DITL^2: Dual-Stage Invariance Transfer Learning for Generalizable Document Image Tampering Localization

Songze Li*
Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen Key Laboratory of Media Security, Shenzhen University
Shenzhen, China
2310433002@email.szu.edu.cn

Yunfei Guo*
Shen Chen
Tencent YouTu Lab
Shanghai, China
rinveyguo@tencent.com
kobeschen@tencent.com

Bin Li†
Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen Key Laboratory of Media Security, Shenzhen University
Shenzhen, China
libin@szu.edu.cn

Kaiqing Lin
Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen Key Laboratory of Media Security, Shenzhen University
Shenzhen, China
linkaiqing2021@email.szu.edu.cn

Changsheng Chen
Haodong Li
Guangdong Provincial Key Laboratory of Intelligent Information Processing, Shenzhen Key Laboratory of Media Security, Shenzhen University
Shenzhen, China
cschen@szu.edu.cn
lihaodong@szu.edu.cn

Taiping Yao
Shouhong Ding
Tencent YouTu Lab
Shanghai, China
taipingyao@tencent.com
ericshding@tencent.com

Abstract

Document Image Tampering Localization (DITL) advances considerably, yet achieving robust cross-dataset generalization remains a formidable challenge for practical applications. Expanding existing document datasets for training is labor-intensive, making it appealing to incorporate data from non-document domains such as natural scene images. However, domain-specific variations, including differences in color distribution and texture, compromise the performance of joint training. To address this issue, we propose DITL^2, a **D**ual-stage **I**nvariance **T**ransfer **L**earning framework for **D**ocument **I**mage **T**ampering **L**ocalization that consists of Cross-Domain Invariance Pre-training (CDIP) and Frequency Decoupling Parameter Adaptation (FDPA). In the pre-training stage, CDIP employs style transfer and texture consistency learning to suppress domain-specific influences from tampered natural scene images, and tampering trace commonality learning to acquire domain-invariant features. In the fine-tuning stage, FDPA adapts the parameters of the pre-trained model, leveraging the general knowledge from the pre-trained model to address DITL tasks while reducing the risk of overfitting. Experiments show that this approach effectively leverages external data resources to boost model performance, achieving state-of-the-art results across a variety of cross-dataset settings.

CCS Concepts

• **Security and privacy** → **Malware and its mitigation**; • **Computing methodologies** → **Artificial intelligence**.

Keywords

Document Image Tampering Localization, Generalization, Parameter-Efficient Fine-Tuning

ACM Reference Format:
Songze Li, Yunfei Guo, Shen Chen, Bin Li, Kaiqing Lin, Changsheng Chen, Haodong Li, Taiping Yao, and Shouhong Ding. 2025. DITL^2: Dual-Stage Invariance Transfer Learning for Generalizable Document Image Tampering Localization. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3754857

1 Introduction

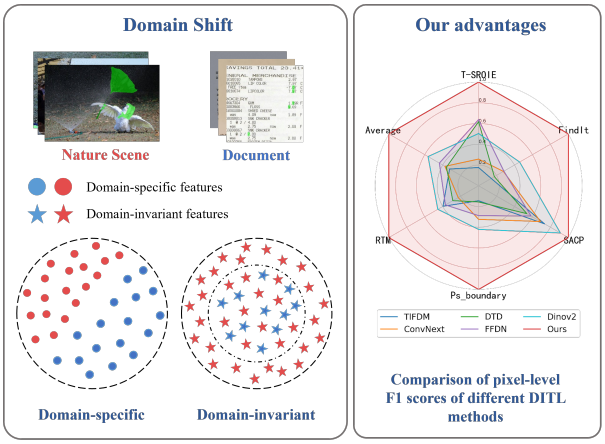
Document images serve as a critical medium for information transmission in the digital era and are extensively utilized in applications such as e-commerce, medical record management, and legal documentation. These images often contain sensitive information, making authenticity essential. Furthermore, the advancement of image editing tools significantly increases the risk of malicious tampering, posing substantial security threats. Consequently, developing robust methods for Document Image Tampering Localization (DITL) is imperative to safeguard the credibility of document images.

To simulate real-world forgery scenarios, several studies [8, 19, 28] propose generating document tampering datasets for training,

*Equal contribution. Work done during internship at Tencent YouTu Lab.
†Corresponding author.

MM '25, Dublin, Ireland

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754857



(a) Enhanced generalization (b) Better generalization

Figure 1: (a) The red region represents natural scene images, while the blue region corresponds to document images. Although the two types of images exhibit significant differences in domain-specific features, they share commonalities in domain-invariant features. The green highlighted areas in the images indicate the tampered regions. (b) Our method demonstrates significantly improved generalization performance for DITL compared to existing approaches.

which greatly advances the development of DITL. However, while existing DITL methods perform well on public datasets, the significant deviation between these simulated datasets and actual tampered data leads to poor generalization in practical applications. Furthermore, given the diverse forms of these documents—such as IDs, contracts, medical records, and other confidential materials—they inherently contain private information. Collecting and constructing high-quality tampered document datasets is both costly and challenging. The scarcity of data impedes performance enhancements in existing models. Considering the availability of high-quality tampering datasets for natural scene images, and inspired by the data scaling law [7, 10, 12], we hypothesize that incorporating these external datasets into training addresses this issue.

As shown in Figure 1(a), tampering traces in natural scene and document images share common characteristics, referred to as domain-invariant features. However, apart from geometry and semantic information, significant domain gaps exist between natural scene images and document images in terms of low-level characteristics such as color and texture. Directly training on both types of images simultaneously negatively impacts DITL performance.

To address these challenges, we propose a dual-stage framework for document image tampering localization. In the pre-training stage, we introduce a novel paradigm called Cross-Domain Invariance Pre-training (CDIP), which leverages a large amount of tampered natural scene images as training data and a small number of unlabeled document images as style references. Stylized images are generated through cross-domain style transfer, reducing color space differences at the image level. Then texture consistency learning is used to minimize texture differences at the feature level and tampering commonality learning is used to capture domain-invariant tampering features shared across image domains. In the fine-tuning stage, we propose a parameter-efficient fine-tuning method called Frequency Decoupled Parameter Adaptation (FDPA). This method introduces a set of trainable parameters to refine feature maps at each layer of the encoder. To enhance the features extracted from image encoders, which encoding spatial and semantic information, we decompose them into high- and low-frequency components, allowing the model to suppress high-frequency noise and focus on low-frequency discontinuities for the DITL task. As shown in Figure 1(b), extensive experiments across multiple cross-dataset scenarios demonstrate that our method consistently outperforms state-of-the-art approaches.

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

- We propose CDIP, a paradigm that incorporates high-quality natural scene image datasets into the DITL task by reducing the model's reliance on domain-specific features and focusing on learning domain-invariant features.
- We introduce a parameter-efficient fine-tuning method, FDPA, which decomposes the features extracted from the pre-trained encoder into high-frequency and low-frequency components, enhancing the tampering features and enabling better adaptation to the DITL task.
- Comprehensive experiments demonstrate that our method surpasses state-of-the-art approaches, and significantly improves detection performance under cross-dataset settings.

2 Related Works

2.1 Document Image Tampering Localization

DITL task aims to precisely locate tampered regions in document images by detecting subtle tampering traces. Several studies [5, 28, 40, 41, 44] have introduced frequency domain information as auxiliary clues to detect invisible tampering traces. DTD [28] designed a frequency-aware head and introduced a curriculum learning mechanism to address the challenges posed by image compression. FFDN [5] preserved high-frequency details during downsampling through wavelet-like frequency enhancement, focusing on subtle but critical tampering clues. Other works [8, 17, 19, 29, 33, 45] have emphasized the importance of text regions in document images.

Despite these advancements, existing methods suffer from significant performance degradation in practical applications. This highlights the need for further research on improving the generalization capabilities of DITL methods.

2.2 Pre-training in Image Forgery Localization

In the field of Image Forgery Localization (IFL), pre-training has been used to acquire general forensic knowledge, which can subsequently be fine-tuned for downstream tasks. DiffForensics [46] employed a self-supervised denoising diffusion pre-training paradigm to focus on mesoscopic image characteristics. SAFIRE [16] pre-trained an image encoder using region-to-region contrastive learning, ensuring embeddings from the same source region are close and the different regions are distant. Okamoto et al. [25] utilized SimCLR [4] as a pre-training framework, performing self-supervised learning to extract features useful for downstream tasks. Qu et al. [30] proposed the Texture Jitter pre-training paradigm that simulates tampered images through simple online augmentations.

Although previous work [25] has primarily focused on introducing additional real images to enhance generalization. However,

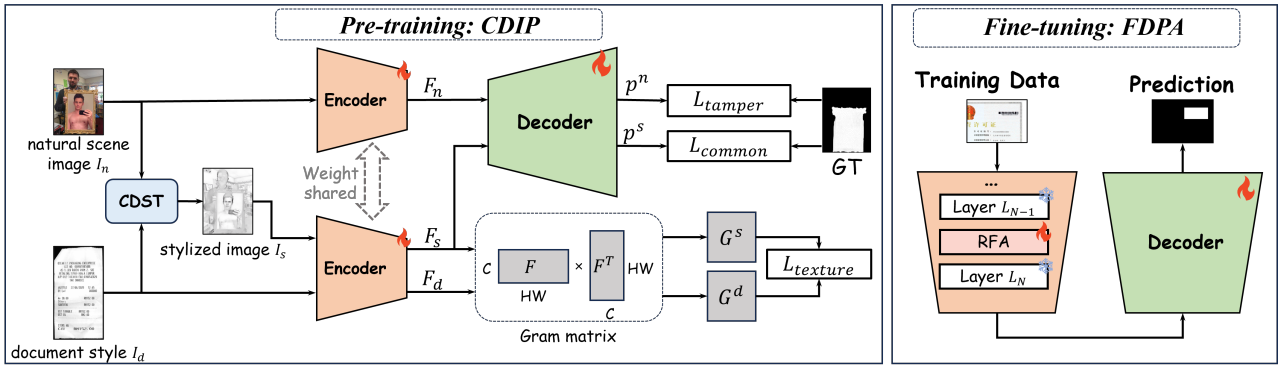


Figure 2: The proposed DITL^2 overall framework divides the training into two stages, called CDIP in the pre-training stage and FDPA in the fine-tuning stage. The example of natural scene image is taken from MIML [31].

there is currently no research on leveraging additional data containing forgeries for training to improve performance. Furthermore, the significant domain gap between data from different fields makes direct joint training problematic, as it may lead the model to learn irrelevant features that are not conducive to the DITL task.

2.3 Domain Generalization

Domain generalization (DG) aims to enhance model performance on unseen target domains by training on multiple source domains. In semantic segmentation, prior works [13, 27, 43, 47] have focused on eliminating domain-specific features to enable the learning of domain-invariant features.

Despite recent advancements, DG remains an underexplored area in the context of DITL. In existing DG tasks, images from the source domain and those from the target domain typically share the same semantics; for example, both may consist of facial images used for anti-spoofing. In contrast, our task involves source domain images that are significantly different from the target domain images in terms of semantics and image structures. However, they do share a commonality in their low-level tampering traces, which could be leveraged for DG. Our approach aims to develop tailored approaches that can effectively bridge this gap and enhance the generalization capabilities of DITL methods.

3 Method

3.1 Overview

Inspired by the data scaling law [7, 10, 12], which states that increasing the amount of training data significantly improves model performance, we incorporate external natural scene tampered image data to enhance the performance of DITL. However, a significant domain gap exists between natural scene images and document images, particularly in terms of color distribution and texture, which hinders the application of external data. To address this issue, as illustrated in Figure 2, our DITL^2 avoids direct joint training with mixed datasets and instead adopts a dual-stage training approach: a pre-training stage for learning generic features related to tampering localization using external natural images, and a fine-tuning stage for adapting the pre-trained model equipped with generic tampering knowledge to DITL tasks.

In the pre-training stage, external natural scene images are utilized along with some unlabeled document images as style references to pre-train the encoder model. We propose a novel paradigm called Cross-Domain Invariance Pre-training (CDIP), which leverages auxiliary training by transforming external data into stylized images resembling document images. This transformation specifically refers to narrowing the stylistic differences between natural scene images and document images, which are typically more structured, rather than reducing the diversity of document image styles, while enabling the model to learn more generalized tampering traces. In the fine-tuning stage, the model adapts to document data through a parameter-efficient fine-tuning method called Frequency Decoupled Parameter Adaptation (FDPA). FDPA operates in the frequency domain, where the features extracted from the pre-trained encoder are decoupled into high-frequency and low-frequency components. Inspired by Rein [42] and Frequency-Adapted (FADA) [3], we ensure efficient adaptability by introducing a minimal set of trainable parameters through Rein Frequency Adaptation (RFA). In the inference phase, we use the pre-trained encoder with RFA and the fine-tuned decoder to predict the input image.

We adopt DINOv2-Mask2Former as the baseline model due to its strong generalization ability in open-world scenarios [13, 42]. The encoder is based on DINOv2-L [26], while the decoder utilizes the Mask2Former framework [6]. Notably, this combination has not yet been explored in the context of DITL.

3.2 Cross-Domain Invariance Pre-training

We intentionally incorporate high-quality natural scene tampered images for the pre-training stage to assist the DITL task. In addition, A small number of unlabeled document images serve as style references, without requiring extra annotations. This approach alleviates the reliance on limited document data and addresses the challenge of improving generalization performance. Specifically, we propose Cross-Domain Invariance Pre-training (CDIP), which comprises three modules: cross-domain style transfer, texture consistency learning, and tampering trace commonality learning. The specific implementation details are introduced below.

3.2.1 Cross-Domain Style Transfer. As shown in Figure 1(a), significant color differences exist between natural scene images and document images. Natural images typically exhibit rich and diverse colors, while document images are characterized by simpler and more uniform background colors. This discrepancy poses challenges for cross-domain learning. To reduce domain differences in the color space, natural scene images are transformed at the image level using Cross-Domain Style Transfer (CDST), as shown in

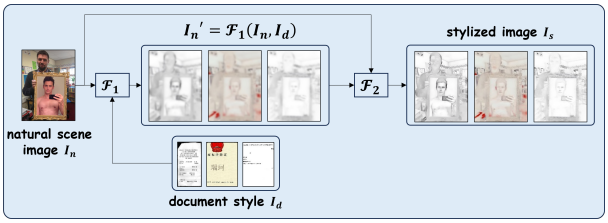


Figure 3: The Cross-Domain Style Transfer framework consists of two components: $\mathcal{F}_1$ and $\mathcal{F}_2$.

Figure 3. However, the style transfer process often introduces structural artifacts and distortions, which destroy the original texture structure of the image and hinder the model's ability to accurately detect tampering traces.

To address this issue, we apply the pre-trained FastPhotoStyle [18], which transfers the style of a document image I_d to a natural scene image I_n, aligning the color space of I_n with that of I_d while preserving its original texture structure. This process consists of two components: stylization and smoothing. The stylization module $\mathcal{F}_1$ adjusts the color distribution of I_n to match the style of I_d, minimizing structural artifacts in the output image I_n' and reducing domain differences at the image level. However, $\mathcal{F}_1$ often introduces structural artifacts in regions with similar semantic features. To address this, the smoothing module $\mathcal{F}_2$ leverages an affinity matrix [18] to eliminate artifacts and preserve structural consistency, generating the doc-stylized natural scene image I_s. This image is produced using a pre-trained frozen FastPhotoStyle without additional learning and can generate different I_s for the same I_n based on different I_d, thus increasing diversity. This process is performed on-the-fly during the pre-training stage and is described as follows:

$$I_n' = \mathcal{F}_1(I_n, I_d), \tag{1}$$

$$I_s = \mathcal{F}_2(I_n', I_n). \tag{2}$$

3.2.2 Texture Consistency Learning. Compared to the diverse content of natural scene images, the layout of document images is highly structured, resulting in regular textures. A significant domain gap in texture exists between the two types of images. To address this issue, Texture Consistency Learning (TCL) is proposed to constrain the DITL learning process for I_s with the help of I_d.

Texture is a global perceptual feature consisting of locally repeated patterns (e.g., stripes, spots). The Gram matrix $G \in \mathbb{R}^{C\times C}$ encodes the global texture distribution by calculating the activation correlation between feature channels, enabling efficient extraction of texture features [9, 13]. Unlike the image level, which focuses on pixel-level transformations, the feature level Gram matrix captures higher-level structural information encoded in the feature maps:

$$G_{ij} = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} F_{i,h,w}\cdot F_{j,h,w}, \tag{3}$$

where $F \in \mathbb{R}^{C\times H\times W}$ represents the feature map output by the encoder, C denotes the number of channels, and H and W represent the height and width, respectively. Each element G_{ij} in the Gram matrix represents the correlation between the i-th and j-th channels, capturing the co-occurrence statistics of feature activations. These correlations encode the global texture information of the image.

As shown in Figure 2, an image encoder with shared weight parameters is utilized to extract features from the input images. Specifically, $F_s = \text{Encoder}(I_s)$ is extracted for the doc-stylized natural scene image, and $F_d = \text{Encoder}(I_d)$ for the document image. The Gram matrix is then applied to compute the texture features $G^s = \text{Gram}(F_s)$ for the I_s and $G^d = \text{Gram}(F_d)$ for the I_d. Inspired by [13], $\|G^s - G^d\|_2$ is used to represent the texture differences between the stylized and document images in the texture feature space. Based on this, a texture consistency loss L_{texture} is proposed to minimize these differences, thereby aligning I_s and I_d in the texture feature space, which serves as a constraint during the training of the encoder:

$$L_{\text{texture}} = \sum_{k=1}^{K}\lambda_k \left\|G^s - G^d\right\|_2, \tag{4}$$

where K represents the number of feature layers, and λ_k denotes the weight coefficient of each layer in the image encoder. As the layers become deeper, the texture features encoded in the feature maps become less prominent and we reduce the value of λ_k as k increases [13].

3.2.3 Tampering Trace Commonality Learning. As mentioned above, the generated doc-stylized nature scene image I_s is aligned with the document image I_d in color through CDST, while the encoded features of I_s can be aligned with I_d in terms of texture through TCL. As different types of images share common microscopic tampering traces, such domain-invariant features are extracted from external natural scene images I_n to enhance the model's generalization ability. The prediction result of I_s from the tampering localization model should ideally be consistent with that of I_n.

To achieve this, we use a shared-weight image encoder, consistent with the encoder structure described earlier, to extract the features of the natural scene image, $F_n = Encoder(I_n)$, and the stylized image, $F_s = Encoder(I_s)$. The extracted features are then input into the same decoding model for prediction, resulting in the prediction outputs $p^n = Decoder(F_n)$ and $p^s = Decoder(F_s)$, respectively. The concept of tampering trace commonality refers to the shared characteristics of tampering traces across different domains, which are preserved during the style transfer process. The objective is to minimize the tampering localization loss L_{tamper} and the tampering trace commonality loss L_{common}.

$$L_{\text{tamper}} = L_{ce}(p^n, GT), \tag{5}$$

$$L_{\text{common}} = L_{ce}(p^s, GT), \tag{6}$$

where GT represents the Ground-Truth forgery region of the image.

The total loss during the pre-training phase is:

$$L_{\text{pre-trained}} = L_{\text{tamper}} + L_{\text{common}} + L_{\text{texture}}. \tag{7}$$

3.3 Frequency Decoupled Parameter Adaptation

The encoder, having learned general semantic representations during pre-training, requires minimal adjustment, as its outputs sufficiently capture the core information of the input for downstream tasks [3, 11, 42]. In contrast, the decoder, which generates task-specific outputs, adapts to the target task's requirements, requiring full fine-tuning. Existing research [5, 28] in DITL shows that frequency domain information is effective for detecting tampering

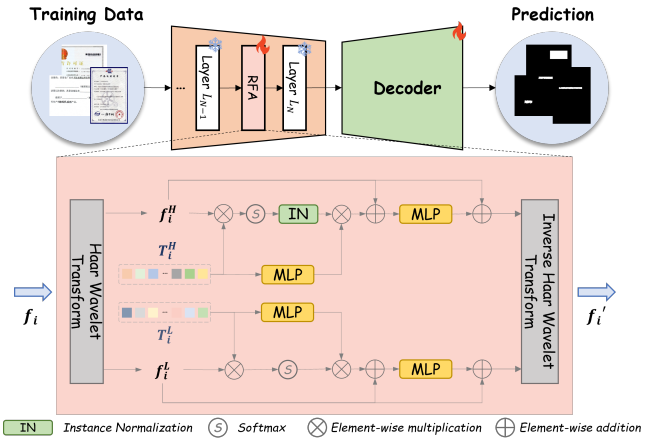


Figure 4: The architecture of Frequency Decoupling Parameter Adaptation.

traces. However, high-frequency details are often degraded by post-processing operations (e.g., Gaussian blur or JPEG compression), leading to the loss of critical tampering clues. Consequently, low-frequency and mid-frequency discontinuities serve as key indicators for tampering detection. To address this, we decouple the encoded features in the frequency domain, enabling the model to capture these clues more effectively during the fine-tuning stage in order to adapt the pre-trained model to the DITL task. To preserve the domain-invariant features learned during pre-training and reduce overfitting to the target data, we adopt parameter-efficient fine-tuning for the image encoder, maintaining the pre-trained model's generalization capability.

We propose the Frequency Decoupled Parameter Adaptation (FDPA) algorithm. As shown in Figure 4, inspired by the frequency decoupling method in [3], we freeze the image encoder obtained during pre-training and insert a lightweight, trainable module called Rein Frequency Adaptation (RFA) [42] after each layer for fine-tuning. This approach refines the features extracted from each layer by introducing a set of trainable tokens.

The features extracted from the frozen image encoder are decomposed into high-frequency f_i^H and low-frequency f_i^L components in the frequency domain using the haar wavelet transform. The RFA module then adaptively learns from each frequency component.

$$f_i = W_i x, \quad f_i \in \mathbb{R}^{c\times n}, \quad i = 1, 2, \ldots, N-1, \tag{8}$$

$$f_i^H, f_i^L = \text{HaarWT}(f_i), \quad f_i^H, f_i^L \in \mathbb{R}^{c\times n}, \quad i = 1, 2, \ldots, N-1, \tag{9}$$

where $W_i \in \mathbb{R}^{c\times c}$ represents pre-trained weight matrix, x denotes the image embedding, c denotes the channel size, n denotes the patch number, N denotes the number of layers, and HaarWT represents the haar wavelet transform.

For each frequency domain branch, the features extracted from each layer are refined using a set of trainable tokens $T_i \in \mathbb{R}^{m\times c}$, where m represents the sequence length of T_i and c denotes the dimension of the frozen image encoder features. First, we compute the similarity maps S_i^H between the tokens T_i^H and the frequency components f_i^H. Similarly, S_i^L is obtained in an analogous manner. These maps measure the correlation between each element in T_i^H (or T_i^L) and each patch embedding represented in f_i^H (or f_i^L), respectively, and are defined as follows:

$$S_i^H = \text{Softmax}\left(\frac{f_i^H \times T_i^H}{\sqrt{c}}\right), S_i^H \in \mathbb{R}^{n\times m}, \tag{10}$$

$$S_i^L = \text{Softmax}\left(\frac{f_i^L \times T_i^L}{\sqrt{c}}\right), S_i^L \in \mathbb{R}^{n\times m}, \tag{11}$$

where Softmax denotes the softmax activation function.

To mitigate the impact of post-processing operations on high-frequency information in the tampering localization task, we apply instance normalization to the high-frequency branch. Since instance normalization normalizes each channel of each sample independently, it suppresses high-frequency differences between samples and emphasizes the low-frequency components:

$$I_i = \text{IN}(S_i^H), I_i \in \mathbb{R}^{n\times m}, \tag{12}$$

where IN represents instance normalization.

The token features T_i^H are projected into the feature space of f_i^H using a multilayer perceptron (MLP) parameterized by the weight matrix $W_{T_i^H}$ and the bias vector $b_{T_i^H}$. This projection is followed by an element-wise multiplication with the similarity map I_i. The process for the high-frequency branch is formally defined as follows:

$$\Delta f_i^H = I_i \times \left[T_i^H \times W_{T_i^H} + b_{T_i^H}\right]. \tag{13}$$

For the low-frequency branch, the process is defined as:

$$\Delta f_i^L = S_i^L \times \left[T_i^L \times W_{T_i^L} + b_{T_i^L}\right], \tag{14}$$

Note that compared to Eq. (13), S_i^L is used without IN, thereby avoiding disruption of structural correlation.

Then, the projected token features Δf_i^H are fused with the low-frequency features f_i^L through another multilayer perceptron (MLP) parameterized by the weight matrix $W_{f_i^H}$ and the bias vector $b_{f_i^H}$. This fusion is followed by a skip connection to preserve the original feature information. The process is formally defined as follows:

$$f_i^{H'} = f_i^H + (f_i^H + \Delta f_i^H) \times W_{f_i^H} + b_{f_i^H}, \tag{15}$$

$$f_i^{L'} = f_i^L + (f_i^L + \Delta f_i^L) \times W_{f_i^L} + b_{f_i^L}. \tag{16}$$

Finally, the high- and low-frequency components are fused and transformed from the frequency domain back to the spatial domain to obtain the updated feature f_i', which is subsequently fed into the next layer of the image encoder.

$$f_i' = \text{InverseHaarWT}\left(f_i^{H'}, f_i^{L'}\right), \tag{17}$$

where InverseHaarWT denotes the inverse haar wavelet transform.

To effectively adapt the pre-trained model to the DITL task, we optimize it using a combination of loss functions that balance tampering localization accuracy and feature consistency. The total loss in the fine-tuning stage is defined as:

$$L_{\text{train}} = L_{\text{cls}} + L_{\text{ce}} + L_{\text{dice}}. \tag{18}$$

Table 1: Detailed description of the experimental datasets. Number represents the total number of tampered images in each dataset. M indicates manual and A refers to automatic.

Datasets	Number	Construction	Stage	Setting
MIML [31]	123,150	M	Pre-trained	I&II
DocTamper [28]	120,000	A	Train	I
DocTamper-T [28]	30,000	A	Test	
FCD [28]	2,000	A		
SCD [28]	18,000	A		
RIFLC [37]	4,000	M&A	Train	II
TTI [23]	4,000	M		
TextTamper [8]	42,465	A		
FindIt [1]	240	M	Test	
Ps_boundary [48]	1,000	M		
SACP [36]	2,005	M&A		
T-SROIE [41]	986	M		
RTM [22]	6,000	M		

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets. The experimental data is summarized in Table 1. In the pre-training stage, we use only a natural scene image dataset and a small number of unlabeled document images for training. The natural scene image dataset is MIML [31], selected for its large scale, diversity, and high quality. Ablation studies were conducted to validate the selection of the natural scene image dataset. For the document images, we select 600 samples from the document image training set based to ensure comprehensive coverage of all texture variations present in the dataset. These images are used as document style references for CDIP. In the training stage, we conduct experiments under two different settings: (I) Following the cross-dataset settings of existing studies [5, 28], we conduct experiments on DocTamper [28], which consists of a training set and a testing set within the same dataset, as well as two cross-dataset benchmarks, FCD and SCD. (II) To better simulate real-world application scenarios, we train on the RIFLC [37], TTI [23], and TextTamper [8] datasets, and test on FindIt [1], Ps_boundary [48], SACP [36], T-SROIE [41], and RTM [22].

4.1.2 Evaluation Metrics. Previous work [2, 5, 28, 34, 45] modeled DITL as a binary semantic segmentation task. Based on established benchmarks, we adopt F1-score, Precision, Recall, and IoU to evaluate the performance.

4.1.3 Implementation Details. For efficient training, images were cropped to a resolution of 512×512, and inference was performed in a sliding window manner during testing. We used AdamW [21] for optimization, with an initial learning rate of 1×10^{-4}, and trained for 40k iterations with a batch size of 48. The learning rate was decayed using the PolyLR strategy [24]. Consistent training configurations were applied across all methods for fair comparison. For Experimental Setting I, we followed the DTD pipeline [28], training on dynamically JPEG-compressed input images, with the quality factor randomly selected between 75 and 100 and the number of compression iterations randomly chosen between 1 and 3. All predictions were binarized with a threshold of 0.5.

4.2 Comparison with State-of-the-art Methods

We compared our DITL[2] with state-of-the-art document image tampering localization methods, including TIFDM [8], ASCFormer [22], DTD [28], and FFDN [5], as well as semantic segmentation methods, including ConvNeXt-Uper [20] and DINOv2-Mask2Former [26]. For a fair comparison, all models were retrained under the same experimental settings as our method, and the results obtained under Settings I are consistent with those reported in the original paper. Additionally, we incorporated CDIP into other methods for comparison by keeping their original training stages unchanged and adding CDIP only in the pre-training stage. The quantitative experimental results were presented in Table 2 and Table 3. Consistent training configurations and data settings were applied across all methods, and these models were retrained for fair comparison. The results demonstrated that our method outperformed other methods in cross-dataset generalization ability under the two experimental settings. Notably, in Setting II, the performance of existing methods in detecting manually tampered datasets was significantly lower than that on SACP, which contained automatic tampered data. By incorporating CDIP, the detection performance on manually tampered datasets was significantly improved across all comparison methods. Furthermore, Figure 5 presented the qualitative results of visual comparisons among different methods, which intuitively highlighted the superior performance of our method.

4.3 Ablation Study

We conduct detailed ablation experiments on each of our proposed components based on Experimental Setting II.

4.3.1 Cross-Domain Invariance Pre-training. We analyzed the impact of different modules on learning domain-invariant features during the pre-training stage. As shown in Table 4, we evaluated the improvement in generalization brought by pre-training under different loss function combinations. Comparing the first two rows shows that natural scene images assist in DITL pre-training, though the domain gap limits improvement. The second, third, and fourth rows demonstrate that reducing domain-specific feature differences in color and texture enabled better utilization of natural scene images to improve the generalization of DITL tasks. Finally, the last three rows confirm that the original natural scene images and the transformed images played complementary roles in learning the commonality of tampering traces.

As shown in Figure 6, we visualized the effect of our method on reducing domain-specific features. Color features are extracted

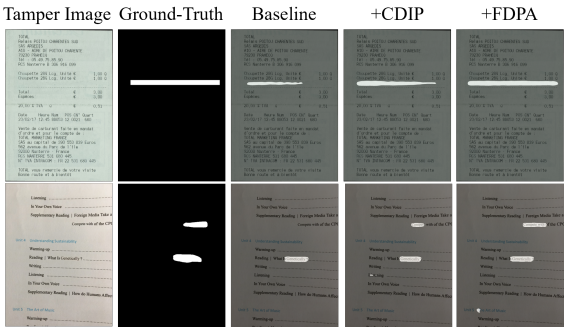


Figure 5: Visualization of experimental results using DINOv2-Mask2Former as the baseline.

Table 2: Comparison on Setting I dataset. Following the process outlined by DTD [28], each image in the test set has been compressed using the quality factor specified by their public repository. P denotes precision, R denotes recall, and F denotes F1-score. The best results are highlighted in bold, and the second-best results are marked with an underline.

Method	DocTamper-TestingSet				DocTamper-FCD				DocTamper-SCD			
	IoU	P	R	F	IoU	P	R	F	IoU	P	R	F
TIFDM [8]	0.360	0.350	0.480	0.405	0.300	0.380	0.500	0.432	0.290	0.330	0.640	0.435
TIFDM+CDIP	0.584	0.667	0.740	0.702	0.507	0.579	0.734	0.647	0.459	0.641	0.503	0.564
ConvNext-Uper [20]	0.706	0.654	0.578	0.614	0.484	0.792	0.533	0.637	0.442	0.543	0.566	0.554
ConvNext-Uper+CDIP	0.728	0.686	0.598	0.639	0.756	0.841	0.816	0.828	0.552	0.737	0.564	0.639
ASCFormer [22]	0.754	0.706	0.654	0.679	0 533	0.670	0.598	0.632	0.528	0.592	0.604	0.598
ASCFormer+CDIP	0.780	0.740	0.671	0.704	0.642	0.756	0.762	0.759	0.605	0.699	0.689	0.694
DTD [28]	0.831	0.827	0.793	0.810	0.790	0.879	0.825	0.851	0.708	0.771	0.789	0.780
DTD+CDIP	0.878	0.842	0.806	0.824	0.891	0.917	0.919	0.918	0.754	0.811	0.801	0.806
FFDN [5]	0.894	0.873	0.840	0.856	0.878	0.927	0.905	0.916	0.748	0.806	0.819	0.812
FFDN+CDIP	**0.920**	**0.895**	**0.861**	**0.878**	0.913	0.948	0.922	0.935	**0.803**	0.846	**0.849**	0.847
DINOv2-Mask2Former [26]	0.821	0.812	0.780	0.796	0.854	0.915	0.885	0.900	0.656	0.769	0.732	0.750
DINOv2-Mask2Former+CDIP	0.863	0.854	0.843	0.848	0.880	0.927	0.920	0.923	0.711	0.811	0.785	0.798
DINOv2-Mask2Former+CDIP+FDPA	0.888	0.884	**0.861**	0.872	**0.930**	**0.954**	**0.953**	**0.953**	0.791	**0.864**	0.842	**0.853**

Table 3: Comparison on Setting II dataset. CDIP significantly improves performance in cross-dataset testing.

Method	FindIt		Ps_boundary		SACP		T-SROIE		RTM		Average	
	F	IoU	F	IoU	F	IoU	F	IoU	F	IoU	F	IoU
TIFDM [8]	0.081	0.048	0.122	0.072	0.571	0.434	0.114	0.068	0.178	0.136	0.213	0.152
TIFDM+CDIP	0.154	0.099	0.684	0.566	0.694	0.549	0.503	0.360	0.376	0.271	0.482	0.369
ConvNext-Uper [20]	0.159	0.086	0.273	0.158	0.536	0.366	0.167	0.091	0.102	0.054	0.247	0.151
ConvNext-Uper+CDIP	0.318	0.189	0.744	0.592	0.641	0.472	0.447	0.288	0.202	0.112	0.470	0.331
ASCFormer [22]	0.113	0.060	0.166	0.090	0.479	0.364	0.327	0.229	0.182	0.139	0.253	0.176
ASCFormer+CDIP	0.223	0.125	0.728	0.574	0.614	0.480	0.460	0.377	0.362	0.261	0.477	0.363
DTD [28]	0.101	0.066	0.132	0.080	0.418	0.310	0.401	0.312	0.129	0.095	0.236	0.173
DTD+CDIP	0.248	0.164	0.714	0.598	0.707	0.563	0.539	0.399	0.383	0.234	0.518	0.400
FFDN [5]	0.162	0.088	0.243	0.138	0.454	0.324	0.414	0.291	0.166	0.090	0.288	0.186
FFDN+CDIP	0.324	0.203	0.716	0.564	0.700	0.539	0.598	0.436	0.290	0.171	0.526	0.383
DINOv2-Mask2Former [26]	0.263	0.152	0.355	0.216	0.708	0.548	0.327	0.195	0.204	0.114	0.371	0.245
DINOv2-Mask2Former+CDIP	0.515	0.347	0.834	0.715	0.764	0.618	0.640	0.471	0.385	0.238	0.628	0.478
DINOv2-Mask2Former+CDIP+FDPA	**0.575**	**0.404**	**0.843**	**0.728**	**0.777**	**0.636**	**0.649**	**0.480**	**0.449**	**0.290**	**0.659**	**0.508**

Table 4: Ablation study of different loss functions in CDIP.

L_{tamper}	L_{common}	$L_{texture}$	FindIt		Ps_boundary		SACP		T-SROIE		RTM		Average	
			F	IoU	F	IoU	F	IoU	F	IoU	F	IoU	F	IoU
			0.263	0.152	0.355	0.216	0.708	0.548	0.327	0.195	0.204	0.114	0.371	0.245
✓			0.277	0.161	0.742	0.589	0.606	0.435	0.489	0.324	0.214	0.112	0.466	0.324
✓	✓		0.475	0.312	0.636	0.467	0.755	0.606	0.517	0.349	0.354	0.215	0.547	0.390
✓		✓	0.467	0.304	0.792	0.655	0.747	0.596	0.535	0.365	0.347	0.210	0.578	0.426
	✓	✓	0.494	0.328	0.767	0.623	0.752	0.602	0.621	0.450	0.366	0.224	0.600	0.445
✓	✓	✓	**0.515**	**0.347**	**0.834**	**0.715**	**0.764**	**0.618**	**0.640**	**0.471**	**0.385**	**0.238**	**0.628**	**0.478**

from the R, G, and B channel histograms and concatenated. Texture features are characterized using feature representations computed from Gram matrices. Through our CDIP, natural scene images and document images became more similar in the color and texture spaces, effectively mitigating the impact of domain-specific differences. Additionally, as shown in Figure 7, we found that 600 style reference images were sufficient to represent the texture distribution of document images in the training set.

Table 5 illustrates the impact of selecting the natural scene image dataset during pre-training on overall performance, showing F1-scores on FindIt [1] and RTM [22]. We compare three manual tampering datasets—CoMoFoD [38], Kours [15], and MIML [31]—with the automatic tampering dataset DIS25K [35]. The results show that the quality of the natural scene image dataset significantly affects model performance, with high-quality manual tampering data being more effective for learning domain-invariant features. Furthermore, pretraining on all datasets results in suboptimal performance

Table 5: Comparison of F1-score results for different natural scene image datasets in CDIP. M indicates manual manipulation, while A denotes automatic manipulation.

Datasets	Construction	Number	FindIt	RTM
None	-	0	0.263	0.204
DIS25K [35]	A	24,964	0.442	0.296
CoMoFoD [38]	M	260	0.495	0.347
Kours [15]		912	0.455	0.351
MIML [31]		123,150	**0.515**	**0.385**
All of the above	M&A	149,286	0.504	0.360

compared to using MIML alone. This is primarily attributed to the noise introduced by low-quality datasets, which contain images and annotations of relatively lower quality.

4.3.2 Frequency Decoupled Parameter Adaptation. Table 6 compared the differences between various fine-tuning methods used in the training phase and presented the test results on FindIt [1] and RTM [22]. Our method required adjusting only about 4% of the training parameters compared to full fine-tuning. Comparing the second and fourth rows, it was observed that the model better captured weak tampering traces in document images by adaptively learning

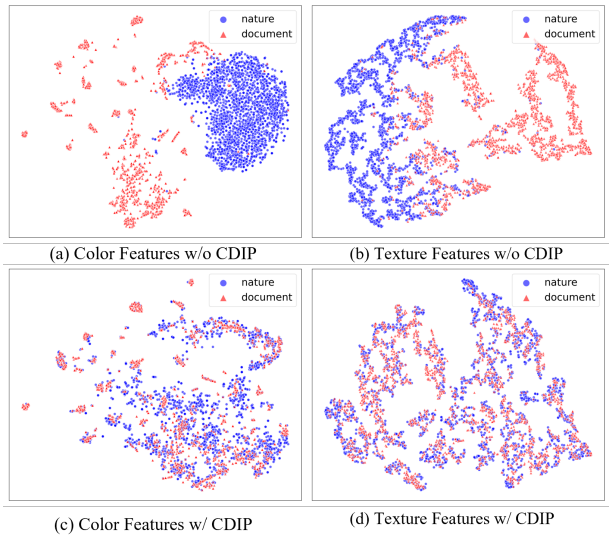


Figure 6: The t-SNE [39] plots the differences in color and texture features between natural scene images and document images.

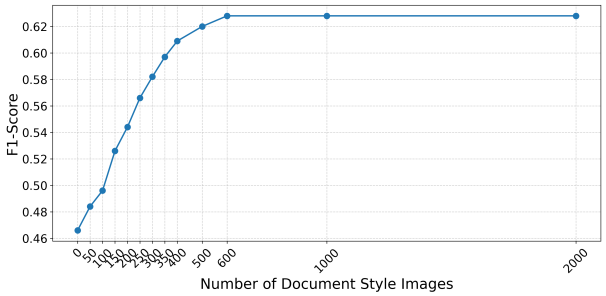


Figure 7: Impact of the number of document style reference images on cross-dataset performance (average F1-score).

Table 6: Ablation of different fine-tuning methods. Parameters refers to the size of trainable parameters.

Method	Parameters	FindIt		RTM	
		F	IoU	F	IoU
Full	304.20M	0.515	0.347	0.385	0.238
w/o FD	2.99M	0.475	0.312	0.398	0.249
w/o IN	11.65M	0.494	0.328	0.418	0.264
Ours	11.65M	**0.575**	**0.404**	**0.449**	**0.290**

Table 7: Ablation study of different encoders.

Method	FindIt		RTM	
	F	IoU	F	IoU
CLIP [32]	0.220	0.124	0.263	0.151
CLIP+CDIP	0.495	0.329	0.406	0.254
CLIP+CDIP+FDPA	0.539	0.369	**0.475**	**0.312**
SAM [14]	0.223	0.125	0.234	0.133
SAM+CDIP	0.484	0.319	0.398	0.249
SAM+CDIP+FDPA	0.510	0.343	0.470	0.307
DINOv2 [26]	0.263	0.152	0.204	0.114
DINOv2+CDIP	0.515	0.347	0.385	0.238
DINOv2+CDIP+FDPA	**0.575**	**0.404**	0.449	0.290

high- and low-frequency features through frequency decoupling (FD). Furthermore, comparing the third and fourth rows highlighted the role of instance normalization (IN) applied to high-frequency components. By standardizing the statistical distribution of high-frequency features, IN effectively reduced the influence of style variations, enabling the model to better capture low-frequency discontinuities where tampering traces were primarily concentrated.

4.3.3 Choice of Backbone. We explored the performance of different Vision Foundation Models (VFMs) in DITL generalization, as shown in Table 7. The table also presented the test results on FindIt and RTM under Setting II. The experimental results demonstrated that DINOv2 [26], based on self-supervised training, outperformed the mainstream models CLIP [32] and SAM [14]. Furthermore, our method exhibited plug-and-play versatility across different models and achieved significant performance improvements for various methods.

5 Conclusion

In this paper, we propose DITL^2, a Dual-stage Invariance Transfer Learning framework for Document Image Tampering Localization, which effectively addresses the challenges of cross-dataset generalization. By leveraging Cross-Domain Invariance Pre-training (CDIP) to suppress domain-specific influences and acquire domain-invariant features, and Frequency Decoupling Parameter Adaptation (FDPA) to adapt pre-trained parameters for fine-tuning, our method demonstrates significant improvements in tampering localization performance. Extensive experiments validate that DITL^2 effectively utilizes external data resources, achieving state-of-the-art results across various cross-dataset settings. These findings highlight the potential of DITL^2 as a robust and generalizable solution for practical DITL applications.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant U23B2022, U22B2047 and 62371301.

References

[1] Chloé Artaud, Antoine Doucet, Jean-Marc Ogier, and Vincent Poulain d'Andecy. 2017. Receipt dataset for fraud detection. In *First International Workshop on Computational Document Forensics*.

[2] Yong-Yeol Bae, Dae-Jea Cho, and Ki-Hyun Jung. 2025. A New Log-Transform Histogram Equalization Technique for Deep Learning-Based Document Forgery Detection. *Symmetry* 17, 3 (2025), 395.

[3] Qi Bi, Jingjun Yi, Hao Zheng, Haolan Zhan, Yawen Huang, Wei Ji, Yuexiang Li, and Yefeng Zheng. 2024. Learning frequency-adapted vision foundation model for domain generalized semantic segmentation. *Advances in Neural Information Processing Systems* 37 (2024), 94047–94072.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PmLR, 1597–1607.

[5] Zhongxi Chen, Shen Chen, Taiping Yao, Ke Sun, Shouhong Ding, Xianming Lin, Liujuan Cao, and Rongrong Ji. 2024. Enhancing Tampered Text Detection Through Frequency Feature Fusion and Decomposition. In *European Conference on Computer Vision*. Springer, 200–217.

[6] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention Mask Transformer for Universal Image Segmentation. *CVPR*.

[7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2818–2829.

[8] Li Dong, Weipeng Liang, and Rangding Wang. 2024. Robust text image tampering localization via forgery traces enhancement and multiscale attention. *IEEE Transactions on Consumer Electronics* (2024).

[9] Leon Gatys, Alexander S Ecker, and Matthias Bethge. 2015. Texture synthesis using convolutional neural networks. *Advances in neural information processing systems* 28 (2015).

[10] Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. 2021. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293* (2021).

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR* 1, 2 (2022), 3.

[12] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361* (2020).

[13] Sunghwan Kim, Dae-hwan Kim, and Hoseong Kim. 2023. Texture learning domain randomization for domain generalized segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 677–687.

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4015–4026.

[15] Paweł Korus and Jiwu Huang. 2016. Multi-scale analysis strategies in PRNU-based tampering localization. *IEEE Transactions on Information Forensics and Security* 12, 4 (2016), 809–824.

[16] Myung-Joon Kwon, Wonjun Lee, Seung-Hun Nam, Minji Son, and Changick Kim. 2024. SAFIRE: Segment Any Forged Image Region. *arXiv preprint arXiv:2412.08197* (2024).

[17] Weixiang Li, Bin Li, Kengtao Zheng, Songze Li, and Haodong Li. 2025. Document image forgery detection and localization in desensitization scenarios. *Signal Processing* (2025), 110123.

[18] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. 2018. A closed-form solution to photorealistic image stylization. In *Proceedings of the European conference on computer vision (ECCV)*. 453–468.

[19] Xin Liao, Siliang Chen, Jiaxin Chen, Tianyi Wang, and Xiehua Li. 2023. CTP-Net: Character texture perception network for document image forgery localization. *arXiv preprint arXiv:2308.02158* (2023).

[20] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11976–11986.

[21] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[22] Dongliang Luo, Yuliang Liu, Rui Yang, Xianjin Liu, Jishen Zeng, Yu Zhou, and Xiang Bai. 2025. Toward real text manipulation detection: New dataset and new solution. *Pattern Recognition* 157 (2025), 110828.

[23] Dongliang Luo, Yu Zhou, Rui Yang, Yuliang Liu, Xianjin Liu, Jishen Zeng, Enming Zhang, Biao Yang, Ziming Huang, Lianwen Jin, et al. 2023. ICDAR 2023 competition on detecting tampered text in images. In *International Conference on Document Analysis and Recognition*. Springer, 587–600.

[24] Purnendu Mishra and Kishor Sarawadekar. 2019. Polynomial learning rate policy with warm restart for deep neural network. In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*. IEEE, 2087–2092.

[25] Yamato Okamoto, Osada Genki, Iu Yahiro, Rintaro Hasegawa, Peifei Zhu, and Hirokatsu Kataoka. 2023. Image Generation and Learning Strategy for Deep Document Forgery Detection. *arXiv preprint arXiv:2311.03650* (2023).

[26] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* (2023).

[27] Duo Peng, Yinjie Lei, Munawar Hayat, Yulan Guo, and Wen Li. 2022. Semantic-aware domain generalized segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2594–2605.

[28] Chenfan Qu, Chongyu Liu, Yuliang Liu, Xinhong Chen, Dezhi Peng, Fengjun Guo, and Lianwen Jin. 2023. 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.

[29] Chenfan Qu, Jian Liu, Haoxing Chen, Baihan Yu, Jingjing Liu, Weiqiang Wang, and Lianwen Jin. 2024. TextSleuth: Towards Explainable Tampered Text Detection. *arXiv preprint arXiv:2412.14816* (2024).

[30] Chenfan Qu, Yiwu Zhong, Fengjun Guo, and Lianwen Jin. 2025. Revisiting tampered scene text detection in the era of generative AI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 694–702.

[31] Chenfan Qu, Yiwu Zhong, Chongyu Liu, Guitao Xu, Dezhi Peng, Fengjun Guo, and Lianwen Jin. 2024. 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.

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

[33] Huiru Shao, Kaizhu Huang, Wei Wang, Xiaowei Huang, and Qiufeng Wang. 2023. Progressive supervision for tampering localization in document images. In *International Conference on Neural Information Processing*. Springer, 140–151.

[34] Huiru Shao, Zhuang Qian, Kaizhu Huang, Wei Wang, Xiaowei Huang, and Qiufeng Wang. 2024. Delving into Adversarial Robustness on Document Tampering Localization. In *European Conference on Computer Vision*. Springer, 290–306.

[35] Eren Tahir and Mert Bal. 2024. Deep image composition meets image forgery. *arXiv preprint arXiv:2404.02897* (2024).

[36] Tianchi. 2020. Security AI Challenger Program. https://tianchi.aliyun.com/competition/entrance/531812/introduction.

[37] Tianchi. 2022. Real-World Image Forgery Localization Challenge. https://tianchi.aliyun.com/competition/entrance/531945/introduction.

[38] Dijana Tralic, Ivan Zupancic, Sonja Grgic, and Mislav Grgic. 2013. CoMoFoD—New database for copy-move forgery detection. In *Proceedings ELMAR-2013*. IEEE, 49–54.

[39] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

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

[41] Yuxin Wang, Boqiang Zhang, Hongtao Xie, and Yongdong Zhang. 2022. Tampered text detection via rgb and frequency relationship modeling. *Chinese Journal of Network and Information Security* 8, 3 (2022), 29–40.

[42] Zhixiang Wei, Lin Chen, Yi Jin, Xiaoxiao Ma, Tianle Liu, Pengyang Ling, Ben Wang, Huaian Chen, and Jinjin Zheng. 2024. Stronger fewer & superior: Harnessing vision foundation models for domain generalized semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 28619–28630.

[43] Qi Xu, Liang Yao, Zhengkai Jiang, Guannan Jiang, Wenqing Chu, Wenhui Han, Wei Zhang, Chengjie Wang, and Ying Tai. 2022. Dirl: Domain-invariant representation learning for generalizable semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 36. 2884–2892.

[44] Wenbo Xu, Junwei Luo, Chuntao Zhu, Wei Lu, Jinhua Zeng, Shaopei Shi, and Cong Lin. 2022. Document images forgery localization using a two-stream network. *International Journal of Intelligent Systems* 37, 8 (2022), 5272–5289.

[45] Zeqin Yu, Bin Li, Yuzhen Lin, Jinhua Zeng, and Jishen Zeng. 2023. Learning to Locate the Text Forgery in Smartphone Screenshots. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[46] Zeqin Yu, Jiangqun Ni, Yuzhen Lin, Haoyi Deng, and Bin Li. 2024. Diffforensics: Leveraging diffusion prior to image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12765–12774.

[47] Xiangyu Yue, Yang Zhang, Sicheng Zhao, Alberto Sangiovanni-Vincentelli, Kurt Keutzer, and Boqing Gong. 2019. Domain randomization and pyramid consistency: Simulation-to-real generalization without accessing target domain data. In *Proceedings of the IEEE/CVF international conference on computer vision*. 2100–2110.

[48] Peiyu Zhuang, Haodong Li, Shunquan Tan, Bin Li, and Jiwu Huang. 2021. Image tampering localization using a dense fully convolutional network. *IEEE Transactions on Information Forensics and Security* 16 (2021), 2986–2999.