SAFIRE: Segment Any Forged Image Region

Myung-Joon Kwon[1*], **Wonjun Lee**[1*], **Seung-Hun Nam**[2], **Minji Son**[1], **Changick Kim**[1]

[1]School of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST)
[2]NAVER WEBTOON AI, Seongnam, South Korea
mjkwon2021@gmail.com, dpenguin@kaist.ac.kr, shnam1520@gmail.com, mjson0103@gmail.com, changick@kaist.ac.kr

Abstract

Most techniques approach the problem of image forgery localization as a binary segmentation task, training neural networks to label original areas as 0 and forged areas as 1. In contrast, we tackle this issue from a more fundamental perspective by partitioning images according to their originating sources. To this end, we propose Segment Any Forged Image Region (SAFIRE), which solves forgery localization using point prompting. Each point on an image is used to segment the source region containing itself. This allows us to partition images into multiple source regions, a capability achieved for the first time. Additionally, rather than memorizing certain forgery traces, SAFIRE naturally focuses on uniform characteristics within each source region. This approach leads to more stable and effective learning, achieving superior performance in both the new task and the traditional binary forgery localization.

Code & Data — https://github.com/mjkwon2021/SAFIRI
Extended version — https://arxiv.org/abs/2412.08197

Introduction

In the era of artificial intelligence (AI), the proliferation of image editing software (Fu et al. 2023; Yu et al. 2023) and sophisticated generative models (Rombach et al. 2022; Ho, Jain, and Abbeel 2020) has made image forgery more accessible and more challenging to detect than ever before (Lin et al. 2024). The ease of image manipulation critically affects areas where the integrity of visual information is crucial, including the spread of fake news in journalism, the use of counterfeit evidence in law enforcement, and the presence of fabricated microscopy images in biomedical research (Verdoliva 2020; Sabir et al. 2021). Therefore, detecting and precisely localizing forgeries within an image is crucial for maintaining trust in digital media.

Currently, most image forensics methods address the problem of **image forgery localization (IFL)** through binary segmentation (Guillaro et al. 2023; Kwon et al. 2022; Liu et al. 2022; Dong et al. 2022; Hu et al. 2020; Wu et al. 2022; Zhou et al. 2023a; Ji et al. 2023a; Sun et al. 2023). That is, within an image, regions that remain unchanged from the camera capture are labeled as 0, and regions that have been manipulated are labeled as 1, to train deep neural networks.

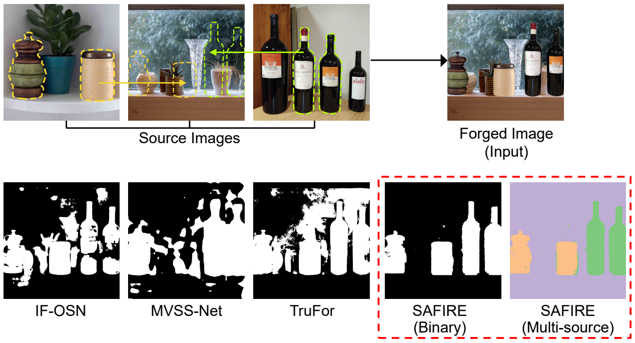

Figure 1: The forged image is composed of three source regions. Previous methods are limited to binary prediction — segmenting forged regions. In contrast, our SAFIRE is also capable of multi-source prediction — distinguishing regions that originate from the same source images.

Instead, we view the IFL from a more fundamental perspective of *partitioning an image into distinct regions based on their origins*. In this context, we define these distinct regions as **source regions**, which are distinct segments of an image that have been independently captured, AI-generated, or manipulated (Fig. 1).

From this perspective, we propose **S**egment **A**ny **F**orged **I**mage **Re**gion (**SAFIRE**), a novel point prompt-based IFL method designed to precisely partition images into regions based on their original sources. SAFIRE employs point prompting, where each point on an image segments the area that shares the same source (Fig. 2).

To achieve this, we capitalize on the Segment Anything Model (SAM)'s (Kirillov et al. 2023) point prompting capability with several differences compared to the original SAM. First, SAFIRE segments a source region containing the given point whereas SAM segments any meaningful chunk around the point. Second, while SAM deals with ambiguous ground truths, SAFIRE has a clear ground truth, where all points on one source region share the same answer. Third, SAFIRE generates and uses point prompts internally,

[*]Equal contribution.

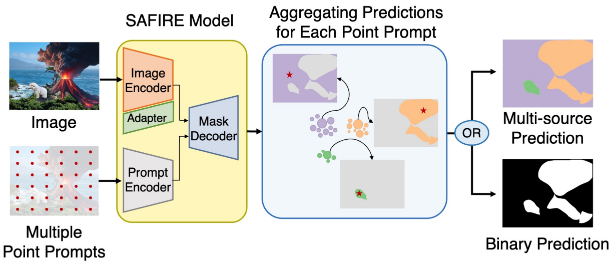

Figure 2: Overview of how SAFIRE conducts IFL. An image along with point prompts are input into the model. The model segments the source region containing each point, and these results are combined to produce the final output

so there is no need for manual input to designate points.

The SAFIRE framework consists of pretraining, training, and inference phases. In the pretraining phase, source region-based contrastive learning is applied to enhance the feature extraction ability of the image encoder. In the training phase, the model is trained to segment the source region corresponding to a given point prompt. Although the model can be trained using forgery datasets with only binary labels, it can perform multi-source prediction. During inference, a grid of points generates multiple masks, which are then combined to produce the final source partitioning result.

SAFIRE is the first method capable of distinguishing each source when an image has been forged twice or more, resulting in three or more sources. Differentiating each source provides a better explanation of the manipulated image than simply locating forged pixels. Additionally, it facilitates subsequent analysis, such as provenance filtering, which involves retrieving the donor image for each source region in a set of candidate images (Pinto et al. 2017; Moreira et al. 2018; Verdoliva 2020). Therefore, multi-source partitioning is particularly beneficial for image forensics in real-world scenarios where multiple manipulations are common.

Furthermore, the novel prompting enables SAFIRE to learn effectively by considering the interchangeable nature of authentic and tampered regions, a characteristic we refer to as label agnosticity in IFL. Tampered areas often lack common traces and are simply different sources within the image compared to authentic areas (Huh et al. 2018). Consequently, attempts to memorize forgery traces lead to confusion and result in unstable learning. In contrast, SAFIRE uses points as references to learn the uniform characteristics of each source region, rather than memorizing forgery traces. This approach results in stable and effective learning, achieving high performance in both traditional binary IFL and new source partitioning tasks.

As the first paper to solve IFL through multi-source partitioning, we have created a dataset, **SafireMS**, composed of multi-source images to promote further research in this area. We plan to make it publicly available.

Our primary contributions can be summarized as follows:

- We introduce a new IFL task that partitions forged images by each originating source. It helps in understanding the composition of the forged image and makes further analysis easier.
- We propose SAFIRE, a novel IFL method that uses point prompting internally. It is the first technique capable of multi-source partitioning, yet it can be trained using traditional binary datasets.
- Extensive experiments show that SAFIRE demonstrates top performance in both the traditional binary IFL and the new task.
- To facilitate the research on the new task, we construct and release a forgery dataset containing images composed of multiple sources.

Related Work

Image Forgery Localization

Effective extraction of forensic clues is essential in IFL. This often involves determining which forensic fingerprints to be utilized. These artifacts, often low-level and inconspicuous, include local CFA artifacts (Bammey, Gioi, and Morel 2020), edge information (Dong et al. 2022; Li et al. 2023), JPEG compression artifacts (Kwon et al. 2021, 2022), unique traces left by different camera models (Guillaro et al. 2023), and explicitly enhanced noise (Zhu et al. 2024).

Designing network architectures specifically tailored for forensics is another key component in solving IFL. This includes the application of steganalysis filter (Zhou et al. 2018), various low-level filters and anomaly-enhancing pooling (Wu, AbdAlmageed, and Natarajan 2019), utilizing both top-down and bottom-up paths (Liu et al. 2022), and efficient modeling of internal relationships using Transformers (Hu et al. 2020; Hao et al. 2021; Wang et al. 2022; Zeng et al. 2024).

Learning the common characteristics of an image and using consistency as a criterion is also a viable approach. The pioneering study (Huh et al. 2018) trains a model using self-supervised learning to determine if two image patches have the same EXIF metadata and uses clustering to check consistency. Subsequent research has utilized the consistency of camera model fingerprints (Cozzolino and Verdoliva 2019) or employed various contrastive learning techniques to utilize consistent features (Zhou et al. 2023a; Wu, Chen, and Zhou 2023; Niloy, Bhaumik, and Woo 2023).

Our method aligns with this research trend of using consistency but stands out in several key ways. Firstly, we employ a symmetric pretraining approach that focuses on source regions without distinguishing between authentic and tampered areas. Secondly, we use point prompting, which enables multi-source partitioning and addresses label agnosticity. Lastly, clustering is performed at the prediction map level, rather than on all patch pairs or individual pixels.

Segment Anything Model

Foundation models have first emerged in the field of Natural Language Processing (NLP), which is pretrained on large datasets and then fine-tuned across a variety of sub-tasks or domains for specific applications (Bommasani et al. 2021). These NLP-based foundation models have demonstrated breakthrough performance in natural language understanding and generation tasks. Their impact has expanded to other AI domains, including computer vision and speech recognition, leading to models like CLIP (Radford et al. 2021) and wav2vec 2.0 (Baevski et al. 2020).

Recently, Meta AI introduced SAM (Kirillov et al. 2023) as the first foundation model for image segmentation. As a prompt-based model, SAM accepts point prompts, bounding boxes, and masks. Furthermore, its design allows for integration with other models to handle text prompts, enabling flexible integration with other systems. SAM has been fine-tuned and applied to various domains such as polyp segmentation (Li, Hu, and Yang 2023; Zhou et al. 2023b), camouflaged object detection (Tang, Xiao, and Li 2023), and others (Ji et al. 2023b), with related research publications keep emerging.

SAM in IFL. Very recently, there have been a few attempts to use SAM in IFL techniques. One approach (Su, Tan, and Huang 2024) constructs an IFL model by adding an SRM filter (Zhou et al. 2018) to SAM. It completely removes the prompt encoder, essentially using SAM as a modern segmentation backbone. Another study (Karageorgiou, Kordopatis-Zilos, and Papadopoulos 2024) fuses various signals using attention, and to assist with this attention, it employs a pretrained and frozen SAM for instance segmentation.

In summary, previous studies have primarily used SAM either as a backbone or solely to obtain segmentation masks. These approaches neglect SAM's most significant feature—its promptable capability—and fail to fully leverage its potential. Meanwhile, we pioneer the application of promptable segmentation models for partitioning images into source regions. By using SAM-based point prompts, we enable each point to serve as a reference, segmenting the source region that contains it. Moreover, inspired by SAM's automatic mask generation process, we propose an inference technique that involves placing points on the image in a grid pattern and aggregating the results. This approach enables, for the first time, multi-source partitioning.

Method

Overview

The core methodology we propose, the SAFIRE framework, refers to the pretraining, training, and inference processes for IFL. The neural network used within this framework is termed the SAFIRE model, which does not require a specific structure and can be freely modified. In this paper, we utilize a slightly altered structure from SAM, adding only adapter layers to the image encoder for enhancing the model's capability to extract forensics features by utilizing low-level signals (outlined in Fig. 2 and detailed in Appendix). It consists of an image encoder $E(\cdot)$, prompt encoder $F(\cdot)$, and mask decoder $D(\cdot,\cdot)$. The model takes an image I and a point prompt P as inputs and outputs a prediction map X and confidence score s for the source region that includes the point.

The upcoming sections will delve into a detailed explanation of the SAFIRE framework. Initially, for effective source image partitioning, the image encoder is pretrained through region-to-region contrastive learning. Subsequently, in the main training phase, the model is trained on source region segmentation using point prompts. In the final inference stage, multiple points are fed into the model in a grid formation, and all the results are aggregated to obtain the final prediction heatmap.

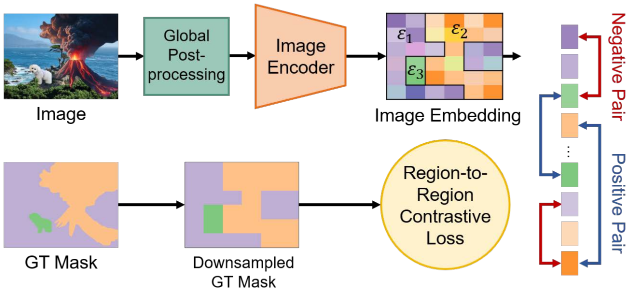

Figure 3: Pretraining. Features originating from the same source region become closer in the feature space, while those from different source regions move apart, enabling the image encoder to learn information that distinguishes source regions.

Pretraining: Region-to-Region Contrastive Learning

We propose **Region-to-Region Contrastive Learning** to pretrain the image encoder for effective source region partitioning (Fig. 3). This approach aims to have embeddings from the same source region close together in the feature space, while those from different source regions are distanced, when an image consists of two or more sources.

Leveraging the proven effectiveness of the InfoNCE loss in contrastive learning (Oord, Li, and Vinyals 2018), we define our loss function as follows. Let $I \in \mathbb{R}^{3\times H\times W}$ be an input image composed of r sources, $E(\cdot)$ the image encoder, and $\mathcal{E} = E(I) \in \mathbb{R}^{V\times\frac{H}{K}\times\frac{W}{K}}$ the image embeddings with downsampling ratio K. With a slight abuse of notation, we treat $\mathcal{E}$ as a set of V-dimensional image embeddings. Then there are $\frac{H}{K}\times\frac{W}{K}$ embeddings $q \in \mathbb{R}^V$ in $\mathcal{E}$. We also let $\{\mathcal{E}_i\}_{i=1}^r$ be the partition of $\mathcal{E}$ which corresponds to source regions in I.

Then we define the region-to-region contrastive loss $\mathcal{L}_{R2R}$ as:

$$InfoNCE(q,p,N) = -\log\left(\frac{\exp\left(\frac{q\cdot p}{\tau}\right)}{\exp\left(\frac{q\cdot p}{\tau}\right)+\sum_{n\in N}\exp\left(\frac{q\cdot n}{\tau}\right)}\right), \quad (1)$$

$$\mathcal{L}_{R2R} = \frac{1}{|\mathcal{E}|}\sum_{i=1}^{r}\sum_{q\in\mathcal{E}_i} InfoNCE\Big(q,\overline{\mathcal{E}_i\backslash\{q\}},\mathcal{E}\backslash\mathcal{E}_i\Big), \quad (2)$$

where τ is a hyperparameter called temperature, $|\cdot|$ returns the number of elements and $\bar{\cdot}$ returns the average over all elements.

Before the image passes through the image encoder, global post-processing such as various blurring, noise addition, or contrast changes are probabilistically applied to it. By doing so, we expect the image encoder to become robust

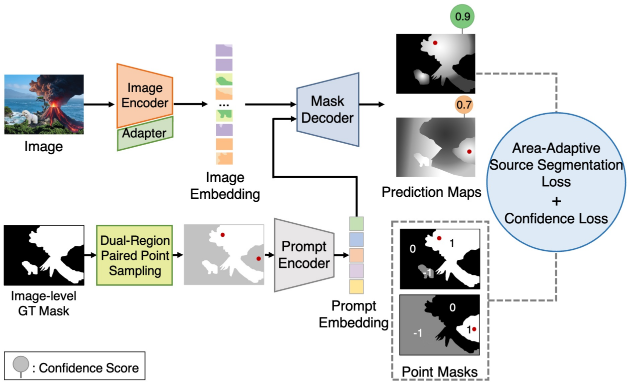

Figure 4: Training. The adapter and mask decoder are trained to segment the source region that includes the given point effectively. Furthermore, it is trained to output a confidence score of this prediction map for inference purposes.

to global common variations and focus more on fine local distinctions.

Given the large size of the image encoder, we have determined that the currently available public forgery datasets are insufficient in scale and noisy. Therefore, we generate and use a large-scale noise-free dataset, SafireMS-Auto. More in Appendix.

Training: Source Region Segmentation Using Point Prompts

Upon completion of the image encoder pretraining, the SAFIRE model undergoes the main training to accurately segment the source region in response to the specified point prompt (Fig. 4). Both the image encoder and prompt encoder are frozen: the image encoder in its pretrained state and the prompt encoder in its original SAM state. The adapter component and mask decoder are trained by feeding image embeddings and prompt embeddings into the mask decoder, ensuring the output aligns with the correct mask.

Point Mask Creation. During the training process, it is necessary to transform the image-level ground truth mask into a mask corresponding to the given point, which we call a point mask. If there is a multi-source mask where different labels are assigned to each source region, then a point mask could be simply created by assigning 1 to the source region that contains the point and 0 otherwise. However, almost all of the datasets currently available for IFL tasks are in only binary form, marking manipulated parts as 1 and unaltered parts as 0.

We introduce a methodology to convert these image-level binary masks into point masks. If a manipulated image uses only two sources, the areas marked as 0 and 1 would each represent a single source region. Taking a step further, we also consider connected components. A connected region containing a given point is marked as 1, and other connected regions neighboring this region are labeled as 0. Regions that are not neighboring it are assigned an ignore label of -1, which is ignored when calculating losses. This transformation allows us to train for multi-source partitioning using only datasets with binary labels.

To be specific, let $Y \in \{0,1\}^{H\times W}$ be the ground truth mask for an image I which contains c connected components, $R = \{(i,j) \in \mathbb{Z}^2 : 0 \leq i < H, 0 \leq j < W\}$ a set of integer coordinates of I, $\{R_i\}_{i=1}^{c}$ the partition of R covering connected components of Y, $P \in R$ a point prompt, and R^P a region contains P which is one of $\{R_i\}_{i=1}^{c}$. Then the point mask $Y_P \in \{-1,0,1\}^{H\times W}$ can be computed as:

$$Y_P[i,j] = \begin{cases} 1, & \text{if } (i,j) \in R^P \\ 0, & \text{if } (i,j) \in neighbors(R^P) \\ -1, & \text{otherwise} \end{cases}, \quad (3)$$

where $neighbors(\cdot)$ returns the union of neighboring regions.

Dual-Region Paired Point Sampling. The image encoder computes image embeddings independently of the point prompts. Maximally leveraging this feature, efficient training can be achieved by simultaneously processing multiple point prompts for a single image. Furthermore, to balance the source regions, points were always sampled in pairs from regions marked as 0 and 1 based on the image-level ground truth.

Area-Adaptive Source Segmentation Loss. For each point, we can define a loss function that minimizes the difference between the prediction map and the point mask (Fig. 4). Here, not all pixels within a point mask contribute equally to the loss because doing so would result in smaller areas being overlooked. Traditional IFL techniques have addressed the similar issue of manipulated areas being small in most images by assigning greater weight to tampered class (Kwon et al. 2022). However, in our point masks, there is no distinction between manipulated and pristine regions; there exist only multiple source regions. Therefore, we use a strategy that assigns greater weight to smaller areas within each point mask, regardless of whether the correct label in those areas is 0 or 1. This differs from the class-specific weights used in most semantic segmentation tasks in that the weights are calculated within a single image (Wang et al. 2020).

Let I be an input image, P a point prompt, $(X, s) = D(E(I), F(P))$ the output of the mask decoder where X is the prediction map and s is the confidence score, and Y_P the ground truth point mask for P. We only compute the loss within the valid label region $R^{Y_P,\{0,1\}}$ by letting $R^{A,B} = \{(i,j) \in R : A[i,j] \in B\}$. Then the Area-Adaptive Source Segmentation Loss $\mathcal{L}_{AASS}$ is defined as:

$$\begin{aligned} \mathcal{L}_{AASS} = -\underset{(i,j)}{\mathbb{E}} \big[& w_1 \cdot Y[i,j] \cdot \log\left(\sigma(X[i,j])\right) \\ & + w_0 \cdot (1 - Y[i,j]) \cdot \log\left(1 - \sigma(X[i,j])\right)\big], \end{aligned}$$

$$w_n = \min\left(\frac{|R^{Y_P,\{0,1\}}|}{|R^{Y_P,\{n\}}|}, C_{AASS} \right) \quad \text{for } n \in \{0,1\}, \quad (4)$$

where the expectation is calculated over $R^{Y_P,\{0,1\}}$, $\sigma(\cdot)$ is a sigmoid function, and C_{AASS} is a hyperparameter to limit the weight.

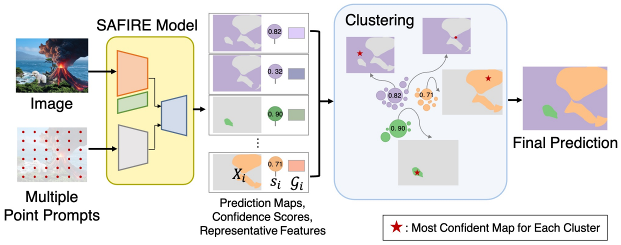

Figure 5: Inference. Multiple points in a grid pattern are input, and a prediction map is obtained for each point. Clustering is performed using the corresponding representative features, and the final prediction is produced.

Confidence Loss. For use in inference, the mask decoder also predicts confidence scores. Unlike SAM which predicts box-level mean intersection over union (mIoU) score to be predicted, our model predicts pixel accuracy to measure the performance globally rather than rectangular mIoU. The confidence score loss $\mathcal{L}_{conf}$ is defined as:

$$\mathcal{L}_{conf} = \underset{R^{Y_P,\{0,1\}}}{MSE} \left(acc \left(bin(X), Y_P \right), s \right), \tag{5}$$

where $bin(\cdot)$ thresholds the input into binary maps, converting values greater than 0 to 1 and values less than or equal to 0 to 0, $MSE(\cdot,\cdot)$ returns the pixel-wise mean squared error, and $acc(\cdot)$ returns the accuracy.

Total Training Loss. Finally, we obtain the total training loss $\mathcal{L}_{train}$ as follows:

$$\mathcal{L}_{train} = \mathcal{L}_{AASS} + \lambda_{conf} \cdot L_{conf}, \tag{6}$$

where λ_{conf} is a hyperparameter balancing the two losses.

Inference: Multiple Points Aggregation

Inference is conducted using multiple point prompts (Fig. 5). Alongside the image to be inferred, points are provided as input to the model in a grid format (for example, 16×16). The output masks are aggregated to obtain the final prediction, which could be a multi-source map or binary map.

Let I be an input image, and $P_1, \cdots, P_N$ point prompts. First, we compute the image embedding $\mathcal{E} = E(I)$ and prompt embeddings $\mathcal{F}_i = F(P_i)$ for all i. Since image embedding extraction is independent of the point prompt, it is performed only once per image. Thus, the total computation does not increase too much even if we use many points.

Thereafter, the image embedding and point embeddings pass through the mask decoder and so a prediction corresponding to each point can be obtained. The output of the mask decoder $D(\cdot,\cdot)$ can be expressed as:

$$(\{X_1, \cdots, X_N\}, \{s_1, \cdots, s_N\}) = D(\mathcal{E}, \{\mathcal{F}_1, \cdots, \mathcal{F}_N\}), \tag{7}$$

where X_i is a prediction map and s_i is a confidence score of X_i.

The next step is to compute a representative feature for each prediction X_i, which is the average of image embeddings corresponding to the prediction area. We define a function $g : \mathbb{R}^{H \times W} \to \mathbb{R}^V$ by:

$$g(X) = \frac{1}{|\mathcal{R}^{bin(\mathcal{X}),\{1\}}|} \sum_{(i,j) \in \mathcal{R}^{bin(\mathcal{X}),\{1\}}} \mathcal{E}[i,j], \tag{8}$$

where $\mathcal{R}$ is a set of integer coordinates of $\mathcal{E}$ and $\mathcal{X}$ is the downsampled prediction map of X to match the resolution with $\mathcal{R}$. Here, $\mathcal{R}^{bin(\mathcal{X}),\{1\}}$ represents the set of coordinates in the embedding space corresponding to the area segmented by the prediction X. The representative features can be expressed as $\mathcal{G}_i = g(X_i)$ for all i.

Subsequently, we cluster the representative features. The clustering is predicated on the assumption that the SAFIRE model accurately extracts features, which results in features from the same source region being gathered together. We cluster $\{\mathcal{G}_1, \cdots, \mathcal{G}_N\}$ into M clusters $C_1, \cdots, C_M$. Any clustering algorithm could be applied and M can be fixed in advance or regressed by the algorithm. For general source region partitioning, we may allow the algorithm to determine the proper M. In situations where the number of sources is known, algorithms with a fixed number of clusters can be used.

Afterward, the most confident mask from each cluster is selected. Each cluster represents one source region of the input image and the most confident mask corresponds to the best prediction of it. We collect indices of the maximum confidence scores for each cluster:

$$j^* = \underset{\mathcal{G}_i \in C_j}{\text{argmax}}\, s_i. \tag{9}$$

Finally, these masks are combined to obtain the final prediction. The simplest method is taking the softmax:

$$X^* = \text{softmax}\{X_{1^*}, \cdots, X_{M^*}\}. \tag{10}$$

For the special case when $M = 2$, to obtain a binary prediction map, the simple average of the two predictions produces an effective output:

$$X^* = \frac{1}{2}\{\sigma(X_{1^*}) + (1 - \sigma(X_{2^*}))\}. \tag{11}$$

Experiments on binary IFL

We begin with the traditional task of localizing forged regions in images. Note that SAFIRE can make binary predictions as well as multi-source predictions.

Experimental Settings

Implementation Details. Our model undergoes pretraining followed by training. The temperature τ for the region-to-region contrastive learning in Eq. (1) is set to 0.1. The weight limit C_{AASS} for the AASS loss in Eq. (4) is set to 10 and $\lambda_{conf} = 0.1$ in Eq. (6). During the inference phase, M is fixed to 2 to obtain predictions in binary form. We use 16×16 point prompts and k-means clustering.

Datasets. We train the network using a commonly adopted setting (Guillaro et al. 2023) that incorporates four datasets (Kniaz, Knyaz, and Remondino 2019; Novozamsky, Mahdian, and Saic 2020; Dong, Wang, and Tan 2013; Kwon et al.

Method	Col.	COV.	CG.	RT.	NC16	Avg.
EXIF-SC	78.8	16.2	29.6	14.4	16.8	31.2
ManTraNet	50.5	31.2	51.6	21.5	19.9	35.0
SPAN	39.7	16.1	29.6	8.7	11.2	21.1
AdaCFA	58.4	17.9	28.7	22.3	11.4	27.7
CAT-Net v2	85.8	37.6	43.3	14.3	28.2	41.8
IF-OSN	74.7	29.9	42.8	33.4	32.5	42.7
MVSS-Net	72.7	50.8	48.6	17.5	32.7	44.5
PSCC-Net	60.0	46.6	51.7	9.7	13.4	36.3
TruFor	85.7	58.7	52.2	**43.2**	41.6	56.3
NCL	47.3	21.3	35.8	14.5	-	-
SAM	40.0	18.1	33.9	8.1	11.2	22.3
SAFIRE (Ours)	**97.9**	**63.4**	**63.5**	39.3	**48.8**	**62.6**

Method	Col.	COV.	CG.	RT.	NC16	Avg.
EXIF-SC	92.9	31.7	42.0	28.6	33.6	45.8
ManTraNet	64.6	47.9	67.3	26.4	27.7	46.8
SPAN	49.9	24.8	38.1	15.6	16.8	29.0
AdaCFA	62.7	21.5	36.3	24.1	14.0	31.7
CAT-Net v2	92.1	57.5	60.3	24.3	43.4	55.5
IF-OSN	82.9	46.4	59.1	46.6	45.6	56.1
MVSS-Net	78.1	64.6	64.2	29.4	44.0	56.0
PSCC-Net	75.7	60.6	68.5	18.3	28.4	50.3
TruFor	91.3	72.1	72.0	**53.3**	54.7	68.7
NCL	62.5	32.4	49.7	27.0	-	-
SAM	47.8	35.7	46.3	17.5	23.3	34.1
SAFIRE (Ours)	**99.7**	**76.9**	**76.9**	49.9	**61.4**	**73.0**

Table 1: IFL results using F1 fixed (%, top) and F1 best (%, bottom). Under both metrics, SAFIRE achieves the highest forgery localization performance in four out of five datasets and is ranked first in terms of average score.

2022) which consists of real and fake images, also known as the CAT-Net (Kwon et al. 2022) setting. We test the performance using five public datasets which have no overlap with the training datasets: Columbia (Ng, Chang, and Sun 2004), COVERAGE (Wen et al. 2016), CocoGlide (Guillaro et al. 2023), RealisticTampering (Korus and Huang 2016), and NC16 (Guan et al. 2019). These consist of various forgery types including splicing, copy-move, removal, and adding objects using generative models. During testing, images are input in their original form, except for the NC16 dataset, where images were scaled down due to memory constraints in some comparative methods.

Comparative Methods. Following the protocol from (Guillaro et al. 2023), we ensure a fair comparison by selecting recent techniques with publicly accessible code and pretrained models, trained on disjoint datasets from test sets. Namely, these include ManTra-Net (Wu, AbdAlmageed, and Natarajan 2019), SPAN (Hu et al. 2020), AdaCFA (Bammey, Gioi, and Morel 2020), CAT-Net v2 (Kwon et al. 2022), IF-OSN (Wu et al. 2022), MVSS-Net (Dong et al. 2022), PSCC-Net (Liu et al. 2022), TruFor (Guillaro et al. 2023), and NCL (Zhou et al. 2023a). Furthermore, we also utilize a pure SAM (Kirillov et al. 2023) model trained on the same datasets.

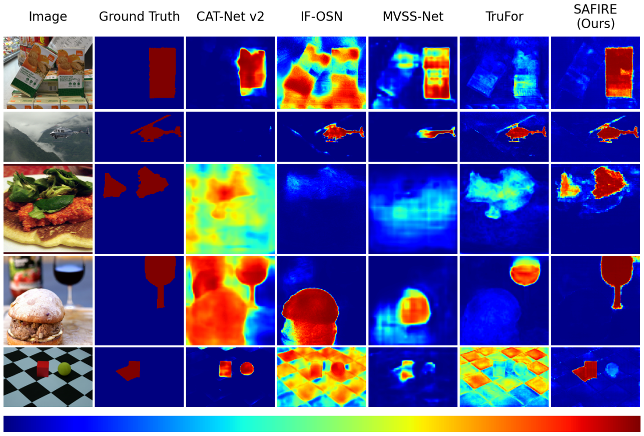

Figure 6: Visualization of IFL. The colors indicate the confidence of forged pixels.

Metrics. We evaluate the localization performance in the same manner as conducted in the TruFor (Guillaro et al. 2023) paper. Specifically, performances are reported in terms of permuted F1 score (Huh et al. 2018; Kwon et al. 2022) using both fixed 0.5 threshold (F1 fixed) and best threshold per image (F1 best).

Evaluation Results

Table 1 shows a comparative analysis of binary IFL performance. A hyphen ('-') indicates that the dataset was used for training, and so excluded. Notably, SAFIRE shows superior performance on both F1 fixed and F1 best, achieving the first rank in four out of five datasets. Moreover, the average score on all datasets reaffirms the superiority of SAFIRE, securing the best position in overall performance. In addition, in the Appendix, we can see that our method also outperforms other techniques under various global post-processing conditions, proving its robustness.

Figure 6 shows the qualitative results of IFL produced by each model. SAFIRE successfully identifies sophisticated and challenging manipulations that other techniques fail to detect, with less false positive detection. In particular, SAFIRE achieves significantly more accurate predictions in sophisticated AI-generated partial manipulations compared to other techniques.

Ablation Study. To ensure the completeness of our study, we conduct an ablation study on the key components of our framework: Region-to-Region Contrastive Loss, area adaptive feature in Area-Adaptive Source Segmentation Loss, point prompting, and Confidence Loss (Table 2). We substitute each with a conventional counterpart for comparison.

The results demonstrate diminished performance in the absence of any single component compared to the full SAFIRE framework, which integrates all four. Furthermore, a baseline model excluding all four key features exhibits significantly inferior results, underscoring the indispensable role of these four components in SAFIRE.

Especially, we observe that source region partitioning based on prompting outperforms binary segmentation. A

Setting	R2R Loss	AASS Loss	Prompting	Conf. Loss	F1 fixed	F1 best
SAFIRE	✓	✓	✓	✓	**62.6**	**73.0**
R2R → SAM pretraining	-	✓	✓	✓	48.6	61.0
No area adaptation	✓	-	✓	✓	56.9	71.5
Prompting → Binary seg.	✓	✓	-	-	28.0	37.4
Using random mask	✓	✓	✓	-	43.5	54.5
Baseline	-	-	-	-	22.3	34.1

Table 2: Ablation study (%). All core components of SAFIRE contribute to the performance of IFL. The usage of prompting is found to be critical to performance.

Method	p_mIoU (%)			ARI (%)		
	2src	3src	4src	2src	3src	4src
CAT-Net v2	53.8	-	-	25.1	-	-
IF-OSN	54.8	-	-	25.2	-	-
MVSS-Net	51.9	-	-	22.5	-	-
TruFor	82.5	-	-	72.3	-	-
SAFIRE (k-means)	**90.3**	**62.2**	**54.0**	**80.7**	57.8	55.2
SAFIRE (DBSCAN)	89.7	57.6	48.9	80.2	**60.2**	**59.4**

Table 3: Multi-source partitioning results of SafireMS-Expert. SAFIRE can accurately partition images into multiple sources based on their origins.

model with the same structure and pretraining achieves only a 28.0% F1 fixed score when using binary segmentation. However, when employing prompting for source partitioning, the model's performance significantly improves, reaching 62.6%. This demonstrates the effectiveness of SAFIRE's prompting approach in enabling the network to understand the characteristics of the same source regions, resulting in stable learning and outstanding performance.

Experiments on Multi-source Partitioning

One of the unique advantages of our method is its ability to partition images composed of three or more sources into each source. To show this capability, we conduct additional experiments on multi-source partitioning.

Experimental Settings

Implementation Details. We use the same model as used in binary IFL without fine-tuning. We consider two settings for inference: the number of sources is given in advance and it is determined by the method. For the former, we use k-means clustering as done in binary IFL. For the latter, we utilize DBSCAN which automatically chooses the number of clusters in the data distribution.

Datasets. Given that traditional forgery datasets lack instances labeled with multiple source regions, we manually constructed a multi-source dataset, dubbed SafireMS-Expert, to assess our framework's effectiveness in handling such scenarios. The forgery types include splicing, removal by AI-based inpainting (Yu et al. 2023), reconstructing some objects using generative models, and adding objects by generative models using text prompts (Zhang, Rao, and Agrawala 2023). More in Appendix.

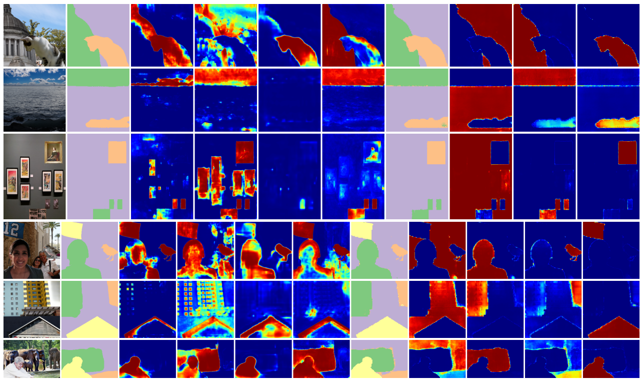

Figure 7: Visualization of multi-source IFL. Each color represents a single source region. Methods other than SAFIRE could only produce binary results. From left to right: Image, Ground Truth, CAT-Net v2, IF-OSN, MVSS-Net, TruFor, SAFIRE, SAFIRE prediction maps for each source.

Metrics. We propose to use permuted mIoU (p_mIoU) and Adjusted Rand Index (ARI) for source region partitioning. We generalize p_mIoU defined on (Huh et al. 2018) for an arbitrary number of source regions. More in Appendix.

Evaluation Results

Table 3 and Figure 7 present the quantitative and qualitative results of multi-source partitioning, respectively. While existing methods based on binary segmentation cannot split images into three or more sources (marked with '-'), SAFIRE can do it, even though it is trained using only binary datasets. These visualizations offer a better interpretation of forged images for humans, as manipulation often occurs multiple times in real-world scenarios. Additionally, SAFIRE's ability to estimate the number of sources is a valuable feature. More results are in the Appendix.

Conclusion

Moving beyond the conventional approach of viewing the IFL tasks through binary segmentation, SAFIRE resolved this issue by partitioning images into multiple originating regions. Through region-to-region contrastive pretraining, we guided the encoder to effectively embed subtle signals for source partitioning. We utilized point prompt-based segmentation to train the SAFIRE model to accurately predict the source region containing each point. During inference, we provided point prompts in a grid format and aggregated the outputs to obtain the final prediction. Through comprehensive evaluation, SAFIRE successfully accommodated label agnosticity issues in IFL and outperformed other state-of-the-art methods. It also opened up possibilities for using point prompting in image partitioning and presented a new challenge of partitioning images into multiple source regions. It aids in comprehending the structure of the forged image and facilitates further analysis. We hope our study contributes to solving the increasingly complex image forgery issues in the era of AI.

References

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33: 12449–12460.

Bammey, Q.; Gioi, R. G. v.; and Morel, J.-M. 2020. An adaptive neural network for unsupervised mosaic consistency analysis in image forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14194–14204.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Cozzolino, D.; and Verdoliva, L. 2019. Noiseprint: A CNN-Based Camera Model Fingerprint. *IEEE Transactions on Information Forensics and Security*, 15: 144–159.

Dong, C.; Chen, X.; Hu, R.; Cao, J.; and Li, X. 2022. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 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*, 422–426. IEEE.

Fu, T.-J.; Hu, W.; Du, X.; Wang, W. Y.; Yang, Y.; and Gan, Z. 2023. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*.

Guan, H.; Kozak, M.; Robertson, E.; Lee, Y.; Yates, A. N.; Delgado, A.; Zhou, D.; Kheyrkhah, T.; Smith, J.; and Fiscus, J. 2019. MFC datasets: Large-scale benchmark datasets for media forensic challenge evaluation. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20606–20615.

Hao, J.; Zhang, Z.; Yang, S.; Xie, D.; and Pu, S. 2021. Transforensics: image forgery localization with dense self-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15055–15064.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, X.; Zhang, Z.; Jiang, Z.; Chaudhuri, S.; Yang, Z.; and Nevatia, R. 2020. 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*, 312–328. Springer.

Huh, M.; Liu, A.; Owens, A.; and Efros, A. A. 2018. Fighting fake news: Image splice detection via learned self-consistency. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 101–117.

Ji, K.; Chen, F.; Guo, X.; Xu, Y.; Wang, J.; and Chen, J. 2023a. Uncertainty-guided Learning for Improving Image Manipulation Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22456–22465.

Ji, W.; Li, J.; Bi, Q.; Li, W.; and Cheng, L. 2023b. Segment anything is not always perfect: An investigation of sam on different real-world applications. *arXiv preprint arXiv:2304.05750*.

Karageorgiou, D.; Kordopatis-Zilos, G.; and Papadopoulos, S. 2024. Fusion Transformer with Object Mask Guidance for Image Forgery Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4345–4355.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Kniaz, V. V.; Knyaz, V.; and Remondino, F. 2019. The point where reality meets fantasy: Mixed adversarial generators for image splice detection. In *Advances in Neural Information Processing Systems*, 215–226.

Korus, P.; and Huang, J. 2016. Multi-scale analysis strategies in PRNU-based tampering localization. *IEEE Transactions on Information Forensics and Security*, 12(4): 809–824.

Kwon, M.-J.; Nam, S.-H.; Yu, I.-J.; Lee, H.-K.; and Kim, C. 2022. Learning JPEG compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, 130(8): 1875–1895.

Kwon, M.-J.; Yu, I.-J.; Nam, S.-H.; and Lee, H.-K. 2021. CAT-Net: Compression Artifact Tracing Network for Detection and Localization of Image Splicing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 375–384.

Li, D.; Zhu, J.; Wang, M.; Liu, J.; Fu, X.; and Zha, Z.-J. 2023. Edge-Aware Regional Message Passing Controller for Image Forgery Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8222–8232.

Li, Y.; Hu, M.; and Yang, X. 2023. Polyp-sam: Transfer sam for polyp segmentation. *arXiv preprint arXiv:2305.00293*.

Lin, L.; Gupta, N.; Zhang, Y.; Ren, H.; Liu, C.-H.; Ding, F.; Wang, X.; Li, X.; Verdoliva, L.; and Hu, S. 2024. Detecting Multimedia Generated by Large AI Models: A Survey. *arXiv preprint arXiv:2402.00045*.

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*, 32(11): 7505–7517.

Moreira, D.; Bharati, A.; Brogan, J.; Pinto, A.; Parowski, M.; Bowyer, K. W.; Flynn, P. J.; Rocha, A.; and Scheirer, W. J. 2018. Image provenance analysis at scale. *IEEE Transactions on Image Processing*, 27(12): 6109–6123.

Ng, T.-T.; Chang, S.-F.; and Sun, Q. 2004. A data set of authentic and spliced image blocks. *Columbia University, ADVENT Technical Report 203-2004-3*.

Niloy, F. F.; Bhaumik, K. K.; and Woo, S. S. 2023. CFL-Net: Image forgery localization using contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 4642–4651.

Novozamsky, A.; Mahdian, B.; and Saic, S. 2020. IMD2020: A Large-Scale Annotated Dataset Tailored for Detecting Manipulated Images. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision Workshops*, 71–80.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pinto, A.; Moreira, D.; Bharati, A.; Brogan, J.; Bowyer, K.; Flynn, P.; Scheirer, W.; and Rocha, A. 2017. Provenance filtering for multimedia phylogeny. In *2017 IEEE international conference on image processing (ICIP)*, 1502–1506. IEEE.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.

Sabir, E.; Nandi, S.; Abd-Almageed, W.; and Natarajan, P. 2021. Biofors: A large biomedical image forensics dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10963–10973.

Su, Y.; Tan, S.; and Huang, J. 2024. A Novel Universal Image Forensics Localization Model Based on Image Noise and Segment Anything Model. In *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*, 149–158.

Sun, Z.; Jiang, H.; Wang, D.; Li, X.; and Cao, J. 2023. Safl-net: Semantic-agnostic feature learning network with auxiliary plugins for image manipulation detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22424–22433.

Tang, L.; Xiao, H.; and Li, B. 2023. Can sam segment anything? when sam meets camouflaged object detection. arXiv 2023. *arXiv preprint arXiv:2304.04709*.

Verdoliva, L. 2020. Media forensics and deepfakes: an overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5): 910–932.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*.

Wang, J.; Wu, Z.; Chen, J.; Han, X.; Shrivastava, A.; Lim, S.-N.; and Jiang, Y.-G. 2022. Objectformer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2364–2373.

Wen, B.; Zhu, Y.; Subramanian, R.; Ng, T.-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)*, 161–165. IEEE.

Wu, H.; Chen, Y.; and Zhou, J. 2023. Rethinking Image Forgery Detection via Contrastive Learning and Unsupervised Clustering. *arXiv preprint arXiv:2308.09307*.

Wu, H.; Zhou, J.; Tian, J.; and Liu, J. 2022. Robust image forgery detection over online social network shared images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13440–13449.

Wu, Y.; AbdAlmageed, W.; and Natarajan, P. 2019. ManTra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9543–9552.

Yu, T.; Feng, R.; Feng, R.; Liu, J.; Jin, X.; Zeng, W.; and Chen, Z. 2023. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*.

Zeng, K.; Cheng, R.; Tan, W.; and Yan, B. 2024. MGQ-Former: Mask-Guided Query-Based Transformer for Image Manipulation Localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 6944–6952.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.

Zhou, J.; Ma, X.; Du, X.; Alhammadi, A. Y.; and Feng, W. 2023a. Pre-training-free Image Manipulation Localization through Non-Mutually Exclusive Contrastive Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22346–22356.

Zhou, P.; Han, X.; Morariu, V. I.; and Davis, L. S. 2018. Learning rich features for image manipulation detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1053–1061.

Zhou, T.; Zhang, Y.; Zhou, Y.; Wu, Y.; and Gong, C. 2023b. Can sam segment polyps? *arXiv preprint arXiv:2304.07583*.

Zhu, J.; Li, D.; Fu, X.; Yang, G.; Huang, J.; Liu, A.; and Zha, Z.-J. 2024. Learning Discriminative Noise Guidance for Image Forgery Detection and Localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7739–7747.