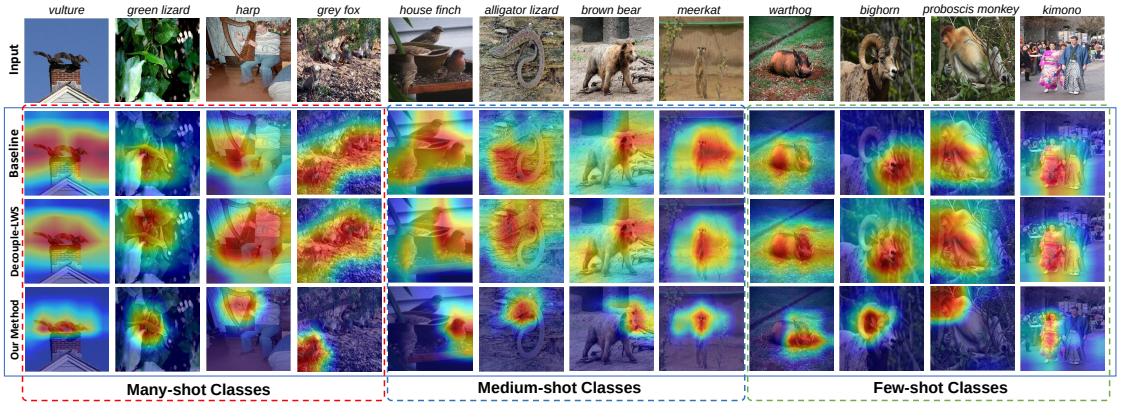


FIGURE 5.7: The visualized activation maps of the linear classifier baseline, Decouple-LWS [11] and the proposed method on ImageNet-LT using the Grad-CAM [12].

Dataset	Long-tailed CIFAR-100			Long-tailed CIFAR-10		
Imbalance ratio	100	50	10	100	50	10
Focal Loss [144]	38.4	44.3	55.8	70.4	76.7	86.7
Mixup [132]	39.5	45.0	58.0	73.1	77.8	87.1
Class-balanced Loss [170]	39.6	45.2	58.0	74.6	79.3	87.1
LDAM [111]	42.0	46.6	58.7	77.0	81.0	88.2
BBN [16]	42.6	47.0	59.1	79.8	82.2	88.3
(Ours) De-confound	40.5	46.2	58.9	71.7	77.8	86.8
(Ours) De-confound-TDE	**44.1**	**50.3**	**59.6**	**80.6**	**83.6**	**88.5**

TABLE 5.3: Top-1 accuracy on Long-tailed CIFAR-10/-100 with different imbalance ratios. All models are using the same ResNet-32 backbone. We further adopted the same warm-up scheduler from BBN [16] for fair comparisons.

5.4.2 Evaluation

For Long-tailed CIFAR-10/-100 [16, 111], we evaluated Top-1 accuracy under three different imbalance ratios: 100/50/10. For ImageNet-LT [10], the evaluation results were reported as the percentage of accuracy on four splits. For LVIS [19], the evaluation metrics are standard segmentation mask AP calculated across IoU threshold 0.5 to 0.95 for all classes. These classes can also be categorized by the frequency and independently reported as AP_r, AP_c, AP_f: subscripts r, c, f stand for rare (appeared in < 10 images), common (appeared in $11 - 100$ images), and frequent (appeared in > 100 images). Since we can use the LVIS to detect bounding boxes, the detection results were reported as AP_{bbox}.

5.4.3 Implementation Details

For image classification on ImageNet-LT, we used ResNeXt-50-32x4d [26] as our backbone for all experiments. All models were trained by using SGD optimizer with momentum $\mu = 0.9$ and batch size 512. The learning rate was decayed by a cosine scheduler [224] from 0.2 to 0.0 in 90 epochs. Hyper-parameters were chosen by the performances on ImageNet-LT val set, and we set $K = 2, \tau = 16, \gamma = 1/32, \alpha = 3.0$. For Long-tailed CIFAR-10/-100, we changed the backbone to ResNet-32 and the training scheduler to warm-up scheduler like BBN [16] for fair comparisons. All parameters except for α are inherited from ImageNet-LT, which was set to 1.0/1.5 for CIFAR-10/-100 respectively. For instance segmentation and object detection on LVIS, we chose Cascade Mask R-CNN framework [17] implemented by [225]. The optimizer was also SGD with momentum $\mu = 0.9$ and we used batch size 16 for a R101-FPN backbone. The models were trained in 20 epochs with learning rate starting at 0.02 and decaying by the factor of 0.1 at the 16-th and 19-th epochs. We selected the top 300 predicted boxes following [19, 39]. The hyper-parameters on LVIS were directly adopted from the ImageNet-LT, except for $\alpha = 1.5$. The main difference between image classification and object detection/instance segmentation is that the latter includes a background class $i = 0$, which is a head class used to make a binary decision between foreground and background. As we discussed in Section. 5.3.3, the Background-Exempted Inference should be used to retain the good background bias. The comparison between with and without Background-Exempted Inference is given in the following section.

5.4.4 Ablation studies

To study the effectiveness of the proposed de-confounded training and TDE inference, we tested a variety of ablation models: 1) the linear classifier baseline (no biased term); 2) the cosine classifier [20, 21]; 3) the capsule classifier [10], where x is normalized by the non-linear function from [22]; 4) the proposed de-confounded model with normal softmax inference; 5) different versions of the TDE. As reported in Table (5.2,5.4), the de-confound TDE achieves the best performance under all settings. The TDE inference improves all three normalized models, because the cosine and capsule classifiers can be considered as approximations to the proposed

Methods	LVIS Version	AP	AP_{50}	AP_{75}	AP_r	AP_c	AP_f	AP_{bbox}
Focal Loss† [144]	V0.5	21.1	32.1	22.6	3.2	21.1	28.3	22.6
(2019 Winner) EQL [39]	V0.5	24.9	37.9	26.7	10.3	27.3	27.8	27.9
Baseline	V0.5	22.6	33.5	24.4	2.5	23.0	30.2	24.3
Cosine† [20, 21]	V0.5	25.0	37.7	27.0	9.3	25.5	30.8	27.1
Capsule† [10, 22]	V0.5	25.4	37.8	27.4	8.5	26.4	**31.0**	27.1
(Ours) De-confound	V0.5	25.7	38.5	27.8	11.4	26.1	30.9	27.7
(Ours) Cosine-TDE	V0.5	28.1	42.6	30.2	20.8	28.7	30.3	30.6
(Ours) Capsule-TDE	V0.5	**28.4**	42.1	**30.8**	21.1	**29.7**	29.6	30.4
(Ours) De-confound-TDE	V0.5	**28.4**	**43.0**	30.6	**22.1**	29.0	30.3	**31.0**
Baseline	V1.0	21.8	32.7	23.2	1.1	20.9	31.9	23.9
(Ours) De-confound	V1.0	23.5	34.8	25.0	5.2	22.7	**32.3**	25.8
(Ours) De-confound-TDE	V1.0	**27.1**	**40.1**	**28.7**	**16.0**	**26.9**	32.1	**30.0**

TABLE 5.4: All models are using the same Cascade Mask R-CNN framework [17] with R101-FPN backbone [18]. The reported results are evaluated on LVIS val set [19].

Methods	BG-Exempted	AP	AP_{50}	AP_{75}	AP_r	AP_c	AP_f	AP_{bbox}
De-confound	✗	25.7	38.5	27.8	11.4	26.1	**30.9**	27.7
De-confound-TDE	False	23.4	35.7	24.9	13.1	23.6	27.1	24.8
De-confound-TDE	True	**28.4**	**43.0**	**30.6**	**22.1**	**29.0**	30.3	**31.0**

TABLE 5.5: The results of the proposed TDE with/without Background-Exempted Inference on LVIS [19] V0.5 val set. The Cascade Mask R-CNN framework [17] with R101-FPN backbone [18] is used.

de-confounded model. To show that the mediation effect removed by TDE indeed controls the preference towards head direction, we changed the parameter α as shown in Figure 5.4, resulting the smooth increasing/decreasing of the performances on tail/head classes, respectively.

5.4.4.1 Background-Exempted Inference

The results with and without Background-Exempted Inference are reported in Table 5.5. As we can see, the Background-Exempted strategy successfully prevents the TDE from hurting the foreground-background selection. It is the key to apply TDE in tasks like object detection and instance segmentation that include one or more legitimately biased head categories, *i.e.*, this strategy allows us to conduct TDE on a selected subset of categories.

5.4.4.2 Selection of Hyper-Parameters

The hyper-parameters used in this chapter are selected according to the performances on ImageNet-LT val set as shown in Table 5.6. To further study the multi-head strategy on different normalized classifiers, we tested the $K = 2$ on cosine classifier [20, 21] and capsule classifier [10, 22] in Table 5.7. It proves that the advantage of the proposed de-confounded model doesn't come from larger K, and the multi-head fine-grained sampling can generally improves the de-confounded training, no matter what kind of normalization function we choose.

5.4.4.3 Evaluation on Different Backbones

As shown in Table 5.8,5.9, we tested the proposed method on different backbones. After equipped with ResNeXt-101-32x4d and ResNeXt-101-64x4d [26] for ImageNet-LT [10] and LVIS [19] V0.5, respectively, the proposed method gains additional improvements. In ImageNet-LT dataset, we changed some hyper-parameters ($K = 4, \gamma = 1/64.0$) and increased the training epochs to 120, because of the significantly increased number of model parameters. The hyper-parameters for LVIS are still the same as the previous.

5.4.4.4 LVIS Performance on Test Server

We also reported the performances of the proposed method on LVIS V0.5 evaluation test server [23] in Table 5.10, where we used ResNeXt-101-64x4d backbone and the original hyper-parameters. It's worth noting that these are single model performances, which neither exploited external dataset nor utilized any model enhancement tricks.

5.4.5 Comparisons with State-of-The-Art Methods

The previous state-of-the-art results on ImageNet-LT are achieved by the two-stage re-balanced training [11] that decouples the backbone and classifier. However, as we discussed in Section 5.3.4, this kind of approaches are less effective or efficient. In Long-tailed CIFAR-10/-100, we outperform the previous methods [16, 111, 170]

K	τ	γ	α	Many-shot	Medium-shot	Few-shot	Overall
1	16.0	1/32.0	✗	69.8	42.8	14.9	49.4
4	16.0	1/32.0	✗	69.0	42.3	13.1	48.6
2	**8.0**	1/32.0	✗	69.5	31.3	1.6	42.0
2	**32.0**	1/32.0	✗	68.6	41.3	13.0	47.9
2	16.0	**1/16.0**	✗	69.3	**44.0**	14.2	49.7
2	16.0	**1/64.0**	✗	**69.9**	43.3	14.7	49.6
2	**16.0**	**1/32.0**	✗	69.5	43.9	**15.2**	**49.8**
2	16.0	1/32.0	**2.5**	**66.2**	49.8	29.4	**53.3**
2	16.0	1/32.0	**3.0**	64.5	**50.0**	32.6	**53.3**
2	16.0	1/32.0	**3.5**	62.5	49.9	**36.0**	52.9

TABLE 5.6: Hyper-parameters selection based on performances of ImageNet-LT val set, where ✗ for α means that TDE inference is not included. The backbone we used here is ResNeXt-50-32x4d.

in all imbalance ratios, which proves that the proposed method can automatically adapt to different data distributions. In LVIS dataset, after a simple adaptation, we beat the champion EQL [39] of LVIS Challenge 2019 in Table 5.4. All reported results in Table 5.4 are using the same Cascade Mask R-CNN framework [17] and R101-FPN backbone [18] for fair comparison. The EQL results were copied from [39], which were trained by 16 GPUs and 32 batch size while the proposed method only used 8 GPUs and half of the batch size. We didn't compare the EQL results on the final challenge test server, because they claimed to exploit external dataset and other tricks like ensemble to win the challenge. Note that EQL is also a re-balanced method, having the same problems as [11]. We also visualized the activation maps using Grad-CAM [12] in Figure 5.7. The linear classifier baseline and decouple-LWS [11] usually activate the entire objects and some context regions to make a prediction. Meanwhile, the de-confound TDE only focuses on the direct effect, *i.e.*, the most discriminative regions, so it usually activates on a more compact area, which is less likely to be biased towards its similar head classes. For example, to classify a "kimono", the proposed method only focuses on the discriminative feature rather than the entire body, which is similar to some other clothes like "dress".

Methods	#heads K	Many-shot	Medium-shot	Few-shot	Overall
Cosine† [20, 21]	1	67.3	41.3	14.0	47.6
Cosine† [20, 21]	2	67.5	42.1	14.1	48.1
Capsule† [10, 22]	1	67.1	40.0	11.2	46.5
Capsule† [10, 22]	2	67.7	41.3	12.6	47.6
(Ours) De-confound	1	67.3	41.8	15.0	47.9
(Ours) De-confound	2	**67.9**	42.7	14.7	48.6
(Ours) Cosine-TDE	1	61.8	47.1	30.4	50.5
(Ours) Cosine-TDE	2	63.0	47.3	31.0	51.1
(Ours) Capsule-TDE	1	62.3	46.9	30.6	50.6
(Ours) Capsule-TDE	2	62.4	47.9	31.5	51.2
(Ours) De-confound-TDE	1	62.5	47.8	**32.8**	51.4
(Ours) De-confound-TDE	2	62.7	**48.8**	31.6	**51.8**

TABLE 5.7: The performances of cosine classifier [20, 21] and capsule classifier [10, 22] under different number of head K on ImageNet-LT test set. Other hyper-parameters are fixed.

Methods	Backbone	Many-shot	Medium-shot	Few-shot	Overall
Baseline	ResNeXt-50	66.1	38.4	8.9	45.0
De-confound	ResNeXt-50	67.9	42.7	14.7	48.6
De-confound-TDE	ResNeXt-50	62.7	48.8	31.6	51.8
Baseline	ResNeXt-101	68.7	42.5	11.8	48.4
De-confound	ResNeXt-101	**68.9**	44.3	16.5	50.0
De-confound-TDE	ResNeXt-101	64.7	**50.0**	**33.0**	**53.3**

TABLE 5.8: The performances of the proposed method under different backbones in ImageNet-LT test set.

Methods	Backbone	AP	AP_{50}	AP_{75}	AP_r	AP_c	AP_f	AP_{bbox}
Baseline	R101-FPN	22.6	33.5	24.4	2.5	23.0	30.2	24.3
De-confound	R101-FPN	25.7	38.5	27.8	11.4	26.1	30.9	27.7
De-confound-TDE	R101-FPN	28.4	43.0	30.6	**22.1**	29.0	30.3	31.0
Baseline	X101-FPN	26.4	39.5	28.4	7.4	28.1	32.0	28.5
De-confound	X101-FPN	28.4	41.9	30.6	13.3	29.5	**32.9**	30.5
De-confound-TDE	X101-FPN	**30.4**	**45.1**	**32.9**	21.1	**31.8**	32.3	**33.1**

TABLE 5.9: The performances of the proposed method under different backbones in LVIS V0.5 val set.

Methods	AP	AP_{50}	AP_{75}	AP_r	AP_c	AP_f
Baseline	19.4	29.8	20.6	3.9	21.9	30.8
De-confound	20.8	31.8	22.1	7.4	22.7	**31.2**
De-confound-TDE	**23.0**	**35.2**	**24.1**	**12.7**	**24.5**	30.7

TABLE 5.10: The single model performances of the proposed method on LVIS V0.5 evaluation test server [23].

5.5 Conclusions

In this chapter, we proposed a causal framework called De-confound-TDE to obtain the general long-tailed robustness in "low-level" computer vision tasks, which fills the gap of multimodal TDE in Chapter 4. The proposed De-confound-TDE pinpoints the causal effect of momentum in the long-tailed classification, which not only theoretically explains the previous methods, but also provides an elegant one-stage training solution to extract the unbiased direct effect of each instance. The detailed implementation consists of de-confounded training and total direct effect inference, which is simple, adaptive, and agnostic to the prior statistics of the class distribution. We achieved the new stage-of-the-arts of various tasks on both ImageNet-LT and LVIS benchmarks. Combining the Chapter 4 with the Chapter 5, we are able to systematically tackle the data bias and obtain the long-tailed robustness in most of the computer vision tasks.

Chapter 6

Adversarial Visual Robustness by Causal Intervention[1]

6.1 Introduction

In the previous chapters, we investigate the long-tailed robustness, which can be viewed as data bias at the category level. However, not all the distributions of features can be explicitly revealed in DNN models. There are a large portion of the hidden features that may contain implicit bias at the pattern level. Such a bias is commonly known as the adversarial vulnerability that hurts the robustness of DNNs even worse than long-tailed bias.

Despite the remarkable progress achieved by Deep Neural Networks (DNNs), adversarial vulnerability [41] keeps haunting the computer vision community since it has been spotted by [47]. Over the years, we have witnessed many defenders, who claim to be "well-rounded", were soon found to lack fair benchmarking, *e.g.*, adaptive adversary [226, 227], or misconduct the attack, *e.g.*, obfuscated gradient [133]. Therefore, the most promising defender remains to be the intuitive Adversarial Training and its variants [76, 228]. Due to the "training" nature, its adversarial robustness is largely dependent on the knowledge of attackers and whether the training set contains sufficient adversarial samples from various attackers as many

[1]The work in this chapter is modified from the paper that is under review by July 13 2021: Kaihua Tang, Mingyuan Tao, Xian-Sheng Hua, Hanwang Zhang. "Adversarial Visual Robustness by Causal Intervention." **arXiv preprint** (**Under Review**). 2021.

as possible [78], yet, brute-forcely enumerating all attackers is prohibitively expensive, making adversarial training mainly over-fitting to known attackers [79]. What's worse, in few-/zero-shot scenarios, it is even impossible to collect enough adversarial training samples based on the out-of-distribution/unseen samples [80].

In other words, adversarial training is a "passive immunization", which cannot react to the ever-evolving attacks responsively. To proactively achieve adversarial robustness, we have to find the "origin" of adversarial perturbations. Previous methods blame adversarial vulnerability on the inherent flaws in fitting models to the limited high-dimensional data [41, 229, 230]. However, simply regarding adversarial samples as "bugs" cannot explain their well-generalizing behaviors [75, 231]. Recent studies [75, 81] show that adversarial examples are not "bugs" but *predictive* features that can only be exploited by machines. Such results urge us to investigate the essential difference between machines and humans.

However, we believe that it is too early for us to shirk responsibility and leave it to the ever-elusive open problem before we answer the following two key questions:

Q1: *What are the non-robust but predictive features?* [75] use adversarial examples to distinguish the robust and non-robust features. However, this will only allow us to recognize the non-robust features as "adversarial perturbations" again, which is, unfortunately, circular reasoning. Therefore, we need a fundamental yet different angle to define the robustness of features beyond the conventional adversarial one.

Q2: *Why do complex systems (human vision) ignore these predictive features that simple systems (DNNs) can capture instead?* Given the fact that biological visions are more complex than machines in terms of both neuron amount [83] and diversity [84, 85], there is no reason for human vision to extract "less feature" than machines. Therefore, there must be a mechanism in human vision that deliberately ignores these features.

In this thesis, we answer the above two questions from a causal perspective [87]—a powerful lens seeing through the generative nature of adversarial attacks. For **Q1**, we postulate that non-robust features are confounding effects, which are spurious correlations established by related but non-causal features. Take Figure 6.1 (a) as an intuitive example, where a large number of vertical edges co-occur with the digit "1". As a result, a model trained by associating samples with labels will recklessly

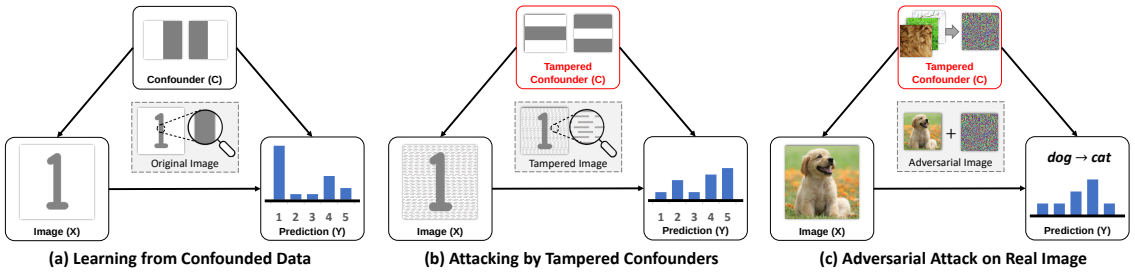


FIGURE 6.1: (a) A digit classifier confounded by counting edges. (b) Attacking the model through tampered confounders. (c) Constructing adversarial perturbations through an ensemble of tampered confounders, *e.g.*, local textures, small edges, and faint shadows.

use the counting of vertical edges—the confounding effect—as the indicator of digit "1" without learning the overall causal structure. Therefore, once tampered edges are constructed, which is much easier than editing the entire digit directly, the confounding effect will mislead the model prediction as shown in Figure 6.1 (b).

In general, any pattern co-occurred with certain labels can constitute confounders. Most of them are even imperceptible, like local textures, small edges, and faint shadows. Since DNN models are based on the statistical association between input and output, they inevitably learn these spurious correlations, which are "predictive" when the distribution of confounders remains the same in training and testing. However, their brittle nature makes them vulnerable to small perturbations as shown in Figure 6.1 (c). In Section 6.3, we will provide a formal revisit for the adversarial attack in the causal viewpoint, where we also design a Confounded-Toy dataset to demonstrate how an adversarial attacker fools the model by exploiting the confounding effect.

Unlike machine vision that scans all the pixels in an image at once, human vision continuously perceives the image using "retinotopic sampling" [86] via non-uniformly distributed retinal photoreceptors at each time frame as shown in Figure 6.2 (a). We conjecture that such a mechanism is the answer to **Q2**, because it can be viewed as causal intervention by using instrumental variable [232], denoted as R in Figure 6.2 (b). With the help of R, the confounded image observation X is no longer dictated only by the confounders. Since the choice of R is designed to be independent of C, as it only depends on the structure of retina, its direct effect on Y can thus be used to mitigate the confounding effect even though C is unobserved. Intuitively, non-robust confounder patterns are local impulses that won't perform

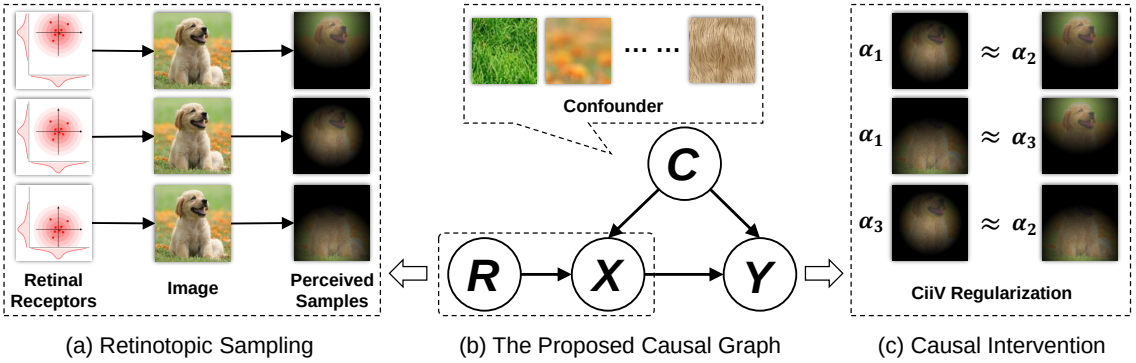


FIGURE 6.2: The proposed CiiV framework (detailed in Section 6.6): (a) the retinotopic augmentation that serves as the instrumental variable; (b) the proposed causal graph; (c) the causal intervention made by the proposed regularization that suppresses non-robust confounding effects.

consistently across different retinotopic centers. They are either captured or not by a retinotopic observation. Meanwhile, causal features are consistent structures. Forcing a model to learn features that linearly vary with the change of R can suppress unstable confounding effects. To this end, in Section 6.6, we propose the Causal intervention by instrumental Variable (CiiV \si:v\) framework that combines a spatial data augmentation through retinotopic sampling with a consistency regularization loss as shown in Figure 6.2 (c).

Our key contributions are as follows:

- We introduce a causal regularization termed CiiV to suppress the learning of non-robust features in DNN models, which not only offers a proactive defender, but also opens a novel yet fundamental viewpoint of adversary research.
- Extensive experiments on a wide range of settings from the adversarial evaluation checklist [40] in CIFAR-10, CIFAR-100, and mini-ImageNet demonstrate that CiiV can withstand adaptive attacks, including the state-of-the-art AutoAttack [226].
- As a general regularization that is orthogonal to most of the previous defenders, the proposed CiiV can be easily plugged into other methods to further boost their adversarial robustness.

6.2 Related Work

Adversarial Examples. Adversarial examples undermine the reliability and interpretability of DNN models in various domains [112–119] and settings [120–125]. Despite of various defenders proposed to improve the adversarial robustness, a universal remedy that can proactively defend against all the known and unknown attackers is still absent. Generally, the existing defenders fall into the following four categories: adversarial training [47, 48, 129], data augmentation [132], denoising [51, 77], and certified defense [130]. In Section 6.5 we will systematically revisit them and compare them to the proposed CiiV from a causal viewpoint.

Causality in Adversarial Robustness. Recently, causality has gradually been accepted as a potential way to explain adversarial robustness. [154, 233] provide a causal perspective to understand the adversarial vulnerability of DNN models; [234] utilize the supervised pixel-wise masking to conduct causal intervention; [235] attempt to unify the adversarial robustness with the distributional shift. However, the solutions they provided are either subject to additional supervisions, complicated causal graph and training strategies, or parallel to the existing adversarial training variants. Meanwhile, this thesis provides a more feasible causal explanation for the adversarial vulnerability, by which we can design an effective plug-and-play causal regularization.

Causal Graph and Intervention. Pearl's graphical model [149] is adopted in this thesis, where directed edges indicate the causality between node variables. The causal graph of the proposed CiiV framework is illustrated in Figure 6.2 (b), where R, C, X, Y indicate retinotopic sampling mask, confounding pattern, image, and prediction, respectively. $X \leftarrow C \rightarrow Y$ denotes that confounder C is a common cause, affecting the distribution of both X and Y, *e.g.*, the edge in Figure 6.1 (a). $X \rightarrow Y$ denotes the desired causality that a robust model is expected to learn. To achieve that, the ultimate goal of causal intervention is to identify the causal effect of $X \rightarrow Y$ by removing all spurious correlations [42], denoted as $P(Y|do(X = x))$. It can be either implemented as active intervention, like the randomized controlled trial, or passive d-separation [149, 236], by which observing the confounder can block the spurious path, *e.g.*, by conditioning on C, the dependency of path $X \leftarrow C \rightarrow Y$ is blocked.

6.3 A Causal View on Adversarial Attack

In causality [149], the total effect and causal effect of a predictive model based on the input X can be defined as $P(Y|X)$, $P(Y|do(X = x))$, respectively. Given the proposed causal graph in Figure 6.1, the confounding path $X \leftarrow C \rightarrow Y$ causes the inequality between the above two, and thus the confounding effect can be represented as their difference.

Meanwhile, the general adversarial attack can be formulated as maximizing the probability of a tampered category $Y = \hat{y}$ within the budget $\mathcal{D}_\epsilon$ [237], denoted as follows:

$$\max_{\delta \in \mathcal{D}_\epsilon} P(Y = \hat{y}|X = x + \delta) \propto \sum_i \hat{y}_i log(p_i), \tag{6.1}$$

where $\hat{y} = y'(y' \neq y)$ for targeted attack, $\hat{y} = -y$ for untargeted attack; $\hat{y}_i$ and p_i are i-th entries of $\hat{y}$ and prediction p, respectively; δ is the additive perturbation; budget $\mathcal{D}_\epsilon$ is usually considered as an enclosing ball under l_2/l_∞ norm within radius ϵ [122, 125].

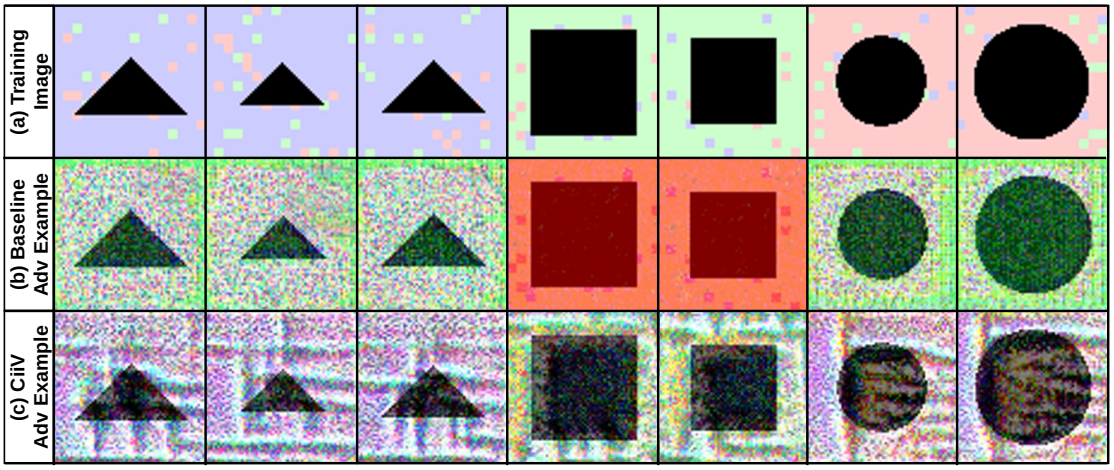


FIGURE 6.3: (a) A Confounded-Toy Dataset with images that are composed of causal geometries and confounding color blocks. The adversarial examples generated by the model (c) w/ and (b) w/o the proposed CiiV.

Notably, a valid $\mathcal{D}_\epsilon$ is not allowed to change the semantic structures, as they are designed to be imperceptible, *i.e.*, the causal effect $P(Y|do(X = x))$ is invariant to δ. Otherwise, the perturbation would become a "poisoning" that is beyond our scope [238]. Therefore, (6.1) essentially equals to maximize a tampered confounding effect through perturbations: $\max_\delta \ P(Y = \hat{y}|X = x+\delta) - P(Y = \hat{y}|do(X = x+\delta))$,

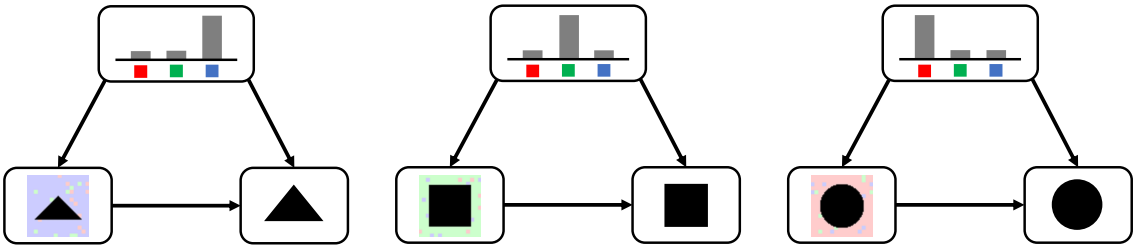
subject to $P(Y = \hat{y}|do(X = x + \delta)) = P(Y = \hat{y}|do(X = x))$, which applies to all kinds of attacks [41, 122, 239–241].

To intuitively demonstrate the above causal theories, we design a Confounded-Toy dataset as shown in Figure 6.3 (a), where images are composed of causal geometries and confounding color blocks. Similar to our example in Figure 6.1, a model directly trained on this dataset will recklessly learn the stochastic color block C that shows statistical correlation with the category as the indicator of Y. As a result, adversarial examples generated by a PGD attacker on this model mainly tamper the confounding patterns (Figure 6.3 (b)). In contrast, the proposed CiiV regularization forces the model to learn causal features instead, so it can only be fooled by poisoning the geometry (Figure 6.3 (c)).

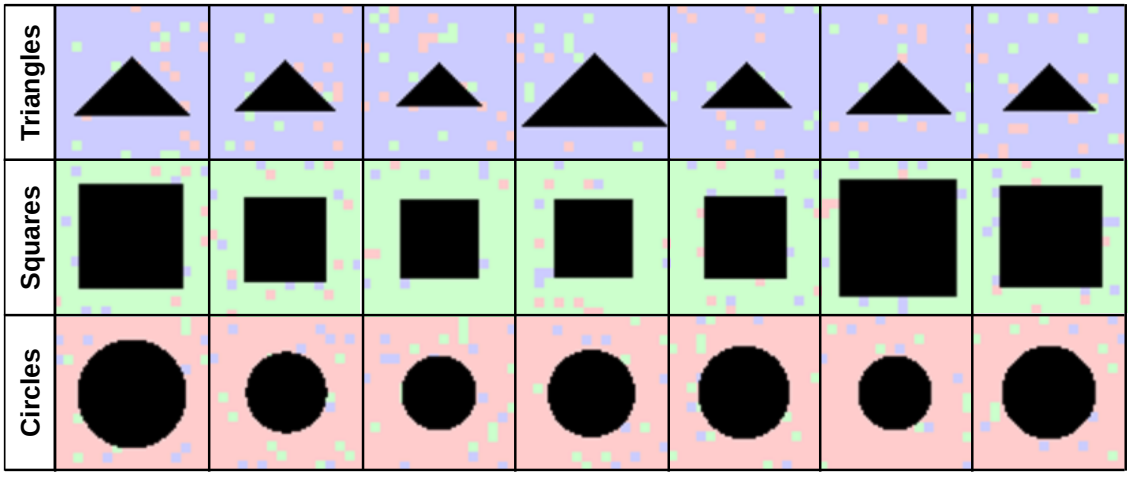
6.4 Details of the Confounded-Toy Dataset

In section 6.3, we introduced a Confounded Toy (CToy) dataset to demonstrate the equivalence between the confounding effect and the adversarial perturbation. The proposed CToy is a three-way classification, containing triangles, squares, and circles. It has 10k/1k/1k images for train/val/test split, respectively. All samples are 64x64 colour images. Except for the causal geometries, there are also confounding patterns, *i.e.*, red/blue/green blocks, with the size of 4x4 pixels. Different from the deterministic geometry, the color of each block is sampled from a biased distribution. For triangle, square, and circle images, each co-occurred block has 80%/10%/10%, 10%/80%/10%, and 10%/10%/80% probability to be blue/green/red, respectively. Therefore, if the confounding distribution stays the same in both training and testing phases, these patterns are indeed "predictive" features. Yet, learning these confounding patterns would significantly reduce the generalization ability of the model, because there will always be samples that are dominated by rare color blocks, and they are also more brittle than geometry structures. The causal graph of the data generation procedure and more examples of CToy dataset are illustrated in Figure 6.4 (a,b).

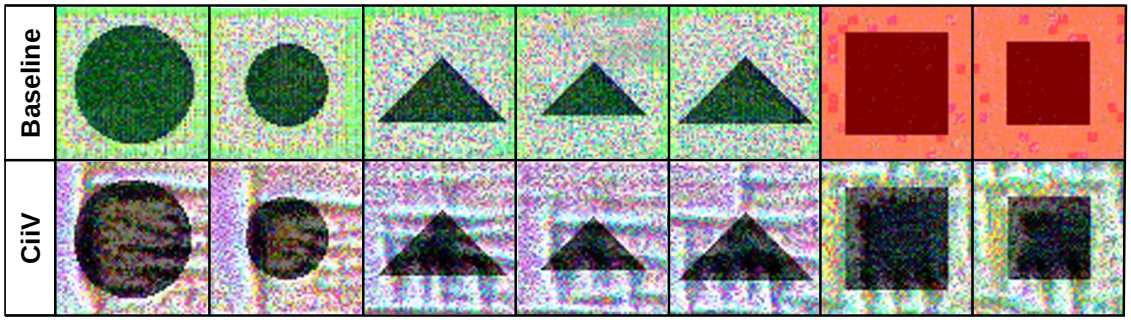
Based on the specifically designed CToy that only contains two patterns, causal shapes and confounding colors, we are able to understand which pattern causes the adversarial vulnerability. As we can see from Figure 6.4 (c), adversarial examples



(a) The Causal Graph of the Confounded-Toy Dataset



(b) Examples of the Confounded-Toy Dataset



(c) More Adversarial Examples on the Confounded-Toy Dataset

FIGURE 6.4: (a) The causal graph of the Confounded-Toy dataset. (b) More examples of the proposed Confounded-Toy dataset. (c) More adversarial examples from the baseline model and CiiV counterpart.

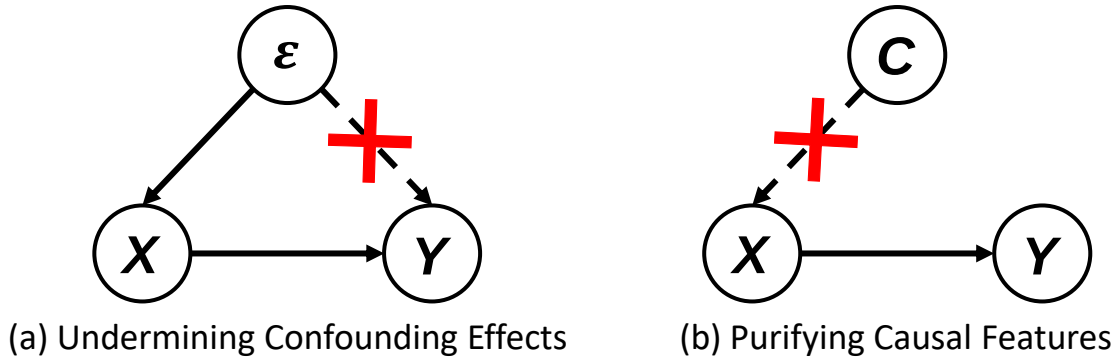


FIGURE 6.5: Two common strategies to increase the adversarial robustness.

of an L_∞ PGD attack (ϵ is set to 128/255 for 100% attacking success rate, so we can understand which pattern can successfully fool the model) that generated from a baseline DNN model were mainly erasing the original color blocks, *i.e.*, the adversarial perturbation is indeed trying to maximize the tampered confounding effect. Specifically, the attacker changed the blue and red blocks in triangle and circle images to the green points. It even painted the entire square images to red. The confounding patterns were obviously tampered in these images while the causal geometries barely changed. On the other hand, the adversarial examples of the proposed CiiV model didn't change the overall colors too much, they directly modified the shapes. It proves that CiiV successfully prevents the model from learning confounding effects, and thus attacker can only poison the causal geometries.

With the help of the CToy dataset, we are not only able to verify the proposed confounding theories for adversarial examples but also visualize the working mechanisms of the proposed CiiV framework, *i.e.*, forcing the model to learn from causal patterns rather than the confounding colors.

6.5 A Causal View on Adversarial Defense

It has been acknowledged that directly adjusting an unknown confounder C for $P(Y|do(X = x))$ without any assumption is impractical in causality field [82]. Due to the fact that adversarial examples are governed by unobserved confounding effects, most of the existing defending methods have to either intuitively assume a generative noise ε to be the underline C or assume C to be certain identifiable noisy features that can be explicitly purified.

Specifically, adversarial training and its variants [47, 48, 129] together with some certified defenders like randomized smoothing [130] design some additive noises ε to imitate adversarial perturbations, then undermine the confounding effect by asking the model to be robust against ε. On the other hand, the de-noising approaches, no matter the pre-network de-noising [49–51] or the in-network de-noising [77, 242] consider confounders to be explicitly removable patterns. Therefore, these common strategies can be summarized by two graphical operations as shown in Figure 6.5.

However, we can neither guarantee the C to be equal to ε, nor ensure all possible C to be disentangled and purified. Relying on the observation of such assumptive C will at best make the above defenders robust against a subset of potential confounders.

Among all existing defenders, mixup [132] is most related to the proposed CiiV. It intervenes an image x_i by linearly fusing with another image x_j, then forces the prediction Y similar to the same combination of their one-hot labels. Yet, a valid instrumental variable is required to be independent of the confounder as we will introduce in the next section. Unfortunately, a new image x_j can still depend on the same confounder of x_i. Recent studies [243, 244] found that universal adversarial perturbations across images also exist, which explains why mixup cannot survive strong attackers.

6.6 Approach

After connecting the adversarial vulnerability to the confounding effect learned by DNN models, the remaining question is how to obtain the pure causal effect, which is equivalent to applying causal intervention $P(Y|do(X = x))$ on the deep learning. Generally, there are four major interventions: randomized controlled trial, backdoor adjustment, front-door adjustment, and instrumental variable estimation. However, the randomized controlled trial requires the control over causal features, the backdoor and front-door adjustments assume confounders or mediators to be observed, which are impractical for imperceptible adversarial perturbations. Therefore, we are interested in the last instrumental variable estimation that does not require such assumptions.

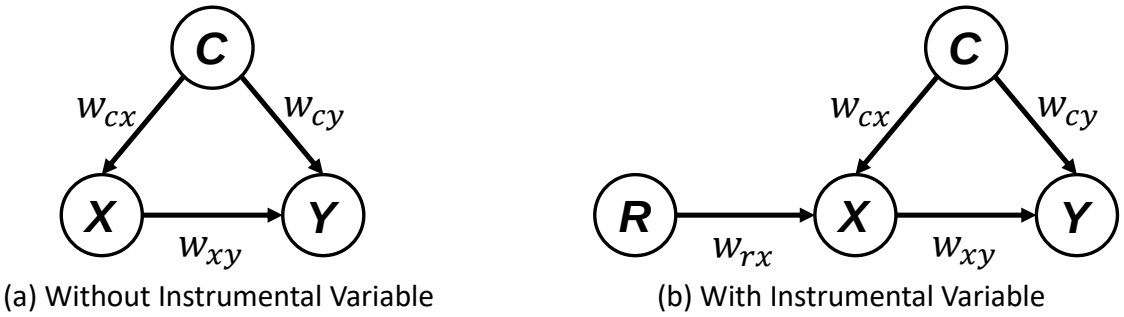


FIGURE 6.6: The causal graphs w/ and w/o the instrumental variable. Nodes are assumed to be linked through linear associations w_*.

6.6.1 Instrumental Variable Estimation

According to the definition [245, 246], a valid instrumental variable should satisfy: 1) it is independent of the confounder variable; 2) it affects Y only through X. The instrumental variable can help to extract the causal effect of $X \to Y$ from $R \to X \to Y$, which is not confounded by C (d-separated).

To better demonstrate the use of the instrumental variable [247], we design two linear confounded models w/ and w/o the instrumental variable as shown in Figure 6.6. All variables are assumed to be linked by linear weights w_*. The confounder is an independent variable sampled from a normal distribution: $C \sim \mathcal{N}(0, 1)$. The total effect and causal effect of $X \to Y$ can be represented as $Y[X = x] = w_{xy}x + w_{cy}c$ and $Y[do(X = x)] = w_{xy}x$, respectively. Note that we slightly abuse the notation of normalized effects $P(Y|X)$ and $P(Y|do(X = x))$, and use the form of unnormalized logits for simplicity.

In the given confounded model of Figure 6.6 (a), since X is dependent on the confounder C as $x = w_{cx}c + b_x$, where b_x is the independent component of X, we cannot directly estimate the causal effect $Y[do(X = x)]$ by simply applying linear regression on (x, y) pairs.

If C is observable, the causal intervention could be conducted using the backdoor adjustment: $P(y|do(x)) = \sum_c P(y|x, c)P(c)$. The causal effect is thus estimated from the total effect by the observed c and its distribution:

$$Y[do(X = x)] = w_{xy}x + w_{cy}\sum_c c \cdot P(c) = w_{xy}x, \tag{6.2}$$

where the confounding effect degrades to a constant as $\sum_c c \cdot P(c) = 0$.

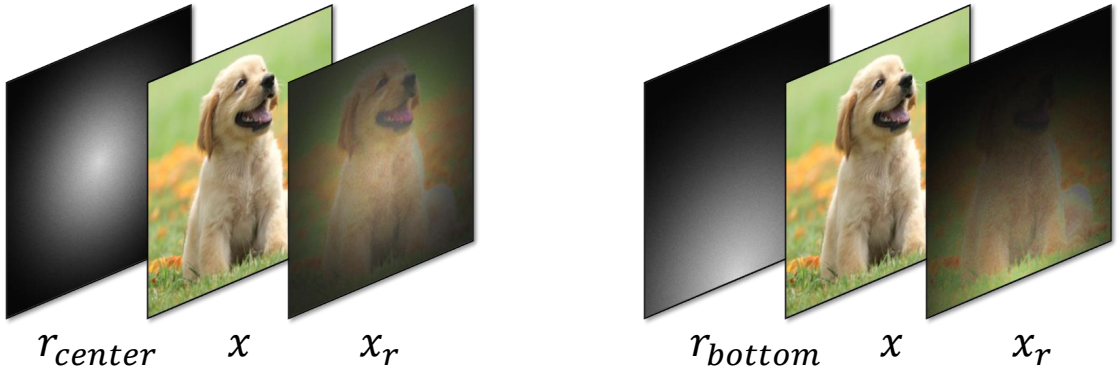


FIGURE 6.7: Examples of retinotopic sampling and how it serves as the instrumental variable.

However, if C is unobservable, both backdoor adjustment and the causal graph in Figure 6.6 (a) cannot remove the confounding effect. To this end, the instrumental variable R is introduced as shown in Figure 6.6 (b), where X is now manipulated by both C and R as $x = w_{cx}c + w_{rx}r + b_x$. Due to the fact that R is independent of C, the weight of causal link $X \to Y$ can be learned by applying different r onto (x, y) pairs, *i.e.*, $y_{r_i} - y_{r_j} = w_{xy}(x_{r_i} - x_{r_j})$. The causal effect is thus estimated as follows:

$$Y[do(X = x)] = \frac{y_{r_i} - y_{r_j}}{x_{r_i} - x_{r_j}} x = w_{xy}x, \tag{6.3}$$

where subscripts $r_i \neq r_j$ indicate the value of X and Y under different instrumental variable R. The d-separated of $R \to X \leftarrow C$ ensures the subtraction eliminating the confounding effect during training.

6.6.2 The Proposed CiiV

With the help of instrumental variable R, the causal effect of the above linear example can be easily estimated. Yet, in practice, the effect of additive R on an image is just as incomprehensible as the additive perturbation C, which doesn't introduce any useful inductive bias. Besides, the above subtraction is also hard to converge during backpropagation, as it may generate confusing gradients with opposite directions of y_{r_i} and y_{r_j}.

In the proposed CiiV framework, we consider the retinotopic sampling mask as a multiplying instrumental variable and use it to augment the original dataset like Figure 6.7. Inspired by the human vision, the retina is known to consist

of photoreceptors and a variety of other neurons [131]. Retinotopic sampling is the result of the non-uniformly spatial distribution of these receptors [86, 248], where the central fovea is significantly denser than the peripheral. It means that human vision is spatially lopsided by a centralized mask, which inspires us to adopt the retinotopic sampling mask with different centers as the instrumental variable R. Luckily, it also satisfies the requirements of a valid instrumental variable discussed in Section. 6.6.1: 1) the pre-defined retinotopic mask is guaranteed to be independent of any confounder in an image; 2) its effect on the prediction Y can only pass through the change of causal features, as the non-robust confounders won't manifest stable patterns under different R. More detailed motivations behind the proposed CiiV will be discussed in Section 6.8.

Therefore, we adopt the multiplying retinotopic mask as our instrumental variable R and design $R \to X$ to be an augmentation function on image $x_r = f(x, r)$, where $f(\cdot)$ applies different retinotopic sampling masks r onto the confounded image x. The function is implemented as a differentiable multiplication layer and proved not to suffer from gradient obfuscation [133] in Section 6.10. Detailed designs and experiments of $f(x, r)$ are investigated in Section 6.9.

Intuitively, when an object moves from the corner of our eyes to the center, the recognizability monotonously increases with the proportion of its captured contour, so we assume that the causal effect is linearly corresponding to the spatial coverage α_r of a retinotopic mask r while the confounding effect is not. It is also consistent with previous findings [249] that visual confounders are usually high-frequency local components unevenly distributed in space. The relationship between the total effect and causal effect can thus be written as follows:

$$Y[X = x_r] = w_{xy} x_r + w_{cy} c \approx \alpha_r Y[do(X = x)] + w_{cy} c. \tag{6.4}$$

Note that we don't need to explicitly observe the above c. We can directly model the $Y[do(X = x)]$ by assigning different r instead. The trick lies in the proposed CiiV regularization loss as follows:

$$L_{CiiV} = \sum_{r_i \neq r_j} \|\alpha_{r_j} Y[X = x_{r_i}] - \alpha_{r_i} Y[X = x_{r_j}]\|, \tag{6.5}$$

where r_i and r_j are two retinotopic sampling masks with spatial coverage α_{r_i} and α_{r_j}, just like r_{center} and r_{bottom} in Figure 6.7. Since $w_{cy} c$ is independent of r, the

above regularization can thus force the model to suppress the confounding effect. In practice, we implement CiiV as an L_1 loss on the feature space extracted by the backbone rather than the logit space, as the classifier weights can be taken out of the above regularization. Otherwise, the L_{CiiV} could hurt the learning of the classifier. The overall training loss would be the combination of the conventional cross-entropy loss and the proposed CiiV loss with a trade-off parameter as $L_{All} = L_{CE} + \beta L_{CiiV}$.

6.7 Details of the Derivation for CiiV Regularization

In this section, we provide a detailed derivation of CiiV Regularization. To begin with, we adopt a more general form of Eq. 6.4 as follows:

$$Y[X = x_r] = f(x_r) + g(c), \tag{6.6}$$

where $f(x_r)$ is the causal effect from the link $R \rightarrow X \rightarrow Y$, and $g(c)$ is the confounding effect from the $C \rightarrow Y$. Without losing the generality, we assume the causal effect and the confounding effect are disentangled as they are indeed two different effects.

By assuming the causal effect being linear interpolated by retinotopic mask R as we discussed, we can have the following equation:

$$Y[X = x_r] \approx \alpha_r Y[do(X = x)] + g(c), \tag{6.7}$$

where $Y[do(X = x)]$ is the overall causal effect, $\alpha_r \in [0, 1]$. Due to the independence between R and C, the confounding effect $g(c)$ won't be affected by α_r. It's worth noting that both no effect or other non-linear random effect from R to C are all acceptable for the following derivation, making the proposed CiiV more general.

After applying two different retinotopic sampling masks r_i and r_j, we will have two equations:

$$\begin{aligned} Y[do(X = x)] &= \frac{1}{\alpha_{r_i}}(Y[X = x_{r_i}] - g(c)) \\ Y[do(X = x)] &= \frac{1}{\alpha_{r_j}}(Y[X = x_{r_j}] - g(c)). \end{aligned} \tag{6.8}$$

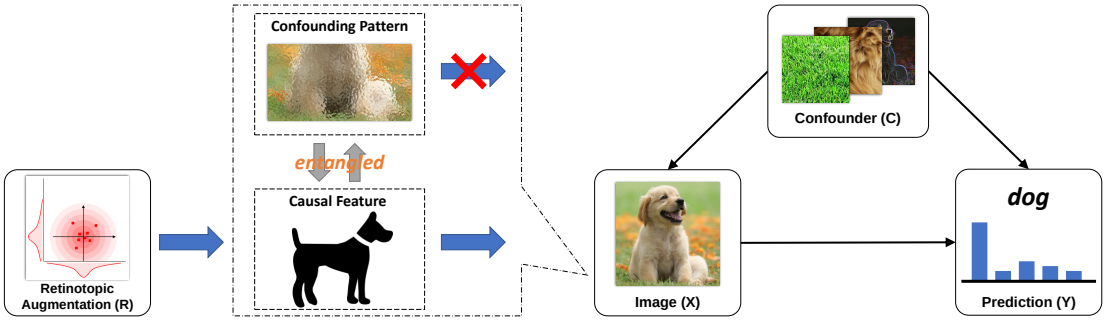


FIGURE 6.8: The details of the proposed causal graph for CiiV regularization and how confounding patterns cause the adversarial vulnerability.

Since the right side of the above two equations are both equal to the same $Y[do(X = x)]$, the proposed CiiV regularization loss can thus be interpreted as directly suppressing the confounding effect $g(c)$:

$$\begin{aligned} L_{CiiV} &= \sum_{r_i \neq r_j} ||\alpha_{r_j} Y[X = x_{r_i}] - \alpha_{r_i} Y[X = x_{r_j}]|| \\ &= \sum_{r_i \neq r_j} ||(\alpha_{r_j} - \alpha_{r_i}) \cdot g(c)||, \end{aligned} \tag{6.9}$$

where $\alpha_{r_j} \neq \alpha_{r_i}$ due to the $r_i \neq r_j$. Since the causal effect is linearly interpolated by the instrumental retinotopic sampling, it can be eliminated in the proposed CiiV regularization, making the CiiV loss suppressing the confounding effect instead during training.

6.8 Details of The Proposed Causal Graph

In this thesis, we firstly attribute the cause of non-robust features, which were originally introduced by [75] as an explanation of adversarial examples, to the ubiquitous confounding effect. But how do confounding patterns affect the learning of causal features and thus hurt the adversarial robustness? We believe the answer is the failure of feature disentanglement [250, 251]. As shown in Figure 6.8, a real-world image is usually composed of both concepts and contexts. Since those contexts often show statistical correlations with the causal concept, it's difficult to disentangle the concept from the context through pure observational data, *e.g.*, the grass feature is usually co-occurred with the dog concept, but it's also shared by other outdoor images and absent in indoor dog images, so it's not a valid causal

feature. Due to the unsuccessful feature disentanglement, adversarial perturbations that simply modify the grass texture would also lead to the collapse of dog feature, which eventually fool the predictor.

However, the feature disentanglement [250, 251] *per se* is still an open question in machine learning. Otherwise, we only need to simply disentangle the robust and non-robust features then learn a classifier based on robust features. To tackle the adversarial vulnerability in practice, we need to bypass the trap of confounder disentanglement and seek help from the causal intervention without confounder observation, *i.e.*, the instrumental variable estimation. As we introduced in section 6.6, there are two requirements for the choice of instrumental variable. The independence of R can be directly guaranteed by the manual design of retinotopic sampling masks. To satisfy the second requirement that the effect of instrumental variable R on Y can only pass through the causal link $X \rightarrow Y$, we assume that causal features are global structures that change consistently across different retinotopic masks while the adversarial patterns are local impulses [249] that simply collapse after applying different retinotopic sampling. Note that this assumption limited the scope of our C to those fragile confounding patterns, which is not trying to disentangle the semantically meaningful confounders. Fortunately, those semantically meaningful confounders brought by the unbalanced dataset also won't be utilized as adversarial perturbations, *e.g.*, the keyboard is usually co-occurred with the monitor and becomes a confounder of the latter, but the adversarial attack is obviously not allowed to create or erase a keyboard for the monitor image based on its definition. Therefore, our assumption still guarantees the proposed retinotopic sampling to be a valid instrumental variable in the adversarial robustness task.

6.9 Details of The Retinotopic Augmentation

In this section, we will introduce the detailed implementation of retinotopic augmentation and the selection of its hyper-parameters. The proposed retinotopic augmentation layer $x_r = f(x, r)$ applies a centralized mask r onto the image x, which imitates the biological retina that the central fovea has significantly denser photoreceptors than the peripheral. The mask r is generated by a non-uniform spatial sampling, whose sampling probability decreases from a given center to the peripheries. We conjecture that human vision benefits from the visual attention

and continuous eye movement [252] to implicitly apply diverse r as the instrumental variable estimation. Intuitively, such a conjecture also explains why human can increase the recognition accuracy by continually gazing at different positions of an object, and why attention or focusing is so important in recognition.

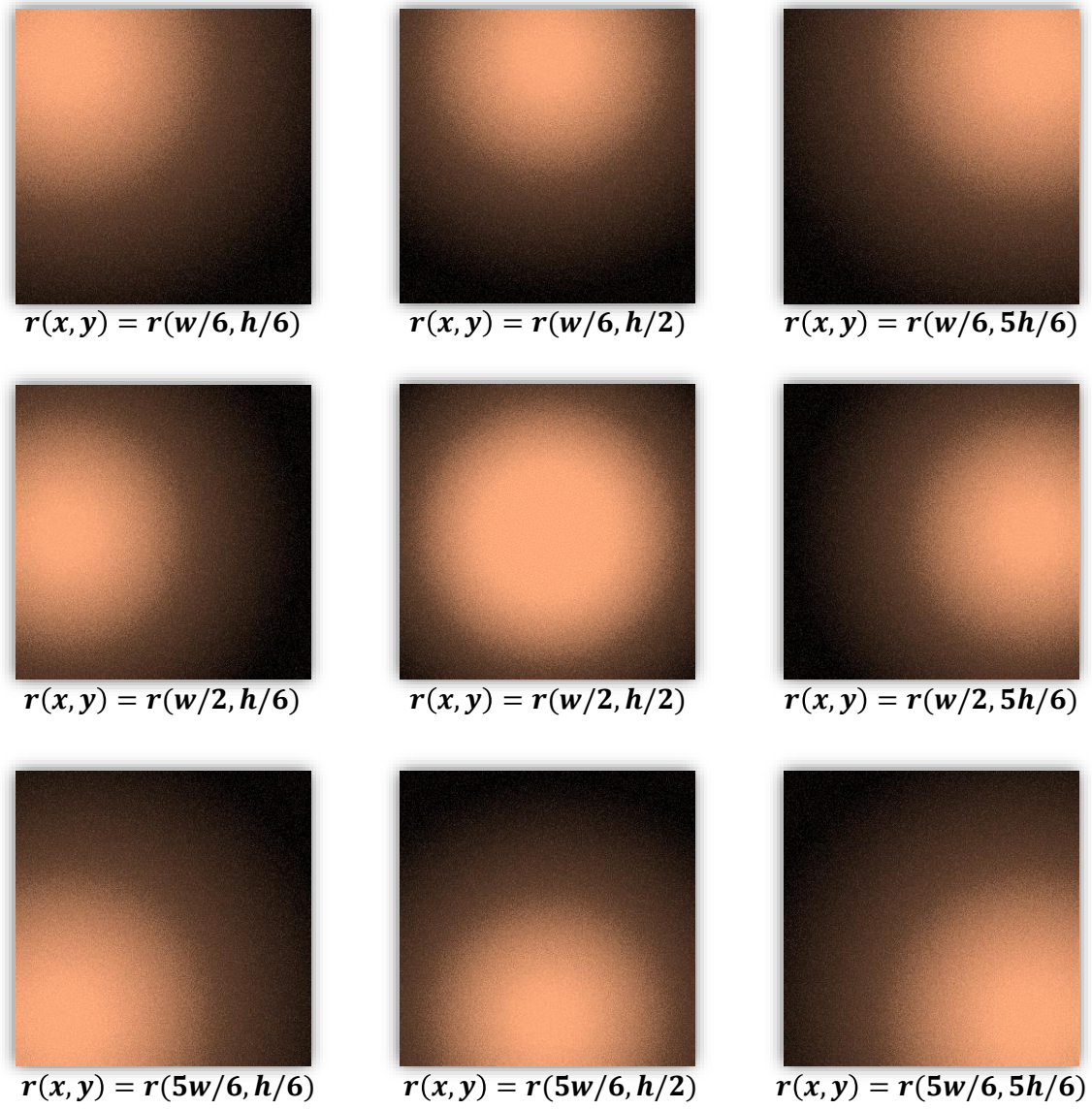
To conduct instrumental variable estimation, we adopt 9 sampling centers (x, y) to generate different r. As illustrated in Figure 6.9 (a), they are $(w/6, h/6)$, $(w/6, h/2)$, $(w/6, 5h/6)$, $(w/2, h/6)$, $(w/2, h/2)$, $(w/2, 5h/6)$, $(5w/6, h/6)$, $(5w/6, h/2)$ and $(5w/6, 5h/6)$, where w and h are the width and height of each corresponding image. Note that the fixed retinotopic centers are only used to ensure the diversity of selected candidates, simply choosing 9 random centers could obtain very similar performances as shown in Table 6.4. Given the retinotopic center (x, y), we define the retinotopic sampling mask r as follows:

$$r_{ij}(x, y) = g(\|(i, j) - (x, y)\|_2) + \varepsilon > \tau, \tag{6.10}$$

where $i \in [0, w], j \in [0, h]$ are the indexes of image pixels, $g(\cdot)$ is a non-linear mapping that can be implemented by various functions, ε is uniformly sampled from $[0, 1]$, $\tau = 0.9$ is the sampling threshold. The spatial coverage α_r used in CiiV is defined as the coverage of the retinotopic mask $\sum r_{ij}/(w * h)$.

Note that the non-linear smoothing function $g(\cdot)$ of $r(x, y)$ can take various implementations, which won't affect the performances of the proposed CiiV too much as long as the sampling frequency decreases from the center (x, y) to the peripheries as shown in Figure 6.9 (a). We intuitively adopt a normalized mapping $g(z) = h((max(z) - z + \alpha)^\gamma)$, $\alpha = 10.0, \gamma = 0.3, h(z) = z/max(z)$ as our default setting, and we further tested two simpler non-linear functions Candidate1: $g_1(z) = 1.0 - z/100$ and Candidate2: $g_2(z) = 2.5/(0.5 \times z^{0.5})$. According to the experimental results in Table 6.5, different $g(\cdot)$ candidates perform very similarly under all attack settings, which proves that CiiV is not sensitive to the detailed implementations of $r(x, y)$. The main reason for us to choose a more complex implementation of $g(\cdot)$ is that it can dynamically fit the image size. The other two simpler functions $g_1(\cdot)$ and $g_2(\cdot)$ have to change parameters for different sizes of images, which is less convenient than our default $g(\cdot)$.

Since the proposed retinotopic augmentation aims to imitate the continuous observations in the human vision. The reaction of biological visual system to different



(a) The Selected Nine Retinotopic Centers

(b) Exposure Parameter ω in Retinotopic Sampling

FIGURE 6.9: (a) The selected 9 retinotopic centers used to generate r in the proposed CiiV. (b) The effect of applying different exposure parameter ω before multiplying with the retinotopic sampling mask r

light intensities is also important, which controls the amount of light absorbed by the retina. Therefore, given the retinotopic mask r generated by $g(\cdot)$, the overall retinotopic augmented image x_r can thus be constructed by $x_r = f(x, r) = 1/N \sum_i (r \odot ReLU(x + \varepsilon_i))$, where ε is the parameter of exposure intensity uniformly sampled from $(-\omega, \omega)$ by N times (N and ω is set to 3 and 0.9, respectively, in our experiments), $\odot$ denotes element-wise multiplication after normalizing the light intensity. The reason we introduce the function $ReLU(x + \varepsilon_i)$ is to find the best exposure ratio for a dataset. As we can see from Figure 6.9 (b), the selection of exposure parameter ω can change the intensity of an observed image. The dark environment requires a smaller ω, so we make it as a hyper-parameter for each dataset. We set ω to 0.9, 0.9, 0.8 for CIFAR-10, CIFAR-100, and mini-ImageNet, respectively. Intuitively, such a function imitates how human eyes react to different light intensities of the environment by controlling the amount of absorbed light. After multiplying with the retinotopic mask r, the proposed $x_r = f(x, r)$ simulates the signals perceived by the biological retina under different environments and focusing points, which continuously "intervene" the images observed by humans.

We also investigated hyper-parameters of retinotopic augmentation. As we can see from Table 6.5, there are trade-offs among different selections. Larger ω can capture more dark details at the cost of light details, and vice versa. To obtain the balanced results between clean images and adversarial examples, we chose the $\omega = 0.9$ as our default setting in CIFAR-10. As to the parameter N that is used to smooth the image after retinotopic augmentation, the larger N we use the less distortion will be in the generated x_r. Therefore, larger N can significantly increase the performance of clean images while smaller N can increase the performance of adversarial examples by suppressing more confounding patterns. Although $N = 1$ would obtain the best overall result, considering the fact that clean images occur more often than adversarial examples in real-world applications, we adopted $N = 3$ as our default setting in all datasets.

Datasets	CIFAR-10						CIFAR-100					
Attackers	Clean	FGSM	PGD-10	AA-l_∞	AA-l_2	Overall	Clean	FGSM	PGD-10	AA-l_∞	AA-l_2	Overall
Baseline	94.42	30.82	0.04	0.0	0.0	25.06	74.53	4.21	0.0	0.0	0.0	15.75
mixup	**95.31**	50.41	2.23	0.0	0.0	29.59	**77.32**	16.60	0.49	0.0	0.0	18.88
BPFC	90.21	24.58	6.19	2.92	35.55	31.89	61.48	17.00	10.23	7.17	29.16	25.01
RS	83.44	53.58	47.06	40.10	75.02	59.84	54.63	26.62	20.21	18.50	47.26	33.44
(ours) CiiV	86.89	64.44	50.75	43.23	82.48	65.56	58.88	32.48	23.63	23.05	55.40	38.69
(ours) CiiV+mixup	87.14	65.28	53.49	**47.24**	81.97	67.02	56.90	35.48	**27.56**	**26.44**	53.14	39.90
(ours) CiiV+RandAug	89.12	**67.96**	**55.01**	47.14	**83.77**	**68.60**	59.26	**36.10**	26.25	25.59	**55.81**	**40.60**
AT_{FGSM}	**84.52**	54.42	43.84	37.94	60.20	56.18	51.99	26.27	22.54	18.31	31.06	30.03
$AT_{PGD\text{-}10}$	83.94	52.90	47.19	43.18	55.46	56.53	**56.48**	25.99	22.56	20.04	28.96	30.81
(ours) CiiV+AT_{FGSM}	83.67	67.28	57.96	50.93	**80.09**	67.99	53.83	**39.00**	32.20	30.48	**50.47**	**41.20**
(ours) CiiV+$AT_{PGD\text{-}10}$	81.35	**68.11**	**59.72**	**54.21**	78.97	**68.47**	51.73	38.59	**33.85**	**32.01**	49.39	41.11

TABLE 6.1: The performances of white-box attack on CIFAR-10 and CIFAR-100. The upper half contains the AT-free defenders while the bottom half reports the AT-involved defenders.

TABLE 6.2: The white-box attack on mini-ImageNet.

Datasets	mini-ImageNet					
Attackers	Clean	FGSM	PGD-10	AA-l_∞	AA-l_2	Overall
Baseline	71.17	1.37	0.01	0.0	0.0	14.51
mixup	**73.88**	2.96	0.0	0.0	0.0	15.37
BPFC	55.34	9.37	3.58	1.74	31.91	20.39
RS	52.15	15.09	13.25	6.93	45.82	26.65
(ours) CiiV	49.18	19.03	9.02	8.73	46.08	26.41
(ours) CiiV+mixup	48.83	23.93	15.12	11.45	45.47	28.96
(ours) CiiV+RandAug	51.65	**32.22**	**24.82**	**18.87**	**48.47**	**35.21**
AT_{FGSM}	45.62	22.12	9.22	3.70	21.39	20.41
$AT_{PGD\text{-}10}$	**49.79**	20.20	16.57	13.52	31.92	26.40
(ours) CiiV+AT_{FGSM}	44.66	30.53	23.83	18.76	**41.57**	**31.87**
(ours) CiiV+$AT_{PGD\text{-}10}$	42.85	**31.72**	**25.46**	**19.30**	39.72	31.81

6.10 Experiments

6.10.1 Datasets and Settings

Datasets. We evaluated the proposed CiiV and other defenders on three benchmark datasets: CIFAR-10, CIFAR-100, and mini-ImageNet [253]. Both CIFAR-10 and CIFAR-100 contain 60K samples with the size of 32x32. mini-ImageNet is originally proposed by [253] for few-shot recognition, which consists of 100 classes and each has 600 images. We scaled the size of images to be 64x64 and split them into train/val/test sets with 42k/6k/12k images.

Training Details. We followed [254]'s project to set all the hyper-parameters and architectures. All models were trained using the SGD optimizer with 0.9 momentum and 5e-4 weight decay. Experiments were conducted on GTX 2080ti GPUs with 128 batch size and 110 total epochs. The learning rate was started with 0.1 and updated by the factor of 0.1 at the following epochs $\{10, 100, 105\}$. The

trade-off parameter β was also initialized by 0.1 then multiplied by 10 at epochs $\{25, 50, 75\}$. Nine retinotopic centers were selected by using the 1/6, 1/2, and 5/6 of width and height for each image. ResNet18 [24] was utilized as the default backbone.

Details of Threat Models. We mainly evaluated the defenders on the clean images together with four threat models: FGSM [41], PGD-10 [239], AA-l_∞ (AutoAttack l_∞), and AA-l_2 (AutoAttack l_2) [226]. For FGSM and PGD-10, the adversarial perturbations were created under l_∞ norm, where the budget radius ϵ was 8/255. PGD-10 ran 10 iterations with step size 2/255. AutoAttack is a recently released parameter-free attack that achieves the state-of-the-art attacking success rate under various defenders. It also prevents the model from gaining a false sense of security from the obfuscated gradients [133]. We set the only parameter ϵ of AA-l_∞ and AA-l_2 to be 8/255 and 0.5, respectively.

Details of Defenders. We divided the defenders into Adversarial Training (AT-involved) and AT-free approaches. For AT-involved, we adopted two popular defenders: AT_{FGSM}, $\text{AT}_{PGD\text{-}10}$, using the same parameters as the corresponding threat models. For AT-free methods, we investigated mixup [132], BPFC [255], and randomized smoothing(RS) [130]. The implementations of mixup and BPFC were directly adopted from their official github repositories. RS was re-implemented in our framework with $\sigma = 0.25$ and the number of test trials $n = 10$. The proposed CiiV itself is also an AT-free method. As a general regularization that is parallel to the above algorithms, we investigated its combination with other defenders as well.

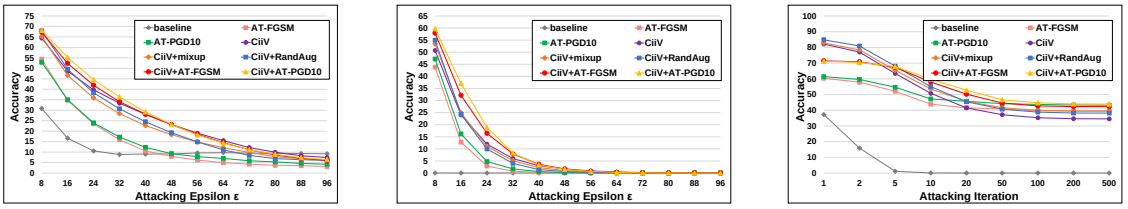
6.10.2 Diagnosis of Adversarial Robustness

The evaluation of adversarial robustness is always controversial as it can easily suffer from flawed or incomplete attack settings. To better eliminate the potential wrong sense of security, we followed [40] to design our experiments and conduct a series of sanity checks at the end of this section.

Adversarial Robustness Against White-box Attack. As reported in Table 6.1 and Table 6.2, we applied multiple white-box attacks on all three datasets. The proposed CiiV and its variants achieved better overall performances among

TABLE 6.3: The performances of targeted PGD-10 under four different targeting settings: untargeted (UT), targeted by most likely / random / least likely categories (T-most, T-random, T-least).

Datasets	CIFAR-10			
Settings	UT	T-most	T-random	T-least
CiiV	50.75	55.62	71.05	74.37
CiiV+mixup	53.49	55.87	73.64	78.48
CiiV+RandAug	55.01	59.18	74.70	77.79
CiiV+AT$_{FGSM}$	57.96	59.44	74.31	77.77
CiiV+AT$_{PGD\text{-}10}$	59.72	60.46	75.21	78.42



(a) Unbounded Attack Using FGSM (b) Unbounded Attack Using PGD-10 (c) Convergence Under Unlimited Iterations

FIGURE 6.10: (a, b) Unbounded attacks on CIFAR-10 that increase the budget ϵ from 8/255 to 96/255. (c) The convergence of defenders under unlimited attacking iterations using PGD.

both AT-free and AT-involved divisions. Note that Random Augmentation(RandAug) [256] is not an adversarial defending method, whose overall performances on three datasets were just 24.38, 16.03, and 15.22, respectively. However, combining CiiV with RandAug worked as well as combining CiiV with AT methods, especially in the real-world mini-ImageNet. It proves that CiiV is indeed a proactive defender that doesn't rely on observing confounders. We also found that AT methods made the model significantly overfit the given attacker in all datasets. Besides, when replacing the training samples of CiiV with AT examples, *i.e.*, CiiV+AT, the robustness came with the price of decreasing clean performances. However, augmenting CiiV with other AT-free methods like mixup and RandAug improved both clean and adversarial performances, which further supported our efforts to design a proactive AT-free defender.

Adversarial Robustness Against AutoAttack. The state-of-the-art AutoAttack is an ensemble of diverse parameter-free attacks [226], including their proposed APGD-CE and APGD-DLR, the black-box Square Attack [257], and the FAB attack [258] that is robust to obfuscated gradients [133]. According to our experiments on AA-l_∞ and AA-l_2, the proposed CiiV performed effectively on all of

the above user-independent attacks. Moreover, combining CiiV with other defenders can further improve their adversarial robustness on both AA settings, proving that the CiiV is a general causal regularization parallel to most of the previous methods.

Adversarial Robustness Against Targeted Attack. The performances of the proposed CiiV under untargeted and targeted PGD-10 attacks were reported in Table 6.3 using CIFAR-10 dataset. The targeted results were further divided into three protocols: 1) most likely category, 2) least likely category, and 3) random category. The results revealed that the confounding effect could also be the cause of ambiguous prediction, as similar categories are easier to be attacked. We also noticed that the performances under untargeted attack would be closer to the most likely targeted attack when the robustness of the model increases. It's probably because the similar categories share the similar confounder distributions, *i.e.*, environments, and thus utilized by the attacker.

Adversarial Robustness Under Unbounded Attack. To evaluate the validity of defenders, we demonstrated the performances of CiiV and its variants together with the baseline and two AT models under unbounded attacking in Figure 6.10 (a, b). When the budget ϵ of the attacker was increased from 8/255 to 96/255, all performances were either converged to 0% accuracy for the strong PGD attack or converged to the random guesses for the weak FGSM attack. Any valid defender shouldn't survive such an unbounded attack, as it allows the attacker to modify the entire image and erasing all causal features. We also tested unlimited iterations of PGD attack, all CiiV and its variants are successfully converged after 100 iterations as shown in Figure 6.10 (c). Note that AT and CiiV+AT are more robust than other defenders in this setting, which is probably caused by the exposure of adversarial examples during training.

Ablation Studies. In this paragraph, we evaluated the effectiveness of different settings and parameters of the proposed CiiV. As reported in Table 6.4, 1) we investigated the L_1 and L_2 versions of the CiiV loss, where the L_1 is slightly better than L_2; 2) we tried random assignments of the retinotopic centers as R=$\{r_{rand}\}$, which is very close to our fixed centers; 3) we also reported the performances of retinotopic augmentation only as RetiAug, which had higher clean performance but worse adversarial robustness than CiiV. Note that RetiAug itself can also be treated as an approximation of CiiV by assigning all α to 1.0. Besides, cross-entropy

TABLE 6.4: Ablation Studies of CiiV on CIFAR-100.

Datasets	CIFAR-100					
Attackers	Clean	FGSM	PGD-10	AA-l_∞	AA-l_2	Overall
L_1 CiiV	58.88	32.48	23.63	23.05	55.40	38.69
L_2 CiiV	57.93	31.78	22.27	22.28	54.51	37.75
R=$\{r_{rand}\}$	58.79	32.20	22.35	22.90	55.28	38.30
RetiAug	61.88	31.69	20.29	18.32	53.19	37.07
$\beta = 0.01$	59.85	32.00	21.52	20.85	54.23	37.69
$\beta = 1.0$	55.26	34.48	26.80	25.02	51.82	38.68
$N_R = 2$	56.45	30.47	21.57	21.03	53.39	36.58
$N_R = 5$	58.34	32.01	22.34	22.61	54.58	37.98

losses under different r also forced the model to ignore the non-robust confounding patterns; 4) other choices of hyper-parameters of CiiV were also reported, we found that β empirically served as a trade-off between clean performance and adversarial robustness, and applying more retinotopic sampling masks (larger N_R) would make a better estimation, yet, its improvements got converged.

Visualization. We visualized the generated PGD perturbations for models w/ and w/o CiiV in Figure 6.11. It demonstrates that the baseline models can be easily fooled by imperceptible confounders while the proposed CiiV forces the model to learn causal features, as the adversarial perturbations have to erase the structural patterns to fool the CiiV model.

The Evaluation Checklist. To verify that the proposed CiiV doesn't suffer from flawed or incomplete evaluations, the above experiments were designed to follow a series of sanity checks introduced by [40]: 1) Iterative attacks are better than single-step attacks, *e.g.*, PGD *vs* FGSM in Table 6.1&6.2 and Figure 6.10. 2) Unbounded adversarial examples become random guessing or 0% accuracy, *e.g.*, Figure 6.10 (a,b). 3) The accuracy converges with the increasing of attack steps: Figure 6.10 (c). 4) Investigating both targeted attacks and untargeted attacks, *e.g.*, Table 6.3. 5) Using black-box attacks and the attacks circumventing obfuscated gradients to avoid the potentially flawed adversarial example generation, *e.g.*, the results under AA-l_∞ and AA-l_2 in Table 6.1&6.2.

6.11 More detailed studies and experiments

In this section, we demonstrate additional studies and experiments on 1) several gradient-free attacks, and 2) more backbones.

Datasets	CIFAR-10					
Attackers	Clean	FGSM	PGD-10	AA-l_∞	AA-l_2	Overall
Default	86.89	64.44	50.75	43.23	82.48	65.56
Candidate1	87.03	64.53	50.37	42.46	82.86	65.45
Candidate2	87.53	63.23	48.88	41.05	82.64	64.67
$\omega = 1.2$	86.50	64.96	52.04	46.33	82.18	66.40
$\omega = 1$	85.96	64.66	50.81	43.88	81.72	65.41
$\omega = 0.8$	87.21	64.19	49.97	41.72	82.74	65.17
$\omega = 0.6$	86.47	62.65	47.70	39.53	81.36	63.54
$N = 1$	82.44	69.40	59.17	57.25	78.16	69.28
$N = 2$	85.84	66.52	55.01	48.36	81.36	64.42
$N = 4$	87.71	63.77	48.39	40.14	83.06	64.61
$N = 5$	88.29	62.22	46.35	37.87	83.92	63.73

TABLE 6.5: The performances of CiiV on CIFAR-10 using different designs of function $g(\cdot)$ to generate retinotopic sampling mask r, and different hyper-parameters ω and N to generate x_r.

Datasets	CIFAR-10					CIFAR-100				
Attackers	Clean	GN	UN	SPSA	BFS	Clean	GN	UN	SPSA	BFS
Baseline	94.42	72.05	74.66	68.60	35.77	74.53	31.44	34.46	27.57	10.39
mixup	95.31	76.02	78.77	71.87	39.82	77.32	40.21	43.69	36.27	18.35
BPFC	90.21	88.90	89.02	88.91	79.48	61.48	60.47	60.53	60.30	52.11
RS	83.44	83.22	83.16	83.35	82.97	54.63	54.46	54.28	54.53	54.41
(ours) CiiV	86.89	86.47	86.61	86.75	85.60	58.88	58.19	58.75	58.63	57.30
(ours) CiiV+mixup	87.14	86.71	87.11	87.07	86.23	56.90	56.41	56.65	56.70	55.91
(ours) CiiV+RandAug	89.12	88.36	88.59	89.00	87.64	59.26	58.82	58.71	59.21	57.85
AT_{FGSM}	84.52	83.02	83.32	82.98	77.86	51.99	51.05	51.27	51.13	45.08
$\text{AT}_{PGD\text{-}10}$	83.94	82.49	82.70	82.59	77.33	56.48	54.64	54.97	54.82	47.83
(ours) CiiV+AT_{FGSM}	83.67	83.10	83.20	83.37	82.34	53.83	53.35	53.77	53.69	52.53
(ours) CiiV+$\text{AT}_{PGD\text{-}10}$	81.35	80.74	81.02	81.09	80.22	51.73	51.16	51.43	51.39	50.12

TABLE 6.6: Gradient-free attacks on CIFAR-10 and CIFAR-100. The upper half contains the AT-free defenders while the bottom half reports the AT-involved defenders.

Datasets	CIFAR-10						CIFAR-100					
Attackers	Clean	FGSM	PGD-10	AA-l_∞	AA-l_2	Overall	Clean	FGSM	PGD-10	AA-l_∞	AA-l_2	Overall
(VGG13) Baseline	90.20	10.48	0.0	0.0	0.21	20.18	66.05	3.14	0.0	0.0	0.05	13.85
(VGG13) CiiV	83.44	58.75	43.62	36.98	79.08	60.37	50.62	32.83	25.58	24.91	47.37	36.26
(VGG13) CiiV+RandAug	83.91	60.60	45.54	40.51	80.63	62.04	52.72	33.24	26.66	25.76	48.64	37.40
(WRN34-10) Baseline	94.93	32.09	0.02	0.0	0.0	25.41	77.74	9.73	0.15	0.0	0.0	17.52
(WRN34-10) CiiV	87.25	59.50	43.89	38.24	82.82	62.34	60.84	34.04	25.58	25.23	57.45	40.63
(WRN34-10) CiiV+RandAug	88.68	64.02	49.23	44.52	83.59	66.01	63.23	38.30	28.82	27.38	59.81	43.51

TABLE 6.7: The performances of Baseline, CiiV, and CiiV+RandAug using different backbones.

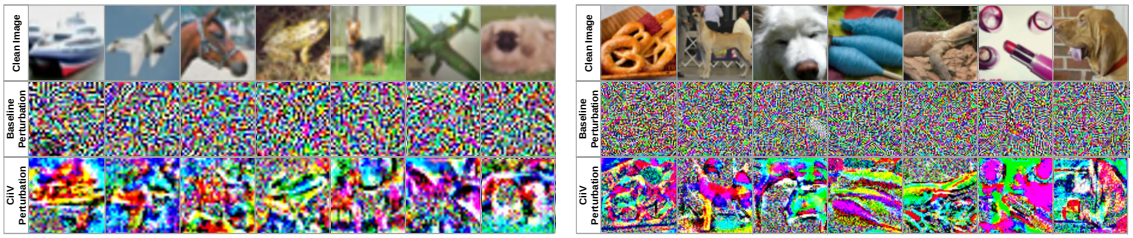


FIGURE 6.11: Generated perturbations of models w/ and w/o CiiV on CIFAR-10 and mini-ImageNet.

According to [40], some flawed defenders may fail in gradient-free attacks. Therefore, we further investigated four gradient-free attackers: 1) GN (Gaussian Noise), 2) UN (Uniform Noise), 3) SPSA [259], and 4) BFS (Brute-Force Search) [40]. Since gradient-free attacks are supposed to be much weaker than gradient-based attacks, we increased the budget ε to 16/255 for all four gradient-free attackers under l_∞ constraint. To be specific, GN and UN add gaussian and uniform noises, respectively, to input images. BFS ran 100 times of GN and reported the most vulnerable adversarial examples. As to the SPSA, it conducted numerical approximation of gradients to circumvent the potential gradient masking, the hyper-parameters were set as δ=0.1, step=20, lr=0.1, batch size=16. According to the experiments in Table 6.6, all the gradient-free attackers were significantly weaker than the gradient-based attackers as we expected even with the doubled attacking budget, proving that the proposed CiiV won't be more vulnerable under gradient-free attacks.

We also applied the proposed CiiV and its combination with Random Augmentation, *i.e.*, the AT-free versions of defenders, into other backbones, *e.g.*, VGG13 [260] and WRN34-10 [261]. As we can see from Table 6.7, the proposed CiiV and its variants consistently increased the adversarial robustness under different backbone models.

6.12 Conclusion

In this Chapter, we presented a CiiV defender that worked as a general causal regularization without the need for adversarial examples. CiiV consists of a spatial data augmentation using different retinotopic sampling masks, and a regularization loss that encourages the model to suppress local confounding patterns by learning features linearly responding to spatial interpolations. We followed the checklist

from [40] to design our evaluation experiments and adopted the user-independent AutoAttack [226] as the main indicator of adversarial robustness. Extensive experiments on all settings proved that CiiV is robust against various adaptive attacks, and it can also serve as a plug-and-play regularization for other defenders. Besides, this Chapter also provides a fundamental viewpoint of the relationship between adversarial robustness and causal intervention, which may guide the design of future defenders.

Chapter 7

Summary

7.1 Conclusion

In this thesis, we systematically studied a variety of potential threats to the robustness of deep neural networks under distribution shifts, including the network architecture that is used to extract the visual context features, the explicitly biased distributions of data annotations, and the implicit confounding bias caused by co-occurred fragile patterns. Due to the vast applications of DNNs in safety-critical systems like face recognition, quality inspection or public safety, obtaining a reliable DNN system that works well on the harshest situations and preventing malicious attackers from modifying the model prediction becomes increasingly important [262, 263].

Therefore, to tackle the emerging and diverse challenges of robustness against distribution shifts in recent computer vision applications, we summarize the robust DNN systems into satisfying the following three demands: 1) architectural robustness, 2) long-tailed robustness and 3) adversarial robustness. The latter two can also be viewed as the explicit and implicit pattern bias on features extracted by DNNs, where the explicit pattern bias follows the distribution of data annotations and the implicit bias is caused by non-observable confounding effects. The proposed methods are evaluated across several popular computer vision tasks, including image classification, object detection, instance segmentation, scene graph generation, and visual question answering.

7.1.1 Architectural Robustness

In Chapter 3, we propose the VCTree that learns to extract robust visual contexts by putting objects into a dynamic tree structure. In computer vision, context features can be captured by either pixel-level convolutions on larger receptive fields or object-level message passing mechanisms between contents. Since the over-fitting problem of large receptive fields has been well treated by methods like dilated convolutions [60] or feature pyramids [18], we mainly focus on the latter situation in this thesis. Existing methods use fixed layouts like a fully connected graph or chain to organize the objects, causing the message passing saturation, *i.e.*, over-fitting to counting objects. Meanwhile, the VCTree architecture varies from content to content, task to task, forcing the DNN model to learn robust context features rather than exploit statistical priors. Experimental results on Scene Graph Generation and Visual Question Answering show that VCTree can significantly increase the architectural robustness in comparison to other message-passing architectures, especially in those fair metrics or test sets.

7.1.2 Long-Tailed Robustness

In Chapter 4 and Chapter 5, we proposed the Total Direct Effect (TDE) that applies a self-critical inference based on the difference between factual and counterfactual outputs of the given causal graph to dynamically remove the biased component in the predict logits. It significantly improves the robustness against long-tailed bias, outperforming previous debiasing methods in scene graph generation, image classification, object detection, and instance segmentation. We proposed two types of TDE algorithms including 1) multi-modal TDE that directly applied to the causal graph of multi-modality during inference and 2) de-confound TDE that incorporates de-confound training and TDE inference. The former resolves the unbalanced contributions between different modalities, using the dynamic difference to cut the shortcuts of memorizing the distributions. The latter jointly combines the de-confound training to eliminate the magnitude bias of feature vectors and the TDE inference to adjust the direction of feature vectors, which works well on most of the common single-modal computer vision tasks.

The advantages of the proposed TDE for long-tailed robustness are three-fold:

- It maintains the instance-balanced training that prevents the re-balancing strategies from hurting the feature extraction backbones.
- It doesn't introduce any additional learning stages or modules, and can be easily applied to a variety of tasks.
- Unlike most of the long-tailed algorithms, it doesn't rely on the accessibility of data distribution, making it applicable to any situations *e.g.*, even working on online and streaming data.

7.1.3 Adversarial Robustness

In Chapter 6, we systematically reviewed the common strategies of defenders in the proposed causal framework. Then, we introduce a Causal intervention by instrumental Variable (CiiV) regularization to train robust DNN against adversarial attacks. It does not only offer a proactive defender avoiding the endless adversarial training, but also opens a novel yet fundamental viewpoint of adversary research. It has a fully differentiable retinotopic sampling layer that augments the input samples and a regularization loss to minimize the confounding effect. The proposed CiiV is stable and guaranteed not to suffer from gradient obfuscation. To correctly verify the adversarial robustness of the proposed CiiV, we strictly follow the adversarial evaluation checklist from Carlini *et al.* [40] for sanity checks on a wide range of attacker settings of CIFAR- 10, CIFAR-100, and mini-ImageNet. This demonstrates that CiiV withstands adaptive attacks.

7.1.4 Causal Inference

Combining the long-tailed robustness with adversarial robustness, we found that the causal inference can be applied to tackle both explicit and implicit bias in computer vision tasks. From Chapter 4 to Chapter 6, we took inspirations from the causal inference to design the proposed algorithms. Due to the various forms of potential bias, the most tricky part of achieving the robust DNN is to identify the source of threats and which part of the algorithm contains the corresponding shortcuts. Thanks to the causality, the causal graph provides us a perfect framework to analyze the interactions between variables of interest, and handy tools to locate

the confounding bias or mediation shortcut. We believe that the key to reaching the robust machine learning system is to replace association-based learning with causation-based learning.

7.2 Future Work

In the future, we are going to keep pushing the boundary of robust inference against distribution shifts, solving new challenges brought by evolving DNN algorithms and computer vision applications.

7.2.1 Architectural Robustness

Recently, the Transformer structure [218] is widely adopted in various natural language processing and computer vision tasks, revealing astonishing abilities to memorize massive data and conduct complicated reasoning. Yet, its robustness is far from being satisfactory. To better understand whether the malicious contents in the training data can train an "evil" model or whether the systematic bias in the real world will hurt the fairness of the model, it's crucial to investigate the architectural robustness for the Transformer structure given the fact that it has been widely used in lots of real-world applications.

7.2.2 Long-Tailed Robustness

Existing long-tailed algorithms formulate the long-tailed problem as overcoming the long-tailed distribution on the category level. However, there could be also implicit long-tailed bias in the context level. For example, as illustrated in Figure 7.1, even if a wild animal dataset is well balanced for all the animal classes, the inherent features may still biased towards the patterns of grass or trees, causing the DNN model hard to generalize to the environment of in-door zoos. Hence, we believe the generalized long-tailed classification should tackble both class-wise imbalance and context-wise imbalance.

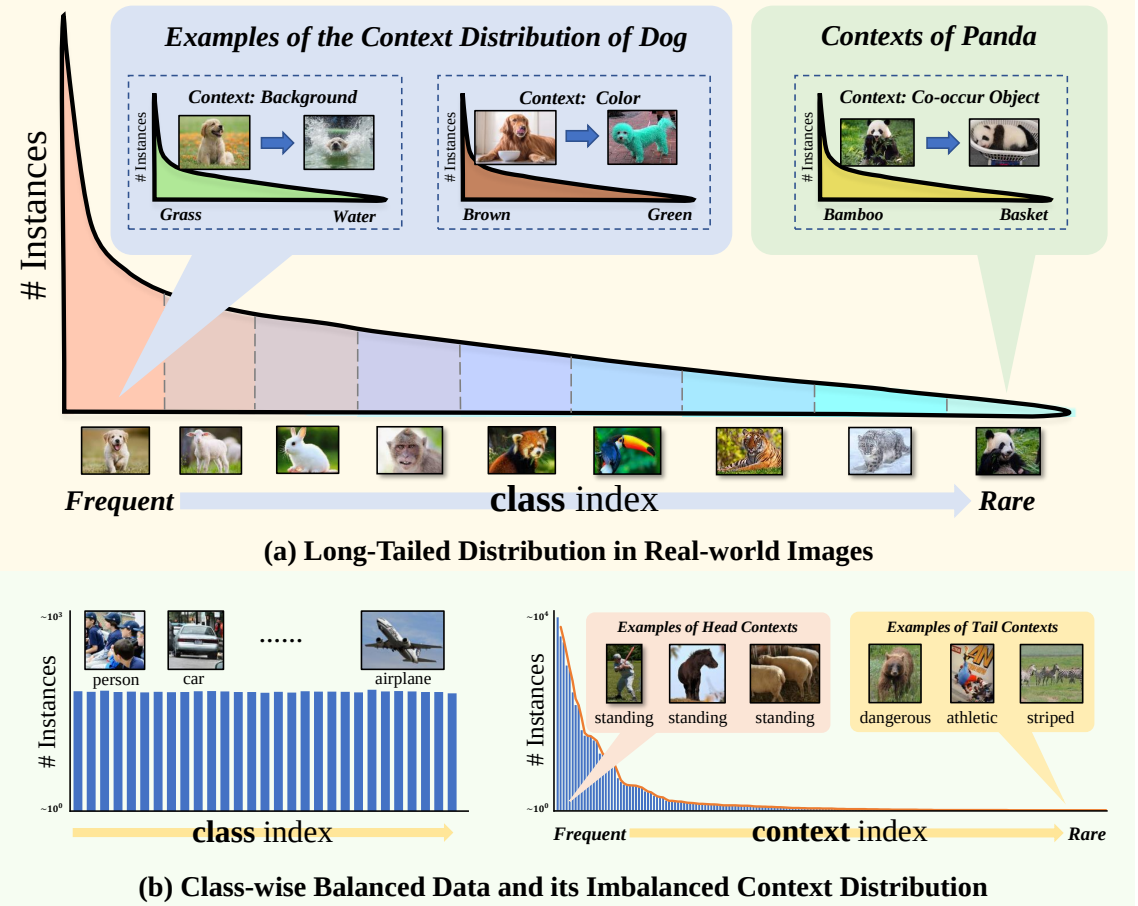


FIGURE 7.1: (a) Long-tailed distribution is both class-wise and context-wise. (b) Even if we balance the class distribution of MSCOCO-Attribute [13], the context (attributes) would still be long-tailed.

7.2.3 Adversarial Robustness

Different from the above two problems, the adversarial attackers keep evolving with the algorithm development. For example, in the popular large-scale Transformer models, researchers even found that some specific inputs, which are meaningless to humans, can trigger the Transformer to leak the private data in the training set [264], *e.g.*, telephone numbers. Meanwhile, due to the wide applications of DNN-based algorithms, there are increasing demands for adversarial robustness to ensure the safety of systems. Therefore, the proposed CiiV is not the end. We still need to face new challenges like the above-mentioned data leaking.

List of Author's Awards, and Publications[1]

Awards

- **2021 PREMIA Best Student Paper Awards**, "The Certificate of Merit (2nd Place)" *PREMIA*.
- **2020 Alibaba Outstanding Interns in Academic Cooperation**, *ALIBABA GROUP*
- **2019 PREMIA Best Student Paper Awards**, "Silver Award (2nd Place)" *PREMIA*.
- **2019 CVPR Best Paper Finalists**, *CVPR Committee*.

Conference Proceedings

- **Kaihua Tang**, Mingyuan Tao, Xian-Sheng Hua, Hanwang Zhang, "Adversarial Visual Robustness by Causal Intervention", *arXiv preprint (Under Review), 2021.* Link: `https://arxiv.org/pdf/2106.09534.pdf`
- Xinting Hu, **Kaihua Tang**, Chunyan Miao, Xian-Sheng Hua, Hanwang Zhang, "Distilling Causal Effect of Data in Class-Incremental Learning", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.* Link: `https://arxiv.org/pdf/2103.01737.pdf`

[1]The superscript * indicates joint first authors

- Yulei Niu, **Kaihua Tang**, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, Ji-Rong Wen, "Counterfactual VQA: A Cause-Effect Look at Language Bias", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.* Link: `https://arxiv.org/pdf/2006.04315.pdf`

- **Kaihua Tang**, Jianqiang Huang, Hanwang Zhang, "Long-Tailed Classification by Keeping the Good and Removing the Bad Momentum Causal Effect", in *Conference on Neural Information Processing Systems (NeurIPS), 2020.* Link: `https://arxiv.org/pdf/2009.12991.pdf`

- **Kaihua Tang**, Yulei Niu, Jianqiang Huang, Jiaxin Shi, Hanwang Zhang, "Unbiased Scene Graph Generation from Biased Training", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020*, **(ORAL Presentation).** Link: `https://arxiv.org/pdf/2002.11949.pdf`

- Xinting Hu, Yi Jiang, **Kaihua Tang**, Hanwang Zhang, Chunyan Miao, Jingyuan Chen, "Learning to Segment the Tail", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.* Link: `https://arxiv.org/pdf/2004.00900.pdf`

- **Kaihua Tang**, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, Wei Liu, "Learning to Compose Dynamic Tree Structures for Visual Contexts, " in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019*, **(ORAL Presentation & Appears on the Best Paper Finalists [45/5160]).** Link: `https://arxiv.org/pdf/1812.01880.pdf`

- Xu Yang, **Kaihua Tang**, Hanwang Zhang, Jianfei Cai, "Auto-Encoding Scene Graphs for Image Captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019*, **(ORAL Presentation).** Link: `https://arxiv.org/pdf/1812.02378.pdf`

Journal Articles

- Mitra Tajrobehkar, **Kaihua Tang**, Hanwang Zhang, Joo-Hwee Lim, "Align R-CNN: A Pairwise Head Network for Visual Relationship Detection," *IEEE Transactions on Multimedia, 2021.*

Bibliography

[1] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. xix, xx, xxv, 4, 16, 24, 25, 26, 30, 36, 37, 38, 39, 40

[2] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. xix, 4, 13, 16, 24, 25, 36, 37, 39, 43, 47, 61, 62, 63, 64

[3] Kai Sheng Tai, Richard Socher, and Christopher D Manning. Improved semantic representations from tree-structured long short-term memory networks. In *ACL*, 2015. xix, 24, 26, 29, 33, 34, 38, 53

[4] Thomas H. Cormen, Clifford Stein, Ronald L. Rivest, and Charles E. Leiserson. *Introduction to Algorithms*. McGraw-Hill Higher Education, 2001. xix, 24, 26

[5] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 2017. xx, 26, 28, 36, 48, 49, 52, 60

[6] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. xx, xxi, 13, 14, 47, 48, 49, 51, 52, 53, 54, 61, 62, 63, 64, 65, 67, 73

[7] Herbert A Simon. Bounded rationality. In *Utility and probability*. Springer, 1990. xxi, 48, 70

[8] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. Learning to compose dynamic tree structures for visual contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. xxi, xxv, 2, 13, 47, 49, 52, 53, 54, 61, 62, 63, 64, 65, 71, 74

[9] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. xxi, 52, 53, 54, 61, 74

[10] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. xxii, xxv, xxvi, 1, 2, 14, 15, 17, 20, 80, 81, 82, 89, 90, 91, 93, 94, 95, 96, 97, 98, 99, 100, 102

[11] 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 (ICLR)*, 2020. xxii, 2, 6, 14, 15, 17, 20, 80, 81, 82, 89, 90, 91, 94, 95, 96, 97, 100, 101

[12] 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 *International Conference on Computer Vision (ICCV)*, 2017. xxii, 97, 101

[13] Genevieve Patterson and James Hays. Coco attributes: Attributes for people, animals, and objects. In *ECCV*. Springer, 2016. xxiii, 137

[14] Yalong Bai, Jianlong Fu, Tiejun Zhao, and Tao Mei. Deep attention neural tensor network for visual question answering. In *European Conference on Computer Vision(ECCV)*, 2018. xxv, 14, 41, 42, 43

[15] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin. Knowledge-embedded routing network for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. xxv, 2, 47, 52, 62, 63, 71

[16] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. Bbn: Bilateral-branch network with cumulative learning for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. xxv, 6, 15, 17, 20, 80, 81, 82, 89, 91, 95, 96, 97, 98, 100

[17] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. xxvi, 82, 89, 98, 99, 101

[18] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2117–2125, 2017. xxvi, 4, 15, 16, 23, 24, 64, 66, 99, 101, 134

[19] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5356–5364, 2019. xxvi, 6, 15, 79, 80, 82, 96, 97, 98, 99, 100

[20] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. xxvi, 88, 89, 90, 91, 94, 95, 98, 99, 100, 102

[21] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. xxvi, 88, 89, 90, 91, 94, 95, 98, 99, 100, 102

[22] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in neural information processing systems*, pages 3856–3866, 2017. xxvi, 91, 95, 98, 99, 100, 102

[23] *LVIS v0.5 Evaluation Server*, 2020. `https://evalai.cloudcv.org/web/challenges/challenge-page/473/overview`. xxvi, 100, 103

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 3, 5, 15, 79, 125

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 3

[26] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 5, 64, 66, 79, 98, 100

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, 2015. 1, 3, 4, 5, 15, 24, 25, 27, 37, 42, 52, 64, 79, 89

[28] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision(ECCV)*, 2016. 4, 24

[29] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *International Conference on Computer Vision (ICCV)*, 2017. 4, 24, 27, 30, 53, 66, 67

[30] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. Panoptic segmentation. *arXiv preprint arXiv:1801.00868*, 2018. 1, 27

[31] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017. 1

[32] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 1

[33] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020. 1

[34] Melanie Mitchell. Why ai is harder than we think. *arXiv preprint arXiv:2104.12871*, 2021. 1

[35] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 1

[36] Drew McDermott, M Mitchell Waldrop, B Chandrasekaran, John McDermott, and Roger Schank. The dark ages of ai: a panel discussion at aaai-84. *AI Magazine*, 6(3):122–122, 1985. 1

[37] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 1

[38] Matthias Hein, Maksym Andriushchenko, and Julian Bitterwolf. Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 41–50, 2019. 1

[39] Jingru Tan, Changbao Wang, Buyu Li, Quanquan Li, Wanli Ouyang, Changqing Yin, and Junjie Yan. Equalization loss for long-tailed object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 6, 14, 20, 81, 82, 89, 96, 98, 99, 101

[40] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv*, 2019. 1, 2, 15, 108, 125, 128, 130, 131, 135

[41] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Machine Learning (ICML)*, 2015. 1, 2, 7, 105, 106, 111, 125

[42] Judea Pearl and Dana Mackenzie. *THE BOOK OF WHY: THE NEW SCIENCE OF CAUSE AND EFFECT*. Basic Books, 2018. 1, 7, 17, 18, 50, 52, 55, 58, 109

[43] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex Smola. Correcting sample selection bias by unlabeled data. *Advances in neural information processing systems*, 19:601–608, 2006. 2

[44] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[45] Daqing Liu, Hanwang Zhang, Feng Wu, and Zheng-Jun Zha. Learning to assemble neural module tree networks for visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4673–4682, 2019. 2

[46] Xu Yang, Hanwang Zhang, and Jianfei Cai. Learning to collocate neural modules for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4250–4260, 2019. 2

[47] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *International Conference on Learning Representations (ICLR)*, 2013. 2, 7, 15, 21, 105, 109, 114

[48] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations (ICLR)*, 2019. 2, 15, 21, 109, 114

[49] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *International Conference on Learning Representations (ICLR)*, 2018. 2, 114

[50] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *International Conference on Learning Representations (ICLR)*, 2018.

[51] Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow. Thermometer encoding: One hot way to resist adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018. 15, 21, 109, 114

[52] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *International Conference on Learning Representations (ICLR)*, 2018.

[53] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations (ICLR)*, 2018. 2

[54] Judea Pearl. *Causality: models, reasoning and inference*. Springer, 2000. 3, 7, 18, 52

[55] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 2018. 4, 16, 23, 24

[56] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 4, 23

[57] Santosh K Divvala, Derek Hoiem, James H Hays, Alexei A Efros, and Martial Hebert. An empirical study of context in object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 4, 23

[58] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to refine object segments. In *European Conference on Computer Vision(ECCV)*, 2016. 4, 24

[59] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 4, 16, 24

[60] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations (ICLR)*, 2016. 4, 16, 24, 134

[61] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. Iterative visual reasoning beyond convolutions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4, 25

[62] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 13

[63] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. Factorizable net: an efficient subgraph-based framework for scene graph generation. In *European Conference on Computer Vision(ECCV)*, 2018. 16

[64] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao, and Chen Change Loy. Zoom-net: Mining deep feature interactions for visual relationship recognition. In *European Conference on Computer Vision(ECCV)*, 2018. 4, 13, 25

[65] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997. 4, 25, 33, 35

[66] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *International Conference on Computer Vision (ICCV)*, 2015. 4, 25

[67] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 4

[68] Takeo Watanabe, Alexander M Harner, Satoru Miyauchi, Yuka Sasaki, Matthew Nielsen, Daniel Palomo, and Ikuko Mukai. Task-dependent influences of attention on the activation of human primary visual cortex. *Proceedings of the National Academy of Sciences*, 1998. 5, 25

[69] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. 5, 15, 79, 96

[70] 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 *European Conference on Computer Vision(ECCV)*, pages 740–755. Springer, 2014. 5, 15, 61, 79

[71] William J Reed. The pareto, zipf and other power laws. *Economics letters*, 74(1):15–19, 2001. 6, 79

[72] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *European Conference on Computer Vision(ECCV)*, 2020. 6, 80

[73] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning (ICML)*, pages 1139–1147, 2013. 6, 81, 83, 85

[74] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999. 6, 81, 82, 83, 85

[75] Andrew Ilyas, Shibani Santurkar, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *NeurIPS*, 2019. 7, 8, 106, 119

[76] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arxiv*, 2018. 7, 105

[77] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7, 15, 21, 109, 114

[78] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *International Conference on Learning Representations (ICLR)*, 2018. 8, 106

[79] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on mnist. In *International Conference on Learning Representations (ICLR)*, 2019. 8, 106

[80] Huan Zhang, Hongge Chen, Zhao Song, Duane Boning, Inderjit S Dhillon, and Cho-Jui Hsieh. The limitations of adversarial training and the blind-spot attack. In *International Conference on Learning Representations (ICLR)*, 2019. 8, 106

[81] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? *arXiv preprint arXiv:2007.08489*, 2020. 8, 106

[82] Alexander D'Amour. On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility, and alternatives. In *AISTATS*, 2019. 8, 113

[83] Suzana Herculano-Houzel. The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost. *Proceedings of the National Academy of Sciences*, 2012. 8, 106

[84] Richard H Masland. The fundamental plan of the retina. *Nature Neuroscience*, 2001. 8, 106

[85] Edward Kim, Jocelyn Rego, Yijing Watkins, and Garrett T Kenyon. Modeling biological immunity to adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 8, 15, 106

[86] Michael J Arcaro, Stephanie A McMains, Benjamin D Singer, and Sabine Kastner. Retinotopic organization of human ventral visual cortex. *Journal of neuroscience*, 2009. 8, 107, 117

[87] Judea Pearl. *Causality*. Cambridge university press, 2009. 8, 106

[88] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 13, 47

[89] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 13, 47, 63

[90] Ranjay Krishna, Ines Chami, Michael Bernstein, and Li Fei-Fei. Referring relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 13

[91] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 13, 47

[92] Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. Linknet: Relational embedding for scene graph. In *Advances in Neural Information Processing Systems*, 2018. 13

[93] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and caption regions. In *International Conference on Computer Vision (ICCV)*, 2017. 13, 25, 36

[94] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *European Conference on Computer Vision(ECCV)*, 2018. 13

[95] Jiuxiang Gu, Handong Zhao, Zhe Lin, Sheng Li, Jianfei Cai, and Mingyang Ling. Scene graph generation with external knowledge and image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[96] Mengshi Qi, Weijian Li, Zhengyuan Yang, Yunhong Wang, and Jiebo Luo. Attentive relational networks for mapping images to scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[97] Wenbin Wang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Exploring context and visual pattern of relationship for scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 13

[98] Alejandro Newell and Jia Deng. Pixels to graphs by associative embedding. In *NIPS*, 2017. 13, 36, 39

[99] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 13, 39

[100] Yuanzhi Liang, Yalong Bai, Wei Zhang, Xueming Qian, Li Zhu, and Tao Mei. Vrr-vg: Refocusing visually-relevant relationships. In *International Conference on Computer Vision (ICCV)*, pages 10403–10412, 2019. 14

[101] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 14, 41

[102] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 14, 31, 42, 43

[103] Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 14, 26, 31, 41, 42, 43

[104] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *European Conference on Computer Vision*, pages 793–811. Springer, 2018. 14, 63

[105] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018.

[106] Xu Yang, Hanwang Zhang, and Jianfei Cai. Deconfounded image captioning: A causal retrospect. *arXiv preprint arXiv:2003.03923*, 2020. 14, 17

[107] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 14, 26, 41

[108] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980, 2018. 14, 63

[109] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1965–1973, 2017. 14

[110] Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 7032–7042, 2017. 14

[111] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, pages 1567–1578, 2019. 15, 20, 80, 82, 89, 91, 96, 97, 100

[112] He Wang, Feixiang He, Zhexi Peng, Yong-Liang Yang, Tianjia Shao, Kun Zhou, and David Hogg. Understanding the robustness of skeleton-based action recognition under adversarial attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 15, 109

[113] Gege Qi, Lijun Gong, Yibing Song, Kai Ma, and Yefeng Zheng. Stabilized medical image attacks. In *International Conference on Learning Representations (ICLR)*, 2021.

[114] Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. Adversarial attacks on medical machine learning. *Science*, 2019.

[115] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *International Conference on Computer Vision (ICCV)*, 2017.

[116] Chong Xiang, Charles R Qi, and Bo Li. Generating 3d adversarial point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[117] Moustapha Cisse, Yossi Adi, Natalia Neverova, and Joseph Keshet. Houdini: Fooling deep structured prediction models. *arXiv*, 2017.

[118] Nicholas Carlini and David Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *Security and Privacy Workshops (SPW)*, 2018.

[119] Sandy Huang, Nicolas Papernot, Ian Goodfellow, Yan Duan, and Pieter Abbeel. Adversarial attacks on neural network policies. *arXiv*, 2017. 15, 109

[120] Yunfeng Diao, Tianjia Shao, Yong-Liang Yang, Kun Zhou, and He Wang. Basar:black-box attack on skeletal action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 15, 16, 109

[121] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[122] Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. Adversarial examples in the physical world, 2016. 110, 111

[123] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *EuroS&P*, 2016.

[124] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Symposium on Security and Privacy (SP)*. IEEE, 2017.

[125] Tianhang Zheng, Changyou Chen, and Kui Ren. Distributionally adversarial attack. In *AAAI*, 2019. 15, 109, 110

[126] Iacopo Masi, Yue Wu, Tal Hassner, and Prem Natarajan. Deep face recognition: A survey. In *2018 31st SIBGRAPI conference on graphics, patterns and images (SIBGRAPI)*, pages 471–478. IEEE, 2018. 15

[127] Jessica Van Brummelen, Marie O'Brien, Dominique Gruyer, and Homayoun Najjaran. Autonomous vehicle perception: The technology of today and tomorrow. *Transportation research part C: emerging technologies*, 89:384–406, 2018. 15

[128] Yingzhen Li, John Bradshaw, and Yash Sharma. Are generative classifiers more robust to adversarial attacks? In *International Conference on Machine Learning (ICML)*, 2019. 15, 21

[129] Xinshuai Dong, Hong Liu, Rongrong Ji, Liujuan Cao, Qixiang Ye, Jianzhuang Liu, and Qi Tian. Api-net: Robust generative classifier via a single discriminator. In *European Conference on Computer Vision(ECCV)*, 2020. 15, 21, 109, 114

[130] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning (ICML)*, 2019. 15, 21, 109, 114, 125

[131] Manish Vuyyuru Reddy, Andrzej Banburski, Nishka Pant, and Tomaso Poggio. Biologically inspired mechanisms for adversarial robustness. *NeurIPS*, 2020. 15, 117

[132] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018. 15, 21, 97, 109, 114, 125

[133] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, 2018. 16, 105, 117, 125, 126

[134] Graham Dunn, Richard Emsley, Hanhua Liu, Sabine Landau, Jonathan Green, Ian White, and Andrew Pickles. Evaluation and validation of social and psychological markers in randomised trials of complex interventions in mental health: a methodological research programme. *NIHR Journals Library*, 2015. 16, 60

[135] Moshe Bar. Visual objects in context. *Nature Reviews Neuroscience*, 2004. 16, 23

[136] Irving Biederman, Robert J Mezzanotte, and Jan C Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 1982.

[137] Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 2007. 16, 23

[138] Antonio Torralba, Alexei A Efros, et al. Unbiased look at dataset bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011. 16

[139] 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*, 2019. 17, 75

[140] Yingwei Li, Yi Li, and Nuno Vasconcelos. Resound: Towards action recognition without representation bias. In *European Conference on Computer Vision(ECCV)*, 2018.

[141] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[142] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 2009. 49, 75

[143] Evgeny Burnaev, Pavel Erofeev, and Artem Papanov. Influence of resampling on accuracy of imbalanced classification. In *ICMV*, 2015. 69

[144] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *International Conference on Computer Vision (ICCV)*, 2017. 17, 20, 49, 69, 70, 75, 91, 97, 99

[145] Yiru Wang, Weihao Gan, Jie Yang, Wei Wu, and Junjie Yan. Dynamic curriculum learning for imbalanced data classification. In *International Conference on Computer Vision (ICCV)*, 2019. 17, 20

[146] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 17, 20

[147] Ishan Misra, C Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 17, 48, 62

[148] Remi Cadene, Corentin Dancette, Hedi Ben-younes, Matthieu Cord, and Devi Parikh. Rubi: Reducing unimodal biases in visual question answering. *arXiv preprint arXiv:1906.10169*, 2019. 17

[149] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 17, 18, 19, 50, 52, 55, 58, 81, 82, 85, 87, 91, 109, 110

[150] David P MacKinnon, Amanda J Fairchild, and Matthew S Fritz. Mediation analysis. *Annu. Rev. Psychol.*, 2007. 17, 60

[151] Luke Keele. The statistics of causal inference: A view from political methodology. *Political Analysis*, 2015.

[152] Lorenzo Richiardi, Rino Bellocco, and Daniela Zugna. Mediation analysis in epidemiology: methods, interpretation and bias. *International journal of epidemiology*, 2013. 17, 60

[153] Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. Two causal principles for improving visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 17

[154] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xiansheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. In *NeurIPS*, 2020. 109

[155] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. *NeurIPS*, 2020.

[156] Yulei Niu, Kaihua Tang, Hanwang Zhang, Zhiwu Lu, Xian-Sheng Hua, and Ji-Rong Wen. Counterfactual vqa: A cause-effect look at language bias. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 17

[157] Xinting Hu, Kaihua Tang, Chunyan Miao, Xian-Sheng Hua, and Hanwang Zhang. Distilling causal effect of data in class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 17

[158] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 2021. 17

[159] Niki Kilbertus, Giambattista Parascandolo, and Bernhard Schölkopf. Generalization in anti-causal learning. *arXiv preprint arXiv:1812.00524*, 2018. 17

[160] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *SSST-8*, 2014. 19

[161] Jihun Choi, Kang Min Yoo, and Sang-goo Lee. Learning to compose task-specific tree structures. In *AAAI*, 2018. 19

[162] Myung Jin Choi, Antonio Torralba, and Alan S Willsky. A tree-based context model for object recognition. *TPAMI*, 2012. 19

[163] Tomasz Malisiewicz and Alyosha Efros. Beyond categories: The visual memex model for reasoning about object relationships. In *NIPS*, 2009. 19

[164] Richard Socher, Cliff C Lin, Chris Manning, and Andrew Y Ng. Parsing natural scenes and natural language with recursive neural networks. In *International Conference on Machine Learning (ICML)*, 2011. 19

[165] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002. 20

[166] Chris Drummond, Robert C Holte, et al. Class imbalance and cost sensitivity: why under-sampling beats over-sampling. In *Workshop on learning from imbalanced datasets II*, volume 11, pages 1–8. Citeseer, 2003.

[167] Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *European Conference on Computer Vision(ECCV)*, pages 467–482. Springer, 2016. 80, 81

[168] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the limits of weakly supervised pretraining. In *European Conference on Computer Vision(ECCV)*, 2018. 80, 81

[169] Xinting Hu, Yi Jiang, Kaihua Tang, Jingyuan Chen, Chunyan Miao, and Hanwang Zhang. Learning to segment the tail. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 20

[170] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9268–9277, 2019. 20, 80, 97, 100

[171] Salman H Khan, Munawar Hayat, Mohammed Bennamoun, Ferdous A Sohel, and Roberto Togneri. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 2017. 20, 80

[172] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. In *Advances in Neural Information Processing Systems*, 2019. 20

[173] Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. *arXiv preprint arXiv:1803.09050*, 2018. 20

[174] Muhammad Abdullah Jamal, Matthew Brown, Ming-Hsuan Yang, Liqiang Wang, and Boqing Gong. Rethinking class-balanced methods for long-tailed visual recognition from a domain adaptation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 20

[175] Yu Li, Tao Wang, Bingyi Kang, Sheng Tang, Chunfeng Wang, Jintao Li, and Jiashi Feng. Overcoming classifier imbalance for long-tail object detection with balanced group softmax. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 20

[176] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *European Conference on Computer Vision(ECCV)*, 2018. 24, 25

[177] Damien Teney, Lingqiao Liu, and Anton van den Hengel. Graph-structured representations for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 24

[178] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 25

[179] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Kate Saenko. Learning to reason: End-to-end module networks for visual question answering. In *International Conference on Computer Vision (ICCV)*, 2017. 26, 32

[180] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 33

[181] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 1992. 26, 32

[182] Robert Clay Prim. Shortest connection networks and some generalizations. *Bell System Technical Journal*, 1957. 28

[183] Yan Zhang, Jonathon Hare, and Adam Prügel-Bennett. Learning to count objects in natural images for visual question answering. In *International Conference on Learning Representations (ICLR)*, 2018. 32, 41, 42, 43

[184] Yang Shi, Tommaso Furlanello, Sheng Zha, and Animashree Anandkumar. Question type guided attention in visual question answering. In *European Conference on Computer Vision(ECCV)*, 2018. 32

[185] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *NIPS*, 2014. 33

[186] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision(ECCV)*, 2016. 36, 37, 39, 48, 49, 63

[187] Seong Jae Hwang, Sathya N Ravi, Zirui Tao, Hyunwoo J Kim, Maxwell D Collins, and Vikas Singh. Tensorize, factorize and regularize: Robust visual relationship learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 36, 39

[188] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *European Conference on Computer Vision(ECCV)*, 2018. 36, 39

[189] Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, and Ahmed M Elgammal. Relationship proposal networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 36

[190] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *International Conference on Computer Vision (ICCV)*, 2017. 41, 43

[191] Jin-Hwa Kim, Kyoung-Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. Hadamard product for low-rank bilinear pooling. In *International Conference on Learning Representations (ICLR)*, 2016. 41, 43

[192] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *European Conference on Computer Vision(ECCV)*, 2018. 47, 63

[193] Damien Teney, Lingqiao Liu, and Anton van den Hengel. Graph-structured representations for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 47, 63

[194] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 47, 63

[195] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 47

[196] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018. 47

[197] Drew A Hudson and Christopher D Manning. Learning by abstraction: The neural state machine. *NeurIPS*, 2019. 47

[198] Long Chen, Hanwang Zhang, Jun Xiao, Xiangnan He, Shiliang Pu, and Shih-Fu Chang. Counterfactual critic multi-agent training for scene graph generation. In *International Conference on Computer Vision (ICCV)*, 2019. 49, 61, 65

[199] Ji Zhang, Kevin J Shih, Ahmed Elgammal, Andrew Tao, and Bryan Catanzaro. Graphical contrastive losses for scene graph parsing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 49

[200] Nicole Van Hoeck, Patrick D Watson, and Aron K Barbey. Cognitive neuroscience of human counterfactual reasoning. *Frontiers in human neuroscience*, 2015. 49

[201] Tyler VanderWeele. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015. 50, 57, 58

[202] Judea Pearl. Direct and indirect effects. In *Proceedings of the 17th conference on uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2001. 58, 81, 82, 85

[203] Tyler J VanderWeele. A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology (Cambridge, Mass.)*, 2013. 50, 57, 58, 85

[204] Neal J Roese. Counterfactual thinking. *Psychological bulletin*, 1997. 55

[205] James M Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 1992. 58

[206] Brayden G King. A political mediation model of corporate response to social movement activism. *Administrative Science Quarterly*, 2008. 60

[207] Varun Manjunatha, Nirat Saini, and Larry S Davis. Explicit bias discovery in visual question answering models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 63

[208] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 64

[209] Sebastian Schuster, Ranjay Krishna, Angel Chang, Li Fei-Fei, and Christopher D Manning. Generating semantically precise scene graphs from textual descriptions for improved image retrieval. In *Proceedings of the fourth workshop on vision and language*, 2015. 64, 66, 68

[210] Azriel Rosenfeld and Mark Thurston. Edge and curve detection for visual scene analysis. *IEEE Transactions on computers*, 1971. 65

[211] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. Bilinear attention networks. In *Advances in Neural Information Processing Systems*, 2018. 66, 68

[212] Francisco Massa and Ross Girshick. maskrcnn-benchmark: Fast, modular reference implementation of Instance Segmentation and Object Detection algorithms in PyTorch, 2018. 66

[213] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 69

[214] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 81

[215] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4): 669–688, 1995. 81, 86

[216] *SGD implementation in PyTorch*, 2021. `https://pytorch.org/docs/stable/_modules/torch/optim/sgd.html`. 83, 85

[217] Douglas C Montgomery and George C Runger. *Applied statistics and probability for engineers*. John Wiley & Sons, 2010. 85

[218] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 87, 136

[219] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 2006. 87

[220] Peter C Austin. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 2011. 87

[221] Judea Pearl. On the consistency rule in causal inference: axiom, definition, assumption, or theorem? *Epidemiology*, 21(6):872–875, 2010. 88

[222] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *UAI*, 2009. 89

[223] Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *International Conference on Learning Representations (ICLR)*, 2019. 95

[224] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 98

[225] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open mmlab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019. 98

[226] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*. PMLR, 2020. 105, 108, 125, 126, 131

[227] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *arxiv*, 2020. 105

[228] Jiequan Cui, Shu Liu, Liwei Wang, and Jiaya Jia. Learnable boundary guided adversarial training. *arXiv preprint arXiv:2011.11164*, 2020. 105

[229] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *Workshop of ICLR*, 2018. 106

[230] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *NeurIPS*, 2018. 106

[231] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 106

[232] Sander Greenland. An introduction to instrumental variables for epidemiologists. *International journal of epidemiology*, 2000. 107

[233] Yonggang Zhang, Mingming Gong, Tongliang Liu, Gang Niu, Xinmei Tian, Bo Han, Bernhard Schölkopf, and Kun Zhang. Adversarial robustness through the lens of causality. *arXiv preprint arXiv:2106.06196*, 2021. 109

[234] Chao-Han Huck Yang, Yi-Chieh Liu, Pin-Yu Chen, Xiaoli Ma, and Yi-Chang James Tsai. When causal intervention meets adversarial examples and image masking for deep neural networks. In *ICIP*. IEEE, 2019. 109

[235] Harvineet Singh, Shalmali Joshi, Finale Doshi-Velez, and Himabindu Lakkaraju. Learning under adversarial and interventional shifts. *arXiv preprint arXiv:2103.15933*, 2021. 109

[236] Jason A. Roy, 2020. URL `https://www.coursera.org/lecture/crash-course-in-causality/conditional-independence-d-separation-CGNIV`. 109

[237] Kui Ren, Tianhang Zheng, Zhan Qin, and Xue Liu. Adversarial attacks and defenses in deep learning. *Engineering*, 2020. 110

[238] Florian Tramer, Jens Behrmann, Nicholas Carlini, Nicolas Papernot, and Jorn-Henrik Jacobsen. Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations. In *International Conference on Machine Learning (ICML)*, 2020. 110

[239] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2018. 111, 125

[240] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *International Conference on Learning Representations (ICLR)*, 2018.

[241] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Workshop on artificial intelligence and security*, 2017. 111

[242] Guanlin Li, Shuya Ding, Jun Luo, and Chang Liu. Enhancing intrinsic adversarial robustness via feature pyramid decoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 114

[243] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *CVPR*, 2017. 114

[244] Chaoning Zhang, Philipp Benz, Tooba Imtiaz, and In So Kweon. Understanding adversarial examples from the mutual influence of images and perturbations. In *CVPR*, 2020. 114

[245] Michael Baiocchi, Jing Cheng, and Dylan S Small. Instrumental variable methods for causal inference. *Statistics in medicine*, 2014. 115

[246] Zijian Guo and Dylan S Small. Control function instrumental variable estimation of nonlinear causal effect models. *The Journal of Machine Learning Research*, 2016. 115

[247] Roger J Bowden and Darrell A Turkington. *Instrumental variables*. Cambridge university press, 1990. 115

[248] Helga Kolb, Eduardo Fernandez, and Ralph Nelson. Webvision: the organization of the retina and visual system. *book*, 1995. 117

[249] Jiawei Du, Hanshu Yan, Vincent YF Tan, Joey Tianyi Zhou, Rick Siow Mong Goh, and Jiashi Feng. Rain: A simple approach for robust and accurate image classification networks. *arXiv preprint arXiv:2004.14798*, 2020. 117, 120

[250] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. Causalvae: disentangled representation learning via neural structural causal models. In *CVPR*, 2021. 119, 120

[251] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017. 119, 120

[252] Michael Land. Eye movements in man and other animals. *Vision research*, 2019. 121

[253] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *NeurIPS*, 2016. 124

[254] Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. *ICLR*, 2021. 124

[255] Sravanti Addepalli, Arya Baburaj, Gaurang Sriramanan, and R Venkatesh Babu. Towards achieving adversarial robustness by enforcing feature consistency across bit planes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 125

[256] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *CVPR Workshops*, 2020. 126

[257] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *ECCV*, 2020. 126

[258] Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *ICML*, 2020. 126

[259] Jonathan Uesato, Brendan O'donoghue, Pushmeet Kohli, and Aaron Oord. Adversarial risk and the dangers of evaluating against weak attacks. In *International Conference on Machine Learning (ICML)*, 2018. 130

[260] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations (ICLR)*, 2015. 130

[261] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 130

[262] BBC News. Facebook apology as ai labels black men 'primates', 2021. URL `https://www.bbc.com/news/technology-58462511`. 133

[263] NBC News. Self-driving uber car that hit and killed woman did not recognize that pedestrians jaywalk, 2019. URL `https://www.nbcnews.com/tech/tech-news/self-driving-uber-car-hit-killed-woman-did-not-recognize-n1079281`. 133

[264] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*, 2020. 137