Towards Better Robustness Against Natural Corruptions in Document Tampering Localization

Huiru Shao[1,2], **Kaizhu Huang**[3], **Wei Wang**[1], **Xiaowei Huang**[2], **Qiufeng Wang**[1✉]
[1]Xi'an Jiaotong-Liverpool University, China
[2]University of Liverpool, UK
[3]Duke Kunshan University, China
Qiufeng.Wang@xjtlu.edu.cn

Abstract

Marvelous advances have been exhibited in recent document tampering localization (DTL) systems. However, confronted with corrupted tampered document images, their vulnerability is fatal in real-world scenarios. While robustness against adversarial attack has been extensively studied by adversarial training (AT), the robustness on natural corruptions remains under-explored for DTL. In this paper, to overcome forensic dependency, we propose the adversarial forensic regularization (AFR) based on min-max optimization to improve robustness. Specifically, we adopt mutual information (MI) to represent forensic dependency between two random variable over tampered and authentic pixels spaces, where the MI can be approximated by Jensen-Shannon-Divergence (JSD) with empirical sampling. To further enable a trade-off between predictive representations in clean tampered document pixels and robust ones in corrupted pixels, an additional regularization term is formulated with divergence between clean and perturbed pixels distribution (DDR). Following min-max optimization framework, our method can also work well against adversarial attacks. To evaluate our proposed method, we collect a dataset (i.e., TSorie-CRP) for evaluating robustness against natural corruptions in real scenarios. Extensive experiments demonstrate the effectiveness of our method against natural corruptions. Without any surprise, our method also achieves good performance against adversarial attack on DTL benchmark datasets.

Datasets — https://github.com/SHR-77/CRP

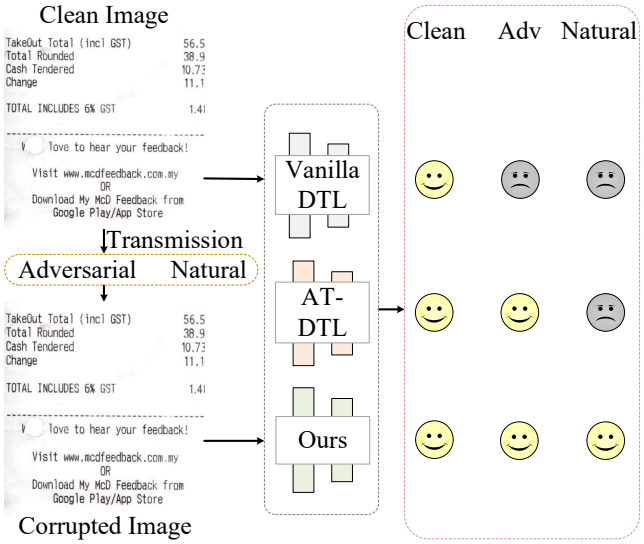

Figure 1: The vulnerability of DTL models against natural corruption. Vanilla DTL model aims to localize tampered pixels in clean tampered document images (TDI) but fails for images after corruptions. DTL with adversarial training (AT-DTL) improves the performance on adversarial attacks while overlooking the natural corruptions like noise and OSN. Our method can improve robustness on natural corruption while maintaining adversarial robustness.

Introduction

To maintain the trustworthiness and integrity of documents, document tampering localization (DTL) has emerged and received significant attention in forensic and security community, which attempts to identify and segment unauthorized alterations made to a document. Meanwhile, rapid advancement of image editing (Wang et al. 2022a; Guo et al. 2023) with less tampered traces further promotes more researchers to devote to document tampering localization and accelerate great progress (Qu et al. 2023; Shao et al. 2023; Kwon et al. 2021; Dong et al. 2022; Wu et al. 2022a).

Although it seems that almost entire tampered pixels are capable to be segmented in recent DTL systems, their vulnerability on corrupted tampered document images is evident in real scenarios, which is depicted in Fig.1. Corruptions are mainly originated from adversarial and natural corruptions. In various tasks, adversarial robustness has attracted much attention and made great progress (Madry et al. 2017; Akhtar et al. 2021; Zhang and Wang 2019). However, very few works are proposed for document tampering localization (DTL). To our knowledge, only one paper (Shao et al. 2024) aims to improve adversarial robustness in DTL by leveraging latent manifolds. However, the robustness against natural corruption remains under-explored in DTL, but these cases are very common in real scenarios. For example, after transmission through online social networks (OSNs), forensic clues captured by vanilla DTL systems tend to be destroyed, resulting in the failure of DTL.

To improve the robustness, one straightforward idea is to add more samples with similar corruption into training set. In practice, there are various corruptions and it is impossible to know the exact corruptions in advance, therefore it is challenging to collect these similar data. To overcome this issue, it is essential to improve robustness fundamentally by delving into DTL for mastering forensic features.

Various patterns in low-level signal within one image exist during the process of image generation and processing, such as mosaic patterns for creating colors (Bammey, Gioi, and Morel 2020) and block artifact grids (BAG) after JPEG. After tampering, various inconsistencies of these patterns between tampered and authentic regions may appear (Dong et al. 2022), such as Local mosaic inconsistencies and discontinuous BAG. DTL tends to align with strong association between such inconsistencies and labels. However, the corruption will obscure such inconsistencies, making larger shared information between authentic and tampered representation (we call it forensic dependency) and resulting in unsatisfactory robustness of current DTL models.

To reduce the forensic dependency in tampered representation from authentic one within the whole data space, we attempt to optimize the worst corrupted tampered document images with the largest forensic dependency to achieve MI minimization on them. After that, the corrupted images with lower forensic dependency will also be optimized. Inspired by this, we form a min-max optimization as AT (Akhtar et al. 2021) by employing MI between tampered and authentic representation as a criterion to reduce forensic information dependency and further improve robustness against natural corruptions with maintained adversarial robustness. To further enable a trade-off between predictive representations in clean tampered document pixels and robust ones in corrupted pixels, an additional regularization term is formulated with divergence between clean and perturbed pixels distribution.

Our contributions can be summarized as follows:

- We are the first to study the robustness against both adversarial and natural corruptions in document tampering localization (DTL).
- We propose a mutual information (MI) based method to represent forensic dependency under a min-max framework, where the MI is approximated by Jensen-Shannon-Divergence (JSD) with empirical sampling.
- We collect a dataset named by TSorie-CRP for evaluating robustness against various natural and adversarial corruptions, which is constructed from one DTL benchmark dataset TSorie.
- We conduct extensive experiments on the proposed dataset, and the results demonstrate the effectiveness of our method against natural corruptions while maintaining adversarial robustness.

Related Work

Document Tampering Localization

Researches based on document images have been focused for many years due to their sensitive and private contents apart from the uniqueness and extensiveness of their application scenarios. Some researchers noticed the importance of content security of document images and attempted to locate tampered characters by traditional typographical features (Zramdini and Ingold 1998; Satkhozhina, Ahmadullin, and Allebach 2013; Bertrand et al. 2015). However, one common strict constrain or limitation of these methods is that the tampered document images need to be tiny, which can not support them to be applied in real scenarios. With development of deep learning networks and techniques, various works have been proposed based on two-dimension images, such as natural images, face images. Although promising performance has been achieved in tampering localization based on these images, different characteristics between natural and document images decide non-generalizability across two kinds of frameworks. For example, tampered document images usually contain extremely unbalanced numbers of pixels in both categories and less color and texture information for forensics. Motivated by this, (Shao et al. 2023) proposed the framework with forgery traces enhancement and progressive supervision modules to capture tampered pixels in document images. (Qu et al. 2023) used frequency information in DCT domain fused with visual features in RGB domain to search for tampered regions with F-score of nearly 100% in one public dataset.

Adversarial Training

Adversarial Training on Image Classification. Adversarial training (AT) has been demonstrated to be one of effective approaches for adversarial robustness improvement. AT adopts a min-max optimization paradigm with various criterion. For example, in the conventional FGSM-based AT (Goodfellow, Shlens, and Szegedy 2014), classification loss, such as the common used cross entropy loss, is used for min-max optimization. Similar to FGSM-based AT, BIM-based AT and PGD-based (Madry et al. 2017) AT were proposed. AT methods based on min-max optimization with other categories of criterion have been proposed. For example, (Qian et al. 2021) tried to ensure local smoothness of distribution and used it as perturbation to generate AEs trained for adversarial robustness. They make use of properties or information of manifold which is a general term for geometric objects, including curves and surfaces in various dimensions. During min-max optimization, adversarial examples are firstly generated by maximizing a criterion and then minimization with such criterion is conduct on corresponding adversarial examples in defensive stage.

Adversarial Training on Semantic Segmentation. Same with document tampering localization, semantic segmentation focus on information in pixel level. We also put attention on works for adversarial robustness in semantic segmentation. SegPGD (Gu et al. 2022) as the variant of PGD was proposed for adversarial robustness by balancing the importance of the cross-entropy loss of different pixels. With this consideration, pixels which have not been perturbed successfully will be focused during the following attacks. Re-training on such effective adversarial examples will be conducted for further improving robustness. Similarly, (Ag-

nihotri and Keuper 2023) presented a weighted attack loss via cosine similarity between one-hot ground truth class and predicted output after activation function to obtain AEs. Meanwhile (Croce, Singh, and Hein 2023) modified Auto-PGD (APGD) with minimal adjustments for AEs generation, yielding consistent improvements over PGD.

Robustness on Tampering Localization.

In tampering localization on natural images, (Zhuo et al. 2022) depends on adversarial examples with much smaller perturbation to improve the performance on clean natural tampered images instead of delving into adversarial robustness. Recently, (Shao et al. 2024) improves adversarial robustness in DTL systems by maximizing distribution between clean and adversarial examples. However, they only focus on adversarial robustness and do not delve into robustness improvement based on the characteristics of DTL. For example, how forensic information affected by corruption will be related to robustness improvement. Instead, our work aims to improve robustness against natural corruption on document tampering localization models by considering forensic dependency in DTL. (Wu et al. 2022b) proposed a method for improving robustness against corruption after OSNs transmission. Apart that they focus on natural images, they attempt to model the OSN-transmitted noise, add the predicted noise to clean tampered natural images and train on images with such predicted noise to improve robustness. However, we explore robustness of DTL against natural corruption from another perspective, by reducing forensic dependency based on min-max optimization.

Methodology

Vanilla DTL models training aims to identify dissimilarities between tampered and authentic regions for predictive forensic clues. However, sometimes these dissimilarities may not be solely originated from the tampered traces and tend to be correlated with specific patterns in authentic pixels. Therefore, confronted with corrupted tampered document images, forensic features with such dependency tend to negatively affect robustness in DTL systems.

Here, a set of the worst corrupted tampered document images are those with the maximized forensic dependency. If forensic dependency minimization on such worst ones for improving robustness is achieved, the corrupted images with lower forensic dependency should be minimized. This insight is coincided with min-max optimization in AT. Therefore we obtain mutual information to represent the forensic dependency and employ min-max optimization with MI to reduce the forensic dependency. Then we can improve robustness against natural corruption with maintained adversarial robustness. Specifically, our method contains two main stage: (1) Adversarial forensic examples generation based on MI and (2) defensive stage with adversarial forensic regularization and distribution divergence regularization. The whole framework of our method is shown in Fig. 2.

Vanilla DTL Introduction

A clean tampered document image, x, can be easily obtained through methods of conventional image editing, such as Copy-Move, Splicing and Inpainting, on an authentic document image. DTL aims to distinguish tampered pixels from authentic ones in x without additional embedded information for forensics, such as watermark.

Generally, vanilla DTL models mainly contain two components: feature extractor for possible forensic information identification, processing and compression and DTL head for discrimination between tampered pixels and authentic ones by utilizing such forensic information. For example, shown in Fig. 2, $x \in \mathbb{R}^{B\times 3\times H\times W}$ is inputted to the feature extractor ε_θ with multiple layers for predictive forensic representation, $r \in \mathbb{R}^{B\times C\times H\times W}$, by shifting representations from the image space to the informative latent space. Then DTL head $\mathcal{H}_\psi$ utilize r for predicted tampering localization maps, $\hat{p} \in \mathbb{R}^{B\times H\times W}$. Finally, the vanilla document tampering localization (DTL) model is trained by optimizing loss function as follows:

$$\min_{\theta} \mathcal{L}_{DTL}(\varepsilon_\theta \circ \mathcal{H}_\psi; x, GT), \tag{1}$$

where $\varepsilon_\theta \circ \mathcal{H}_\psi$ represents the whole DTL network withε_θ and $\mathcal{H}_\psi$. θ and ψ are parameters of the feature extractor and DTL head respectively. GT represents the ground truth mask of tampered document image x, in which 0 represents authentic pixels and 1 means tampered pixels.

$\mathcal{L}_{DTL}(\cdot)$ can be any classification loss in pixel level. For example Cross-Entropy (CE) is widely used and losses for solving class-imbalance problem may also be adopted to put more attention on tampered pixels, such as Lovasz loss.

Adversarial Examples Generation with MI

We generate the worst corrupted examples with forensic dependency by maximizing mutual information between tampered and authentic latent representation, which is decomposed in two steps: (1) MI estimation for DTL; (2) MI maximization for AEs generation.

MI Estimation for DTL. As defined in Eq. 2, mutual information equals to Kullback-Leibler (KL) divergence between joint probability distribution of two random variables and the product of two random variables' marginal probability distribution. However, it is not tractable for MI computation due to the unknown forms of $\mathbb{J}$ and $\mathbb{M}$ in Eq. 2. Therefore, to find a tractable method and samples from $\mathbb{J}$ and $\mathbb{M}$ are two challenges for MI estimation.

$$\mathcal{I}(Z_1; Z_2) = \mathcal{D}_{KL}(\mathbb{J} \| \mathbb{M}) \tag{2}$$

where $\mathbb{J}$ and $\mathbb{M}$ represent the joint and the product of marginals of random variables, Z_1 and Z_2. Z_1 and Z_2 are over data spaces, Ω_{z_1} and Ω_{z_2} respectively. In the following, we will introduce how to calculate such KL divergence by a tractable method, where the empirical sampling is adopted.

(1) Utilizing a tractable method. In our case, the two variables are $\mathcal{R}_T^{i,j}$ and $\mathcal{R}_A^{i,j}$ over tampered representation space, Ω_T, and authentic representation space, Ω_A, respectively in pixel level. MI is estimated by maximizing a lower bound based on Jensen–Shannon divergence (JSD). Borrowed from (Hjelm et al. 2019), we use a JSD-based MI

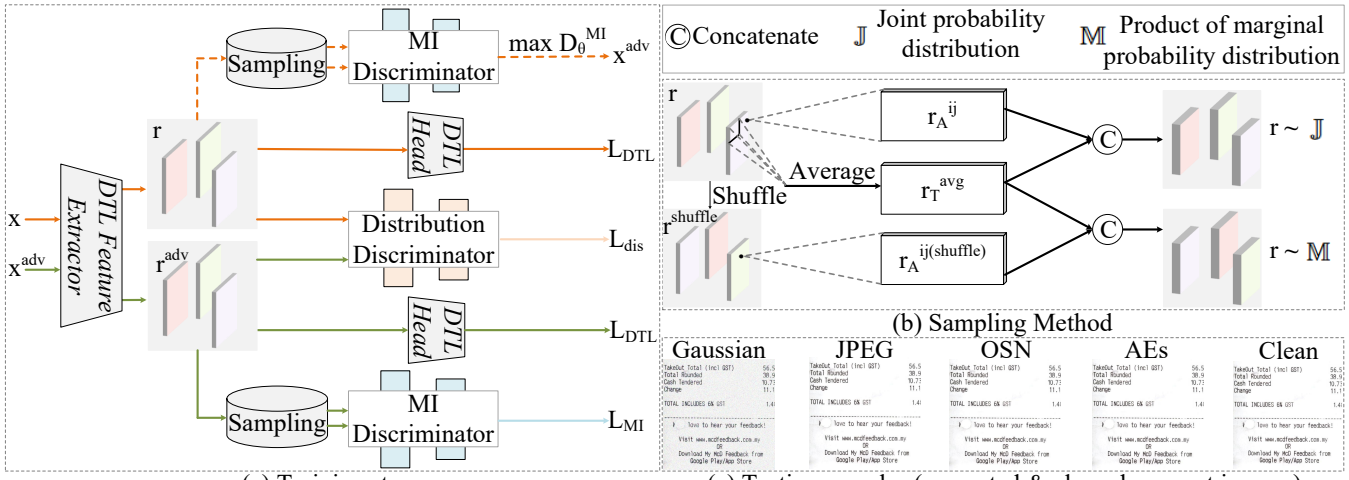

Figure 2: The framework of the proposed method. (a) Training stage: AEs are obtained by MI maximization and update parameters of DTL by MI minimization and distribution divergence minimization with $\mathcal{L}_{MI}$ and $\mathcal{L}_{dis}$ besides $\mathcal{L}_{DTL}$. (b) Sampling method for MI estimation with such empirical samples (Hjelm et al. 2019). (c) Testing samples w/ and w/o corruption.

estimator defined as

$$\mathcal{D}_{KL}(\mathbb{J}\|\mathbb{M}) \geq \hat{\mathcal{I}}_\omega^{JSD}(\mathcal{R}_A^{i,j};\mathcal{R}_T^{i,j}) \triangleq \mathbb{E}_{\mathbb{J}}\{g_f(D_\omega(r^{i,j}))\} - \mathbb{E}_{\mathbb{M}}\{f^*(g_f(D_\omega(r^{i,j})))\}, \tag{3}$$

where $g_f(u) = \log 2 - log(1+e^{-u})$ and $f^*(t) = -\log(2-e^t)$ for JSD estimation (Nowozin, Cseke, and Tomioka 2016). $r^{i,j}$ is the representation from the joint probability distribution between $\mathcal{R}_A^{i,j}$ and $\mathcal{R}_T^{i,j}$, $\mathbb{J}$, or the product of marginals of $\mathcal{R}_A^{i,j}$ and $\mathcal{R}_T^{i,j}$, $\mathbb{M}$. D_ω is a discriminator function modeled by a neural network with parameters ω to distinguish $r^{i,j} \sim \mathbb{J}$ from $r^{i,j} \sim \mathbb{M}$.

(2) Empirical sampling $r^{i,j}$ from $\mathbb{J}$ and $\mathbb{M}$. To face the second challenge of obtaining samples from $\mathbb{J}$ and $\mathbb{M}$, we conduct sampling within one batch of latent representation maps, which is inspired by (Hjelm et al. 2019). One example of sampling on the original location of authentic pixel is depicted in Fig. 2(b).

Given a batch of latent representation maps, $r \in \mathbb{R}^{B\times C\times H\times W} \triangleq \{r_b \in \mathbb{R}^{C\times H\times W}, b = 0,1,...B-1\}$, the averaged tampered representation, $r_T^{avg} \in \mathbb{R}^{B\times C} \triangleq \{r_{b,T}^{avg} \in \mathbb{R}^{1\times C}, b = 0,1,...B-1\}$ and $r_A^{avg} \in \mathbb{R}^{B\times C} \triangleq \{r_{b,A}^{avg} \in \mathbb{R}^{1\times C}, b = 0,1,...B-1\}$ are first obtained. By concatenating $r_{b,T}^{avg}$ with representation of each authentic pixel, $r_{b,A}^{i,j} \in \mathbb{R}^{1\times C}$, and concatenating $r_{b,A}^{avg}$ with representation of each tampered pixel, $r_{b,T}^{i,j} \in \mathbb{R}^{1\times C}$, we can obtain $r_{\mathbb{J}} \in \mathbb{R}^{B\times 2C\times W\times H} \triangleq \{[r_{b,A}^{i,j}, r_{b,T}^{avg}] \cup [r_{b,A}^{avg}, r_{b,T}^{i,j}] \in \mathbb{R}^{2C\times W\times H}, b = 0,1,...B-1\}$ from $\mathbb{J}$. Then, $\widetilde{r}$ is obtained after we shuffle r according to batch dimension. By concatenating previous $r_{b,T}^{avg}$ with representation of each authentic pixel, $\widetilde{r}_{b,A}^{i,j} \in \mathbb{R}^{1\times C}$, and concatenating previous $r_{b,A}^{avg}$ with representation of each tampered pixel, $\widetilde{r}_{b,T}^{i,j} \in \mathbb{R}^{1\times C}$, we can obtain $r_{\mathbb{M}} \in \mathbb{R}^{B\times 2C\times W\times H} \triangleq \{[\widetilde{r}_{b,A}^{i,j}, r_{b,T}^{avg}] \cup [r_{b,A}^{avg}, \widetilde{r}_{b,T}^{i,j}] \in \mathbb{R}^{2C\times W\times H}, b = 0,1,...B-1\}$ from $\mathbb{M}$. Finally, Eq. 3 can be rewritten as

$$\hat{\mathcal{I}}_\omega^{JSD}(\mathcal{R}_A^{i,j};\mathcal{R}_T^{i,j}) \triangleq \mathbb{E}_{\mathbb{J}}\{g(D_\omega(r_{\mathbb{J}}^{i,j}))\} - \mathbb{E}_{\mathbb{M}}\{g(D_\omega(r_{\mathbb{M}}^{i,j}))\}, \tag{4}$$

where $r_{\mathbb{J}}^{i,j}$ and $r_{\mathbb{M}}^{i,j}$ are empirical samples from $\mathbb{J}$ and $\mathbb{M}$ respectively. $[\cdot]$ means concatenate operator.

MI Maximization. By further generating data that are more similar between tampered and authentic pixels in latent space, the worst-case adversarial data with respect to the entire representation dependency are obtained by maximizing

$$x^{adv} = \underset{x^{adv}\in\mathcal{B}(x^{adv},\epsilon)}{\arg\max}\ \hat{\mathcal{I}}_\omega^{JSD}(\varepsilon_\theta(X_T^{0;i,j});\varepsilon_\theta(X_A^{0;i,j})), \tag{5}$$

where $X_T^{0;i,j}$ and $X_A^{0;i,j}$ are two random variables over tampered pixel space, $\mathcal{X}_T^0$ and authentic pixels space, $\mathcal{X}_A^0$ in initially sampled process; $\mathcal{B}(x^{adv},\epsilon) \triangleq x^{adv} : \|x^{adv} - x\|_p \leq \epsilon$; ε_θ is the feature extractor in DTL model. Specifically, in our experiments, adversarial examples are generated by maximizing $\mathcal{L}_{MI}$ in Eq. 8.

Defensive Regularization

Adversarial Forensic Regularization. Correspondingly, in order to further improve robustness over representation dependency, the model is regularized by minimizing the mutual information between the tampered and the adversarial representation. The key insight is to enable the model to exploit robustness well to the representation dependency worsen by perturbations, rather than only considering the error rate at a particular data point (commonly used in the label-guided methods). The optimization problem of the

proposed adversarial forensics regularization is shown as follows:

$$\min_{\omega} \hat{\mathcal{I}}_{\omega}^{JSD}(\varepsilon_\theta(X_T^{adv;i,j}); \varepsilon_\theta(X_A^{adv;i,j}))$$
$$\text{s.t.} x^{adv} = \underset{x^{adv} \in B(x^{adv}, \epsilon)}{\arg\max} \hat{\mathcal{I}}_{\omega}^{JSD}(\varepsilon_\theta(X_T^{0;i,j}); \varepsilon_\theta(X_A^{0;i,j})). \tag{6}$$

Distribution Divergence Regularization. In order to further obtain better robust generalization over unseen corruption and promote a better trade-off between clean data and the adversarial data, the DTL model is regularized by minimizing the distribution distance between the clean representation $\varepsilon_\theta(x)$ for $x \sim \mathbb{P}$ and the adversarial representation $\varepsilon_\theta(x)$ for $x \sim \mathbb{Q}$ by a discriminator D_ω termed as distribution regularizer. The key insight is to enable the model to generalize well to the new distribution where the adversarial data are located. The optimization problem of the proposed regularization term is shown as

$$\begin{aligned}\mathcal{D}_{JS}(\mathbb{P}\|\mathbb{Q}) \geq J\hat{S}D(\mathbb{P}\|\mathbb{Q}) \triangleq \mathbb{E}_{x\sim\mathbb{P}}\{g_f(D_\omega(\varepsilon_\theta(x)))\} \\ - \mathbb{E}_{x\sim\mathbb{Q}}\{f^*(g_f(D_\omega(\varepsilon_\theta(x))))\}.\end{aligned} \tag{7}$$

Loss function. According to the derivation in Supplementary Materials, the minimization of $\hat{\mathcal{I}}_{\omega}^{JSD}$ in Eq. 3 for adversarial forensic regularization can be achieved by minimizing

$$\begin{aligned}\mathcal{L}_{MI} = &\frac{1}{W \times H} \sum_{i,j} \{\log \sigma(D_\omega(r_{\mathbb{J}}^{i,j})) \\ &+ \log \sigma(1 - D_\omega(r_{\mathbb{M}}^{i,j}))\},\end{aligned} \tag{8}$$

and $J\hat{S}D(\mathbb{P}\|\mathbb{Q})$ minimization in Eq. 7 can be achieved by

$$\begin{aligned}\min_{\theta} \mathcal{L}_{dis} = &\frac{1}{W \times H} \sum_{i,j} \{\log \sigma(D_\omega(\varepsilon_\theta(x))) \\ &+ \log \sigma(1 - D_\omega(\varepsilon_\theta(x^{adv})))\},\end{aligned} \tag{9}$$

where σ is the Sigmoid function.

In summary, the DTL model is trained by minimizing the total loss function given by

$$\mathcal{L} = \mathcal{L}_{DTL} + \lambda_1 \mathcal{L}_{MI} + \lambda_2 \mathcal{L}_{dis} \tag{10}$$

where $\mathcal{L}_{DTL}$ is the classification loss in pixel level, $\mathcal{L}_{MI}$ denotes MI loss between tampered and authentic representation in DTL, λ_1 and λ_2 are balancing parameters of the loss.

Experiments

Experimental Setting

Datasets. We conduct experiments on the benchmark TSroie dataset (Wang et al. 2022b) for delving into robustness in DTL. Further, We collect a testing dataset with various corruption based on the testing set in TSroie for robustness evaluation in DTL system, which is termed as TSroie-CRP.

TSroie contains 626 training images and 360 testing images with larger sizes. It is collected in 2022 with only the type of bill document images. They firstly use SRNet to automatically modify original words with target contents and then refine some hard samples with low-quality visualization by photoshop (PS). It serves as clean dataset in the following experiments.

Corruption Types			Number
Natural corruption	Noise	Gaussian noise	360
	Compression	JPEG	360
	OSN	WeChat	360
Adversarial corruption	Single-step	FGSM	360
	Multi-step	BIM	360
	Multi-step	PGD	360
	Multi-step	APGD	360

Table 1: Details of TSroie-CRP Dataset.

TSroie-CRP contains natural and adversarial corruption. For natural corruption, we consider three categories: noise, compression and online social networks (OSNs). We select the representative in each category, such as gaussian noise, JPEG compression and WeChat (performance on other types of OSN is shown in Supplementary Materials.). More details are indicated in Table. 1.

Evaluation Metrics. In this paper, we utilize F-score to evaluate the performance of document tampering localization. We firstly report the clean F-score which measure the performance on the clean samples without any corruption. In addition, we report F-scores to evaluate model robustness under natural corruption with hand-crafted types, such as JPEG and gaussian noise, and one type of OSNs. Adversarial robustness is also evaluated on different attacking methods including single step attack FGSM and several iterative attacks, such as BIM, PGD and APGD.

Implementation Details. We conduct experiments based on one valilla DTL model and use input images with size 512×512. We use SGD optimizer for updating parameters in discriminator networks. In min-max optimization, we set attack step to 1 when generating AEs based on MI. We leverage several convolutional layers as the basic network structure of our MI discriminator and distribution discriminator in this paper.

Main Results

Robustness Against Natural Corruption. We choose another five AT methods for comparing on robustness against natural corrupted tampered document images on TSroie-CRP dataset. Two methods are proposed for the task of image classification, PGD-AT and Trades-AT and the other two methods, DDC-AT and SegPGD-AT, are proposed for semantic segmentation. Additionally, we also compare robustness against natural corruption with one recent AT method in DTL for adversarial robustness, LM-AT. The results are shown in Tab. 2. We use bold for the first best performance and underline the second best performance in each column.

Trades-AT forms min-max optimization with sample-wise KL-divergence between adversarial and clean data. LM-AT considers the relationship between clean and AEs

Methods	Clean	JPEG (Quality)			GN (σ)			OSN
		95	85	75	10	20	30	WeChat
Vanilla DTL	**95.2**	56.9	7.2	4.3	62.4	33.3	2.5	46.9
PGD-AT	83.4	79.4	67.2	63.3	45.7	32.5	17.1	45.1
Trades-AT	92.9	87.2	77.9	74.6	70.8	41.5	6.5	64.8
DDC-AT	89.9	86.7	**82.0**	81.0	63.2	68.7	38.5	66.1
SegPGD-AT	83.9	43.4	49.2	56.9	22.2	10.2	8.6	34.8
LM-AT	85.9	80.9	80.7	80.7	77.8	60.6	27.9	65.7
Ours	90.9	**87.3**	81.9	**82.4**	**83.0**	**81.4**	**77.6**	**69.8**

Table 2: Comparing with AT Methods against natural corruption (F1 (%)).

for perturbation generation and minimizes DTL loss on such AEs. PGD-AT, DDC-AT and SegPGD-AT are based on min-max optimization with DTL loss. They all ignore the forensic dependency between tampered and authentic pixels. Our method based on min-max optimization with MI regularization can promote forensic information more robust and independent. The results show its validity. With larger σ in gaussian noise, the performance becomes worse for all AT method. Comparing with other AT method, our method indicate more robustness against gaussian noise. For example, with $\sigma = 30$, our method achieves 77.6% which increases 101.6% comparing with the second largest F_1 in DDC-AT. For robustness against JPEG compression, our result shows a competitive performance. For OSN-transmitted corruption, our method achieves the best performance with F_1, 69.8%, which is improved by 46.9% than vanilla DTL model. Moreover, the trade-off of performance between clean and corrupted tampered document images is better than the other methods.

Robustness Against Adversarial Corruption. Adversarial examples (AEs) are another important threats faced by DTL models. To evaluate the adversarial robustness of our MI-based min-max optimization, we conduct experiments by comparing with other AT methods on TSroie under four adversarial attacks: FGSM, BIM, PGD and APGD.

We use bold to indicate the first best result and underline the second best result in Tab. 3. Our method indicate the best robustness against adversarial attacks compared with the other method, and the performance in clean tampered document images is also better maintained. For example, for FGSM attack, our method is larger 2.6% than the second largest F_1 in LM-AT, but our method is improved 5% compared with LM-AT in clean tampered document images.

Ablation Study

Data Augmentation. Noise and compression are two common categories of simple natural corruption. Among them, we select Gaussian noise and JPEG compression as representative hand-crafted natural corruption based on TSroie. Images after transmission through OSNs will be corrupted with various known image processing operations for fast transmission, such as JPEG, resize, and other unknown operations. To evaluate robustness against natural corruption in real scenarios, we also test corrupted images after OSNs

Methods	Clean	FGSM	BIM	PGD	APGD
Vanilla DTL	**95.2**	21.9	16.7	14.6	4.7
PGD-AT	83.4	77.3	78.7	77.8	49.2
Trades-AT	92.9	34.7	46.4	42.8	37.6
DDC-AT	89.9	68.0	75.0	69.0	26.0
SegPGD-AT	83.9	78.1	75.8	75.2	51.0
LM-AT	85.9	82.5	81.4	81.5	51.1
Ours	90.9	**85.1**	**83.4**	**83.5**	**53.2**

Table 3: Comparing with AT Methods against adversarial corruption (F1 (%)).

Methods	Clean	JPEG (Quality)			GN (σ)			OSN
		95	85	75	10	20	30	WeChat
Vanilla DTL	95.2	56.9	7.2	4.3	62.4	33.3	2.5	46.9
+JPEG-95	66.1	95.3	61.7	36.4	1.4	0.0	0.0	50.3
+JPEG-85	81.1	**86.9**	91.4	**81.8**	0.1	0.0	0.0	56.8
+JPEG-75	78.0	79.3	**82.0**	85.4	71.2	57.8	42.9	**62.2**
+GN-10	**83.7**	70.0	79.3	79.2	85.6	79.0	63.5	55.7
+GN-20	9.4	71.9	28.9	79.0	**85.5**	84.8	**82.7**	8.5
+GN-30	29.3	40.6	53.3	29.7	**84.4**	**84.5**	84.3	28.8
Ours	**90.9**	**87.3**	**81.9**	**82.4**	83.0	**81.4**	**77.6**	**69.8**

Table 4: Ablation study of various data augmentation against natural corruption (F1 (%)).

transmission, such as WeChat which is extensively used in real scenarios.

The results are shown in Tab. 4. Apart that we underline the results on the known corruption during training, we also use bold on the other two best performance. The robustness against specific natural corruption can be improved by data augmentation on such natural corruption with specific parameters. But when evaluated with unknown natural corruption during training, the robustness may not be maintained. For example, 95.3% can be achieved on testing data after JPEG compression with quality factor (QF), 95, after training on document tampered images compressed with the same value. But the performance on JPEG images with different QF and other type of natural corruption is out of satisfactory, such as 36.6% on JPEG images (QF=75), and 0% on images with gaussian noise ($\sigma = 20, 30$). Compared with such methods for robustness, Our method shows consistent robustness on natural corruption with various parameters. For example, 87.3% on JPEG images (QF=95) and 82.4% on JPEG images (QF=75) and 83.0% on corrupted images with gaussian noise ($\sigma = 10$).

Additionally, for tampered document images with OSN-transmitted corruption, fine-tuning with data augmentation on specific natural corruption indicates instability. For example, 8.5% is achieved by fine tuning on corrupted tampered document images with gaussian noise (σ=20). 62.2% is achieved by fine tuning on corrupted tampered document images with JPEG tampered document images (QF=75). F_1 is 69.8% in our method which is the best performance compared with other methods. The results in Tab. 4 indicate the effectiveness of our method.

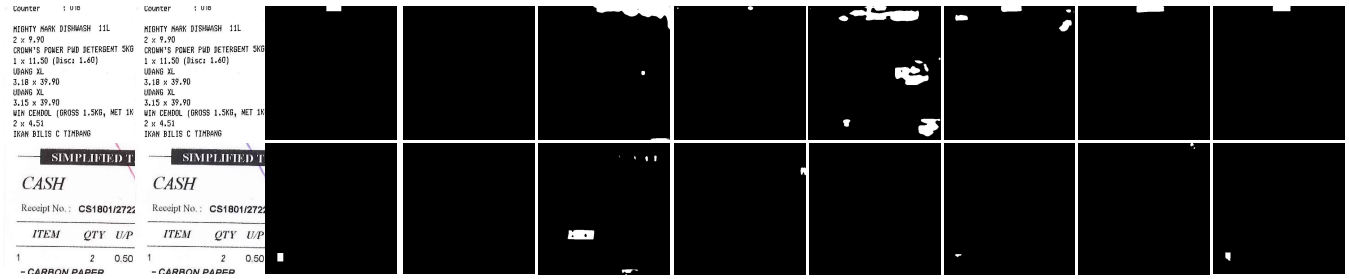

Figure 3: Visualization of predicted localization maps against PGD-attacked corruption in the first row and OSN-transmitted corruption in the second row with different models. From left to right: clean tampered images, corrupted tampered images, ground-truth masks and predicted maps by Vanilla, PGD-AT, DDC-AT, SegPGD-AT, Trades-AT, LM-AT, Our-AT models.

Methods	Clean	WeChat
Vanilla DTL	95.2	46.9
+ MI maximization	86.9	67.8
+ AFR (+MI min-max)	89.7	69.7
+ MI maximization + DDR	92.6	52.2
+ AFR + DDR	90.9	69.8

Table 5: Ablation study against OSN-wise natural corruption (F1 (%)).

Network Components. To evaluate the effectiveness of adversarial forensic regularization (AFR) with min-max optimization for robustness against OSN-wise natural corruption and distribution divergence regularization (DDR) for better trade-off between clean and corrupted pixels. We conduct experiments on five frameworks based on VF: (1) vanilla DTL model; (2) vanilla DTL model only with MI maximization; (3) vanilla DTL model only with AFR; (4) vanilla DTL model only with DDR; and (5) vanilla DTL model only with AFR and DDR (ours).

The results in the five frameworks shown in Tab. 5 indicate the effectiveness of both AFR and DDR. In the vanilla DTL model, the best F_1 is achieved with 95.2%, but the performance against OSN-transmitted corruption largely drops to 46.9%. After training on AEs with MI maximization, the performance on OSN-transmitted images is improved by 46.9%, but F_1 drops 8.3%. After training on framework (3) with AFR, robustness against OSN-transmitted corruption is consistently improved. Under training on AEs with MI maximization and with DDR, the performance on clean images is largely improved. By optimization with AFR and DDR can achieve the best robustness against OSN-transmitted corruption with maintained clean performance.

Visualization

We visualize the predicted localization maps with different AT methods for indicating the robustness against PGD adversarial attack and OSN-transmitted corruption.

In general, we expected the predicted localization maps to be as similar with GT as possible, and we can see that the prediction of our method matches the GT much better than others. During adversarial examples generation in the other five AT methods. AEs are obtained without considering forensic information. Therefore, the forensic dependency is not considered and the results are unstable during the defensive process for parameters update of tampering localization models. For example, in Fig. 3, the qualitative result of LM-AT against adversarial corruption is good, but the result in the OSN-transmitted corruption are bad with more false negatives. In our method, we consider forensic dependency between tampered and authentic representation during min-max optimization, therefore, the parameters update in defensive process will reduce forensic dependency for more robust and independent forensic representation.

Conclusion and Limitations

In this paper, we uncover the vulnerability of existing document tampering localization models against natural corruption, like OSN transmission. In pursuit of robust DTL, we propose a mutual information (MI) based method to represent forensic dependency between tampered and authentic pixels. We adopt Jensen-Shannon- Divergence (JSD) with empirical sampling to calculate MI efficiently under the min-mix optimization framework. To maintain the adversarial robustness, we also integrate the distribution divergence between clean and adversarial samples. To promote the research on the robustness of DTL, we collect a dataset named by TSorie-CRP including various both natural and adversarial corruptions, which is constructed from one DTL benchmark dataset TSorie. In the experiments, we analyze results and verify its effectiveness on the benchmark dataset and state-of-the-art DTL model, where the robustness is improved for both various natural and adversarial corruptions. In our dataset TSorie-CRP, we only considered limited number of natural corruptions. In the future, we will extend our dataset with more corruptions and more samples.

Acknowledgments

The work was partially supported by the following: National Natural Science Foundation of China under No. 92370119, 62276258 and 62376113, XJTLU Funding REF-22-01-002, and Suzhou Municipal Key Laboratory for Intelligent Virtual Engineering (SZS2022004).

References

Agnihotri, S.; and Keuper, M. 2023. CosPGD: a unified white-box adversarial attack for pixel-wise prediction tasks. *arXiv preprint arXiv:2302.02213*.

Akhtar, N.; Mian, A.; Kardan, N.; and Shah, M. 2021. Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access*, 9: 155161–155196.

Bammey, Q.; Gioi, R. G. v.; and Morel, J.-M. 2020. An adaptive neural network for unsupervised mosaic consistency analysis in image forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14194–14204. IEEE.

Bertrand, R.; Terrades, O. R.; Gomez-Krämer, P.; Franco, P.; and Ogier, J.-M. 2015. A conditional random field model for font forgery detection. In *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, 576–580. IEEE.

Croce, F.; Singh, N. D.; and Hein, M. 2023. Robust Semantic Segmentation: Strong Adversarial Attacks and Fast Training of Robust Models. *arXiv preprint arXiv:2306.12941*.

Dong, C.; Chen, X.; Hu, R.; Cao, J.; and Li, X. 2022. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3539–3553.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Gu, J.; Zhao, H.; Tresp, V.; and Torr, P. H. 2022. Segpgd: An effective and efficient adversarial attack for evaluating and boosting segmentation robustness. In *European Conference on Computer Vision*, 308–325. Springer.

Guo, X.; Liu, X.; Ren, Z.; Grosz, S.; Masi, I.; and Liu, X. 2023. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3155–3165.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2019. Learning deep representations by mutual information estimation and maximization. In *International Conference on Learning Representations*.

Kwon, M.-J.; Yu, I.-J.; Nam, S.-H.; and Lee, H.-K. 2021. 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*, 375–384.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Nowozin, S.; Cseke, B.; and Tomioka, R. 2016. f-gan: Training generative neural samplers using variational divergence minimization. *NeurIPS*, 29.

Qian, Z.; Zhang, S.; Huang, K.; Wang, Q.; Zhang, R.; and Yi, X. 2021. Improving model robustness with latent distribution locally and globally. *arXiv preprint arXiv:2107.04401*.

Qu, C.; Liu, C.; Liu, Y.; Chen, X.; Peng, D.; Guo, F.; and Jin, L. 2023. Towards Robust Tampered Text Detection in Document Image: New Dataset and New Solution. In *CVPR*, 5937–5946.

Satkhozhina, A.; Ahmadullin, I.; and Allebach, J. P. 2013. Optical font recognition using conditional random field. In *Proceedings of the 2013 ACM symposium on Document engineering*, 119–122.

Shao, H.; Huang, K.; Wang, W.; Huang, X.; and Wang, Q. 2023. Progressive Supervision for Tampering Localization in Document Images. In *ICONIP*, 140–151. Springer.

Shao, H.; Qian, Z.; Huang, K.; Wang, W.; Huang, X.; and Wang, q. 2024. Delving into Adversarial Robustness on Document Tampering Localization. In *European Conference on Computer Vision*. ECVA.

Wang, J.; Li, Z.; Zhang, C.; Chen, J.; Wu, Z.; Davis, L. S.; and Jiang, Y.-G. 2022a. Fighting Malicious Media Data: A Survey on Tampering Detection and Deepfake Detection. *arXiv preprint arXiv:2212.05667*.

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

Wu, H.; Zhou, J.; Tian, J.; and Liu, J. 2022a. Robust image forgery detection over online social network shared images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13440–13449.

Wu, H.; Zhou, J.; Tian, J.; Liu, J.; and Qiao, Y. 2022b. Robust image forgery detection against transmission over online social networks. *IEEE Transactions on Information Forensics and Security*, 17: 443–456.

Zhang, H.; and Wang, J. 2019. Defense against adversarial attacks using feature scattering-based adversarial training. *Advances in Neural Information Processing Systems*, 32.

Zhuo, L.; Tan, S.; Li, B.; and Huang, J. 2022. Self-adversarial training incorporating forgery attention for image forgery localization. *IEEE Transactions on Information Forensics and Security*, 17: 819–834.

Zramdini, A.; and Ingold, R. 1998. Optical font recognition using typographical features. *IEEE Transactions on pattern analysis and machine intelligence*, 20(8): 877–882.