Detect Any AI-Counterfeited Text Image

Chenfan Qu[1,2],Yiwu Zhong[3],Xuekang Zhu[2] ,Junchi Li[4,2],Changjiang Jiang[5,2],Jian Liu[2 *],Lianwen Jin[1 *]
[1]South China University of Technology,[2]Ant Group,[3]Peking University,
[4]Zhejiang University,[5]Wuhan University
202221012612@mail.scut.edu.cn, rex.lj@antgroup.com, eelwjin@scut.edu.cn

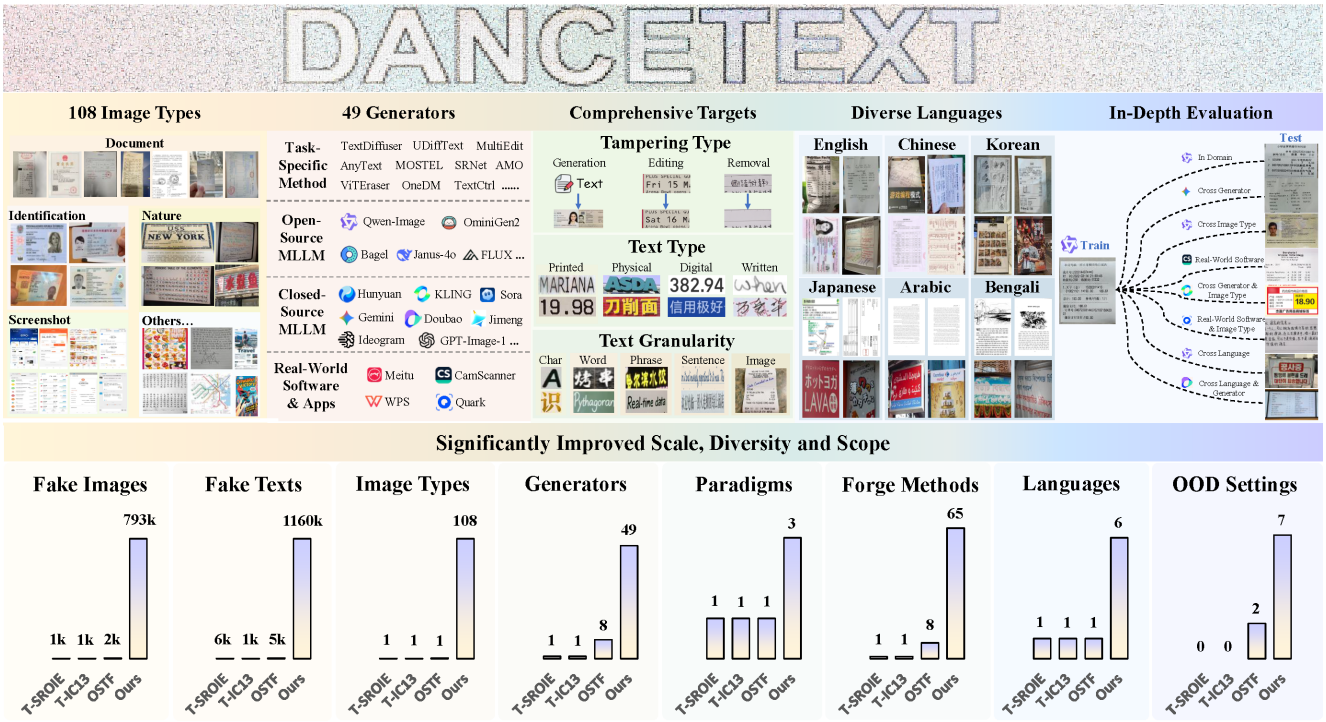

Figure 1. We propose the DanceText dataset to tackle the detection of any AI-counterfeited text image. It establishes a new standard by delivering unprecedented scale and comprehensive scope, spanning a vast array of image types, generators, languages, counterfeit paradigms and evaluation protocols, significantly outperforming previous works in almost all dimensions.

Abstract

The rapid advancement of generative AI enables the creation of highly realistic text images, posing significant security risks from fraud and disinformation. However, research into robust detection is critically hampered by existing datasets that lack scale, diversity, and updated counterfeit techniques, as well as by models that fail to generalize. To address these deficiencies, we introduce DanceText, a large-scale, comprehensive dataset for AI-counterfeited text image detection. Constructed using our novel Creative Proposer pipeline, which automates the generation of diverse and realistic counterfeits, DanceText surpasses previous works by over 100-fold in multiple dimensions. It is the first to include counterfeits from multimodal large models, commercial software, and mobile apps, covering all major paradigms, including full-image generation, regional removal, and editing. Building on this dataset, we propose DS-Net, a novel and effective detection model. It features two key components: a Forensic Decoupling Encoder to extract generator-agnostic artifacts, and a Synergy Denoising Decoder that synergizes image-level classification with instance-level localization. Extensive experiments demonstrate that DS-Net achieves state-of-the-art performance, advancing the field toward robust and unified detection of AI-counterfeited text images in real-world scenarios. Our code and dataset are publicly available at https://github.com/qcf-568/DanceText.

1. Introduction

Generative AI models have achieved surprisingly rapid progress and they can now generate realistic text images of any type effortlessly. These endlessly erupted fake text images can be maliciously used for fraud, rumors, posing unprecedented risks to social security. It is urgently critical to develop robust methods that can detect any type of fake text images generated by any possible generator. However, developing such a robust model is challenging in both data and task aspects.

First, existing datasets [80, 117, 118] cannot support a robust detector due to their limited diversity, scope, scale and quality. Their dataset lacks diversity, only covering single image type, single language or very few image generators. And their scope is specialized, overlooking different counterfeit tools (*e.g.*, commercial apps, generative MLLMs such as Qwen-Image) and different editing targets (*e.g.*, image region removal, full image generation). Furthermore, these datasets are small in scale, fewer than 2,000 samples, with outdated quality. They usually use visual generators developed before 2023 and lag notably behind recent AI advancements.

Second, detecting AI-counterfeited text images is challenging. Each image generator produces distinct artifacts that are often tightly coupled with image content and style [40, 99]. Consequently, detection models tend to overfit to the generator-specific traces and spurious content correlations in the training data, which compromises their ability to generalize to unseen generators and image types.

In this paper, we aim to address these deficiencies through two innovations: (1) a large-scale dataset that is diverse and comprehensive in its coverage of image types, generators, counterfeit paradigms, and languages; (2) a robust detector designed to generalize across unseen image styles and generators.

However, a primary challenge is encountered in scaling up the dataset: How can we control the generators to create text images meeting our requirements? Current methods for generating fake text images suffer from a significant domain gap with real-world scenarios, often yielding unrealistic counterfeits due to simplistic prompts or implausible random edits. To address this, we propose the Creative Proposer, a novel pipeline that automatically generates diverse, high-quality instructions for image generators. Our key insight is to leverage multi-modal large language models that can describe visual details in rich, semantically-correct language, which in turn can be naturally digested by generators. This approach produces diverse text images that closely align with real-world scenarios, encompassing both full image synthesis and semantically plausible regional counterfeits.

Using our Creative Proposer pipeline and 48 distinct generators, we created 793,731 realistic counterfeited images based on a large and diverse collection of real text images from 108 categories, covering a wide range of real-world scenarios. The resulting images are split into a training set and eight distinct test sets for in-depth evaluation. As shown in Figure 1, our DanceText dataset significantly surpasses previous works in scale, diversity, and comprehensiveness. Compared to prior datasets, it provides hundreds of times more real and counterfeited text images, 100 times more image types, eight times more counterfeit methods, and six times more generators and languages. Notably, DanceText is the first dataset to incorporate text images generated by open-source and closed-source multimodal large models, commercial software or mobile applications. It is the first dataset to include AI-counterfeits of real-world photographed documents. It is also the first to cover both full-image generation and regional erasure counterfeiting techniques.

With such a large scale, comprehensiveness and diversity, our DanceText significantly bridges the gap between real-world applications while inspiring further research. For the first time, it creates the opportunity to train and evaluate a truly robust counterfeit detector, one that generalizes across any image type, any text style, any counterfeit type, any level of granularity, any language, and any generator.

Building upon the DanceText dataset, we further propose the DS-Net, a novel model for unified and generalizable AI-counterfeited text image detection. DS-Net features two key innovations. The Forensic Decoupling Encoder learns to disentangle generator-specific artifacts from semantic content and leverages vast and diverse counterfeited data from non-text domains. The Synergy Denoising Decoder then mimics the reasoning of a human expert by creating explicit synergy between image-level classification and instance-level localization. Extensive experiments on both the DanceText dataset and public datasets have verified the effectiveness of the proposed method, significantly outperforming previous methods.

Our main contributions are summarized as follows:

- **DanceText Dataset**: A large-scale, comprehensive, and high-quality dataset for detecting AI-counterfeited text images. It surpasses previous datasets by over 100-fold in multiple dimensions such as scale and diversity.
- **Creative Proposer Pipeline**: A novel pipeline that automatically and efficiently generates high-quality instructions for image generators. This process produces realistic and diverse synthetic text images with close alignment to real-world counterfeiting characteristics.
- **DS-Net**: A unified and generalizable detection model that incorporates artifact–content decoupling and explicit task synergy between image-level classification and instance-level localization. This novel and effective design leads to significant performance gains over prior approaches.

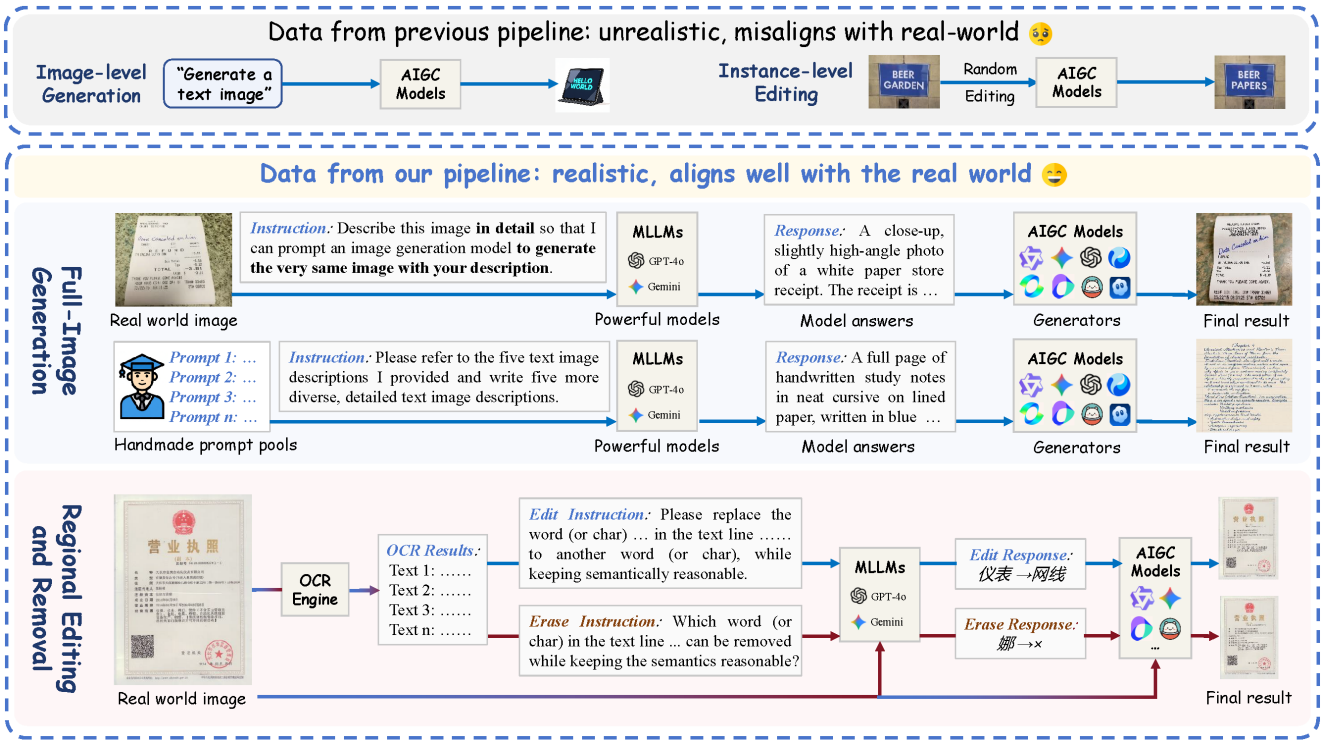

Figure 2. The proposed Creative Proposer pipeline produces diverse high-quality prompts for generating realistic text images.

2. Related Works

2.1. Text Image Generation, Editing and Removal

Recent advancements in generative models, including GANs, Diffusion models, and multimodal large language models like Qwen-Image [123], have enabled the high-quality synthesis of text images [20]. These models facilitate not only full-image generation [34, 128] but also sophisticated regional editing and removal [50, 116]. The accessibility and realism of these AI-based techniques, which require minimal expertise compared to traditional methods, pose a significant security threat [14, 29, 30, 54, 65, 67, 77–79, 81, 82, 90, 93, 95, 122, 139, 144].

2.2. AI-Counterfeited Text Image Detection

Several benchmarks have been proposed to address this threat, but they suffer from critical limitations. For example, DocTampered [76] contains no AI-counterfeited texts. Datasets like T-SROIE [118], T-IC13 [117], and OSTF [80] are restricted in scale, diversity (often one image type), and modernity, relying on outdated generators and overlooking real-world counterfeits from MLLMs, software and apps. Existing detection methods also fall short. Models such as S3R [117], FFDN [22], and DAF [80] show promise in localizing edits but struggle with image-level classification and detecting text removal. Furthermore, robustness to unseen generators and image types remains a key unsolved challenge. General DeepFake detectors are often limited to image-level classification and cannot perform the necessary regional localization [4, 7, 8, 11–13, 19, 21, 23–26, 31, 33, 36, 39, 40, 43–45, 47, 48, 51, 53, 55–57, 59, 60, 66, 68, 70, 84, 86, 88, 89, 99–102, 110, 111, 113, 119, 121, 127, 129, 131, 133, 136, 137, 140, 141, 143, 145]. A pressing need, therefore, exists for a comprehensive dataset and a generalizable detection model that can address the full spectrum of modern AI-based text forgeries.

3. Creative Proposer Pipeline

A critical step in constructing a large-scale dataset for AI-counterfeited text detection is obtaining high-quality, diverse prompts to guide image generators. Current methods for generating synthetic text images suffer from a significant domain gap: simplistic prompts yield unrealistic full images (Fig. 2, top), while random regional edits lack the semantic plausibility of real-world forgeries. To address this, we introduce the Creative Proposer, a novel pipeline that automatically generates diverse, high-quality instructions for both full-image generation and regional editing.

3.1. Full-Image Generation

Our pipeline employs two sub-pipelines to generate fully synthetic, high-quality text images (Fig. 2 middle).

1. **Image-to-Text-to-Image**: We prompt Multimodal Large

Dataset	Fake Image	Fake Text	Real Image	Image Types	Tamp. Methods	Latest Tamp.	Lang. Num.	Image Generators			Paradigms			Evaluation		
								T.S.	MLLM	R.S.	R.E.	R.R.	I.G.	C-G.	C-T.	C-L.
T-SROIE [118]	986	6,225	0	1	1	2019	1	1	0	0	✓	×	×	×	×	×
T-IC13 [117]	378	995	84	1	1	2019	1	1	0	0	✓	×	×	×	×	×
OSTF [80]	1,980	5,018	838	1	8	2023	1	8	0	0	✓	×	×	✓	×	×
Ours	**793,731**	**1,160,762**	**144,657**	**108**	**65**	**2025**	**6**	**26**	**19**	**4**	✓	✓	✓	✓	✓	✓

Table 1. Comparison between **publicly available** datasets for AI-counterfeited text image detection. 'Tamp.': Tampering, 'Lang.': Language, 'T.S.': Task-Specific generator, 'MLLM': Multimodal Large Model, 'R.S.': Real-world Software, 'R.E.': Regional Editing, 'R.R.': Regional Removal, 'I.G.': Image-level Generation. 'C-G.': Cross-Generator, 'C-T.': Cross-Image-Type, 'C-L.': Cross-Language.

Language Models (MLLMs), such as Gemini-2.5-pro and GPT-4o, to produce a detailed textual description of a real-world text image. Our prompt, "Describe this image in detail so that I can prompt an image generation model to generate the very same image," elicits rich descriptions that capture subtle visual elements. These descriptions are then fed to text-to-image generators (e.g., Qwen-Image, HunYuan3) to synthesize realistic counterfeits. Inherent MLLM recognition errors are leveraged as a natural form of text content counterfeiting, further enhancing data diversity.

2. **Text-to-Text-to-Image**: To augment diversity and accommodate generators with prompt length limitations, we use a text-only approach. Starting with a curated seed pool of high-quality prompts, we iteratively instruct MLLMs to generate new, distinct prompts that maintain a similar level of detail. This expanding pool of prompts is then used to generate a wide variety of realistic synthetic images.

3.2. Regional Editing and Removal

For regional counterfeits, our pipeline generates semantically plausible edits (Fig. 2). The process is as follows:

1. **Text Extraction**: An OCR engine extracts text and coordinates from an input image. These are then grouped into semantically meaningful text segments (char / word / phrase) using NLP and rule-based methods.

2. **Plausible Edit Proposal**: For each text segment, we provide its content, surrounding context, position, and the original image to an MLLM. The MLLM is queried to propose a semantically plausible replacement or removal.

3. **Controlled Inpainting**: If a valid edit is proposed, we mark the region with a blue bounding box for disambiguation and instruct inpainting models to perform the edit while removing the visual marker. A rigorous post-processing pipeline filters out low-quality results, with further details provided in the Appendix.

4. DanceText Dataset

We introduce DanceText, a large-scale, comprehensive, and high-quality dataset designed to fundamentally advance the Detection of any AI-counterfeited Text image.

4.1. Motivation

The development of DanceText is motivated by the critical limitations of prior works. For example:

• **Limited Scale**: Fewer than 2,000 counterfeit images (Table 1), too small for robust model training and evaluation.

• **Limited Image Types**: Restricted to a single type (e.g., scanned receipts or signboards), ignoring the diversity of real-world text images.

• **Limited Generators**: Counterfeited with only task-specific open-source models, excluding the more accessible general-purpose MLLMs (e.g., Qwen-Image, Gemini-2.5-flash-image) and real-world software.

• **Limited Counterfeit Paradigms**: Focusing solely on regional editing and neglecting full-image generation and regional removal paradigms, which can also pose threats.

• **Outdated Forgery**: All generators developed before 2023, producing outdated counterfeits that lag significantly behind the current generative-AI capabilities.

• **Limited Language Scope**: Monolingual, limited to English, hindering cross-lingual evaluation.

Our DanceText dataset overcomes **all** these constraints, significantly expanding scale, scope, and diversity to match real-world detection demands.

4.2. Construction

We collected 144,657 authentic text images from 108 categories, counterfeited them using the Creative Proposer and 45 state-of-the-art text image generators, including:

26 Task-Specific Models: SR-Net [125], MOSTEL [83], STEFANN [87], DST [91], DiffSTE [42], AnyText [108], UDiffText [138], TextDiffuser [16], TextCtrl [134], TextFlux [126], FluxText [49], DiffUTE [15], STRIVE [98], RSSTE [32], AnyText2 [109], TextDiffuser2 [17], AMO [38], GlyphByT5 [63], Type-R [92], BizGen [75], MultiEdit [52], ViTEraser [74], CTRNet++ [58], TMIM [120], WordStylist [69], OneDM [27].

8 Open-source MLLMs: Qwen-Image [123], OmniGen2 [124], Bagel [28], Qwen-Image-Edit [123], BagelCoT [28], Janus-4o [18], Flux.1-Kontext [2], Flux.1-Kera [3].

11 Closed-source MLLMs: Gemini-2.5-flash-image [94], KLING [107], Doubao4-i2i [5], HunYuanImage3 [10],

Jimeng4-t2i [6], Sora-image [72], Doubao3-i2i [5], GPT-Image-1 [71], Jimeng3.1 [5], Wan-2.5 [1], Ideogram [41].

Open-source MLLMs contribute 60% of the counterfeited images, which are primarily allocated to the training set (with Qwen-Image-Edit-2509 accounting for 45%). For evaluation purposes, closed-source MLLMs and task-specific generators each provide 20% of the images, ensuring a comprehensive test of model robustness.

The dataset underwent a rigorous cleaning process, first using automated methods and then human inspection to ensure quality. We also created 12,000 real-world counterfeits **manually** using **4 commercial AI-powered applications**, Meitu [104], CamScanner [103], WPS [106], Quark [105], to capture the real-world counterfeit characteristics.

The dataset is divided into a training set (counterfeited by Qwen-Image, Qwen-Image-Edit-2509, TextDiffuser, SRNet, ViTEraser) and eight distinct test subsets to facilitate a thorough evaluation of model robustness and generalization under various cross-domain conditions:

(1) DanceText-Test: An in-distribution test set with data from the same distribution as the training set. **(2) DanceText-CT**: Evaluates robustness to unseen image Types. **(3) DanceText-CG**: Evaluates robustness to unseen Generators. **(4) DanceText-CTG**: Tests generalization to both unseen image types and generators. **(5) DanceText-CL**: Assesses cross-lingual generalization to unseen Languages. **(6) DanceText-CLG**: Combines unseen Languages and unseen Generators. **(7) DanceText-RW**: Contains images manually counterfeited by Real-World software and Apps. **(8) DanceText-RWT**: Combines unseen image Types with handcrafted counterfeits from Real-World software. Statistics are presented in the Appendix.

4.3. Highlights

As shown in Table 1, our DanceText establishes a new standard for research in AI-counterfeited text detection, eclipsing all previous datasets in every critical dimension.

• **Unprecedented Scale**: With over **400 times** more counterfeit and real images than all prior datasets combined, DanceText provides the data scale necessary for training and evaluating powerful deep learning models.

• **Unmatched Diversity**: DanceText features a **100-fold** increase in image types (108 total), an **8-fold** increase in forgery methods, and a **6-fold** increase in generators (49 total). Crucially, it is **the first** to be multilingual, spanning six languages and enabling cross-lingual generalization studies.

• **First-of-its-Kind Realism**: DanceText is **the first** and only dataset to include forgeries from modern MLLMs and, most importantly, 12,000 expert-crafted forgeries from commercial software and apps. This unique subset, complete with real-world post-processing like social media transmission, provides an invaluable, "in-the-wild" benchmark for practical model performance. All other synthetic images were produced by our Creative Proposer pipeline specifically designed to ensure high realism, effectively bridging the gap to real-world scenarios.

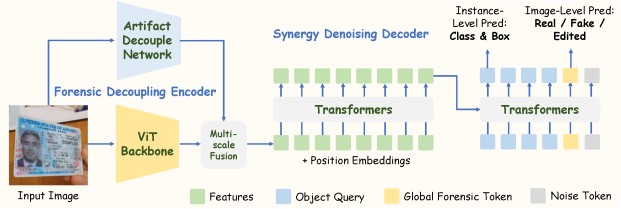

Figure 3. The proposed Decouple and Synergy Network.

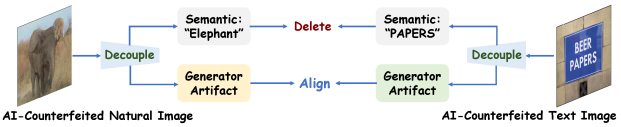

Figure 4. The key idea of our Forensic Decoupling Encoder.

• **Comprehensive Counterfeit Coverage**: DanceText has unmatched comprehensiveness, as it is **the first** to span all three major forgery paradigms (regional editing, removal, and full-image generation), all four major text types (printed, physical, digital, and handwritten), and all five levels of text granularity (character, word, phrase, sentence, and image), ensuring models are prepared for the complete spectrum of AI-based attacks.

• **State-of-the-Art Modernity**: With over half of its generators (29 of 49) developed in 2025, DanceText directly reflects the current, rapidly evolving threat landscape of generative-AI. This ensures that evaluations are truly representative of contemporary, real-world attacks, making our benchmark both relevant and future-proof.

• **Pioneering Evaluation Protocol**: DanceText provides **the first** comprehensive suite of out-of-domain tests, including robustness against unseen image types, generators, languages, real-world software, and their combinations. This enables an unprecedented, in-depth analysis of models, fostering insights that will drive future research.

In summary, DanceText is not merely an incremental update; it is a **fundamental leap forward**, providing the community with the first truly comprehensive resource to build and validate the next generation of robust counterfeit detection systems for text images.

A few samples from the DanceText dataset are shown in Fig. 1. Additional samples and details, including per-generator statistics, are available in the Appendix.

5. Decouple and Synergy Network

To achieve unified and generalized detection of AI-counterfeited text images, we propose the Decouple and Synergy Network (DS-Net). As illustrated in Fig. 3, DS-

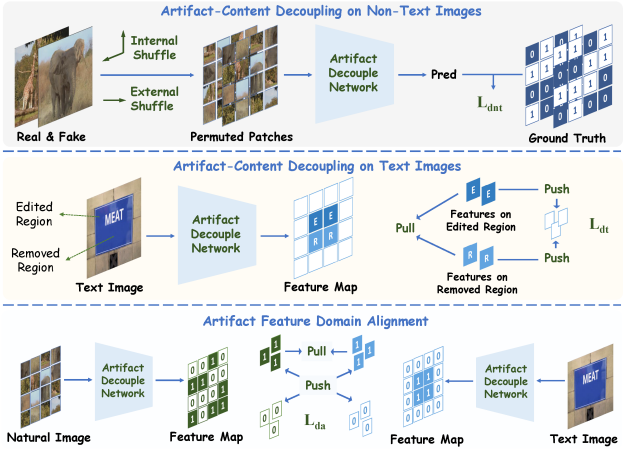

Figure 5. Artifact Decouple Network training pipeline.

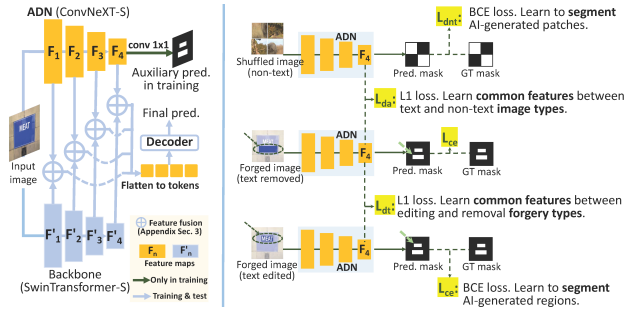

Figure 6. Overall optimization pipeline of our Encoder.

Net consists of a Forensic Decoupling Encoder and a Synergy Denoising Decoder. The encoder is designed to extract generalizable forensic features through artifact-content decoupling and leveraging diverse data from other domains. The decoder improves performance through an explicit synergy between the classification and localization tasks.

5.1. Forensic Decoupling Encoder

Motivation. The relatively limited number of available text image generators provides insufficient artifact diversity, leading to model overfitting. In contrast, non-text domains offer vast fake data; for instance, the Community Forensics dataset [73] contains 2.7 million images from over 4,800 generators. Inspired by this, we propose to leverage the diversity of non-text forgeries to mitigate overfitting and improve generalization for text image detection.

Key Idea. Directly training on non-text images is suboptimal due to content and task misalignments (e.g., text vs. non-text content; full-image generation vs. regional edits). To overcome this, we introduce the Forensic Decoupling Encoder. As shown in Fig. 4, its core idea is to first decouple generator-produced artifacts from semantic content within both text and non-text domains, and then align these

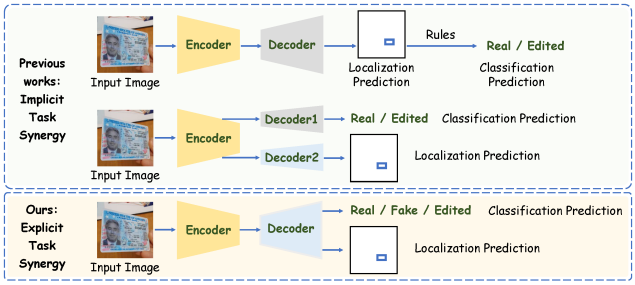

Figure 7. The key idea of our Synergy Denoising Decoder.

purified artifact features in a common latent space.

Method. Our encoder consists of a ViT [61] backbone and a parallel Artifact Decouple Network (ADN) with a ConvNeXt [62] backbone. While ViT features can overfit, the ADN is designed to extract generalized, semantic-agnostic artifact features, which then enhance the primary ViT features. The ADN is trained with three novel loss functions:

1. L_{dnt} is for decoupling on non-text images, Fig. 5 top. It is calculated on the ADN's auxiliary output mask from its top feature map F4 (Fig. 6). We apply internal (spatial) and external (batch-wise) patch shuffling to real and fake non-text images. Internal shuffling disrupts semantics, forcing the ADN to focus on local artifacts; External shuffling enables patch-level independent prediction on these image-level fully generated data, bridging the gap between image-level classification and regional localization tasks. L_{dnt} is a binary cross-entropy (BCE) loss on these shuffled patches.

2. L_{dt} for decoupling on text images, Fig. 5 middle. It is calculated on the ADN's top output feature map F4 (Fig. 6). Using an L2 loss, we attract the feature representations of edited E and removed R text regions while repelling them from authentic regions. Since the commonality between E and R is the presence of artifacts, while **their difference is the absence of text content for** R, this forces the ADN to learn content-agnostic artifact features.

3. L_{da} for domain alignment, Fig. 5 bottom. It is calculated on the ADN's top feature map F4 (Fig. 6). To bridge the distributional gap between the two domains, L_{da} aligns their feature maps in the latent space. It attracts features from fake regions in both domains while repelling them from authentic ones, enabling the model to learn a unified representation of diverse artifacts.

The ADN is trained end-to-end with the loss $L_{ADN} = L_{dnt} + L_{dt} + L_{da} + L_{ce}$, where L_{ce} is a standard pixel-level BCE loss for text images. It is calculated on the ADN's auxiliary output mask (Fig. 6). Finally, multi-scale features from the ADN and ViT are fused via channel attention [37].

5.2. Synergy Denoising Decoder

Motivation. Human experts synergize image-level assessment with region-level investigation. Global anomalies

prompt a closer inspection of local regions, while the discovery of a local forgery informs the final image-level decision. In contrast, previous methods fail to model this synergistic reasoning. As shown in Fig. 7, they either derive a global decision from a noisy localization map or use separate, non-interacting heads for each task [80, 117]. To address this limitation, we propose a novel decoding paradigm that explicitly restores this crucial, bidirectional interaction.

Key Idea. The core of our decoder is a learnable Global Forensic Query. As shown in Fig. 3, this query iteratively interacts with instance-level localization queries within the transformer decoder layers. This ensures its final image-level prediction is informed by local evidence, while better rectifying the instance queries with global information, mimicking an expert's reasoning process.

Method. Our decoder adapts the DINO's denoising architecture [135]. Encoder features are fed into transformer decoder layers along with two types of queries: Object queries that predict bounding boxes and the forgery type (removal or editing) for suspected regions. Global Forensic Query (GFQ) that produces a three-way image-level classification (real, fully generated, or regionally edited). The decoder is trained end-to-end using DINO's loss functions, augmented with a cross-entropy loss for the image-level classification of our GFQ. Further details are provided in the Appendix.

6. Experiments

6.1. Implementation Details

We adopted the small versions of Swin-Transformer [61] and ConvNeXt [62] as the backbones for our main model and the Artifact Decouple Network (ADN), respectively. All models were trained with a batch size of 32. We used the AdamW [64] optimizer, with a learning rate that decayed from 8e-6 to 0. The specific training iterations varied by dataset: DanceText: 120,000 iterations. T-IC13: 15,000 iterations. OSTF: 80,000 iterations of Texture Jitter pre-training [80], followed by 10,000 iterations of fine-tuning. Input images were resized to 1024×1536 for text images and 512×512 for non-text images. For non-text images, a patch size of 32 was used. Image-level performance for the three-way classification task (real, fully-synthesized, or regionally edited) was evaluated using the balanced accuracy metric [35, 132]. Instance-level counterfeit text detection performance was evaluated using F1-score, calculated with the ICDAR2015 DetEval Protocol [46]. This evaluation methodology adheres to standard practices in both text detection [96, 97, 130] and AI-counterfeited text detection research [80, 117, 118].

6.2. Comparison Study

For a fair comparison, all models in Table 2 were re-trained on the DanceText-Train dataset using the same backbone and training configurations. Since previous models originally lack image-level classification capabilities, we added a fully-connected layer to the top of their backbones to serve as a classification head.

As shown in Table 2, all models perform well on the in-domain test set. Their performance on the cross-image-type set (CT) is comparable, demonstrating that our DanceText dataset contains sufficient diversity to support generalization to unseen image types. However, all models exhibit a significant performance drop in cross-generator evaluations (CG vs. Test, CTG vs. CT), which indicates that deep models tend to overfit to specific generator artifacts. Performance also degrades in cross-language evaluations (CL vs. Test), particularly for the localization F1-score. This suggests that the learned features are highly coupled with the text's visual content, as merely changing the language degrades performance even when the generator is unchanged. When both the language and generator are changed (CLG vs. Test), this degradation becomes even more pronounced. Finally, all models performed worst on the counterfeits produced manually by real-world software and apps (RW and RWT). This difficulty arises because these real-world tools use proprietary generators trained on in-house data, which are optimized to produce highly realistic forgeries. Furthermore, they often integrate post-processing steps that conceal generator artifacts.

Despite these challenges, our model demonstrates superior robustness. The Forensic Decoupling Encoder effectively decouples generator artifacts from visual content, alleviating performance loss in cross-language scenarios. By leveraging diverse non-text data and aligning artifact features, our model also achieves better cross-generator generalization. Furthermore, the Synergy Denoising Decoder facilitates more robust analysis by synergizing the classification and localization tasks. These designs significantly mitigate cross-domain performance degradation, enabling our model to considerably outperform existing methods. As shown in Table 2, our model also achieves state-of-the-art results on both T-IC13 [117] and OSTF [80] datasets when trained and tested under the same configurations as previous work, further validating the effectiveness of our designs.

The qualitative comparison provided in Figure 8 further confirms the effectiveness of the proposed method. Due to page limitations, additional comparison studies are provided in the Appendix, including per-generator and per-language breakdowns, results against segmentation-based models, and cross-dataset analyses.

6.3. Ablation Study

Ablation results for DS-Net are presented in Table 4. Setting (1) corresponds to the DINO baseline [135] without any proposed modules. Setting (9) represents the full DS-Net.

Key findings: Setting (9) outperforms (2): Internal patch

Method	Test		CT		CG		CTG		CL		CLG		RW		RWT		AVG	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
RFRM [118]	92.8	73.1	93.8	75.8	76.2	52.9	59.4	39.7	92.5	58.7	58.7	33.5	59.1	5.2	61.6	12.0	74.3	43.9
EAST [142]-S3R [117]	92.2	63.3	93.4	67.2	74.8	47.4	58.7	34.4	92.1	50.2	58.4	27.4	60.3	5.1	62.4	9.5	74.0	38.1
PSENet [112]-S3R [117]	92.4	67.4	93.4	71.9	74.9	49.8	59.0	36.1	92.2	53.4	58.4	29.9	60.4	5.7	62.6	11.4	74.2	40.7
ATRR [114]-S3R [117]	92.5	72.9	93.6	76.6	75.5	53.2	59.2	40.3	92.2	58.8	58.5	32.0	60.8	6.4	63.0	13.7	74.4	44.2
CounterNet [115]-S3R [117]	92.3	74.2	93.5	78.5	75.3	56.5	59.3	41.6	92.2	60.1	58.6	33.8	60.7	7.6	63.0	14.1	74.4	45.8
FRCNN [85]-DAF [80]	92.6	77.3	93.9	80.3	75.5	58.7	59.2	43.5	92.5	62.9	59.2	36.1	61.0	9.5	63.2	16.0	74.6	48.0
CRCNN [9]-DAF [80]	92.6	78.5	93.8	81.2	75.6	60.4	58.9	44.8	92.4	64.0	59.1	38.6	60.8	9.8	63.5	17.9	74.6	49.4
DS-Net (Ours)	**93.2**	**83.6**	**94.0**	**86.1**	**80.8**	**68.7**	**67.9**	**45.3**	**92.6**	**72.1**	**63.5**	**39.4**	**61.7**	**12.3**	**66.2**	**24.3**	**77.4**	**53.9**

Table 2. Comparison on the eight test subsets of DanceText dataset. "Acc.": Category-balanced accuracy of image-level classification. "F1": Instance-level F1-score of regional editing and removal detection, this score is only evaluated on regionally counterfeited images.

Method	T-IC13 [117]			OSTF [80]		
	P	R	F	P	R	F
RFRM [118]	83.3	90.2	86.6	56.4	40.3	42.3
EAST [142]-S3R [117]	70.2	70.0	69.9	52.2	36.8	38.4
PSENet [142]-S3R [117]	79.9	79.4	79.7	53.7	38.6	40.0
ATRR [142]-S3R [117]	84.6	90.6	87.5	56.8	41.4	42.9
CounterNet [142]-S3R [117]	86.7	91.5	89.0	58.6	42.6	44.5
FRCNN [142]-DAF [80]	91.4	96.3	93.8	77.6	72.6	73.7
CRCNN [142]-DAF [80]	92.4	96.7	94.4	80.4	72.4	75.0
DS-Net (Ours)	**93.6**	**97.4**	**95.5**	**82.9**	**76.7**	**78.2**

Table 3. Comparison study on the public Tampered-IC13 and OSTF datasets. "P": Precision. "R": Recall. "F": F1-score.

Set.	Ablation	T-IC13	OSTF	DanceText	
		F1	F1	Acc.	F1
(1)	Baseline	88.5	72.5	74.2	45.1
(2)	w.o. Internal Shuffle	93.8	75.5	75.6	51.6
(3)	w.o. External Shuffle	91.4	74.9	75.9	50.5
(4)	w.o. L_{dnt}	89.3	73.8	75.8	47.9
(5)	w.o. L_{dt}	-	-	76.7	51.1
(6)	w.o. L_{da}	90.8	75.0	76.3	49.0
(7)	w.o. ADN (Sec. 5.1)	89.3	73.8	75.4	47.3
(8)	w.o. Synergy (Sec. 5.2)	95.3	77.3	75.9	51.8
(9)	DS-Net (Ours)	95.5	78.2	77.4	53.9

Table 4. Ablation study of DS-Net. "Acc.": Balanced accuracy.

shuffle breaks semantics in non-text images, enabling effective artifact-content decoupling. (9) outperforms (3): External shuffle bridges task gaps, allowing use of image-level synthesized and labeled fake data to improve regional counterfeit localization. (9) outperforms (4): L_{dnt} significantly improves utilization of non-text fake images. (9) outperforms (5): L_{dt} reduces overfitting to forgery-irrelevant visual content. (9) outperforms (6): Without L_{da}, artifact features from text and non-text domains reside in different latent spaces; alignment is essential for bridging domain gaps. (7) under-performs (9) by removing the ADN module and the three losses from (9): ADN extracts more generalized artifact features that are essential for cross-domain generalization. (8) under-performs (9) by removing the Global Forensic Query from (9): The Synergy Denoising Decoder enables a "$1 + 1 > 2$" synergistic effect between image-level classification and regional localization. Each proposed module contributes to performance improvements. Due to page constraints, further ablation experiments, including robustness evaluations, are provided in the Appendix.

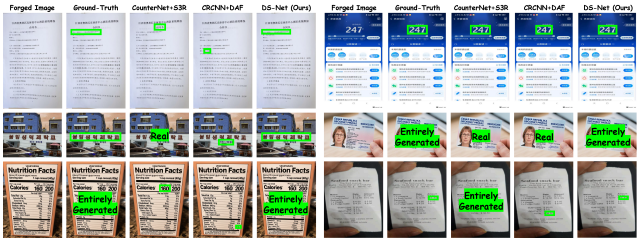

Figure 8. Visual qualitative comparison on DanceText.

7. Conclusion

In this paper, we addressed the urgent challenge of detecting AI-counterfeited text images by introducing two primary contributions: the DanceText dataset and the DS-Net detection model. We demonstrated that existing datasets are inadequate for real-world scenarios due to severe limitations in scale, diversity, and forgery realism. To overcome this, we developed the Creative Proposer pipeline to construct DanceText, a comprehensive benchmark that is the first to include forgeries from modern MLLMs and commercial applications, thereby bridging the gap to real-world threats. Furthermore, our proposed DS-Net architecture introduces novel forensic decoupling and task synergy mechanisms, achieving state-of-the-art performance and demonstrating remarkable generalization to unseen forgeries. Together, DanceText and DS-Net not only set a new standard for research in this domain but also provide a robust foundation for developing next-generation security tools. We believe this work will significantly accelerate progress in combating text counterfeiting and foster a more secure information ecosystem.

Acknowledgments

This research is supported in part by the National Natural Science Foundation of China (Grant No.:62476093) and Ant Group Research Intern Program.

References

[1] Alibaba Cloud. Wan ai: Leading ai video generation model. https://tongyi.aliyun.com/wan/, 2025. 5

[2] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025. 4

[3] black-forest labs. Flux.1-krea-dev. https://huggingface.co/black-forest-labs/FLUX.1-Krea-dev, 2025. 4

[4] Jonathan Brokman, Amit Giloni, Omer Hofman, Roman Vainshtein, Hisashi Kojima, and Guy Gilboa. Manifold induced biases for zero-shot and few-shot detection of generated images. *arXiv preprint arXiv:2504.15470*, 2025. 3

[5] ByteDance Inc. Doubao. https://www.doubao.com/chat/create-image, 2025. 4, 5

[6] ByteDance Inc. Jimeng ai. https://jimeng.jianying.com/, 2025. 5

[7] Qianshu Cai, Chao Wu, Yonggang Zhang, Jun Yu, and Xinmei Tian. Towards generalizable detector for generated image. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 3

[8] Yu Cai, Shan Jia, Jiahe Tian, Jiao Dai, Jizhong Han, and Siwei Lyu. Cataid: Category-guided ai-generated image detection via vision-language model adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 1553–1563, 2025. 3

[9] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 8

[10] Siyu Cao, Hangting Chen, Peng Chen, Yiji Cheng, Yutao Cui, Xinchi Deng, Ying Dong, Kipper Gong, Tianpeng Gu, Xiusen Gu, et al. Hunyuanimage 3.0 technical report. *arXiv preprint arXiv:2509.23951*, 2025. 4

[11] George Cazenavette, Avneesh Sud, Thomas Leung, and Ben Usman. Fakeinversion: Learning to detect images from unseen text-to-image models by inverting stable diffusion, 2024. 3

[12] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize, 2020.

[13] Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. DRCT: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-first International Conference on Machine Learning*, 2024. 3

[14] Changsheng Chen, Liangwei Lin, Yongqi Chen, Bin Li, Jishen Zeng, and Jiwu Huang. Cma: a chromaticity map adapter for robust detection of screen-recapture document images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15577–15586, 2024. 3

[15] Haoxing Chen, Zhuoer Xu, Zhangxuan Gu, jun lan, , Yaohui Li, Changhua Meng, Huijia Zhu, and Weiqiang Wang. Diffute: Universal text editing diffusion model. In *Advances in Neural Information Processing Systems*, pages 63062–63074. Curran Associates, Inc., 2023. 4

[16] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. In *Advances in Neural Information Processing Systems*, pages 9353–9387. Curran Associates, Inc., 2023. 4

[17] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. In *European Conference on Computer Vision*, pages 386–402. Springer, 2024. 4

[18] Junying Chen, Zhenyang Cai, Pengcheng Chen, Shunian Chen, Ke Ji, Xidong Wang, Yunjin Yang, and Benyou Wang. Sharegpt-4o-image: Aligning multimodal models with gpt-4o-level image generation. *arXiv preprint arXiv:2506.18095*, 2025. 4

[19] Ruoxin Chen, Junwei Xi, Zhiyuan Yan, Ke-Yue Zhang, Shuang Wu, Jingyi Xie, Xu Chen, Lei Xu, Isabel Guan, Taiping Yao, et al. Dual data alignment makes ai-generated image detector easier generalizable. *arXiv preprint arXiv:2505.14359*, 2025. 3

[20] Xinhong Chen, Bangdong Chen, Chenfan Qu, Dezhi Peng, Chongyu Liu, and Lianwen Jin. Dtsm: Toward dense table structure recognition with text query encoder and adjacent feature aggregator. In *International Conference on Document Analysis and Recognition*, pages 438–452. Springer, 2024. 3

[21] Yingjian Chen, Lei Zhang, and Yakun Niu. Forgelens: Data-efficient forgery focus for generalizable forgery image detection, 2025. 3

[22] Zhongxi Chen, Shen Chen, Taiping Yao, Ke Sun, Shouhong Ding, Xianming Lin, Liujuan Cao, and Rongrong Ji. Enhancing tampered text detection through frequency feature fusion and decomposition. In *European Conference on Computer Vision*, pages 200–217. Springer, 2024. 3

[23] Siyuan Cheng, Lingjuan Lyu, Zhenting Wang, Xiangyu Zhang, and Vikash Sehwag. Co-spy: Combining semantic and pixel features to detect synthetic images by ai, 2025. 3

[24] Sungik Choi, Sungwoo Park, Jaehoon Lee, SeungHyun Kim, Stanley Jungkyu Choi, and Moontae Lee. Training-free detection of ai-generated images via high-frequency influence.

[25] Beilin Chu, Xuan Xu, Xin Wang, Yufei Zhang, Weike You, and Linna Zhou. Fire: Robust detection of diffusion-generated images via frequency-guided reconstruction error, 2025.

[26] Davide Cozzolino, Giovanni Poggi, Matthias Nießner, and

Luisa Verdoliva. Zero-shot detection of ai-generated images, 2024. 3

[27] Gang Dai, Yifan Zhang, Quhui Ke, Qiangya Guo, and Shuangping Huang. One-dm: One-shot diffusion mimicker for handwritten text generation. In *European Conference on Computer Vision*, pages 410–427. Springer, 2024. 4

[28] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025. 4

[29] Li Dong, Weipeng Liang, and Rangding Wang. Robust text image tampering localization via forgery traces enhancement and multiscale attention. *IEEE Transactions on Consumer Electronics*, 70(1):3495–3507, 2024. 3

[30] Bo Du, Xuekang Zhu, Xiaochen Ma, Chenfan Qu, Kaiwen Feng, Zhe Yang, Chi-Man Pun, Jian Liu, and Jizhe Zhou. Forensichub: A unified benchmark codebase for all-domain fake image detection and localization, 2025. 3

[31] Mingqi Fang, Ziguang Li, Lingyun Yu, Quanwei Yang, Hongtao Xie, and Yongdong Zhang. Forensic-moe: Exploring comprehensive synthetic image detection traces with mixture of experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17772–17782, 2025. 3

[32] Zhengyao Fang, Pengyuan Lyu, Jingjing Wu, Chengquan Zhang, Jun Yu, Guangming Lu, and Wenjie Pei. Recognition-synergistic scene text editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13104–13113, 2025. 4

[33] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of the 37th International Conference on Machine Learning*, pages 3247–3258. PMLR, 2020. 3

[34] Yifan Gao, Zihang Lin, Chuanbin Liu, Min Zhou, Tiezheng Ge, Bo Zheng, and Hongtao Xie. Postermaker: Towards high-quality product poster generation with accurate text rendering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8083–8093, 2025. 3

[35] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20606–20615, 2023. 7

[36] Fabrizio Guillaro, Giada Zingarini, Ben Usman, Avneesh Sud, Davide Cozzolino, and Luisa Verdoliva. A bias-free training paradigm for more general ai-generated image detection, 2025. 3

[37] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 6

[38] Xixi Hu, Keyang Xu, Bo Liu, Qiang Liu, and Hongliang Fei. Amo sampler: Enhancing text rendering with overshooting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13157–13166, 2025. 4

[39] Yingsong Huang, Hui Guo, Jing Huang, Bing Bai, and Qi Xiong. Diffusion epistemic uncertainty with asymmetric learning for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17097–17107, 2025. 3

[40] Zhenglin Huang, Jinwei Hu, Xiangtai Li, Yiwei He, Xingyu Zhao, Bei Peng, Baoyuan Wu, Xiaowei Huang, and Guangliang Cheng. Sida: Social media image deepfake detection, localization and explanation with large multimodal model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28831–28841, 2025. 2, 3

[41] Inc. Ideogram. Ideogram. https://ideogram.ai/, 2025. 5

[42] Jiabao Ji, Guanhua Zhang, Zhaowen Wang, Bairu Hou, Zhifei Zhang, Brian Price, and Shiyu Chang. Improving diffusion models for scene text editing with dual encoders. *arXiv preprint arXiv:2304.05568*, 2023. 4

[43] Zexi Jia, Chuanwei Huang, Yeshuang Zhu, Hongyan Fei, Xiaoyue Duan, Zhiqiang Yuan, Ying Deng, Jiapei Zhang, Jinchao Zhang, and Jie Zhou. Secret lies in color: Enhancing ai-generated images detection with color distribution analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13445–13454, 2025. 3

[44] Hengrui Kang, Siwei Wen, Zichen Wen, Junyan Ye, Weijia Li, Peilin Feng, Baichuan Zhou, Bin Wang, Dahua Lin, Linfeng Zhang, and Conghui He. Legion: Learning to ground and explain for synthetic image detection, 2025.

[45] Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. Any-resolution ai-generated image detection by spectral learning, 2025. 3

[46] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *2015 13th international conference on document analysis and recognition (ICDAR)*, pages 1156–1160. IEEE, 2015. 7

[47] Sohail Ahmed Khan and Duc-Tien Dang-Nguyen. Clipping the deception: Adapting vision-language models for universal deepfake detection, 2024. 3

[48] Christos Koutlis and Symeon Papadopoulos. Leveraging representations from intermediate encoder-blocks for synthetic image detection, 2024. 3

[49] Rui Lan, Yancheng Bai, Xu Duan, Mingxing Li, Dongyang Jin, Ryan Xu, Lei Sun, and Xiangxiang Chu. Flux-text: A simple and advanced diffusion transformer baseline for scene text editing. *arXiv preprint arXiv:2505.03329*, 2025. 4

[50] Hyeonsu Lee and Chankyu Choi. The surprisingly straightforward scene text removal method with gated attention and region of interest generation: A comprehensive prominent model analysis. In *European Conference on Computer Vision*, pages 457–472. Springer, 2022. 3

[51] Chunxiao Li, Xiaoxiao Wang, Meiling Li, Boming Miao, Peng Sun, Yunjian Zhang, Xiangyang Ji, and Yao Zhu. Bridging the gap between ideal and real-world evaluation:

Benchmarking ai-generated image detection in challenging scenarios, 2025. 3

[52] Mingsong Li, Lin Liu, Hongjun Wang, Haoxing Chen, Xijun Gu, Shizhan Liu, Dong Gong, Junbo Zhao, Zhenzhong Lan, and Jianguo Li. Multiedit: Advancing instruction-based image editing on diverse and challenging tasks. *arXiv preprint arXiv:2509.14638*, 2025. 4

[53] Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Fuli Feng. Improving synthetic image detection towards generalization: An image transformation perspective, 2025. 3

[54] Songze Li, Yunfei Guo, Shen Chen, Bin Li, Kaiqing Lin, Changsheng Chen, Haodong Li, Taiping Yao, and Shouhong Ding. Ditl2: Dual-stage invariance transfer learning for generalizable document image tampering localization. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 82–91, 2025. 3

[55] Ziqiang Li, Jiazhen Yan, Ziwen He, Kai Zeng, Weiwei Jiang, Lizhi Xiong, and Zhangjie Fu. Is artificial intelligence generated image detection a solved problem? *arXiv preprint arXiv:2505.12335*, 2025. 3

[56] Shuqiao Liang, Jian Liu, Renzhang Chen, and Quanlong Guan. Ferretnet: Efficient synthetic image detection via local pixel dependencies. *arXiv preprint arXiv:2509.20890*, 2025.

[57] Yachao Liang, Min Yu, Gang Li, Jianguo Jiang, Fuqiang Du, Li Jingyuan, Lanchi Xie, Zhen Xu, and Weiqing Huang. Denoising trajectory biases for zero-shot ai-generated image detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 3

[58] Chongyu Liu, Dezhi Peng, Yuliang Liu, and Lianwen Jin. Ctrnet++: Dual-path learning with local-global context modeling for scene text removal. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(1):1–22, 2024. 4

[59] Huan Liu, Zichang Tan, Chuangchuang Tan, Yunchao Wei, Yao Zhao, and Jingdong Wang. Forgery-aware adaptive transformer for generalizable synthetic image detection, 2023. 3

[60] Zhengzhe Liu, Xiaojuan Qi, and Philip Torr. Global texture enhancement for fake face detection in the wild, 2020. 3

[61] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6, 7

[62] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 6, 7

[63] Zeyu Liu, Weicong Liang, Zhanhao Liang, Chong Luo, Ji Li, Gao Huang, and Yuhui Yuan. Glyph-byt5: A customized text encoder for accurate visual text rendering, 2024. 4

[64] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[65] Dongliang Luo, Yuliang Liu, Rui Yang, Xianjin Liu, Jishen Zeng, Yu Zhou, and Xiang Bai. Toward real text manipulation detection: New dataset and new solution. *Pattern Recognition*, 157:110828, 2025. 3

[66] Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. Lareˆ2: Latent reconstruction error based method for diffusion-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17006–17015, 2024. 3

[67] Xiaochen Ma, Xuekang Zhu, Lei Su, Bo Du, Zhuohang Jiang, Bingkui Tong, Zeyu Lei, Xinyu Yang, Chi-Man Pun, Jiancheng Lv, et al. Imdl-benco: A comprehensive benchmark and codebase for image manipulation detection & localization. *Advances in Neural Information Processing Systems*, 37:134591–134613, 2025. 3

[68] Tai D. Nguyen, Aref Azizpour, and Matthew C. Stamm. Forensic self-descriptions are all you need for zero-shot detection, open-set source attribution, and clustering of ai-generated images, 2025. 3

[69] Konstantina Nikolaidou, George Retsinas, Vincent Christlein, Mathias Seuret, Giorgos Sfikas, Elisa Barney Smith, Hamam Mokayed, and Marcus Liwicki. Wordstylist: styled verbatim handwritten text generation with latent diffusion models. In *International Conference on Document Analysis and Recognition*, pages 384–401. Springer, 2023. 4

[70] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee. Towards universal fake image detectors that generalize across generative models, 2024. 3

[71] OpenAI. Gpt-4o-image-1. https://openai.com/index/image-generation-api/, 2025. 5

[72] OpenAI. Sora. https://sora.com, 2025. 5

[73] Jeongsoo Park and Andrew Owens. Community forensics: Using thousands of generators to train fake image detectors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8245–8257, 2025. 6

[74] Dezhi Peng, Chongyu Liu, Yuliang Liu, and Lianwen Jin. Viteraser: Harnessing the power of vision transformers for scene text removal with segmim pretraining. *arXiv preprint arXiv:2306.12106*, 2023. 4

[75] Yuyang Peng, Shishi Xiao, Keming Wu, Qisheng Liao, Bohan Chen, Kevin Lin, Danqing Huang, Ji Li, and Yuhui Yuan. Bizgen: Advancing article-level visual text rendering for infographics generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23615–23624, 2025. 4

[76] Chenfan Qu, Chongyu Liu, Yuliang Liu, Xinhong Chen, Dezhi Peng, Fengjun Guo, and Lianwen 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*, pages 5937–5946, 2023. 3

[77] Chenfan Qu, Jian Liu, Haoxing Chen, Baihan Yu, Jingjing Liu, Weiqiang Wang, and Lianwen Jin. Textsleuth: Towards explainable tampered text detection. *arXiv preprint arXiv:2412.14816*, 2024. 3

[78] Chenfan Qu, Yiwu Zhong, Fengjun Guo, and Lianwen Jin. Omni-iml: towards unified image manipulation localization. *arXiv preprint arXiv:2411.14823*, 2024.

[79] Chenfan Qu, Yiwu Zhong, Chongyu Liu, Guitao Xu, Dezhi Peng, Fengjun Guo, and Lianwen Jin. Towards modern image manipulation localization: A large-scale dataset and novel methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10781–10790, 2024. 3

[80] Chenfan Qu, Yiwu Zhong, Fengjun Guo, and Lianwen Jin. Revisiting tampered scene text detection in the era of generative ai. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 694–702, 2025. 2, 3, 4, 7, 8

[81] Chenfan Qu, Yiwu Zhong, Huiguo He, Bin Li, and Lianwen Jin. Webly-supervised image manipulation localization via category-aware auto-annotation. *arXiv preprint arXiv:2508.20987*, 2025. 3

[82] Chenfan Qu, Yiwu Zhong, Jian Liu, Xuekang Zhu, Bohan Yu, and Lianwen Jin. Textshield-r1: Reinforced reasoning for tampered text detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8621–8629, 2026. 3

[83] Yadong Qu, Qingfeng Tan, Hongtao Xie, Jianjun Xu, Yuxin Wang, and Yongdong Zhang. Exploring stroke-level modifications for scene text editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2119–2127, 2023. 4

[84] Anirudh Sundara Rajan, Utkarsh Ojha, Jedidiah Schloesser, and Yong Jae Lee. Aligned datasets improve detection of latent diffusion-generated images. *arXiv preprint arXiv:2410.11835*, 2024. 3

[85] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 8

[86] Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9130–9140, 2024. 3

[87] Prasun Roy, Saumik Bhattacharya, Subhankar Ghosh, and Umapada Pal. Stefann: Scene text editor using font adaptive neural network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[88] Ayush Sarkar, Hanlin Mai, Amitabh Mahapatra, Svetlana Lazebnik, D. A. Forsyth, and Anand Bhattad. Shadows don't lie and lines can't bend! generative models don't know projective geometry...for now, 2024. 3

[89] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models, 2023. 3

[90] Huiru Shao, Zhuang Qian, Kaizhu Huang, Wei Wang, Xiaowei Huang, and Qiufeng Wang. Delving into adversarial robustness on document tampering localization. In *European Conference on Computer Vision*, pages 290–306. Springer, 2024. 3

[91] Wataru Shimoda, Daichi Haraguchi, Seiichi Uchida, and Kota Yamaguchi. De-rendering stylized texts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1076–1085, 2021. 4

[92] Wataru Shimoda, Naoto Inoue, Daichi Haraguchi, Hayato Mitani, Seiichi Uchida, and Kota Yamaguchi. Type-r: Automatically retouching typos for text-to-image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2745–2754, 2025. 4

[93] Yalin Song, Wenbin Jiang, Xiuli Chai, Zhihua Gan, Mengyuan Zhou, and Lei Chen. Cross-attention based two-branch networks for document image forgery localization in the metaverse. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(2): 1–24, 2025. 3

[94] Google AI Studio. Gemini 2.5 flash image (nano banana), 2025. 4

[95] Lei Su, Xiaochen Ma, Xuekang Zhu, Chaoqun Niu, Zeyu Lei, and Ji-Zhe Zhou. Can we get rid of handcrafted feature extractors? sparsevit: Nonsemantics-centered, parameter-efficient image manipulation localization through spare-coding transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7024–7032, 2025. 3

[96] Yuchen Su, Zhineng Chen, Zhiwen Shao, Yuning Du, Zhilong Ji, Jinfeng Bai, Yong Zhou, and Yu-Gang Jiang. Lranet: Towards accurate and efficient scene text detection with low-rank approximation network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4979–4987, 2024. 7

[97] Yuchen Su, Zhineng Chen, Yongkun Du, Zhilong Ji, Kai Hu, Jinfeng Bai, and Xieping Gao. Explicit relational reasoning network for scene text detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7069–7077, 2025. 7

[98] Jeyasri Subramanian, Varnith Chordia, Eugene Bart, Shaobo Fang, Kelly Guan, Raja Bala, et al. Strive: Scene text replacement in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14549–14558, 2021. 4

[99] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12105–12114, 2023. 2, 3

[100] Chuangchuang Tan, Renshuai Tao, Huan Liu, Guanghua Gu, Baoyuan Wu, Yao Zhao, and Yunchao Wei. C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection, 2024.

[101] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28130–28139, 2024.

[102] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake

detection: Improving generalizability through frequency space domain learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(5):5052–5060, 2024. 3

[103] Camscanner Team. Camscanner. https://www.camscanner.com/, 2025. 5

[104] Meitu Team. Meitu-xiuxiu. https://pc.meitu.com/, 2025. 5

[105] Quark Team. Quark scanner. https://www.quark.cn/, 2025. 5

[106] WPS Team. Wps image. https://www.wps.cn/, 2025. 5

[107] Kuaishou Technology. Kling ai: Next-generation ai creative studio. https://klingai.com/, 2025. 4

[108] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. *arXiv preprint arXiv:2311.03054*, 2023. 4

[109] Yuxiang Tuo, Yifeng Geng, and Liefeng Bo. Anytext2: Visual text generation and editing with customizable attributes. *arXiv preprint arXiv:2411.15245*, 2024. 4

[110] Hongsong Wang, Renxi Cheng, Yang Zhang, Chaolei Han, and Jie Gui. Lota: Bit-planes guided ai-generated image detection, 2025. 3

[111] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A. Efros. Cnn-generated images are surprisingly easy to spot... for now, 2020. 3

[112] Wenhai Wang, Enze Xie, Xiang Li, Wenbo Hou, Tong Lu, Gang Yu, and Shuai Shao. Shape robust text detection with progressive scale expansion network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9336–9345, 2019. 8

[113] Xi Wang and Vicky Kalogeiton. Your diffusion model is an implicit synthetic image detector. In *European Conference on Computer Vision*, pages 418–434. Springer, 2024. 3

[114] Xiaobing Wang, Yingying Jiang, Zhenbo Luo, Cheng-Lin Liu, Hyunsoo Choi, and Sungjin Kim. Arbitrary shape scene text detection with adaptive text region representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6449–6458, 2019. 8

[115] Yuxin Wang, Hongtao Xie, Zheng-Jun Zha, Mengting Xing, Zilong Fu, and Yongdong Zhang. Contournet: Taking a further step toward accurate arbitrary-shaped scene text detection. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11753–11762, 2020. 8

[116] Yuxin Wang, Hongtao Xie, Shancheng Fang, Yadong Qu, and Yongdong Zhang. Pert: a progressively region-based network for scene text removal. *arXiv preprint arXiv:2106.13029*, 2021. 3

[117] Yuxin Wang, Hongtao Xie, Mengting Xing, Jing Wang, Shenggao Zhu, and Yongdong Zhang. Detecting tampered scene text in the wild. In *European Conference on Computer Vision*, pages 215–232. Springer, 2022. 2, 3, 4, 7, 8

[118] Yuxin Wang, Boqiang Zhang, Hongtao Xie, and Yongdong Zhang. Tampered text detection via rgb and frequency relationship modeling. *Chinese Journal of Network and Information Security*, 8(3):29–40, 2022. 2, 3, 4, 7, 8

[119] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection, 2023. 3

[120] Zixiao Wang, Hongtao Xie, YuXin Wang, Yadong Qu, Fengjun Guo, and Pengwei Liu. Leveraging text localization for scene text removal via text-aware masked image modeling. In *European Conference on Computer Vision*, pages 357–373. Springer, 2024. 4

[121] Siwei Wen, Junyan Ye, Peilin Feng, Hengrui Kang, Zichen Wen, Yize Chen, Jiang Wu, Wenjun Wu, Conghui He, and Weijia Li. Spot the fake: Large multimodal model-based synthetic image detection with artifact explanation. *arXiv preprint arXiv:2503.14905*, 2025. 3

[122] Kahim Wong, Jicheng Zhou, Haiwei Wu, Yain-Whar Si, and Jiantao Zhou. 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 (ICCV)*, pages 19280–19289, 2025. 3

[123] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 3, 4

[124] Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. Omnigen2: Exploration to advanced multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025. 4

[125] Liang Wu, Chengquan Zhang, Jiaming Liu, Junyu Han, Jingtuo Liu, Errui Ding, and Xiang Bai. Editing text in the wild. In *Proceedings of the 27th ACM International Conference on Multimedia*, page 1500–1508, New York, NY, USA, 2019. Association for Computing Machinery. 4

[126] Yu Xie, Jielei Zhang, Pengyu Chen, Ziyue Wang, Weihang Wang, Longwen Gao, Peiyi Li, Huyang Sun, Qiang Zhang, Qian Qiao, et al. Textflux: An ocr-free dit model for high-fidelity multilingual scene text synthesis. *arXiv preprint arXiv:2505.17778*, 2025. 4

[127] Shilin Yan, Ouxiang Li, Jiayin Cai, Yanbin Hao, Xiaolong Jiang, Yao Hu, and Weidi Xie. A sanity check for ai-generated image detection, 2025. 3

[128] Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. Glyphcontrol: Glyph conditional control for visual text generation. *Advances in Neural Information Processing Systems*, 36: 44050–44066, 2023. 3

[129] Junyan Ye, Baichuan Zhou, Zilong Huang, Junan Zhang, Tianyi Bai, Hengrui Kang, Jun He, Honglin Lin, Zihao Wang, Tong Wu, Zhizheng Wu, Yiping Chen, Dahua Lin, Conghui He, and Weijia Li. Loki: A comprehensive synthetic data detection benchmark using large multimodal models, 2025. 3

[130] Maoyuan Ye, Jing Zhang, Shanshan Zhao, Juhua Liu, Bo Du, and Dacheng Tao. Dptext-detr: Towards better scene text detection with dynamic points in transformer. In *Pro-*

ceedings of the AAAI conference on artificial intelligence, pages 3241–3249, 2023. 7

[131] Jiaruo Yu, Dagong Lu, Xingyue Shi, Chenfan Qu, and Fengjun Guo. Unified face attack detection with micro disturbance and a two-stage training strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 960–969, 2024. 3

[132] Zeqin Yu, Haotao Xie, Jian Zhang, Jiangqun Ni, Wenkang Su, and Jiwu Huang. Toward real-world text image forgery localization: Structured and interpretable data synthesis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 7

[133] Lin Yuan, Xiaowan Li, Yan Zhang, Jiawei Zhang, Hongbo Li, and Xinbo Gao. Mlep: Multi-granularity local entropy patterns for generalized ai-generated image detection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. 3

[134] Weichao Zeng, Yan Shu, Zhenhang Li, Dongbao Yang, and Yu Zhou. Textctrl: Diffusion-based scene text editing with prior guidance control. *Advances in Neural Information Processing Systems*, 37:138569–138594, 2024. 4

[135] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022. 7

[136] Haifeng Zhang, Qinghui He, Xiuli Bi, Weisheng Li, Bo Liu, and Bin Xiao. Towards universal ai-generated image detection by variational information bottleneck network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23828–23837, 2025. 3

[137] Yanran Zhang, Bingyao Yu, Yu Zheng, Wenzhao Zheng, Yueqi Duan, Lei Chen, Jie Zhou, and Jiwen Lu. $\mathbf{D^3}$qe: Learning discrete distribution discrepancy-aware quantization error for autoregressive-generated image detection, 2025. 3

[138] Yiming Zhao and Zhouhui Lian. Udifftext: A unified framework for high-quality text synthesis in arbitrary images via character-aware diffusion models. *arXiv preprint arXiv:2312.04884*, 2023. 4

[139] Zhen Zhao, Jingqun Tang, Binghong Wu, Chunhui Lin, Shu Wei, Hao Liu, Xin Tan, Zhizhong Zhang, Can Huang, and Yuan Xie. Harmonizing visual text comprehension and generation. *Advances in Neural Information Processing Systems*, 37:97499–97522, 2024. 3

[140] Chende Zheng, Chenhao Lin, Zhengyu Zhao, Hang Wang, Xu Guo, Shuai Liu, and Chao Shen. Breaking semantic artifacts for generalized ai-generated image detection. *Advances in Neural Information Processing Systems*, 37: 59570–59596, 2024. 3

[141] Nan Zhong, Haoyu Chen, Yiran Xu, Zhenxing Qian, and Xinpeng Zhang. Beyond generation: A diffusion-based low-level feature extractor for detecting ai-generated images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8258–8268, 2025. 3

[142] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560, 2017. 8

[143] Ziyin Zhou, Yunpeng Luo, Yuanchen Wu, Ke Sun, Jiayi Ji, Ke Yan, Shouhong Ding, Xiaoshuai Sun, Yunsheng Wu, and Rongrong Ji. Aigi-holmes: Towards explainable and generalizable ai-generated image detection via multimodal large language models, 2025. 3

[144] Xuekang Zhu, Xiaochen Ma, Lei Su, Zhuohang Jiang, Bo Du, Xiwen Wang, Zeyu Lei, Wentao Feng, Chi-Man Pun, and Ji-Zhe Zhou. Mesoscopic insights: Orchestrating multi-scale & hybrid architecture for image manipulation localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11022–11030, 2025. 3

[145] Wanyi Zhuang, Qi Chu, Tao Gong, Changtao Miao, and Nenghai Yu. Towards good generalizations for diffusion generated image detection using multiple reconstruction contrastive learning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, page 5431–5440, New York, NY, USA, 2025. Association for Computing Machinery. 3