

ForensiText: ROI-Guided Multimodal Forensics for Scene Text Tampering Detection and Localization

1st Abhineet Kumar Pandey

Computer Science

University At Albany, SUNY

Albany, USA

apandey@albany.edu, 0009-0002-9449-3186

2nd Ming-Ching Chang

Computer Science

University at Albany, SUNY

Albany, USA

mchang2@albany.edu, 0000-0001-9325-5341

Abstract—Scene text manipulation is a subtle yet consequential form of visual tampering, where small edits to signs, receipts, labels, or prices can significantly alter image meaning while leaving minimal forensic traces. Existing forensic methods either analyze the full image, diluting evidence from small text regions, or adopt OCR-centric approaches that prioritize textual content over visual authenticity. This paper presents *ForensiText*, an ROI-guided forensic feature learning framework for scene text tampering detection and localization. Given an RGB image, a scene-text detector estimates text-bearing regions containing both pristine and manipulated text; ground-truth manipulation masks are used only for pixel-level supervision and evaluation. Within the detected text ROIs, the framework extracts complementary forensic cues from ROI-gated RGB appearance, SRM residuals, CFA inconsistencies, JPEG error-level analysis, and optional Noiseprint++ features. These streams are fused with an explicit text-ROI channel and processed by a compact dual-head encoder-decoder to jointly predict an image-level tampering score and a pixel-level manipulation mask. By concentrating forensic reasoning on automatically detected text regions, *ForensiText* preserves fine-grained localization while suppressing irrelevant background noise. Under a detector-guided text-ROI evaluation protocol, the method achieves strong pixel-level localization performance across multiple benchmarks, with Dice scores of 0.8684 on RealTextManipulation, 0.8364 on OSTF, 0.9843 on Tampered-IC13, and 0.8269 on TextSleuth. These results demonstrate the value of combining text-region priors with low-level forensic evidence for scene text manipulation detection.

I. INTRODUCTION

Text embedded in natural images often shapes human interpretation and decision making. Small edits to road signs, receipts, price tags, storefronts, product labels, document photographs, or public notices can substantially alter image meaning while affecting only localized regions. This makes scene text tampering an important visual forensics problem. Unlike most generic manipulations, text edits are often spatially compact, semantically consequential, and visually subtle, sometimes spanning only a few strokes, characters, or words.

Scene text tampering differs from conventional splicing or object-level manipulation because forensic evidence is concentrated in structured text regions, where stroke geometry, character boundaries, background texture, color, perspective, and compression artifacts interact at a fine spatial scale. Realistic edits must also preserve typographic consistency, including stroke width, alignment, spacing, color, and blending.

Consequently, full-image forensic models risk being distracted by irrelevant background content or missing fine-grained text-level artifacts, while OCR-centered systems recognize textual content without assessing visual authenticity.

Recent work has introduced specialized benchmarks and protocols for text-region manipulation detection and localization [1]–[5]. These studies suggest that scene-text forensics requires sensitivity to fine-grained traces from editing, compositing, resampling, demosaicing, and compression. As modern text editing and generative synthesis methods render manipulation cues increasingly subtle and localized, forensic analysis must concentrate on text-bearing regions while maintaining dense pixel-level localization.

In this paper, we propose *ForensiText*, an ROI-guided forensic cue fusion framework for scene text tampering detection and localization. Given an RGB image and a detected text-region ROI, the framework constructs complementary evidence streams from ROI-gated RGB appearance, SRM residuals, demosaicing/CFA consistency cues, JPEG error-level analysis, and an explicit ROI channel, with optional Noiseprint++ support. These cues are processed by a compact dual-head encoder-decoder to jointly predict an image-level tampering score and a dense manipulation mask. By using the ROI as both a feature-gating mechanism and a spatial prior, the model concentrates on text-relevant forensic evidence while suppressing unrelated background variation.

We evaluate *ForensiText* on multiple scene-text tampering benchmarks under a detector-guided text-ROI protocol. CRAFT [6] is used to estimate scene-text regions for both training-time feature construction and inference-time analysis. The manipulation mask is not used to construct the ROI; it serves only for pixel-level supervision and evaluation. Under this protocol, *ForensiText* achieves strong pixel-level localization and outperforms generic manipulation detection baselines on text-region tampering.

Contributions of this paper include the following:

- We formulate scene text tampering as a detector-guided forensic localization problem, where CRAFT-derived ROIs define candidate text regions and manipulation masks are used only for supervision and evaluation.
- We introduce a compact forensic cue pipeline that fuses ROI-gated RGB, SRM residuals, demosaicing/CFA cues,

JPEG ELA, and an explicit ROI channel, optionally augmented with Noiseprint++.

- We present a dual-head encoder-decoder that jointly predicts image-level tampering scores and pixel-level manipulation masks from the fused forensic representation.
- We evaluate ForensiText across multiple scene-text tampering benchmarks and analyze its dependence on text-detector quality and sensitivity to text-region manipulation.

II. RELATED WORK

Scene text tampering detection. Scene text tampering has recently emerged as a specialized problem in image forensics because text edits are often spatially small but semantically significant. Early work showed that edit targeting words, characters, and text-line regions cannot be treated simply as generic image splicing, since the forensic evidence is tightly coupled with text geometry, typography, local background, and semantic content [1]. TextSleuth [2] further emphasized the importance of text-region priors and introduced dedicated benchmarks for this task. More recent datasets and protocols have expanded the task toward realistic and generative-era text manipulation, where edited text is often visually well harmonized with the surrounding scene [3]–[5]. These works motivate models that explicitly reason over text-bearing regions rather than relying on full-image forensic cues alone.

Generic image manipulation localization. A broad line of image-forensics research has studied manipulation detection and localization using residual artifacts, compression inconsistencies, noise traces, and multi-scale feature fusion. Methods such as ManTra-Net, CAT-Net, MVSS-Net, TruFor, and MMFusion have demonstrated strong performance for general manipulation localization across splicing, copy-move, resampling, and compression-based manipulations [7]–[10]. However, these methods are typically designed for full-frame analysis. In scene text tampering, forensic evidence may occupy only a few strokes or characters, and irrelevant background content can dominate the forensic response. Our work differs by constraining forensic cue extraction to text-region support and fusing multiple low-level evidence streams within that region.

Forensic residual cues. Low-level residual features remain central to detecting subtle image edits. SRM-style high-pass filters emphasize local noise and high-frequency artifacts [11], Bayer/CFA inconsistency cues expose disruptions in demosaicing patterns [12], and JPEG error-level analysis highlights recompression mismatch from localized edits [13]. Noiseprint++ captures camera-related noise statistics and has proven effective in broader forensic settings [14]. Since no single cue is reliable under all acquisition and post-processing conditions, ForensiText combines complementary residual streams with an explicit ROI prior for text-focused tamper localization.

OCR, text detection, and text-centric forensics. Scene text detectors such as CRAFT provide useful spatial priors for identifying where text appears in an image [6]. However, OCR and text-recognition systems are designed to read text,

not to verify whether the observed text is visually authentic. A manipulated sign or receipt may be recognized correctly while still containing forensic inconsistencies. Related document-forensics methods also show that character appearance, font consistency, stroke texture, and layout regularity are informative for text manipulation analysis [15]–[17]. Unlike structured document images, scene text appears under perspective distortion, uneven illumination, occlusion, low resolution, and cluttered backgrounds. This motivates combining text-region localization with robust forensic residual analysis.

Position of this work. The proposed method lies at the intersection of scene text analysis and image forensics. Unlike OCR-centered approaches, it does not infer authenticity from recognized text alone. Unlike generic manipulation-localization networks, it does not treat the entire image as an unconstrained forensic target. Instead, ForensiText uses a text-region prior to guide the extraction of complementary forensic cues, including RGB appearance, SRM residuals, Bayer/CFA inconsistency, JPEG error-level analysis, and an explicit ROI channel. This design targets the small, structured, and semantically consequential manipulations characteristic of scene text tampering.

III. METHOD

A. Problem Formulation

Given an RGB image $I \in [0, 1]^{3 \times H \times W}$, the goal is to estimate both an image-level tampering label and a dense pixel-level manipulation mask. Let $y \in \{0, 1\}$ denote the binary image label, where $y = 1$ indicates a manipulated image, and let $M \in \{0, 1\}^{1 \times H \times W}$ denote the ground-truth tamper mask used for supervision. We use a scene-text ROI mask $T \in \{0, 1\}^{1 \times H \times W}$ produced by CRAFT [6]. ROI T identifies candidate text-bearing regions and may include both pristine and manipulated text. Importantly, T is not derived from the manipulation mask M .

ForensiText predicts a scalar image-level logit z and a dense mask-logit map $A \in \mathbb{R}^{1 \times H \times W}$. The corresponding probabilities are: $\hat{y} = \sigma(z)$, $\hat{M} = \sigma(A)$, where $\sigma(\cdot)$ denotes the sigmoid function. The CRAFT-derived text ROI guides feature construction and restricts pixel-level supervision to detected text regions.

B. ROI-Guided Image Preparation

The CRAFT text-confidence map or detected text boxes are converted into a binary scene-text ROI, resized to the network resolution, and binarized. When enabled, a small morphological dilation is applied to include neighboring stroke boundaries and local background context around the detected text. We denote the resulting binary ROI by T . To suppress irrelevant background variation while preserving local text evidence, the input image is converted into an ROI-gated image: $\tilde{I} = I \odot T + B(I) \odot (1 - T)$, where $\odot$ denotes element-wise multiplication and $B(I)$ is a background fill image. In our default setting, $B(I)$ is the image-wise mean RGB value

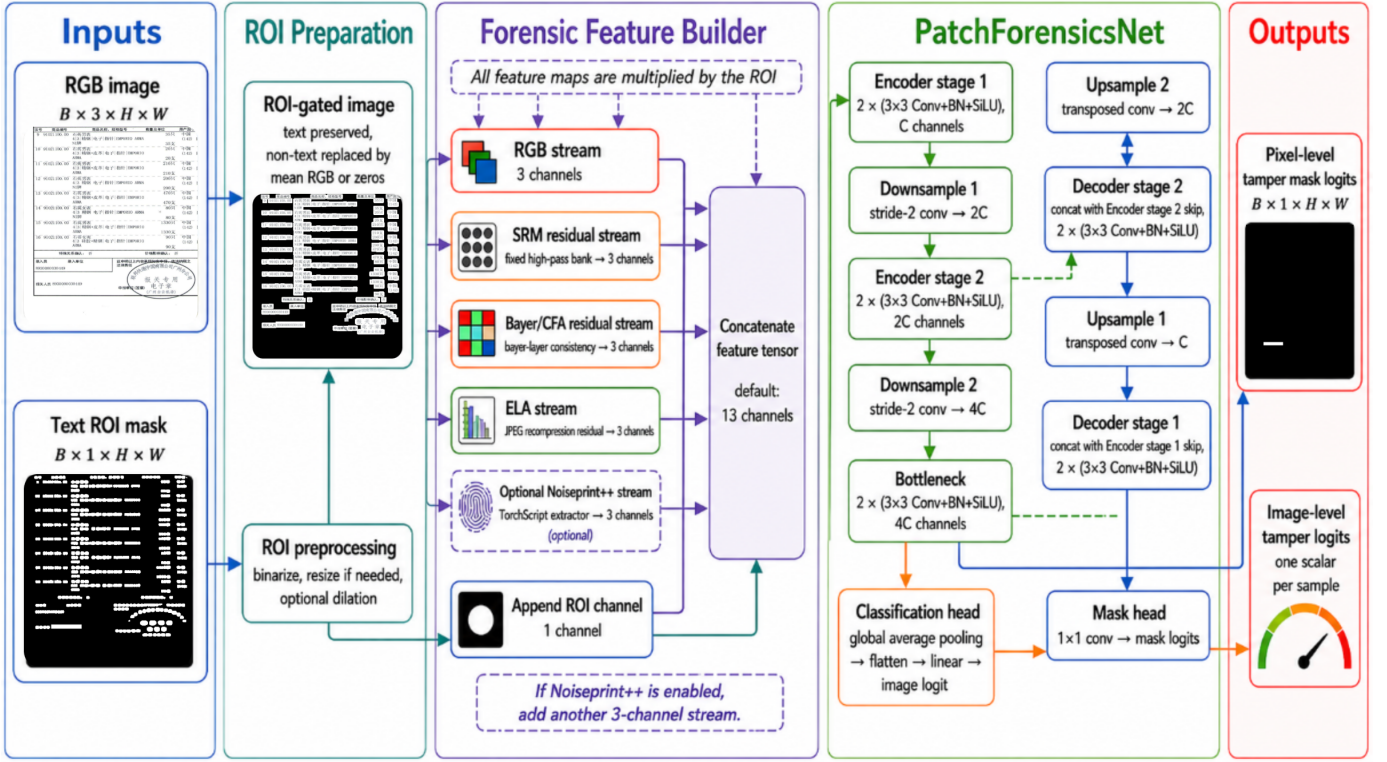


Fig. 1. Overview of ForensiText. Given an RGB image, CRAFT [6] produces a scene-text ROI mask covering candidate text regions; the ground-truth manipulation mask is used only for supervision. The feature builder constructs complementary forensic streams from the ROI-gated image: RGB appearance, SRM residuals, CFA inconsistency, ELA recompression artifacts, and an optional Noiseprint++ stream. These are concatenated with an explicit ROI channel to form a 13-channel input tensor. PatchForensicsNet then jointly predicts a pixel-level manipulation mask via an encoder-decoder branch and an image-level tampering score via a bottleneck classification head.

broadcast over the spatial dimensions. This gating keeps text-region pixels unchanged while replacing non-text pixels with a low-information background.

C. Forensic Feature Construction

Scene text edits may leave different types of forensic traces depending on the manipulation process, compression history, and image acquisition pipeline. We therefore construct a compact set of complementary forensic streams from the ROI-gated image $\tilde{I}$. The RGB stream preserves direct appearance evidence: $X_{\text{RGB}} = T \odot \tilde{I}$. The SRM stream applies fixed high-pass filters to emphasize local residual artifacts: $X_{\text{SRM}} = T \odot R_{\text{SRM}}(\tilde{I})$.

The Bayer/CFA stream captures demosaicing inconsistency by evaluating phase-dependent color-filter-array residuals: $X_{\text{Bayer}} = T \odot R_{\text{Bayer}}(\tilde{I})$. The ELA stream measures the reconstruction difference between the ROI-gated image and a JPEG-recompressed version:

$$R_{\text{ELA}}(\tilde{I}) = \left| \tilde{I} - J_q(\tilde{I}) \right|, \quad (1)$$

where $J_q(\cdot)$ denotes JPEG recompression at quality factor q . The corresponding ROI-restricted ELA feature is $X_{\text{ELA}} = T \odot R_{\text{ELA}}(\tilde{I})$.

When Noiseprint++ [14] features are enabled, an additional stream $X_{\text{NP}} = T \odot N(\tilde{I})$ is appended, where $N(\cdot)$ denotes the

Noiseprint++ feature extractor. The ROI itself is also appended as an explicit spatial prior, $X_{\text{ROI}} = T$.

The final feature tensor is formed by concatenating the active streams along the channel dimension:

$$X = \text{concat}(X_{\text{RGB}}, X_{\text{SRM}}, X_{\text{Bayer}}, X_{\text{ELA}}, X_{\text{ROI}}), \quad (2)$$

with X_{NP} appended when Noiseprint++ [14] is enabled. In the default configuration, the active streams are RGB, SRM, Bayer/CFA, ELA, and the ROI channel, yielding a 13-channel input tensor. For multi-channel streams, the one-channel ROI mask T is broadcast along the channel dimension before element-wise multiplication.

D. PatchForensicsNet

The fused feature tensor X is processed by a compact dual-head encoder-decoder network, denoted as **PatchForensicsNet**. The encoder extracts progressively lower-resolution forensic representations, while the decoder restores spatial detail for dense mask prediction. Let C denote the base channel width. The encoder is defined as: $e_1 = E_1(X)$, $e_2 = E_2(D_1(e_1))$, $e_3 = E_3(D_2(e_2))$, where E_1 , E_2 , and E_3 are two-layer Conv-BatchNorm-SiLU blocks, and D_1 and D_2 are stride-2 convolutional downsampling layers. The feature widths are C , $2C$, and $4C$, respectively.

The decoder reconstructs a dense localization representation using transpose-convolution upsampling and skip connections:

$$d_2 = G_2([U_2(e_3), e_2]), \quad d_1 = G_1([U_1(d_2), e_1]), \quad (3)$$

where U_2 and U_1 are transpose-convolution upsampling layers, G_2 and G_1 are Conv-BatchNorm-SiLU blocks, and $[\cdot, \cdot]$ denotes channel-wise concatenation. The skip connections preserve text-boundary and stroke-level detail that may be weakened by downsampling.

The localization head predicts a dense manipulation logit map from the final decoder representation: $A = h_{\text{mask}}(d_1)$, where h_{mask} is a 1×1 convolution. The image-level head pools the bottleneck representation and applies a linear classifier:

$$z = w_{\text{cls}}^\top \text{GAP}(e_3) + b_{\text{cls}}, \quad (4)$$

where $\text{GAP}(\cdot)$ denotes global average pooling. This dual-head design encourages the shared encoder to learn forensic representations useful for both image-level detection and pixel-level localization.

E. Training Objective

Pixel-level supervision is restricted to CRAFT-detected text regions. We define the ROI-restricted target and prediction as

$$M^* = M \odot T, \quad \hat{M}^* = \sigma(A) \odot T. \quad (5)$$

The mask loss is a weighted binary cross-entropy over ROI pixels:

$$\mathcal{L}_{\text{mask}} = \frac{\sum_{u,v} T_{uv} \ell_{\text{BCE}}(A_{uv}, M_{uv}^*)}{\sum_{u,v} T_{uv} + \epsilon}, \quad (6)$$

where ℓ_{BCE} denotes binary cross-entropy with logits and ϵ prevents division by zero.

To encourage spatial overlap, we also use an ROI-restricted Dice loss:

$$\mathcal{L}_{\text{dice}} = 1 - \frac{2 \sum_{u,v} \hat{M}_{uv}^* M_{uv}^* + \epsilon}{\sum_{u,v} \hat{M}_{uv}^* + \sum_{u,v} M_{uv}^* + \epsilon}. \quad (7)$$

The image-level prediction is supervised using binary cross-entropy with logits $\mathcal{L}_{\text{cls}} = \ell_{\text{BCE}}(z, y)$. The final objective combines localization, overlap, and classification losses:

$$\mathcal{L} = \lambda_{\text{mask}} \mathcal{L}_{\text{mask}} + \lambda_{\text{dice}} \mathcal{L}_{\text{dice}} + \lambda_{\text{cls}} \mathcal{L}_{\text{cls}}. \quad (8)$$

All three loss weights default to one unless otherwise specified.

F. ROI Acquisition

The proposed framework uses scene-text ROIs produced by an external text detector. In our implementation, CRAFT [6] is applied to each input image to obtain text confidence maps and word-level text boxes, which are converted into a binary ROI mask T by thresholding the text heatmap or rasterizing detected text boxes. A small dilation is optionally applied to include stroke boundaries and nearby background context.

The ROI T covers detected scene text regions and is independent of the manipulation mask; it may contain both pristine and manipulated text regions. During training, the ground-truth manipulation mask M is used only as pixel-level

TABLE I
SOURCE-LEVEL DATASET STATISTICS REPORTED BY THE ORIGINAL DATASET SOURCES.

Dataset	Total	Tampered	Authentic
RTM	9,000	6,000	3,000
OSTF	4,418	1,980	2,438
Tampered-IC13	462	378	84
TextSleuth-Total-Text	970	970	-

supervision inside the detected text ROI. During inference, the same CRAFT-based ROI extraction is applied, and *ForensiText* predicts manipulation probability for each pixel within the detected text regions.

For legacy pair-file datasets, CRAFT-derived text ROIs can be precomputed offline to ensure consistent ROI generation across training, validation, and test splits. This detector-guided protocol evaluates manipulation localization within automatically detected scene-text regions, rather than full-image manipulation discovery.

IV. EXPERIMENTAL EVALUATION

A. Data and Protocol

Table I reports source-level dataset statistics. Our experiments use a processed image-mask pair split containing samples with valid RGB images, pixel-level masks, compatible annotations, and CRAFT-derived text ROIs. The final split contains 10,915 training pairs, 1,587 validation pairs, and 1,587 held-out test pairs. This also includes unmodified Total-Text [18] to complement TextSleuth Authentic samples.

Each sample has an RGB image, a pixel-level manipulation mask, and an image-level authentic/manipulated label. Image labels are taken from explicit annotations when available and otherwise inferred from mask non-emptiness. CRAFT [6] generates a binary scene-text ROI T , which may include both pristine and manipulated text. The ROI serves for feature construction and ROI-restricted supervision. The ground-truth manipulation mask M is not used to construct T ; it serves only for training supervision and evaluation. The reported results therefore correspond to detector-guided text-region tamper localization rather than full-image manipulation discovery.

B. Implementation Details

All images and masks are resized to 512×512 . Patch-ForensicsNet uses base width $C = 32$ and a 13-channel input tensor formed from ROI-gated RGB, SRM residuals, demosaicing/CFA consistency cues, JPEG error-level analysis, and an explicit ROI channel. Noiseprint++ [14] is disabled unless otherwise specified.

CRAFT ROIs are binarized and dilated with a 9×9 kernel. Non-text pixels are replaced by the image-wise mean RGB value, and ELA features are computed with JPEG quality factor $q = 75$. The model is trained with AdamW for 250 epochs using learning rate 3×10^{-4} , weight decay 10^{-4} , batch size 60, and four data-loading workers. All loss weights are default to $\lambda_{\text{mask}} = \lambda_{\text{dice}} = \lambda_{\text{cls}} = 1$, and binary masks are obtained at a threshold of 0.5.

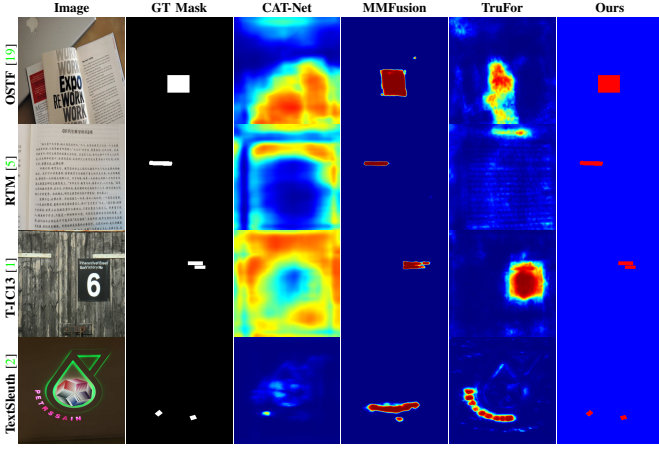


Fig. 2. Qualitative comparison under the detector-guided ROI protocol. Each row shows a representative example from one benchmark subset. ForensiText produces compact, text-focused responses within detected text regions, while generic full-image baselines frequently activate on irrelevant background content.

C. Evaluation Metrics

We report image-level accuracy, F1 score, and IoU for authentic-versus-manipulated classification. Pixel-level localization is evaluated by Dice and IoU between the predicted mask and ground-truth manipulation mask within the detected text ROI.

Let $P = \mathbf{1}[\sigma(A) > 0.5] \odot T$ be the ROI-restricted predicted mask and $G = M \odot T$ be the ROI-restricted ground-truth mask. Pixel metrics are computed as:

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G| + \epsilon}, \quad \text{IoU} = \frac{|P \cap G|}{|P \cup G| + \epsilon},$$

where ϵ prevents division by zero.

D. Results and Analysis

Table II reports held-out performance on the 1,587-sample final-pairs test split. We evaluate both image-level (manipulated vs. authentic) detection and pixel-level tamper localization. Since the default setting uses CRAFT-derived scene-text ROIs, the reported localization scores should be interpreted as detector-guided text-region localization rather than full-image tamper discovery. Overall, ForensiText performs strongly across all datasets. On RealTextManipulation, it achieves 0.8615 image accuracy, 0.8162 image F1, 0.8684 pixel Dice, and 0.7674 pixel IoU. On OSTF, image accuracy rises to 0.9249 and pixel IoU to 0.7928. The strongest result is obtained on Tampered-IC13, with 0.9693 image accuracy, 0.9632 image F1, 0.9843 pixel Dice, and 0.9610 pixel IoU. This subset is relatively small, so the results should be interpreted cautiously.

Compared with CAT-Net, MMFusion, and TruFor, ForensiText provides substantially stronger pixel-level localization on all evaluated subsets. This indicates that generic image-forensics models may detect suspicious images in some cases, but often struggle to localize small, structured scene-text edits. In contrast, the proposed ROI-guided formulation focuses the

model on text-relevant regions and reduces interference from unrelated background content.

The results also show that performance is not uniform across datasets. TextSleuth is the most challenging subset, where ForensiText obtains lower image-level F1 and pixel IoU while still maintaining reasonable pixel-level localization. This suggests that diverse text layouts, manipulation styles, and scene conditions remain difficult, particularly for image-level classification.

Qualitative results in Fig. 2 are consistent with the quantitative findings. ForensiText produces compact responses around manipulated text regions, whereas generic baselines often generate broader or less focused heatmaps. This supports forensic interpretation by providing localized evidence rather than only an image-level decision.

These results support the value of combining text-region priors with complementary forensic cues, including RGB appearance, SRM residuals, Bayer/CFA inconsistencies, and ELA features. Since the current protocol depends on CRAFT-derived text ROIs, the results should be viewed as evidence for text-guided forensic cue fusion rather than arbitrary full-image manipulation discovery. A practical deployment will depend on the robustness of the text detector, particularly for small, blurred, curved, or low-contrast scene text.

V. CONCLUSION

This paper presented *ForensiText*, an ROI-guided forensic cue fusion framework for scene text tampering detection and localization. By concentrating forensic analysis on automatically detected text regions, the framework combines ROI-gated RGB appearance, SRM residuals, CFA consistency cues, JPEG error-level analysis, and an explicit ROI channel within a compact dual-head encoder-decoder for joint image-level detection and pixel-level localization. Across RealTextManipulation, OSTF, Tampered-IC13, and TextSleuth, ForensiText achieves substantially stronger pixel-level localization than generic manipulation-localization baselines under the detector-guided text-ROI protocol, demonstrating the effectiveness of combining text-region priors with complementary low-level forensic cues. TextSleuth remains the most challenging subset, particularly at the image level, reflecting greater variation in text layout, scene complexity, and manipulation style.

Limitations: The reported results should be interpreted within the detector-guided setting: localization is evaluated within CRAFT-derived text regions rather than as full-image tamper discovery. Performance therefore depends on scene-text ROI quality, and missed or inaccurate detections can degrade localization, particularly for small, blurred, curved, occluded, or low-contrast text. The contribution of each individual forensic stream has also not been fully quantified.

Future work: Immediate priorities include systematic ablations of the RGB, SRM, CFA, ELA, ROI, and Noiseprint++ streams, and evaluation of stronger text detectors such as DBNet with alternative ROI-generation strategies. Broader evaluation across diverse compression conditions, text layouts,

TABLE II

PER-DATASET COMPARISON ON THE SCENE-TEXT TAMPERING BENCHMARKS (TABLE I). * DENOTES RESULTS FROM THE AUTHORS' BEST RELEASED CHECKPOINT; ALL OTHERS ARE FROM MODELS RETRAINED ON OUR DATASET SPLIT. IMAGE-LEVEL METRICS MEASURE AUTHENTIC-VERSUS-MANIPULATED CLASSIFICATION; PIXEL-LEVEL METRICS MEASURE MANIPULATION MASK LOCALIZATION.

Method	Dataset	Samples	Manip.	Clean	Img. Acc. $\uparrow$	Img. F1 $\uparrow$	Img. IoU $\uparrow$	Pix. Dice $\uparrow$	Pix. IoU $\uparrow$
ForensiText (Our)	RealTextManipulation	898	594	304	0.8615	0.8162	0.6895	0.8684	0.7674
ForensiText (Our)	OSTF	413	192	221	0.9249	0.8772	0.7813	0.8364	0.7928
ForensiText (Our)	Tampered-IC13	54	41	13	0.9693	0.9632	0.9293	0.9843	0.9610
ForensiText (Our)	TextSleuth	222	104	118	0.7748	0.5881	0.4165	0.8269	0.7049
CAT-Net	RealTextManipulation	898	594	304	0.6448	0.7831	0.6436	0.0211	0.0107
CAT-Net	OSTF	413	192	221	0.4939	0.6261	0.4557	0.1175	0.0624
CAT-Net	Tampered-IC13	54	41	13	0.7407	0.7444	0.5929	0.1254	0.0669
CAT-Net	TextSleuth	222	104	118	0.5315	0.4157	0.2624	0.0157	0.0079
CAT-Net*	RealTextManipulation	898	594	304	0.4310	0.4581	0.2971	0.0192	0.0097
CAT-Net*	OSTF	413	192	221	0.4818	0.3631	0.2218	0.0339	0.0172
CAT-Net*	Tampered-IC13	54	41	13	0.5926	0.6562	0.4884	0.1480	0.0799
CAT-Net*	TextSleuth	222	104	118	0.4550	0.6254	0.4550	0.0262	0.0133
MMFusion	RealTextManipulation	898	594	304	0.6693	0.7992	0.6655	0.2251	0.1268
MMFusion	OSTF	413	192	221	0.6223	0.7079	0.5478	0.8099	0.6806
MMFusion	Tampered-IC13	54	41	13	0.7889	0.7318	0.5770	0.7885	0.7594
MMFusion	TextSleuth	222	104	118	0.5631	0.6367	0.4670	0.3985	0.2488
MMFusion*	RealTextManipulation	898	594	304	0.6537	0.7889	0.6513	0.0696	0.0360
MMFusion*	OSTF	413	192	221	0.4818	0.6373	0.4677	0.4025	0.2519
MMFusion*	Tampered-IC13	54	41	13	0.7593	0.7602	0.6132	0.6417	0.4725
MMFusion*	TextSleuth	222	104	118	0.5090	0.6472	0.4785	0.1842	0.1014
TruFor	RealTextManipulation	898	594	304	0.484	0.439	0.281	0.070	0.036
TruFor	OSTF	413	192	221	0.668	0.593	0.422	0.310	0.184
TruFor	Tampered-IC13	54	41	13	0.704	0.771	0.628	0.651	0.483
TruFor	TextSleuth	222	104	118	0.5315	0.2464	0.1405	0.1427	0.0768
TruFor*	RealTextManipulation	898	594	304	0.452	0.328	0.196	0.069	0.036
TruFor*	OSTF	413	192	221	0.770	0.731	0.576	0.362	0.221
TruFor*	Tampered-IC13	54	41	13	0.633	0.677	0.5117	0.517	0.349
TruFor*	TextSleuth	222	104	118	0.5766	0.3088	0.1826	0.2916	0.1707

and modern generative text-editing methods will further clarify the robustness and generalization of the proposed approach.

REFERENCES

- [1] Y. Wang, H. Xie, M. Xing, J. Wang, S. Zhu, and Y. Zhang, "Detecting tampered scene text in the wild," in *ECCV*. Springer, 2022, pp. 215–232.
- [2] A. K. Pandey, M.-C. Chang, and X. Li, "Textsleuth: A new dataset and baseline for scene text manipulation detection," in *MIPR*, 2024.
- [3] C. Qu, Y. Zhong, F. Guo, and L. Jin, "Generalized tampered scene text detection in the era of generative ai," 2024.
- [4] D. Luo, Y. Liu, R. Yang, X. Liu, J. Zeng, Y. Zhou, and X. Bai, "Toward real text manipulation detection: New dataset and new solution," *Pattern Recognition*, vol. 157, p. 110828, 2025.
- [5] J. Duan, H. Sun, F. Ji, K. Zhou, Z. Wang, H. Huang, and L. Jin, "Realdtt: Towards a comprehensive real-world dataset for tampered text detection," *IJCV*, vol. 133, no. 10, pp. 6993–7011, 2025.
- [6] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," in *CVPR*, 2019, pp. 9365–9374.
- [7] Y. Wu, W. AbdAlmageed, and P. Natarajan, "Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features," in *Proceedings of the IEEE/CVF CVPR*, 2019, pp. 9543–9552.
- [8] M.-J. Kwon, S.-H. Nam, I.-J. Yu, H.-K. Lee, and C. Kim, "Cat-net: Compression artifact tracing network for detection and localization of image splicing," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 375–384.
- [9] X. Chen, C. Dong, J. Ji, J. Cao, and X. Li, "Image manipulation detection by multi-view multi-scale supervision," in *Proceedings of the IEEE/CVF ICCV*, 2021, pp. 14 185–14 193.
- [10] K. Triaridis and V. Mezaris, "Exploring multi-modal fusion for image manipulation detection and localization," in *International Conference on Multimedia Modeling*. Springer, 2024, pp. 269–281.
- [11] J. J. Fridrich and J. Kodovský, "Rich models for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 7, pp. 868–882, 2012.
- [12] B. Bayar and M. C. Stamm, "Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 11, pp. 2691–2706, 2018.
- [13] N. Krawetz, "A picture's worth... digital," 2007.
- [14] D. Cozzolino and L. Verdoliva, "Noiseprint: A cnn-based camera model fingerprint," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 144–159, 2018.
- [15] C. Qu, C. Liu, Y. Liu, X. Chen, D. Peng, F. Guo, and L. Jin, "Towards robust tampered text detection in document image: New dataset and new solution," in *Proceedings of the IEEE/CVF*, 2023, pp. 5937–5946.
- [16] X. Liao, S. Chen, J. Chen, T. Wang, and X. Li, "Ctp-net: Character texture perception network for document image forgery localization," 2023.
- [17] Z. Chen, S. Chen, T. Yao, K. Sun, S. Ding, X. Lin, L. Cao, and R. Ji, "Enhancing tampered text detection through frequency feature fusion and decomposition," in *ECCV*. Springer, 2024, pp. 200–217.
- [18] C. K. Ch'ng, C. S. Chan, and C. Liu, "Total-text: Towards orientation robustness in scene text detection," (*IJDAR*), vol. 23, pp. 31–52, 2020.
- [19] C. Qu, Y. Zhong, F. Guo, and L. Jin, "Revisiting tampered scene text detection in the era of generative ai," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 1, 2025, pp. 694–702.