Formatting Instructions For NeurIPS 2022

Anonymous Author(s)
Affiliation
Address
email

Abstract

The abstract paragraph should be indented 1/2 inch (3 picas) on both the left- and right-hand margins. Use 10 point type, with a vertical spacing (leading) of 11 points. The word **Abstract** must be centered, bold, and in point size 12. Two line spaces precede the abstract. The abstract must be limited to one paragraph.

1 Constraint-based causal discovery algorithm

Constraint-based causal discovery methods determine the causal structure through statistical dependencies between variables. Independence tests and Conditional Independence (CI) tests are key tools for testing statistical dependencies. As a start, we first give a brief introduction to them.

Let X, Y, Z be the random variables, P_X, P_Y, P_Z be their probability density functions and P_{XY} be the joint probability function of (X, Y). Then $X \perp\!\!\!\perp Y$ is equivalent to $P_{XY} = P_X P_Y$. From an information-theoretic point of view, the independence of X, Y is equivalent to the mutual information $I(X, Y)$ between them being zero. Conditional Independence tests are more general than independence tests. Let $P_{X|Z}, P_{Y|Z}$ be the conditional probability of X, Y and $P_{XY|Z}$ be the conditional joint probability function of (X, Y) with Z as the conditional variable, then $X \perp\!\!\!\perp Y|Z$ is equivalent to $P_{XY|Z} = P_{X|Z} P_{Y|Z}$. For practical purposes, conditional independence tests are usually written as hypothesis tests:

$$\mathcal{H}_0 : X \perp\!\!\!\perp Y|Z \quad \text{versus} \quad \mathcal{H}_1 : X \not\perp\!\!\!\perp Y|Z.$$

where we refer to $\mathcal{H}_0$ as the null hypothesis and $\mathcal{H}_1$ as the corresponding alternative hypothesis. The independence test can be written as a hypothesis test by simply replacing the condition set with the empty set. Next, we describe some specific (conditional) independence testing methods.

Independence Test According to the type of data, the data can be divided into discrete data and continuous data. For discrete (categorical) data, the commonly used independent tests are $\mathcal{F}$-test [49] and Chi-squared test [14]. These type of test methods determine independence by the specific distribution (e.g. $\mathcal{F}$-distribution, χ^2-distribution) of the statistic under the null hypothesis. Independence testing for the continuous variables is more challenging than for the discrete case. The Pearson correlation coefficient [8] is often used to measure the correlation between variables. This coefficient is widely used but reflects only linear dependence. In order to measure richer dependence, a class of distance correlation [47] metrics was proposed. Such methods detect nonlinear dependence between variables by means of different metric functions (e.g. energy distance [34]). After that, a class of kernel-based independence testing [6, 18, 17, 16] was proposed. These methods are mainly based on the framework proposed by Rényi [33] to measure the nonlinear dependence of variables by sufficiently adequate mappings under function classes. Under the reproducing kernel Hilbert space (RKHS) [9] space, the kernel function is defined as a distance metric induced by the inner product of the feature mapping. Under the RKHS space, the kernel function is defined as a distance metric induced by the inner product of feature mappings. One widely used class of kernel-based independence tests is the Hilbert Schmidt Independence Criterion (HSIC) [16], which measures dependence by the squared *Hilbert-Schmidt norm* induced by the cross-covariance operators in

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

the RKHS space. From the perspective of distance metric, HSIC can be regarded as the squared Maximum Mean Discrepancy [15] (MMD) distance between distribution $P_X P_Y$ and P_{XY}. After that, the random dependence coefficient [25] (RDC) was proposed. Compared to the kernel-based method, RDC is computationally efficient. This method ensures the marginal invariance by copula transformation, and measures the dependence between variables by maximizing the correlation under random projection.

Conditional Independence Test Conditional independent tests are generally more difficult than independent tests due to the hardness of estimating the conditional density distribution compared to the marginal distribution. A class of metric-based CI test [45] employs a number of kernel smoothers to estimate conditional characteristic functions. This type of kernel smoothing estimation has a large computational cost when the condition set is high-dimensional. Another widely used method is kernel-based conditional independence testing such as KCIT [57, 58]. KCIT is based on the partial association framework proposed by Daudin [12] and uses conditional cross-correlation operators to identify conditional independence. KCIT is easy to implement in practice and works well, but due to its kernel regression steps (e.g. kernel ridge regression and Gaussian regression), the computational complexity increases rapidly as the dimension of the condition set grows. Later, the approximate kernel-based method RCIT [44] was proposed. RCIT uses random fourier features to approximate the Gaussian kernel, resulting in an improvement in the computational efficiency of KCIT. Another class of regression-based [40, 56] methods for testing conditional independence. ReCIT [56] tests conditional independence by measuring the independence between two residuals $X - \mathbf{E}[X|Z]$ and $Y - \mathbf{E}[Y|Z]$, where the two expectation terms were estimated by regression. Another class of methods [13, 7, 41, 35, 39, 27] obtains the distribution of the statistic under the null hypothesis by estimating the conditional density function or conditional mutual information, followed by a hypothesis test to determine conditional independence. The permutation-based method [13] obtains resampled samples (X, PY, Z) of factorized distribution $P_{X|Z} P_{Y|Z} P_Z$ by performing permutations on the samples that satisfy a specific structure ($PZ \approx Z$). Some other methods [7, 41], using generative models (e.g. generative adversarial network) to estimate the conditional density. The method [35, 27] uses k-nearest-neighbor (KNN) to obtain factorized distribution or estimate Kullback–Leibler (KL) divergence by classification to estimate conditional mutual information for judging conditional independence. Some model-based method [39] introduce powerful models to discriminate conditional independence by classifying to identifythe difference in distribution between $P_{X|Z} P_{Y|Z}$ and $P_{XY|Z}$.

PC Algorithm We describe how to identify causal networks using (conditional) independence tests. In 1990, the well-known IC (Inductive Causation) algorithm [43] and the PC (Peter-Clark) algorithm [42] were proposed. The PC algorithm starts with a fully connected undirected graph and consists of three key steps. In the first step, the skeleton of the causal graph is obtained by performing an edge deletion operation according to the results of the independence and conditional independence tests. In the second step, the orientation of some edges is determined based on the V-structure. In the third step, constraint propagation is performed based on structural constraints such as acyclicity to determine the direction of some of the remaining undirected edges.

2 Causal function model-based causal discovery algorithm

3 Causal Inference on grneralization

Generalizability is a critical problem in deep learning. Currently, many deep learning-based researches have achieved good performance on various tasks, but they assume the training dataset and test dataset are independent identically distribution. However, in many real tasks, this assumption is not valid. Take the famous image classification dataset ImageNet as an example, it has two popular versions, ImageNet-1k and Imagenet-21k. The previous one has smaller samples and categories, but the labels have better quality. ImageNet-21k is very big and includes many categories, however, the number of categories is still far from the real situation, and the distribution of samples is not the same as in the real environment. With the above shortcomings, algorithms learned from such datasets will contain biases from the dataset itself, which makes the algorithm performs well on a trained dataset, but is difficult to use on other datasets, as their performance may drop dramatically. Instead of improving performance when training and testing on the same dataset, many works are trying to improve the generalizability of the model and make the model still works well when training and testing data are out of distribution. Since causal inference provides a principled framework for modeling structural

invariances, it is more interpretable and easier to identify selection bias. Therefore, using causal inference to improve model generalizability is a hot research area. Main research directions for causal inference on generalization are the following: out-of-distribution detection (OOD), open-set recognition (OSD), few-shot learning (FSL), and zero-shot learning (ZSL). In this section, we mainly discuss applications of above research directions on computer vision (CV).

Out-of-distribution Detection and Open-set Recognition. For conventional classification algorithms, they will always give a label to any input sample. However, the category of a sample may not exist in training, e.g. a model is trained to classify cats and dogs, but when an image of a bird is fed to the model, it will always give a wrong result, no matter the result is cat or dog. In OOD, the algorithm should recognize if the input does not belong to any category. For OOD generalization in CV, there are two major types of domain generalization: the multi-source domain generation and the single-source domain generation. In multi-source domain generation, the model knows the domain index [1, 4, 23]. In single-source domain generation, the model treats all input is extracted from the same single domain [3, 20, 32, 50]. As it lacks domain information, the problem is rather challenging. Causal inference is a novel principle for OOD detection, as it is more interpretable. OSR problem is very similar to OOD, the major difference is that OOD only needs to identify whether input does not belong to any category, i.e. binary classification problem. But OSR needs to give the correct category if the input is not an outlier, so it is more difficult than OOD. A simple implementation for OSR first detects OOD, then followed by a normal classifier. However, this two-stage method not always performs well, as it cannot use the detailed classification results. As the result, some methods predict OOD and the category for in-distribution inputs jointly [36, 19, 37]. [26] focuses on out-of-distribution generalization problem, and extract the robust features from observational data through both causal and deep representations with some mild assumptions. [31] studied out-of-distribution on Visual Question Answering (VQA) problem, it designed a two-stage learning method, doing causal inference in the first stage, and distillation in the second stage. [54] solves open-set recognition problem using consistency rule to improve the counterfactual faithfulness, if the input doesn't fit the consistency rule, the input is considered an outlier.

Few-shot Learning and Zero-shot Learning. In FSL and ZSL, algorithms will be trained on training categories and tested on categories that are different from training categories. They can also receive attribute descriptions of inputs, and predict unseen categories based on these attributes [22, 53]. In FSL, algorithms can fine-tune results on a very limited number of new categories (e.g. three samples). And in ZSL, algorithms should predict the category directly, without any additional training. The main difference compared with OOD and OSR is that FSL and ZSL do not need to distinguish outliers (all test inputs are outliers), but they should predict their categories. To solve the problem, instead of memorizing features of existing categories, algorithms should learn a generalized attributes representation, so they can classify unknown categories. Some algorithms solve the problem by inferring a sample's attribute and finding the closest match [2, 11, 21]. Another way to solve the problem is generating features using the attributes [24, 28, 30]. Recently, many algorithms use causal inference to solve FSL and ZSL. [55] proposed interventional few-shot learning, it begins with a structural causal model, and uses causal intervention to solve the FSL problem. [5] considered the compositional zero-shot recognition problem and solve it with a causal perspective. It described a new embedding-based architecture that infers causally stable representations for compositional recognition.

4 Causal Inference in Solving the Fairness of Recommender Systems

The wide application of recommender systems in the industry has brought far-reaching significance and great practical value. However, Recommender systems usually amplify the biases in the data [51]. The model learned from historical interactions with imbalanced item distribution will amplify the imbalance by over-recommending items from the majority groups. Addressing this issue is essential for a healthy ecosystem of recommendation in the long run. The latest work can be roughly summarize as the following four aspects.

Debiasing Learning and Evaluation Most data for evaluating and training recommender systems is subject to selection biases, either through self-selection by the users or through the actions of the recommendation system itself. [38] provide a principled approach to handle selection biases by adapting models and estimation techniques from causal inference.this is one of the early papers that

introduced causality into recommender systems, and it is very enlightening. [46] propose several debiasing algorithms during this chain of events, and evaluate how these algorithms impact the predictive behavior of the recommender systems, as well as trends in the popularity distribution of items over time. they also propose a novel blind-spot-aware matrix factorization (MF) algorithm to debias the recommender systems and achieved a higher debiasing effect on recommendations.[60] uses causal embedding to decouple user interests and popular products, and use the collision effect in causal inference to de-bias.

Selection Bias To bridge the gap between the final recommendation objective and the classical setup [10] propose a new domain adaptation algorithm that learns from logged data containing outcomes from a biased recommendation policy and predicts recommendation outcomes according to random exposure. Using the method of causal inference to solve the problem of data selection bias in recommender systems has great enlightening significance. [29] pay attention to correcting for selection bias, which occurs because clicked documents are reflective of what documents have been shown to the user in the first place. they propose new counterfactual approaches which adapt Heckman's two-stage method and accounts for selection and position bias in LTR systems. Their empirical evaluation shows that our proposed methods are much more robust to noise and have better accuracy compared to existing unbiased LTR algorithms, especially when there is moderate to no position bias.

Popularity Bias [61] theoretically prove that a popular choice of contrastive loss is equivalent to reducing the exposure bias via inverse propensity weighting. Based on the theoretical discovery, they design CLRec to improve DCG in terms of fairness, effectiveness and efficiency in Recommender systems.[51] propose a Deconfounded Recommender System (DecRS), which models the causal effect of user representation on the prediction score. The key to eliminating the impact of the confounder lies in backdoor adjustment, which is however difficult to do due to the infinite sample space of the confounder. For this challenge, they contribute an approximation operator for backdoor adjustment which can be easily plugged into most recommender models.[59] studies an unexplored problem in recommendation — how to leverage popularity bias to improve the recommendation accuracy. The key lies in two aspects: how to remove the bad impact of popularity bias during training, and how to inject the desired popularity bias in the inference stage that generates top-K recommendations. And they propose a new training and inference paradigm for recommendation named Popularity-bias Deconfounding and Adjusting (PDA). It removes the confounding popularity bias in model training and adjusts the recommendation score with desired popularity bias via causal intervention.

Counterfactual Fairness and Explainable Recommendation the author of[52] explore the popularity bias issue from a novel and fundamental perspective — cause-effect. They describes the causal relationship between some variables in the recommender system from the perspective of causal inference, and solves the influence of Popularity Bias on the model from the perspective of counterfactual reasoning. To eliminate popularity bias, it is essential to answer the counterfactual question. So they formulate a causal graph to describe the important cause-effect relations in the recommendation process and use counterfactual reasoning during model testing to eliminate the influence of popularity on recommendations. By providing explanations for users and system designers to facilitate better understanding and decision making, explainable recommendation has been an important research problem. [48] propose Counterfactual Explainable Recommendation(CountER), which takes the insights of counterfactual reasoning from causal inference for explainable recommendation.

References

[1] Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014.

[2] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1425–1438, 2015.

[3] Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H Falk, and Ioannis Mitliagkas. Generalizing to unseen domains via distribution matching. *arXiv preprint arXiv:1911.00804*, 2019.

[4] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[5] Yuval Atzmon, Felix Kreuk, Uri Shalit, and Gal Chechik. A causal view of compositional zero-shot recognition. *Advances in Neural Information Processing Systems*, 33:1462–1473, 2020.

[6] Francis R Bach and Michael I Jordan. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48, 2002.

[7] Alexis Bellot and Mihaela van der Schaar. Conditional independence testing using generative adversarial networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[8] Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

[9] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[10] Stephen Bonner and Flavian Vasile. Causal embeddings for recommendation. In *Proceedings of the 12th ACM conference on recommender systems*, pages 104–112, 2018.

[11] Long Chen, Hanwang Zhang, Jun Xiao, Wei Liu, and Shih-Fu Chang. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1043–1052, 2018.

[12] Ingrid Daubechies. *Ten lectures on wavelets*. SIAM, 1992.

[13] Gary Doran, Krikamol Muandet, Kun Zhang, and Bernhard Schölkopf. A permutation-based kernel conditional independence test. In *UAI*, pages 132–141. Citeseer, 2014.

[14] Priscilla E Greenwood and Michael S Nikulin. *A guide to chi-squared testing*, volume 280. John Wiley & Sons, 1996.

[15] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[16] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[17] Arthur Gretton, Ralf Herbrich, and Alexander J Smola. The kernel mutual information. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 4, pages IV–880. IEEE, 2003.

[18] Arthur Gretton, Alexander Smola, Olivier Bousquet, Ralf Herbrich, Andrei Belitski, Mark Augath, Yusuke Murayama, Jon Pauls, Bernhard Schölkopf, and Nikos Logothetis. Kernel constrained covariance for dependence measurement. In *International Workshop on Artificial Intelligence and Statistics*, pages 112–119. PMLR, 2005.

[19] Manuel Gunther, Steve Cruz, Ethan M Rudd, and Terrance E Boult. Toward open-set face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 71–80, 2017.

[20] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349, 2021.

[21] Huajie Jiang, Ruiping Wang, Shiguang Shan, and Xilin Chen. Transferable contrastive network for generalized zero-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9765–9774, 2019.

[22] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE conference on computer vision and pattern recognition*, pages 951–958. IEEE, 2009.

[23] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[24] Jingjing Li, Mengmeng Jing, Ke Lu, Zhengming Ding, Lei Zhu, and Zi Huang. Leveraging the invariant side of generative zero-shot learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7402–7411, 2019.

[25] David Lopez-Paz, Philipp Hennig, and Bernhard Schölkopf. The randomized dependence coefficient. *Advances in neural information processing systems*, 26, 2013.

[26] Chengzhi Mao, Kevin Xia, James Wang, Hao Wang, Junfeng Yang, Elias Bareinboim, and Carl Vondrick. Causal transportability for visual recognition. *arXiv preprint arXiv:2204.12363*, 2022.

[27] Sudipto Mukherjee, Himanshu Asnani, and Sreeram Kannan. Ccmi: Classifier based conditional mutual information estimation. In *Uncertainty in artificial intelligence*, pages 1083–1093. PMLR, 2020.

[28] Sanath Narayan, Akshita Gupta, Fahad Shahbaz Khan, Cees GM Snoek, and Ling Shao. Latent embedding feedback and discriminative features for zero-shot classification. In *European Conference on Computer Vision*, pages 479–495. Springer, 2020.

[29] Zohreh Ovaisi, Ragib Ahsan, Yifan Zhang, Kathryn Vasilaky, and Elena Zheleva. Correcting for selection bias in learning-to-rank systems. In *Proceedings of The Web Conference 2020*, pages 1863–1873, 2020.

[30] Ayyappa Pambala, Titir Dutta, and Soma Biswas. Generative model with semantic embedding and integrated classifier for generalized zero-shot learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1237–1246, 2020.

[31] Yonghua Pan, Zechao Li, Liyan Zhang, and Jinhui Tang. Causal inference with knowledge distilling and curriculum learning for unbiased vqa. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(3):1–23, 2022.

[32] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.

[33] Alfréd Rényi. On measures of dependence. *Acta mathematica hungarica*, 10(3-4):441–451, 1959.

[34] Maria L Rizzo and Gábor J Székely. Energy distance. *wiley interdisciplinary reviews: Computational statistics*, 8(1):27–38, 2016.

[35] Jakob Runge. Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 938–947. PMLR, 2018.

[36] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772, 2012.

[37] Walter J Scheirer, Lalit P Jain, and Terrance E Boult. Probability models for open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2317–2324, 2014.

[38] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *international conference on machine learning*, pages 1670–1679. PMLR, 2016.

[39] Rajat Sen, Ananda Theertha Suresh, Karthikeyan Shanmugam, Alexandros G Dimakis, and Sanjay Shakkottai. Model-powered conditional independence test. *Advances in neural information processing systems*, 30, 2017.

[40] Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

[41] Chengchun Shi, Tianlin Xu, Wicher Bergsma, and Lexin Li. Double generative adversarial networks for conditional independence testing. *Journal of Machine Learning Research*, 22(285):1–32, 2021.

[42] Peter Spirtes and Clark Glymour. An algorithm for fast recovery of sparse causal graphs. *Social science computer review*, 9(1):62–72, 1991.

[43] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000.

[44] Eric V Strobl, Kun Zhang, and Shyam Visweswaran. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference*, 7(1), 2019.

[45] Liangjun Su and Halbert White. A consistent characteristic function-based test for conditional independence. *Journal of Econometrics*, 141(2):807–834, 2007.

[46] Wenlong Sun, Sami Khenissi, Olfa Nasraoui, and Patrick Shafto. Debiasing the human-recommender system feedback loop in collaborative filtering. In *Companion Proceedings of The 2019 World Wide Web Conference*, pages 645–651, 2019.

[47] Gábor J Székely and Maria L Rizzo. Partial distance correlation with methods for dissimilarities. *The Annals of Statistics*, 42(6):2382–2412, 2014.

[48] Juntao Tan, Shuyuan Xu, Yingqiang Ge, Yunqi Li, Xu Chen, and Yongfeng Zhang. Counterfactual explainable recommendation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1784–1793, 2021.

[49] ML Tiku. Tables of the power of the f-test. *Journal of the American Statistical Association*, 62(318):525–539, 1967.

[50] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019.

[51] Wenjie Wang, Fuli Feng, Xiangnan He, Xiang Wang, and Tat-Seng Chua. Deconfounded recommendation for alleviating bias amplification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1717–1725, 2021.

[52] Tianxin Wei, Fuli Feng, Jiawei Chen, Ziwei Wu, Jinfeng Yi, and Xiangnan He. Model-agnostic counterfactual reasoning for eliminating popularity bias in recommender system. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1791–1800, 2021.

[53] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2251–2265, 2018.

[54] Zhongqi Yue, Tan Wang, Qianru Sun, Xian-Sheng Hua, and Hanwang Zhang. Counterfactual zero-shot and open-set visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15404–15414, 2021.

[55] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. *Advances in neural information processing systems*, 33:2734–2746, 2020.

[56] Hao Zhang, Shuigeng Zhou, and Jihong Guan. Measuring conditional independence by independent residuals: Theoretical results and application in causal discovery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[57] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. 804–813. corvallis, or, 2011.

[58] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775*, 2012.

[59] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. Causal intervention for leveraging popularity bias in recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 11–20, 2021.

[60] Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. Disentangling user interest and conformity for recommendation with causal embedding. In *Proceedings of the Web Conference 2021*, pages 2980–2991, 2021.

[61] Chang Zhou, Jianxin Ma, Jianwei Zhang, Jingren Zhou, and Hongxia Yang. Contrastive learning for debiased candidate generation in large-scale recommender systems. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 3985–3995, 2021.

A Appendix

Optionally include extra information (complete proofs, additional experiments and plots) in the appendix. This section will often be part of the supplemental material.