SARIF: Segment Anything for Robust Image Forensics

Dong-Hyun Moon*[1], Ju-Hyeon Nam*[1], and Sang-Chul Lee†[1]

Department of Electrical and Computer Engineering, Inha University,
100 Inha-ro, Michuhol-gu, Incheon, 22212, Republic of Korea
{22261207, jhnam0514}@inha.edu, sclee@inha.ac.kr

Abstract. Image forgery localization remains challenging due to diverse manipulation techniques and distribution shifts. Existing recent forgery localization models achieve high accuracy on benchmarks but often struggle with cross-domain generalization and robustness. In this paper, we propose **SARIF (Segment Anything for Robust Image Forensics)**, a framework that leverages Segment Anything Model (SAM), which has a promptable architecture and generalization ability to overcome these limitations. SARIF introduces a feedback-guided mask decoder and a dual-encoder design that extracts forgery-specific information to capture forensic traces while exploiting SAM's architecture. To localize manipulated regions, we design a block-wise prompting mechanism that derives forgery-specific cues from residual features between an adapted encoder and its frozen counterpart. These features are fused with the previous mask prompt to drive a feedback-based mask refinement process, enabling automatic forgery segmentation without manual input. Extensive experiments on standard forgery-localization benchmarks show that SARIF achieves strong average cross-dataset performance and robustness to common image corruptions. Our SARIF code is available in GitHub Link.

Keywords: Forgery Localization · Segment Anything Model · Corruption Robustness

1 Introduction

Large-scale generative models [36, 46, 47, 64, 80] and widely available commercial editing software [3, 5] have made image editing ubiquitous, enabling non-experts to create convincing image forgeries with minimal effort [11, 29, 54]. These advances have heightened concerns, as forged images fuel misinformation, trigger legal disputes, and erode public trust, undermining societal stability [1, 56, 73]. To meet these demands, early image forgery detection relied on traditional hand-crafted features such as illumination inconsistency traces [10, 34, 38], JPEG compression artifacts [66, 81], sensor pattern noise (SPN) [8, 42, 50], and color

* Equal contribution.
† Corresponding author.

filter array (CFA) interpolation patterns [12, 16, 67]. However, these methods focus on *image-level detection* and exhibit poor generalization to unseen forgery types, often requiring specific conditions [28, 78].

Motivated by these limitations, deep learning models, which are now the dominant paradigm across diverse vision tasks [15, 21, 24, 33, 84], enable *pixel-level forgery localization*. Models such as MantraNet [78] and SPAN [28] demonstrate strong performance in automatic forgery-trace extraction. However, these approaches are often computationally intensive and have limited representational capacity, constraining generalization to unseen forgeries [7, 86]. Recently, attention mechanisms [9, 27, 76] have mitigated these issues, improving both efficiency and accuracy in UNet-like encoder–decoder architectures [23, 48, 86]. Nevertheless, performance remains inadequate in localization accuracy, out-of-distribution generalization, and robustness to common corruptions.

To move beyond these limitations, we focus on the Segment Anything Model (SAM) [41], a vision foundation model trained on more than a billion mask annotations across millions of images; it provides strong zero- and few-shot generalization and a promptable interface. Recent image forgery localization [44, 83, 85] methods adapt SAM to exploit this capacity. However, many SAM-based pipelines still depend on manual prompts, which is labor-intensive, sensitive to prompt design, and often biased toward object boundaries rather than manipulation evidence, limiting practicality for forgery localization [70]. To address these issues, IMDPrompter [85] adopts multi-view prompt learning to automate prompt generation with optimal prompt selection, yet the method remains computationally heavy and structurally complex due to multiple view-specific encoders and auxiliary modules. Additionally, many recent studies rely solely on SAM's final embeddings, overlooking hierarchical cues and, critically, the *task-specific information*, which SAM trained on natural images typically does not provide, needed for accurate forgery localization.

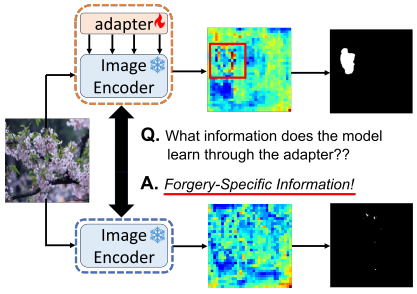

Fig. 1: ***Motivation***. Without the adapter, the encoder focuses on semantic content and misses subtle manipulation artifacts, causing imprecise masks. With the adapter, it learns forgery-specific cues that guide the decoder to produce sharper and more accurate localization. Concretely, this domain gap between the adapted and the frozen encoder represents the adapter-learned forgery-specific information.

Building on these observations, we present ***S****egment* ***A****nything for* ***R****obust* ***I****mage* ***F****orensics (****SARIF****)*, a SAM-based forgery localization framework that extracts task-specific residual cues from an adapted/frozen dual-encoder pair and refines masks through automatic feedback. This design improves sensitivity to subtle manipulation traces while preserving the strong generalization ability

of SAM. Additionally, we introduce a feedback-guided mask decoder to fully exploit SAM's lightweight decoder and promptable interface. At each feedback stage, we fuse the previous mask prediction with forgery-specific information to yield the next refined mask, updating it via a lightweight decoder with shared weights. With feedback-guided prompting, boundaries are progressively sharpened, spurious regions suppressed, and subtle *task-specific* cues recovered, while fully exploiting SAM's lightweight mask decoder and preserving the base encoder's strong generalization. Extensive experiments on standard public benchmarks show that SARIF achieves strong average performance, improved cross-domain generalization, and competitive robustness to common corruptions. The contributions of this paper can be summarized as follows:

- **Task-specific prompting.** We compute block-wise task-specific features from the differences between the fine-tuned and frozen encoders at selected global-attention blocks and at the final embedding. This design lets us progressively probe what each fine-tuned layer learns during training, producing a coarse-to-fine hierarchy of task-specific information that spans broad semantic context down to fine-grained manipulation cues.
- **Feedback-guided mask decoding with prompt–mask fusion.** We fuse the previous mask and forgery-specific feature to form the task-specific prompt and feed it to SAM's mask decoder with shared weights to obtain the current prediction. This stage-wise refinement accumulates evidence across refinement stages, stabilizes boundaries, and fully exploits SAM's promptable design.
- **Extensive experiments.** Across diverse benchmarks, our method shows strong localization accuracy, cross-domain generalization, and robustness under common post-processing distortions such as JPEG Compression, Gaussian Noise and Gaussian Blur.

2 Related Work

Deep Learning Based Forgery Localization.

1) CNN approaches: Representative CNN methods include RGB Noise [87], which extracts SRM based noise for cross type detection, and MantraNet [78] and SPAN [28], which use no pooling for pixel level localization. Several early models still return only bounding boxes, are computationally heavy or require extra fine tuning. To mitigate these issues, RRUNet [4] introduces feedback learning in a UNet style encoder decoder, MT-SENet [86] adds SE blocks [27] with multitask learning to sharpen boundaries, and PSCCNet [48], HDFNet [22], and PIMNet [2] adopt multiscale attention for adaptive fusion.

2) Transformer approaches: To supply global context missing in prior CNN methods, TransForensics [23] uses Vision Transformer self attention [15] to capture long range dependencies and improve pixel level localization.

3) Frequency approaches: Frequency features enhance cross domain generalization in various vision applications [57–60, 62]. For forgery localization,

ObjectFormer [74] and FBINet [18] apply the 2D DCT to expose subtle tampering cues. M2SFormer [61] fuses multiscale and multispectral signals for better generalization and corruption robustness, while DNet [82], AFENet [79], and EITLNet [20] further strengthen generalization using a Haar wavelet transform, a 2D discrete Fourier transform, and a high pass noise filter with RGB inputs, respectively.

4) Segment Anything approaches: Vision foundation models such as SAM [41,68] provide promptable mask generation with strong zero-shot and few-shot generalization, motivating their use in forgery localization. IMDPrompter [85] freezes the encoder and learns multi-view prompts with a noise-oriented CNN and optimal selection, while SAMIF [83] augments the image encoder with adapters and an SRM branch to emphasize high-frequency cues. SAFIRE [44] reframes forgery localization as source-region partitioning: it performs point-prompted source-region segmentation and aggregates predictions from a grid of automatically generated points to recover source partitions and forged regions. Nevertheless, many SAM-based pipelines remain prompt-dependent or reduce SAM to a single final embedding, leaving hierarchical signals and task-specific manipulation cues underexplored. These limitations motivate our design, which injects hierarchical cues through dual-encoder and automatic feedback refinement while preserving SAM's promptable decoder.

Segment Anything for Computer Vision. The Segment Anything Model (SAM) [37, 41, 68] comprises an image encoder and a promptable, lightweight mask decoder that accepts points, boxes, or masks as prompts and generalizes in zero-shot across diverse distributions. Enabled by this generalization, SAM is increasingly adopted as a downstream backbone across tasks. In medical imaging, MedSAM [51], AutoSAM [70], SAM-Med2D [77], and SAM-Med3D [88] extend SAM to diverse 2D and 3D modalities, delivering versatile organ and lesion segmentation with minimal prompting. In camouflaged object detection, COD-SAM [17] and SAM-COD+ [6] employ adapters, distillation, and task aware prompting to sharpen subtle boundary cues. For forgery localization, ***SARIF*** combines parameter-efficient adaptation, dual-encoder residual extraction, and feedback-guided decoding to inject task-specific cues without relying on manual prompts or a single final embedding.

3 Proposed Method

3.1 Overall Architecture

To fully leverage SAM's hierarchical representations and promptable mask decoder for forgery localization, we propose our overall framework, **S**egment **A**nything for **R**obust **I**mage **F**orensics (SARIF), illustrated in Fig. 2. SARIF comprises three modules: *(1) a dual SAM image encoder*, *(2) the Forgery-Specific Information Extractor (FSIE; Sec. 3.2)*, and *(3) the Feedback-Guided Mask Decoder (FGMD; Sec. 3.3)*. As illustrated in Fig. 2 (a) and (b), we insert LoRA adapters [26] into one image-encoder branch and update only these adapters and mask decoder for forgery localization, while keeping the backbone weights

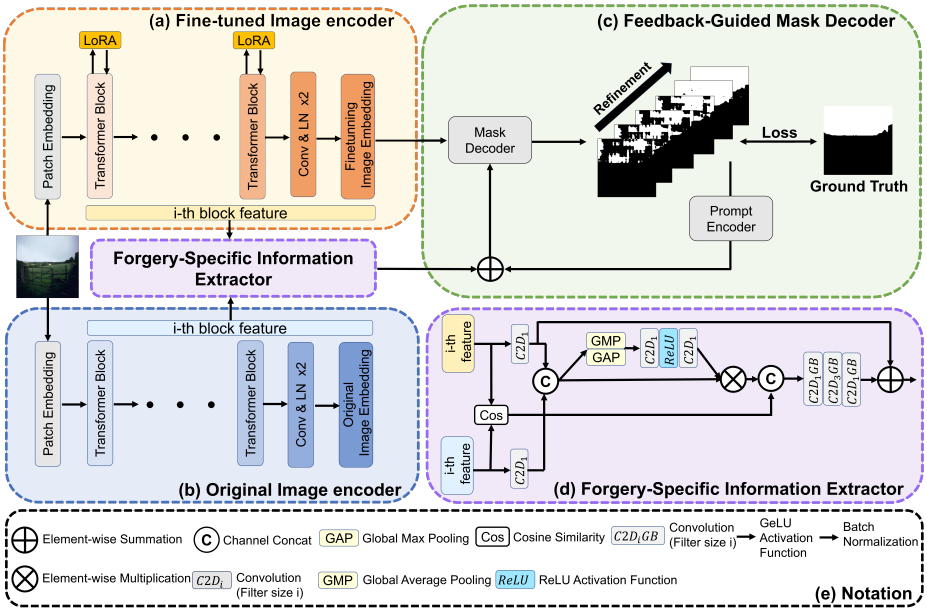

Fig. 2: The overall architecture of the proposed **SARIF**. (a) Fine-tuned SAM Image Encoder. (b) Original SAM Image Encoder. (c) Feedback-Guided Mask Decoder. (d) Forgery-Specific Information Extractor. (e) Notation description used in this paper.

frozen. At each selected transformer block, FSIE computes block-wise residual cues between the adapted and frozen branches to distill forgery-specific information, which is then fused with the previous mask prompt to produce the next task-specific prompt. At each refinemen stage, the decoder takes the fine-tuned SAM embedding together with this prompt to predict and refine the mask. The final mask is obtained by leveraging a task-specific prompt derived from the fine-tuned and original embeddings. By capturing what SAM learns through LoRA and fully using SAM's promptable design and lightweight mask decoder, the framework precisely adapts SAM to the forgery localization task.

3.2 Forgery Specific Information Extractor

Motivation. We observe that a single branch adapter setting is parameter efficient but provides no explicit understanding of what the adapter learns for the forgery specific task, limiting sensitivity to subtle manipulations. Motivated by primate biological vision theories [32, 43, 53], where opponent channels and center surround circuitry encode relative contrast rather than absolute intensity, we pair the pretrained SAM encoder with its LoRA adapted counterpart and compute blockwise residuals at transformer block to form compact forgery specific cues. To realize this, we introduce the ***Forgery Specific Information Extractor (FSIE)***, which aggregates these features with a lightweight resid-

ual block to produce low cost features that preserve manipulation sensitivity while keeping the dual branch design practical. This design is also inspired by HQ-SAM [37], which taps features immediately after global-attention blocks; accordingly, we therefore compute residual cues only at global-attention blocks, where long-range interactions are explicitly enabled, capturing global inconsistencies from forgeries with minimal overhead. FSIE operates in three stages: *1) Feature Similarity Calculation*, *2) Feature Enhancement*, and *3) Fusion*.

Feature Similarity Calculation. Let $\mathbf{O}^t, \mathbf{F}^t \in \mathbb{R}^{C\times H\times W}$ denote the features extracted at refinement stage t from the original and fine-tuned SAM encoders, respectively. We first compute the cosine similarity map $\mathbf{cos}^t$ between the input feature maps, which is utilized later in subsequent stages.

$$\mathbf{cos}^t_{ij} \;=\; \frac{\langle \mathbf{O}^t_{:,ij}, \mathbf{F}^t_{:,ij} \rangle}{\|\mathbf{O}^t_{:,ij}\|_2 \, \|\mathbf{F}^t_{:,ij}\|_2}, \quad \mathbf{cos}^t \in \mathbb{R}^{1\times H\times W}. \tag{1}$$

Where $\langle\cdot,\cdot\rangle$ inner product, $\|\cdot\|_2$ ℓ_2-norm. Then we project both branches to $C^{'} < C$ using 1×1 convolution, which reduces computation and aligns the two branches for subsequent enhancement and fusion.

Feature Enhancement. To estimate the channel importance within the feature map of each stage t, we apply a channel attention mechanism. Specifically, we apply global average pooling and global max pooling over the spatial dimensions to obtain two channel-wise descriptors, processed by a two 1×1 convolution layer with an intervening ReLU.

$$\mathbf{C}^t = \mathrm{Cat}\big(\mathbf{F}^t, \mathbf{O}^t\big) \in \mathbb{R}^{2C'\times H\times W}, \tag{2}$$

$$\mathbf{g}^t_{\max} = \mathrm{Conv}_{1\times1}\big(\phi(\mathrm{Conv}_{1\times1}(\mathrm{GMP}(\mathbf{C}^t)))\big), \tag{3}$$

$$\mathbf{g}^t_{\mathrm{avg}} = \mathrm{Conv}_{1\times1}\big(\phi(\mathrm{Conv}_{1\times1}(\mathrm{GAP}(\mathbf{C}^t)))\big). \tag{4}$$

Denote ϕ is ReLU activation function. Then aggregate their outputs, and apply a sigmoid function to generate channel-wise attention weights. These weights are multiplied with the original feature map to emphasize informative channels while suppressing less relevant ones.

$$\mathbf{G}^t = \sigma\big(\mathbf{g}^t_{\max} + \mathbf{g}^t_{\mathrm{avg}}\big), \tag{5}$$

$$\mathbf{T}^t_{\mathrm{ca}} = \mathbf{C}^t \odot \mathbf{G}^t \in \mathbb{R}^{2C'\times H\times W}. \tag{6}$$

Denote $\sigma(\cdot)$ is the sigmoid function, $\odot$ denotes channel-wise reweighting with spatial broadcasting.

Fusion. We use the previously computed cosine similarity map as a hint map that guides the network toward forgery-specific information and fuse it with the channel-attended feature map.

$$\mathbf{Z}^t = \mathrm{Cat}\big(\mathbf{T}^t_{\mathrm{ca}}, \mathbf{cos}^t\big) \in \mathbb{R}^{(2C'+1)\times H\times W}. \tag{7}$$

To extract forgery-specific information, we employ a lightweight convolutional layer, GeLU activation, batch normalization with a skip connection. For computational efficiency, we reduce the channel dimension to $r < C'$, apply a 3×3

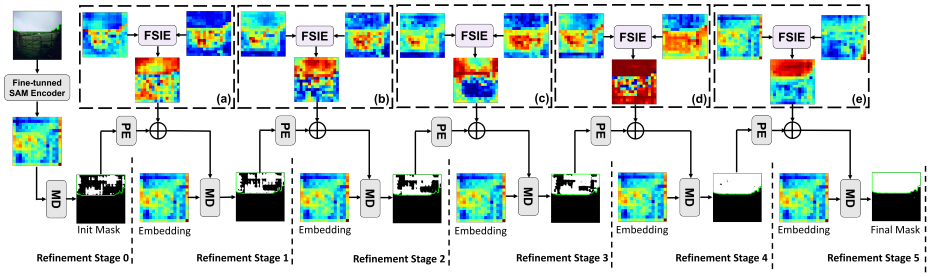

Fig. 3: More details about Feedback-Guided Mask Decoder. (a–e) show features from the 5th, 11th, 17th, and 23rd blocks and the final embedding. We also visualize forgery-specific information maps produced by FSIE. For each FSIE input pair, the left feature map comes from the fine-tuned encoder and the right one comes from the original frozen encoder. At each refinement stage, the previous mask is passed to the prompt encoder (PE) to generate a mask prompt, which is fused with the forgery-specific information produced by FSIE to form a task-specific prompt. This task-specific prompt, together with the fine-tuned image embedding, is fed into the mask decoder (MD) to progressively refine the mask. **Green** lines denote the boundaries of the ground truth.

convolution in this reduced space, and then project the features back to the original embedding dimension.

$$\mathbf{T}_{\text{fs}}^{t} = \mathbf{F}^{t} \oplus \mathbf{CGB}_{1\times1}^{C'}\Big(\mathbf{CGB}_{3\times3}\big(\mathbf{CGB}_{1\times1}^{r}(\mathbf{Z}^{t})\big)\Big), \tag{8}$$

We extract forgery-specific features at the global-attention blocks with indices {5, 11, 17, 23}, and at final embedding. These stage-wise cues are sequentially injected into the Feedback-Guided Mask Decoder together with the image embedding and previous mask prompt to progressively refine masks.

3.3 Feedback-Guided Mask Decoder

Motivation. SAM provides a promptable architecture that can be reused at inference, which we leverage to progressively recover fine mask detail [41]. However, most automatic SAM based forgery localization methods do not reuse previous predictions as feedback [83,85]. We therefore design a ***Feedback-Guided Mask Decoder (FGMD)*** that, at each refinement stage, fuses the previous mask after SAM prompt encoding with FSIE cues to form a new prompt and refine boundaries in a coarse to fine hierarchy, consistent with human vision system where center and surround interactions encode relative contrast and recurrent processing sharpens perception [31,40,45]. This design also accords with evidence that iterative refinement improves segmentation accuracy and boundary fidelity at modest computational cost. FGMD operates in two stages: *1) Forgery-specific Prompt Generation*, and *2) Iterative Mask Refinement.*

Forgery-specific Prompt Generation. Prior to mask refinement, an initial mask is required to serve as the reference for subsequent updates. We define an

initial mask as the mask decoder's output given only the fine-tuned SAM image embeddings, without external prompts. Finally, the previous mask is defined as:

$$\mathbf{M}^{(t-1)} = \begin{cases} \textbf{None} & \text{if } t = 1, \\ \mathbf{M}^{(t-1)} & \text{if } t > 1, \end{cases} \tag{9}$$

For refinement stage t, we encode the previous mask $\mathbf{M}^{t-1}$ with the prompt encoder and fuse the resulting mask prompt with the t-th forgery-specific cue extracted by FSIE.

$$\mathbf{Mask}^{t}_{prompt} = \text{PromptEncoder}\big(\mathbf{M}^{t-1}\big). \tag{10}$$

$$\mathbf{T}^{t}_{\mathbf{fs}} = \text{FSIE}\big(\mathbf{F}^{t}, \mathbf{O}^{t}\big). \tag{11}$$

Where PromptEncoder denotes SAM's prompt encoder. To generate the task-specific prompt, the task-specific features are merged with the mask prompt via element-wise summation.

$$\mathbf{T}^{t}_{\mathbf{pt}} = \big(\mathbf{Mask}^{t}_{Prompt} \oplus \mathbf{T}^{t}_{\mathbf{fs}}\big). \tag{12}$$

Stage-wise Mask Refinement. Let $\mathbf{T}^{t}_{\text{pt}}$ denote the task-specific prompt at refinement stage t. The SAM mask decoder $\mathcal{D}$ takes the final fine-tuned image embedding $\mathbf{E}_{\text{fine}}$ and $\mathbf{T}^{t}_{\text{pt}}$ as inputs and produces the refined prediction:

$$\mathbf{M}^{t} = \sigma\big(\mathcal{D}\big(\mathbf{E}_{\text{fine}}, \mathbf{T}^{t}_{\text{pt}}\big)\big)\,. \tag{13}$$

Here, $\sigma(\cdot)$ denotes the sigmoid function. The predicted mask is thresholded and passed to the next refinement stage as a mask prompt. Each refinement stage is supervised with the BCE loss against the ground-truth mask:

$$\mathcal{L}_{\text{BCE}} = \sum_{t=1}^{6} \text{BCE}\big(\mathbf{M}^{t}, \mathbf{Y}\big)\,, \tag{14}$$

where $\mathbf{Y}$ denotes the ground-truth mask. This design injects task cues through the prompt path, preserves accumulated spatial context through the mask path, and performs progressive refinement with a lightweight decoder. Further details of the refinement are shown in Figure 3.

4 Experiment Results

4.1 Experiment Settings

To evaluate model generalization, we train under CASIAv2 [65], then assess cross domain performance on seven external datasets: CASIAv1 [14], Columbia [25], IMD2020 [63], CoMoFoD [72], In the Wild [30], MISD [35] and DIS25k [71].

Table 1: Segmentation results (DSC) on CASIAv2 [65] training scheme. (·) denotes the standard deviations of five-fold cross-validation experiment results.

Type	Method	Seen Domain	Unseen Domain						
		CASIAv2 [65] DSC	DIS25k [71] DSC	CASIAv1 [14] DSC	Columbia [25] DSC	IMD2020 [63] DSC	CoMoFoD [72] DSC	In the Wild [30] DSC	MISD [35] DSC
Convolution	UNet [69]	32.3 (10.9)	8.7 (1.2)	25.0 (1.9)	19.8 (2.5)	14.9 (1.0)	12.2 (1.0)	18.6 (1.4)	47.3 (2.5)
	MantraNet [78]	18.8 (8.0)	12.7 (0.8)	19.8 (1.0)	25.0 (1.8)	14.1 (0.5)	11.2 (0.6)	18.2 (1.1)	30.7 (3.5)
	RRUNet [4]	21.8 (10.4)	10.8 (2.2)	26.7 (2.4)	21.3 (7.2)	14.5 (1.9)	13.7 (2.2)	18.3 (4.2)	32.8 (3.4)
	TransForensic [23]	40.1 (15.7)	32.3 (2.9)	44.2 (1.6)	35.9 (5.7)	27.2 (2.1)	21.7 (2.0)	31.9 (3.8)	60.0 (1.8)
	FBINet [18]	35.3 (14.3)	26.3 (2.3)	37.5 (2.1)	18.2 (2.7)	24.2 (1.6)	22.3 (0.8)	25.0 (2.7)	48.0 (2.2)
	MT-SENet [86]	19.9 (9.9)	7.8 (1.0)	18.4 (1.6)	9.4 (1.1)	11.4 (1.3)	10.6 (2.1)	10.6 (2.1)	22.3 (2.5)
	MVSSNet [13]	31.2 (13.4)	24.9 (3.6)	36.6 (1.6)	33.8 (3.7)	22.8 (2.4)	17.2 (1.2)	27.0 (3.1)	53.9 (3.6)
Trans.	PIMNet [2]	55.8 (15.1)	37.5 (2.4)	49.7 (0.8)	32.5 (5.2)	29.6 (2.7)	24.7 (1.6)	31.2 (2.5)	61.1 (0.9)
	EITLNet [20]	54.0 (14.7)	30.8 (2.8)	52.9 (1.7)	28.0 (4.6)	25.3 (2.6)	18.1 (1.8)	24.3 (3.6)	58.8 (1.8)
	M2SFormer [61]	***58.8 (12.8)***	38.5 (2.4)	***58.4 (0.7)***	42.4 (5.8)	32.6 (2.2)	24.9 (1.3)	35.0 (1.8)	**69.1 (0.7)**
SAM	SAM [41]	27.1 (12.2)	18.7 (3.4)	33.4 (6.4)	24.6 (11.3)	22.3 (2.7)	19.7 (3.0)	29.6 (8.1)	31.3 (8.1)
	AutoSAM [70]	49.0 (19.5)	31.4 (1.4)	45.1 (4.1)	28.2 (9.0)	34.0 (3.0)	23.8 (1.7)	36.5 (12.0)	61.3 (5.5)
	IMDPromter [85]	32.5 (16.9)	28.6 (2.2)	46.1 (2.8)	33.9 (6.6)	28.0 (2.2)	23.0 (1.1)	33.7 (3.1)	47.4 (2.7)
	SAFIRE [44]	56.8 (10.4)	**50.7 (0.9)**	**61.6 (0.9)**	***53.1 (4.9)***	**53.1 (0.6)**	***39.5 (1.0)***	**63.3 (2.2)**	48.3 (1.5)
	SARIF (Ours)	**63.1 (12.9)**	***47.8 (2.5)***	***58.4 (1.3)***	**58.4 (3.4)**	***48.4 (1.7)***	**65.4 (2.0)**	***55.7 (3.5)***	***66.2 (1.9)***

Table 2: Segmentation results (mIoU) on CASIAv2 [65] training scheme. (·) denotes the standard deviations of five-fold cross-validation experiment results.

Type	Method	Seen Domain	Unseen Domain						
		CASIAv2 [65] mIoU	DIS25K [71] mIoU	CASIAv1 [14] mIoU	Columbia [25] mIoU	IMD2020 [63] mIoU	CoMoFoD [72] mIoU	In the Wild [30] mIoU	MISD [35] mIoU
Convolution	UNet [69]	25.8 (9.2)	5.5 (0.8)	19.5 (1.8)	12.2 (1.7)	9.7 (0.6)	7.6 (0.7)	11.9 (1.0)	34.8 (2.2)
	MantraNet [78]	11.9 (5.7)	7.4 (0.6)	12.1 (0.8)	14.9 (1.2)	8.2 (0.3)	6.5 (0.4)	10.6 (0.8)	19.2 (2.5)
	RRUNet [4]	15.8 (8.4)	6.9 (1.5)	18.9 (1.7)	13.9 (5.3)	9.4 (3.3)	9.0 (1.7)	11.7 (3.0)	21.4 (2.5)
	TransForensic [23]	32.0 (13.9)	24.1 (2.5)	35.0 (1.7)	25.0 (4.6)	19.1 (1.6)	14.3 (1.5)	22.4 (3.0)	46.5 (2.1)
	FBINet [18]	29.2 (12.8)	20.3 (1.9)	30.8 (1.9)	11.7 (2.1)	17.5 (1.3)	15.2 (0.5)	17.8 (2.1)	35.1 (2.0)
	MT-SENet [86]	14.6 (7.6)	4.8 (0.7)	12.8 (1.3)	5.3 (0.6)	7.2 (1.0)	6.5 (1.5)	6.5 (1.5)	14.0 (1.8)
	MVSSNet [13]	23.4 (11.1)	17.4 (2.9)	27.3 (1.5)	23.3 (3.0)	15.2 (1.8)	10.9 (0.9)	18.3 (2.2)	17.4 (2.9)
Trans.	PIMNet [2]	48.5 (14.6)	30.1 (2.3)	42.2 (1.0)	23.1 (4.3)	22.2 (2.3)	16.8 (1.4)	22.9 (2.2)	48.2 (0.9)
	EITLNet [20]	47.9 (14.7)	25.6 (2.6)	46.5 (1.5)	20.9 (4.0)	19.7 (2.2)	12.4 (1.4)	19.0 (3.1)	45.9 (1.8)
	M2SFormer [61]	***50.8 (12.8)***	31.3 (2.3)	***50.1 (0.6)***	32.4 (5.3)	24.9 (1.9)	16.8 (1.0)	27.4 (1.6)	**56.9 (0.8)**
SAM	SAM [41]	21.0 (10.5)	14.0 (3.1)	26.9 (6.1)	15.9 (8.0)	16.1 (2.1)	12.8 (2.2)	20.3 (6.9)	20.0 (5.7)
	AutoSAM [70]	41.7 (19.0)	24.6 (1.6)	38.9 (4.4)	18.9 (7.6)	25.6 (2.3)	16.3 (1.4)	25.6 (11.0)	46.1 (5.9)
	IMDPromter [85]	25.9 (14.9)	21.3 (2.0)	37.9 (3.3)	22.9 (4.5)	19.7 (2.0)	15.4 (0.8)	23.4 (2.9)	33.1 (2.5)
	SAFIRE [44]	45.5 (10.9)	***40.9 (1.2)***	49.6 (0.9)	***40.5 (4.7)***	**41.1 (0.6)**	***26.4 (0.8)***	**51.0 (2.2)**	33.9 (1.3)
	SARIF (Ours)	**56.7 (13.3)**	**41.5 (2.4)**	**52.0 (1.1)**	**49.7 (3.1)**	***40.4 (1.6)***	**55.8 (2.0)**	***48.6 (3.6)***	***53.9 (2.0)***

We emphasize that none of these benchmarks are used for training, in line with the evaluation protocol followed by the recent state of the art M2SFormer [61]. We also provide detailed dataset summaries and the exact split definitions for each training scheme in the Supplementary Material. To measure performance, we report Dice Similarity Coefficient (DSC) and mean Intersection over Union (mIoU), which are established metrics for segmentation.

We compare the proposed **SARIF (Ours)** against fourteen representative forgery localization models, including eight convolutional-based approaches (UNet [69], ManTraNet [78], RRU-Net [4], TransForensic [23], FBI-Net [18], MT-SENet [86], MVSS-Net [13], PIM-Net [2]), two Transformer-based approaches (EITL-Net [20], M2SFormer [61]), and four SAM-based approaches (SAM [41] used without prompts, AutoSAM [70], IMDPrompter [85], SAFIRE [44]). For a fair comparison, all methods are trained on the same training set with matched pre-processing and hyperparameters where applicable, and evaluated on an identical held out test set to assess generalization. We report the mean over five-fold cross-validation for reliability. In all tables, **Red** and ***Blue*** denote the best and second best results.

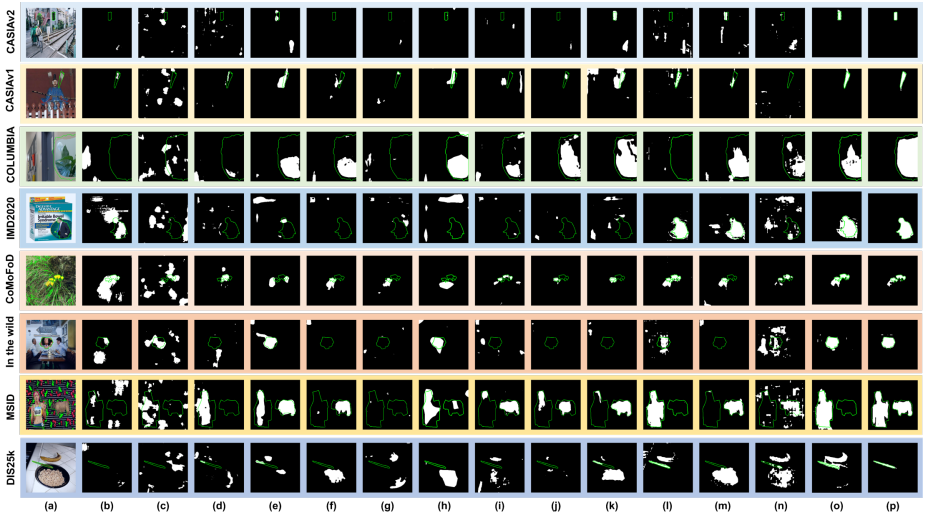

Fig. 4: Qualitative comparison of other methods and SARIF under the CASIAv2 training scheme. (a) Input images with ground truth. (b) UNet [69]. (c) MantraNet [78]. (d) RRUNet [4]. (e) TransForensic [23]. (f) FBINet [18]. (g) MT-SENet [86]. (h) MVSS-Net [13]. (i) PIMNet [2]. (j) EITLNet [20]. (k) M2SFormer [61]. (l) SAM [41]. (m) AutoSAM [70]. (n) IMDprompter [85]. (o) SAFIRE [44]. (p) **SARIF(ours)**. **Green** lines denote the boundaries of the ground truth.

4.2 Implementation Details

Training Settings. All models were optimized with Adam [39] starting from an initial learning rate of 10^{-4}, and the learning rate was annealed to 10^{-6} using a cosine schedule [49]. We trained for 100 epochs with a batch size of 32. Because the datasets provide images at heterogeneous native resolutions, we resized all inputs to 256×256 for the unified setting. Additionally, to ensure a controlled comparison, we used identical preprocessing, optimizer, and scheduler configurations for all models across all experiments.

Hyperparameters of SARIF. We adopt SAM's ViT-L pretrained weights and set LoRA rank as 32 in this paper. We extract intermediate embeddings at global attention block {5, 11, 17, 23} and the final embedding from both the fine-tuned SAM image encoder and the original frozen encoder.

4.3 Comparison With State-of-the-art Method

We evaluate our method against a diverse set of representative forgery localization baselines that span three backbone families: (i) CNN-based architectures such as UNet [69], MantraNet [78], RRUNet [4], TransForensic [23], MT-SENet [86], MVSSNet [13] and FBI-Net [18] which rely on convolutional feature extractors; (ii) Transformer-based architectures such as EITLNet [20], M2SFormer [61],

Table 3: Ablation study on Adapter, FSIE and FGMD. (·) denotes the standard deviations of five-fold cross-validation experiment results.

Setting	Adapter	FSIE	FGMD	Params(M)	FLOPs(G)	Seen			Unseen		
						DSC	mIoU	AUC	DSC	mIoU	AUC
1				13.90	490.65	23.6 (9.8)	18.1 (8.4)	63.0 (4.6)	23.3 (6.6)	16.7 (5.0)	59.9 (3.5)
2	✓			34.90	503.79	41.2 (14.5)	35.0 (13.6)	71.4 (6.9)	33.0 (11.0)	26.0 (9.7)	64.6 (6.1)
3	✓	✓		41.04	994.88	62.4 (12.9)	56.0 (13.3)	83.0 (5.4)	44.7 (15.2)	38.1 (14.8)	70.2 (8.6)
4	✓		✓	41.34	506.13	62.0 (12.1)	55.5 (12.4)	82.5 (5.1)	42.6 (15.2)	35.4 (14.5)	69.1 (8.6)
5	✓	✓	✓	52.62	998.43	**63.1 (12.9)**	**56.7 (13.3)**	**83.3 (5.4)**	**57.9 (7.1)**	**49.8 (6.1)**	**78.3 (3.6)**

PIMNet [2] which leverage global self-attention to capture long-range manipulation cues beyond local texture inconsistencies; and (iii) SAM-based architectures such as autoSAM [70], IMDPROMPTER [85], SAFIRE [44] and the original SAM [41], which adapt the Segment Anything Model for manipulation mask prediction. By comparing against all three categories, we ensure that our evaluation covers both traditional CNN-style forensic detectors, modern vision transformer models, and recent SAM-based models, demonstrating that our approach remains competitive across fundamentally different backbone designs.

Quantitative Result. Tabs. 1-2 report the cross-domain localization results. SARIF achieves the best seen-domain performance and strong average generalization across unseen datasets. On DSC, SARIF performs best on several unseen benchmarks, while SAFIRE remains stronger on DIS25K, IMD2020, and In the Wild, and M2SFormer is strongest on MISD. On mIoU, SARIF performs best on most datasets, whereas SAFIRE remains stronger on IMD2020 and In the Wild. These results indicate that SARIF is highly competitive across domains and achieves the strongest overall average performance.

Qualitative Result. Fig. 4 presents qualitative comparisons under the CASIAv2-only training scheme. Compared with CNN- and transformer-based baselines, SARIF more often suppresses background leakage and produces sharper boundaries. Overall, SARIF tends to produce sharper and more spatially coherent forgery masks than the compared methods.

4.4 Ablation Studies on FSIE and FGMD

Under the CASIAv2 training protocol, we conduct ablation studies to isolate the contributions of the dual-encoder Forgery-Specific Information Extractor (FSIE) and the Feedback-Guided Mask Decoder (FGMD). Unless otherwise noted, all ablation settings use the same data split, preprocessing, training schedule, and evaluation protocol as the main experiments.

Tab. 3 summarizes the results. Starting from vanilla SAM, adding adapters already provides a substantial improvement, showing that lightweight adaptation is effective for forgery localization. Building on this adapter-based baseline, FSIE and FGMD each provide further gains from different perspectives: FSIE improves the quality of forgery-aware features delivered to the decoder, whereas FGMD improves the quality of the predicted masks through iterative refinement. Combining both modules yields the best overall performance, sug-

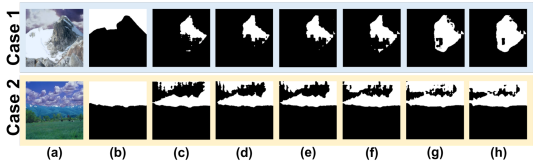

Fig. 5: Failure cases. (a) Input image. (b) Ground truth. (c) initial mask. (d−h) refinement stages 1 to 5. (h) indicates final prediction mask.

Fig. 6: Authentic Image Prediction. Left: Input image, Right: Model prediction.

gesting that forgery-specific cue extraction and feedback-based mask refinement are complementary.

Setting 1 (Vanilla SAM). Directly applying SAM to forgery localization results in poor performance. Because of the domain gap between natural-object segmentation and image forensics, SAM often fails to attend to manipulation-specific traces, leading to inaccurate localization and noisy segmentation masks.

Setting 2 (SAM + Adapters). Adding lightweight adapters to the SAM image encoder significantly improves performance over the vanilla SAM baseline. This result indicates that parameter-efficient adaptation helps the model encode forgery-relevant patterns and partially bridge the domain gap.

Setting 3 (SAM + Adapters + FSIE). When FSIE is added on top of the adapter-based SAM, performance improves further. FSIE extracts forgery-sensitive residual cues from the dual encoder, organizes them into stage-wise prompts, and delivers them to the decoder. Exposing these multi-level cues helps the decoder better localize subtle manipulation artifacts.

Setting 4 (SAM + Adapters + FGMD). Adding FGMD to the adapter-based SAM also improves performance. By feeding the predicted mask back as a mask prompt, FGMD progressively refines the output and reduces boundary ambiguity, resulting in cleaner masks. However, since FGMD mainly improves mask quality given the available cues, its gain is naturally limited when the underlying forgery-specific features are weak.

Setting 5 (SARIF). The full model achieves the best overall performance. FSIE supplies hierarchical forgery-aware cues from the dual encoder, while FGMD refines the prediction through feedback-guided decoding. Their combination leads to more robust localization and cleaner segmentation, confirming that the two components are complementary.

5 Discussion

In this section, we focus on various questions that naturally arise when assessing SARIF's practical usability.

Q1. How does SARIF behave when provided pristine images? Analyzing the behavior on pristine images is critical for practical forensic deployment. As shown in Fig. 6, when a fully authentic image is provided, SARIF doesn't

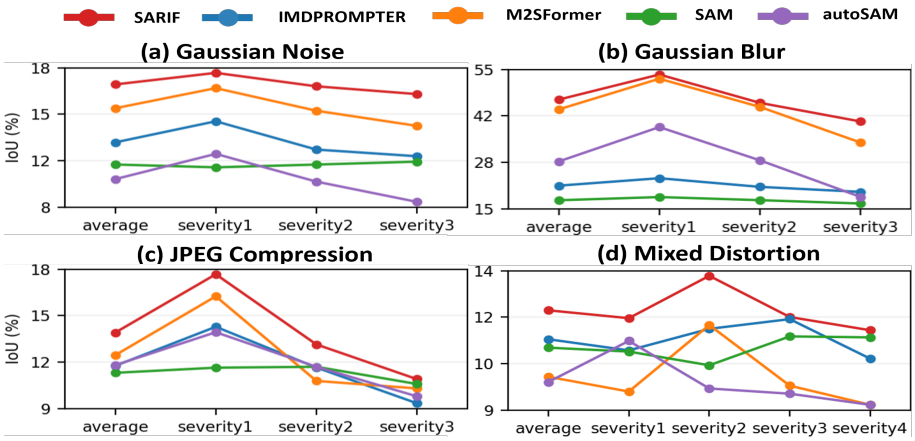

Fig. 7: Various distortion tests on CASIAv2. (a) Gaussian Noise, severity 1, 2, 3 means σ = 0.1, 0.3, 0.5. (b) Gaussian Blur, severity 1, 2, 3 means σ = 3,5,9. (c) JPEG Compression, severity 1, 2, 3 means q = 100, 50, 10. (d) Mixed robustness test, severity 1-4 correspond to (blur & JPEG), (noise & blur), (noise & JPEG), (noise & blur & JPEG), respectively. For each mixed setting, severity level 2 was used for every individual distortion component.

predict an all-zero mask; instead, it may produce sparse low-level responses. Importantly, these responses tend to be ***spatially scattered***, which contrasts with the ***spatially coherent masks*** typically observed on manipulated regions. This observation indicates that SARIF may produce sparse responses on pristine images; we therefore report pristine-image false-positive statistics in the Supplementary Material to quantify this behavior.

Q2. What are the limitations and typical failure cases of SARIF? Fig. 5 presents failure cases of SARIF. **Case 1**. When the initial mask localizes an incorrect region, subsequent refinement tends to reinforce the error and continue refining the wrong area. This indicates that the adapted encoder may capture unreliable forgery cues, causing refinement to amplify an erroneous initial mask. As a result, overall performance largely depends on how effectively the adapter captures forgery cues. **Case 2**. The SAM-based architecture can introduce blocky artifacts because the prompt encoder first projects the mask prediction into a low-resolution embedding space, after which the mask decoder upsamples it with transposed convolution. This design can produce staircase-like artifacts.

Q3. Can SARIF be reliably applied in real-world scenarios? In real-world scenarios, images often suffer from hybrid and variable distortions [75]. To assess practical robustness, we evaluate SARIF under Gaussian noise, Gaussian blur, JPEG compression, and mixed distortions. As shown in Fig. 7, SARIF remains competitive across these settings and maintains strong performance under mixed distortions. These results suggest promising robustness in degraded environments.

Table 4: Model complexity and evaluation on recent forgery datasets

Model	Model Complexity			CocoGlide [19]			TGIF [52]		
	Params(M)	FLOPs(G)	Inference(ms)	DSC	mIoU	AUC	DSC	mIoU	AUC
SAM [41]	13.90	490.65	18.8	17.4 (2.1)	11.7 (1.6)	54.9 (0.9)	13.4 (1.6)	9.1 (1.2)	58.8 (1.2)
autoSAM [70]	158.53	1087.99	21.4	21.8 (1.0)	14.9 (0.5)	56.8 (0.3)	16.9 (0.9)	11.9 (0.8)	61.2 (0.9)
IMDPROMPTER [85]	605.07	989.11	22.5	26.0 (1.4)	17.9 (1.4)	58.8 (0.7)	18.6 (1.5)	12.8 (1.3)	61.6 (1.1)
M2SFormer [61]	107.38	15.16	2.5	12.2 (1.6)	7.7 (1.1)	52.0 (0.6)	14.6 (0.8)	10.3 (0.7)	59.8 (0.3)
SARIF (Ours)	52.62	998.43	47.5	**34.0 (3.7)**	**27.3 (3.1)**	**63.7 (1.8)**	**30.8 (2.6)**	**23.7 (2.1)**	**68.7 (1.7)**

Q4. What is SARIF's computational complexity, and how does it perform on recent challenging forgery datasets? We further evaluated SARIF on more recent and challenging forgery benchmarks, including an inpainting dataset (CocoGlide [19]) and AI-generated forgery dataset (TGIF [52]) in Tab. 4. SARIF remains competitive on recent and challenging forgery benchmarks, including CocoGlide [19] and TGIF, indicating strong reliability and generalization to modern forgery patterns. In addition, we report the overall model complexity in Table 4 and the complexity of each module in Table 3. Although SARIF is computationally more expensive than the recent transformer-based M2SFormer, it consistently achieves clear accuracy gains across multiple datasets. We therefore explicitly present the accuracy–efficiency trade-off.

6 Conclusion

In this paper, we attempted to overcome the limitations of existing forgery localization methods, which are generalization performance and robustness by introducing SARIF—a fully automatic forgery localization framework that leverages SAM foundation-model for broad generalization and channels it into manipulation-aware prompting and refinement. The pipeline initializes from a prompt-free base mask produced by the mask decoder conditioned only on the fine-tuned SAM embeddings. Forgery-specific features are then derived by integrating the fine-tuned and original image embeddings, and combined with a mask prompt encoded from the previous prediction to drive the lightweight SAM mask decoder. Iterative, feedback-guided refinement with deep supervision at every stage improves boundary fidelity and stabilizes training, yielding consistent gains without manual points, boxes or masks. Experiments across heterogeneous datasets indicate strong cross-domain robustness, highlighting the benefit of forgery-specific prompting and feedback-guided mask decoder. Ultimately, by converting task-specific signal into automatic prompts and driving a feedback-guided refinement loop on SAM's lightweight decoder, our method preserves SAM's structural advantages while amplifying manipulation cues and cross-domain generalization and establishing a practical foundation for reliable, large-scale automated forgery localization.

Acknowledgements

This work was supported in part by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-16065822) and in part by Inha University Research Grant.

References

1. Ahmed, S., Chua, H.W.: Perception and deception: Exploring individual responses to deepfakes across different modalities. Heliyon **9**(10) (2023)
2. Bai, N., Wang, X., Han, R., Hou, J., Wang, Y., Pang, S.: Pim-net: Progressive inconsistency mining network for image manipulation localization. Pattern Recognition **159**, 111136 (2025)
3. Barnes, C., Shechtman, E., Finkelstein, A., Goldman, D.B.: Patchmatch: A randomized correspondence algorithm for structural image editing. ACM Trans. Graph. **28**(3), 24 (2009)
4. Bi, X., Wei, Y., Xiao, B., Li, W.: Rru-net: The ringed residual u-net for image splicing forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (2019)
5. Caruso, R.D., Postel, G.C.: Image editing with adobe photoshop 6.0. Radiographics **22**(4), 993–1002 (2002)
6. Chen, H., Wei, P., Guo, G., Gao, S.: Sam-cod+: Sam-guided unified framework for weakly-supervised camouflaged object detection. IEEE Transactions on Circuits and Systems for Video Technology (2024)
7. Chen, X., Dong, C., Ji, J., Cao, J., Li, X.: Image manipulation detection by multi-view multi-scale supervision. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14185–14193 (2021)
8. Chierchia, G., Poggi, G., Sansone, C., Verdoliva, L.: A bayesian-mrf approach for prnu-based image forgery detection. IEEE Transactions on Information Forensics and Security **9**(4), 554–567 (2014)
9. Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y., Barnard, K.: Attentional feature fusion. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 3560–3569 (2021)
10. De Carvalho, T.J., Riess, C., Angelopoulou, E., Pedrini, H., de Rezende Rocha, A.: Exposing digital image forgeries by illumination color classification. IEEE Transactions on Information Forensics and Security **8**(7), 1182–1194 (2013)
11. De Ruiter, A.: The distinct wrong of deepfakes. Philosophy & Technology **34**(4), 1311–1332 (2021)
12. Dirik, A.E., Memon, N.: Image tamper detection based on demosaicing artifacts. In: 2009 16th IEEE International Conference on Image Processing (ICIP). pp. 1497–1500. IEEE (2009)
13. Dong, C., Chen, X., Hu, R., Cao, J., Li, X.: Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 3539–3553 (2022)
14. Dong, J., Wang, W., Tan, T.: CASIA image tampering detection evaluation database. In: 2013 IEEE China Summit and International Conference on Signal and Information Processing. IEEE (Jul 2013). https://doi.org/10.1109/chinasip.2013.6625374, https://doi.org/10.1109/chinasip.2013.6625374

15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=YicbFdNTTy
16. Ferrara, P., Bianchi, T., De Rosa, A., Piva, A.: Image forgery localization via fine-grained analysis of cfa artifacts. IEEE Transactions on Information Forensics and Security **7**(5), 1566–1577 (2012)
17. Gao, D., Zhou, Y., Yan, H., Chen, C., Hu, X.: Cod-sam: Camouflage object detection using sam. Pattern Recognition p. 111826 (2025)
18. Gu, A.R., Nam, J.H., Lee, S.C.: Fbi-net: Frequency-based image forgery localization via multitask learning with self-attention. IEEE access **10**, 62751–62762 (2022)
19. Guillaro, F., Cozzolino, D., Sud, A., Dufour, N., Verdoliva, L.: 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 (CVPR). pp. 20606–20615 (June 2023)
20. Guo, K., Zhu, H., Cao, G.: Effective image tampering localization via enhanced transformer and co-attention fusion. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4895–4899. IEEE (2024)
21. Guo, M.H., Lu, C.Z., Hou, Q., Liu, Z., Cheng, M.M., Hu, S.M.: Segnext: Rethinking convolutional attention design for semantic segmentation. Advances in neural information processing systems **35**, 1140–1156 (2022)
22. Han, R., Wang, X., Bai, N., Wang, Y., Hou, J., Xue, J.: Hdf-net: Capturing homogeny difference features to localize the tampered image. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
23. Hao, J., Zhang, Z., Yang, S., Xie, D., Pu, S.: Transforensics: image forgery localization with dense self-attention. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15055–15064 (2021)
24. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
25. Hsu, Y.F., Chang, S.F.: Detecting image splicing using geometry invariants and camera characteristics consistency. In: International Conference on Multimedia and Expo (2006)
26. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
27. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
28. Hu, X., Zhang, Z., Jiang, Z., Chaudhuri, S., Yang, Z., Nevatia, R.: Span: Spatial pyramid attention network for image manipulation localization. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16. pp. 312–328. Springer (2020)
29. Huang, Y., Huang, J., Liu, Y., Yan, M., Lv, J., Liu, J., Xiong, W., Zhang, H., Cao, L., Chen, S.: Diffusion model-based image editing: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
30. Huh, M., Liu, A., Owens, A., Efros, A.A.: Fighting fake news: Image splice detection via learned self-consistency. In: Proceedings of the European conference on computer vision (ECCV). pp. 101–117 (2018)

31. Hupé, J., James, A., Payne, B., Lomber, S., Girard, P., Bullier, J.: Cortical feedback improves discrimination between figure and background by v1, v2 and v3 neurons. Nature **394**(6695), 784–787 (1998)
32. Hurvich, L.M., Jameson, D.: An opponent-process theory of color vision. Psychological review **64**(6p1), 384 (1957)
33. Jain, Y., Behl, H., Kira, Z., Vineet, V.: Damex: Dataset-aware mixture-of-experts for visual understanding of mixture-of-datasets. Advances in Neural Information Processing Systems **36**, 69625–69637 (2023)
34. Johnson, M.K., Farid, H.: Exposing digital forgeries by detecting inconsistencies in lighting. In: Proceedings of the 7th workshop on Multimedia and security. pp. 1–10 (2005)
35. Kadam, K.D., Ahirrao, S., Kotecha, K.: Multiple image splicing dataset (misd): a dataset for multiple splicing. Data **6**(10), 102 (2021)
36. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6007–6017 (2023)
37. Ke, L., Ye, M., Danelljan, M., Tai, Y.W., Tang, C.K., Yu, F., et al.: Segment anything in high quality. Advances in Neural Information Processing Systems **36**, 29914–29934 (2023)
38. Kee, E., O'brien, J.F., Farid, H.: Exposing photo manipulation from shading and shadows. ACM Transactions on Graphics **33**(5), 1–21 (2014)
39. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
40. Kirchberger, L., Mukherjee, S., Schnabel, U.H., van Beest, E.H., Barsegyan, A., Levelt, C.N., Heimel, J.A., Lorteije, J.A., van der Togt, C., Self, M.W., et al.: The essential role of recurrent processing for figure-ground perception in mice. Science advances **7**(27), eabe1833 (2021)
41. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
42. Korus, P., Huang, J.: Multi-scale analysis strategies in prnu-based tampering localization. IEEE Transactions on Information Forensics and Security **12**(4), 809–824 (2016)
43. Kuffler, S.W.: Discharge patterns and functional organization of mammalian retina. Journal of neurophysiology **16**(1), 37–68 (1953)
44. Kwon, M.J., Lee, W., Nam, S.H., Son, M., Kim, C.: Safire: Segment any forged image region. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4437–4445 (2025)
45. Lamme, V.A., Roelfsema, P.R.: The distinct modes of vision offered by feedforward and recurrent processing. Trends in neurosciences **23**(11), 571–579 (2000)
46. Lee, H., Kang, M., Han, B.: Diffusion-based conditional image editing through optimized inference with guidance. arXiv preprint arXiv:2412.15798 (2024)
47. Liu, C., Li, X., Ding, H.: Referring image editing: Object-level image editing via referring expressions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13128–13138 (2024)
48. Liu, X., Liu, Y., Chen, J., Liu, X.: Pscc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. IEEE Transactions on Circuits and Systems for Video Technology **32**(11), 7505–7517 (2022)
49. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)

50. Lukáš, J., Fridrich, J., Goljan, M.: Detecting digital image forgeries using sensor pattern noise. In: Security, steganography, and watermarking of multimedia contents VIII. vol. 6072, pp. 362–372. SPIE (2006)
51. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
52. Mareen, H., Karageorgiou, D., Wallendael, G.V., Lambert, P., Papadopoulos, S.: Tgif: Text-guided inpainting forgery dataset (2024). `https://doi.org/10.48550/arXiv.2407.11566`, accepted at IEEE WIFS 2024
53. McMahon, M.J., Packer, O.S., Dacey, D.M.: The classical receptive field surround of primate parasol ganglion cells is mediated primarily by a non-gabaergic pathway. Journal of Neuroscience **24**(15), 3736–3745 (2004)
54. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.Y., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. arXiv preprint arXiv:2108.01073 (2021)
55. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
56. Mubarak, R., Alsboui, T., Alshaikh, O., Inuwa-Dutse, I., Khan, S., Parkinson, S.: A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. Ieee Access **11**, 144497–144529 (2023)
57. Nam, J.H., Kang, D.H., Lee, S.C.: M3fpolypsegnet++: Adaptive gaussian frequency decomposition with multi-task learning for polyp segmentation. Neurocomputing p. 131134 (2025)
58. Nam, J.H., Lee, S.C.: Random image frequency aggregation dropout in image classification for deep convolutional neural networks. Computer Vision and Image Understanding **232**, 103684 (2023)
59. Nam, J.H., Lee, S.C.: Fsda: Frequency re-scaling in data augmentation for corruption-robust image classification. Pattern Recognition **150**, 110332 (2024)
60. Nam, J.H., Lee, S.C.: Frequency-based boundary-guided attention network for domain generalizable polyp segmentation from colonoscopy images. Artificial Intelligence in Medicine p. 103288 (2025)
61. Nam, J.H., Moon, D.H., Lee, S.C.: M2sformer: Multi-spectral and multi-scale attention with edge-aware difficulty guidance for image forgery localization. arXiv preprint arXiv:2506.20922 (2025)
62. Nam, J.H., Syazwany, N.S., Kim, S.J., Lee, S.C.: Modality-agnostic domain generalizable medical image segmentation by multi-frequency in multi-scale attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11480–11491 (2024)
63. Novozamsky, A., Mahdian, B., Saic, S.: Imd2020: A large-scale annotated dataset tailored for detecting manipulated images. In: 2020 IEEE Winter Applications of Computer Vision Workshops (WACVW). pp. 71–80 (March 2020)
64. Pan, X., Tewari, A., Leimkühler, T., Liu, L., Meka, A., Theobalt, C.: Drag your gan: Interactive point-based manipulation on the generative image manifold. In: ACM SIGGRAPH 2023 conference proceedings. pp. 1–11 (2023)
65. Pham, N.T., Lee, J.W., Kwon, G.R., Park, C.S.: Hybrid image-retrieval method for image-splicing validation. Symmetry **11**(1), 83 (2019)
66. Popescu, A.C., Farid, H.: Statistical tools for digital forensics. In: International workshop on information hiding. pp. 128–147. Springer (2004)
67. Popescu, A.C., Farid, H.: Exposing digital forgeries in color filter array interpolated images. IEEE Transactions on Signal Processing **53**(10), 3948–3959 (2005)

68. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
69. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
70. Shaharabany, T., Dahan, A., Giryes, R., Wolf, L.: Autosam: Adapting sam to medical images by overloading the prompt encoder. arXiv preprint arXiv:2306.06370 (2023)
71. Tahir, E., Bal, M.: Deep image composition meets image forgery (2024)
72. Tralic, D., Zupancic, I., Grgic, S., Grgic, M.: Comofod—new database for copy-move forgery detection. In: Proceedings ELMAR-2013. pp. 49–54. IEEE (2013)
73. Vaccari, C., Chadwick, A.: Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social media+ society **6**(1), 2056305120903408 (2020)
74. Wang, J., Wu, Z., Chen, J., Han, X., Shrivastava, A., Lim, S.N., Jiang, Y.G.: Objectformer for image manipulation detection and localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2364–2373 (2022)
75. Wang, Y., Li, H., Hou, S., Dong, Z., Gao, R.: Distortion information and edge features guided network for real-world image restoration. Knowledge-Based Systems **314**, 113159 (2025). https://doi.org/10.1016/j.knosys.2025.113159
76. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV). pp. 3–19 (2018)
77. Wu, J., Ji, W., Liu, Y., Fu, H., Xu, M., Xu, Y., Jin, Y.: Medical sam adapter: Adapting segment anything model for medical image segmentation (2023)
78. Wu, Y., AbdAlmageed, W., Natarajan, P.: Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9543–9552 (2019)
79. Xu, X., Chen, J., Lv, W., Wang, W., Zhang, Y.: Image tampering detection with frequency-aware attention and multi-view fusion. IEEE Transactions on Artificial Intelligence (2024)
80. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18381–18391 (2023)
81. Yang, J., Zhang, Y., Zhu, G., Kwong, S.: A clustering-based framework for improving the performance of jpeg quantization step estimation. IEEE transactions on circuits and systems for video technology **31**(4), 1661–1672 (2020)
82. Yang, Z., Liu, B., Bi, X., Xiao, B., Li, W., Wang, G., Gao, X.: D-net: A dual-encoder network for image splicing forgery detection and localization. Pattern Recognition **155**, 110727 (2024)
83. Zhang, L., Zhu, X., He, D., Liao, X., Sun, B.: Samif: Adapting segment anything model for image inpainting forensics. In: Proceedings of the Asian Conference on Computer Vision. pp. 3605–3621 (2024)
84. Zhang, L., Lu, J., Zheng, S., Zhao, X., Zhu, X., Fu, Y., Xiang, T., Feng, J., Torr, P.H.: Vision transformers: From semantic segmentation to dense prediction. International Journal of Computer Vision **132**(12), 6142–6162 (2024)

85. Zhang, Q., Qi, Y., Tang, X., Fang, J., Lin, X., Zhang, K., Yuan, C.: Imdprompter: Adapting sam to image manipulation detection by cross-view automated prompt learning. In: International Conference on Learning Representations (2025), https://openreview.net/forum?id=vl7kf0YHwj
86. Zhang, Y., Zhu, G., Wu, L., Kwong, S., Zhang, H., Zhou, Y.: Multi-task se-network for image splicing localization. IEEE Transactions on Circuits and Systems for Video Technology **32**(7), 4828–4840 (2021)
87. Zhou, P., Han, X., Morariu, V.I., Davis, L.S.: Learning rich features for image manipulation detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1053–1061 (2018)
88. Zhu, J., Hamdi, A., Qi, Y., Jin, Y., Wu, J.: Medical sam 2: Segment medical images as video via segment anything model 2 (2024)