TextSleuth: A New Dataset and Baseline for Scene Text Manipulation Detection

Abhineet Kumar Pandey Ming-Ching Chang Xin Li
Department of Computer Science, University at Albany, State University of New York, NY, USA
{apandey, mchang2, xli48}@albany.edu

***Abstract*—With the rise of digital content on social media and the advancement of image editing tools, tampering with scene text has become a serious concern. Scene text manipulation detection (STMD) is a kind of image manipulation detection (IMD) with focus on the tampering of scene text pixels, which is crucial for image content integrity and media forensics. In this paper, we present TextSleuth, a novel benchmark dataset specifically designed for STMD, by integrating three public datasets with newly introduced manipulation and annotations. We introduce professional edits on the Total-Text dataset (∼1K images) with four levels of manipulated region perceptibility, and a large synthetic manipulation set (858K images) on the SynthText dataset, as well the integration of the Tampered-IC13 dataset (378 images). We established a new STMD baseline based on TextSleuth using MMFusion-IML, the state-of-the-art image manipulation detection model. We performed extensive experiments, reporting the AUC from ROC analysis and the balanced accuracy (bACC) metrics to maintain a balanced performance evaluation. The MMFusion-IML baseline achieves 0.641 AUC and 0.588 bACC on the Total-Text subset. In comparison, it achieves 0.89 AUC and 0.8272 bACC on the Tampered-IC13 subset. This showcases the real-world STMD challenges reflected in our new dataset. TextSleuth is a valuable resource for future research in scene text manipulation detection and forensics. The dataset is available at https://github.com/abhineet-pandey/Text-Sleuth.**

***Index Terms*—Scene text, image manipulation detection, media forensics, MMFusion, Total-Text, SynthText, Tampered-IC13.**

I. Introduction

In an era of digital contents, the authenticity of visual information is increasingly susceptible to manipulation [1], [2]. Scene text, often found in photos of documents and real-world images, is particularly vulnerable to tampering. The ease of altering text pixels with advanced editing tools poses a significant threat media content trustworthiness [3]. Detecting scene text forgery is crucial, as manipulated text can spread misinformation, lead to legal disputes, and compromise personal and institutional integrity. Ensuring the reliability of scene text is essential for maintaining the credibility of digital information in various critical domains.

Scene text manipulation detection (STMD) is a specialized task of image manipulation detection (IMD) focusing on identifying tampered texts within images. STMD has not been extensively investigated, mainly due to a lack of adequate data and the inherent challenges. While notable research exists in the broader field of image-based tampering detection [4], IMD methods [5]–[9] typically targets semantic objects like spliced people, face swaps, or other generic objects, supported by benchmark datasets [10]–[13]. In contrast, STMD requires deeper semantic and contextual understanding, with manipulation areas being relatively smaller and less diverse. Current state-of-the-art IMD methods [6], [8], [9] excel in detecting general image manipulations but often fail to accurately identify and localize altered text within images. This gap highlights the need for targeted research and specialized datasets to address the unique challenges posed by scene text forgery and manipulation detection.

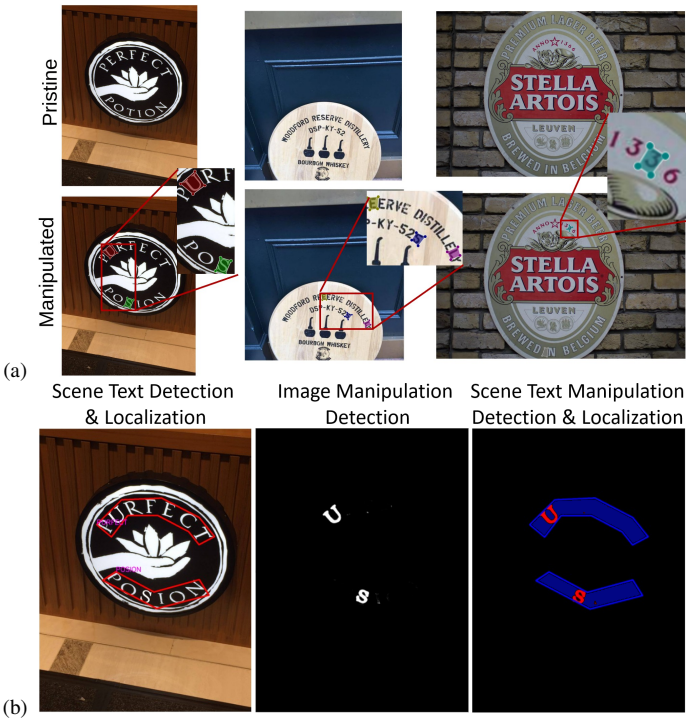

Fig. 1. (a) Example scene text manipulation and groundtruth from the proposed **TextSleuth** dataset. The manipulation region is typically tiny *w.r.t.* the whole image. (b) We provide a STMD baseline by intersecting the scene text detection and image manipulation detection masks.

To address the gaps in existing solutions, we introduce **TextSleuth**, an innovative dataset specifically designed for detecting scene text manipulations. Fig. 1 shows examples of the TextSleuth dataset and a STMD baseline. We incorporate and integrate three public STMD datasets, with newly introduced scene text manipulation and annotations. (1) We enhance the Total-Text dataset [14] by incorporating professional image-based edits on selected scene texts, creating realistic scene

text manipulations with ground truth. This subset includes approximately 1,000 images, with about 1,151 manipulation instances. To reflect real-world STMD scenarios, the manipulations are categorized into four difficulty levels, ranging from subtle alterations to drastic changes based on human visual perception. (2) We extract an extensive set of synthetic text manipulations from the SynthText dataset [15], resulting in 858,000 high-quality scene text manipulation images with ground truth. (3) We incorporate the Tampered-IC13 dataset [16] from ICDAR 2013, which includes synthetically manipulated images annotated with bounding boxes for each manipulated instance. The existing annotations are enriched with manipulated word-level masks as ground truth.

In addition to constructing a benchmark dataset, we establish a baseline for STMD by integrating state-of-the-art IMD methods with scene text detection and localization models. Specifically, we combine the MMFusion IMD [6] with the CRAFT [17] scene text detection model. Model training and evaluation is performed following standard train-test split on TextSleuth. By intersecting the STMD and scene text detection masks as shown in Fig. 1(b), we achieve accurate detection and localization of scene text forgery.

We employ the standard ROC AUC analysis for performance evaluation, measuring the performance STMD in distinguishing between manipulated and original images, independent of threshold settings. We also used balanced accuracy (bACC) to maintain a fair measure of classification accuracy, accounting for both *sensitivity* (True Positive Rate, TPR) and *specificity* (True Negative Rate, TNR), addressing imbalance between pristine and manipulated images.

The new TextSleuth STMD dataset offers a comprehensive training set and benchmark for the research community, featuring realistic and diverse scene text manipulation cases that closely resemble real-world scenarios. This work advances the study of STMD by providing targeted resources to address the unique challenges of scene text forgery, ultimately enhancing the reliability and trustworthiness of visual information. Our contributions are summarized as follows:

- *Development of a new high-quality dataset:* We created a new tampered version of the Total-Text dataset [14], featuring meticulously created manipulated scene text images with tampering ground truth masks done by three professional editors, covering four difficulty levels and manipulation types. Additionally, masks were extracted the large SynthText dataset [15] and the Tampered-IC13 [16] dataset to improve scene text manipulation detection capabilities.
- *Baseline establishment:* We introduced a baseline method that combines CRAFT scene text detection with MMFusion image manipulation detection, enhancing the detection of manipulated scene text.
- *Evaluation benchmark:* We established an evaluation benchmark for scene text manipulation detection, utilizing metrics such as ROC AUC and balanced accuracy (bACC) analysis. Our baseline achieves 0.641 AUC and 0.588 bACC on the Total-Text subset. It achieves 0.89 AUC and 0.8272 bACC on the Tampered-IC13 subset. On the SynthText subset, a score of 0.966 for AUC and 0.918 bACC was achieved.

II. Related Work

A. Image Manipulation Detection

Image manipulation detection (IMD) techniques have advanced significantly, with several state-of-the-art methods developed to identify and localize tampered regions in images. Below we survey some prominent IMD algorithms.

CAT-Net [9] uses fully convolutional neural network combined with attention mechanisms to learn the forensics features of compression artifacts on RGB and DCT domains. Multiple resolutions of the each stream is considered to deal with artifacts brought by the spliced object.

ManTra-Net [5] employs a fully convolutional network for a self-supervised learning task aimed at comprehensively learning 385 types of manipulations. Manipulation localization is treated as anomaly detection, handled via a LSTM-based approach to assess local anomalies using a z-score feature.

MVSS-Net [7] introduces a multi-view approach to capture subtle inconsistencies in the structural and textural properties of an image. Multi-view feature learning is designed to extract semantic-agnostic which yields more generalizable features. This method also learns form authentic images rather than just forged images. By utilizing multiple views, MVSS-Net enhances its ability to detect fine-grained manipulations that are often imperceptible to the human eye. This method excels in scenarios where high precision is required for identifying small, localized tampering.

The MMFusion-IML [6] combines information from various forensic artifacts and traces produced by noise-sensitive filters such as SRM [18], Bayar convolution [19] and Noiseprint++ [20]. This fusion-based approach leverages multiple modalities to improve detection accuracy, making it robust against a wide range of manipulation techniques. Two distinct approaches are proposed for combining the outputs of different forensic filters for image manipulation localization and detection.

TrueFor [8] integrates several forensic features to provide a robust solution for detecting and localizing manipulated areas in images, aiming for high accuracy and reliability.

Image manipulation detection datasets: There are several datasets developed for image manipulation detection, including CASIA-v2 [10], Coverage [11], Columbia [21], DSO-1 [22], VIPP [23], NIST16 [24], and OpenForensics [25], among others. A recent IMD benchmark dataset is developed in [4]. These datasets primarily focus on manipulations at the object level and include techniques such as copy-move, splicing, removal, and enhancement.

B. Scene Text Image Tampering Datasets

While numerous IMD datasets have been reviewed in § II-A, there are only a few datasets focusing specifically on scene text tampering. Although some publicly available and mostly private datasets for document forgery detection exist, such as those in [26]–[30], these forgeries are typically created through random copy-move, splicing, or content synthesis

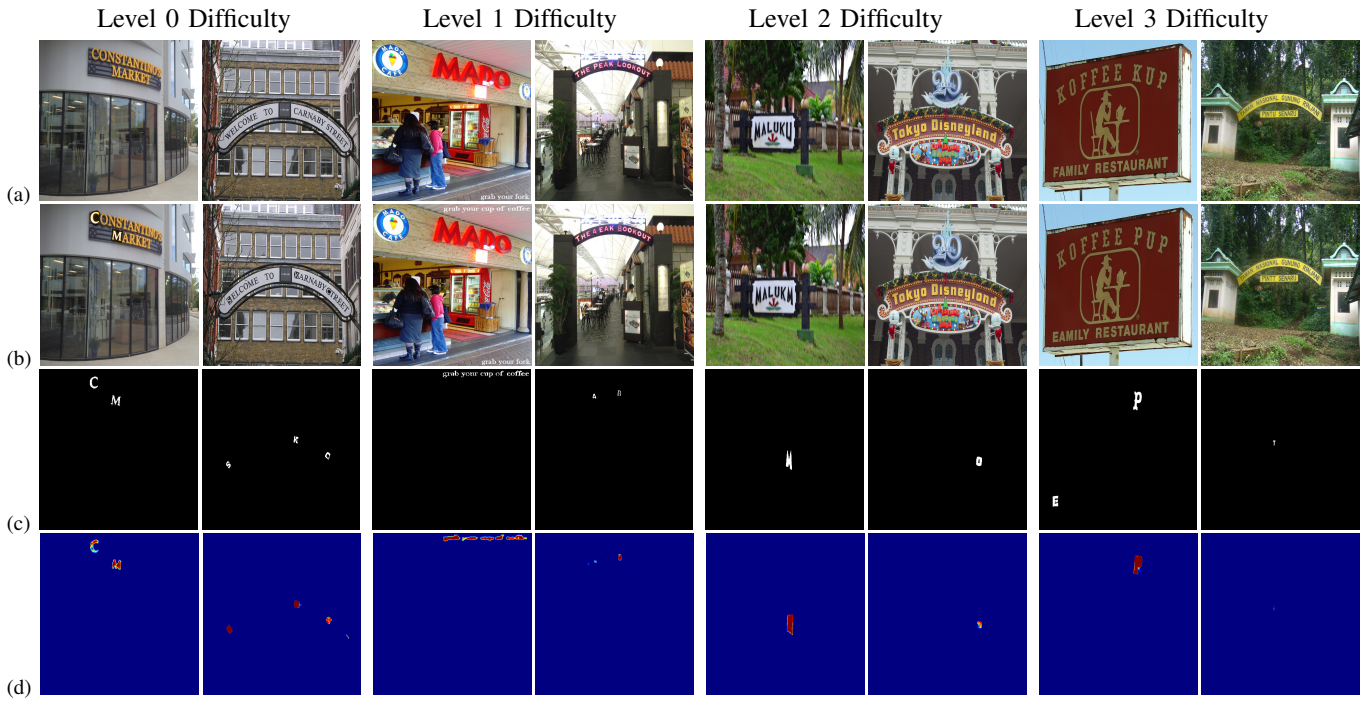

Fig. 2. Example of tampering images in our TextSleuth scene text manipulation detection (STMD) dataset: (a) original real images, (b) scene text tampered images, (c) the manipulation groundtruth mask, (d) detected manipulated region.

techniques. These datasets do not reflect real-world scene text forgery scenarios, which are addressed in this paper. It is important to note that document and scene text tampering differ significantly, with each presenting unique challenges and requiring specialized approaches.

Tampered-IC13 [16] is an extension of ICDAR 2013 dataset for addressing scene text image tampering. The manipulations are generated algorithmically using class-specific texture learning. A parallel-branch feature extractor captures both RGB and frequency domain features. The dataset contains high-quality tampered textures, featuring text-free areas and smooth transitions in the contour regions. However, GAN-based text synthesis produces unnecessary blurry artifacts in the samples.

C. Scene Text Detection

There exists extensive works on scene text detection, including CRAFT [17], DB [31], EAST [32], TextFuseNet [33], CharNet [34]. CRAFT [17] is effectively in detecting text area by exploring each character with region awareness and modeling the affinity between characters, without the need of individual character level annotations. Differentiable Binarization (DB) [31] achieves realtime performance by integrating differentiable binarization and an adaptive scale fusion technique. TextFuseNet [33] uses multiple feature maps at different levels of abstraction to enhance the representation of text regions in images. The integration of CNN features from both shallow and deep layers captures richer spatial and contextual information for accurately detecting text in complex scenes.

III. The TextSleuth Benchmark Dataset

The proposed TextSleuth dataset integrates scene text images from three publicly available datasets: Total-Text [14], SynthText [15], and Tampered-IC13 [16]. We introduce new types of scene text editing and manipulation into images of these parent datasets, with ground truth annotations. Unlike most scene text manipulation datasets that generate manipulations by simply rendering text pixels on top of scenes, resulting in visually fake and easily identifiable text, our approach involves professional scene text insertion and editing on the Total-Text dataset, performed by three experts skilled in image and media editing. Additionally, we incorporate the extensive synthetic dataset from SynthText. Lastly, we include the Tampered-IC13 dataset, which comes with its own scene text manipulation and ground truth. Table I provides the properties and features of the TextSleuth dataset across these three parent datasets.

TABLE I
TextSleuth Dataset Compilation Breakdown.

Features Details	Parent Datasets		
	Total-Text	*Tampered-IC13*	*SynthText*
# Tampered images	970	378	858,749
# Instances	1,151	995	6M
Groundtruth Mask†	✓	✓ (ours)	✓ (ours)
Groundtruth Polygon	✓	✓	✓
Tampering Style	Manual	Synthetic	Synthetic
Extra Features‡	✓	-	-

†Ground truth mask for the manipulated region. ‡See § III-A for explanation.

In the following subsections, we describe the process of generating or incorporating scene text manipulations and curating ground truth annotations for these three parent datasets.

A. The Tampered Total-Text Subset

Finding authentic, non-synthetic tampered datasets for scene text manipulation poses a significant challenge. To address

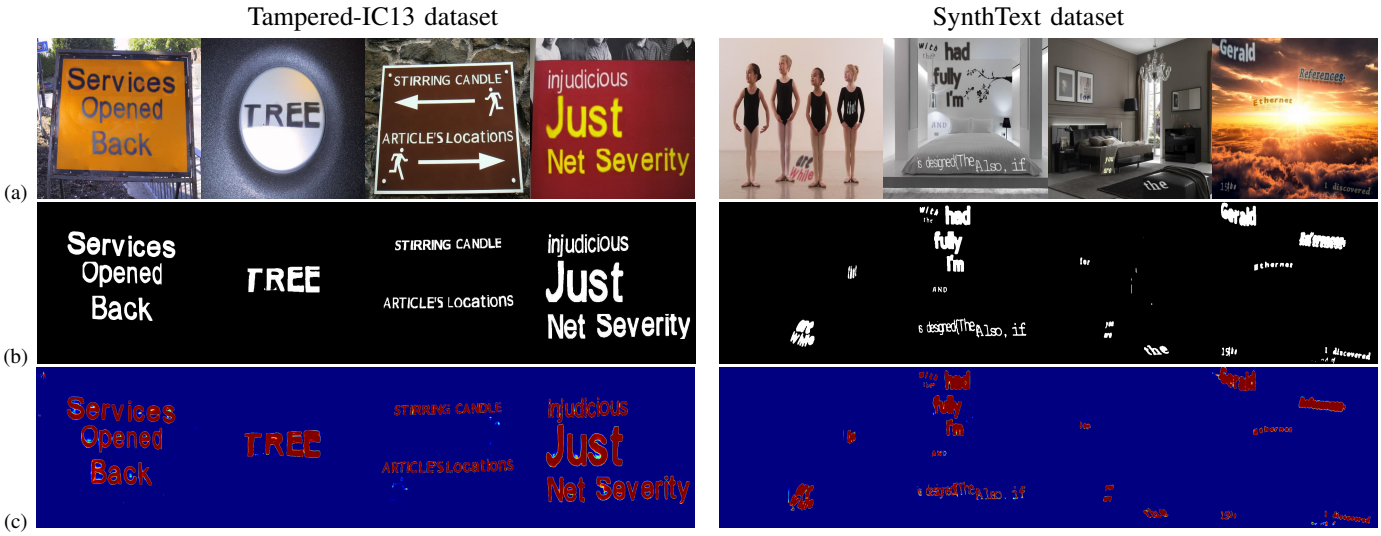

Fig. 3. Results on Tampered-IC13 and SynthText images: (a) the tampered images, (b) ground truth mask, (c) the predicted mask.

this gap, we emphasize the importance of manually tampered datasets that feature genuine examples of image manipulation. These datasets demonstrate a wide array of tampering scenarios and techniques, including subtle alterations that are challenging to detect. This diversity enhances the capability of detection algorithms to handle real-world manipulation techniques effectively. Accurate ground truth annotations are crucial in manually tampered datasets, specifying which parts of images are tampered and how. Models trained on such datasets are better equipped to generalize to unseen manipulation techniques and variations in real-world images. In contrast, synthetic datasets such as the SynthText dataset often lack the complexity and variability of real-world manipulation, which can limit their effectiveness in training models for authentic image scenarios.

To address this need, we developed a manually manipulated scene text dataset. Three professional annotators performed tampering on the Total-Text dataset [14] at the character level. We altered a total of 970 images, creating 1,151 instances of manipulation. Each modification is carefully documented with precise ground truth annotations. Sample images illustrating these changes are shown in Fig. 2.

The process of generating this **Tampered Total-Text Subset** was entirely manual to ensure unbiased data with no specific patterns, thereby maintaining data quality. Annotators had the flexibility to modify, add, or remove words in scene text using various tools such as Adobe Photoshop, Illustrator, Paint, and mobile devices. Each annotator worked independently, employing techniques such as text removal, addition, modification, replacement, distortion, font and size changes, as well as copy-paste and splicing.

Our manipulations and annotations provide diverse insights into different forms of text manipulation, significantly enhancing the dataset's robustness and applicability for training and evaluating scene text manipulation detection algorithms.

Ground truth types: The dataset includes detailed annotations for each modified word or character, including the location if a new character or word is added. Each modification is outlined with polygons, accompanied by modification masks. Additionally, the dataset includes information on difficulty levels (explained below), modification type (color, font, content), specific font used, software version, and the modifier's identity. Polygon annotations are available for each manipulated text along with its content, and masks are provided for each manipulated character.

Difficulty levels: We categorized the quality of manipulation or tampering on a scale from 0 to 3, indicating the difficulty in visually identifying the manipulation. Five independent assessors participated in this classification process, each assigning difficulty levels based on their assessment. The final rating for each instance was determined by reaching a majority consensus among the assessors. Examples of scene text tampered images categorized by these levels are shown in Fig. 2: row (a) shows the pristine image, row (b) displays all corresponding modified images, and row (c) presents the mask for the modified text region. The four difficulty levels are defined as follows:

- *Level 0*: Easily detectable by human eyes; typically noticeable at first glance.
- *Level 1*: Detectable by human eyes, but requires some effort beyond a cursory look.
- *Level 2*: Difficult to detect by human eyes.
- *Level 3*: Nearly impossible to detect by human eyes; these manipulations are executed with high precision and time.

The dataset includes other details about the font and modification type, along with information on the software used for modification and the individual who made the modifications. This aid in the identification of their signature style of the modification.

B. The SynthText Subset

Generating synthetic data for scene text is challenging due to the limited area available for image modification. Overcoming these challenges requires the development of

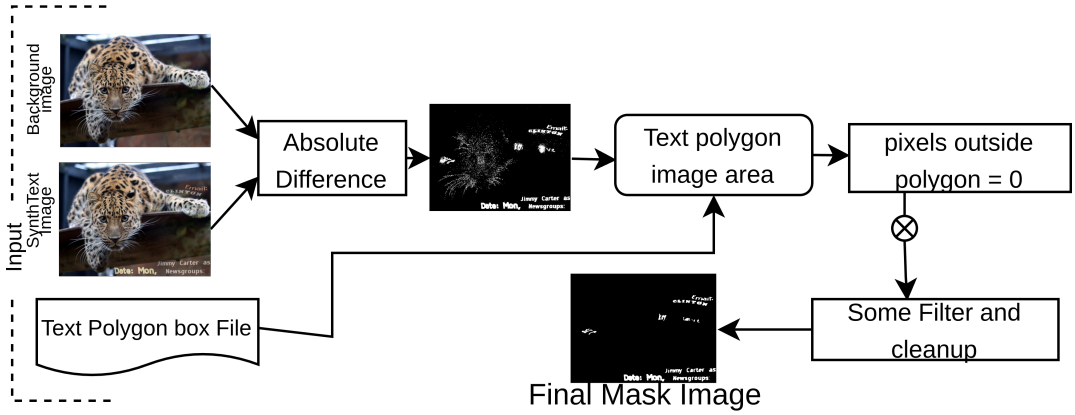

Fig. 4. Flow diagram illustrating the extraction of text masks from the SynthText dataset.

effective tools and the application of creative techniques to ensure high-quality synthetic data generation. One approach is to leverage existing datasets like SynthText [15], which provides rendered text images and masks. By saving the rendered text masks generated in each step, we can effectively generate new synthetic data. This method allows flexibility in generating scene text in multiple languages, thereby expanding the applicability of the dataset.

The SynthText dataset comprises approximately 858,749 synthetic images, containing a total of 6 million words and approximately 29 million characters. Sample images and their corresponding masks are illustrated in Fig. 3 (a,b).

Extracting scene text tampering masks: The SynthText dataset involves rendering scene text onto natural images using various fonts, colors, and orientations to mimic real-world scenarios. To extract the groundtruth tampering masks, traditional background subtraction methods struggle to accurately extract masks for the rendered text due to noise, artifacts, complex backgrounds, shadows, and variations in text characteristics such as font, size, color, orientation, and lighting conditions. To address these challenges and obtain precise masks for the rendered text, we developed a sequential approach illustrated in Fig. 4. Initially, we compute the absolute difference between the background and the rendered text image, which often results in significant noise and artifacts. To refine the masks, we utilize the provided text polygon ground truth to mask out non-text areas. Subsequently, we apply a median-blur filter with a kernel size of 3 to further enhance the text polygons and eliminate residual noise or artifacts. This systematic process consistently delivers accurate masks for each image in the SynthText dataset.

C. The Tampered-IC13 Subset

The dataset comprises 378 images featuring tampering, with a total of 995 instances of manipulation, each accompanied by its corresponding bounding box. To produce scene text tampering groundtruth masks for these images, we employed scene text segmentation as outlined in [35]. This involved initially extracting all potential scene text from the image and then removing any pristine text to produce masks that specifically highlight the tampered text regions.

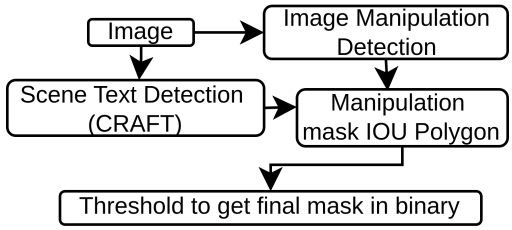

Fig. 5. The proposed scene text forgery detection pipeline.

IV. A STMD Baseline and Experimental Evaluation

In § IV-A, we develop a scene text manipulation detection (STMD) baseline by integrating a popular scene text detection model with a state-of-the-art image manipulation detection method. § IV-B describes the evaluation metric, and § IV-C presents experimental results with discussions.

A. Scene Text Manipulation Detection Baseline

The construction of the STMD baseline is flexible in selecting an image manipulation detection (IMD) model to combine with a scene text detection model. There exists several prominent IMD algorithms including [5], [7]–[9] as surveyed in § II. After conducting thorough evaluation experiments, we select the MMFusion-IML [6] due to its outstanding IMD performance. We employ the CRAFT scene text detector [17] to exclude the *non-textual* image pixels. Fig. 5 illustrates the step-by-step pipeline for this process. The process initiates with image manipulation detection and scene text detection. Subsequently, all undesired non-textual areas are systematically removed. This procedure yields the MMFusion-IML prediction heatmap specifically highlighting text regions. Following the application of a threshold, a binary mask representing the prediction is derived.

For the pristine images used for model training, we use multiple sources of datasets including Total-Text [14], Tampered-IC13 [16], and ICDAR-15 [36] datasets.

We partition the STMD data samples following the 70-15-15 split for training, validation and testing. An A100 GPU is used for model training, with a learning rate of 0.005.

For model training and test in the experiments, we train the MMFusion-IML model using the TextSleuth training set with

TABLE II
EVALUATION OF THE MMFUSION-IML SCENE TEXT MANIPULATION DETECTION. THE TOP ROW SPECIFIES THE DATASET THE MODEL IS TRAINED ON. THE LEFT COLUMN SHOWS THE EVALUATION DATASET TO PERFORM TESTING.

Training on →	**Stock Model**		**Tampered-IC13 and †Total-Text**		**†SynthText, Tampered-IC13, †Total-Text**	
Test on ↓	AUC	bACC	AUC	bACC	AUC	bACC
Tampered-IC13	0.7983	0.7270	0.9402	0.7121	0.8899	0.8272
SynthText	0.6214	0.6430	0.5822	0.5290	0.9665	0.9181
Tampered Total-Text	0.6388	0.6138	0.6985	0.6578	0.6409	0.5880
All combined	0.6318	0.6444	0.6160	0.5589	0.8950	0.8118

†These are part of the TextSleuth dataset, see § III-A and III-B.

multiple experimental runs, each with different setups. The setup includes: (1) model trained on real tampered data and tested on either real or synthetic data, (2) model trained on the mixed real/synthetic data and tested on either real, synthetic, or mixed data.

B. Evaluation Metric

For IMD, distinguishing between pristine and manipulated images often encounters significant class imbalance. To address this, we use the standard ROC AUC analysis over simple accuracy to mitigate threshold selection issues. The True Positive Rate (TPR), or *sensitivity*, measures the proportion of actual positives correctly identified, while the True Negative Rate (TNR), or *specificity*, assesses the proportion of actual negatives correctly identified. Additionally, we employ balanced accuracy (bACC), which computes the mean of TPR and TNR, offering a balanced evaluation of classification performance across both classes. These metrics are particularly suited for image manipulation detection tasks, where the number of manipulated images may be considerably fewer than authentic ones.

C. Results and Discussion

Evaluation of the MMFusion-IML with CRAFT scene text detection baseline is performed on the TextSleuth test set, including the Tampered Total-Text, SynthText, and Tampered-IC13 test sets. Furthermore, we report the performance of MMFusion-IML on the TextSleuth test set using the original stock model described in the paper [6], which is trained on datasets [10], [13], [37], [38]. Visual results on SynthText and Tampered-IC13 datasets are presented in Fig. 2 and Fig. 3, respectively.

Table II shows the outcomes of our evaluation. The AUC and bACC scores clearly demonstrate the difference between running MMFusion-IML with the CRAFT baseline and using MMFusion-IML for IMD alone, without scene text detection. One notable distinction is the finer granularity of manipulation at the character or word level, which results in a low ratio of manipulated to untouched areas. This aspect poses challenges, particularly in detecting small spliced objects where noise features and artifacts can be difficult to identify.

It is noteworthy that in Table II, the model performs better when trained on the entire TextSleuth dataset (including both real and synthetic tampering), compared to training on smaller subsets such as Tampered-IC13 and Tampered Total-Text alone, which comprises only about 1,000 images. This improvement underscores the benefit of larger dataset sizes in enhancing accuracy. Additionally, results on manually tampered or Total-Text tampered datasets show slightly lower performance, emphasizing the challenges posed by real-world manipulations compared to synthetic ones encountered in controlled environments.

Limitation: Despite its utility, TextSleuth faces limitations in its current iteration. The AUC score of 0.64 on the TextSleuth dataset suggests room for enhancing detection capabilities in real-world scenarios. While comprehensive, the dataset may not encompass the full spectrum of scene text manipulations encountered across diverse contexts. Furthermore, existing STMD models may not adequately address novel or sophisticated tampering techniques that evade detection. The synthetic components of the dataset, though beneficial, might not capture all nuances of real-world tampering, thereby limiting model generalizability.

V. CONCLUSION

We address the growing concern of scene text tampering in digital content, particularly on social media, by introducing TextSleuth, a specialized dataset for scene text tampering detection and localization. We establish a new baseline with state-of-the-art image manipulation detection models to assess the effectiveness of our dataset. We evaluate this baseline trained with TextSleuth for detecting various levels of scene text tampering.

Future Work includes enhancing the effectiveness and robustness of STMD methods. Expanding the dataset to cover a wider range of manipulation techniques and real-world scenarios will improve model generalizability. Developing more sophisticated algorithms to better handle subtle and complex manipulations is crucial. Exploring the integration of multimodal data, such as combining visual information with metadata and contextual analysis, could provide more comprehensive detection capabilities.

Acknowledgments: This work is part of the DARPA Semantic Forensics (SemaFor) Program under contract HR001120C0123. The authors appreciate the computational resource provided by the University at Albany – SUNY.

Declaration: The TextSleuth dataset is solely intended for research purposes. It is designed to aid in the development and evaluation of methods and techniques for detecting scene text manipulation. The creators emphasize that the dataset is not intended to cause offense, harm, or discrimination against any individual, group, or entity. The creators of the TextSleuth dataset disclaim any responsibility for the misuse or misinterpretation of the dataset by users.

REFERENCES

[1] L. Zhao, C. Chen, and J. Huang, "Deep learning-based forgery attack on document images," *IEEE Transactions on Image Processing*, vol. 30, pp. 7964–7979, 2021.

[2] F. Z. Mehrjardi, A. M. Latif, M. S. Zarchi, and R. Sheikhpour, "A survey on deep learning-based image forgery detection," *Pattern Recognition*, vol. 144, p. 109778, 2023.

[3] F. Cruz, N. Sidère, M. Coustaty, V. Poulain d'Andecy, and J.-M. Ogier, "Categorization of document image tampering techniques and how to identify them," in *Pattern Recognition and Information Forensics*, Z. Zhang, D. Suter, Y. Tian, A. Branzan Albu, N. Sidère, and H. Jair Escalante, Eds. Cham: Springer International Publishing, 2019, pp. 117–124.

[4] Z. Zhang, M. Li, and M.-C. Chang, "A new benchmark and model for challenging image manipulation detection," in *AAAI*, 2024.

[5] Y. Wu, W. AbdAlmageed, and P. Natarajan, "Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9535–9544.

[6] K. Triaridis and V. Mezaris, "Exploring multi-modal fusion for image manipulation detection and localization," in *Conference on Multimedia Modeling*, 2023.

[7] C. Dong, X. Chen, R. Hu, J. Cao, and X. Li, "Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, pp. 3539–3553, 2021.

[8] F. Guillaro, D. Cozzolino, A. Sud, N. Dufour, and L. Verdoliva, "Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization," *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20 606–20 615, 2022.

[9] M.-J. Kwon, I.-J. Yu, S.-H. Nam, and H.-K. Lee, "Cat-net: Compression artifact tracing network for detection and localization of image splicing," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 375–384.

[10] J. Dong, W. Wang, and T. Tan, "Casia image tampering detection evaluation database," in *2013 IEEE China Summit and International Conference on Signal and Information Processing*, 2013, pp. 422–426.

[11] B. Wen, Y. Zhu, R. Subramanian, T.-T. Ng, X. Shen, and S. Winkler, "Coverage — a novel database for copy-move forgery detection," in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 161–165.

[12] H. Guan, M. Kozak, E. Robertson, Y. Lee, A. N. Yates, A. Delgado, D. Zhou, T. Kheyrkhah, J. Smith, and J. Fiscus, "Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation," in *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 2019, pp. 63–72.

[13] A. Novozámský, B. Mahdian, and S. Saic, "Imd2020: A large-scale annotated dataset tailored for detecting manipulated images," in *2020 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 2020, pp. 71–80.

[14] C.-K. Ch'ng, C. S. Chan, and C.-L. Liu, "Total-text: toward orientation robustness in scene text detection," *Int. J. Doc. Anal. Recognit.*, vol. 23, no. 1, p. 31–52, mar 2020.

[15] A. Gupta, A. Vedaldi, and A. Zisserman, "Synthetic data for text localisation in natural images," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2315–2324, 2016.

[16] Y. Wang, H. Xie, M. Xing, J. Wang, S. Zhu, and Y. Zhang, "Detecting tampered scene text in the wild," in *Computer Vision – ECCV 2022*, S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds. Cham: Springer Nature Switzerland, 2022, pp. 215–232.

[17] Y. Baek, B. Lee, D. Han, S. Yun, and H. Lee, "Character region awareness for text detection," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9357–9366, 2019.

[18] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 868–882, 2012.

[19] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, ser. IHMMSec '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 5–10.

[20] 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.

[21] T.-T. Ng, J. Hsu, and S.-F. Chang, "Columbia image splicing detection evaluation dataset," *DVMM lab. Columbia Univ CalPhotos Digit Libr*, 2009.

[22] T. J. de Carvalho, C. Riess, E. Angelopoulou, H. Pedrini, and A. de Rezende Rocha, "Exposing digital image forgeries by illumination color classification," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 7, pp. 1182–1194, 2013.

[23] T. Bianchi and A. Piva, "Image forgery localization via block-grained analysis of jpeg artifacts," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 3, pp. 1003–1017, 2012.

[24] H. Guan, M. Kozak, E. Robertson, Y. Lee, A. N. Yates, A. Delgado, D. Zhou, T. Kheyrkhah, J. Smith, and J. Fiscus, "Mfc datasets: Large-scale benchmark datasets for media forensic challenge evaluation," in *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 2019, pp. 63–72.

[25] T.-N. Le, H. H. Nguyen, J. Yamagishi, and I. Echizen, "Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild," *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10 097–10 107, 2021.

[26] N. Sidere, F. Cruz, M. Coustaty, and J.-M. Ogier, "A dataset for forgery detection and spotting in document images," in *2017 Seventh International Conference on Emerging Security Technologies (EST)*, 2017, pp. 26–31.

[27] C. Artaud, A. Doucet, J.-M. Ogier, and V. Poulain d'Andecy, "Receipt Dataset for Fraud Detection," in *First International Workshop on Computational Document Forensics*, Kyoto, Japan, Nov. 2017.

[28] C. H. Lampert, L. Mei, and T. M. Breuel, "Printing technique classification for document counterfeit detection," in *2006 International Conference on Computational Intelligence and Security*, vol. 1, 2006, pp. 639–644.

[29] M. Bibi, A. Hamid, M. Moetesum, and I. Siddiqi, "Document forgery detection using printer source identification—a text-independent approach," in *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, vol. 8, 2019, pp. 7–12.

[30] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 5937–5946.

[31] M. Liao, Z. Wan, C. Yao, K. Chen, and X. Bai, "Real-time scene text detection with differentiable binarization," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2019, pp. 11 474–11 481.

[32] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, and J. Liang, "East: An efficient and accurate scene text detector," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2642–2651, 2017.

[33] J. Ye, Z. Chen, J. Liu, and B. Du, "Textfusenet: Scene text detection with richer fused features," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, C. Bessiere, Ed. International Joint Conferences on Artificial Intelligence Organization, 7 2020, pp. 516–522, main track.

[34] L. Xing, Z. Tian, W. Huang, and M. R. Scott, "Convolutional character networks," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9125–9135, 2019.

[35] X. Xu, Z. Zhang, Z. Wang, B. L. Price, Z. Wang, and H. Shi, "Rethinking text segmentation: A novel dataset and a text-specific refinement approach," *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12 040–12 050, 2020.

[36] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu, F. Shafait, S. Uchida, and E. Valveny, "Icdar 2015 competition on robust reading," in *Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, ser. ICDAR '15. USA: IEEE Computer Society, 2015, p. 1156–1160.

[37] V. V. Kniaz, V. A. Knyaz, and F. Remondino, "The point where reality meets fantasy: Mixed adversarial generators for image splice detection," in *Neural Information Processing Systems*, 2019.

[38] M.-J. Kwon, S.-H. Nam, I.-J. Yu, H.-K. Lee, and C. Kim, "Learning jpeg compression artifacts for image manipulation detection and localization," *International Journal of Computer Vision*, vol. 130, no. 8, p. 1875–1895, May 2022.