SCIENCE CHINA
Information Sciences

• LETTER •

June 2025, Vol. 68, Iss. 6, 160108:1–160108:2
https://doi.org/10.1007/s11432-024-4347-6

Special Topic: Enabling Techniques and Cutting-Edge Applications of Foundation Models

Trustworthy forgery detection with causal inference

Junxian DUAN[1], Fan JI[2], Yi LI[2], Hao SUN[1] & Ran HE[1*]

[1]*State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*
[2]*National Key Laboratory of Space Integrated Information System, Institute of Software, Chinese Academy of Sciences, Beijing 100190, China*

Received 30 November 2024/Revised 9 February 2025/Accepted 14 March 2025/Published online 21 May 2025

Citation Duan J X, Ji F, Li Y, et al. Trustworthy forgery detection with causal inference. Sci China Inf Sci, 2025, 68(6): 160108, https://doi.org/10.1007/s11432-024-4347-6

Advancements in vision-language models have made the generation of highly realistic visual media, such as images and videos, increasingly widespread. Meanwhile, as the accessibility of the generative foundation models technology increases, the risk of misuse escalates, which could contribute to the spread of fake information. Thus, it is crucial for developing visual media forgery detection models [1]. Recent research primarily focuses on identifying authenticity and fake location, with an emphasis on improving accuracy and generalization. However, it often overlooks the significance of explaining why certain results are classified as fake or real, which is crucial for ensuring reliable traceability and forensic analysis of forgeries. Thus, we aim to develop a trustworthy and explainable forgery detection model.

Causal inference, which is the process of discovering the causal relationship by eliminating spurious correlations introduced by confounding factors, has gained significant attention recently. In forgery detection, spurious correlations and confounding variables can mislead detection models, reducing their reliability. Causal learning provides a principled approach to addressing these challenges by explicitly modeling the influence of confounders and disentangling misleading correlations. This makes causal inference a powerful tool for enhancing the robustness and interpretability of visual media forgery detection. In fake news detection, CLIMB [2] employs a causal framework to mitigate image-text matching bias, while Li et al. [3] utilized soft-prompt learning to integrate textual and numerical covariates for effective confounder representation. Inspired by these, we can leverage causal inference to detect visual media forgery.

We propose a structural causal mode (SCM) to elucidate the inferent objective of forgery detection from a causal perspective, thereby improving the accuracy of vanilla forgery detection methods in estimating the desired causal effect. We aim to enhance the trustworthiness of the detection by providing transparency in the processing of identifying and justifying potential forgeries. Specifically, we introduce a lightweight, feature-decoupled plugin, the forgery bias eliminating module (FBEM), which is designed to be seamlessly integrated into existing systems in a plug-and-play manner. Building on FBEM, we further propose a forgery bias that eliminates the loss function and a conditional mutual information regularization term. These components collectively improve the interpretability and robustness of other methods while ensuring that they are deployed locally with minimal computational cost. The flexibility of our approach enables effortless adoption across various applications, allowing existing models to achieve enhanced performance without requiring significant architectural modifications.

Problem formulation. To comprehensively understand the intrinsic objective of forgery detection methods, we re-evaluate the learning paradigm of forgery detection through a causal framework and derive the corresponding SCM. Specifically, the process of forgery detection can be described as follows. Given the input image X, we utilize a backbone network to obtain the deep feature representation Z, which is then used to predict the label Y via a classification head. Meanwhile, the feature Z can be decomposed into forgery pattern features Z_n and forgery-irrelevant features Z_p. In addition, domain knowledge K, including information such as forgery methods and image semantics, impacts both X and Y. Therefore, we construct the corresponding SCM of the forgery detection process in Figure 1(a). Refer to Appendix A for an SCM introduction.

Examining the SCM presented in Figure 1(a), we can expound the inferent objective of forgery detection from a causal perspective as follows: capturing the causal effect between the input image X and the prediction label Y, which is implemented by computing $P(Y \mid \mathrm{do}(x))$ rather than modeling $P(Y \mid X)$ in vanilla detection approaches [4]. Nevertheless, the undesired backdoor path $X \leftarrow K \rightarrow Y$ renders a gap between $P(Y \mid X)$ and $P(Y \mid \mathrm{do}(x))$. To estimate $P(Y \mid \mathrm{do}(x))$, we can apply the back-door criterion as follows:

$$P(Y \mid \mathrm{do}(X)) = \sum_K P(K \mid X) \cdot P(Y \mid K). \tag{1}$$

However, K as domain knowledge represents a latent variable that cannot be directly measured. The back-door

* Corresponding author (email: rhe@nlpr.ia.ac.cn)

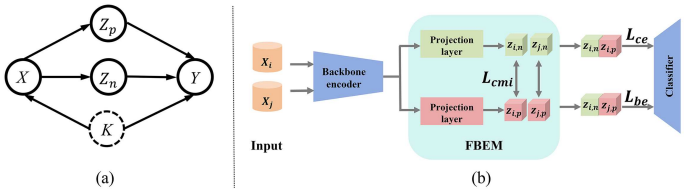

Figure 1 (Color online) (a) The SCM diagram of the input image X, predicted label Y, forgery pattern features Z_n, forgery-irrelevant features Z_p, and domain knowledge K. The dashed circle represents the variable that cannot be measured. (b) The framework of the proposed method.

adjustment may not be sufficient, as we cannot condition on K directly. To disclose this focal issue, with front-door criterion, we perform causal intervention via front-door adjustment with the following formula:

$$P(Y \mid \mathrm{do}(X)) = \sum_{Z_n, Z_p} P(Z_n, Z_p \mid X) \cdot P(Y \mid Z_n, Z_p). \quad (2)$$

Adhering to the front door adjustment formula, the accuracy of the causal effect estimate of X on Y after the intervention is contingent upon the reliability of Z_n and Z_p. Therefore, to improve the accuracy of vanilla forgery detection methods in estimating the desired causal effect, we propose a plug-and-play feature decoupled module and a conditional mutual information regularization term.

Methodology. Since mainstream forgery detection networks typically consist of an encoder and a classifier, we propose FBEM to decouple the features derived from the encoder into Z_n and Z_P using two projection layers, rather than employing additional encoders with a large number of parameters. The overview of the method is shown in Figure 1(b). The two projection layers are represented by φ and ψ, respectively. For a given image x_i, the outputs of φ and ψ can be interpreted as the forgery pattern features $z_{i,n}$ and forgery-irrelevant features $z_{i,p}$, respectively, as expressed by the following equations:

$$z_{i,n} = \varphi(z_i), \quad z_{i,p} = \psi(z_i). \quad (3)$$

Subsequently, we concatenate these two features and use the combined representation as the input for the classifier σ. To ensure φ and ψ can effectively capture the corresponding features, we propose forgery bias eliminating loss $\mathcal{L}_{be}$ to expand the available value set of the adjustment variables. Specifically, after concatenating $z_{i,n}$ with $z_{j,p}$ from other samples, the resulting feature should still be identifiable as belonging to the y_i class.

$$\mathcal{L}_{\mathrm{ce}} = \mathrm{BCELoss}(\sigma([z_{i,n}; z_{i,p}]), y_i), \quad (4)$$

$$\mathcal{L}_{\mathrm{be}} = \mathrm{BCELoss}(\sigma([z_{i,n}; z_{j,p}]), y_i). \quad (5)$$

Furthermore, to make sure that these two features can be decoupled, we formulate a regularization term to minimize the conditional mutual information (CMI) between the forgery pattern features and forgery-irrelevant features:

$$\mathcal{L}_{\mathrm{cmi}} = I(z_n; z_p \mid Y), \quad (6)$$

where $I(\cdot)$ denotes the Shannon mutual information. In practice, we estimate the mutual information using the implementation provided in [5].

$$I(\boldsymbol{z}_n; \boldsymbol{z}_p \mid Y) := \left\| \frac{1}{N} \sum_{i=1}^{N} z_{i,n} \left(z_{i,p} - \sum_{j=1}^{N} \frac{q_j^i}{\sum_{j=1}^{N} q_j^i} z_{j,p} \right) \right\|_1, \quad (7)$$

where $q_j^i = 1$ if and only if $y_i = y_j$; otherwise $q_j^i = 0$.

Overall, the final objective function for training is

$$\mathcal{L} = \mathcal{L}_{\mathrm{ce}} + \alpha \mathcal{L}_{\mathrm{be}} + \beta \mathcal{L}_{\mathrm{cmi}}. \quad (8)$$

Experiments. We evaluate our method on the FaceForensics++, DFDC, and CelebDF-v2 datasets. In Appendix B, the experimental results show that by incorporating the proposed module into Xception and F3Net, the model's performance, in terms of area under curve (AUC), improved significantly by 0.48% and 0.46%, and Accuracy increased by 0.76% and 0.45%, while their FLOPs only increased by 0.31% and 0.26%, respectively. The outcome indicates that the addition of FBEM does not result in a significant increase in model complexity, but rather improves performance without overwhelming the computational cost. Meanwhile, we also conduct ablation experiments and cross-dataset testing, as detailed in Appendixes C and D2, including the visualization results in Appendix D1.

Conclusion and future work. In this study, we present a trustworthy forgery detection scheme based on the structural causal mode. The proposed method introduces a lightweight plug-and-play module, which incorporates forgery bias, eliminating loss, and a conditional mutual information regularization term. The experimental performance demonstrates the effectiveness of the causal learning approach and the lightweight effect of the plugin. The potential of causal inference can be further leveraged by developing more sophisticated and reliable SCMs that better capture the underlying causal structure of forgery detection. Additionally, integrating advanced causal inference techniques, such as instrumental variable methods or counterfactual reasoning, could further improve the generalization of emerging forgery techniques and unseen data.

Acknowledgements This work was partially supported by National Natural Science Foundation of China (Grant Nos. U21B2045, U20A20223, 32341009, 62206277)

Supporting information Appendixes A–E. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Xiao J, Yin Q, Lu W, et al. Deepfake detection based on video flow spectrum feature space (in Chinese). Sci Sin Inform, 2024, 54: 2572–2588
2 Hu L, Chen Z, Zhao Z, et al. Causal inference for leveraging image-text matching bias in multi-modal fake news detection. IEEE Trans Knowl Data Eng, 2023, 35: 11141–11152
3 Li Y C, Lee K, Kordzadeh N, et al. What boosts fake news dissemination on social media? A causal inference view. In: Advances in Knowledge Discovery and Data Mining. Cham: Springer, 2023. 234–246
4 Judea P. Causal Inference in Statistics: A Primer. Hoboken: John Wiley & Sons, 2016
5 Jiang Y B, Veitch V. Invariant and transportable representations for anti-causal domain shifts. In: Proceedings of the International Conference on Neural Information Processing Systems, 2022. 20782–20794