Revisiting Tampered Scene Text Detection in the Era of Generative AI

Chenfan Qu[1], **Yiwu Zhong**[2], **Fengjun Guo**[3, 4], **Lianwen Jin**[1, 4 *]

[1]South China University of Technology
[2]The Chinese University of Hong Kong
[3]Intsig Information Co., Ltd
[4]INTSIG-SCUT Joint Lab on Document Analysis and Recognition
202221012612@mail.scut.edu.cn, eelwjin@scut.edu.cn

Abstract

The rapid advancements of generative AI have fueled the potential of generative text image editing, meanwhile escalating the threat of misinformation spreading. However, existing forensics methods struggle to detect unseen forgery types that they have not been trained on, underscoring the need for a model capable of generalized detection of tampered scene text. To tackle this, we propose a novel task: open-set tampered scene text detection, which evaluates forensics models on their ability to identify both seen and previously unseen forgery types. We have curated a comprehensive, high-quality dataset, featuring the texts tampered by eight text editing models, to thoroughly assess the open-set generalization capabilities. Further, we introduce a novel and effective pre-training paradigm that subtly alters the texture of selected texts within an image and trains the model to identify these regions. This approach not only mitigates the scarcity of high-quality training data but also enhances models' fine-grained perception and open-set generalization abilities. Additionally, we present DAF, a novel framework that improves open-set generalization by distinguishing between the features of authentic and tampered text, rather than focusing solely on the tampered text's features. Our extensive experiments validate the remarkable efficacy of our methods. For example, our zero-shot performance can even beat the previous state-of-the-art full-shot model by a large margin.

Code — https://github.com/qcf-568/OSTF
Datasets — https://github.com/qcf-568/OSTF
Extended version — https://arxiv.org/abs/2407.21422

Introduction

The rapid development of deep models sparks a generative AI revolution in computer vision, demonstrating remarkable progress in controllable editing (Sun et al. 2023b; Qu et al. 2024b). However, the advancement of generative AI also leads to the spread of malicious fake information on text images, posing serious risks to social information security (Wang et al. 2022; Qu et al. 2023a). Consequently, the detection of AI-tampered text has become a vital topic in recent years (Qu et al. 2024a). It is crucial to develop effective methods for detecting AI-tampered text.

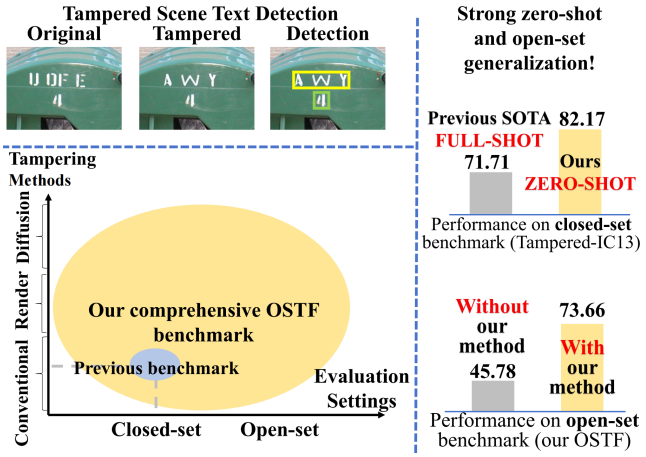

Figure 1: Tampered scene text detection aims to simultaneously detect real text (green box) and tampered text (yellow box) in the given image. In this paper, we introduce a novel task: open-set tampered scene text detection, where models are tested on both seen and unseen types of forgery. We also manually construct a comprehensive high-quality benchmark for this task. Moreover, we propose a simple-yet-effective method for this task, which shows strong zero-shot and open-set generalization ability.

Recently, the Tampered-IC13 dataset (Wang et al. 2022) has been introduced to benchmark the detection methods for tampered scene text. Several promising methods have been proposed for this task, such as frequency domain feature extraction (Wang et al. 2022; Qu et al. 2023a) and masked image modelling pre-training (Peng et al. 2023a,b). Despite significant progress that has been achieved, the existing techniques are far from sufficient for real-world scenarios. We summarize the limitations as follows:

First, **the failure of the existing dataset** to reflect model performance in real-world scenarios. The generative AI revolution has led to the continuous emergence of new text editing methods, producing increasingly realistic forgeries. The tampered texts in the Tampered-IC13 dataset were all forged using the outdated text editing model, SRNet (Wu et al. 2019), which is unlikely to be the real-world preference due to its relatively worse performance. Moreover,

*Corresponding author.

as the development of generative scene text editing is quite rapid, it is impractical to fit the forensics model to all types of the generative tampering methods. Therefore, the ability to generalize across unseen tampering methods and unseen scenarios becomes a crucial indicator of a model's practicality in real-world applications. Unfortunately, this open-set generalization ability cannot be adequately evaluated using the Tampered-IC13 benchmark.

Second, **the challenging nature of the tampered scene text detection task**. As introduced by pioneering work (Wang et al. 2022), this task faces two main challenges. (1). The lack of high-quality training data. Existing text editing methods have a high probability of producing unsatisfactory outputs, requiring costly manual refinement in post-processing for visual consistency (Wang et al. 2022). Training models on the purely generated samples without manual refinement will lead them to overfit to the obvious visual inconsistency, thereby making them difficult to generalize to real-world forgeries. (2). Enabling fine-grained perception. Tampered scene texts, after being generated by text editing models and further refined by human effort, become visually consistent with authentic texts (Qu et al. 2023b; Tuo et al. 2023). Only subtle texture anomalies may remain, which are challenging to detect. These two challenges have not been sufficiently addressed in previous works.

Third, **the poor open-set generalization of existing methods** for detecting tampered scene text. Forensic models oftentimes suffer from significant performance degradation on unseen types of forgeries. Such degradation is widely observed in other related fields, such as face anti-spoofing (Sun et al. 2023a; Zhou et al. 2023), deepfake detection (Dong et al. 2023; Wang et al. 2023), image manipulation localization (Sun et al. 2023c). Undoubtedly, no exception for the tampered scene text detection. However, none of the previous works pay attention to the open-set generalization of tampered scene text detection methods, resulting in a huge gap between the real-world applications.

To address these issues and to bridge the gap to real-world scenarios, we propose the following techniques:

1) To address the limitation of the existing benchmark, we manually construct a comprehensive high-quality benchmark for open-set tampered scene text detection, termed as Open-set Scene Text Forensics (OSTF). As shown in Figure 1, the tampered texts in our OSTF benchmark are tampered by a comprehensive set of text editing models, covering all the three types of scene text editing methods (conventional deep models, font rendering model, and diffusion models), successfully lining up with the recent development of generative AI. Except for the test setting of cross tampering methods, our OSTF benchmark also includes the test setting of cross source data, thus can better evaluate open-set generalization ability.

2) To simultaneously address the scarcity of high-quality training data, facilitate fine-grained perception, and improve open-set generalization, we propose a novel pre-training paradigm termed as Texture Jitter. In this paradigm, we subtly change the texture of the randomly selected text regions with elaborately designed operations to produce diverse types of texture anomalies, and train the model to localize these processed texts. Since the proposed Texture Jitter does not change the macro appearance, it can be applied directly to any type of image and will always produce perfectly visually consistent results. Therefore, the scarcity of high-quality training data can be significantly mitigated. In addition, since models trained with our Texture Jitter are forced to capture the subtle anomaly in the texture, the ability of fine-grained perception can be considerably improved. Moreover, by guiding the model to detect the texture anomaly rather than a specific type of tampering clue left by a particular tampering method, our Texture Jitter can also notably improve the model's open-set generalization.

3) To further facilitate the models' ability of open-set generalization, we propose a novel pluggable framework, termed as Difference-Aware Forensics, inspired by the development of unsupervised anomaly detection (Li et al. 2021b; Pang et al. 2021). The key idea is to learn a compact, robust representation for authentic text, and identify whether the input text is tampered or not by comparing its features with the learned authentic representation. By paying attention to the feature differences rather than only the features of the input text, models can better generalize to unseen forgeries.

We conduct extensive experiments on both the proposed OSTF benchmark and the widely-used Tampered-IC13 (Wang et al. 2022) benchmark. Our method demonstrates strong generalization ability in these experiments. For example, the proposed method leads to a gain of 27.88 mean F-score on the open-set generalization ability in the OSTF benchmark. Moreover, the zero-shot version of our method even outperforms the full-shot version of the previous SOTA method UPOCR (Peng et al. 2023b) by 10.46 mean IoU on the Tampered-IC13 benchmark.

In summary, the contributions of this paper are as follows:

- We propose a novel task, open-set tampered scene text detection, to meet the crucial demands. We manually construct a comprehensive high-quality benchmark for it.
- We propose a novel pre-training paradigm for tampered scene text detection. It significantly mitigates the scarcity of high-quality training data, and notably improves capabilities for both fine-grained perception and open-set generalization.
- We propose a novel pluggable framework that further improves open-set generalization on unseen forgery.
- In-depth analysis and extensive experiments have verified the effectiveness of the proposed method.

Related Works

Scene Text Editing. Since scene texts are sparse and have complex styles, various sizes, it is difficult to edit them to target content by copy-paste (Wang et al. 2022). The actual scene text editing is achieved by neural networks. Existing scene text editing methods can be divided into three types. (1) Conventional deep models, which edit scene text in an E2E manner without utilizing diffusion models. SRNet (Wu et al. 2019) is the first model to achieve E2E scene text editing. STEFANN (Roy et al. 2020) proposed to edit

scene text at char-level. MOSTEL (Qu et al. 2023b) improved the visual quality of text editing with stroke-level masks and self-supervised learning on non-synthetic data. (2) Font-rendering based methods, which edit text with digital font files. DST (Shimoda et al. 2021) inpainted the original text region and rendered new text on it with the corresponding digital font file. This is very similar to the process of manually editing the target text using image processing software. (3) Diffusion methods, which leverage the power of diffusion models for realistic text tampering. TextDiffuser (Chen et al. 2023b) generated text images with the given prompts, while DiffSTE (Ji et al. 2023) manipulated the specific parts of the images with target texts. AnyText (Tuo et al. 2023) introduced Auxiliary Latent Module for higher visual quality. UDiffText (Zhao and Lian 2023) improved scene text editing with large-scale training data and text embedding. The rapid development of generative scene text editing techniques brings huge challenges and risks to social security (Chen et al. 2023b). Therefore, it is essential to develop forensics models that can achieve open-set generalization on the text tampered by unseen methods.

Tampered scene text detection aims to localize tampered text on the given image. Wang. et al. (Wang et al. 2022) did the first work for tampered scene text detection, they proposed the first tampered scene text detection benchmark Tampered-IC13, they also proposed the S3R strategy and frequency domain modelling. However, this type of frequency domain modelling is likely to suffer from significant performance degradation on unseen types of forgery (Tan et al. 2023). Qu. et al. (Qu et al. 2023a) introduced Selective Tampering Synthesis method and Document Tampering Detector model to improve tampered text detection in documents. However, these methods are not suitable for scene text due to the board variety in sizes and appearance. They also introduced the DocTamper synthetic dataset for documents, but it unable to benchmark model performance across unseen tampering methods and on AIGC-based tampering. Benefiting from more training data and more lenient evaluation metrics, Peng. et al. (Peng et al. 2023b) achieved the highest mean F-score on the Tampered-IC13 benchmark by averaging real text detection scores. However, due to the lack of a specialized design, their performance for tampered text detection is still unsatisfactory. None of the existing work has explored open-set tampered scene text detection. To meet with the real-life requirements, we manually construct a comprehensive high-quality benchmark for open-set tampered scene text detection and propose novel, simple-yet-effective methods.

OSTF Dataset and Benchmark

Motivation. In this era of generative AI, numerous new text editing models continuously emerge (Chen et al. 2023b; Zhao and Lian 2023). However, the existing Tampered-IC13 benchmark only covers text tampered by the oldest text editing method SRNet, which can hardly be a real-world preference due to its relatively inferior performance. Moreover, the ability to detect the text tampered by unseen text editing model and that on unseen scenario is essential for forensics model in this era of generative AI. However, this ability totally cannot be evaluated in the Tampered-IC13 benchmark. To address the above issues, we manually construct a comprehensive high-quality new benchmark for tampered scene text detection, termed as OSFT, which includes text tampered by various latest text editing methods and cross source-dataset evaluation settings.

Dataset Construction

Tampering Methods. We take all the three types of generative tampering methods (conventional, font rendering, diffusion) into account, and select eight text editing methods, including SRNet (Wu et al. 2019), STEFANN (Roy et al. 2020), MOSTEL (Qu et al. 2023b), DST (Shimoda et al. 2021), DiffSTE (Ji et al. 2023), AnyText (Tuo et al. 2023), UDiffText (Zhao and Lian 2023), Textdiffuser (Chen et al. 2023a), as shown in Table 2. Given that the Tampered-IC13 dataset (Wang et al. 2022) already has reasonable forgeries tampered by SRNet, we take it as our 'SRNet' part.

Data Source. We forge the text images from the ICDAR2013 (Karatzas et al. 2013) with the selected eight text editing methods. To enable the cross-source dataset evaluation, we further edit the text images from the TextOCR (Singh et al. 2021) validation set with UDiffText, and the text images from the ICDAR2017 (Nayef et al. 2017), ReCTS (Zhang et al. 2019) validation sets with TextDiffuser.

Manual Improvement. To ensure the high quality of the tampered texts in our dataset, we edit all text instances using the chosen methods, and manually pick the most successful outputs. We further manually and elaborately improve the visual consistency of the picked texts using PhotoShop and GIMP. To ensure the high quality of the annotations, we manually update the bounding box for each tampered text to match its new appearance. We also check the labels via visualization to avoid errors.

Dataset Statistics, Benchmark Settings, Highlights

Dataset Statistics. As shown in Table 1, there are a total of 5018 tampered texts and 1980 tampered images in our OSTF dataset. The detailed statistics of it are shown in Table 2.

Evaluation Settings. As shown in Table 2, there are 9 sessions in our dataset (ICDAR2013 tampered by 7 methods, TextOCR tampered by UDiffText, ICDAR2017 and ReCTS tampered by TextDiffuser). To evaluate both closed-set performance and open-set generalization, the models are **trained on one session of the training set** and **tested on all nine sessions of the testing set**. As a result, there are 9×9=81 test settings, enabling three evaluation protocols: cross tampering methods, cross source dataset, and cross both tampering methods and source datasets.

Dataset Highlights. The comparison between our OSTF dataset and previous dataset is shown in Table 1. Some of the samples in our dataset are shown in Figure 2. The main highlights of our dataset are as follows:

(1) **Comprehensive**. Our dataset includes all three types of generative text editing methods, keeping pace with the generative AI revolution. Our benchmark includes both the cross-AI-model and cross-dataset evaluation settings.

Name	Cross Domain	Number of images		Number of texts		Tampering methods			
		All	Tampered	All	Tampered	Types	Conv.	Rend.	Diff.
Tampered-IC13	×	462	378	1944	995	1	✓	×	×
OSTF (Ours)	✓	**4418**	**1980**	**64858**	**5018**	**8**	✓	✓	✓

Table 1: Comparison between our OSTF dataset and the previous dataset of Tampered Scene Text Detection. "Conv." denotes "Conventional deep models", "Rend." denotes "Rendering based methods", "Diff." denotes "Diffusion models".

Tampering type	Tampering method	Images				Text instances				Data Source
		Authentic		Tampered		Authentic		Tampered		
		train	test	train	test	train	test	train	test	
Rendering	DST	72	82	157	151	382	588	467	507	IC-13
Conventional deep models	SRNet	29	55	200	178	342	607	507	488	IC-13
	STEFANN	182	181	47	52	721	946	128	149	IC-13
	MOSTEL	168	172	61	61	628	882	221	213	IC-13
Diffusion models	DiffSTE	174	181	55	52	683	943	166	152	IC-13
	AnyText	181	191	48	42	715	974	134	121	IC-13
	UDiffText	129	132	100	101	471	772	378	323	IC-13
	UDiffText	196	233	218	211	23737	22886	419	399	T-OCR
	TextDiffuser	40	40	123	123	2048	1515	123	123	IC-17

Table 2: The detailed statistics of the proposed OSTF benchmark. "IC-13" denotes ICDAR2013 (Karatzas et al. 2013) dataset, "T-OCR" denotes TextOCR (Singh et al. 2021) dataset, "IC-17" denotes ICDAR2017 (Nayef et al. 2017) dataset.

Figure 2: Samples in the proposed Open-set Scene Text Forensics (OSTF) dataset.

(2) **High Quality**. The tampered texts in our OSTF dataset are manually selected and elaborately enhanced for better visual consistency. The bounding boxes are manually adjusted to match the tampered texts.

(3) **Board Diversity**. The images in our OSTF dataset have various styles and resolutions. The texts in these images have various fonts and backgrounds, and are tampered by 8 different generative methods.

Pre-training with Texture Jitter

Motivation. Existing text editing methods often generate noticeable visual inconsistencies (Wang et al. 2022), especially when dealing with complex fonts, unfamiliar languages, and diverse styles. Such failure to generate vivid tampered text leads to overfitting to obvious visual inconsistency. As the result, the trained models struggle to generalize their applicability to real-world scenarios where the visual appearance of the manipulated text is refined through human intervention and has minimized visual inconsistency.

Method. Based on the above observations, we propose a **simple-yet-effective** method called Textual Jitter, as shown in Figure 3. It slightly changes the texture of randomly picked texts while keeping their macro appearance unchanged, that is, the processed texts are almost the same as the original ones. Once this data processing is done, we train the models to localize the text regions and to identify whether their texture has been changed or not. Without changing the macro appearance, this approach can be applied directly to any text image and can always produce high-quality training data that is visually consistent and similar to the elaborately tampered text. Despite the simplicity, the proposed Texture Jitter can **simultaneously address three major issues** in tampered scene text detection:

(1) **The scarcity of high-quality training data** is significantly mitigated. Our Texture Jitter can automatically output tampered text that is visually consistent with the macro appearance. It can be directly applied to any type of image and any foreign language, and produce diverse and high-quality data, thereby addressing the data scarcity issue.

(2) **The ability of fine-grained perception** is considerably improved. The text processed by our Texture Jitter is

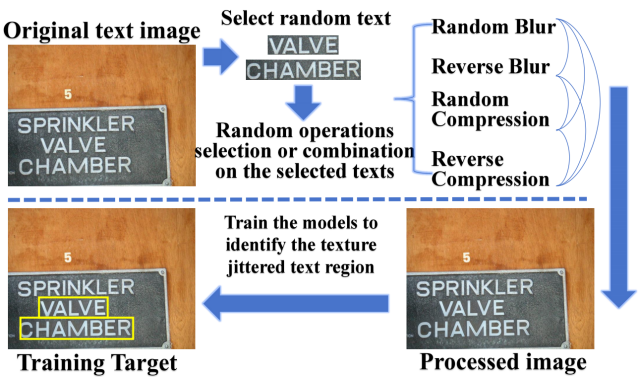

Figure 3: The pipeline of the proposed Texture Jitter.

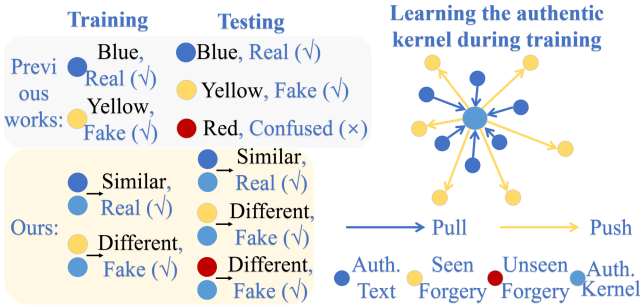

Figure 4: The key idea of our DAF is to model the feature difference rather than the input features themselves.

visually consistent with surrounding texts and its anomaly is obscure and difficult to capture. By training forensics models with our Texture Jitter, the models are encouraged to detect subtle anomalies and the ability of fine-grained perception can be improved.

(3) **The ability of open-set generalization** is notably enhanced. By training with our Texture Jitter, the models learn to identify the tampered text by judging whether the texture is abnormal, rather than relying on a specific feature of a tampering method, and thus achieve better open-set generalization on unseen forgery.

Implementation. The pipeline of our Texture Jitter is shown in Figure 3. Given an input text image, we randomly select some text instances, and apply a random texture processing operation or the combination of multiple operations. The texture processing operations include random blur, reverse blur, random image compression, and reverse image compression. By doing this, various types of texture anomalies are created. The processed text instances are then spliced into the same positions as the original ones, with a smooth transition at the edges to avoid obvious visual anomalies. Further, to ensure a balance between the learning difficulty and a natural fusion between the original and processed regions, we also propose to adaptively adjust the intensity of jittering and the edge smoothing with rules, based on the size of the target text. With the proposed adaptive intensity, Texture Jitter can always produce satisfactory outputs with minimized visual inconsistency. More details are given in the appendix. The full implementation will be open-source.

Difference Aware Forensics

Motivation. Forensics models often suffer from significant performance degradation on unseen forgeries (Zhou et al. 2023; Sun et al. 2023c). Typically in the era of generative AI, advanced generative models are rapidly emerging and they can edit the texts in images in ways that the forensics models have never seen during training. Therefore, constructing a robust forensics model that can generalize across unseen forgeries is critical for real-life scenarios.

Analysis. The performance degradation on unseen forgeries is mostly caused by the training objective, a common binary classification task (Perez-Cabo et al. 2019). As shown in the top left of Figure 4, during the training process, the models simply learn the specific features for the seen authentic class (blue circle) and the tampered class (yellow circle). When text is tampered by an unseen text editing method, the editing styles are never seen before and thus the features of the tampered text (red circle) are different from the seen ones. Consequently, the classifier is confused by new features of unseen forgeries, leading to poor performance.

Key idea. Although diverse text editing models will produce various tampering styles, the features extracted from the real text should remain similar. Moreover, the features of the tampered text can be regarded as anomalies since they are different from those of the authentic text. Inspired by this, we propose to model the difference between the input feature and the authentic feature for text forensics, instead of just relying on the individual input feature, as shown in Figure 4. Both the seen and unseen forgery features can be distinguished from the authentic feature, so that the confusion caused by the unseen forgery can be alleviated.

Implementation. Specifically, we propose a novel framework DAF, as shown in Figure 5. Our DAF builds on the widely-used detection models (e.g., Faster R-CNN (Ren et al. 2015)) **by simply adding a forensics branch**, and is trained and tested in an E2E manner. The forensics branch shares the same backbone model as the original detection head, and extracts features for real/fake classification with an extra FPN network (Lin et al. 2017). This follows the S3R (Wang et al. 2022), which performs text localization and forensics with separate networks. In the forensics branch, an authentic kernel is learned by pulling together the features of authentic text and pushing away the features of tampered text, which is supervised by L2 loss during training. More details are given in the appendix. **The intuition behind our design** is that the authentic kernel is approximately equivalent to the average of the features extracted from the real texts in the training set, and it can be regarded as the common representation for the real texts. For each input image, the learned authentic kernel is modulated by the global representation of the image with a linear layer to adapt to the image style. We subtract between the modulated authentic kernel and the RoI feature vectors. Finally, the subtracted features are fed into the classifier for final prediction.

Method	Real Text		Tampered Text		Average	
	IoU	F	IoU	F	mIoU	mF
DeepLabV3+ (Chen et al. 2018) (full-shot)	48.1	65.0	72.2	83.7	60.2	74.4
HRNetv2 (Wang et al. 2020a) (full-shot)	43.3	60.4	73.1	84.5	58.2	72.4
BEiT-UPer (Bao et al. 2021) (full-shot)	57.1	72.7	70.9	83.0	64.0	77.8
SegFormer (Xie et al. 2021) (full-shot)	53.2	69.5	77.8	87.5	65.5	79.0
Swin-UPer (Liu et al. 2021) (full-shot)	61.8	76.4	77.3	87.2	69.6	81.8
UPOCR (Peng et al. 2023b) (full-shot)	71.8	83.6	71.6	83.5	71.7	83.5
Ours+Faster R-CNN (Ren et al. 2015) **(zero-shot)**	79.9	88.8	84.5	91.6	82.2	90.2
Ours+Cascade R-CNN (Cai and Vasconcelos 2018) **(zero-shot)**	79.1	88.3	84.3	91.5	81.7	89.9
Ours+Faster R-CNN (Ren et al. 2015) (full-shot)	**82.9**	**90.6**	89.5	94.4	**86.2**	**92.5**
Ours+Cascade R-CNN (Cai and Vasconcelos 2018) (full-shot)	79.4	88.5	**90.0**	**94.7**	84.7	91.6

Table 3: Comparison study on the Tampered-IC13 dataset, 'mIoU' denotes mean IoU and 'mF' denotes mean F1.

Method	Real Text			Tampered Text			Avg.
	P	R	F	P	R	F	mF
S3R+EAST (Zhou et al. 2017) (full-shot)	50.5	27.3	35.5	70.2	70.0	69.9	52.7
S3R+PSENet (Wang et al. 2019a) (full-shot)	61.6	41.9	49.9	79.9	79.4	79.7	64.8
S3R+ATRR (Wang et al. 2019b) (full-shot)	76.7	54.6	63.8	84.6	90.6	87.5	75.7
S3R+CounterNet (Wang et al. 2020b) (full-shot)	77.9	54.8	64.3	86.7	91.5	89.0	76.7
DTD (Qu et al. 2023a) (full-shot)	-	-	-	92.1	89.3	90.7	45.4
Ours+Faster R-CNN **(zero-shot)**	71.4	78.4	74.7	86.64	82.38	84.45	79.6
Ours+Cascade R-CNN **(zero-shot)**	75.6	76.1	75.9	88.44	81.56	84.86	80.4
Ours+Faster R-CNN (full-shot)	80.5	81.7	81.1	91.44	96.31	93.81	87.5
Ours+Cascade R-CNN (full-shot)	**83.0**	**82.5**	**82.7**	**92.37**	**96.72**	**94.49**	**88.6**

Table 4: Comparison study on the Tampered-IC13 dataset, 'S3R' represents the method proposed in (Wang et al. 2022).

Number	Ablation settings							Performance		
	Pre-training			Framework		Base detector		Mean F-score		
	COCO Det	Text Det	Ours	SCL	DAF (Ours)	Faster R-CNN	Cascade R-CNN	Real text	Fake text	Mean score
(1)	✓					✓		59.92	34.82	47.37
(2)	✓				✓		✓	63.38	37.66	50.52
(3)		✓				✓		71.34	45.78	58.56
(4)		✓			✓		✓	72.66	48.01	60.34
(5)			✓			✓		74.87	71.96	73.41
(6)			✓	✓		✓		75.26	72.32	73.79
(7)			✓		✓	✓		75.98	73.66	74.82
(8)			✓		✓		✓	**76.74**	**74.96**	**75.85**

Table 5: Ablation study on the OSTF dataset. 'mP', 'mR', 'mF' denote mean precision, mean recall and mean F1-score respectively. 'SCL' denotes using single-center-loss (Li et al. 2021a) in training. 'DAF' denotes our Difference-Aware Forensics.

Experiments

Implement Details We conduct experiments on both the well-acknowledged Tampered-IC13 benchmark (Wang et al. 2022) and our proposed OSTF benchmark. We first pre-train our model for 12 epochs with the proposed Texture Jitter on the same training sets as UPOCR (Peng et al. 2023b), including LSVT (Sun et al. 2019), ReCTS (Zhang et al. 2019), ICDAR2013 (Karatzas et al. 2013), ICDAR2015 (Karatzas et al. 2015), ICDAR2017 (Nayef et al. 2017), TextOCR (Singh et al. 2021), ArT (Chng et al. 2019). The AdamW optimizer (Loshchilov and Hutter 2017) with a learning rate initialized at 6e-5 and decaying to 1e-6 is used in the experiments. We then fine-tune the model using also the training sets of the Tampered-IC13 (Wang et al. 2022) and the OSTF datasets respectively for 15k iterations with a batch size of 8. We adopt Swin-Transformer (Small) (Liu et al. 2021) as the backbone following previous work (Peng et al. 2023b). The input image is resized to ensure that the shortest edge ≤ 1024 and the longest edge ≤ 1536.

Evaluation Metric. For a fair comparison with the segmentation-based methods, we convert the bounding box predictions of our model into segmentation maps (detailed in the appendix) and calculate the Precision (P), Recall (R) and F1-score (F) at the pixel-level with the built-in functions of mmsegmentation (Contributors 2020) following the previous works (Peng et al. 2023b). For a fair comparison with

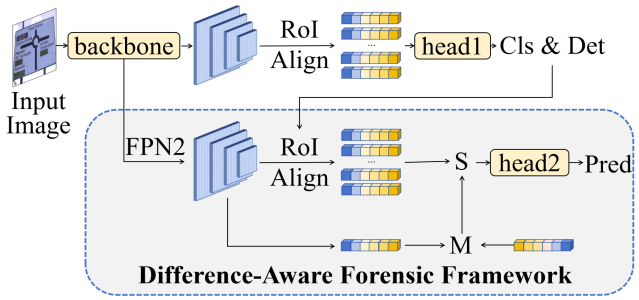

Figure 5: The proposed Difference-Aware Forensics (DAF). It builds on two-stage detectors such as Faster R-CNN.

the detection-based methods, we adopt P, R, F calculated at the instance-level with the official scripts provided by the Tampered-IC13 (Wang et al. 2022). We use these detection-based metrics for the ablation experiments.

Comparison Study

Comparison with segmentation models. We compare the performance between our models and the SOTA segmentation-based forensic models, including DeepLabV3+ (Chen et al. 2018), HRNetV2 (Wang et al. 2020a), BEiT-UPer (Bao et al. 2021), SegFormer (Xie et al. 2021), Swin-UPer (Liu et al. 2021) and UPOCR (Peng et al. 2023b). The comparison results on the Tampered-IC13 benchmark (Wang et al. 2022) are shown in Table 3. The **zero-shot variants** of our models can even notably outperform the **full-shot version** of the previous SOTA model UPOCR by **more than 10 points** mIoU. The "zero-shot" represents our pre-trained model that has never seen any tampered text used in fine-tuning. Its strong zero-shot generalization ability is attributed to the effectiveness of our methods. Our full-shot models achieve higher performance.

Comparison with detection models. We compare the performance between our models and the SOTA detection-based forensic models, including S3R+EAST (Zhou et al. 2017), S3R+PSENet (Wang et al. 2019a), S3R+ATRR (Wang et al. 2019b), S3R+CounterNet (Wang et al. 2020b) and DTD (Qu et al. 2023a). As shown in Table 4, the zero-shot variants of our models also outperform the full-shot version of the state-of-the-art (SOTA) model by about 3 points mean F-score. The full-shot variants of our models further significantly outperform the state-of-the-art model (e.g. the full-shot version of 'Ours+Cascade R-CNN' achieves 88.62 mean F-score, about **12 points higher** than the SOTA of 76.66), demonstrating strong generalization.

Ablation Study

The ablation results on the OSFT dataset are shown in Table 5, where the 'mP', 'mR', 'mF' are calculated by averaging the P, R, F of all the 81 test settings in the OSTF dataset.

In Table 5, settings (1), (3), (5) and (2), (4), (8) are designed to verify the effectiveness of our Texture Jitter paradigm. The model pre-trained with our Texture Jitter (setting (5)) gets 71.96 mF-score on tampered text detection, 26.18 points higher than the baseline pre-trained with the same training configuration but only with the common text detection task (setting (3)), and even 37.14 points higher than the baseline initialized with the official COCO detection pretrained weights (setting (1)). Similarly, huge improvement on Cascade R-CNN (Cai and Vasconcelos 2018) can also be observed by comparing between settings (2), (4), (8). This demonstrates the surprising effectiveness of the proposed Texture Jitter. Settings (5), (6), (7) are designed to verify the effectiveness of our DAF framework. By adding single-center-loss (Li et al. 2021a) to setting (5), setting (6) gets only tiny improvements. However, by equipping setting (5) with our DAF framework, setting (7) achieves much larger improvements. Setting (7) and setting (8) show that our methods are robust to different base detectors, and that a better base detector can mostly lead to better performance.

Discussion. Models have relatively worse open-set performance when being tested on 'STEFANN' and 'TextOCR', for the following reasons: (1) For STEFANN, there is a huge difference in texture appearance between the output of STEFANN and other editing methods. Texts edited by STEFANN have almost binary texture appearance, while texts edited by other methods have richer and more realistic texture details (Figure 2). Therefore, generalization is much more challenging. (2) For TextOCR, it has many tiny and fuzzy texts that are challenging to detect. When models are trained on other source data, they are adapted to big clear text and thus have performance degradation on such tiny and fuzzy texts. The tiny and fuzzy texts are also the main reason for false positives in other subsets. Despite these challenges, our proposed methods still significantly improve the performance (e.g. 33.07 points improvement for 'STEFANN' and 28.31 points improvement for 'TextOCR' on average).

Conclusion

In the era of generative AI, this paper introduces a novel task of open-set tampered scene text detection, designed to meet the demands for generalized forensics analysis. This task challenges forensics models with both seen and previously unseen forgeries, aiming to thoroughly evaluate their generalization. To facilitate this, we have manually developed a comprehensive high-quality benchmark, named OSTF, which stands out by including text images tampered by various editing models. Further, we introduce an innovative pre-training paradigm, termed Texture Jitter, which effectively mitigates the lack of high-quality training data and significantly enhances the model capabilities in fine-grained perception and open-set generalization. The proposed Texture Jitter is the first fine-grained perception pre-training paradigm, which differs from previous works that are coarse-grained perception. To further enhance open-set generalization, we also present a novel pluggable framework DAF, which focuses on identifying the feature difference between authentic and tampered texts. This differs from previous works that rely only on the specific input features. Extensive experiments validate the effectiveness of our methods. With these advances, our work achieves a significant step forward in the field of tampered text detection.

Acknowledgments

This research is supported in part by the National Natural Science Foundation of China (Grant No.: 62441604, 62476093) and IntSig-SCUT Joint Lab Foundation.

References

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*.

Cai, Z.; and Vasconcelos, N. 2018. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6154–6162.

Chen, J.; Huang, Y.; Lv, T.; Cui, L.; Chen, Q.; and Wei, F. 2023a. TextDiffuser-2: Unleashing the Power of Language Models for Text Rendering. *arXiv preprint arXiv:2311.16465*.

Chen, J.; Huang, Y.; Lv, T.; Cui, L.; Chen, Q.; and Wei, F. 2023b. TextDiffuser: Diffusion Models as Text Painters. In Oh, A.; Neumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 9353–9387. Curran Associates, Inc.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Chng, C. K.; Liu, Y.; Sun, Y.; Ng, C. C.; Luo, C.; Ni, Z.; Fang, C.; Zhang, S.; Han, J.; Ding, E.; et al. 2019. Icdar2019 robust reading challenge on arbitrary-shaped text-rrc-art. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1571–1576. IEEE.

Contributors, M. 2020. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark.

Dong, S.; Wang, J.; Ji, R.; Liang, J.; Fan, H.; and Ge, Z. 2023. Implicit identity leakage: The stumbling block to improving deepfake detection generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3994–4004.

Ji, J.; Zhang, G.; Wang, Z.; Hou, B.; Zhang, Z.; Price, B.; and Chang, S. 2023. Improving diffusion models for scene text editing with dual encoders. *arXiv preprint arXiv:2304.05568*.

Karatzas, D.; Gomez-Bigorda, L.; Nicolaou, A.; Ghosh, S.; Bagdanov, A.; Iwamura, M.; Matas, J.; Neumann, L.; Chandrasekhar, V. R.; Lu, S.; et al. 2015. ICDAR 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, 1156–1160. IEEE.

Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; i Bigorda, L. G.; Mestre, S. R.; Mas, J.; Mota, D. F.; Almazan, J. A.; and De Las Heras, L. P. 2013. ICDAR 2013 robust reading competition. In *2013 12th international conference on document analysis and recognition*, 1484–1493. IEEE.

Li, J.; Xie, H.; Li, J.; Wang, Z.; and Zhang, Y. 2021a. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6458–6467.

Li, T.; Wang, Z.; Liu, S.; and Lin, W.-Y. 2021b. Deep Unsupervised Anomaly Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 3636–3645.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Nayef, N.; Yin, F.; Bizid, I.; Choi, H.; Feng, Y.; Karatzas, D.; Luo, Z.; Pal, U.; Rigaud, C.; Chazalon, J.; et al. 2017. Icdar2017 robust reading challenge on multi-lingual scene text detection and script identification-rrc-mlt. In *2017 14th IAPR international conference on document analysis and recognition (ICDAR)*, volume 1, 1454–1459. IEEE.

Pang, G.; Shen, C.; Cao, L.; and Hengel, A. V. D. 2021. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2): 1–38.

Peng, D.; Liu, C.; Liu, Y.; and Jin, L. 2023a. ViTEraser: Harnessing the Power of Vision Transformers for Scene Text Removal with SegMIM Pretraining. *arXiv preprint arXiv:2306.12106*.

Peng, D.; Yang, Z.; Zhang, J.; Liu, C.; Shi, Y.; Ding, K.; Guo, F.; and Jin, L. 2023b. UPOCR: Towards Unified Pixel-Level OCR Interface. arXiv:2312.02694.

Perez-Cabo, D.; Jimenez-Cabello, D.; Costa-Pazo, A.; and Lopez-Sastre, R. J. 2019. Deep Anomaly Detection for Generalized Face Anti-Spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Qu, C.; Liu, C.; Liu, Y.; Chen, X.; Peng, D.; Guo, F.; and Jin, L. 2023a. Towards robust tampered text detection in document image: new dataset and new solution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5937–5946.

Qu, C.; Zhong, Y.; Guo, F.; and Jin, L. 2024a. Omni-IML: Towards Unified Image Manipulation Localization. *arXiv preprint arXiv:2411.14823*.

Qu, C.; Zhong, Y.; Liu, C.; Xu, G.; Peng, D.; Guo, F.; and Jin, L. 2024b. Towards Modern Image Manipulation Localization: A Large-Scale Dataset and Novel Methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10781–10790.

Qu, Y.; Tan, Q.; Xie, H.; Xu, J.; Wang, Y.; and Zhang, Y. 2023b. Exploring stroke-level modifications for scene text

editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2119–2127.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Roy, P.; Bhattacharya, S.; Ghosh, S.; and Pal, U. 2020. STEFANN: Scene Text Editor Using Font Adaptive Neural Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shimoda, W.; Haraguchi, D.; Uchida, S.; and Yamaguchi, K. 2021. De-Rendering Stylized Texts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1076–1085.

Singh, A.; Pang, G.; Toh, M.; Huang, J.; Galuba, W.; and Hassner, T. 2021. Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8802–8812.

Sun, Y.; Liu, Y.; Liu, X.; Li, Y.; and Chu, W.-S. 2023a. Rethinking domain generalization for face anti-spoofing: Separability and alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24563–24574.

Sun, Y.; Ni, Z.; Chng, C.-K.; Liu, Y.; Luo, C.; Ng, C. C.; Han, J.; Ding, E.; Liu, J.; Karatzas, D.; et al. 2019. ICDAR 2019 competition on large-scale street view text with partial labeling-RRC-LSVT. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, 1557–1562. IEEE.

Sun, Z.; Fang, H.; Zhao, X.; Wang, D.; and Cao, J. 2023b. Rethinking Image Editing Detection in the Era of Generative AI Revolution. arXiv:2311.17953.

Sun, Z.; Fang, H.; Zhao, X.; Wang, D.; and Cao, J. 2023c. Rethinking Image Editing Detection in the Era of Generative AI Revolution. arXiv:2311.17953.

Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; and Wei, Y. 2023. Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12105–12114.

Tuo, Y.; Xiang, W.; He, J.-Y.; Geng, Y.; and Xie, X. 2023. AnyText: Multilingual Visual Text Generation And Editing. *arXiv preprint arXiv:2311.03054*.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. 2020a. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3349–3364.

Wang, W.; Xie, E.; Li, X.; Hou, W.; Lu, T.; Yu, G.; and Shao, S. 2019a. Shape robust text detection with progressive scale expansion network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9336–9345.

Wang, X.; Jiang, Y.; Luo, Z.; Liu, C.-L.; Choi, H.; and Kim, S. 2019b. Arbitrary shape scene text detection with adaptive text region representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6449–6458.

Wang, Y.; Xie, H.; Xing, M.; Wang, J.; Zhu, S.; and Zhang, Y. 2022. Detecting tampered scene text in the wild. In *European Conference on Computer Vision*, 215–232. Springer.

Wang, Y.; Xie, H.; Zha, Z.-J.; Xing, M.; Fu, Z.; and Zhang, Y. 2020b. Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11753–11762.

Wang, Z.; Bao, J.; Zhou, W.; Wang, W.; and Li, H. 2023. Altfreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4129–4138.

Wu, L.; Zhang, C.; Liu, J.; Han, J.; Liu, J.; Ding, E.; and Bai, X. 2019. Editing Text in the Wild. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, 1500–1508. New York, NY, USA: Association for Computing Machinery. ISBN 9781450368896.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090.

Zhang, R.; Zhou, Y.; Jiang, Q.; Song, Q.; Li, N.; Zhou, K.; Wang, L.; Wang, D.; Liao, M.; Yang, M.; et al. 2019. Icdar 2019 robust reading challenge on reading chinese text on signboard. In *2019 international conference on document analysis and recognition (ICDAR)*, 1577–1581. IEEE.

Zhao, Y.; and Lian, Z. 2023. UDiffText: A Unified Framework for High-quality Text Synthesis in Arbitrary Images via Character-aware Diffusion Models. *arXiv preprint arXiv:2312.04884*.

Zhou, Q.; Zhang, K.-Y.; Yao, T.; Lu, X.; Yi, R.; Ding, S.; and Ma, L. 2023. Instance-aware domain generalization for face anti-spoofing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20453–20463.

Zhou, X.; Yao, C.; Wen, H.; Wang, Y.; Zhou, S.; He, W.; and Liang, J. 2017. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 5551–5560.