

Zooming In on Fakes: A Novel Dataset for Localized AI-Generated Image Detection with Forgery Amplification Approach

Lypan Cai^{1*}, Haowei Wang^{2*}, Jiayi Ji^{1†}, Yanshu Zhoumen¹,
Shen Chen², Taiping Yao², Xiaoshuai Sun^{1†}

¹Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, School of Informatics, Xiamen University, Xiamen 361005, China

²Youtu Lab, Tencent, China

Abstract

The rise of AI-generated image tools has made localized forgeries increasingly realistic, posing challenges for visual content integrity. Although recent efforts have explored localized AIGC detection, existing datasets predominantly focus on object-level forgeries while overlooking broader scene edits in regions such as sky or ground. To address these limitations, we introduce **BR-Gen**, a large-scale dataset of 150,000 locally forged images with diverse scene-aware annotations, which are based on semantic calibration to ensure high-quality samples. BR-Gen is constructed through a fully automated “Perception-Creation-Evaluation” pipeline to ensure semantic coherence and visual realism. In addition, we further propose **NFA-ViT**, a Noise-guided Forgery Amplification Vision Transformer that enhances the detection of localized forgeries by amplifying subtle forgery-related features across the entire image. NFA-ViT mines heterogeneous regions in images, *i.e.*, potential edited areas, by noise fingerprints. Subsequently, attention mechanism is introduced to compel the interaction between normal and abnormal features, thereby propagating the traces throughout the entire image, allowing subtle forgeries to influence a broader context and improving overall detection robustness. Extensive experiments demonstrate that BR-Gen constructs entirely new scenarios that are not covered by existing methods. Take a step further, NFA-ViT outperforms existing methods on BR-Gen and generalizes well across current benchmarks.

Code — <https://github.com/clpbc/BR-Gen>.

Introduction

The rapid advancement of generative models, such as Generative Adversarial Networks (GANs) (Karras et al. 2018a) and Diffusion Models (DMs) (Dang et al. 2025), enable fine-grained image modifications through simple user interactions like masks, and prompts (Ju et al. 2024). These techniques raise serious concerns about image authenticity (Ferreira et al. 2020), especially on social media platforms, while democratizing the creation of creative content. Consequently, the ability to detect whether visual content has been altered is becoming increasingly critical.

*These authors contributed equally.

†Corresponding authors.

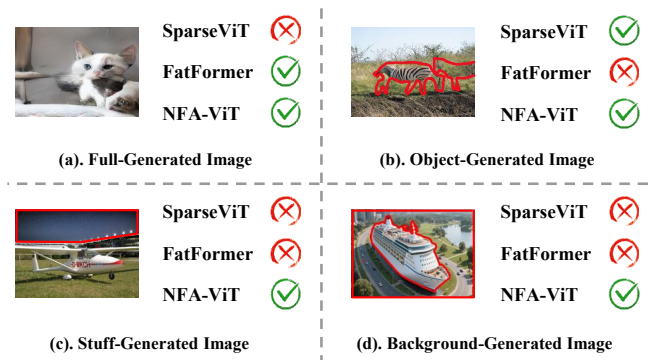


Figure 1: Comparison of four forgery scenarios in existing datasets. They mainly cover full-generated images and object-level forgeries, while forgeries in stuff and background regions remain largely unaddressed. Red regions show ground-truth forgeries. State-of-the-art models (FatFormer (Liu et al. 2024) and SparseViT (Su et al. 2025)) struggle with these new cases. Our proposed NFA-ViT achieves robust detection across all four scenarios. The source data comes from open-source datasets GRE (Sun et al. 2024), COCO (Lin et al. 2014), and ImageNet (Deng et al. 2009).

Around the AI-generated content (AIGC), a few early researches construct various datasets (Wang et al. 2023; Zhu et al. 2023) and benchmarks (Wang et al. 2020; Ojha et al. 2023) to improve the detection performance on fully synthesized images. However, these efforts often overlook localized generation scenarios, where only specific regions are modified. Although some recent works (Guillaro et al. 2023; He et al. 2025) have attempted to localize manipulations, their progress is constrained by limitations in existing datasets and detection paradigm.

Specifically, current localized AIGC datasets (Guillaro et al. 2023) suffer from two major limitations. (1) **Pervasive forgery region bias**. Previous datasets focus on salient objects or synthetic rectangular patches while defining forged areas, neglecting complex scene-level elements like sky, ground, vegetation, or structural background. Detectors trained on such data tend to overfit to object-centric artifacts and fail to generalize to more subtle or spatially

Dataset	Dataset Scale		Gen. Category		Mask Type		Gen. Area
	Real Images	Gen. Images	GAN-based	DM-based	Stuff	Background	
NIST16 (Guan et al. 2019)	0	564	1	-	✗	✗	Small
DEFACTO (Mahfoudi et al. 2019)	-	149,000	1	-	✗	✗	Small
IMD2020 (Novozamsky, Mahdian, and Saic 2020)	35,000	35,000	1	-	✗	✗	Small
CocoGLIDE (Guillaro et al. 2023)	512	512	-	1	✗	✗	Small
AutoSplice (Jia et al. 2023)	2,273	3,621	-	1	✗	✗	Small/Medium
TGIF (Mareen et al. 2024)	3,124	74,976	-	2	✗	✗	Small
GRE (Sun et al. 2024)	-	228,650	2	3	✗	✗	Small
SID-Set (Huang et al. 2025)	70,000	140,000	-	2	✗	✗	Small/Full
BR-Gen (Ours)	15,000	150,000	2	3	✓	✓	Small/Medium/Large

Table 1: Summary of the attributes for localized AIGC detection datasets. Numbers indicate category quantities, and “Area” denotes the distribution range of forged regions.

distributed forgeries. (2) **Uncontrollable editing quality**, which further limits their effectiveness. Many generated samples exhibit unrealistic textures, compression artifacts, or visible boundary seams due to low-quality generation pipelines and a lack of quality control. These flaws not only reduce visual plausibility but also make detection easier, masking the true difficulty of localized forgery detection in real-world scenarios. Fig. 1 illustrates how data limitations affect model detection. The model fails to detect when faced with out-of-distribution data or complex scene-level elements.

To address these challenges, we present the Broader Region Generation (**BR-Gen**) dataset, a large-scale and high-quality benchmark containing 150,000 locally forged images with diverse region coverage. BR-Gen targets underrepresented “stuff” and “background” categories—including sky, ground, wall, grass, and vegetation—substantially broadening the scope of localized forgeries beyond objects. The dataset is constructed through a fully automated “Perception-Creation-Evaluation” pipeline that ensures semantic integrity and visual realism. Specifically, we use grounding and segmentation models to guide localized editing (Ravi et al. 2024; Liu et al. 2023), diffusion-based generative models for content synthesis (Podell et al. 2023; Zhuang et al. 2024; Wang et al. 2025), and multi-stage perceptual evaluation metrics (Fu et al. 2023; Radford et al. 2021) to validate image quality. Compared to prior datasets, BR-Gen offers broader region diversity, more realistic forgeries, and stronger alignment with real-world editing patterns.

Despite recent advances (Su et al. 2025; Guillaro et al. 2023; Liu et al. 2022) in localized forgery detection, existing methods exhibit two major limitations when evaluated on BR-Gen. First, they tend to detect generic discrepancies between regions without the capability to identify which parts are truly forged. This often leads to overfitting to dataset-specific biases, such as the typical size, shape, or location of forged areas, and results in significant errors when these patterns vary. See Case 3 in Fig.5 for an example. Second, when forgeries are small or embedded in visually inconspicuous regions (e.g., within complex backgrounds or “stuff” areas), the manipulation signals become extremely weak and are easily overshadowed by surrounding authentic content, rendering them difficult to detect. This issue is illustrated in

Case 4 of Fig.5.

To address these challenges, we introduce **NFA-ViT**, a novel architecture designed for robust localized forgery detection. NFA-ViT introduces a *forgery amplification* mechanism that enhances weak manipulation cues through a dual-branch framework. Specifically, a dedicated noise fingerprint branch extracts subtle discrepancies between authentic and forged regions, while a visual transformer backbone incorporates these signals via noise-guided attention queries. This allows authentic regions to absorb and propagate discriminative forgery features across the image, thereby amplifying weak signals and improving detection sensitivity, particularly in sparse or spatially diffuse cases. Moreover, when forgery cues are diffused across the entire image, the classifier is no longer constrained to detecting discrepancies between isolated regions. Instead, it learns to distinguish globally between authentic and manipulated content based on the presence or absence of generative attributes. This adaptive learning paradigm significantly improves the model’s ability to localize forgeries, regardless of their size, shape, or location. Extensive experiments on BR-Gen and other benchmarks demonstrate that NFA-ViT not only achieves state-of-the-art performance but also exhibits strong cross-dataset generalization, establishing a new foundation for scene-aware, fine-grained forgery detection.

In summary, our contributions are three-fold:

- We identify key limitations in existing localized AIGC datasets, including region bias and low visual quality, and introduce BR-Gen, a large-scale benchmark with diverse and realistic scene-level forgeries.
- We propose NFA-ViT, a noise-guided forgery amplification transformer that leverages a dual-branch architecture to diffuse forgery cues into real regions through modulated self-attention, significantly improving the detectability of small or spatially subtle forgeries.
- We conduct extensive experiments showing that BR-Gen is a more challenging datasets, and that NFA-ViT achieves strong detection and localization performance.

Related Work

Generation Datasets

In recent years, Artificial Intelligence Generated Content (AIGC) has become widely used. Datasets (Wang et al.

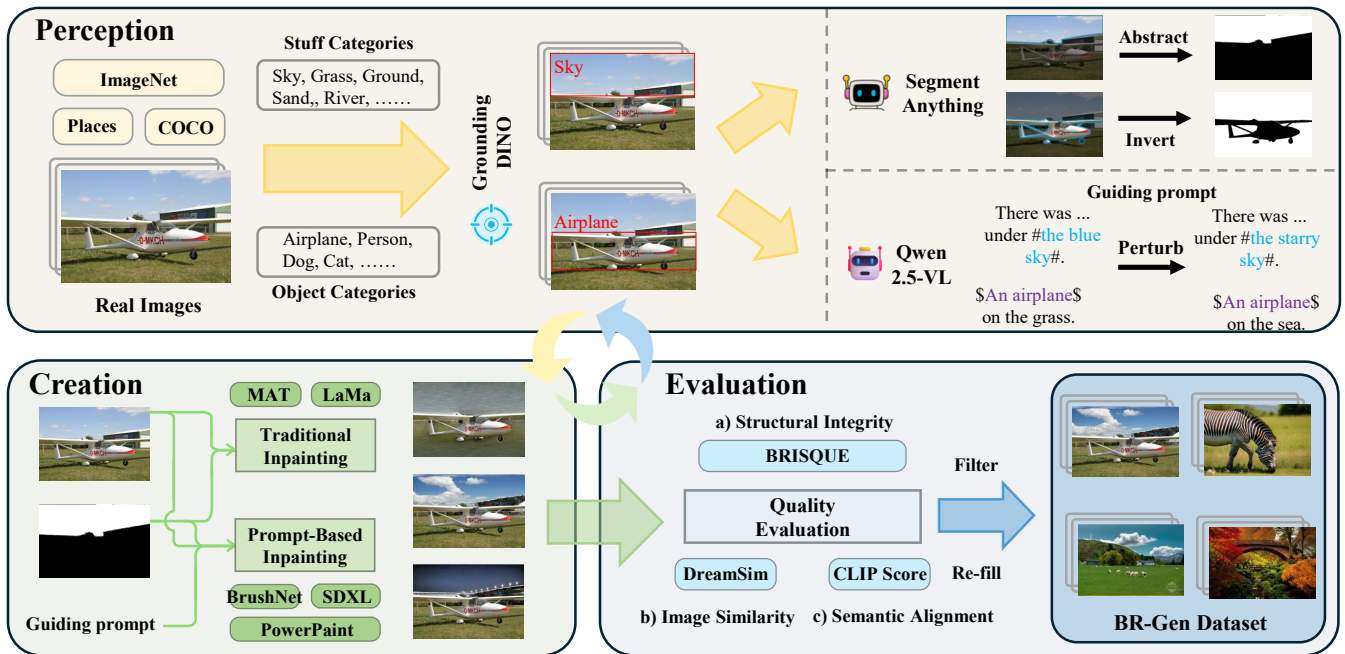


Figure 2: The automated pipeline for the BR-Gen dataset consists of three iterative stages: **Perception**, **Creation**, and **Evaluation**. These stages are applied to produce high-quality localized generation datasets through progressive refinement. All samples are sourced from publicly available datasets (Zhou et al. 2017; Lin et al. 2014; Deng et al. 2009).

2020; Ojha et al. 2023) containing both real and generated images have been organized for training and evaluating detection systems. CNNSpot (Wang et al. 2020) uses GAN-generated images (Karras et al. 2018a; Chen et al. 2024), and Chameleon (Yan et al. 2025) provides highly realistic test cases. However, these support only image-level tasks. For localized detection, existing datasets (Tab. 1) often use object masks from COCO (Lin et al. 2014) or SAM (Kirillov et al. 2023), focusing on countable objects. This ignores large regions like sky or background, leading to bias and poor model generalization on such areas.

Generation Detection

The need for detecting AI-generated content has existed since the rise of deep learning. Early detection used spatial cues like color and reflection (McCloskey and Albright 2018; O’Brien and Farid 2012). As GANs improved, methods like CNNSpot (Wang et al. 2020) used data augmentation for better generalization (Karras et al. 2018b; Chen et al. 2025; Yang et al. 2025; Lin et al. 2025). Recent works exploit frequency or reconstruction patterns (Wang et al. 2023; Qian et al. 2020), but still focus on whole images. For localization, ManTra-Net (Wu et al. 2019) uses LSTM analysis, Trufor (Guillaro et al. 2023) uses noise patterns, and SparseViT (Su et al. 2025) uses sparse attention for top performance. Yet, they fail on complex or non-object regions, highlighting the need for our approach.

BR-Gen Dataset

Recent datasets for detecting localized forgery based on generative models (Sun et al. 2024; Guillaro et al. 2023) have emerged. To address the gaps in existing datasets, we have taken into account the neglected local edits in Stuff and Background, proposing a high-quality, scene-based local generation dataset named the Broader Region Generation (**BR-Gen**). We propose an automated pipeline with open-source models (Ravi et al. 2024; Liu et al. 2023; Bai et al. 2025), generating local edited images from unannotated ones. As shown in Tab. 1, BR-Gen takes into full account diverse generation methods and tampering areas, addressing the shortcomings in the types of previous datasets.

Real Image Collection

We sampled images from three large-scale datasets like previous works (Jia et al. 2023): ImageNet (Deng et al. 2009), COCO (Lin et al. 2014), and Places (Zhou et al. 2017). These datasets provide diverse scenes with rich semantic content, enhancing the diversity of dimensions.

Localized Generation Pipeline

To simulate real-world image editing processes while maintaining content and semantic consistency, we designed an automated pipeline that integrates multiple open-source models (Ravi et al. 2024; Liu et al. 2023; Bai et al. 2025), as illustrated in Fig. 2. The pipeline comprises three stages: **“Perception-Creation-Evaluation”**. Below, we briefly describe each stage, with detailed process descriptions and dataset statistics provided in the appendix.

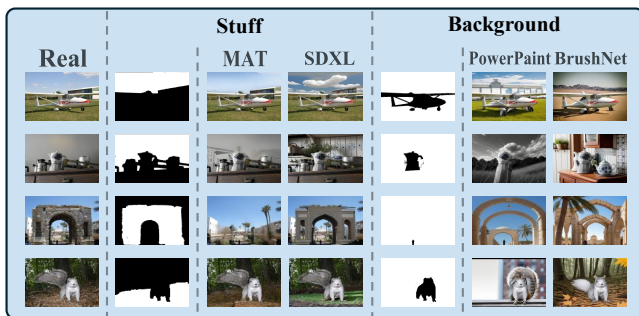


Figure 3: Partial examples from BR-Gen dataset. The real data comes from the open datasets Places (Zhou et al. 2017), COCO (Lin et al. 2014), and ImageNet (Deng et al. 2009).

Preception. This stage identifies regions of interest in real images and extracts semantic information to guide subsequent creation stages. We select the candidate categories by “thing category” and “stuff category” from COCO. GroundingDINO (Liu et al. 2023) detects object bounding boxes for these categories, which are then converted into masks by SAM2 (Ravi et al. 2024). “Stuff” masks are obtained directly, while “background” masks are derived by inverting “thing” masks. To balance category distribution, we control the number of instances per category.

For prompt-based inpainting, we use Qwen2.5-VL (Bai et al. 2025) to generate image descriptions. Then, we apply Semantic Perturbation to increase semantic diversity: for “stuff”, we replace content within “#” symbols (e.g., “the blue sky” → “the starry sky”); for “background”, we replace text outside “\$” symbols. Extra semantically similar object lists ensure plausible edits.

Creation. This stage uses region masks and guiding prompts to generate locally forged images. We employ five diversity state-of-the-art inpainting methods: two GAN-based (LaMa (Suvorov et al. 2022), MAT (Li et al. 2022)) and three diffusion-based (SDXL (Podell et al. 2023), BrushNet (Ju et al. 2024), PowerPaint (Zhuang et al. 2024)). GAN methods require only the image and mask, while diffusion methods also require the guiding prompt.

Evaluation. This stage filters high-quality generated images using image quality assessment methods. Image quality is assessed based on: (a) BRISQUE (Mittal, Moorthy, and Bovik 2012) for structural integrity; (b) DreamSim (Fu et al. 2023) for image similarity; and (c) CLIP scores for prompt alignment. Low-quality samples are removed. Missing data is replenished through iterative generation to maintain dataset size.

Showcase

To visually illustrate the effectiveness of our automated pipeline and the quality of the resulting dataset, we present examples in Fig. 3, covering diverse data sources, mask types, and inpainting methods. Additional detailed visualization samples are provided in the appendix.

Dataset Splits

Following the standard dataset partitioning approach for localized detection, the dataset is randomly divided into training, validation, and test sets using an 8:1:1 ratio. This division is applied to the subset of real images within the dataset. Regardless of the data source (ImageNet, COCO, Places), the partitioning ratios remain consistent. As a result, the training set includes 12,000 real images, while both the validation and test sets each contain 1,500 real images.

To prevent data leakage that could compromise the evaluation of model performance on the dataset, the partitioning of region masks and generated image sets is synchronized with the pre-divided real fig/image set. Thus, each triplet (real image, mask, forged image) is assigned to a single dataset partition, ensuring data integrity.

The proposed NFA-ViT

Overview

To further enhance the performance of local AIGC detection, we propose the Noise-guided Forgery Amplification Vision Transformer (NFA-ViT) to take advantages of the non-homologous between the generated regions and real regions. Noise fingerprints and related features exhibit discrepancies across different regions (Guillaro et al. 2023; Wang et al. 2023). However, related studies suffer from significant limitations, such as cannot determine which region is forged and being insensitive to subtle and tiny-area forgeries. We propose a novel detection mechanism called *forgery amplification*. NFA-ViT leverages noise information as guidance to amplify and diffuse localized forgery features across the entire image, making forgery features more distinguishable while ensuring that the judgment of real images remains unaffected.

Fig. 4(a) illustrates the overall framework. For an input RGB image x , we first use the noise extractor Noiseprint++ (Guillaro et al. 2023) to extract the noise trace n of the image. We verify the effectiveness of the features and provide detailed information in the appendix. Subsequently, x and n are jointly fed into a dual-branch network. For each stage, the Noise-guided Mask M_i^n from the noise branch is used to guide the learning of the proposed **Noise-guided Amplification Attention** (NAA) in the image branch. With the NAA, real regions directly focus on the differential forgery regions, gradually diffusing forgery features into real regions:

$$P_{l+1}(i, j) = \alpha \cdot P_l(i, j) + \beta \cdot \frac{\sum_{(m, n) \in \mathcal{N}(i, j)} P_l(m, n)}{|\mathcal{N}(i, j)|}, \quad (1)$$

where $P_l(i, j)$ represents the vector of real features at position (i, j) in layer l , and $\mathcal{N}(i, j)$ represents the set of neighboring forgery features. Through layer-by-layer diffusion operation, forgery features expand from local areas to global areas.

Meanwhile, residual connections in the Noise-guided Mask Attention maintain the original image features. Finally, outputs from all stages are fed into a light-weight weighted decoder to generate the final results.

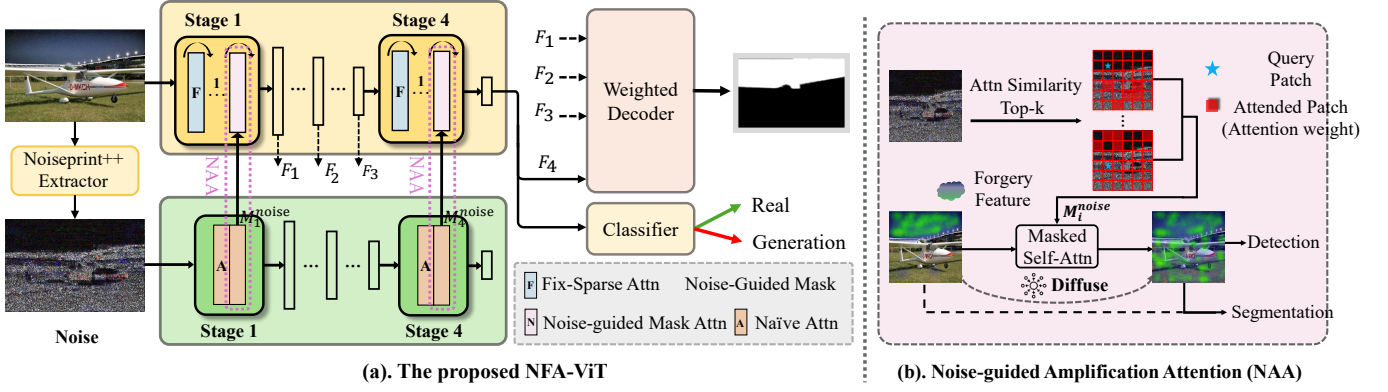


Figure 4: The proposed NFA-ViT framework, which contains dual branches of noise and image, uses noise information to guide the focus area of the image. For the image encoder, a sparse attention mechanism is introduced.

Noise-guided Amplification Attention

For simplicity, we describe the attention workflow for a single attention head. We first introduce Fix-Sparse Attention (Su et al. 2025), which helps refine localization by eliminating irrelevant semantic information. However, Fix-Sparse Attention disrupts semantics from a global perspective to learn non-semantic features, lacking the ability to aggregate and recognize local information.

Based on it, the proposed Noise-guided Amplification Attention (NAA) using noise signals to guide the amplification of forged features in images. Specifically, each stage of the noise branch is composed with vanilla self-attention. In the last layer of the vanilla attention, we take the noise as query matrix Q^{noise} and key matrix K^{noise} to compute the attention matrix A^{noise} as follows:

$$A^{noise} = \text{Softmax}\left(\frac{Q^{noise}K^{noiseT}}{\sqrt{d}}\right). \quad (2)$$

To amplify and diffuse features of forged regions toward real regions, we identify the k most dissimilar K values corresponding to each Q in A^{noise} , forming a Noise-guided Mask M^{noise} , which represents the forged regions corresponding to real regions:

$$M^{noise} = \mathbf{1} [\text{Top-}k(-A^{noise})]. \quad (3)$$

Since the number of heads in corresponding layers of the two branches is identical, it is feasible to use noise information to guide image processing. The mask M^{noise} is then inserted into the last layer in the each stage of image branch. Taking image feature as Q^{image} and K^{image} , the output features F is:

$$F_{ij} = \text{Softmax}\left(\frac{Q^{image}K^{imageT}}{\sqrt{d}}\right)_{ij} \text{ iff } M_{ij}^{noise} = 1, \quad (4)$$

where i and j are the pixel location. In this way, real-region features learn from generated regions, integrating traces of forgery.

Weighted Decoder

Current multi-level feature fusion methods often use addition or concatenation (Lin et al. 2017), producing feature maps through fixed linear aggregation without accounting for the varying contributions of hierarchical features to final maps. To improve region mask prediction, we propose a simple yet efficient decoder design. This approach introduces learnable scaling parameters γ_i ($1 \leq i \leq 4$) for each hierarchical feature map, enabling adaptive weighted fusion by modulating layer-wise contributions.

The decoder processes four hierarchical features F_i ($1 \leq i \leq 4$) extracted from the encoder. Each feature map F_i is first projected uniformly to 512 channels using linear layers. Features $F_{2,3,4}$ are then upscaling to $\frac{1}{4}$ of the original resolution to align with the spatial dimensions of F_1 . Each feature map F_i is scaled by its corresponding parameter γ_i before being summed. The aggregated features are compressed via linear projection to produce \hat{M} , which is subsequently upscaling to the original image resolution for mask prediction \hat{M} . The process is defined as follows:

$$\hat{F}_i = \text{Upscale}(MLP(F_i)), \quad 1 \leq i \leq 4, \quad (5)$$

$$\hat{M} = \text{Upscale}(MLP(\sum_i^4 (\hat{F}_i \times \gamma_i))). \quad (6)$$

This design effectively balances multi-scale features by suppressing irrelevant information and emphasizing critical features via parameterized hierarchical integration.

Loss Function

For detection tasks, we use a lightweight backbone network (He et al. 2016) to extract features F_4 , which generate predictions \hat{y} . For localization tasks, predicted masks \hat{M} from the Weighted Decoder are utilized. Given ground truth labels y and ground truth region masks M , the NFA-ViT model is trained using the following objective function:

$$\mathcal{L} = \mathcal{L}_{\text{cls}}(y, \hat{y}) + \mathcal{L}_{\text{seg}}(M, \hat{M}), \quad (7)$$

where both \mathcal{L}_{cls} and \mathcal{L}_{seg} are binary cross-entropy loss.

Task	Method	Real Recall@50	BR-Gen dataset				Split A		Split B	
			F1	AUC	Recall@50	IoU	GAN R@50	Diffusion R@50	Background R@50	Stuff R@50
Localized Detection	ManTranet (Wu et al. 2019)	0.822	0.123	-	0.069	0.008 (\downarrow 0.133)	0.074	0.061	0.077	0.058
	MVSS-Net (Dong et al. 2022)	0.862	0.183	0.344	0.122	0.029 (\downarrow 0.424)	0.154	0.098	0.154	0.092
	PSCC-Net (Liu et al. 2022)	0.806	0.253	0.284	0.164	0.052 (\downarrow 0.426)	0.166	0.161	0.170	0.155
	Trufor (Guillaro et al. 2023)	0.881	0.295	0.319	0.194	0.048 (\downarrow 0.630)	0.195	0.194	0.199	0.187
	SparseViT (Su et al. 2025)	0.735	0.277	-	0.203	0.049 (\downarrow 0.649)	0.214	0.186	0.205	0.202
AIGC Detection	LGrad (Tan et al. 2023)	0.974	0.165	0.635	0.088	-	0.101 (\downarrow 0.762)	0.057	0.093	0.085
	FreqNet (Tan et al. 2024a)	0.767	0.360	0.472	0.231	-	0.244 (\downarrow 0.671)	0.228	0.240	0.229
	NPR (Tan et al. 2024b)	0.894	<u>0.443</u>	0.501	<u>0.300</u>	-	<u>0.323</u> (\downarrow 0.602)	<u>0.290</u> (\downarrow 0.662)	<u>0.318</u>	0.289
	FatFormer (Liu et al. 2024)	0.989	0.493	<u>0.606</u>	0.331	-	0.358 (\downarrow 0.626)	0.321 (\downarrow 0.629)	0.349	<u>0.310</u>

Table 2: The cross-domain results on BR-Gen. The evaluation methods include AIGC detection and localization detection. We **bold** the best result and mark the second-best result with an underline. The red decline values indicate the level of performance decrease compared to the original dataset (Ma et al. 2025; Liu et al. 2024; Tan et al. 2024b). Since some methods don’t provide corresponding indicators, some values are missing.

Task	Method	Real Recall@50	BR-Gen dataset				Split A		Split B	
			F1	AUC	Recall@50	IoU	GAN R@50	Diffusion R@50	Background R@50	Stuff R@50
Localized Detection	ManTranet (Wu et al. 2019)	0.853	0.774	-	0.760	0.632	0.772	0.760	0.778	0.754
	MVSS-Net (Dong et al. 2022)	0.903	0.892	0.924	0.883	0.671	0.913	0.846	0.889	0.856
	PSCC-Net (Liu et al. 2022)	0.935	0.894	0.937	0.861	0.705	0.898	0.840	0.867	0.844
	Trufor (Guillaro et al. 2023)	0.944	0.918	0.942	0.896	0.779	0.915	0.865	0.903	0.871
	SparseViT (Su et al. 2025)	0.984	0.946	-	0.911	<u>0.824</u>	<u>0.958</u>	0.872	0.931	0.907
AIGC Detection	LGrad (Tan et al. 2023)	0.937	0.831	0.872	0.755	-	0.801	0.732	0.775	0.738
	DIRE (Wang et al. 2023)	0.939	0.823	0.825	0.742	-	0.750	0.744	0.762	0.739
	FreqNet (Tan et al. 2024a)	0.825	0.699	0.702	0.631	-	0.659	0.614	0.648	0.622
	NPR (Tan et al. 2024b)	0.946	0.922	0.933	0.902	-	0.938	0.884	0.921	0.893
	FatFormer (Liu et al. 2024)	<u>0.990</u>	0.961	0.971	0.935	-	0.955	0.913	0.949	0.915
	AIDE (Yan et al. 2025)	0.986	<u>0.964</u>	<u>0.973</u>	<u>0.941</u>	-	0.950	<u>0.932</u>	0.965	<u>0.937</u>
NFA-ViT (ours)		0.992	0.972	0.979	0.953	0.907	0.972	0.941	<u>0.961</u>	0.948

Table 3: The evaluation results of BR-Gen in-domain testing. After training the model on the BR-Gen training set, in-domain evaluation was conducted on the test set.

Experiment

Experimental Setup

Protocols and Evaluation Metrics. We conducted a comprehensive evaluation of the BR-Gen dataset using methods that include AIGC detection and local AIGC detection. First, we assessed the generalization ability of current models on BR-Gen. Next, we performed in-domain testing to evaluate model performance within the BR-Gen. Additionally, we tested the models on existing traditional benchmark. For fairness, the tampered images were downsampled to match the number of the authentic images when training. We used several metrics for a thorough evaluation from detection to localization: (1) Recall@50, classification metrics, which measures the model’s ability to correctly identify categories; (2) F1 and AUC, classification metrics, are used to evaluate the overall performance and stability of the model; (3) IoU, localization metrics, which measures segmentation accuracy at the localization level.

Implementation Details. For NFA-ViT, we use SegFormer (Xie et al. 2021) as the backbone, with the image and noise encoders being the b2 and b0 versions, respectively. During training, the model is optimized using the Adam optimizer (Kinga, Adam et al. 2015), with an initial learning rate of 5×10^{-3} and a weight decay of 1×10^{-6} . Using the Warmup and CosineAnnealing to help models achieve better convergence. In the Noise-Guided Amplification Attention mechanism, the Top- k ratio is set to 25%. Equal weights are given to all parts of the loss function. All experiments are

run for 30 epochs with a batch size of 64.

Experimental Results

Cross-domain on BR-Gen. To evaluate data bias in current generated image detection tasks, we conducted cross-domain testing on the BR-Gen dataset using detection models from two tasks: AIGC detection and local AIGC detection. We directly test the released trained models of those models, which are trained on data (Novozamsky, Mahdian, and Saic 2020; Karras et al. 2018b) that shares the same source as BR-Gen. We divided the BR-Gen’s into “Split A” and “Split B” according generation method and mask sources. Among them, “Split A” is categorized based on the generation method, specifically GAN and diffusion, while “Split B” is categorized based on the source of the masks, *i.e.*, stuff and background. For each model, we also compared its performance drop relative to its original report (Ma et al. 2025; Liu et al. 2024; Tan et al. 2024b), under the corresponding data distribution. The experimental results are shown in Tab. 2.

The results show that local AIGC detection models perform worse overall compared to AIGC detection models. Meanwhile, all methods show a clear drop in generalization performance compared to their original reports. Although these models maintain high recall for real images, they show very low recall for partially generated content, indicating a consistent misclassification of tampered images. In terms of localization ability, the highest IoU value is only

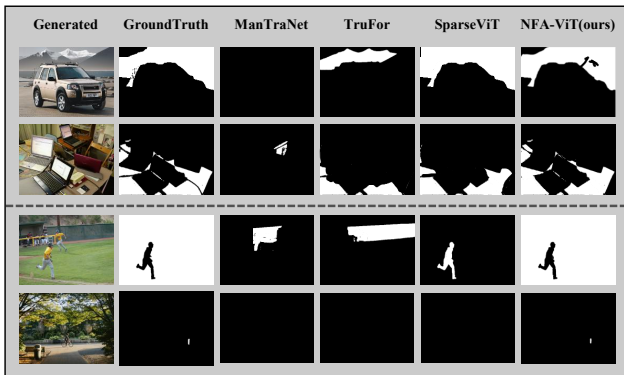


Figure 5: Localization results of different models. We compared images generated by two types of masks. All samples are sourced from publicly available datasets (Zhou et al. 2017; Lin et al. 2014; Deng et al. 2009).

0.052, showing a broad failure to correctly identify tampered regions. These findings confirm the presence of data bias in current tasks and support the improved balance of our dataset, providing a strong base for improving detection performance in this field.

In-domain on BR-Gen. To evaluate performance on BR-Gen, we trained and tested detection models on the same dataset. Results, shown in Tab. 3, indicate clear performance improvements post-training. For detection, AIDE achieved a recall rate of 94.1% for generated images in the dataset, showing the complexity and difficulty of BR-Gen data during in-domain testing. For localization, SparseViT achieved an IoU of 82.4%. A detailed analysis of dataset subtypes showed that in “Split A”, GAN-based detection performed better than Diffusion-based methods. This is partly because GAN-generated images in the dataset were of lower quality and had more visible forgery features. In “Split B”, background types were easier to detect than stuff types, and larger forged areas gave models more useful information.

We carried out the same evaluation on NFA-ViT, and the results showed that NFA-ViT achieved better performance across all metrics. The F1 score reached **0.972**, surpassing AIDE by 0.8%. For localization, the IoU reached **0.907**, outperforming SparseViT by 8.3%, showing the value of amplifying forgery signals. Localization results from several models were visualized in Fig. 5. By effectively addressing regional discrepancies and subtle forgeries, NFA-ViT achieves the best segmentation performance, while other models showed clear gaps.

Ablation Studies

Effectiveness of each component. We conducted structure ablations to systematically evaluate the contribution of each component within NFA-ViT and the results are presented in the Tab. 4. We observed that each proposed component contributes positively to both detection and localization performance: The addition of the Noise branch improved localization by about 4%, showing that noise differences between authentic and forged regions provide valuable cues.

Noise	NAA	Weighted Decoder	Mani R@50	Real R@50	IoU
×	×	×	0.880	0.925	0.759
✓	×	×	0.892	0.938	0.798
✓	✓	×	0.937	0.972	0.833
✓	×	✓	0.913	0.959	0.867
✓	✓	✓	0.953	0.992	0.907

Table 4: Components of the NFA-ViT.

	Gen. R@50	Real R@50	IoU
10%	0.945	0.989	0.887
25%	0.953	0.992	0.907
50%	0.947	0.993	0.897

Table 5: Different value of Top- k in Noise-Guided Amplification Attention.

The inclusion of the NAA module further enhanced detection accuracy, demonstrating its efficacy in amplifying forgery traces across the entire image. And the weighted decoder significantly improved localization accuracy by dynamically adjusting the contribution of features from different layers, thereby refining the final localization output. Ultimately, combining all proposed components achieved the best overall performance.

Ablation of Top- k .

We systematically examined the effects of various Top- k strategies on model detection performance. As shown in Tab. 5, setting k to 25% gave the best performance across multiple metrics, suggesting that this value provides a good balance between accuracy and information retention. The value of k affects the model’s focus area; when k is too small, the model lacks enough information, leading to a drop in performance.

Additional ablation experiment results are provided in the appendix, including more hyperparameter analysis, the impact of different augmentation information and training losses, and the effects of varying tampered area and subclass distributions on model performance.

Conclusion

This paper addresses the limitations of existing AIGC detection datasets, which largely focus on full-generated or object-level forgeries. We introduce BR-Gen, a high-quality dataset with 150,000 locally forged images, covering under-represented stuff and background regions. To better detect subtle and spatially scattered forgeries, we propose NFA-ViT, a noise-guided transformer that amplifies forgery features across the image through attention modulation. Experimental results show that BR-Gen poses significant challenges to current methods, while NFA-ViT achieves strong and consistent performance. Our work provides a new foundation for advancing localized forgery detection in more diverse and realistic settings.

Acknowledgments

This work was supported by National Key R&D Program of China (No.2023YFB4502804), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U22B2051, No. 62302411) and China Postdoctoral Science Foundation (No. 2023M732948).

References

- Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; Zhong, H.; Zhu, Y.; Yang, M.; Li, Z.; Wan, J.; Wang, P.; Ding, W.; Fu, Z.; Xu, Y.; Ye, J.; Zhang, X.; Xie, T.; Cheng, Z.; Zhang, H.; Yang, Z.; Xu, H.; and Lin, J. 2025. Qwen2.5-VL Technical Report. *CoRR*.
- Chen, R.; Wang, Z.; Zhang, K.-Y.; Wu, S.; Sun, J.; Wang, S.; Yao, T.; and Ding, S. 2024. Decoupled data augmentation for improving image classification. *arXiv preprint arXiv:2411.02592*.
- Chen, R.; Xi, J.; Yan, Z.; Zhang, K.-Y.; Wu, S.; Xie, J.; Chen, X.; Xu, L.; Guan, I.; Yao, T.; and Ding, S. 2025. Dual Data Alignment Makes AI-Generated Image Detector Easier Generalizable. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.
- Dang, J.; Chen, L.; Wu, J.; Lin, R.; Wang, B.; Wang, Y.; Wang, L.; Zhu, N.; and Wang, T. 2025. Diff-LMM: Diffusion Teacher-Guided Spatio-Temporal Perception for Video Large Multimodal Models. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*.
- Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. Ieee.
- Dong, C.; Chen, X.; Hu, R.; Cao, J.; and Li, X. 2022. Mvssnet: Multi-view multi-scale supervised networks for image manipulation detection. *TPAMI*.
- Ferreira, W. D.; Ferreira, C. B.; da Cruz Júnior, G.; and Soares, F. 2020. A review of digital image forensics. *Computers & Electrical Engineering*, 85: 106685.
- Fu, S.; Tamir, N.; Sundaram, S.; Chai, L.; Zhang, R.; Dekel, T.; and Isola, P. 2023. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*.
- Guan, H.; Kozak, M.; Robertson, E.; Lee, Y.; Yates, A. N.; Delgado, A.; Zhou, D.; Kheyrkhan, T.; Smith, J.; and Fiscus, J. 2019. MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*.
- Guillaro, F.; Cozzolino, D.; Sud, A.; Dufour, N.; and Verdoliva, L. 2023. 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*.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.
- He, S.; Ji, P.; Yang, Y.; Wang, C.; Ji, J.; Wang, Y.; and Ding, H. 2025. A survey on 3d gaussian splatting applications: Segmentation, editing, and generation. *arXiv preprint arXiv:2508.09977*.
- Huang, Z.; Hu, J.; Li, X.; He, Y.; Zhao, X.; Peng, B.; Wu, B.; Huang, X.; and Cheng, G. 2025. SIDA: Social Media Image Deepfake Detection, Localization and Explanation with Large Multimodal Model.
- Jia, S.; Huang, M.; Zhou, Z.; Ju, Y.; Cai, J.; and Lyu, S. 2023. Autosplice: A text-prompt manipulated image dataset for media forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.
- Ju, X.; Liu, X.; Wang, X.; Bian, Y.; Shan, Y.; and Xu, Q. 2024. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *ECCV*.
- Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018a. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *ICLR*.
- Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018b. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *ICLR*.
- Kinga, D.; Adam, J. B.; et al. 2015. A method for stochastic optimization. In *ICLR*.
- Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.
- Li, W.; Lin, Z.; Zhou, K.; Qi, L.; Wang, Y.; and Jia, J. 2022. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10758–10768.
- Lin, K.; Yan, Z.; Chen, R.; Ye, J.; Zhang, K.-Y.; Zhou, Y.; Jin, P.; Li, B.; Yao, T.; and Ding, S. 2025. Seeing Before Reasoning: A Unified Framework for Generalizable and Explainable Fake Image Detection. *arXiv preprint arXiv:2509.25502*.
- Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.
- Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*.
- Liu, H.; Tan, Z.; Tan, C.; Wei, Y.; Wang, J.; and Zhao, Y. 2024. Forgery-aware adaptive transformer for generalizable synthetic image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*.
- Liu, X.; Liu, Y.; Chen, J.; and Liu, X. 2022. PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*.

- Ma, X.; Zhu, X.; Su, L.; Du, B.; Jiang, Z.; Tong, B.; Lei, Z.; Yang, X.; Pun, C.-M.; Lv, J.; et al. 2025. Imdl-benco: A comprehensive benchmark and codebase for image manipulation detection & localization. *Advances in Neural Information Processing Systems*, 37: 134591–134613.
- Mahfoudi, G.; Tajini, B.; Retraint, F.; Morain-Nicolier, F.; Dugelay, J. L.; and Marc, P. 2019. DEFACTO: image and face manipulation dataset. In *EUSIPCO*.
- Mareen, H.; Karageorgiou, D.; Van Wallendael, G.; Lambert, P.; and Papadopoulos, S. 2024. TGIF: Text-guided inpainting forgery dataset. In *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*.
- McCloskey, S.; and Albright, M. 2018. Detecting gan-generated imagery using color cues. *arXiv preprint arXiv:1812.08247*.
- Mittal, A.; Moorthy, A. K.; and Bovik, A. C. 2012. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Trans. Image Process.*, 21(12): 4695–4708.
- Novozamsky, A.; Mahdian, B.; and Saic, S. 2020. IMD2020: A large-scale annotated dataset tailored for detecting manipulated images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision workshops*.
- O’Brien, J. F.; and Farid, H. 2012. Exposing photo manipulation with inconsistent reflections. *ACM Trans. Graph.*
- Ojha, U.; Li, Y.; Lee, Y. J.; and Ojha, U. 2023. Towards Universal Fake Image Detectors that Generalize Across Generative Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on Learning Representations*.
- Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; and Shao, J. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.
- Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollár, P.; and Feichtenhofer, C. 2024. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*.
- Su, L.; Ma, X.; Zhu, X.; Niu, C.; Lei, Z.; and Zhou, J. 2025. Can We Get Rid of Handcrafted Feature Extractors? SparseViT: Nonsemantics-Centered, Parameter-Efficient Image Manipulation Localization Through Sparse-Coding Transformer. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI-25*.
- Sun, Z.; Fang, H.; Cao, J.; Zhao, X.; and Wang, D. 2024. Rethinking Image Editing Detection in the Era of Generative AI Revolution. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 3538–3547.
- Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*.
- Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; and Wei, Y. 2024a. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; Liu, P.; and Wei, Y. 2024b. 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*, 28130–28139.
- Tan, C.; Zhao, Y.; Wei, S.; Gu, G.; and Wei, Y. 2023. Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection. In *CVPR*.
- Wang, J.; Deng, X.; Wei, W.; Zhang, D.; Wang, S.; Sun, Q.; Zhang, J.; Liu, H.; Xie, N.; and Zhang, M. 2025. Training-free ann-to-snn conversion for high-performance spiking transformer. *arXiv preprint arXiv:2508.07710*.
- Wang, S.-Y.; Wang, O.; Zhang, R.; Owens, A.; and Efros, A. A. 2020. CNN-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Wang, Z.; Bao, J.; Zhou, W.; Wang, W.; Hu, H.; Chen, H.; and Li, H. 2023. Dire for diffusion-generated image detection. In *CVPR*.
- Wu, Y.; Wael, A.; Premkumar, N.; and Wu, Y. 2019. ManTra-Net: Manipulation Tracing Network for Detection and Localization of Image Forgeries With Anomalous Features. In *CVPR*.
- Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*.
- Yan, S.; Li, O.; Cai, J.; Hao, Y.; Jiang, X.; Hu, Y.; and Xie, W. 2025. A Sanity Check for AI-generated Image Detection. In *The Thirteenth International Conference on Learning Representations*.
- Yang, Z.; Chen, R.; Yan, Z.; Zhang, K.-Y.; Fu, X.; Wu, S.; Shu, X.; Yao, T.; Ding, S.; and Li, X. 2025. All Patches Matter, More Patches Better: Enhance AI-Generated Image Detection via Panoptic Patch Learning. *arXiv preprint arXiv:2504.01396*.
- Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- Zhu, M.; Chen, H.; Yan, Q.; Huang, X.; Lin, G.; Li, W.; Tu, Z.; Hu, H.; Hu, J.; and Wang, Y. 2023. Genimage: A million-scale benchmark for detecting ai-generated image. *NeurIPS*.
- Zhuang, J.; Zeng, Y.; Liu, W.; Yuan, C.; and Chen, K. 2024. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. In *ECCV*.