Towards Robust Tampered Text Detection in Document Image: New dataset and New Solution

Chenfan Qu[1] , Chongyu Liu[1] , Yuliang Liu[2] , Xinhong Chen[1] , Dezhi Peng[1] , Fengjun Guo[3] , Lianwen Jin[1,*]
[1]South China University of Technology, [2]Huazhong University of Science and Technology,
[3]IntSig Information Co., Ltd

202221012612@mail.scut.edu.cn

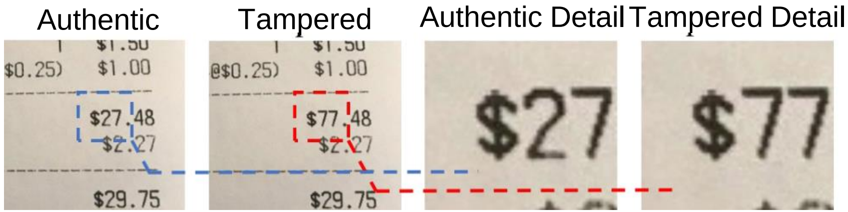
https://github.com/qcf-568/DocTamper

Background

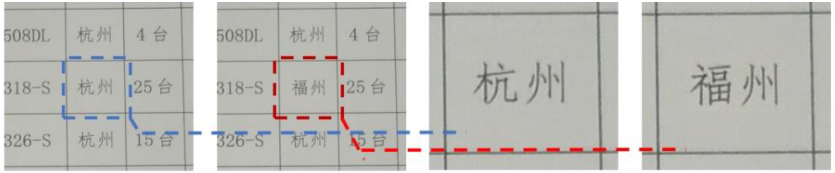
Document Tampering Methods

The three most commonly used document image tampering methods are: Copy-Move、Splicing、Generation.

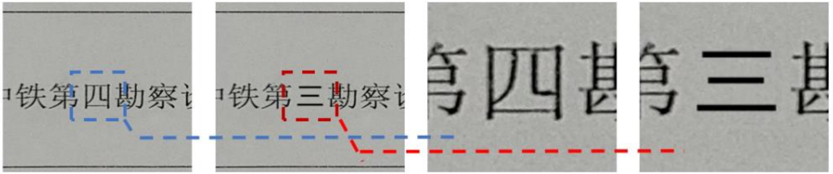
- **Copy-move means shifting the spatial locations of texts within images.**



- **Splicing means copying text regions from one image and paste to other images.**



- **Generation means replacing regions of images with visually plausible but different contents.**



Main Features of Tampered Text Detection on Documents

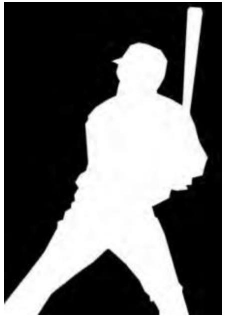
Comparing to manipulation on images, document tampering has two main features：

1. The area of tampered text are small.

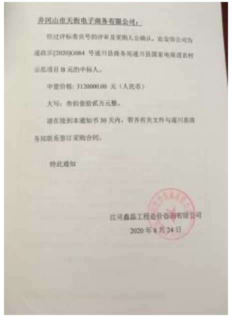
   Manipulation on images with natural objects usually covers whole object. Therefore the area of tampered regions are usually bigger. While texts usually are small thus the area of tampered regions are small.

Manipulated Image

Manipulated Region



Tampered document



井冈山市天街电子商务有限公司：

经过评标委员号的评审及采购人公确认，兹宣布公司为遂政示[2020]G084 号遂川县商务局遂川县国家电商进农村示范项目 B 元的中标人。

中壹价格：3120000.00 元（人民币）

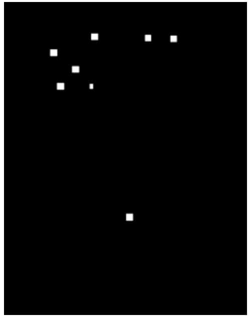
大写：叁佰壹拾贰万元整。

请在接到本通知书 30 天内，带齐有关文件与遂川县商务局联系签订采购合同。

特此通知

江司鑫磊工程造价咨询有限公司
2020 年 8 月 24 日

Tampered region

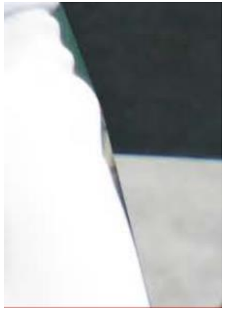


2. Less likely to leave visual tampering clues.

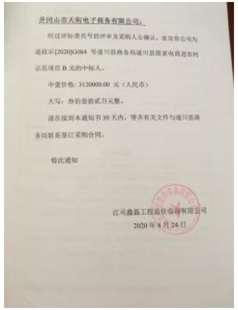
   In natural object images, the luminance and background among objects have larger difference. Therefore more likely to leave visual tampering clues. While in documents, due to the high consistency of fonts and background, tampering are less likely to leave visual clues.

Manipulated Image

Detail



Tampered document



井冈山市天街电子商务有限公司：

经过评标委员号的评审及采购人公确认，兹宣布公司为遂政示[2020]G084 号遂川县商务局遂川县国家电商进农村示范项目 B 元的中标人。

中壹价格：3120000.00 元（人民币）

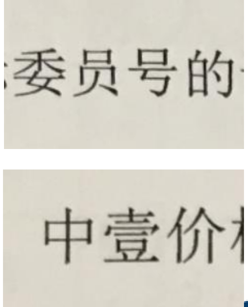
大写：叁佰壹拾贰万元整。

请在接到本通知书 30 天内，带齐有关文件与遂川县商务局联系签订采购合同。

特此通知

江司鑫磊工程造价咨询有限公司
2020 年 8 月 24 日

Detail



委员号的

中壹价

Motivation

1. **In the aspect of data, there lack a public dataset that have enough samples (>1k) and close enough to real-world scenarios.**

   There are significant differences between existing public datasets for document tampering detection research and the real-world demands. Existing public datasets mostly focus on simple scenarios such as scanned documents, while the real-world demands often include photographed contracts, notifications, and receipt. Moreover, due to manual tampering and annotation process are too time-consuming and labor-intensive, the size of existing public datasets are very small.

2. **In the aspect of method, previous methods can't effectively detect document tampering that lack visual tampering clues.**

   Existing research often requires a clean and tidy document layout, which is difficult to meet the demand for tampering detection on various photographed document images.

   Existing methods often rely on visual cues to detect tampered text, making it difficult to achieve satisfactory performance for Copy-Paste tampering in complex scenarios.

DocTamper dataset

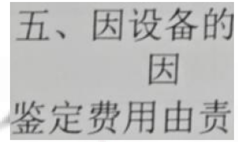
Selective Tampering Generation-Motivation

1. **Reduce annotation costs.**

The process of manually tampering and labeling tampered regions is very time-consuming and labor-intensive. Considering that there are no natural tampered document images, manually tampered images also need to be synthesized through digital image processing programs such as photoshop. It is reasonable to automatically generate tampering and annotations through algorithms, which can effectively alleviate the data-hunger of deep learning algorithms.

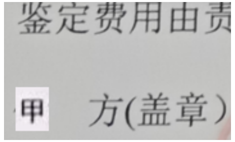
2. **Mimic manual tampering process.**

The majority of training data in the field of natural image tamper detection is also automatically synthesized, usually by directly using an object mask to extract an object and randomly paste it to other locations [1] . If we also randomly paste or generate text on document images in this way, it will cause significant distortion (obvious disharmony in text position, font, and background), resulting in significant differences from manual tampering in the real world. This will in turn make the trained model unable to learn how to detect careful manual tampering, making it difficult to effectively meet real-world demands. The proposed STG can solve these problems by carefully mimic the manual tampering process.



五、因设备的
因
鉴定费用由责

Avoid obvious inconsistency of position like this.

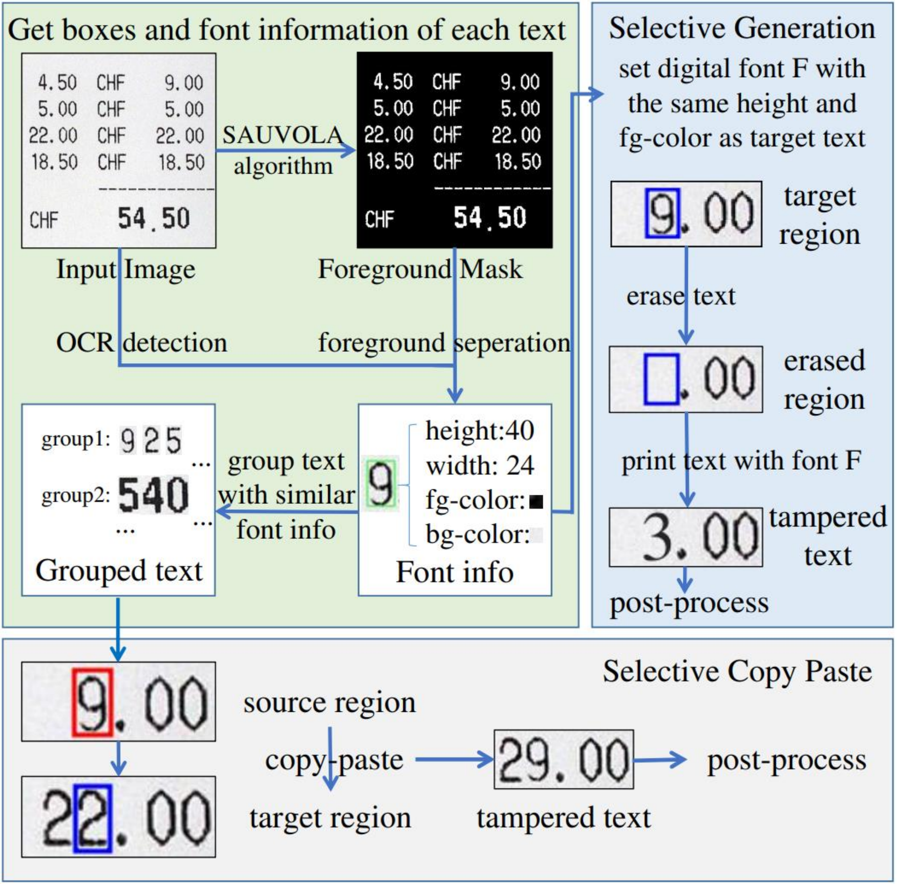


鉴定费用由责
甲 方(盖章)

Avoid obvious inconsistency of style like this.

[1] Wang, J., Wu, Z., Chen, J., Han, X., Shrivastava, A., Lim, S. N., & Jiang, Y. G. (2022). Objectformer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2364-2373).

Selective Tampering Generation-Method



1、**Preparation stage**：

- ➢ Get stroke level text mask with SAUVOLA algorithm.
- ➢ Get OCR detection boxes for each text.
- ➢ Use the height, width, foreground color (mean, variance) within stroke level mask, and background color (m, v)of the text boxes as font information to approximate the representation of the font.
- ➢ Group the texts that share similar font information.

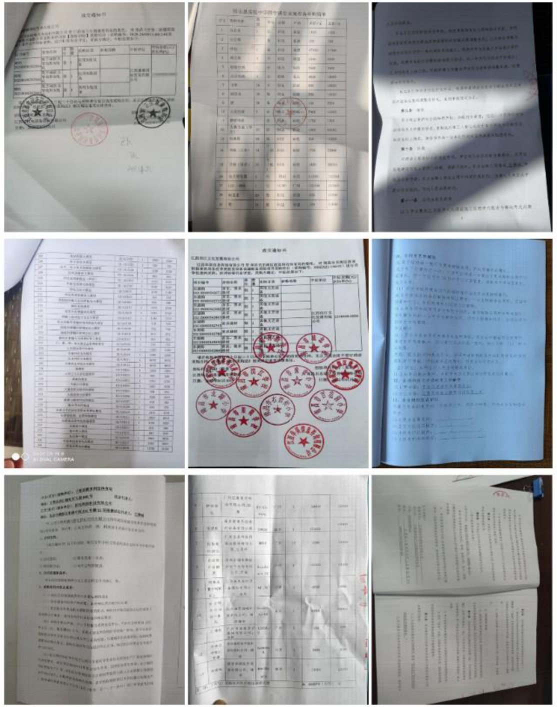
2、**Selective copy-paste**：Randomly swap the positions of the text within each group, random post-processing.

3、**Selective generation**：Set similar TTF font with the recorded font info, erase out the origin text in target region and print new text on it.

Core concept：Use sizes, foreground and background statistics of texts to approximate font information that cannot be directly obtained from complex photographed documents in complex scenarios. Tampering text with similar style, foreground and background color and texture to reduce font and position distortion.

DocTamper dataset

We collect 50562 images of contracts, notifications, receipts, invoices, sheets, notes and pages from various scenarios and build DocTamper dataset to address the lack of sufficient data that can reflect real-world demands.



Contracts, notifications, sheets

Receipts, notes

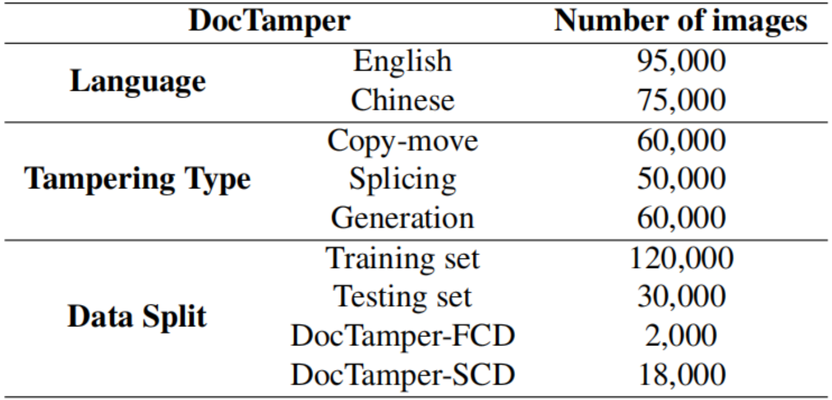
Invoices, normal pages

DocTamper dataset

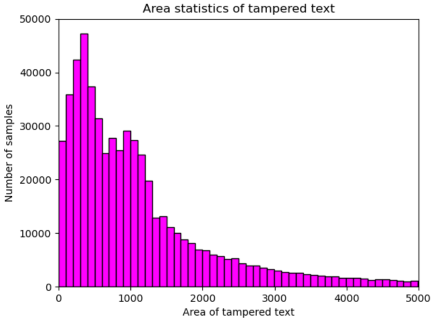
1. Cross source image domain testing subsets

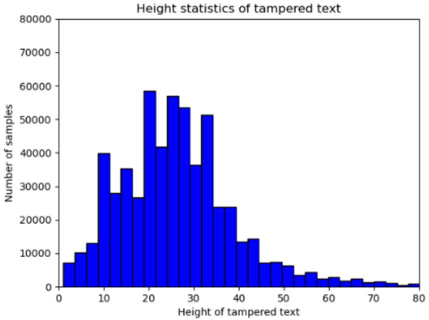
Inspired by the common practice of using image materials from one image source as training data and materials from another image source as testing data in the field of natural image manipulation detection, we introduced two cross domain testing sets (DocTamper-FCD and DocTamper-SCD). The image materials not only do not include the image materials involved in the training data, but also the image sources between them and training set are completely different, so the layout style of the document differs greatly from the training set. More in line with real-world demands.

2. Basic Statistics of DocTamper dataset



DocTamper		Number of images
Language	English	95,000
	Chinese	75,000
Tampering Type	Copy-move	60,000
	Splicing	50,000
	Generation	60,000
Data Split	Training set	120,000
	Testing set	30,000
	DocTamper-FCD	2,000
	DocTamper-SCD	18,000





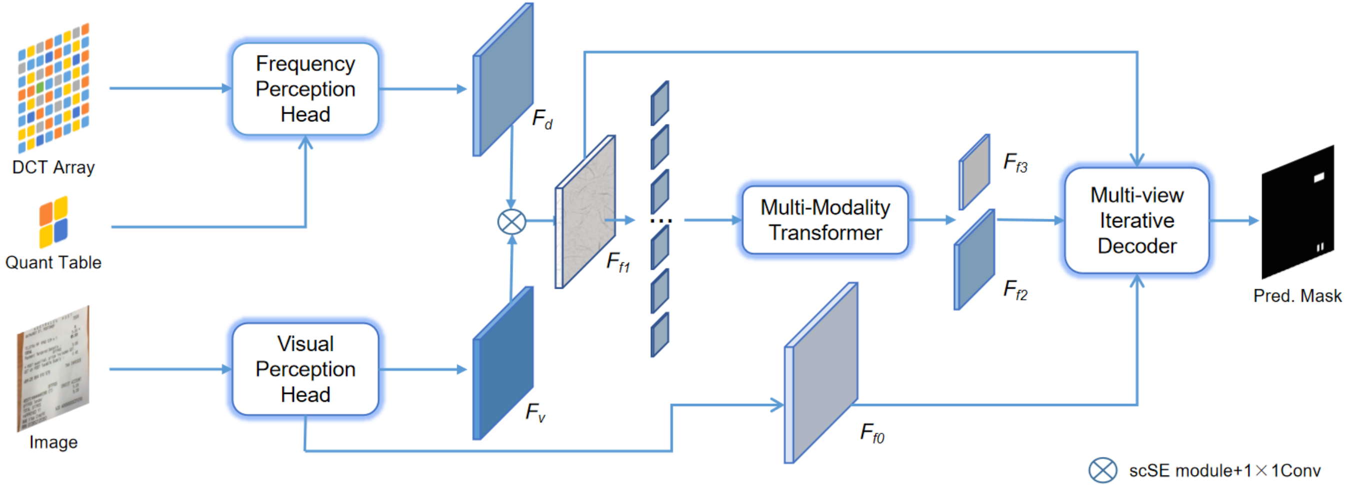
DocTamper dataset

The main features of DocTamper dataset are summarized as：

- ✓ **Large Scale**. The public datasets in previous works only have less than 1k images, while Doc-Tamper has total 170k images.
- ✓ **Board Diversity**. To build the DocTamper Dataset, we collect 50,562 document images from various publicly available websites and document image datasets. Various bilingual real-world document images including contracts, invoices, receipts, etc., are included in the source images of our dataset.
- ✓ **Comprehensiveness**. All the three commonly used text tampering methods are included in our dataset to imitate the real-world applications. In Addition, we introduce two cross-domain testing subsets to fully evaluate the generalization ability of different methods.

Document Tampering Detector

Document Tampering Detector-Framework



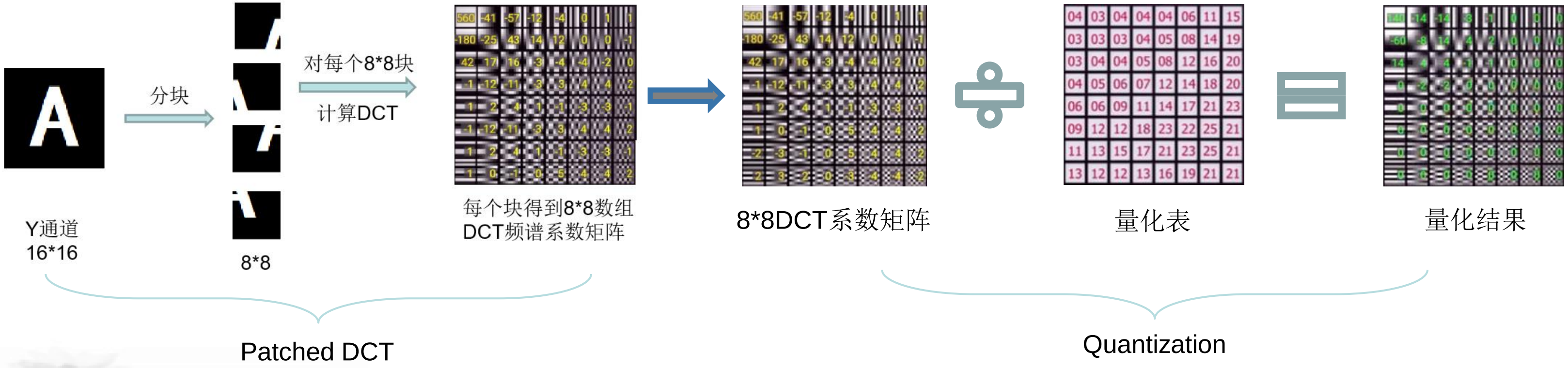
The proposed DTD can detect the tampered texts that have seldom visual tampering clues since the tampering operation can cause the discontinues of block artifacts grid.

The overall framework of the proposed model (Document Tampering Detector, DTD) are shown in the figure above. It extracts visual features through visual perception heads and extracts frequency features through frequency perception head, then fuse them in an early fusion manner before fed them into multi-modality transformer. Finally a multi-view iterative decoder is utilized to get the predictions with the multi-modal features.

Theoretical Analysis

Block Artifacts Grid

The photographed document images will undergo slight lossy compression on almost any device after being captured. Currently, most lossy compression of images will generate block effects. Taking the most common JPEG image as an example, it independently quantizes each 8 * 8 image block in the Y-channel in the frequency domain after imaging to remove high-frequency components and reduce storage occupation, as shown in the following figure.



04	03	04	04	04	06	11	15
03	03	03	04	05	08	14	19
03	04	04	05	08	12	16	20
04	05	06	07	12	14	18	20
06	06	09	11	14	17	21	23
09	12	12	18	23	22	25	21
11	13	15	17	21	23	25	21
13	12	12	13	16	19	21	21

The independent frequency domain quantization of 8 * 8 blocks will cause the numerical characteristics of the DCT coefficient matrix of the image to exhibit periodic changes of 8 cycles. Presenting a grid like distribution, which is known as the Block Artifacts Grid (BAG) [2] 。

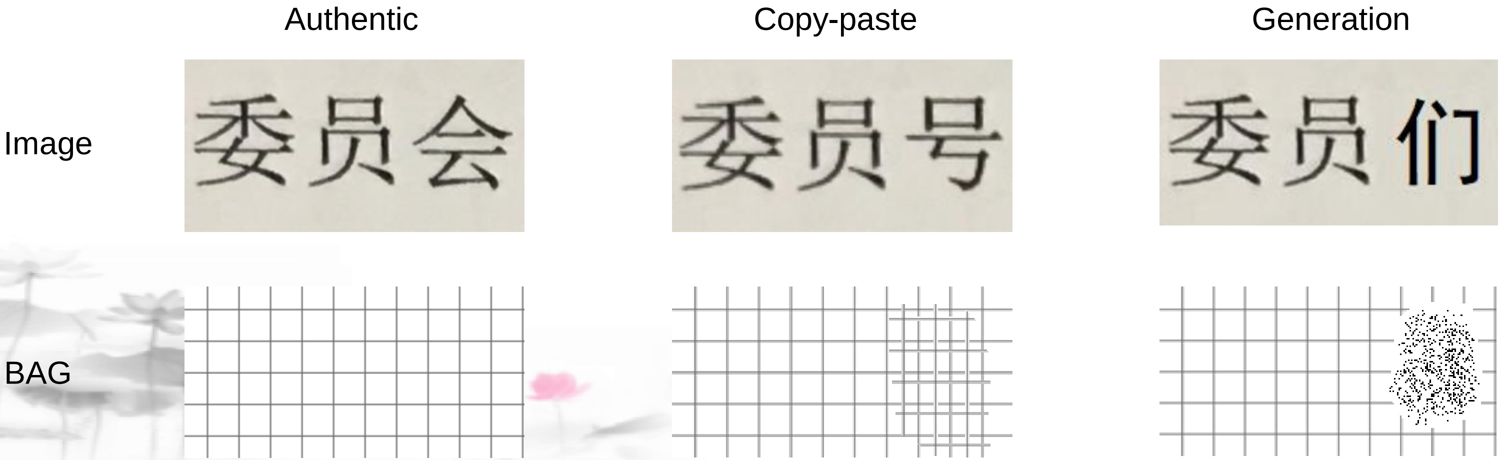
[2] Li, W., Yuan, Y. and Yu, N., 2009. Passive detection of doctored JPEG image via block artifact grid extraction. Signal Processing, 89(9), pp.1821-1829.

Theoretical Analysis

Tampering on photographed document image mostly will break the continuity of BAG

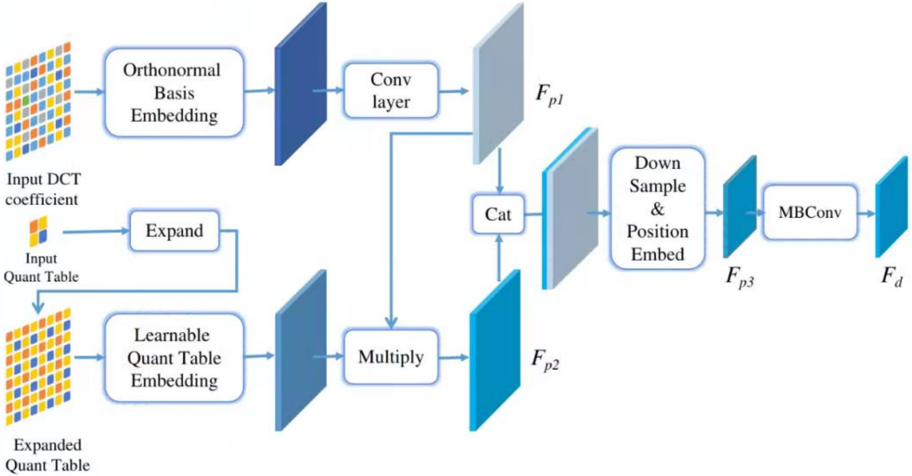
Tampering operations typically disrupt the continuity of the block artifacts grid, resulting in fracture of the block artifacts grid at the edge of the tampered area and abnormal distribution of DCT coefficients within the tampered area. Although careful document image tampering is difficult to capture at the visual domain, if frequency domain input is processed in an appropriate way, local anomalies in the block artifacts grid can help locate the tampered area.

On the other hand, various types of tampering can lead to discontinuity and distribution anomalies in the block artifacts grid, which can be captured through appropriate methods. Being able to unify different types of tampering methods into an approximate pattern is beneficial for helping the model achieve stronger generalization.



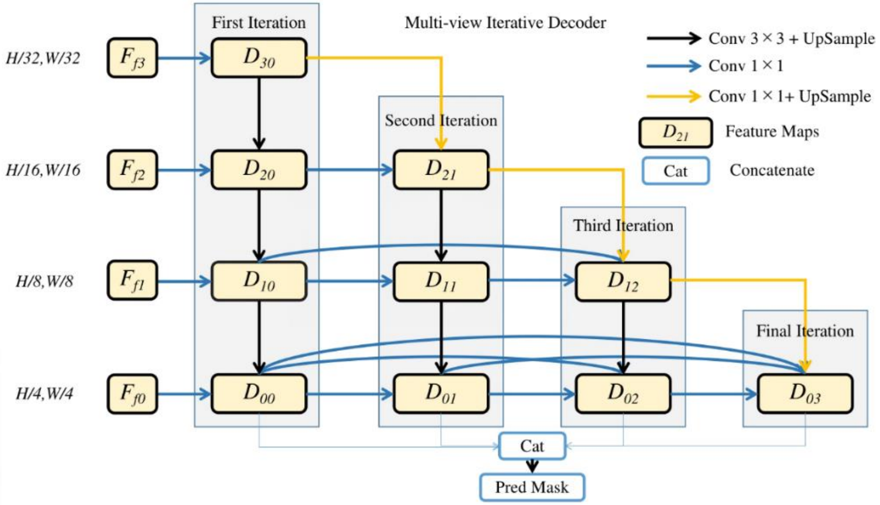
Frequency Perception Head

1. **Motivation**： Assisting in locating tampered text that have seldom visual traces by capturing abnormal in block artifacts grid.
2. **Core concepts**： Encode the discrete DCT coefficients with orthogonal basis embedding to sensitively capture local anomalies in BAG; Automatically learn the optimal features under different compression settings through learnable quantization table embedding; Down-sampling with a conv layer of kernel size 8 to align the period of the BAG and aligns visual features through position encoding.

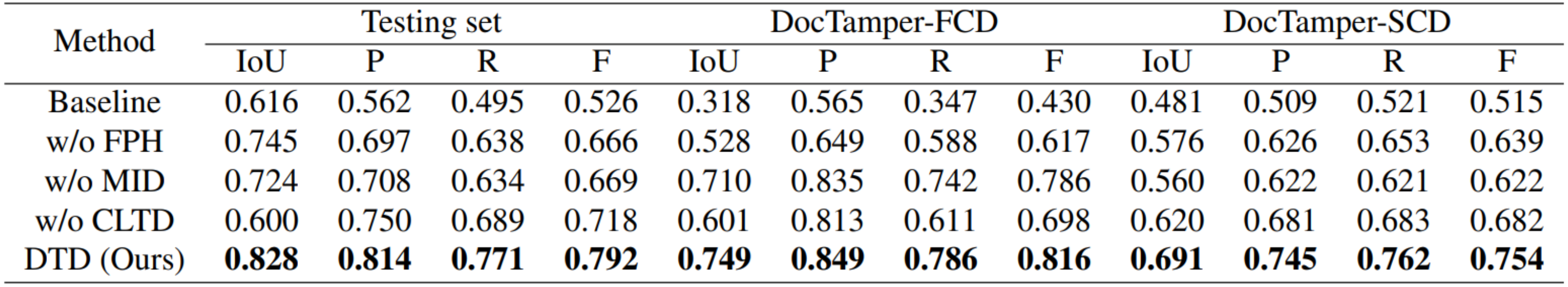


Multi-view Iterative Decoder

1. **Motivation**：To mimic the process people zoom in and out the image over and over again, combining clues of different views and analysis them repeatedly to do careful tampering detection.
2. **Core concept**：Alternately ① Fuse high-level semantic information and low-level detail in vertical direction. ② Continuously refine the features in horizontal direction.

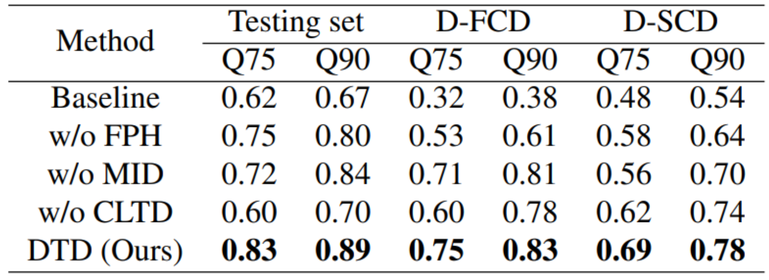


Ablation experiments



Method	Testing set				DocTamper-FCD				DocTamper-SCD			
	IoU	P	R	F	IoU	P	R	F	IoU	P	R	F
Baseline	0.616	0.562	0.495	0.526	0.318	0.565	0.347	0.430	0.481	0.509	0.521	0.515
w/o FPH	0.745	0.697	0.638	0.666	0.528	0.649	0.588	0.617	0.576	0.626	0.653	0.639
w/o MID	0.724	0.708	0.634	0.669	0.710	0.835	0.742	0.786	0.560	0.622	0.621	0.622
w/o CLTD	0.600	0.750	0.689	0.718	0.601	0.813	0.611	0.698	0.620	0.681	0.683	0.682
DTD (Ours)	**0.828**	**0.814**	**0.771**	**0.792**	**0.749**	**0.849**	**0.786**	**0.816**	**0.691**	**0.745**	**0.762**	**0.754**

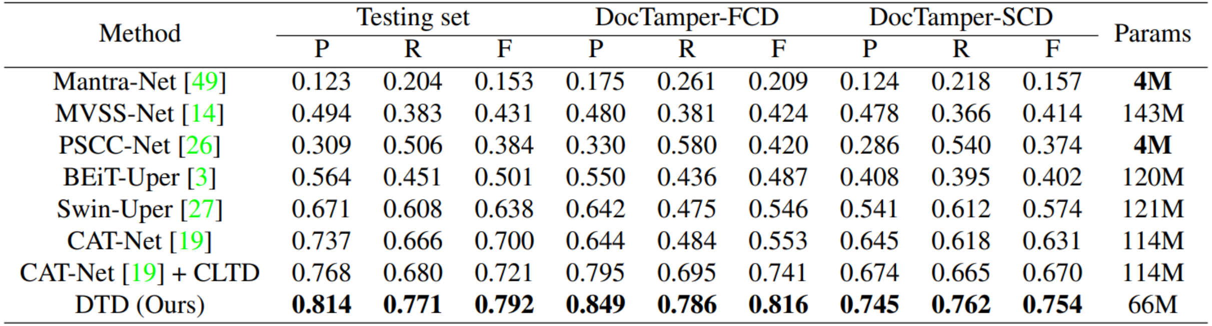
IoU metric



Method	Testing set		D-FCD		D-SCD	
	Q75	Q90	Q75	Q90	Q75	Q90
Baseline	0.62	0.67	0.32	0.38	0.48	0.54
w/o FPH	0.75	0.80	0.53	0.61	0.58	0.64
w/o MID	0.72	0.84	0.71	0.81	0.56	0.70
w/o CLTD	0.60	0.70	0.60	0.78	0.62	0.74
DTD (Ours)	**0.83**	**0.89**	**0.75**	**0.83**	**0.69**	**0.78**

Ablation experiments shows the effectiveness of each proposed modules.

Comparison experiments

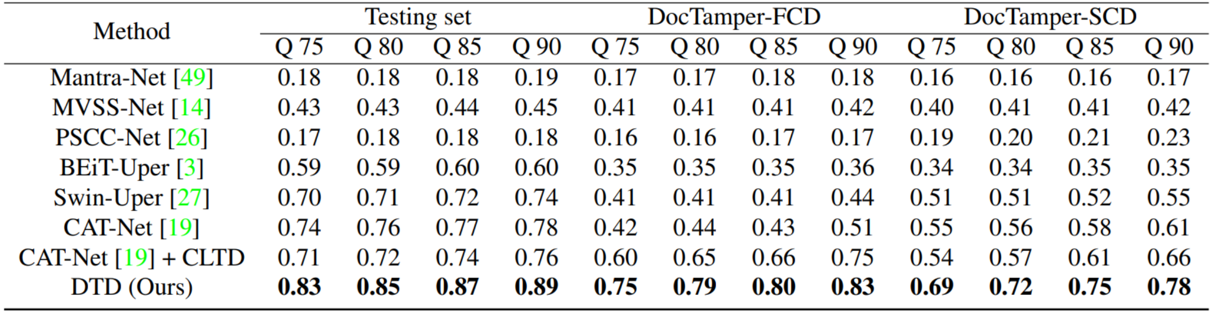


Method	Testing set			DocTamper-FCD			DocTamper-SCD			Params
	P	R	F	P	R	F	P	R	F	
Mantra-Net [49]	0.123	0.204	0.153	0.175	0.261	0.209	0.124	0.218	0.157	**4M**
MVSS-Net [14]	0.494	0.383	0.431	0.480	0.381	0.424	0.478	0.366	0.414	143M
PSCC-Net [26]	0.309	0.506	0.384	0.330	0.580	0.420	0.286	0.540	0.374	**4M**
BEiT-Uper [3]	0.564	0.451	0.501	0.550	0.436	0.487	0.408	0.395	0.402	120M
Swin-Uper [27]	0.671	0.608	0.638	0.642	0.475	0.546	0.541	0.612	0.574	121M
CAT-Net [19]	0.737	0.666	0.700	0.644	0.484	0.553	0.645	0.618	0.631	114M
CAT-Net [19] + CLTD	0.768	0.680	0.721	0.795	0.695	0.741	0.674	0.665	0.670	114M
DTD (Ours)	**0.814**	**0.771**	**0.792**	**0.849**	**0.786**	**0.816**	**0.745**	**0.762**	**0.754**	66M

In the DocTamper dataset, experiments are conducted under the same training/testing settings.

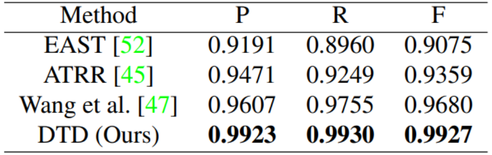
Comparison experiments shows that the proposed method outperforms previous methods.

IoU metric



Method	Testing set				DocTamper-FCD				DocTamper-SCD			
	Q 75	Q 80	Q 85	Q 90	Q 75	Q 80	Q 85	Q 90	Q 75	Q 80	Q 85	Q 90
Mantra-Net [49]	0.18	0.18	0.18	0.19	0.17	0.17	0.18	0.18	0.16	0.16	0.16	0.17
MVSS-Net [14]	0.43	0.43	0.44	0.45	0.41	0.41	0.41	0.42	0.40	0.41	0.41	0.42
PSCC-Net [26]	0.17	0.18	0.18	0.18	0.16	0.16	0.17	0.17	0.19	0.20	0.21	0.23
BEiT-Uper [3]	0.59	0.59	0.60	0.60	0.35	0.35	0.35	0.36	0.34	0.34	0.35	0.35
Swin-Uper [27]	0.70	0.71	0.72	0.74	0.41	0.41	0.41	0.44	0.51	0.51	0.52	0.55
CAT-Net [19]	0.74	0.76	0.77	0.78	0.42	0.44	0.43	0.51	0.55	0.56	0.58	0.61
CAT-Net [19] + CLTD	0.71	0.72	0.74	0.76	0.60	0.65	0.66	0.75	0.54	0.57	0.61	0.66
DTD (Ours)	**0.83**	**0.85**	**0.87**	**0.89**	**0.75**	**0.79**	**0.80**	**0.83**	**0.69**	**0.72**	**0.75**	**0.78**

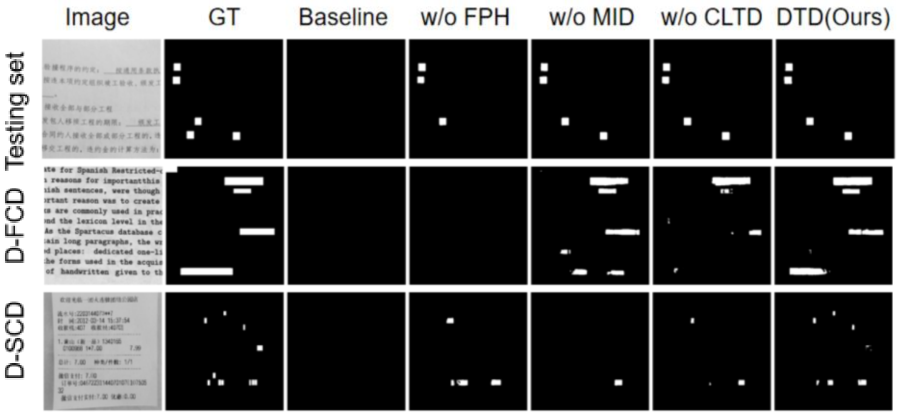
T-SROIE实验结果



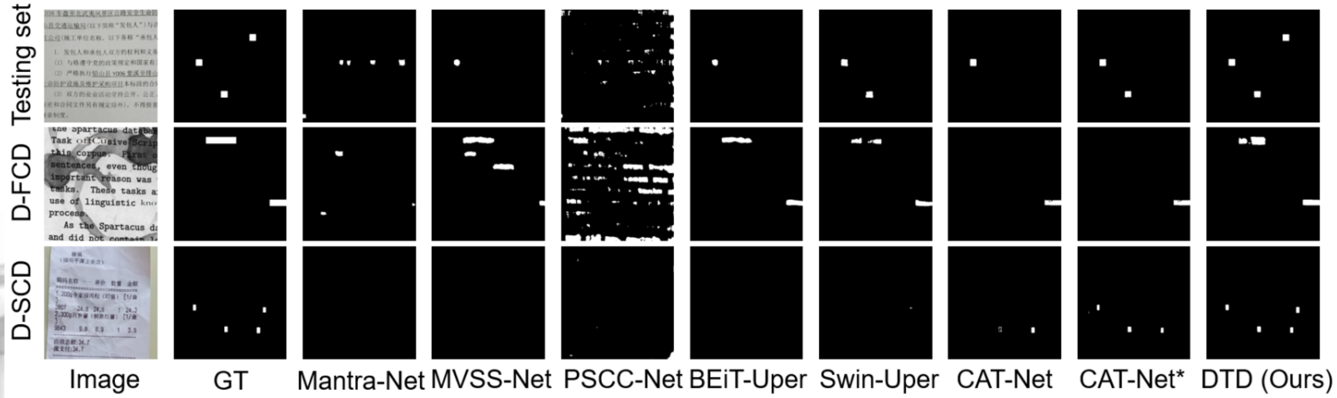
Method	P	R	F
EAST [52]	0.9191	0.8960	0.9075
ATRR [45]	0.9471	0.9249	0.9359
Wang et al. [47]	0.9607	0.9755	0.9680
DTD (Ours)	**0.9923**	**0.9930**	**0.9927**

Visualization

Visualization of Ablation experiments



Visualization of Comparison experiments



Thank you