

Token-to-Pixel Semantic Guidance for Document Image Tampering Localization

Tiancheng Wang^{1,3} Haoyu Huang^{2,3} Zichao Feng¹

Linlin Yang^{4*} Sheng Xu⁴ Baochang Zhang^{1,3,5}

¹Institute of Artificial Intelligence, Beihang University, Beijing, China

²National College for Excellent Engineers, Beihang University, Beijing, China

³Zhongguancun Laboratory, Beijing, China

⁴State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing, China

⁵Beihang University Hangzhou Innovation Research Institute

Abstract

Document image tampering poses a serious threat to information authenticity in security-critical documents. Existing document tampering detection methods mainly rely on RGB-domain and frequency-domain artifacts, following a paradigm similar to natural image forensics, and often overlook a key property of document images: dense and semantically rich textual content. In this work, we propose TSPG, a plug-and-play Textual Semantic Prior Guidance module for document image tampering detection. TSPG uses a pre-trained language model with a lightweight token classification head to estimate token-level tampering-aware semantic scores from OCR-recognized text. It then projects these scores back to the image space through OCR bounding boxes to construct a spatially aligned Textual Semantic Prior (TSP) map. To inject this prior into existing detectors, we design a Gated Semantic Residual Modulation (GSRM) module that adaptively modulates multi-scale detector features through residual gated guidance. TSPG can be inserted into existing document tampering detectors without modifying their original backbone or decoder. Experiments on DocTamper with two representative baselines, DTD and FFDN, together with additional external-dataset and OCR robustness evaluations, demonstrate the effectiveness and generalizability of TSPG and the complementary value of language-derived textual semantics. Code is available at <https://github.com/tcwangbuaa/TSPG>.

1 Introduction

Document images, such as contracts, invoices, certificates, and receipts, are widely used in security-critical scenarios. With the advance of image editing tools and generative models, malicious users can easily manipulate key textual elements. Such

*Corresponding author: lyang@cuc.edu.cn.

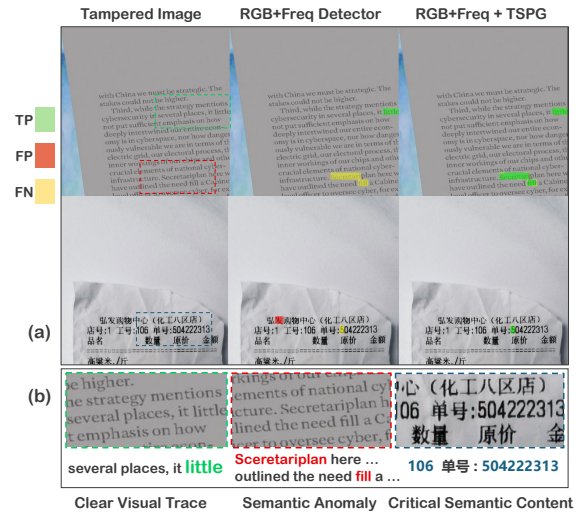


Figure 1: Motivating examples of textual semantic guidance for document tampering localization. (a) RGB-frequency detectors can localize tampered regions when visual traces are clear, but may produce false detections or miss subtle manipulations when such traces are weak. TSPG complements these forensic cues with textual semantic guidance. (b) The zoomed examples illustrate clear visual traces, contextual semantic anomalies, and critical semantic content such as numbers and identifiers, showing how semantic cues can help localize subtle tampered text. TP, FP, and FN denote correctly detected, falsely detected, and missed regions, respectively.

tampering may alter the legal, financial, or administrative validity of documents, making accurate document tampering localization an important forensic task.

Existing document and text tampering localization methods largely inherit the paradigm of natural image manipulation localization, treating the task as pixel-level forgery segmentation and mainly relying on RGB-domain artifacts, frequency-domain inconsistencies, or transformed-domain cues such as DCT, SRM, and ELA (Qu et al., 2023; Chen et al., 2024; Luo

et al., 2024). Although recent studies improve robustness by adaptively controlling unreliable frequency features or disentangling irrelevant document factors (Wong et al., 2025; Zheng et al., 2026), the textual semantic nature of document images remains insufficiently explored.

Unlike natural image manipulation, document image tampering usually targets text regions that carry critical information. Since document images often have simple backgrounds and structured layouts, tampered regions are typically small, concealed, and leave fewer visual traces, requiring fine-grained localization. More importantly, text manipulation often aims to change the contextual information conveyed by text, suggesting that semantic changes and visual traces are usually spatially correlated. As illustrated in Figure 1, RGB-frequency detectors can localize tampered regions when visual traces are clear, but may produce false detections or miss tampered text when manipulation traces are subtle. In such cases, textual semantics can provide complementary guidance by highlighting semantically high-risk regions, such as contextual semantic anomalies and critical semantic content. Moreover, visual and frequency-domain traces can be weakened by post-processing operations such as JPEG compression (Qu et al., 2023; Chen et al., 2024), whereas textual semantics are relatively more stable as long as the text remains recognizable. Therefore, textual semantic information can provide valuable complementary evidence for more robust document tampering localization.

Although some recent studies have explored textual semantics for tampered text detection, they are usually designed for specific settings, such as bilingual semantic consistency checking or MLLM-based explainable reasoning (Li et al., 2025; Xu et al., 2025). Different from them, our goal is to leverage the textual semantics of document images to guide tampering localization models and improve detection performance.

In this work, we propose **TSPG**, a plug-and-play **Textual Semantic Prior Guidance** module for document image tampering detection. Existing document tampering detectors typically extract multi-scale visual or frequency-domain representations and aggregate them for pixel-level localization. To introduce textual semantics into this pixel-level process, TSPG first uses a frozen pretrained language model with a lightweight token classification head to estimate token-level

tampering-aware semantic scores from OCR-recognized text. These scores are then projected back to the image plane through OCR bounding boxes, forming a spatially aligned **Textual Semantic Prior** (TSP) map. This token-to-pixel grounding bridges sequential language modeling and dense visual forensic localization, allowing textual semantic cues to guide existing detectors in the image space.

To make the TSP compatible with multi-scale detector features, we construct a lightweight CNN-based prior pyramid that generates single-channel TSP maps at different resolutions. Since the TSP may be affected by imperfect token scores and coarse box-level spatialization, it is used as soft guidance rather than a hard mask. Specifically, we design a **Gated Semantic Residual Modulation** (GSRM) module to inject each scale of TSP into the corresponding detector feature through residual gated modulation. In this way, TSPG can guide existing detectors at the feature level while preserving their original backbone and decoder structures. Experiments on DocTamper with DTD (Qu et al., 2023) and FFDN (Chen et al., 2024) show that TSPG consistently improves both baselines, demonstrating the complementary value of language-derived textual semantics.

Our main contributions are summarized as follows:

- We propose TSPG, a plug-and-play Textual Semantic Prior Guidance module for document image tampering detection, which enhances existing pixel-level tampering localization models through feature-level semantic modulation without modifying their original backbone or decoder.
- We introduce a Textual Semantic Prior construction strategy that projects token-level tampering-aware semantic scores from a frozen pretrained language model back to the image space through OCR bounding boxes, bridging sequential language modeling and pixel-level tampering localization.
- We design a lightweight prior pyramid and a Gated Semantic Residual Modulation mechanism to achieve scale-aware semantic guidance, allowing textual priors to modulate detector features while preserving the original forensic representations.

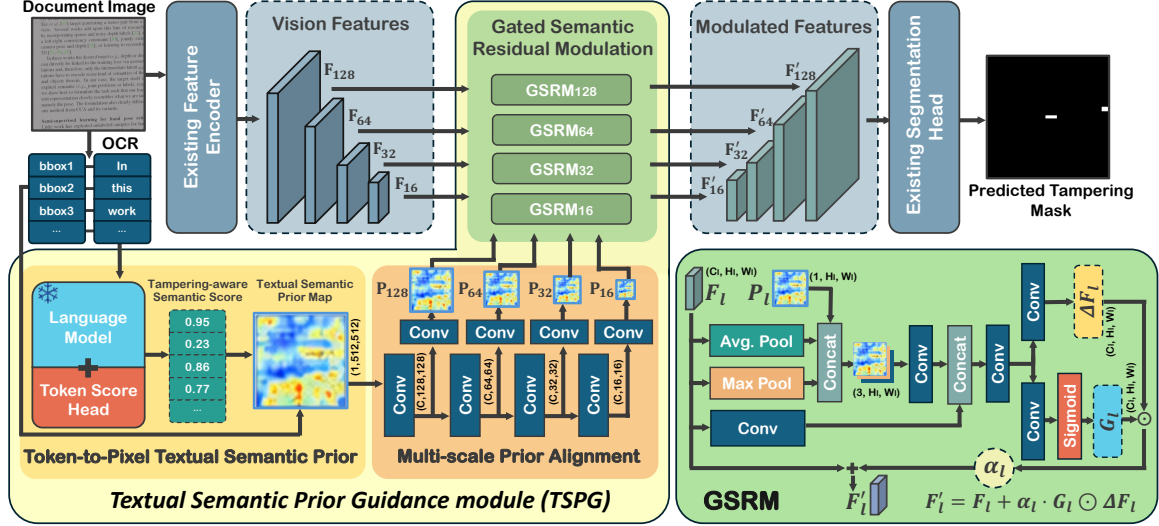


Figure 2: Overall framework of the proposed plug-and-play Textual Semantic Prior Guidance (TSPG) module. TSPG preserves the detector components (blue regions) and introduces three components. OCR-recognized text is encoded by a frozen language model with a lightweight token score head, and token-level semantic scores are projected back to the image space through OCR bounding boxes to construct a Textual Semantic Prior map (yellow region). The prior is then aligned with multi-scale detector features (orange region) and injected via Gated Semantic Residual Modulation (green region), without modifying the original encoder or segmentation head.

- We evaluate TSPG with two representative document tampering detectors, DTD and FFDN, and further conduct external-dataset and OCR robustness evaluations, demonstrating consistent improvements across different settings.

2 Related Work

2.1 Document Image Tampering Localization

Image manipulation localization has been widely studied in multimedia forensics, where the goal is to localize forged regions caused by splicing, copy-move, inpainting, or other editing operations. General image forgery localization methods usually exploit RGB artifacts, noise residuals, compression traces, boundary inconsistencies, or multi-scale visual representations (Liu et al., 2022a; Kwon et al., 2021; Guillaro et al., 2023).

Early document forgery detection methods mainly relied on handcrafted cues, such as font inconsistency, text-line geometry, image quality measures, and local texture patterns (Bertrand et al., 2015; van Beusekom et al., 2013; Cruz et al., 2017). With the development of deep learning, document tampering localization has been increasingly formulated as a pixel-level segmentation task. Qu et al. (2023) proposed DocTamper, a large-scale benchmark for tampered text detection, and DTD, which combines RGB features

with DCT-domain forensic cues for document images. Following this RGB-frequency paradigm, FFDN (Chen et al., 2024) improves tampered text detection through frequency feature fusion and decomposition, while ADCD-Net (Wong et al., 2025) adaptively modulates DCT contributions to enhance robustness against block-misalignment distortions such as resizing, cropping, and pixel shifting. In another line, CD-SD (Zheng et al., 2026) disentangles background color and local visual semantic factors, such as font, location, and shape, from forensic features to improve cross-dataset generalization.

Despite these advances, the above methods still primarily rely on visual and frequency-domain forensic evidence, leaving the textual semantics of document images underexplored.

2.2 Textual Semantics for Document Tampering Detection

Unlike generic image manipulation, text tampering is inherently semantic, as it often aims to distort the contextual information conveyed by text (Luo et al., 2024). This has motivated recent attempts to introduce textual semantics into tampering detection. Tornés et al. (2023) mainly rely on OCR-derived textual representations and domain-specific ontology-based entities and relations for forged receipt detection, without model-

ing RGB or frequency-domain forensic cues. Li et al. (2025) detect tampered text through bilingual semantic consistency, which requires paired multilingual content. Xu et al. (2025) use multi-modal large language models for explainable tampered text detection, focusing on semantic reasoning and interpretation.

However, existing semantic-related methods either rely mainly on textual reasoning or are designed for specific scenarios, without effectively combining textual semantics with RGB-frequency forensic cues, both of which are important for document tampering localization. In contrast, we estimate token-level tampering-aware semantic scores from OCR-recognized text and project them back to the image space through OCR bounding boxes, forming Textual Semantic Prior maps that guide existing visual forensic detectors through TSPG.

3 Method

3.1 Overview

Given a document image $I \in \mathbb{R}^{H \times W \times 3}$, document image tampering detection aims to predict a binary tampering mask $\hat{Y} \in [0, 1]^{H \times W}$. Existing detectors usually extract visual or frequency-domain features and perform pixel-level localization with multi-scale representations. Here, we propose TSPG, a plug-and-play Textual Semantic Prior Guidance module that introduces textual semantics into existing tampering detectors.

As shown in Figure 2, TSPG preserves the original detector components (blue region) and introduces three main steps: (i) OCR-recognized tokens are fed into a frozen pretrained language model with a lightweight token classification head to estimate token-level tampering-aware semantic scores. These scores are projected back to the image plane through OCR bounding boxes, forming a spatially aligned Textual Semantic Prior (TSP) map (yellow region); (ii) a lightweight prior pyramid aligns the TSP with multi-scale detector features (orange region); (iii) Gated Semantic Residual Modulation (GSRM) injects the aligned priors into intermediate detector features through soft residual gating (green region).

Note that TSPG only modulates intermediate feature representations of existing methods and does not require modifying the original backbone or decoder. Therefore, it can be seamlessly integrated into current pixel-level document tampering localization models.

3.2 Token-to-Pixel Textual Semantic Prior

To align token-level textual semantics with pixel-level visual features, we construct a token-to-pixel Textual Semantic Prior (TSP). Given a document image I , OCR produces a sequence of recognized tokens and their bounding boxes:

$$\mathcal{T} = \{(t_i, b_i)\}_{i=1}^N, \quad (1)$$

where t_i denotes the i -th token and $b_i = (x_i^1, y_i^1, x_i^2, y_i^2)$ denotes its bounding box in the image coordinate system.

We feed the OCR-recognized text into a frozen pretrained language model and attach a lightweight trainable token classification head $g_\phi(\cdot)$ to estimate token-level tampering-aware semantic scores:

$$s_i = \sigma(g_\phi(\mathbf{h}_i)), \quad \mathbf{h}_i = \text{LM}(t_i), \quad (2)$$

where $\text{LM}(\cdot)$ denotes the language model, and $\sigma(\cdot)$ is the sigmoid function, $s_i \in [0, 1]$ and $\mathbf{h}_i$ is the contextual representation of t_i .

Since DocTamer provides pixel-level masks rather than token-level labels, we derive soft token supervision labels from the overlap between each OCR box and the ground-truth mask Y :

$$y_i^{\text{tok}} = \frac{1}{|b_i|} \sum_{(x,y) \in b_i} Y(x, y). \quad (3)$$

After obtaining token-level scores, we project them back to the image plane according to OCR bounding boxes:

$$P(x, y) = \begin{cases} \max_{i:(x,y) \in b_i} s_i, & (x, y) \in \bigcup_{i=1}^N b_i, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

The resulting $P \in [0, 1]^{H \times W}$ converts sequential token scores into a spatially aligned TSP map, which serves as soft semantic guidance rather than a hard tampering mask.

3.3 Multi-Scale Prior Alignment

Pixel-level tampering detectors rely on multi-scale feature maps, while the initial TSP map P is constructed in the original image space. Directly resizing P to each feature scale is simple but cannot adapt the coarse box-level prior to detector features. Therefore, we introduce a lightweight CNN-based prior pyramid $\mathcal{E}_p$ to generate scale-aligned prior maps:

$$\{P_l\}_{l=1}^L = \mathcal{E}_p(P), \quad P_l \in \mathbb{R}^{1 \times H_l \times W_l}. \quad (5)$$

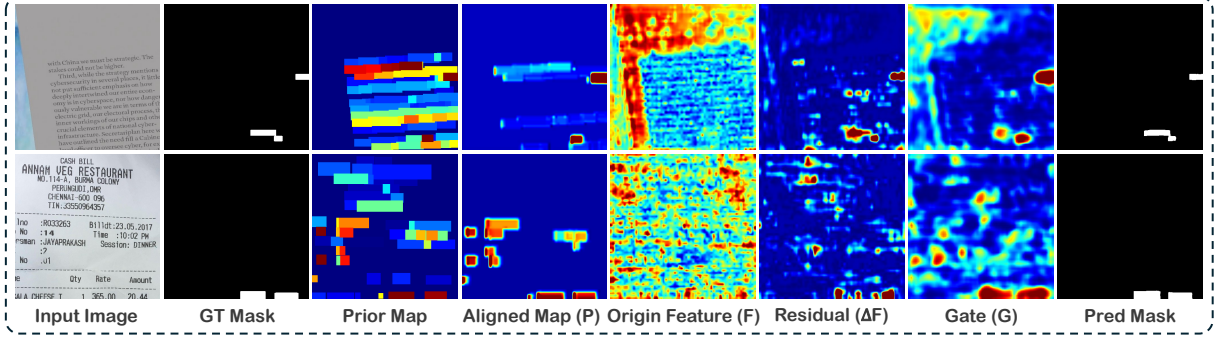


Figure 3: Visualization of intermediate maps in TSPG. The token-to-pixel textual prior is aligned to detector feature scales, and GSRM generates residual and gate maps to softly modulate the original features. The modulated features lead to a prediction that better matches the ground-truth tampering mask.

Table 1: Comparison on the DocTammer dataset. ‘P’ denotes precision, ‘R’ denotes recall, and ‘F1’ denotes F1-score. Results in the upper block are taken from ADCD-Net (Wong et al., 2025). Methods marked with * are reproduced under our implementation. The best results are highlighted in **bold**, and the second-best results are underlined.

Method	Test (30,000)			FCD (2,000)			SCD (18,000)			AVG F1
	P	R	F1	P	R	F1	P	R	F1	
MVSS-Net	0.494	0.383	0.431	0.480	0.381	0.424	0.478	0.366	0.414	0.423
CAT-Net	0.768	0.680	0.721	0.795	0.695	0.741	0.674	0.665	0.670	0.711
TruFor	0.673	0.710	0.691	0.659	0.675	0.667	0.528	0.678	0.594	0.651
ConvNeXt-UPer	0.676	0.591	0.630	0.606	0.450	0.517	0.550	0.521	0.535	0.561
MA-Net	0.459	0.473	0.466	0.335	0.405	0.367	0.349	0.508	0.414	0.416
PS-Net	0.698	0.972	0.798	0.716	0.987	0.827	0.659	0.909	0.741	0.789
ADCD-Net	0.789	0.823	0.806	0.866	0.770	0.815	0.690	0.799	0.740	0.787
DTD*	0.811	0.766	0.788	0.866	0.787	0.824	0.739	0.754	0.746	0.786
DTD* + TSPG	0.832	0.794	0.812	0.878	0.824	0.850	0.767	0.774	0.770	0.811
FFDN*	0.851	0.811	0.830	0.902	0.870	0.886	0.799	0.784	0.791	0.836
FFDN* + TSPG	0.882	<u>0.863</u>	0.873	0.914	<u>0.901</u>	0.907	<u>0.793</u>	<u>0.821</u>	0.807	0.862

The prior pyramid progressively downsamples the single-channel TSP map with lightweight convolutional blocks and projects each scale back to a single-channel prior through a 1×1 convolution, as illustrated by an example in Figure 2.

3.4 Gated Semantic Residual Modulation

After obtaining multi-scale TSP maps, we inject them into detector features through Gated Semantic Residual Modulation. Let

$$F_l \in \mathbb{R}^{C_l \times H_l \times W_l}, \quad P_l \in \mathbb{R}^{1 \times H_l \times W_l}, \quad (6)$$

denote the detector feature and its corresponding TSP map at scale l . To combine the textual prior with detector responses, we first compute channel-wise average and maximum maps from F_l :

$$A_l = \text{AvgPool}_c(F_l), \quad M_l = \text{MaxPool}_c(F_l), \quad (7)$$

where $A_l, M_l \in \mathbb{R}^{1 \times H_l \times W_l}$. They are then concatenated with P_l to form:

$$U_l = \text{Concat}(A_l, M_l, P_l). \quad (8)$$

The prior-aware spatial map U_l and the original detector feature F_l are then projected and fused:

$$Z_l = f_m(\text{Concat}(f_s(U_l), f_v(F_l))), \quad (9)$$

where $f_s(\cdot)$, $f_v(\cdot)$, and $f_m(\cdot)$ are lightweight convolutional transformations. Based on Z_l , GSRM predicts a residual feature and a gate:

$$\Delta F_l = f_r(Z_l), \quad G_l = \sigma(f_g(Z_l)). \quad (10)$$

where $f_r(\cdot)$ and $f_g(\cdot)$ are lightweight convolutional transformations, and $\sigma(\cdot)$ is the sigmoid function. The final modulated feature is:

$$F'_l = F_l + \alpha_l \cdot G_l \odot \Delta F_l, \quad (11)$$

where $\odot$ denotes element-wise multiplication and α_l is a learnable scalar controlling the modulation

strength. In practice, we set $\alpha_l = \alpha_{\max}\sigma(a_l)$ to constrain its range.

This residual gated formulation preserves the original detector feature while selectively injecting textual semantic guidance. The residual branch provides tampering-aware feature compensation, and the gate branch controls where such compensation should be activated. Their combination helps highlight tampering-relevant regions and suppress irrelevant responses. The enhanced features from all selected scales are then fed into the original neck or decoder for final mask prediction.

3.5 Training Objective

TSPG keeps the original pixel-level localization loss of the target detector. In our experiments, DTD and FFDN use the combination of cross-entropy and Lovasz loss:

$$\mathcal{L}_{seg} = \mathcal{L}_{ce}(\hat{Y}, Y) + \mathcal{L}_{lovasz}(\hat{Y}, Y). \quad (12)$$

Based on Eq. (2) and Eq. (3), We only introduce an additional token-level supervision loss for the lightweight token classification head:

$$\mathcal{L}_{tok} = \text{BCE}(\mathbf{s}, \mathbf{y}^{tok}). \quad (13)$$

The final objective is:

$$\mathcal{L} = \mathcal{L}_{seg} + \lambda_{tok}\mathcal{L}_{tok}. \quad (14)$$

During inference, TSP is constructed from OCR results and predicted token scores, without using ground-truth masks.

4 Experiments

4.1 Experimental Setup

Dataset. We conduct experiments on DocTammer (Qu et al., 2023), which contains 170K tampered Chinese and English document images generated by copy-move, splicing, and generation-based manipulations. Models are trained on 120K images and evaluated on DocTammer-Test with 30K same-domain samples, as well as FCD and SCD with 2K and 18K cross-domain samples. Each DocTammer-Test sample has a fixed JPEG quality factor, enabling robustness evaluation under different compression levels. To further evaluate generalization beyond DocTammer, we additionally conduct experiments on T-SROIE (Wang et al., 2022), a tampered receipt dataset. We evaluate both cross-dataset transfer, where models trained on DocTammer are directly tested on T-SROIE, and in-domain training on T-SROIE.

Evaluation Metrics. Following prior document tampering detection (Qu et al., 2023; Wong et al., 2025), we report pixel-level precision (P), recall (R), and F1-score (F) in the main comparison. For ablation studies, we additionally report Intersection-over-Union (IoU) together with F1-score to evaluate pixel-level localization quality.

Baselines. We compare our method with representative localization methods, including MVSS-Net (Dong et al., 2022), CAT-Net (Kwon et al., 2021), TruFor (Guillaro et al., 2023), ConvNeXt-UPer (Liu et al., 2022b; Xiao et al., 2018), MA-Net (Xu et al., 2023), PS-Net (Shao et al., 2023), and ADCD-Net (Wong et al., 2025). We reproduce DTD (Qu et al., 2023) and FFDN (Chen et al., 2024), and insert TSPG into their intermediate feature stages to evaluate its plug-and-play compatibility.

Implementation Details. Following DTD (Qu et al., 2023), all models are trained on DocTammer for 100K iterations with a batch size of 12 and a cosine learning rate schedule. The learning rate decays from 3×10^{-4} to 1×10^{-5} for DTD/DTD+TSPG, and from 7×10^{-5} to 5×10^{-6} for FFDN/FFDN+TSPG. We set $\lambda_{tok} = 0.1$ for DTD+TSPG and $\lambda_{tok} = 0.3$ for FFDN+TSPG. We use frozen mDeBERTa (He et al., 2021) with a trainable token classification head, and keep PP-OCRv5 (Cui et al., 2025) OCR results fixed for token supervision and TSP construction during training and inference. All experiments are conducted on three NVIDIA RTX 4090 GPUs.

4.2 Results and Analysis

As shown in Table 1, TSPG consistently improves both reproduced baselines across all three test sets. For DTD, the average F1 score increases from 0.786 to 0.811, with consistent gains on Test, FCD, and SCD. For the stronger FFDN baseline, TSPG further improves the average F1 score from 0.836 to 0.862 and achieves the best overall performance among all compared methods. These results indicate that textual semantic guidance can effectively complement visual and frequency-domain forensic cues under both same-domain and cross-domain settings. The improvements also show a better balance between precision and recall. Compared with methods such as PS-Net, which achieve high recall but relatively lower precision, TSPG improves localization performance while maintaining competitive

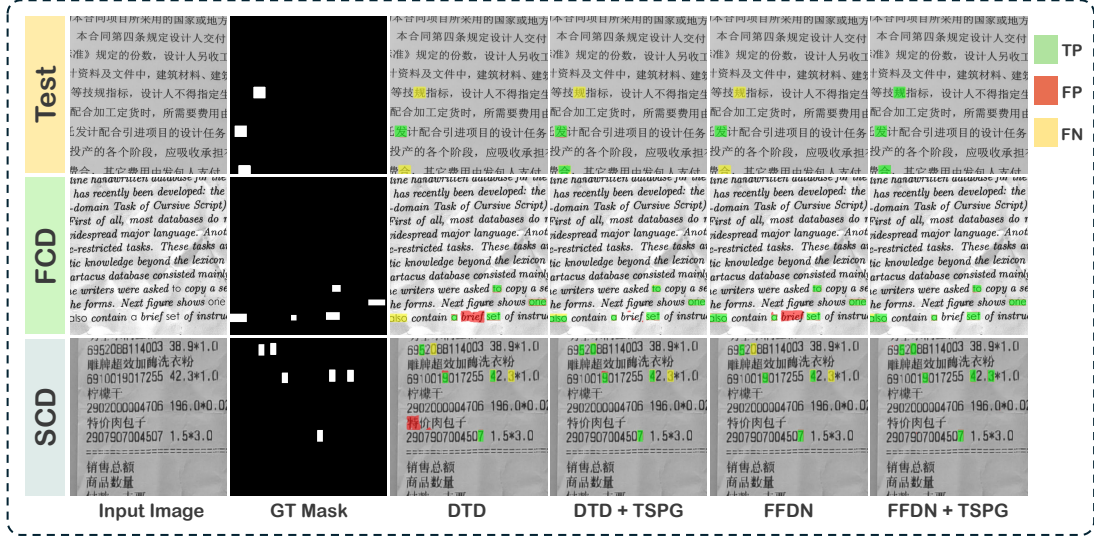


Figure 4: Qualitative comparison on the DocTammer Test, FCD, and SCD sets. TP, FP, and FN denote correctly detected, falsely detected, and missed regions, respectively. Adding TSPG improves the localization results of both DTD and FFDN, especially for subtle tampered regions.

Table 2: Robustness under different JPEG compression qualities on the DocTammer-Test set. We report precision (P), recall (R), and F1-score (F1) under fixed quality factors $Q = 100, 90, 75$. Δ_{90} and Δ_{75} denote the absolute F1 drops from $Q = 100$ to $Q = 90$ and $Q = 75$, respectively.

Method	Q100			Q90			Q75			F1 Drop ↓	
	P	R	F1	P	R	F1	P	R	F1	Δ_{90}	Δ_{75}
DTD*	0.939	0.951	0.945	0.863	0.820	0.841	0.580	0.473	0.521	0.104	0.424
DTD* + TSPG	0.947	0.952	0.949	0.877	0.829	0.852	0.617	0.507	0.556	0.097	0.393
FFDN*	0.950	0.961	0.956	0.908	0.887	0.898	0.731	0.753	0.690	0.058	0.266
FFDN* + TSPG	0.957	0.963	0.960	0.916	0.905	0.911	0.805	0.767	0.786	0.049	0.174

Table 3: Generalization results on T-SROIE. “Cross-dataset” denotes training on DocTammer and direct evaluation on T-SROIE.

Setting	IoU	F1
Cross-dataset FFDN	0.1809	0.3064
Cross-dataset FFDN + TSPG	0.2726	0.4284
In-domain FFDN	0.8318	0.9082
In-domain FFDN + TSPG	0.8625	0.9262

precision, suggesting that OCR-grounded textual semantics help suppress false detections and recover subtle tampered text regions.

Figure 3 shows the intermediate maps and features of two examples. It shows that the learnable alignment refines the coarse box-level TSP into a smoother and more feature-compatible guidance map. Moreover, Figure 4 provides qualitative comparisons on the Test, FCD, and SCD sets. Compared with the original DTD and FFDN predictions, TSPG-enhanced models produce masks that better match the ground truth. In particular,

TSPG reduces scattered false responses in non-tampered text areas and recovers subtle tampered regions missed by the baselines. These visual results are consistent with the quantitative improvements and further demonstrate the usefulness of textual semantic priors for pixel-level document tampering localization.

4.3 Generalization to External Dataset

To examine whether the proposed semantic guidance generalizes beyond DocTammer, we further evaluate TSPG on T-SROIE under both cross-dataset and in-domain settings. The results are shown in Table 3.

Under cross-dataset evaluation, where models trained on DocTammer are directly tested on T-SROIE, TSPG improves the F1 score of FFDN from 0.3064 to 0.4284 and the IoU from 0.1809 to 0.2726. When trained directly on T-SROIE, TSPG further improves the F1 score from 0.9082 to 0.9262 and the IoU from 0.8318 to 0.8625. These results suggest that textual semantic guid-

Table 4: Robustness of TSPG under different OCR settings on DocTammer-Test.

Setting	F1
Vanilla FFDN	0.8302
TSPG w/ PP-OCRv5	0.8726
TSPG w/ docTR	0.8587
TSPG w/ Tesseract	0.8542
TSPG w/ 50% word dropping	0.8518

ance remains beneficial under both distribution shift and in-domain training.

4.4 Robustness Analysis

Robustness under JPEG Compression. We evaluate robustness on the DocTammer-Test set using fixed JPEG quality factors of $Q = 100, 90$, and 75 . JPEG compression can weaken low-level RGB-frequency artifacts, while the textual meaning of recognizable document content is relatively more stable. We also examine the OCR confidence under different compression qualities. The average OCR confidence remains stable, decreasing only from 0.957 at $Q = 100$ to 0.956 at $Q = 90$ and 0.954 at $Q = 75$, suggesting that textual cues remain reliable under the tested compression levels.

Table 2 shows that TSPG improves both DTD and FFDN under all compression qualities. When Q decreases from 100 to 75 , TSPG reduces the F1 drop of DTD from 0.424 to 0.393 . For FFDN, TSPG improves the F1 score from 0.690 to 0.786 under $Q = 75$, and reduces the F1 drop from 0.266 to 0.174 . These results show that textual semantic guidance can complement compression-sensitive forensic cues and improve robustness under JPEG degradation.

Robustness to OCR Variations. Since TSPG relies on OCR outputs to construct the textual semantic prior, we further investigate its robustness to different OCR systems and imperfect OCR results. As shown in Table 4, replacing PP-OCRv5 with docTR (Mindée, 2021) or Tesseract (Tesseract OCR, 2026) still consistently improves over the vanilla FFDN baseline. Moreover, TSPG retains clear improvements when OCR words are randomly dropped. These results indicate that the benefit of TSPG is not tied to a specific OCR engine and remains effective under imperfect OCR conditions.

Table 5: Ablation study of TSPG on DTD using the DocTammer-Test set.

Variant	IoU	P	R	F1
Original	0.827	0.811	0.766	0.788
Full TSPG	0.850	0.832	0.794	0.812
w/o Sem. Scores	0.830	0.815	0.770	0.792
w/o Prior Pyramid	0.823	0.813	0.766	0.789
w/o GSRM	0.829	0.816	0.765	0.790

Table 6: Ablation study of TSPG on FFDN using the DocTammer-Test set.

Variant	IoU	P	R	F1
Original	0.870	0.851	0.811	0.830
Full TSPG	0.901	0.882	0.864	0.873
w/o Sem. Scores	0.875	0.863	0.826	0.844
w/o Prior Pyramid	0.889	0.875	0.839	0.857
w/o GSRM	0.872	0.861	0.840	0.850

4.5 Ablation Study

Tables 5 and 6 ablate key components on the DocTammer-Test set to verify their effectiveness.

Effect of Tampering-Aware Semantic Scores. The variant *w/o Sem. Scores* here removes the mDeBERTa-based token scoring process and uses a binary OCR-region map as the prior. Although this variant improves over the original baselines, it is still inferior to the full TSPG, indicating that tampering-aware semantic scores provide more discriminative guidance than bbox-only priors. To further evaluate the discriminative ability of the token scoring branch, we visualize the ROC curve on the DocTammer-Test set. As shown in Figure 5, the token scores achieve an ROC-AUC of 0.876 for distinguishing high-risk and low-risk tokens, further supporting the effectiveness of the proposed token-level prior.

Effect of Prior Alignment. The variant *w/o Prior Pyramid* removes the learnable prior pyramid and directly resizes the prior map to different feature scales. Its performance is lower than the full TSPG on both DTD and FFDN, showing that learnable prior alignment helps refine the coarse box-level prior and adapt it to multi-scale detector features. Figure 3 also shows that the aligned map is smoother and more concentrated than the original prior map.

Effect of Gated Residual Modulation. The variant *w/o GSRM* replaces the proposed gated residual modulation with simple prior-based feature modulation. The performance drop demonstrates that directly multiplying features with the

Table 7: Ablation study of the token-level loss weight λ_{tok} on FFDN using the DocTamper-Test set.

λ_{tok}	IoU	P	R	F1
0.1	0.897	0.879	0.857	0.868
0.3	0.901	0.882	0.864	0.873
0.5	0.895	0.877	0.854	0.865

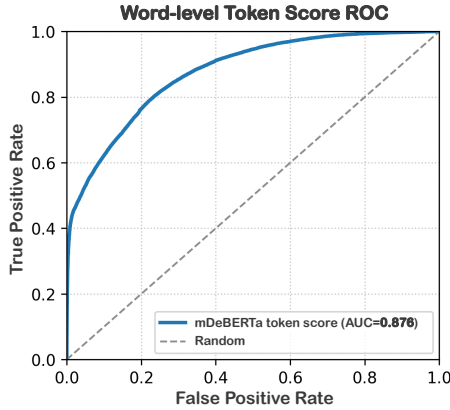


Figure 5: ROC curve of token-level tampering-aware scores on the DocTamper-Test set.

prior is less effective than adaptive residual gating. As shown in Figure 3, the residual and gate maps jointly highlight tampering-related regions and suppress irrelevant responses, enabling selective semantic guidance.

Overall, the full TSPG achieves the best results on both DTD and FFDN. It improves the F1 score from 0.788 to 0.812 on DTD and from 0.830 to 0.873 on FFDN, showing that semantic score estimation, multi-scale prior alignment, and gated residual modulation all contribute to performance. **Effect of λ_{tok} .** We further study the influence of the token-level loss weight λ_{tok} on FFDN. As shown in Table 7, TSPG achieves stable performance across different λ_{tok} values, with the best result at $\lambda_{tok} = 0.3$. This confirms that our method is not sensitive to the selection of λ_{tok} .

4.6 Efficiency Analysis

TSPG introduces only 0.95M additional trainable parameters to FFDN, increasing the number of trainable parameters from 140.30M to 141.25M. The larger total parameter count mainly comes from the frozen language model, which does not participate in parameter updates.

In terms of inference efficiency, vanilla FFDN requires 39.9 ms per image, while FFDN+TSPG requires 69.3 ms per image. When PP-OCRv5 is

Table 8: Inference efficiency of TSPG with FFDN. Parallel batch inference reports the effective per-image processing time rather than single-image end-to-end latency.

Setting	Time (ms/image)
FFDN	39.9
FFDN + TSPG	69.3
PP-OCRv5	45.9
Serial complete pipeline	115.2
Parallel batch inference	75.6

executed serially with FFDN+TSPG, the complete pipeline requires 115.2 ms per image, approximately $2.89\times$ the latency of vanilla FFDN. Since OCR processing is independent of the tampering localization network, it can be executed in parallel with FFDN+TSPG during batch inference. Under this setting, the effective per-image processing time is 75.6 ms, corresponding to approximately $1.89\times$ that of vanilla FFDN and only 6.3 ms more than FFDN+TSPG alone. Nevertheless, the complete pipeline still introduces additional computational and memory overhead.

5 Conclusion

We presented TSPG, a plug-and-play Textual Semantic Prior Guidance module for document image tampering localization. The core idea is to bridge token-level language modeling and pixel-level forensic localization by projecting tampering-aware semantic scores from OCR-recognized text back to the image space through OCR bounding boxes. The Textual Semantic Prior is then aligned with multi-scale detector features and injected through Gated Semantic Residual Modulation, enabling existing detectors to leverage textual semantic cues while preserving their original feature representations.

Experiments on DocTamper show that TSPG consistently improves both DTD and FFDN and that textual semantics provide complementary cues when low-level forensic traces are degraded. Additional evaluations on an external dataset, different OCR engines, and degraded OCR inputs demonstrate the generalizability and robustness of the proposed semantic guidance. These results suggest that incorporating OCR-grounded textual semantics is a promising direction for more accurate and robust document tampering localization.

Limitations

While TSPG provides an effective way to incorporate textual semantics into document tampering localization, it still has several limitations.

First, the proposed method relies on OCR results to construct the Textual Semantic Prior. If OCR fails to recognize tampered text, misses text regions, or produces inaccurate bounding boxes, the projected semantic prior may become incomplete or spatially misaligned, which can weaken its guidance effect. Although TSPG uses the prior as soft feature-level guidance rather than a hard mask, improving robustness to OCR errors remains an important direction for future work.

Second, the Textual Semantic Prior is an auxiliary cue rather than an independent detector. It provides high-risk textual regions to guide visual and frequency-domain features, but the final localization decision still relies on the underlying detector. Future work will explore more reliable prior estimation and tighter integration between textual semantics and visual forensic evidence.

Third, although TSPG introduces only a small number of additional trainable parameters, the complete pipeline still requires OCR and language-model inference. While parallel execution of OCR and tampering localization during batch inference can partially reduce the latency, the overall pipeline still incurs additional inference latency, computational cost, and memory overhead.

Acknowledgments

This work was supported by National Cyber Security-National Science and Technology Major Project (2025ZD1503700), the Beijing Natural Science Foundation under Grant No.L244043 and L2607011, and the Zhejiang Provincial Natural Science Foundation of China under Grant No.LD24F020007.

References

Romain Bertrand, Oriol Ramos Terrades, Petra Gomez-Krämer, Patrick Franco, and Jean-Marc Ogier. 2015. [A conditional random field model for font forgery detection](#). *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 576–580.

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*.

Francisco Cruz, Nicolas Sidère, Mickaël Coustaty, Vincent Poulain d’Andecy, and Jean-Marc Ogier. 2017. [Local binary patterns for document forgery detection](#). *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 01:1223–1228.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025. [Paddleocr 3.0 technical report](#). *Preprint*, arXiv:2507.05595.

Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. 2022. [Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection](#). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–14.

Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. 2023. [Tru-for: Leveraging all-round clues for trustworthy image forgery detection and localization](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20606–20615.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. [Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing](#). *Preprint*, arXiv:2111.09543.

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

Zhenjiang Li, Jingzhe Sun, and Shu Wang. 2025. [Semantic-based conflict detection: Tampering detection research in bilingual scene images containing textual content](#). *Symmetry*, 17:536.

Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. 2022a. [Pscn-net: Progressive spatio-channel correlation network for image manipulation detection and localization](#). *IEEE Transactions on Circuits and Systems for Video Technology*.

Zhuang Liu, Hanzi Mao, Chaozheng Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022b. [A convnet for the 2020s](#). *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11966–11976.

Dongliang Luo, Yuliang Liu, Rui Yang, Xianjin Liu, Jishen Zeng, Yu Zhou, and Xiang Bai. 2024. [Toward real text manipulation detection: New dataset and new solution](#). *Pattern Recognition*, page 110828.

- Mindee. 2021. doctr: Document text recognition. <https://github.com/mindee/doctr>.
- Chenfan Qu, Chenglong Liu, Yuliang Liu, Xinjian 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*, pages 5937–5946.
- 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*.
- Tesseract OCR. 2026. Tesseract open source ocr engine. <https://github.com/tesseract-ocr/tesseract>. Accessed: 2026.
- Beatriz Martínez Tornés, Emanuela Boros, Antoine Doucet, Petra Gomez-Krämer, and Jean-Marc Ogier. 2023. [Detecting forged receipts with domain-specific ontology-based entities & relations](#). In *IEEE International Conference on Document Analysis and Recognition*.
- Joost van Beusekom, Faisal Shafait, and Thomas M. Breuel. 2013. [Text-line examination for document forgery detection](#). *International Journal on Document Analysis and Recognition (IJDAR)*, 16:189–207.
- 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):29.
- Kahim Wong, Jicheng Zhou, Haiwei Wu, Yain-Whar Si, and Jiantao Zhou. 2025. Adcd-net: Robust document image forgery localization via adaptive dct feature and hierarchical content disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*.
- Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. 2018. Unified perceptual parsing for scene understanding. In *European Conference on Computer Vision*. Springer.
- Guitao Xu, Ziqi Yi, Peirong Zhang, Jiahuan Cao, Shih-Ting Wu, and Lianwen Jin. 2025. [From pixels to semantics: A novel mllm-driven approach for explainable tampered text detection](#). *Proceedings of the 33rd ACM International Conference on Multimedia*.
- Yanzhi Xu, Muhammad Irfan, Aiqing Fang, and Jiangbin Zheng. 2023. [Multiscale attention network for detection and localization of image splicing forgery](#). *IEEE Transactions on Instrumentation and Measurement*, 72:1–15.
- Shiqiang Zheng, Changsheng Chen, Shen Chen, Taiping Yao, Shouhong Ding, B. Li, and Jiwu Huang. 2026. [Generalized document tampering localization via color and semantic disentanglement](#). *IEEE Transactions on Circuits and Systems for Video Technology*, 36:1279–1292.