EARG-Net: Edge-Aware Reconstruction-Guided Network for Image Manipulation Detection and Localization

Yanpu Yu[1], Zhaoxin Shi[1], Hanqing Zhao[2], Tianyi Wei[2,✉], Wenbo Zhou[1,✉], Nenghai Yu[1]

[1]University of Science and Technology of China
[2]Nanyang Technological University
{yyp3334522@mail., welbeckz@}ustc.edu.cn , tianyi.wei@ntu.edu.sg

Abstract

Recent advances in image editing tools, particularly those used in content-aware retouching and object-level manipulation, have raised significant concerns regarding the authenticity of digital images. While many Image Manipulation Detection and Localization (IMDL) methods have been proposed, they often struggle with subtle forgeries, intricate boundary artifacts, and manipulations generated by unseen editing techniques. In this work, we propose a novel edge-aware framework that leverages the strong natural image priors of pre-trained inpainting models to harmonize manipulated regions. By guiding the inpainting process with generated edge-aware masks, our method reconstructs tampered areas using surrounding context, yielding perceptually coherent results. The pixel-wise residual between the original and reconstructed images reveals manipulation-sensitive inconsistencies—particularly around editing boundaries—thereby enabling accurate and generalizable detection and localization. Extensive experiments across multiple benchmarks demonstrate that our approach achieves state-of-the-art performance, especially in challenging scenarios involving realistic and finely retouched image forgeries.

Code — https://github.com/YYP3334522/EARG-Net

Introduction

Early approaches to Image Manipulation Detection and Localization (IMDL) predominantly relied on handcrafted features, such as sensor noise patterns (Zhu and Li 2018; Lyu, Pan, and Zhang 2014; Pan, Zhang, and Lyu 2012) and JPEG compression artifacts (Korus and Huang 2016; Iakovidou et al. 2018). However, with the rapid evolution of image editing technologies, manipulated regions have become increasingly sophisticated, often exhibiting intricate and seamless boundaries. The powerful post-processing capabilities of tools like Adobe Photoshop, combined with the remarkable image synthesis quality of Diffusion Generative Models (DGMs), have substantially narrowed the visual gap between tampered and authentic regions. As a result, modern IMDL methods must contend with three key challenges:

A major challenge in IMDL lies in detecting **subtle and localized manipulations**, where visual differences between tampered and authentic regions are marginal. In such cases, models often default to empty predictions due to weak supervisory signals and conservative loss formulations. Compounding this, real-world forgeries frequently exhibit **complex and high-curvature boundaries** that differ significantly from the smooth masks common in existing datasets, making fine-grained localization difficult. Additionally, the presence of **unseen editing tools and post-processing operations**—such as blending, color correction, and lighting harmonization—further obscures manipulation traces, reducing the reliability of both pixel-level and deep-feature-based cues.

The convergence of these challenges creates a perfect storm for manipulation detection systems, demanding novel methodologies capable of simultaneously: (1) detecting increasingly imperceptible tampering traces, (2) adapting to rapidly evolving and diverse editing techniques, and (3) distinguishing manipulated content from authentic regions despite intentional statistical obfuscation. Recent research has explored directions such as physics-based invariant features and self-supervised learning paradigms to mitigate these issues. However, the pace at which image editing technologies advance—particularly in terms of realism and post-processing sophistication—continues to outstrip the capabilities of existing detection frameworks.

In response to the aforementioned challenges, we propose a novel framework, termed Edge-Aware Reconstruction-Guided Network (EARG-Net), for effective image manipulation detection and localization. Our approach is built upon the insight that common editing operations—such as copy-move, splicing, and object removal—often introduce subtle, high-frequency artifacts along manipulation boundaries, even when synthesized by advanced deep generative models (DGMs). These boundary-level inconsistencies are frequently obscured by post-processing techniques, rendering them difficult to detect within the original image domain. To address this, we hypothesize that explicitly amplifying such inconsistencies can significantly enhance a model's ability to detect and localize tampered regions.

EARG-Net adopts a masking-based reconstruction strategy that targets high-frequency details critical for forensic analysis. Unlike traditional reconstruction pipelines that prioritize visual fidelity, our objective is to expose manipulation traces rather than suppress them. To this end, we intro-

✉ Corresponding Author.

duce the Edge-Aware Targeted Mask (EATM), which selectively masks regions with suspected boundary inconsistencies. Compared to random masking, EATM more effectively focuses the reconstruction process on semantically relevant areas, resulting in residual signals that are both more discriminative and more stable for manipulation detection. Pretrained image inpainting model is then applied to reconstruct the masked areas. Due to its reliance on surrounding context and strong natural image priors, the inpainting model produces visually coherent outputs that effectively harmonize manipulated regions. By computing the pixel-wise residual between the original and reconstructed images, we obtain a difference map that highlights manipulation-induced inconsistencies along object boundaries. To further refine these forensic cues, we incorporate a learnable Adaptive Edge Operator (AEO). Unlike fixed edge detectors, the AEO dynamically adapts its kernel weights to varying manipulation patterns, improving sensitivity to diverse boundary-level artifacts. The resulting Edge-aware Inconsistency Feature (EIF) is fused with spatial features from the original image through a dual-stream architecture, where cross-modal features are rectified and integrated via transformer-based interaction, enabling robust manipulation detection and localization.

Extensive experiments on benchmark datasets demonstrate that EARG-Net consistently outperforms state-of-the-art methods, particularly in challenging scenarios involving subtle manipulations and intricate tampering boundaries.

Our contributions can be summarized as follows:

- We propose a novel forensic feature called Edge-aware Inconsistency Feature (EIF), which reveals manipulation artifacts by applying targeted edge masking followed by image inpainting. To further enhance the inconsistency signal, we introduce a learnable Adaptive Edge Operator (AEO) for feature refinement.
- We design a dual-stream framework, EARG-Net, that integrates spatial features and edge-aware inconsistency representations through attention-driven cross-modal interaction, enabling precise and robust manipulation detection and localization.
- Extensive experiments demonstrate that our method outperforms state-of-the-art approaches across multiple benchmarks, particularly in localizing subtle manipulations and handling complex tampering boundaries.

Related Work

Image Manuplation Detection and Localization

Handcrafted Cues Based Method. Early IMDL approaches relied on intrinsic artifacts introduced during image acquisition or basic editing operations, such as inconsistencies in sensor noise (Zhu and Li 2018; Lyu, Pan, and Zhang 2014; Pan, Zhang, and Lyu 2012; Mahdian and Saic 2009), illumination variations (Riess et al. 2017; de Carvalho et al. 2013; Yao et al. 2012), Color Filter Array (CFA) demosaic artifacts (Ferrara et al. 2012; Singh, Singh, and Singh 2018), and JPEG compression traces (Korus and Huang 2016; Iakovidou et al. 2018; Li, Yuan, and Yu 2009). Although conceptually straightforward, these methods suffer from limited applicability, high computational cost, and vulnerability to simple post-processing, rendering them increasingly ineffective against modern editing tools.

Deep Learning Based Method. With the advent of deep learning, data-driven IMDL techniques have become predominant. For example, MVSS-Net (Dong et al. 2023) employs multi-view and multi-scale supervision to enhance localization precision. CAT-Net (Kwon et al. 2021) integrates RGB-domain and DCT-domain features for end-to-end forgery detection. TruFor (Guillaro et al. 2023) combines global signals (e.g., illumination consistency) with localized noise analysis to identify generative manipulations. Recent advances in large language models (LLMs) have spurred growing interest in GPT-based approaches for IMDL tasks. These methods leverage the multimodal capabilities of foundation models to achieve breakthroughs in forensic analysis. For example, ForgeryGPT (Li et al. 2024) establishes a novel paradigm by capturing high-level forensic correlations between visual artifacts and their linguistic representations across diverse feature spaces, enabling unprecedented interactive capabilities. Similarly, FakeShield (Xu et al. 2024) integrates GPT-4o's textual reAEOning with visual evidence, generating tamper-descriptive narratives that provide human-interpretable detection results while maintaining forensic accuracy.

Despite these advancements, existing methods still struggle with subtle artifacts, complex boundaries, and generalization to unseen manipulations—challenges our proposed EARG-Net is specifically designed to address. Early IMDL methods primarily relied on spatial or frequency domain cues, but often struggled to generalize to unseen manipulation types. To improve robustness, feature fusion has emerged as a key strategy, evolving through three main stages. The first stage used direct concatenation of low-level features (e.g., RGB, noise), which lacked the ability to model cross-modal relationships. The second stage introduced dual-stream networks that process spatial and frequency (or edge) features in parallel, often combined through attention mechanisms. For instance, PSCC-Net (Liu et al. 2022) enforces cross-modal consistency to highlight tampering traces. However, fusion designs at this stage were typically task-specific and lacked generalizability. In the current stage, fusion itself has become a central focus. Recent works adopt more flexible strategies—such as multi-scale fusion, pyramid integration, and cross-modal transformers—to learn richer representations. ObjectFormer (Wang et al. 2022) and GIMFormer (Chen et al. 2025) exemplify this shift toward general-purpose fusion frameworks. Building on this trend, our method integrates targeted forensic features into a unified pipeline for accurate and generalizable manipulation localization.

Method

Overview

Figure 1 provides an overview of our proposed method. For the input image, we first apply Edge-Aware Targeted Mask (EATM) to deliberately obscure the edges of candidated tampered regions. Next, we employ the inpainting model to reconstruct the image by removing the masked ar-

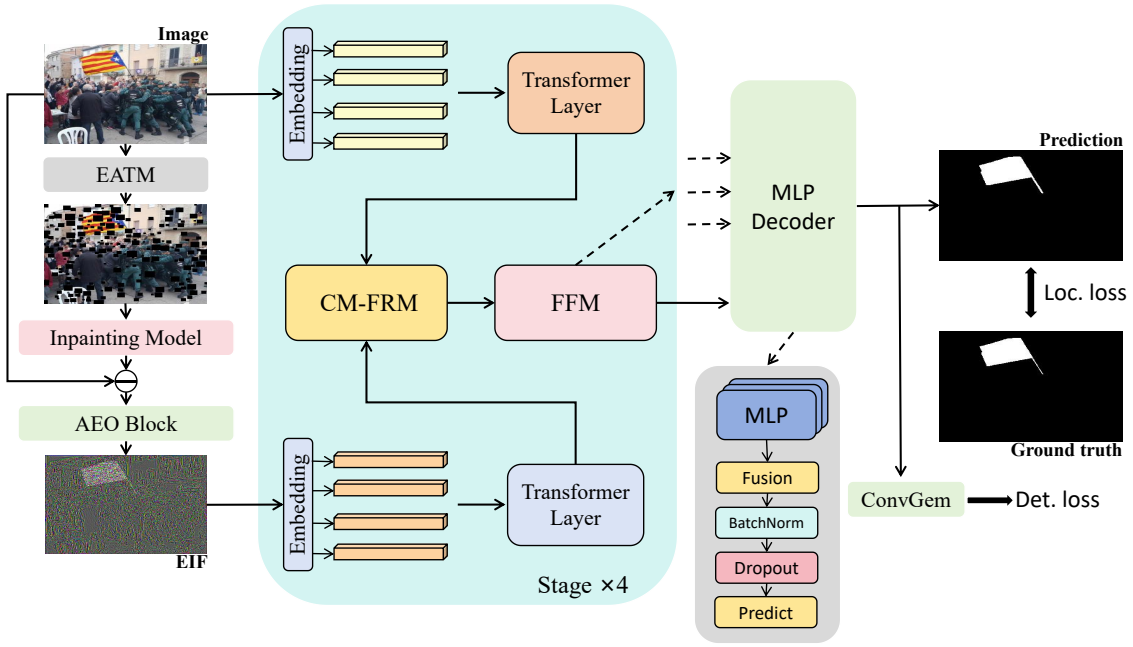

Figure 1: **Overall architecture of the proposed EARG-Net.** Given an input image, the Edge-Aware Targeted Mask (EATM) first obscures potential tampered boundaries, followed by reconstruction using a pre-trained inpainting model. The residual between the original and reconstructed images is processed by the Adaptive Edge Operator (AEO Block) to extract the Edge-aware Inconsistency Feature (EIF). The original image and the extracted Edge-aware Inconsistency Feature (EIF) are processed through two parallel transformer branches to encode spatial and forensic representations, respectively. These features are first refined via the Cross-Modal Feature Rectification Module (CM-FRM), which leverages global average pooling and MLP layers to align and calibrate modality-specific features. Subsequently, the Feature Fusion Module (FFM) employs a cross-attention mechanism to effectively integrate spatial and edge-aware cues. The fused representation is then decoded into a pixel-wise manipulation localization map under the supervision of ground-truth masks.

eas. We then perform pixel-by-pixel subtraction between the source image and the reconstructed image and feed this difference into the Adaptive Edge Operator(AEO) to extract forensic features - called the Edge-Aware Inconsistency Feature (EIF) . Finally, the EIF and spatial features are independently embedded and passed through parallel transformer layers. The resulting representations are fused via Cross-Modal Feature Rectification Module (CM-FRM) and Feature Fusion Module (FFM), and subsequently decoded to produce a precise localization map of manipulated regions.

Observation

Recent advances in image inpainting have demonstrated remarkable capabilities in reconstructing missing content with high visual plausibility, guided by strong natural image priors and spatial coherence. These models effectively exploit surrounding contextual cues to fill occluded regions in a perceptually harmonious manner. As illustrated in Figure 2, when a manipulated region (e.g., an inserted object such as a spider) is masked and restored by a state-of-the-art inpainting model, the tampered content is replaced with semantically consistent and visually coherent background, seamlessly integrating into the surrounding scene.

This behavior reveals a key insight: inpainting models inherently "disagree" with manipulated content, particularly when the manipulation introduces semantic or structural inconsistencies that deviate from natural image statistics. In contrast, authentic regions—due to their inherent contextual consistency—tend to be faithfully reconstructed without introducing significant deviations. By comparing the original image with its inpainted counterpart, we can extract meaningful discrepancies that serve as strong and semantically grounded forensic cues. Unlike traditional low-level artifacts, these inpainting-induced inconsistencies emerge from contextual or compositional mismatch, making them especially discriminative for detecting subtle, manually retouched forgeries.

Building on this observation, our approach explicitly leverages the reconstruction residual between the input and its inpainted version to extract edge-aware forensic signals, forming the core of the proposed EARG-Net framework.

Edge-aware Inconsistency Feature

In the IMDL domain, a variety of forensic features have been proposed to capture manipulation traces, such as the Bayar filter, SRM filters, and DCT-based representations. While these handcrafted or low-level statistical features have shown effectiveness in detecting certain types of manipulations, they also exhibit several limitations.First, many of these features are hand-designed and fixed, making them

Figure 2: **Illustration of the motivation behind our approach.** The top row shows an authentic image and its inpainting result, while the bottom row shows a manipulated image with its corresponding inpainting and ground-truth mask. For authentic images, the inpainting result remains visually consistent with the original, reflecting strong contextual priors. In contrast, for tampered images, the inpainting model removes the inserted object and restores semantically coherent background, revealing clear inconsistencies. This contrast highlights the potential of using reconstruction-based differences to expose manipulation traces.

less adaptive to diverse or unseen manipulation patterns, especially those introduced by generative models (e.g., DGM-based forgeries). Second, they typically operate in a global or patch-wise manner, which may overlook fine-grained inconsistencies along tampered boundaries—a critical cue for accurate localization. Moreover, most of these features lack the ability to capture semantic or context-aware inconsistencies, which are increasingly common in modern image forgeries (Fang et al. 2025).

To address these issues, we introduce a new forensic feature termed EIF (Edge-aware Inconsistency Feature). Unlike traditional descriptors, EIF is derived from the residual difference between the original image and an inpainted reconstruction guided by Edge-Aware Targeted Mask (EATM) . This residual emphasizes boundary-level inconsistencies introduced by manipulations. Additionally, we apply a learnable Adaptive Edge Operator (AEO) to further enhance these signals, enabling our method to better adapt to subtle and diverse forgery patterns.

Edge-Aware Targeted Mask. In this work, we introduce a novel edge-aware targeted masking strategy to guide self-supervised learning via high-frequency semantic regions. Unlike random or uniform masking approaches, our method exploits low-level visual cues—such as edges and corners—to apply fine-grained occlusions in a spatially informative manner.

Specifically, we propose a dense boundary-aware masking technique that synthesizes Sobel and Laplacian operators to identify salient regions characterized by strong gradients or corner responses. Based on the resulting composite edge map, we sample high-activation areas to apply small, non-overlapping masks, subject to constraints on both patch size and total coverage ratio. Compared to conventional random masking, this strategy not only guides the model toward reconstructing semantically critical structures like object boundaries but also promotes robustness to structured occlusions, enabling more discriminative and generalizable feature learning.

Given an input image $I \in \mathbb{R}^{C\times H\times W}$, we compute a composite edge heatmap E by combining Sobel and Laplacian responses. Based on the binarized high-activation regions in E, we sample multiple small, non-overlapping patches $\mathcal{R}_i$ such that the total covered area does not exceed a predefined ratio ρ. This yields a binary mask $M \in \{0,1\}^{1\times H\times W}$.The masked image $\tilde{I}$ is computed as $\tilde{I} = I \odot (1 - M)$, where all pixels within the mask are replaced with zero.

By focusing sampling near edge-rich regions—often aligned with object boundaries—our masking strategy guides the model to reconstruct structurally meaningful content. This targeted design ensures high coverage over potential manipulation areas, enhancing the forensic value of the residual. While the overall mask ratio has minimal impact, spatially aware masking proves critical for exposing subtle tampering cues.

Adaptive Edge Operators. Traditional edge operators exhibit limited adaptability to the diverse and subtle artifacts introduced by modern image manipulation techniques. To address this limitation, we introduce a differentiable module designed to selectively enhance forensic-relevant edges while suppressing irrelevant or noisy gradients. The core innovation of this module lies in integrating trainable feature transformations with classical gradient-based edge extraction, enabling adaptive emphasis on manipulation-specific boundaries within an end-to-end learning framework. Our edge enhancement module consists of two components:

- **Edge Normalization Unit:** We first apply fixed Sobel filters in the horizontal and vertical directions to compute gradient responses:

$$\begin{aligned} E_x &= \text{GroupNorm}(\text{Sobel}_x(I)) \\ E_y &= \text{GroupNorm}(\text{Sobel}_y(I)) \end{aligned} \tag{1}$$

  The normalized gradients are then combined to compute the gradient magnitude:

$$G = \sigma\left(\frac{\sqrt{E_x^2 + E_y^2 + \epsilon}}{T}\right) \tag{2}$$

  where $\sigma(\cdot)$ denotes the sigmoid function, ϵ is a small constant for numerical stability, and T is a learnable temperature parameter controlling edge sharpness.
- **Gradient-Guided Residual Block:** To propagate and enhance edge-aware features, we introduce a residual block guided by the gradient map G :

$$\begin{aligned} F' &= \text{Conv}_{1\times1}(F) \\ F'' &= \text{ResBlock}(F' \odot (G + \epsilon)) \\ F_{\text{out}} &= F'' + F' \end{aligned} \tag{3}$$

  Here, F is the residual between the input image and the reconstructed image, $\odot$ denotes element-wise multiplication, and the additive skip connection ensures preservation of identity information.

More details can be found in the supplementary material.

Edge-Aware Reconstruction-Guided Network

Building upon the proposed Edge-aware Inconsistency Feature (EIF), we develop EARG-Net, a dual-stream architecture designed for precise manipulation detection and localization. As illustrated in Figure 1, the network comprises an RGB stream that captures semantic and structural cues from the original image, and an EIF stream that highlights boundary-level inconsistencies induced by tampering. Inspired by CMX (Zhang et al. 2023), both streams follow a four-stage hierarchical transformer backbone with shared configurations to extract multi-scale representations ranging from low-level textures to high-level semantics.

To strengthen intra- and inter-modal representations, we introduce two key components: the Cross-Modal Feature Rectification Module (CM-FRM) and the Feature Fusion Module (FFM). FRM enhances each modality by applying channel-wise and spatial-wise rectification, using attention-guided scaling and MLP-based refinement to suppress redundant features while preserving manipulation-sensitive signals. FFM is designed to align and integrate the RGB and EIF features through bidirectional cross-attention, followed by residual fusion and lightweight convolutional mixing. This process adaptively balances global context and local edge information, bridging the semantic gap between modalities.

The fused features are then passed to a multi-layer decoder to generate pixel-wise manipulation masks, supervised by ground-truth annotations. In parallel, ConvGem classification head predicts image-level manipulation labels, providing global context to complement the localization task. This dual-task formulation encourages the network to learn both fine-grained and holistic forensic cues, improving robustness and generalization.

Loss Function

Our training objective consists of two main branches: manipulation mask prediction and tampering label classification. We design tailored loss functions for both tasks to ensure accurate localization and robust classification, particularly under challenging cases such as small tampered regions or subtle manipulations.

For pixel-wise manipulation localization, we employ a combination of Binary Cross-Entropy (BCE) loss and Dice loss:

$$\mathcal{L}_{\text{mask}} = \mathcal{L}_{\text{bce}}^{\text{mask}} + \lambda_1 \cdot \mathcal{L}_{\text{dice}} \tag{4}$$

where λ_1 is a balancing coefficient for the Dice loss, default set to be 0.1. To address the class imbalance between tampered and untampered regions, we introduce an adaptive weighting factor into the BCE term:

$$\mathcal{L}_{\text{bce}}^{\text{mask}} = -\left[\beta \cdot M \cdot \log(\hat{M}) + (1 - M) \cdot \log(1 - \hat{M})\right] \tag{5}$$

β is defined as follows:

$$\beta = \left(1 + \alpha \cdot (1 - \frac{|M|}{|I|})\right) \tag{6}$$

where M and $\hat{M}$ denote the ground truth and predicted manipulation masks, respectively, $|M|$ is the size of tampered field, $|I|$ is the size of image, and α controls the strength of the dynamic weight. This adaptive term encourages the model to pay more attention to samples with small manipulated regions, which are typically more difficult to detect.

For image-level tampering classification, we use a weighted Binary Cross-Entropy loss:

$$\mathcal{L}_{\text{label}} = -\left[\gamma \cdot y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})\right] \tag{7}$$

where $y \in \{0, 1\}$ is the ground truth label, and $\hat{y}$ is the predicted probability. The coefficient $\gamma > 1$ emphasizes penalizing false negatives, i.e., tampered images misclassified as authentic, which is critical in forensic scenarios.

The final optimization objective combines both branches:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{mask}} + \lambda_2 \cdot \mathcal{L}_{\text{label}} \tag{8}$$

where λ_2 is a balancing coefficient for the label loss, default set to be 0.3. This joint loss allows the network to simultaneously learn fine-grained localization and holistic tampering classification in an end-to-end fashion.

Experiment

Experiment Setup

Training Details. All images were resized to 512x512 pixels. We conducted the training over 50 epochs, utilizing a batch size of 8 on four A6000 graphics cards. The learning rate followed a cosine schedule, starting at 1e-4 and decreasing to a minimum of 5e-7.

Training Protocol. Following the IMDL-Benco (Ma et al. 2024) configuration, we employ two training protocols: Protocol-MVSS and Protocol-CAT.

- **Protocol-MVSS:** The model is trained exclusively on CASIAv2 (Dong, Wang, and Tan 2013) and evaluated on other datasets.
- **Protocol-CAT:** The model is trained on a composite dataset comprising CASIAv2, Fantastic Reality (Kniaz, Knyaz, and Remondino 2019), IMD2020 (Novozámský, Mahdian, and Saic 2020), tampered COCO, and tampered RAISE(Kwon et al. 2021), then tested on other datasets.

Testing Datasets. For benchmark comparisons, we utilize datasets commonly adopted in the IMDL research community:

- **Traditional manipulation datasets:** COVERAGE (Wen et al. 2016), Columbia (Hsu and Chang 2006), NIST16 (Guan et al. 2019), IMD2020 and CASIAv1 (Dong, Wang, and Tan 2013).
- **AI-generated datasets:** AutoSplice (Jia et al. 2023) and CocoGlide(Guillaro et al. 2023).

Since Protocol-CAT's training data already includes IMD2020, we exclude IMD2020 from testing when evaluating models trained under this protocol to prevent data leakage. Meanwhile, following IMDL-Benco's recommendations, we cleaned the NIST16 dataset.

Benchmark Models. We select 6 state-of-the-art methods in the IMDL field as benchmark models: MVSS-Net(Dong et al. 2023), CAT-Net(Kwon et al. 2021), ObjectFormer(Wang et al. 2022), PSCC-Net(Liu et al. 2022), Trufor(Guillaro et al. 2023), IML-ViT (Ma et al. 2023), Mesorch (Zhu et al. 2025) and SparseViT (Su et al. 2025).

Protocol	Method	Editing					DGM		Average
		COVERAGE	Columbia	NIST16	CASIAv1	IMD2020	AutoSplice	CocoGlide	
MVSS	MVSS-Net	0.304	0.361	0.095	0.274	0.162	0.167	0.214	0.225
	CAT-Net	0.259	0.584	0.271	0.581	0.221	0.217	0.084	0.316
	ObjectFormer	0.197	0.116	0.030	0.226	0.083	0.238	0.357	0.178
	PSCC-Net	0.206	0.604	0.201	0.378	0.245	**0.386**	0.226	0.321
	Trufor	0.322	**0.864**	**0.307**	**0.721**	0.287	0.196	0.202	0.414
	IML-ViT	0.354	0.746	0.268	0.718	**0.316**	0.137	0.208	0.393
	Ours	**0.403**	0.806	0.269	0.563	0.242	0.334	**0.387**	**0.429**
CAT	CAT-Net	0.230	0.915	0.055	0.442	–	0.043	0.139	0.352
	PSCC-Net	0.402	0.864	0.301	0.403	–	0.306	**0.470**	0.457
	Trufor	0.480	0.875	0.320	**0.820**	–	0.200	0.280	0.495
	IML-ViT	0.612	0.944	0.173	0.792	–	0.154	0.207	0.480
	Mesorch	0.594	0.867	0.302	0.820	–	0.211	0.326	0.520
	SparseViT	0.487	**0.956**	0.011	0.819	–	0.233	0.278	0.464
	Ours	**0.755**	0.924	**0.589**	0.498	–	**0.463**	0.401	**0.605**

Table 1: Comparative performance of state-of-the-art methods on benchmark datasets (values indicate F1 scores). The best and secondbest performances for each dataset under each protocol are highlighted in **bold** and underlined.

Protocol	Method	Editing					DGM		Average
		COVERAGE	Columbia	NIST16	CASIAv1	IMD2020	AutoSplice	CocoGlide	
MVSS	MVSS-Net	0.729	0.670	0.670	0.835	0.704	0.397	0.641	0.664
	CAT-Net	0.684	0.800	0.789	0.910	0.657	0.455	0.684	0.708
	ObjectFormer	0.337	0.336	0.173	0.546	0.147	0.410	0.697	0.378
	PSCC-Net	0.697	0.814	0.693	0.833	0.775	**0.692**	0.550	0.722
	Trufor	0.847	0.927	0.800	**0.945**	**0.831**	0.528	0.743	0.803
	IML-ViT	0.837	0.898	0.767	0.939	0.819	0.457	0.726	0.777
	Ours	**0.908**	**0.967**	**0.871**	0.836	0.761	0.526	**0.802**	**0.810**
CAT	CAT-Net	0.918	0.945	0.817	**0.980**	–	0.523	0.866	0.841
	PSCC-Net	0.887	0.945	0.789	0.890	–	**0.548**	0.873	0.817
	Trufor	0.927	0.899	0.805	0.974	–	0.489	0.867	0.826
	IML-ViT	0.922	0.955	**0.887**	0.976	–	0.501	0.824	0.844
	Mesorch	0.929	0.375	0.878	0.928	–	0.510	0.894	0.752
	SparseViT	0.914	0.484	0.832	0.966	–	0.504	0.861	0.760
	Ours	**0.988**	**0.995**	0.881	0.840	–	0.528	**0.912**	**0.853**

Table 2: Comparative performance of state-of-the-art methods on benchmark datasets (values indicate AUC scores). The best and secondbest performances for each dataset under each protocol are highlighted in **bold** and underlined.

Quantitative and Qualitative Comparison

As shown in Table 1 and Table 2, our method consistently outperforms state-of-the-art approaches under both the MVSS and CAT evaluation protocols. This improvement is largely attributed to the proposed Edge-aware Inconsistency Feature (EIF), which captures fine-grained manipulation cues by amplifying contextual and structural inconsistencies. By providing complementary forensic signals to spatial features, EIF enhances the model's ability to localize subtle or boundary-level tampering. In addition, our overall architecture—featuring targeted fusion and cross-modal interaction—ensures efficient integration of multi-modal information, further supporting accurate and robust manipulation detection.

Beyond quantitative metrics, qualitative visualizations of the predicted manipulation masks of various models, as provided in Figure 3, further highlight the efficacy of our method. As demonstrated in these results, our approach not only accurately localizes tampered regions (even those with subtle manipulations) , but also reconstructs complex and irregular tampering boundaries with remarkable fidelity. One potential area for future refinement lies in further calibrating the response of the model to edge-aware signals, as strong boundary inconsistencies may occasionally lead to slight overextension in predicted regions.

Ablation Study

Impact of Masking Strategy and Ratio. We first analyze the impact of the Edge-Aware Targeted Masking (EATM) strategy used to guide reconstruction. As shown in Table 3, EATM consistently outperforms random masking baselines across various masking ratios on both CASIAv1 and NIST16. This confirms that masking semantically and structurally relevant regions (e.g., along manipulation bound-

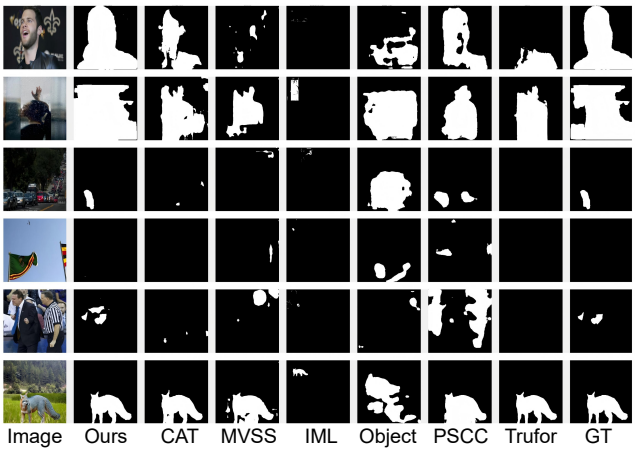

Figure 3: Qualitative result on some image from different datasets. Our model is compared with CAT-Net, MVSS-Net, IML-ViT, ObjectFormer, PSCC-Net and TruFor.

Mask choice	CASIAv1		NIST16	
	F1 ↑	AUC ↑	F1 ↑	AUC ↑
Random (ratio=0.25)	0.161	0.570	0.129	0.593
Random (ratio=0.50)	0.163	0.575	0.131	0.607
Random (ratio=0.75)	0.161	0.551	0.131	0.613
EATM (ratio=0.25)	0.487	0.836	0.584	0.879
EATM (ratio=0.50)	**0.498**	**0.840**	**0.589**	0.881
EATM (ratio=0.75)	0.498	0.839	0.597	**0.882**

Table 3: Ablation study on the contributions of EATM on CASIAv1 and NIST16 (CAT protocol). EATM demonstrates significantly better performance than random masking under all masking ratio settings.

aries) significantly enhances the informativeness of the reconstruction residual and leads to better localization performance, validating the core intuition behind EATM's design.

Interestingly, while different masking ratios yield very similar results under the EATM setting, random masking leads to both lower and less stable performance, suggesting that mask quality is more critical than mask size. We also observe that models trained with EATM typically converge more slowly than those using random masking—likely due to the increased semantic difficulty of the reconstruction task—but achieve better generalization on unseen manipulations. Within a reasonable range, the masking ratio does not significantly affect final performance; however, an overly small ratio may fail to produce meaningful EIF features, while an excessively large ratio increases task difficulty, hindering convergence. This trade-off between training complexity and cross-domain robustness further highlights the advantages of our targeted masking strategy in promoting meaningful and transferable feature learning.

Modality Contributions and Fusion Effectiveness. To further understand the roles of different feature streams, we conduct ablation experiments under three configurations: the baseline model using only RGB features, a variant without the RGB stream relying solely on EIF, and the full EARG-Net with both branches and feature fusion. The results on NIST16 and AutoSplice are shown in Table 4. We observe that removing either stream leads to a significant drop in performance, confirming the complementary nature of RGB semantics and EIF.

Model Variant	NIST16		AutoSplice	
	F1 ↑	AUC ↑	F1 ↑	AUC ↑
w/o RGB (EIF only)	0.158	0.783	0.184	0.506
w/o EIF (RGB only)	0.135	0.626	0.193	0.363
Full Setup	**0.269**	**0.871**	**0.334**	**0.526**

Table 4: Ablation study on the contributions of RGB features and EIF on NIST16 and AutoSplice (MVSS protocol). The exclusion of either RGB features or EIF individually compromises the overall effectiveness of the model.

Beyond validating the utility of each modality, this study also highlights the effectiveness of the proposed fusion mechanism. The full model consistently achieves the best results in both F1 and AUC metrics, indicating that the fusion strategy is not merely aggregating redundant information, but instead enables synergistic learning between appearance and edge-aware cues. The performance gap between unimodal and full settings further underscores the importance of integrating global semantics with local inconsistency signals for robust manipulation localization. Moreover, exploring how EIF can be more deeply integrated into the RGB representation stream may further enhance the model's discriminative capacity.

Conclusion

In this paper, we proposed EARG-Net, a novel framework for Image Manipulation Detection and Localization (IMDL) that addresses three key challenges: subtle tampering traces, complex manipulation boundaries, and generalization to unseen editing techniques. By integrating edge-aware targeted masking (EATM) with a pre-trained inpainting model, our method effectively amplifies boundary-level inconsistencies. A learnable adaptive edge operator further enhances forensic feature extraction by improving sensitivity to manipulation artifacts. Extensive experiments across multiple benchmarks and standardized evaluation protocols (Protocol-MVSS and Protocol-CAT) demonstrate that EARG-Net achieves state-of-the-art performance in localization accuracy and robustness, showing strong generalization to conventional and AI-generated forgeries. Future work may explore the integration of multiple inpainting models to capture stronger and more diverse natural image priors, enabling broader coverage of manipulation types and artifacts. Such model ensembles may provide richer contextual representations and further enhance detection accuracy and generalization. Overall, EARG-Net takes a solid step toward interpretable, generalizable manipulation detection, with practical value for digital forensics and content authentication.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grants 62121002,62372423,62072421, and was also supported by the Fundamental Research Funds for the Central Universities WK2100250070.

References

Chen, Y.; Huang, X.; Zhang, Q.; Li, W.; Zhu, M.; Yan, Q.; Li, S.; Chen, H.; Hu, H.; Yang, J.; Liu, W.; and Hu, J. 2025. GIM: A Million-scale Benchmark for Generative Image Manipulation Detection and Localization. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 2311–2319. AAAI Press.

de Carvalho, T. J.; Riess, C.; Angelopoulou, E.; Pedrini, H.; and de Rezende Rocha, A. 2013. Exposing Digital Image Forgeries by Illumination Color Classification. *IEEE Trans. Inf. Forensics Secur.*, 8(7): 1182–1194.

Dong, C.; Chen, X.; Hu, R.; Cao, J.; and Li, X. 2023. MVSS-Net: Multi-View Multi-Scale Supervised Networks for Image Manipulation Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(3): 3539–3553.

Dong, J.; Wang, W.; and Tan, T. 2013. CASIA Image Tampering Detection Evaluation Database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing, ChinaSIP 2013, Beijing, China, July 6-10, 2013*, 422–426. IEEE.

Fang, Z.; Zhao, H.; Wei, T.; Zhou, W.; Wan, M.; Wang, Z.; Zhang, W.; and Yu, N. 2025. UniForensics: Face Forgery Detection via General Facial Representation. *IEEE Transactions on Dependable and Secure Computing*.

Ferrara, P.; Bianchi, T.; Rosa, A. D.; and Piva, A. 2012. Image Forgery Localization via Fine-Grained Analysis of CFA Artifacts. *IEEE Trans. Inf. Forensics Secur.*, 7(5): 1566–1577.

Guan, H.; Kozak, M.; Robertson, E.; Lee, Y.; Yates, A. N.; Delgado, A.; Zhou, D.; Kheyrkhah, T.; Smith, J.; and Fiscus, J. G. 2019. MFC Datasets: Large-Scale Benchmark Datasets for Media Forensic Challenge Evaluation. In *IEEE Winter Applications of Computer Vision Workshops, WACV Workshops 2019, Waikoloa Village, HI, USA, January 7-11, 2019*, 63–72. IEEE.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 20606–20615. IEEE.

Hsu, Y.; and Chang, S. 2006. Detecting Image Splicing using Geometry Invariants and Camera Characteristics Consistency. In *Proceedings of the 2006 IEEE International Conference on Multimedia and Expo, ICME 2006, July 9-12 2006, Toronto, Ontario, Canada*, 549–552. IEEE Computer Society.

Iakovidou, C.; Zampoglou, M.; Papadopoulos, S.; and Kompatsiaris, Y. 2018. Content-aware detection of JPEG grid inconsistencies for intuitive image forensics. *J. Vis. Commun. Image Represent.*, 54: 155–170.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, 893–903. IEEE.

Kniaz, V. V.; Knyaz, V. A.; and Remondino, F. 2019. The Point Where Reality Meets Fantasy: Mixed Adversarial Generators for Image Splice Detection. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 215–226.

Korus, P.; and Huang, J. 2016. Multi-Scale Fusion for Improved Localization of Malicious Tampering in Digital Images. *IEEE Trans. Image Process.*, 25(3): 1312–1326.

Kwon, M.; Yu, I.; Nam, S.; and Lee, H. 2021. CAT-Net: Compression Artifact Tracing Network for Detection and Localization of Image Splicing. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, 375–384. IEEE.

Li, J.; Zhang, F.; Zhu, J.; Sun, E.; Zhang, Q.; and Zha, Z. 2024. ForgeryGPT: Multimodal Large Language Model For Explainable Image Forgery Detection and Localization. *CoRR*, abs/2410.10238.

Li, W.; Yuan, Y.; and Yu, N. 2009. Passive detection of doctored JPEG image via block artifact grid extraction. *Signal Process.*, 89(9): 1821–1829.

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*.

Lyu, S.; Pan, X.; and Zhang, X. 2014. Exposing Region Splicing Forgeries with Blind Local Noise Estimation. *Int. J. Comput. Vis.*, 110(2): 202–221.

Ma, X.; Du, B.; Liu, X.; Hammadi, A. Y. A.; and Zhou, J. 2023. IML-ViT: Image Manipulation Localization by Vision Transformer. *CoRR*, abs/2307.14863.

Ma, X.; Zhu, X.; Su, L.; Du, B.; Jiang, Z.; Tong, B.; Lei, Z.; Yang, X.; Pun, C.; Lv, J.; and Zhou, J. 2024. IMDL-BenCo: A Comprehensive Benchmark and Codebase for Image Manipulation Detection & Localization. *CoRR*, abs/2406.10580.

Mahdian, B.; and Saic, S. 2009. Using noise inconsistencies for blind image forensics. *Image Vis. Comput.*, 27(10): 1497–1503.

Novozámský, A.; Mahdian, B.; and Saic, S. 2020. IMD2020: A Large-Scale Annotated Dataset Tailored for Detecting Manipulated Images. In *IEEE Winter Applications of Computer Vision Workshops, WACV Workshops 2020, Snowmass Village, CO, USA, March 1-5, 2020*, 71–80. IEEE.

Pan, X.; Zhang, X.; and Lyu, S. 2012. Exposing image splicing with inconsistent local noise variances. In *2012 IEEE International Conference on Computational Photography, ICCP 2012, Seattle, WA, USA, April 28-29, 2012*, 1–10. IEEE Computer Society.

Riess, C.; Unberath, M.; Naderi, F.; Pfaller, S.; Stamminger, M.; and Angelopoulou, E. 2017. Handling multiple materials for exposure of digital forgeries using 2-D lighting environments. *Multim. Tools Appl.*, 76(4): 4747–4764.

Singh, A.; Singh, G.; and Singh, K. 2018. A Markov based image forgery detection approach by analyzing CFA artifacts. *Multim. Tools Appl.*, 77(21): 28949–28968.

Su, L.; Ma, X.; Zhu, X.; Niu, C.; Lei, Z.; and Zhou, J.-Z. 2025. 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*, volume 39, 7024–7032.

Wang, J.; Wu, Z.; Chen, J.; Han, X.; Shrivastava, A.; Lim, S.; and Jiang, Y. 2022. ObjectFormer for Image Manipulation Detection and Localization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 2354–2363. IEEE.

Wen, B.; Zhu, Y.; Subramanian, R.; Ng, T.; Shen, X.; and Winkler, S. 2016. COVERAGE - A novel database for copy-move forgery detection. In *2016 IEEE International Conference on Image Processing, ICIP 2016, Phoenix, AZ, USA, September 25-28, 2016*, 161–165. IEEE.

Xu, Z.; Zhang, X.; Li, R.; Tang, Z.; Huang, Q.; and Zhang, J. 2024. FakeShield: Explainable Image Forgery Detection and Localization via Multi-modal Large Language Models. *CoRR*, abs/2410.02761.

Yao, H.; Wang, S.; Zhao, Y.; and Zhang, X. 2012. Detecting Image Forgery Using Perspective Constraints. *IEEE Signal Process. Lett.*, 19(3): 123–126.

Zhang, J.; Liu, H.; Yang, K.; Hu, X.; Liu, R.; and Stiefelhagen, R. 2023. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*.

Zhu, N.; and Li, Z. 2018. Blind image splicing detection via noise level function. *Signal Process. Image Commun.*, 68: 181–192.

Zhu, X.; Ma, X.; Su, L.; Jiang, Z.; Du, B.; Wang, X.; Lei, Z.; Feng, W.; Pun, C.-M.; and Zhou, J.-Z. 2025. Mesoscopic insights: orchestrating multi-scale & hybrid architecture for image manipulation localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 11022–11030.