SAMIF: Adapting Segment Anything Model for Image Inpainting Forensics*

Lan Zhang[1], Xinshan Zhu[1], Di He[1], Xin Liao[2], and Biao Sun**[1]

[1] School of Electrical and Information Engineering, Tianjin University, Tianjin, China
{zl2022,xszhu,di_he,sunbiao}@tju.edu.cn
[2] College of Computer Science and Electronic Engineering, Hunan University, Hunan, China
xinliao@hnu.edu.cn

Abstract. Image inpainting technologies pose increasing threats to the security of image data through malicious use. Therefore, image inpainting forensics is crucial. The Segment Anything Model (SAM) is a powerful universal image segmentation model for various downstream tasks. However, the performance of SAM in inpainting forensics is significantly degraded due to the substantial disparity between natural and inpainted image domains. In this paper, we propose SAMIF, a SAM-based model for image inpainting forensics. First, based on SAM, a parallel convolutional neural network (CNN) branch is introduced to assist the SAM in extracting local noise information. Second, the cross-domain alignment fusion module (CAFM) is designed to better fuse the features of the two branches. Third, the artifact features generator (AFG) is designed between the encoder and decoder to disentangle the features extracted by the encoder. The auxiliary loss is introduced in AFG, which shortens the backpropagation path and guides the SAM branch to learn artifact features, thus enhancing the adaptability of SAM for the inpainting forensics task. Extensive experiments demonstrate that the proposed model achieves state-of-the-art results on five inpainting forensics datasets and exhibits excellent robustness and generalization capabilities.

Keywords: Image inpainting forensics · Segment anything model · Cross-domain alignment fusion · Artifact feature generator

1 Introduction

With the rapid development of image editing technology, digital image forgery is becoming increasingly prevalent [6, 25]. In recent years, concerns about multimedia information security have grown significantly. Accurately detecting and

* This work was supported in part by the National Natural Science Foundation of China under Grants 61972282, 61971303, U22A2030, T2322020, and 62371329 and in part by the Hunan Provincial Funds for Distinguished Young Scholars under Grant 2024JJ2025.
** Corresponding author.

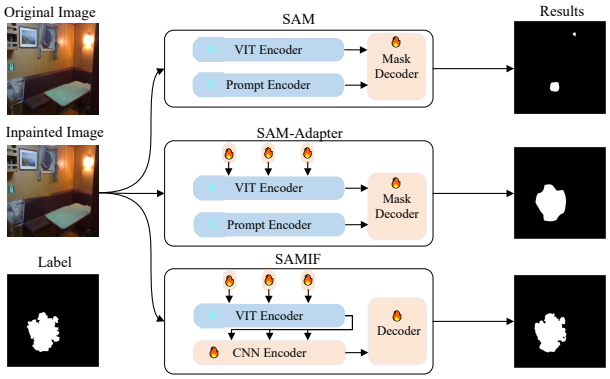

Fig. 1: Detection results of Different SAM Models for inpainting forensics.

locating tampered areas in images is crucial, prompting an increasing number of scholars to study image tampering forensics [17–19, 27, 28, 35, 39]. Image inpainting is a common and effective image editing technique that aims to remove parts of the content or fill missing areas in a visually plausible way using the existing information in the image [1]. Image inpainting has a powerful ability to edit images. Compared to other image manipulation techniques, such as splicing, inpainting is more complex, resulting in more hidden traces. This makes inpainting forensics more challenging.

In recent years, to improve detection efficiency and performance, deep learning-based image inpainting forensics algorithms have been continuously developing. Some algorithms introduce convolutional neural networks (CNNs) for image inpainting forensics to automatically extract inpainting features and locate the inpainted regions [14, 30]. However, CNNs primarily extract local features, particularly in the lower layers of the network. Although the higher layers can capture global features, resolution degradation can cause the loss of critical information, which is detrimental to inpainting forensics tasks. Consequently, this results in poor robustness and generalization of CNN-based forensic models.

To address these issues, we turn our attention to the vision foundation model SAM [11]. SAM, a universal segmentation network trained on a large visual corpus, has demonstrated strong segmentation abilities in various scenarios. The addition of adapters further improves its performance in downstream tasks [3]. The introduction of SAM is expected to improve the generalisation and robustness of image inpainting forensics. However, when SAM is directly applied to image inpainting forensics, as shown in Fig. 1, the results are often unsatisfactory. This is because the region of image inpainting is not necessarily a complete object. Image inpainting forensics requires the network to extract anomalous manipulation features instead of semantic features of the object. We refer to these anomalous manipulated features as artifact features. Existing SAM-based methods do not effectively guide SAM to extract these artifact features, resulting in its inability to adapt to inpainting forensics.

In this paper, we are motivated by the above observation. We aim to transition the powerful segmentation and generalization capabilities of SAM to image inpainting forensics to build a forensic model with superior and robust forensic performance. We believe that guiding SAM to extract such artifact features is crucial. Under this motivation, we propose SAMIF. As shown in Fig. 1, compared to other SAM-based models, SAMIF can accurately predict the inpainted regions in tampered images. We introduce local noise information and high-frequency information into the encoder to expose tampering artifacts. Specifically, we retain the encoder part of SAM and add a CNN branch in parallel to assist SAM in extracting local information. High-frequency and local noise information is introduced into the encoder by adding filters before the two branches. A cross-domain alignment fusion module is then designed to integrate features extracted by SAM into various stages of the CNN encoder. To further guide SAM in extracting artifact features, we design an artifact feature generator (AFG) between the encoder and decoder. The AFG disentangles the features extracted by the encoder and predicts tamper masks at multiple scales. We compute multi-scale auxiliary loss in AFG. This auxiliary loss shortens the backpropagation path and guides the encoder to learn the tampering artifacts, enhancing SAM's adaptability for the inpainting forensics task. Finally, we constructed five inpainting forensic datasets using different image inpainting methods. Extensive experiments demonstrate that the proposed model achieves higher location accuracy for various inpainting manipulations and is more robust against JPEG compression and additive noise attacks compared to state-of-the-art methods for inpainting forensics. Moreover, SAMIF exhibits desirable generalization for typical unseen inpainting methods.

In summary, our main contributions are as follows:

- We propose SAMIF, a foundation model specifically designed for image inpainting forensics. SAMIF transitions the powerful segmentation and generalization capabilities of SAM to the field of image inpainting forensics.
- In the dual-branch encoder comprising SAM and CNN, we designed the cross-domain alignment fusion module (CAFM) to integrate the rich forensic information extracted by the ViT encoder of SAM into the CNN branch.
- We designed an artifact feature generator (AFG) to disentangle the features extracted by the encoder and guide the encoder to extract artifact features specific to the inpainting forensics.
- Extensive experiments on various datasets have demonstrated that SAMIF compares favorably with state-of-the-art methods and exhibits excellent robustness and generalization capabilities.

2 Related Work

2.1 Inpainting Forensics

The purpose of image inpainting forensics is to detect regions of an image tampered with using inpainting techniques. As multimedia information security gains

attention, research on image inpainting forensics advances. Zhu et al. [38] proposed a CNN-based encoder-decoder network that extracts tampering traces via neural networks, achieves pixel-level image prediction, and identifies inpainted regions. Li et al. [12] designed an HP-FCN network that preprocesses the input image using a high-pass filtering module. The residual map is then transferred to a CNN-based feature extraction module, upsampled to the input image size, and results in a pixel-level localization map. Wu et al. [31] proposed IID-Net, a novel end-to-end image inpainting detection network, aiming to enhance the detection performance of existing algorithms for invisible image inpainting methods. The network utilizes the NAS algorithm to search for suitable network architectures and integrates a newly proposed attention module to optimize inpainting features. Zhu et al. [37] introduced the Transformer, known for its capability in long-distance modelling, into inpainting forensics. They integrated it with CNN to create a dual-stream inpainting forensics network aimed at capturing global tampering features from the inpainting process. While VIT enhances forensic accuracy, it concurrently increases computational load in the network.

2.2 SAM

In recent years, the foundation model has undergone significant advancements in the field of computer vision [21, 29, 40]. Among them, SAM has attracted considerable attention due to its excellent segmentation performance for natural images. The model can achieve 'few-shot' and even 'zero-shot' learning across diverse image categories through fine-tuning. However, SAM often performs poorly in downstream image segmentation tasks. To address this issue, some scholars have introduced adapters into SAM [3]. Instead of fine-tuning SAM, they incorporate domain-specific information or visual prompts into the segmentation network using simple yet effective adapters, improving the performance of SAM in various downstream tasks. Some scholars have designed network structures incorporating domain-specific knowledge and adapted SAM to domain-specific image segmentation, such as medical image segmentation [9, 15, 20, 33] and remote sensing image segmentation [2, 8, 16, 22]. Our work aims to adapt SAM to image inpainting forensic tasks.

3 Method

3.1 Architecture Overview

The architecture of the proposed SAMIF framework is illustrated in Fig. 2. The network is designed based on the encoder-decoder architecture. Specifically, SAMIF consists of two parallel branches: the CNN branch and the SAM branch. The CNN branch comprises four stages, each consisting of two ResNet blocks. The embeddings output from the SAM branch are fused with the embeddings of each stage of the CNN branch through the cross-domain alignment fusion module (CAFM) to integrate forensic features from different domains. The output

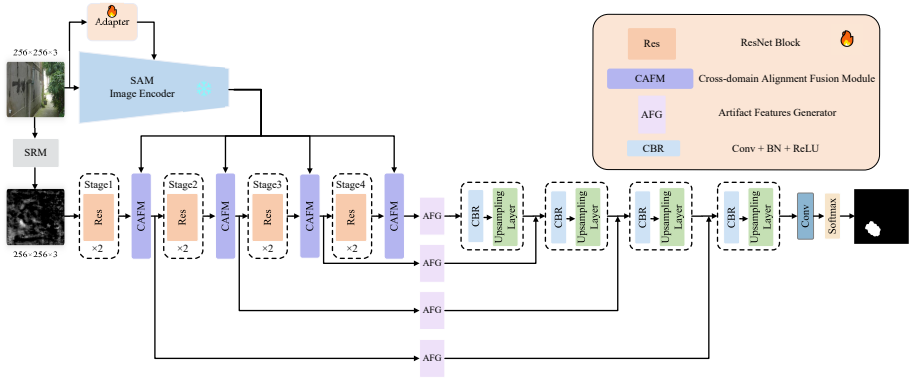

Fig. 2: The architecture of our proposed SAMIF. Modules with flame icons indicate that parameters are trainable, and modules with snowflake icons indicate that parameters are frozen during training.

embeddings of each stage of the encoder are disentangled by the artifacts feature generator (AFG), retaining the artifact features relevant to inpainting forensics. These embeddings are then delivered to the corresponding stage of the decoder. The decoder comprises a repeated stack of convolutional blocks and upsampling layers. Finally, the feature maps are fed into the Sigmoid layer to generate the predictions.

3.2 Encoder

The role of the encoder is to receive the inpainted image and extract artifact features. Inpainting methods usually leave faint traces in the image, and traditional CNNs and SAM primarily learn features related to the image content. As a result, it is difficult to extract effective features directly from the RGB image for forensics. Therefore, we designed the dual-stream encoder with filters to extract artifact features from different domains.

The SAM branch The overall architecture of the SAM branch in the encoder is inherited from the SAM-Adapter [3] with some minor adjustments. The adapter in this architecture can integrate task-specific knowledge with general knowledge learnt by the large model to improve the performance of SAM in a specific downstream task. The adapter consists of only two MLPs and an activate function within two MLPs. Specially, the adapter receives the task-specific knowledge F to obtain the prompt P. As tampering artifacts are more visible in the high frequency component, we extract the task-specific knowledge F from the frequency domain. Taking the image X as input, we first transforms it from the RGB domain to the frequency domain using Discrete Cosine Transform (DCT):

$$X_q = \mathrm{DCT}\left(X\right) \tag{1}$$

$$\frac{1}{4}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 2 & -4 & 2 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \quad \frac{1}{12}\begin{bmatrix} -1 & 2 & -2 & 2 & -1 \\ 2 & -6 & 8 & -6 & 2 \\ -2 & 8 & -12 & 8 & -2 \\ 2 & -6 & 8 & -6 & 2 \\ -1 & 2 & -2 & 2 & -1 \end{bmatrix} \quad \frac{1}{2}\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Fig. 3: SRM filter kernels

where X_q is the frequency domain representation and DCT$(\cdot)$ denotes DCT. We then filter the frequency domain representation using a high pass filter to capture the high-frequency components and transform it back to the RGB domain using the Inverse Discrete Cosine Transform (IDCT):

$$F = \text{IDCT}\left(H\left(X_q, \alpha\right)\right) \tag{2}$$

where $H(\cdot)$ denotes the high pass filter and a is a manually set parameter used to control the filtering threshold of the high pass filter. IDCT$(\cdot)$ denotes IDCT. Then, the task-specific knowledge F is used to generate the prompt P, which is attached to each transformer layer of the SAM model:

$$P = \text{MLP}_{\text{up}}\left(\text{GELU}\left(\text{MLP}_{\text{tune}}\left(F\right)\right)\right) \tag{3}$$

in which $\text{MLP}_{\text{tune}}(\cdot)$ are linear layers used to generate task-specific prompts for each Adapter. $\text{MLP}_{\text{up}}(\cdot)$ is an up-projection layer shared across all Adapters that adjusts the dimensions of transformer features. P refers to the output prompt that is attached to each transformer layer of SAM model. GELU$(\cdot)$ is the GELU activation function. A more detailed description of the SAM-adpter structure can be found in the literature [3]. SAM is bootstrapped to extract frequency features through the prompt. The minor adjustments made to the SAM-Adapter include retaining only the ViT encoder part and reducing the input spatial resolution from 1024×1024 pixels to 256×256 pixels, thereby reducing the GPU memory cost. We made a corresponding scaling of the positional embedding of VIT in the pretrained weights to ensure that the pretrained weights can be loaded correctly.

The CNN branch During image inpainting, new pixels are generated to fill the missing regions, and previous research [5] demonstrates that these new elements differ in noise distribution from real regions. To address this, we first filter the input RGB image using a spatial rich model (SRM) [7] to convert it from the RGB domain to the noise domain. Specifically, we employ three SRM filter kernels identified from the literature [36] as convolution kernels, as depicted in Fig. 3. An RGB image of size $256 \times 256 \times 3$ is filtered to produce a noise representation of the same dimensions. Following the SRM module, a feature extraction backbone based on a residual CNN is implemented, as shown in Fig. 2. This backbone consists of four stages, each containing two bottleneck units. This minimalist and lightweight design of the CNN branch aims to prevent overfitting during training. The output embeddings of the CNN branch specifically match the resolution of the embeddings from the SAM branch. Specifically,

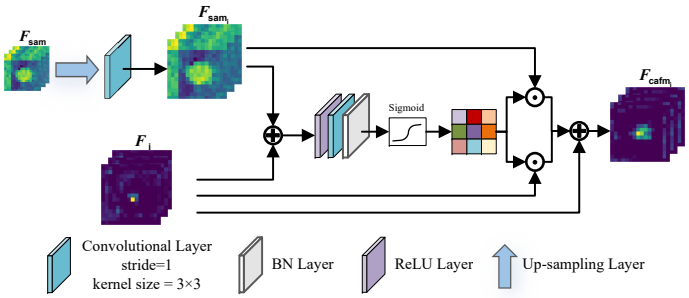

Fig. 4: The architecture of cross-domain alignment fusion module (CAFM).

the image undergoes 16-fold downsampling through the CNN branch after four stages.

The cross-domain alignment fusion module The purpose of the CAFM is to enrich artifact features by fusing information from different domains. The embeddings from the SAM branch are delivered to each stage of the CNN branch through the CAFM. The structure of CAFM is illustrated in Fig. 4. Let F_{sam} denote the feature embeddings output from the SAM branch, and F_i denote the embeddings output from the ith stage of the CNN branch. At first, except for the CAFM following the fourth stage of CNN, F_{sam} will be adjusted to the same size as F_i after the upsampling layer and a 3×3 convolution layer, which is denoted as F_{sam_i}:

$$F_{sam_i} = \begin{cases} \text{Conv}(\text{Upsam}(F_{sam})) & i \in \{1, 2, 3\} \\ F_{sam} & i = 4 \end{cases} \tag{4}$$

Where Upsam$(\cdot)$ and Conv$(\cdot)$ denote the upsampling layer and a 3×3 convolution layer, respectively. Next, we combine the feature maps of the two branches through element-wise addition. The rectified linear unit (ReLU) layer, 3×3 convolution, and batch normalization (BN) layer, are then used to further process the feature embeddings. After that, the sigmoid function is applied to adaptively generate the attention weight maps A_i. Finally, the features F_{sam_i} and F_i are multiplied by the attention weight maps respectively, and F_i is added element-wise to obtain the final output F_{casm_i}. The calculation process of the ith CAFM can be represented as:

$$A_i = \text{Sigmod}(\text{Conv}(\text{BN}(\text{Relu}(F_{sam_i} + F_i)))) \tag{5}$$

$$F_{cafm_i} = F_i + F_i \odot A_i + F_{sam_i} \odot A_i \tag{6}$$

where Sigmoid$(\cdot)$ and Relu$(\cdot)$ denote the Sigmoid and ReLU functions, respectively, BN denotes the BN layer and $\odot$ denotes elementwise multiplication.

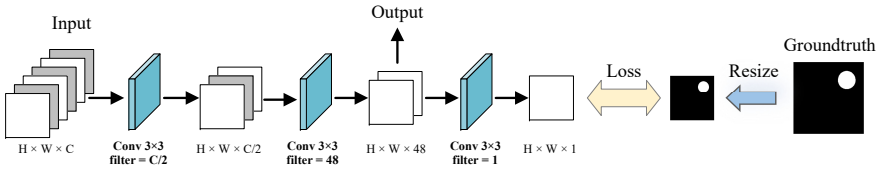

Fig. 5: The architecture of artifact feature generator (AFG). The AFG consists of three convolutional layers. The output feature maps of the second convolutional layer are delivered to the decoder, and the output feature map of the third convolutional layer is used to compute the auxiliary loss with the downsampled ground truth.

3.3 Decoder

As shown in Fig. 2, we employ a decoder constructed by repeatedly stacking convolutional blocks and upsampling layers, instead of selecting the mask decoder in SAM. One reason for selecting this architecture is that it strikes an optimal balance between model complexity and detection accuracy. The skip connections between corresponding resolutions of the encoder and decoder reduces the loss of high-resolution details in manipulation artifacts, which is crucial for forensic tasks. Specifically, the decoder structure involves a repeated stack of convolutional blocks and upsampling layers four times, followed by a convolutional layer to reduce the feature map channels to 1. Finally, the feature maps are fed into the Sigmoid layer to obtain the final predictions.

3.4 The artifact features generator

The AFGs are positioned between the corresponding stages of the encoder and decoder. As shown in Fig. 5, each artifact feature generator consists of three 3×3 convolutional layers and a Sigmoid function. After receiving the encoder's output feature maps, the first convolutional layer halves the number of channels, the second adjusts the channels to a fixed 48, and the third reduces the channels to 1, which is then passed through the Sigmoid function to obtain the prediction mask. This lightweight design introduces auxiliary loss while avoiding redundant computation. Four artifact feature generators are placed at different resolutions, producing four prediction masks. The ground truth is downsampled to match the prediction masks of different resolutions to compute the cross-entropy loss, which serves as an auxiliary loss for network training. Notably, the final output feature maps of the AFG are derived from the second convolution layer, not the last convolutional layer, ensuring that the output feature maps conform to the actual artifact distribution rather than overfitting to the labels. The primary purpose of the AFG is to disentangle the features extracted by the encoder and retain the artifact features. Furthermore, the auxiliary loss in the AFG shortens the backpropagation path, accelerating network convergence and guiding the network to learn artifact features, thereby enhancing the SAM encoder's adaptation to the inpainting forensics. The calculation process of the ith AFG can be represented as:

$$F_{afg_i} = \text{Conv}(\text{Conv}(F_{casm_i})) \tag{7}$$

$$\hat{M}_i = \text{Sigmoid}(\text{Conv}(F_{afg_i})) \tag{8}$$

where F_{afg_i} denotes the final output feature maps of the AFG, and $\hat{M}_i$ denotes the prediction masks, which participate in the calculation of auxiliary loss.

3.5 Loss function

The SAMIF training process integrates three loss functions: the auxiliary loss from the AFGs, the cross-entropy loss between the final output and ground truth, and the IOU loss computed between the final output $\hat{M}_l$ and ground truth G. These terms are weighted to form the overall optimization objective:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{bce}(\hat{M}_l, G) + \lambda_2 \mathcal{L}_{iou}(\hat{M}_l, G) + \lambda_3 \sum_{i=1}^{4} \mathcal{L}_{bce}(\hat{M}_i, G_i) \tag{9}$$

where G_i denotes the ground truth that is downsampled to the same size as $\hat{M}_i$. Parameters λ_1, λ_2, and λ_3 balance the importance of the influence of these loss terms. Finally, the model parameters are trained by optimizing the total loss.

4 Discussion

Here, we will discuss how SAMIF guides SAM in extracting artifact features, enabling adaptation to inpainting forensics.

Introduction of task-specific information: In image inpainting forensics, inpainting leaves more artifact information in the high-frequency components of the frequency domain. Therefore, based on the structure of the SAM-adapter, we transform the input image using DCT, apply a high-pass filter to remove low-frequency components, and feed the high-frequency components into each transformer layer of the SAM model.

Integration of multi-domain information: To assist the SAM encoder, we add a parallel CNN branch combined with an SRM filter to extract local noise features related to tampering. The noise features from the CNN branch complement the features from the SAM branch, and the integration of multi-domain information provides comprehensive artifact features for the forensic task.

Introduction of auxiliary loss: Since SAM has already acquired strong segmentation and generalization performance on a large corpus, it is crucial to fine-tune SAM. The backpropagation path can be effectively shortened by adding auxiliary loss to the AFG. Multi-scale supervision motivates SAM to learn artifact features, facilitating its adaptation to the downstream task of image inpainting forensics.

5 Experiments

5.1 Experimental Settings

Image inpainting forensics datasets We randomly selected 20,250 images of size 256×256 from the Place 365 dataset [34]. We considered five representative inpainting methods, three of which are DL-based and were proposed in recent years: DeepfillV2 [32], ICT [26], and Lama [23]. The remaining two methods are traditional (non-DL-based): Diffusion [24] and Exemplar [4]. We first removed a part of the image area using a mask with a random shape, size, and position. The random shapes included circles, rectangles, and irregular shapes. Since diffusion-based inpainting method is more suitable for small missing areas, we set the mask size to 0.1%, 0.4%, 1.56%, or 6.25% for this method. The mask size for the other inpainting methods was set to 1%, 5%, or 10%. The mask location was randomly selected within the image. We then inpainted the images using these five inpainting methods to create five distinct datasets. The following article will refer to the datasets by the names of the five inpainting methods. Finally, we divided all the inpainted datasets into three parts: 18,000 images for training, 1,350 images for validation, and 900 images for testing.

Implementation Details and Evaluation Metrics We built our SAMIF on the PyTorch framework and trained it on NVIDIA GeForce RTX 3090 GPUs. The input size for training was set to 256×256. We used the Adam optimizer [10] with a batch size of 24. The initial learning rate was set to 1e-3, and cosine decay was applied to the learning rate. We set the weight parameters in the loss function as follows: $\lambda_1 = 1$, $\lambda_2 = 1$, and $\lambda_3 = 0.5$. Additionally, to prevent network overfitting, we performed data enhancement. The data enhancement included random horizontal or vertical flipping, random rotation by 90°, and JPEG compression enhancement with a quality factor randomly chosen between 70 and 100.

To compare with other state-of-the-art methods, we selected deep learning-based inpainting forensics methods, including FCNet [38], HPFCN [12], IID-Net [31], GLFFNet [37], and other general forensics methods, including PSCC-Net [13] and MVSS-Net [5]. For a fair comparison, these methods were based on their open-source codes. All methods strictly followed the training procedures and parameter settings presented in their papers, and were retrained on our dataset. We selected Intersection over Union (IoU), F1 score, and AUC as evaluation metrics.

5.2 Evaluation in the absence of attacks

Quantitative Assessment Comparison In order to quantitatively evaluate the comparison between other methods and our proposed method, we trained and tested them on the five datasets. Table 1 summarises the performance of each forensics method. Moreover, our SAMIF have the best performance among the tested methods. Notably, SAMIF achieves an IoU of 94.54% on the Diffusion

Table 1: Quantitative comparisons with other state-of-the-arts on different inpainting methods in the absence of attacks. The best results are shown in bold.

Method	Diffusion			Exemplar			ICT			Lama			DeepfillV2		
	IoU	F1	AUC	IoU	F1	AUC	IoU	F1	AUC	IoU	F1	AUC	IoU	F1	AUC
FCNet [38]	61.93	68.38	96.74	79.96	87.56	99.54	69.45	76.97	99.07	76.28	84.26	99.63	49.69	57.18	94.57
HPFCN [12]	61.25	66.88	96.60	80.61	88.02	99.65	72.99	81.34	99.62	80.89	88.18	99.82	68.75	76.87	99.60
IID-Net [31]	59.97	64.29	95.61	82.63	89.63	99.67	76.66	84.30	99.63	81.40	88.75	99.83	54.42	61.92	96.38
PSCC-Net [13]	63.32	68.75	97.99	85.07	91.52	99.87	86.14	92.13	99.90	87.50	93.11	99.93	82.28	89.61	99.83
MVSS-Net [5]	71.88	77.64	96.72	88.79	93.82	99.93	86.52	92.33	99.90	87.04	92.79	99.93	82.60	89.61	99.78
GLFFNet [37]	82.03	85.59	98.97	91.10	95.22	99.97	91.02	95.06	99.97	92.50	96.00	99.98	88.85	93.92	99.96
SAMIF	**94.54**	**97.21**	**99.98**	**93.48**	**96.66**	**99.97**	**94.44**	**97.15**	**99.98**	**94.61**	**97.24**	**99.98**	**91.52**	**95.59**	**99.97**

dataset, which is about 12.51% higher than the second-best GLFFNet. This indicates that SAMIF can effectively detect small inpainted regions. The superior performance on different datasets is attributed to the superior segmentation capabilities of SAM.

Qualitative Assessment Comparison To compare our method more comprehensively with others, we conducted a qualitative assessment of different forensic models, as shown in Fig. 6. Specifically, we plotted the tampered images and ground truth for DeepfillV2 datasets, as well as the detection results of various forensic methods. In the first row, we present the detection results of different models for regular inpainted regions. Compared to other models, our model accurately detects the shape of the inpainted regions. The last two rows display the detection results of different models for irregular inpainted regions. For irregular regions, most forensic methods are ineffective, and the detection results tend to be regular shapes because they cannot effectively capture the details (e.g., boundary information) of the inpainted regions. When the inpainted region is small, some forensic methods even miss detections. Conversely, our proposed SAMIF not only fits the shape of irregular regions better but also maintains superior detection results for smaller areas. This can be attributed to the well-designed network structure.

5.3 Evaluation under typical attacks

In practice, malicious individuals may perform post-processing operations after inpainting to hide tampering traces, reduce file size, or evade forensic detection. Thus, we investigate the robustness of the proposed method against image compression and additive white Gaussian noise (AWGN) using DeepfillV2 datasets. The quantitative evaluation results are shown in Tab. 2. Specifically, we applied JPEG compression and noise addition to the DeepfillV2 datasets. The JPEG compression quality factors (QF) were set to 90 and 70, and the signal-to-noise ratios (SNRs) of AWGN were set to 50 dB and 40 dB, respectively. It can be observed that as the compression factors and signal-to-noise ratios decrease, the forensic performance of all models degrades, while our method consistently exhibits the best performance. On the challenging DeepfillV2 dataset, our network

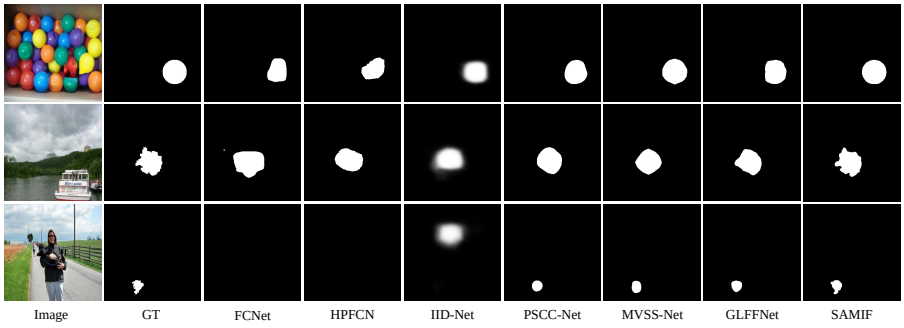

Fig. 6: Qualitative comparisons for detection of inpainting forgeries on the DeepfillV2 dataset. Our proposed model exhibits the best detection performance, with detection results for the inpainted regions being closest to the ground truth.

Table 2: Quantitative comparisons of robustness with other state-of-the-arts on Deepfillv2 Dataset. The best results are shown in bold.

Attack	FCNet [38]			IID-Net [31]			HPFCN [12]			PSCC-Net [13]			MVSS-Net [5]			GLFFNet [37]			SAMIF		
	IoU	F1	AUC	IoU	F1	AUC	IoU	F1	AUC	IoU	F1	AUC	IoU	F1	AUC	IoU	F1	AUC	IoU	F1	AUC
JPEG 90	45.25	53.02	92.87	44.80	52.26	94.45	47.60	55.62	98.62	76.21	84.59	99.54	76.36	84.40	99.38	86.01	92.09	99.93	**88.86**	**94.12**	**99.94**
JPEG 70	42.17	49.69	91.75	37.38	44.61	91.88	29.39	36.32	94.76	51.02	59.28	96.03	50.45	59.00	94.60	58.96	66.92	97.74	**74.46**	**85.39**	**99.05**
AWGN 50	49.10	56.65	94.41	52.90	60.32	96.02	65.38	73.50	99.10	75.85	84.30	99.60	81.18	88.46	99.71	88.41	93.61	99.95	**90.91**	**95.26**	**99.93**
AWGN 40	46.82	54.53	93.47	45.56	52.83	93.97	32.85	41.88	90.65	39.06	49.99	96.08	68.13	76.24	98.30	81.94	88.36	99.66	**88.95**	**94.22**	**99.90**

achieves an IoU of 74.46% at QF 70. The performance of GLFFNet, which is also composed of VIT and CNN, is not satisfactory, with the IoU of only 58.96% at QF 70. The experimental results demonstrate that the introduction of the SAM model effectively enhances the network's robustness against post-processing.

5.4 Evaluation of unseen inpainting methods

In practice, the inpainting methods for tampered images are often unknown, which requires that the inpainting forensics model be applicable to unknown inpainting methods. Therefore, we evaluate the forensics performance of the model when facing unseen inpainting methods. The evaluation results are shown in Fig. 7. Specifically, we train the model using one inpainting dataset and test the model with other image inpainting datasets, i.e., the inpainting methods in the test set are unknown during model training. The Diffusion dataset was excluded due to the small size of the inpainted regions, creating a domain gap with other datasets. As can be seen from the experimental results, our model successfully recognizes unseen inpainting methods. Among these datasets, the model trained on the DeepfillV2 dataset demonstrates the highest performance in cross-dataset evaluation. The excellent detection performance of SAMIF for unseen inpainting methods illustrates that SAMIF transitions the powerful generalization capabilities of SAM to the field of image inpainting forensics.

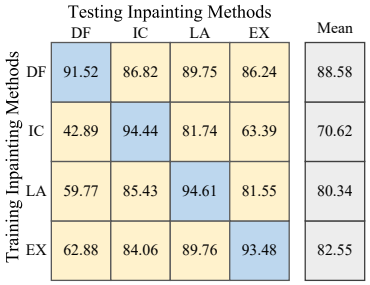

Fig. 7: Detection performance of SAMIF on unseen inpainting methods. Each row shows the results (IOU values) of the model trained on the dataset generated by a particular inpainting method and tested on several datasets generated by seen and unseen inpainting methods. The rightmost column provides the average IoU for each row.

Table 3: Ablation study results of SAMIF on the DeepfillV2 Dataset.

Components					No attacks		
F-Adapter	SAM	CNN	CAFM	AFG	IoU	F1	AUC
	✓				24.20	40.07	85.28
✓	✓				66.34	79.85	98.51
✓	✓	✓			80.27	89.11	99.51
✓	✓	✓	✓		85.26	92.07	99.78
✓	✓	✓		✓	88.58	93.67	99.86
✓	✓	✓	✓	✓	91.52	95.59	99.97

5.5 Ablation Study

The five components of SAMIF, SAM encoder, frequency-adapter (F-adapter), CNN branch, CAFM, and AFG, are sequentially integrated into the network and assessed using the DeepFillV2 dataset. The quantitative experimental results are presented in Tab. 3. Each component of SAMIF effectively improves the forensic performance of the model. Comparing the first two rows, the introduction of frequency-domain information into SAM through DCT significantly improved the forensic performance, resulting in a 42.14% improvement in IoU, indicating that task-specific information (high-frequency components) can effectively guide SAM's adaptation to inpainting forensics. Furthermore, the incorporation of the CNN branch introduces noise-domain information to the encoder, which further enhances performance, demonstrating that the integration of multi-domain information provides comprehensive forensic artifact features. Comparing the third and fourth rows, it is evident that the CAFM fuses multi-domain features more effectively than simple channel concatenation, resulting in a 4.99% improvement in IoU. The introduction of AFG further guides SAM to adapt to the inpainting forensics task. SAMIF achieves the best forensic performance by integrating these four components.

In addition to quantitative analysis, we visualize the output feature maps of the encoder when the network comprises different components. As shown in Fig. 8, we present the inpainted image, label, and the encoder's output feature maps. The third column represents the case where SAM is used directly as an encoder. It can be seen that the encoder focuses more on the image content

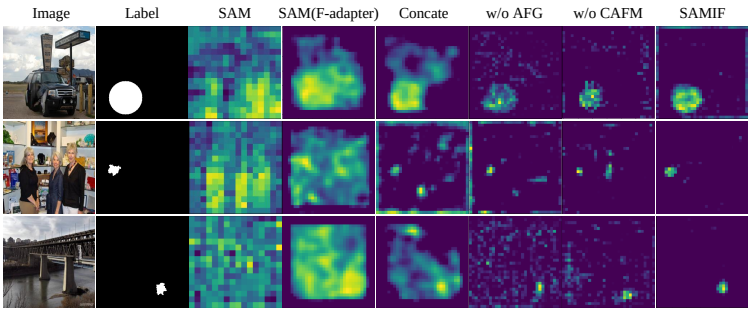

Fig. 8: Visualization of averaged output feature maps of the encoder, brighter color indicating higher responses.

rather than the inpainted region. The fourth column represents the case where the frequency-adapter is introduced into SAM through the DCT. It can be seen that the introduction of frequency domain information directs SAM to focus on inpainted regions. However, it also shows some concern for uninpainted regions, suggesting that the features extracted by SAM contain redundant information. The fifth column represents cases with a dual-stream encoder using channel concatenation for feature fusion. Compared to the results in the fourth column, the introduction of multi-domain information makes the encoder focus more on the tampered region, though some redundant information remains. The sixth and seventh columns represent cases where SAMIF has AFG and CAFM removed, respectively. Compared to the fifth column, either improving the feature fusion method or introducing auxiliary loss guides the encoder in extracting artifact features. When we combine all the proposed components, the encoder exhibits a strong response to the inpainted region while filtering out irrelevant information.

6 Conclusions

In this paper, we propose SAMIF for image inpainting forensics. This model transitions the powerful segmentation and generalization capabilities of SAM to the field of image inpainting forensics. Our approach adds a parallel CNN branch to SAM. The two branches with filters extract frequency-domain information and noise-domain information, respectively. We propose a cross-domain alignment fusion module (CAFM) to better fuse features from different branches. To make SAM more adaptable to inpainting forensics, we propose the artifact feature generator (AFG). The features extracted by the encoder are disentangled for the decoder. Additionally, the auxiliary loss introduced in the AFG shortens the backpropagation path and guides the SAM branch to learn artifact features. Extensive experiments show that the proposed model outperforms current forensic methods in terms of forensic performance and robustness while maintaining excellent performance for unseen inpainting datasets.

References

1. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proc. 27th Annu. Conf. Comput. Graph. Interact. Techn. pp. 417–424. New York, NY , USA (2000)
2. Chen, K., Liu, C., Chen, H., Zhang, H., Li, W., Zou, Z., Shi, Z.: Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. IEEE Trans. Geosci. Remote Sens. (2024)
3. Chen, T., Zhu, L., Deng, C., Cao, R., Wang, Y., Zhang, S., Li, Z., Sun, L., Zang, Y., Mao, P.: Sam-adapter: Adapting segment anything in underperformed scenes. In: Proc. IEEE Int. Conf. Comput. Vis. pp. 3367–3375 (2023)
4. Criminisi, A., Perez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Trans. Image Process. **13**(9), 1200–1212 (2004)
5. Dong, C., Chen, X., Hu, R., Cao, J., Li, X.: Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. IEEE Trans. Pattern Anal. Mach. Intell. **45**(3), 3539–3553 (Mar 2022)
6. Elharrouss, O., Almaadeed, N., Al-Maadeed, S., Akbari, Y.: Image inpainting: A review. Neural Process. Lett. **51**, 2007–2028 (2020)
7. Fridrich, J., Kodovsky, J.: Rich models for steganalysis of digital images. IEEE Trans. Inf. Forensics Security **7**(3), 868–882 (2012)
8. Hetang, C., Xue, H., Le, C., Yue, T., Wang, W., He, Y.: Segment anything model for road network graph extraction. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 2556–2566 (2024)
9. Kim, S., Kim, K., Hu, J., Chen, C., Lyu, Z., Hui, R., Kim, S., Liu, Z., Zhong, A., Li, X., et al.: Medivista-sam: Zero-shot medical video analysis with spatio-temporal sam adaptation. arXiv preprint arXiv:2309.13539 (2023)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. 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: Proc. IEEE Int. Conf. Comput. Vis. pp. 4015–4026 (2023)
12. Li, H., Huang, J.: Localization of deep inpainting using high-pass fully convolutional network. In: Proc. IEEE Int. Conf. Comput. Vis. pp. 8301–8310. Seoul, Korea (October 2019)
13. Liu, X., Liu, Y., Chen, J., Liu, X.: PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. IEEE Trans. Circuits Syst. Video Technol. **32**(11), 7505–7517 (Nov 2022)
14. Lu, M., Liu, S.: A detection approach using lstm-cnn for object removal caused by exemplar-based image inpainting. Electronics **9**(5), 1–22 (May 2020)
15. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024)
16. Ma, X., Wu, Q., Zhao, X., Zhang, X., Pun, M.O., Huang, B.: Sam-assisted remote sensing imagery semantic segmentation with object and boundary constraints. arXiv preprint arXiv:2312.02464 (2023)
17. Ma, X., Zhu, X., Su, L., Du, B., Jiang, Z., Tong, B., Lei, Z., Yang, X., Pun, C.M., Lv, J., et al.: Imdl-benco: A comprehensive benchmark and codebase for image manipulation detection & localization. arXiv preprint arXiv:2406.10580 (2024)
18. Miao, C., Chu, Q., Gong, T., Tan, Z., Jin, Z., Zhuang, W., Luo, M., Hu, H., Yu, N.: Mixture-of-noises enhanced forgery-aware predictor for multi-face manipulation detection and localization. arXiv preprint arXiv:2408.02306 (2024)

19. Miao, C., Chu, Q., Tan, Z., Jin, Z., Zhuang, W., Wu, Y., Liu, B., Hu, H., Yu, N.: Multi-spectral class center network for face manipulation detection and localization. arXiv preprint arXiv:2305.10794 (2023)
20. Na, S., Guo, Y., Jiang, F., Ma, H., Huang, J.: Segment any cell: A sam-based auto-prompting fine-tuning framework for nuclei segmentation. arXiv preprint arXiv:2401.13220 (2024)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proc. Int. Conf. Mach. Learn. pp. 8748–8763. PMLR (2021)
22. Sultan, R.I., Li, C., Zhu, H., Khanduri, P., Brocanelli, M., Zhu, D.: Geosam: Fine-tuning sam with sparse and dense visual prompting for automated segmentation of mobility infrastructure. arXiv preprint arXiv:2311.11319 (2023)
23. Suvorov, R., Logacheva, E., Mashikhin, A., Remizova, A., Ashukha, A., Silvestrov, A., Kong, N., Goka, H., Park, K., Lempitsky, V.: Resolution-robust large mask inpainting with fourier convolutions. In: Proc. IEEE Winter Conf. Appl. Comput. Vis. pp. 2149–2159. Waikoloa, HI, USA (Jan 2022)
24. Tschumperlé, D., Fourey, S.: G'mic(greyc's magic for image computing): A full-featured open-source framework for image processing. `http://www.gmic.eu/`
25. Verdoliva, L.: Media forensics and deepfakes: An overview. IEEE J. Sel. Topics Signal Process. **14**(5), 910–931 (Aug 2020)
26. Wan, Z., Zhang, J., Chen, D., Liao, J.: High-fidelity pluralistic image completion with transformers. In: Proc. IEEE Int. Conf. Comput. Vis. pp. 4672–4681. Montreal, QC, Canada (Oct 2021)
27. Wang, C., Huang, Z., Qi, S., Yu, Y., Shen, G., Zhang, Y.: Shrinking the semantic gap: spatial pooling of local moment invariants for copy-move forgery detection. IEEE Trans. Inf. Forensics Security **18**, 1064–1079 (2023)
28. Wang, J., Wu, Z., Chen, J., Han, X., Shrivastava, A., Lim, S.N., Jiang, Y.G.: Objectformer for image manipulation detection and localization. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 2364–2373. New Orleans, LA, USA (Jun 2022)
29. Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., Huang, T.: Seggpt: Towards segmenting everything in context. In: Proc. IEEE Int. Conf. Comput. Vis. pp. 1130–1140 (2023)
30. Wang, X., Wang, H., Niu, S.: An image forensic method for ai inpainting using faster r-cnn. In: Proc. Int. Conf. Artif. Intell. Secur. pp. 476–487. New York, NY, USA (July 2019)
31. Wu, H., Zhou, J.: IID-Net: Image inpainting detection network via neural architecture search and attention. IEEE Trans. Circuits Syst. Video Technol. **32**(3), 1172–1185 (Mar 2022)
32. Yu, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.: Free-form image inpainting with gated convolution. In: Proc. IEEE Int. Conf. Comput. Vis. pp. 4470–4479. Seoul, Korea (Oct 2019)
33. Zhang, J., Ma, K., Kapse, S., Saltz, J., Vakalopoulou, M., Prasanna, P., Samaras, D.: Sam-path: A segment anything model for semantic segmentation in digital pathology. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 161–170. Springer (2023)
34. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. **40**(6), 1452–1464 (2017)

35. Zhou, J., Ma, X., Du, X., Alhammadi, A.Y., Feng, W.: Pre-training-free image manipulation localization through non-mutually exclusive contrastive learning. In: Proc. IEEE Int. Conf. Comput. Vis. pp. 22346–22356 (2023)
36. Zhou, P., Han, X., Morariu, V.I., Davis, L.S.: Learning rich features for image manipulation detection. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. pp. 1053–1061 (2018)
37. Zhu, X., Lu, J., Ren, H., Wang, H., Sun, B.: A transformer–cnn for deep image inpainting forensics. The Visual Computer **39**(10), 4721–4735 (2023)
38. Zhu, X., Qian, Y., Zhao, X., Sun, B., Sun, Y.: A deep learning approach to patch-based image inpainting forensics. Signal Process. Image Commun. **67**, 90–99 (Jun 2018)
39. Zhuo, L., Tan, S., Li, B., Huang, J.: Self-adversarial training incorporating forgery attention for image forgery localization. IEEE Trans. Inf. Forensics Security **17**, 819–834 (2022)
40. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. Adv. Neural Inform. Process. Syst. **36** (2024)