Towards Generalization of Scene Text Tampering Localization via Causal Invariance

Supplementary Material

1. Appendix.A: The Overall Objective Loss Functions

This section supplements details of the overall objective loss functions.

Refined semantic and in-distribution(ID) forensic representation spaces construction is necessary for confounding environment generation. Therefore, we use a lightweight linear layers as the semantic adapter on the pretrained foundation model for further semantic spaces refinement in the downstream task. Meanwhile, we adopt additional lightweight linear layers as the forensic adapter for the ID forensic space construction. Additionally, the association between the latent semantic/forensic representation spaces and the corresponding label spaces is modeled by the semantic and forensic classifiers. In this process, trainable parameters are updated with Eq. (1) and Eq. (2).

$$\mathcal{L}_{\mathcal{S}} = \mathcal{L}_{CE}(X_T, Y_S), \tag{1}$$

where, L_S refers to the loss function for refined semantics; X_T and Y_S are input images and their semantic annotations to indicate the positions of text regions, respectively.

$$\mathcal{L}_{\mathcal{F}} = \mathcal{L}_{CE}(X_T, Y_F), \tag{2}$$

where, L_F refers to the loss function for ID forensics; X_T and Y_F are input images and their 3-class forensic annotations where 0 means background regions, 1 refers to the authentic text regions, and 2 indicates the tampered text regions. Additionally, we leverage **??** for causal invariant representation optimization and rewrite it as

$$\begin{aligned}\mathcal{L}_F^{causal} = &\underbrace{\sum_{e\in\mathcal{E}} \mathcal{L}_{CE}(X_T, Y; e) +}_{\mathcal{L}_F^{CE}} \\ &\underbrace{\sum_{j\in\{at,tt\}} var(\mathcal{L}_{CE}^{e,j}(X_T, Y))\,.}_{\mathcal{L}_F^{Inv}}\end{aligned} \tag{3}$$

where, $\mathcal{L}_F^{CE}$ and $\mathcal{L}_F^{Inv}$ refer to L_{CE} and L_{Inv} in **??**; data from the set of environments, $\mathcal{E}$, include the synthesized interventioal data which are generated from confounding environments in case(1)-(3) and the ID observable data in training dataset.

The overall objective loss function is indicated as

$$\mathcal{L}_{total} = \mathcal{L}_S + \mathcal{L}_F + \mathcal{L}_F^{causal}, \tag{4}$$

where, we utilize three optimizers for independently updating the corresponding parameters. For example, semantic optimizer is used to update the parameters of semantic adapter and semantic classifier.

2. Appendix.B: Experiments for Multi-Type Generalization

The results of our comparative study evaluating multi-type generalization against benchmark methods of natural image tampering localization and text image tampering localization are supplemented in Tab. 1 and Tab. 2. To evaluate the multi-type generalization, we collect the subsets which are manipulated by SRNet[23], DST[16], and DiffSTE[6] as the training sets, and evaluate the testing sets which are tampered with STEFANN[15], MOSTEL[13], AnyText[19], UdiffText[25], and TextDiff[1].

The expansion in the diversity of seen manipulation types enriches the forensic feature space, subsequently improving the performance of previously unseen forensic artifacts. This mechanism offers a notable advantage for the performance of current natural/text image tampering localization models. For example, IML-vit demonstrates the most promising forensic efficacy on ID/OOD testing sets except the sets of AnyText and TextOCR[25]. Meanwhile, for among image tampering localization methods, CAFTB presents the most competitive forensic performance besides the sets of TPIC[22], STEFANN, AnyText, UdiffText, and TextOCR. However, this mechanism is also beneficial for our method, leading to significant progress in OOD performance. For example, the performance achieved on AnyText and TextOCR is superior to that of all compared methods. However, a decline in performance is observed on ID data, such as 79.7% for ID TPIC in single-type setting to 71.4% for ID TPIC in multi-type setting. Furthermore, while the generalizability on OOD sets remains satisfactory in the single-type setting, it diminishes when facing the multi-type setting, such as TextDiff and MOSTEL sets. Additionally, the overall performance of the current tamper localization networks still requires further improvement. Consequently, we intend to pursue continued refinement in our future work.

3. Appendix.C: Ablation Studies

We supplement other ablation studies in this section. If taking a foundation model with the linear probing as the base model, several additional modules are composed of

Table 1. The comparison with natural image tampering localization methods for multi-type generalization.

Methods	ID				OOD					
	TP-IC	SRNet	DST	DiffSTE	STEFANN	MOSTEL	AnyText	UDiffText	TextOCR	TextDiff
MVSS[3]	64.7	65.8	67.1	59.7	42.4	52.3	48.0	58.6	25.5	45.9
PSCC[8]	63.4	64.4	66.2	59.2	46.0	53.5	40.6	57.2	21.6	54.5
Object-Former[21]	41.9	41.7	43.5	39.8	41.1	29.2	28.9	45.3	19.1	28.6
CAT [7]	70.2	71.8	74.7	61.1	40.4	59.9	37.8	57.2	27.9	56.8
IML-vit[9]	**76.0**	**76.0**	**79.3**	**64.0**	**65.2**	**64.4**	49.1	**65.2**	34.5	51.6
Mesorch [26]	73.7	74.2	77.2	63.3	52.0	56.2	48.0	64.0	36.3	57.8
Sparse [18]	74.1	74.6	76.6	61.9	54.3	54.9	46.9	63.2	34.7	**59.5**
Trufor [5]	64.6	64.3	68.0	57.2	44.7	46.1	43.0	55.7	33.5	49.6
Ours	71.4	68.6	70.6	63.0	43.7	47.9	**52.1**	62.7	**42.7**	34.0

Table 2. The comparison with text image tampering localization methods for multi-type generalization.

Methods	ID				OOD					
	TP-IC	SRNet	DST	DiffSTE	STEFANN	MOSTEL	AnyText	UDiffText	TextOCR	TextDiff
DTD[11]	61.9	61.9	67.1	51.3	40.1	44.6	36.4	58.7	28.9	51.1
CAFTB [17]	71.2	**70.5**	**73.9**	**63.8**	38.9	**58.4**	40.8	62.6	34.9	**56.6**
FFDB [2]	72.0	72.3	74.3	57.6	29.4	51.8	37.7	56.9	30.7	48.4
TIFDM [4]	57.1	58.4	57.8	42.2	26.1	41.0	22.7	48.7	25.6	46.1
DAF [12]	10.1	10.2	9.4	9.5	8.6	12.5	9.8	11.8	2.9	4.3
Ours	**71.4**	68.6	70.6	63.0	**43.7**	47.9	**52.1**	**62.7**	**42.7**	34.0

our method. We conduct experiments on these modules to demonstrate the validity of the proposed method in Sec. 3.1. Additionally, we conduct ablation studies about data augmentation, efficient-parameter finetuning techniques, and foundation models in Sec. 3.2, Sec. 3.3, and Sec. 3.4, respectively.

3.1. Modules

In the proposed method, there are three main modules: linear probe, adapter, and casual invariance. To evaluate the validity of each module, we conduct experiments in the three settings: (1) frozen foundation model + linear probe, (2) linear adapters + frozen foundation model + linear probe, (3) confounding + ERM: linear adapters + frozen foundation model + augmented confounding environments + linear probe, and (4) confounding + IRM: linear adapters + frozen foundation model + augmented confounding environments + invariance regularization + linear probe. The performance indicated in Tab. 3 could demonstrate the effectiveness of each module in the proposed method.

3.2. Data Augmentation

Confounding environment generation in the proposed method could also be viewed as a form of data augmentation, specifically one that implements augmentation by intervening on the data from the perspective of causal diagram. We conduct the compared experiment between one random data augmentation technique, manifold mixup[20] (MM) and ours, causally interventional data augmentation in the latent space. The results presented in Tab. 4 could indicate the effectiveness of our confounding environment generation for ID and OOD evaluation. The performance of manifold mixup and our proposed confounding environment generation on the ID datasets dose not differ significantly, but a substantial performance discrepancy could be observed on the OOD data. Under a strategy of random data augmentation, the performance variance across different OOD subsets is high, which highlights the uncontrolled nature of random data augmentation in capturing useful information.

3.3. Parameter-Efficient Adaptation Methods

Parameter-Efficient Adaptation (PEA) methods are the effective techniques to finetune the pretrained foundation models for downstream-task feature adaptation and retaining the favorable properties of the pre-trained space. Throughout this paper, and unless otherwise specified, the exemplified PEA method is light-weight adapter which is applied directly on the representation. There are also strategies that perform Parameter-Efficient Adaptation (PEA)

Table 3. The ablation study of modules for single-type generalization.

Methods	ID		OOD							
	TP-IC	SRNet	DST	STEFANN	MOSTEL	DiffSTE	AnyText	UDiffText	TextOCR	TextDiff
(1)	41.5	38.4	34.3	**32.0**	26.4	25.9	32.5	37.7	15.4	33.9
(2)	74.6	71.8	39.1	11.8	53.7	48.3	**44.6**	47.3	26.6	53.8
(3)	**79.0**	**77.3**	32.8	14.0	65.6	53.6	40.9	51.8	**42.5**	**66.8**
(4)	78.1	77.0	**52.3**	25.0	**69.4**	**58.6**	43.8	**58.1**	41.6	60.6

Table 4. The ablation study of data augmentation methods for single-type generalization. MM means manifold mixup.

Methods	ID		OOD							
	TP-IC	SRNet	DST	STEFANN	MOSTEL	DiffSTE	AnyText	UDiffText	TextOCR	TextDiff
MM	70.2	63.6	16.1	3.8	48.9	36.4	24.9	35.3	25.2	57.6
Ours	**78.1**	**77.0**	**52.3**	**25.0**	**69.4**	**58.6**	**43.8**	**58.1**	**41.6**	**60.6**

fine-tuning on the model parameters, such as SVD in [24].

The fundamental principle of SVD involves decomposing the model parameters into principal components and residual components. The principal components which represent semantic information and are kept frozen during fine-tuning, while the residual components are updated during training. These residual components thus encode information relevant to the downstream task, which, in the context of our work, refers to the forensic information. Based on this decomposition, both the content and the forensic information can be effectively quantified. This methodology bears a striking similarity to the principle underlying quantification approaches based on lightweight adapters which take the features captured by the original weights of foundation model as the content information and take the offset induced by the forensic adapters as the ID type of forensic information. We present the results in Tab. 5. The integration of the proposed causal invariance mechanism led to performance improvements on both light-weight adapters and SVD, which collectively demonstrate the efficacy of our method.

3.4. Foundation Models

Throughout this paper, and unless otherwise specified, our illustrative foundation model is OWL-ViT. In this subsection, we conduct a comparative analysis against CLIP.

As shown in Tab. 6, the performance achieved by using a linear probe on a frozen CLIP[14] model is superior to the performance observed under the setting where a linear probe is applied to a frozen OWL-ViT. Intriguingly, when we integrate our proposed causal invariance method with CLIP, the performance significantly declines. Conversely, applying causal invariance to OWL-ViT[10] leads to a substantial performance increase. This phenomenon aligns with the prerequisite we established for the efficacy of causal invariance, namely that the semantic information within the latent space must be relatively accurate and robust.

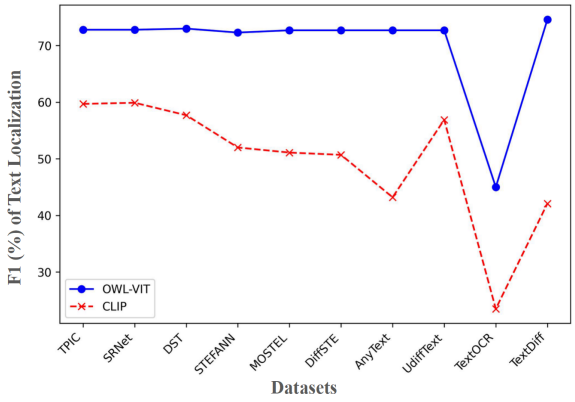

Figure 1. The performance of text localization by linear probing on CLIP and OWL-VIT. CLIP could not provide more accurate semantics (or content) than OWL-VIT.

OWL-ViT is a model designed for object detection, fine-tuned on the pre-aligned CLIP space. However, CLIP is generally insensitive to fine-grained information, a point substantiated by Fig. 1. Specifically, when linear probing is performed on both frozen CLIP and OWL-ViT for the task of text localization, CLIP's performance is markedly inferior to that of OWL-ViT. Therefore, optimizing the model on data generated using imprecise content information in the proposed causal invariance will naturally lead to suboptimal performance. The refinement of the confounding environment generation mechanism for improved precision remains a key objective for our future endeavors.

Table 5. The ablation study of Parameter-Efficient Adaptation methods for single-type generalization. Cau-Inv refers to the proposed causal invariance, including confounding environment generation and invariance optimization.

Methods	ID		OOD							
	TP-IC	SRNet	DST	STEFANN	MOSTEL	DiffSTE	AnyText	UDiffText	TextOCR	TextDiff
SVD	54.2	52.8	50.3	39.5	**41.1**	35.8	35.8	47.6	21.2	34.1
SVD + Cau-Inv	**55.0**	**53.6**	**51.8**	**40.9**	40.7	**36.0**	**36.4**	**48.3**	**23.2**	**36.4**
Adapter	74.6	71.8	39.1	11.8	53.7	48.3	**44.6**	47.3	26.6	53.8
Adapter + Cau-Inv	**78.1**	**77.0**	**52.3**	**25.0**	**69.4**	**58.6**	43.8	**58.1**	**41.6**	**60.6**

Table 6. The ablation study of foundation models for single-type generalization.

Methods	ID		OOD							
	TP-IC	SRNet	DST	STEFANN	MOSTEL	DiffSTE	AnyText	UDiffText	TextOCR	TextDiff
CLIP	**50.9**	**47.5**	**49.4**	**37.1**	**39.9**	**42.1**	**33.0**	**42.4**	**20.0**	**35.1**
CLIP + Cau-Inv	24.7	22.6	22.7	19.8	13.9	15.1	14.4	15.5	14.3	24.0
OWL-VIT	41.5	38.4	34.3	**32.0**	26.4	25.9	32.5	37.7	15.4	33.9
OWL-VIT + Cau-Inv	**78.1**	**77.0**	**52.3**	25.0	**69.4**	**58.6**	**43.8**	**58.1**	**41.6**	**60.6**

4. Appendix.D: Visualization

We visualize the localization masks, which are predicted in our method in Fig. 2. The exemplified tampered images, the corresponding predictions, and the ground truth masks are indicated in the first row, the second row, and the third row. The visualization suggests the promising performance of the proposed method.

5. Appendix.E: Others

We conduct the ablation study on the gradient detachment of observed data during confounding environment generation. Demonstrated in Tab. 7, the observed data should remain detached to block the gradient interference from the invariant regularization term on the observed data stream. This separation is critical for the improvement of generalization performance.

References

[1] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *Advances in Neural Information Processing Systems*, 36:9353–9387, 2023. 1

[2] Zhongxi Chen, Shen Chen, Taiping Yao, Ke Sun, Shouhong Ding, Xianming Lin, Liujuan Cao, and Rongrong Ji. Enhancing tampered text detection through frequency feature fusion and decomposition. In *European Conference on Computer Vision*, pages 200–217. Springer, 2024. 2

[3] Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3539–3553, 2022. 2

[4] Li Dong, Weipeng Liang, and Rangding Wang. Robust text image tampering localization via forgery traces enhancement and multiscale attention. *IEEE Transactions on Consumer Electronics*, 70(1):3495–3507, 2024. 2

[5] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20606–20615, 2023. 2

[6] Jiabao Ji, Guanhua Zhang, Zhaowen Wang, Bairu Hou, Zhifei Zhang, Brian Price, and Shiyu Chang. Improving diffusion models for scene text editing with dual encoders. *arXiv preprint arXiv:2304.05568*, 2023. 1

[7] Myung-Joon Kwon, In-Jae Yu, Seung-Hun Nam, and Heung-Kyu Lee. Cat-net: Compression artifact tracing network for detection and localization of image splicing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 375–384, 2021. 2

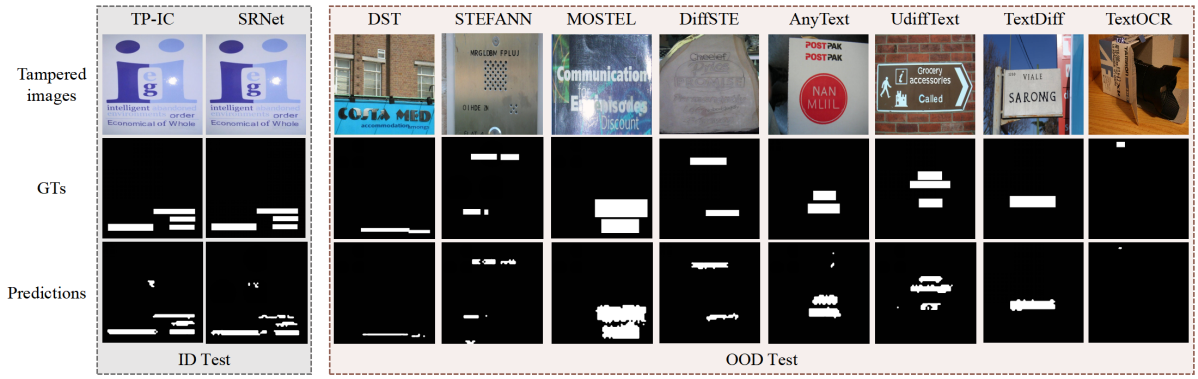

Figure 2. The visualization of the localization masks in our method.

Table 7. Ablation study on the gradient detachment of observed data during confounding environment generation.

Methods	ID		OOD							
	TP-IC	SRNet	DST	STEFANN	MOSTEL	DiffSTE	AnyText	UDiffText	TextOCR	TextDiff
Ours-w/o detach	**79.7**	71.8	48.9	24.7	61.5	55.3	**49.3**	55.0	30.4	53.0
Ours -w/ detach	78.1	**77.0**	**52.3**	**25.0**	**69.4**	**58.6**	43.8	**58.1**	**41.6**	**60.6**

[8] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32 (11):7505–7517, 2022. 2

[9] Xiaochen Ma, Bo Du, Zhuohang Jiang, Ahmed Y Al Hammadi, and Jizhe Zhou. Iml-vit: Benchmarking image manipulation localization by vision transformer. *arXiv preprint arXiv:2307.14863*, 2023. 2

[10] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *European conference on computer vision*, pages 728–755. Springer, 2022. 3

[11] Chenfan Qu, Chongyu Liu, Yuliang Liu, Xinhong Chen, Dezhi Peng, Fengjun Guo, and Lianwen Jin. Towards robust tampered text detection in document image: New dataset and new solution. In *CVPR*, pages 5937–5946, 2023. 2

[12] Chenfan Qu, Yiwu Zhong, Fengjun Guo, and Lianwen Jin. Revisiting tampered scene text detection in the era of generative ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 694–702, 2025. 2

[13] Yadong Qu, Qingfeng Tan, Hongtao Xie, Jianjun Xu, Yuxin Wang, and Yongdong Zhang. Exploring stroke-level modifications for scene text editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2119–2127, 2023. 1

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 3

[15] Prasun Roy, Saumik Bhattacharya, Subhankar Ghosh, and Umapada Pal. Stefann: scene text editor using font adaptive neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13228–13237, 2020. 1

[16] Wataru Shimoda, Daichi Haraguchi, Seiichi Uchida, and Kota Yamaguchi. De-rendering stylized texts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1076–1085, 2021. 1

[17] Yalin Song, Wenbin Jiang, Xiuli Chai, Zhihua Gan, Mengyuan Zhou, and Lei Chen. Cross-attention based two-branch networks for document image forgery localization in the metaverse. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(2):1–24, 2025. 2

[18] Lei Su, Xiaochen Ma, Xuekang Zhu, Chaoqun Niu, Zeyu Lei, and Ji-Zhe Zhou. Can we get rid of handcrafted feature extractors? sparsevit: Nonsemantics-centered, parameter-efficient image manipulation localization through spare-coding transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7024–7032, 2025. 2

[19] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. *arXiv preprint arXiv:2311.03054*, 2023. 1

[20] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. Manifold mixup: Better representations by interpolating hidden states. In *International conference on machine learning*, pages 6438–6447. PMLR, 2019. 2

[21] Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang.

Objectformer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2364–2373, 2022. 2

[22] Yuxin Wang, Hongtao Xie, Mengting Xing, Jing Wang, Shenggao Zhu, and Yongdong Zhang. Detecting tampered scene text in the wild. In *European Conference on Computer Vision*, pages 215–232. Springer, 2022. 1

[23] Liang Wu, Chengquan Zhang, Jiaming Liu, Junyu Han, Jingtuo Liu, Errui Ding, and Xiang Bai. Editing text in the wild. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1500–1508, 2019. 1

[24] Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable ai-generated image detection. *arXiv preprint arXiv:2411.15633*, 2024. 3

[25] Yiming Zhao and Zhouhui Lian. Udifftext: A unified framework for high-quality text synthesis in arbitrary images via character-aware diffusion models. In *European conference on computer vision*, pages 217–233. Springer, 2024. 1

[26] Xuekang Zhu, Xiaochen Ma, Lei Su, Zhuohang Jiang, Bo Du, Xiwen Wang, Zeyu Lei, Wentao Feng, Chi-Man Pun, and Ji-Zhe Zhou. Mesoscopic insights: orchestrating multi-scale & hybrid architecture for image manipulation localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11022–11030, 2025. 2