Learning Discriminative Noise Guidance for Image Forgery Detection and Localization

Jiaying Zhu*, Dong Li*, Xueyang Fu†, Gang Yang, Jie Huang, Aiping Liu, Zheng-Jun Zha

University of Science and Technology of China
{zhujy53, dongli6, yg1997, hj0117}@mail.ustc.edu.cn, {xyfu, aipingl, zhazj}@ustc.edu.cn

Abstract

This study introduces a new method for detecting and localizing image forgery by focusing on manipulation traces within the noise domain. We posit that nearly invisible noise in RGB images carries tampering traces, useful for distinguishing and locating forgeries. However, the advancement of tampering technology complicates the direct application of noise for forgery detection, as the noise inconsistency between forged and authentic regions is not fully exploited. To tackle this, we develop a two-step discriminative noise-guided approach to explicitly enhance the representation and use of noise inconsistencies, thereby fully exploiting noise information to improve the accuracy and robustness of forgery detection. Specifically, we first enhance the noise discriminability of forged regions compared to authentic ones using a de-noising network and a statistics-based constraint. Then, we merge a model-driven guided filtering mechanism with a data-driven attention mechanism to create a learnable and differentiable noise-guided filter. This sophisticated filter allows us to maintain the edges of forged regions learned from the noise. Comprehensive experiments on multiple datasets demonstrate that our method can reliably detect and localize forgeries, surpassing existing state-of-the-art methods.

Introduction

Forged images present risks in numerous areas, such as copyright watermark removal, fake news generation, and even evidence falsification in court (Zhang et al. 2023a,b; Lin et al. 2023). Consequently, **image forgery detection and localization (IFDL)** is of paramount importance. However, with the widespread use of techniques like GAN (Isola et al. 2017; Zhu et al. 2017), VAE (Kingma and Welling 2013; Van Den Oord, Vinyals et al. 2017), and homogeneous manipulation (Cong et al. 2022; Ling et al. 2021), the manipulation traces of forged images become visually invisible, making IFDL challenging. Thus, it is crucial to devise effective methods for accurate manipulation trace capture.

Conversely, the noise distribution of authentic and forged regions is inconsistent (Zhou et al. 2018; Wang et al. 2022a), leading to manipulation traces in the noise domain. Many researchers have used this noise information to aid IFDL, achieving significant results. For instance, RGB-N (Zhou et al. 2018) leverages noise features extracted from a steganalysis rich model filter layer to identify the noise inconsistency between authentic and tampered regions. SPAN (Hu et al. 2020) extracts anomalous local noise features from noise maps using CNNs to differentiate heterologous regions. MVSS-Net (Chen et al. 2021) learns multi-view features by utilizing both noise views and boundary artifacts. These methods directly build the end-to-end mapping of noise features to masks and adopt fusion strategies to integrate RGB and noise information to enhance forgery detection accuracy. However, as tampering and post-processing techniques evolve, the difference between the two regions in the noise domain becomes less noticeable or even hidden.

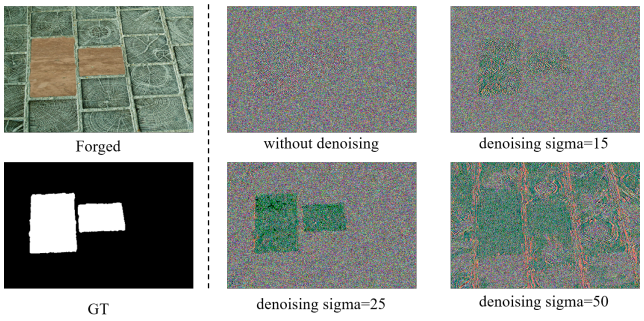


Figure 1: We process the forged image using denoising networks with different standard deviations (CBDNet (Guo et al. 2019) trained with noise with standard deviations of 15, 25, and 50), and then extract the noise separately. For this image, the denoiser of 25 standard deviation can help to obtain discriminative noise. (Best viewed on screen.)

Given these findings, we propose that explicitly learning and leveraging noise inconsistencies can further improve IFDL performance. Therefore, we introduce a novel two-step noise-guided scheme. The first step involves training a noise extractor to explicitly enlarge the noise distribution difference between authentic and forged regions. We use a denoising network followed by a Bayar convolution (Wu, AbdAlmageed, and Natarajan 2019) to construct the noise extractor, optimized using a statistics-based constraint. The rationale for using a denoiser stems from the observation

*Co-first authors contributed equally, † corresponding author.

that a suitable denoising network can amplify the noise distribution difference between two regions. As shown in Figure 1, the denoising network with a standard deviation of 25 maximizes the difference between the two regions, i.e., authentic and forged regions, in the noise domain, while the directly extracted noise cannot achieve the same effect. In order to adaptively tune the denoising network, we impose the customized constraint on the processed image noise, designed based on the Jensen–Shannon (JS) divergence of the Gaussian distribution.

The second phase involves the integration of noise inconsistency and RGB data for forgery detection and localization. Unlike previous fusion strategies, we utilize noise explicitly to guide the RGB branch, significantly enhancing effectiveness. We merge the hand-crafted guided filtering (He, Sun, and Tang 2012) and data-driven attention mechanism (Wang et al. 2018) to create the Cross-Attention-Based Guided Filter (CAGF). Thanks to the local linearity and edge preservation of guided filtering, CAGF not only fully integrates the complementary information in the RGB and noise domains but also ensures the transfer of structural information from the noise domain to the RGB domain. In essence, our method explicitly learns the noise inconsistency in the first phase and utilizes this representation in the second phase. This approach allows us to effectively mine the noise prior and use it to model forgery inconsistency. Our contributions are as follows:

- We propose a novel discriminative noise-guided scheme that explicitly enhances the representation and exploitation of noise inconsistencies.
- We develop a method to highlight noise inconsistencies in forged regions, using a denoising network to process images and a statistics-based constraint to optimize the noise extraction.
- We design a cross-attention-based guided filter that combines model-driven and data-driven technologies to explicitly enhance the guiding effect of noise inconsistencies on the RGB branch, fully utilizing the forgery-informed noise representations.

Extensive experiments on several representative benchmarks show that our method is superior to state-of-the-art methods, especially on the real-life dataset IMD20 (Novozamsky, Mahdian, and Saic 2020).

Related Work

Noise-unrelated IFDL

Most early works tend to focus on a specific type of forgery, including splicing (Huh et al. 2018), copy-move (Cozzolino, Poggi, and Verdoliva 2015), and removal (Aloraini, Sharifzadeh, and Schonfeld 2020). While the above works demonstrate satisfactory performance, the practical application of these methods encounters challenges due to the unpredictability of forgery types. Therefore, recent studies emphasize the need for an approach that employs one model to address multiple forgery types. ManTra-net (Wu, AbdAlmageed, and Natarajan 2019) leverages an end-to-end network, which extracts image manipulation trace features and identifies anomalous regions by assessing how different a local feature is from its reference features. PSCCNet (Liu et al. 2022) uses a progressive spatial-channel correlation module that uses features at different scales and dense cross-connections to generate masks in a coarse-to-fine fashion. Besides, some methods (Liu et al. 2024) explore forgery detection in the frequency domain, such as ObjectFormer (Wang et al. 2022b) captures forged traces from the high-frequency parts of images. Different from the above, we focus on fully mining and exploiting noise inconsistencies.

Noise-assisted IFDL

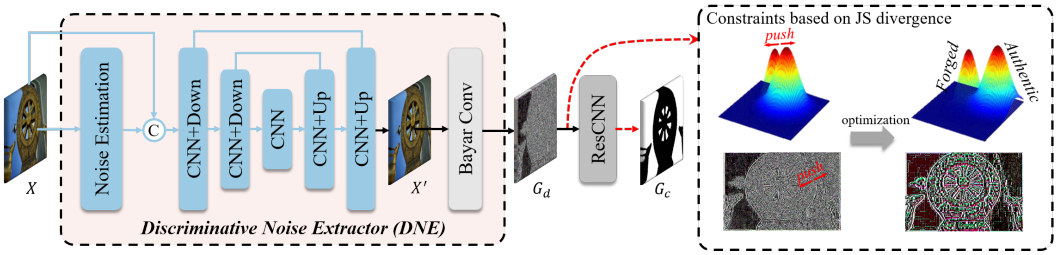
A series of methods use noise information to assist IFDL. Mahdian et al. (Mahdian and Saic 2009) detects changes in noise standard deviations for blind image forensics. Lyu et al. (Lyu, Pan, and Zhang 2014) expose region splicing by revealing inconsistencies in local noise levels. These methods focus on specific tampering artifacts and are limited to specific forgeries. Recently, some deep learning-based IFDL also utilize noise information as assistance. RGB-N (Zhou et al. 2018) explores to leverage noise features to model the inconsistency between tampered and untouched regions. NoiseDF (Wang and Chow 2023) extracts noise traces and features from the video image frames' cropped face and background squares. ERMPC (Li et al. 2023) utilizes noise branches as auxiliary information to facilitate further refinement of forgery localization. TruFor (Guillaro et al. 2023) learns the noise-sensitive fingerprint by training on real data in a self-supervised manner. In contrast, our method explicitly learns the noise inconsistency by statistical constraints and exploits the noise in the form of explicit guidance.

Methodology

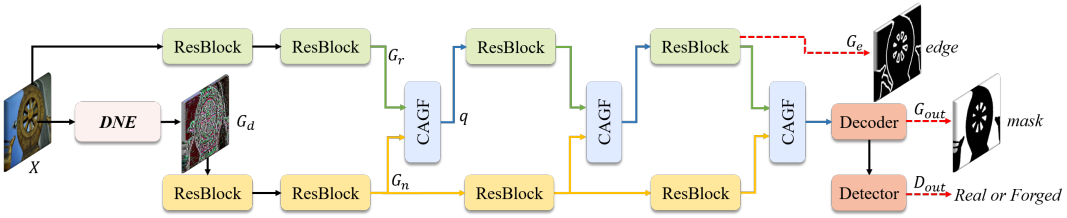
Overview

Noise features between the source and target images are unlikely to match (Zhou et al. 2018), and the tampering operation destroies the natural noise distribution (Wang et al. 2022a). Besides, using noise can suppress the content information, which is beneficial to extract semantic-agnostic features for IFDL (Chen et al. 2021). However, with the development of tampering and post-processing techniques, the inconsistency of the noise is not obvious or even hidden. We argue that explicitly mining and exploiting noise inconsistencies can further improve accuracy and robustness.

Therefore, we propose a two-step strategy to make full use of noise inconsistencies for IFDL, including noise representation learning and noise guided network. First, we propose a learning scheme using a denoising network and a customized constraint. The input image is represented as $X \in \mathbb{R}^{H\times W\times 3}$, where H and W represent the height and width of the image. X is input to the denoising network to obtain the image $X^{'}$, and then the noise $G_d \in \mathbb{R}^{H\times W\times 3}$ is extracted by BayarConv (Wu, AbdAlmageed, and Natarajan 2019). The choice of denoising networks is not the focus of this work, so we use the widely used CBDNet (Guo et al. 2019) for the trade-off of performance and computation. And we impose a statistics-based constraint on G_d to ensure that the noise distributions of the two regions are pulled



(a) Noise Representation Learning Network. The blue components represent the denoising network CBDNet.



(b) Noise Guided Network. The well-trained DNE comes from the Noise Representation Learning Network.

Figure 2: An overview of our two-step discriminative noise guided scheme. It contains noise representation learning and noise guided network. The dashed red lines denote the constraints we imposed.

apart. Second, we train the noise-guided network (NGNet) to explicitly apply noise inconsistency. The NGNet is a dual-branch network that contains multiple cross-attention-based guided filters that progressively guide RGB information with noise inconsistencies for more accurate IFDL.

Noise Representation Learning

To explicitly represent the noise inconsistency, we design a noise representation learning network (NRLNet) as in Figure 2a, which uses a denoising network to process images and statistical losses to optimize the network. Considering that the noise distribution of a forged image is unknown, we use the blind denoising network CBDNet (Guo et al. 2019) and load the weights of blind denoising as initialization. The specific architecture and training strategy is described below.

We process the input image X using a CBDNet-based network. Specifically, we first predict a noise level map $\hat{l} \in \mathbb{R}^{H\times W\times 3}$, which can be viewed as weights related to the noise distribution. Then we feed $\hat{l}$ together with the input X into the encoder-decoder structure to get the image $X^{'} \in \mathbb{R}^{H\times W\times 3}$, which is expressed as:

$$\hat{l} = \mathrm{NE}\left(X\right), \tag{1}$$

$$X^{'} = \mathrm{D}\left(\mathrm{Concat}\left(X, \hat{l}\right)\right), \tag{2}$$

where Concat denotes the concatenate operation. NE is the noise estimation module implemented by a five-layer fully convolutional network, and the convolution kernel size is 3×3. D is a U-Net architecture that obtains images with discriminative noise. Then, following (Chen et al. 2021), we adopt BayarConv to extract the noise $G_d \in \mathbb{R}^{H\times W\times 3}$ from $X^{'}$. Besides, in order to make the learned noise more conducive to IFDL, we feed the noise into the Res-CNN to predict coarse localization result $G_c \in \mathbb{R}^{H\times W\times 1}$. Res-CNN contains ten res-blocks, and one block consists of two 3×3 convolution and ReLU function.

Optimization. To explicitly pull apart the noise distribution of the two regions (authentic and forged), we introduce the JS divergence to constrain G_d. First, we divide G_d into the noise of the authentic region N_a and the noise of the forged region N_f with the help of the mask of the ground truth. The stationary disturbances in images can be modeled as Gaussian (Guo et al. 2019), and both N_a and N_f can be regarded as sampled values of noise. Therefore, we utilize the JS divergence of the continuous Gaussian to measure the distance between the noise distributions of two regions:

$$JSD\left(P_a \parallel P_f\right) = \frac{1}{2}KL\left(P_a \parallel M\right) + \frac{1}{2}KL\left(P_f \parallel M\right), \tag{3}$$

where P_a and P_f are the distributions of N_a and N_f respectively, and M is $\frac{(P_a+P_f)}{2}$. The KL divergence of two Gaussian distributions is calculated as follows:

$$KL\left(P_1 \parallel P_2\right) = log\sigma_2 - log\sigma_1 + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}, \tag{4}$$

where σ_1, σ_2 are the standard deviations of P_1 and P_2, and μ_1, μ_2 are the means respectively. Then, Equation 3 is calculated as follows:

$$JSD = \log\frac{\sqrt{{\sigma_a}^2 + {\sigma_f}^2}}{2} - \frac{\log\sigma_a + \log\sigma_f}{2} + \frac{(\mu_a - \mu_f)^2}{{\sigma_a}^2 + {\sigma_f}^2} + \frac{1}{2}, \tag{5}$$

where σ_a, σ_b are the standard deviations of N_a and N_f, and μ_1, μ_2 are the mean values of N_a and N_f. In addition, if only JS divergence is used as the loss function, the optimization process of the network will oscillate. Therefore, we adopt the loss of forgery localization to assist the optimization of the network, which is also more conducive to the final image forgery localization. We combine the assisted loss and

Algorithm 1: Cross-attention-based Guided Filter (CAGF)

Input: G_n: guidance image(Noise), G_r: input image(RGB),
Output: q: CAGF filtering output.

$mean_n$ = MFConv(G_n)
$mean_r$ = MFConv(G_r)
 # Calculate variance and covariance (CMA: Figure 3)
var_n = CMA(G_n, G_n)
cov_{nr} = CMA(G_n, G_r)
 # Calculate coefficients of local linear relationship
a = ResBlock(Concat(cov_{nr}, var_n))
b = $mean_r$ - $a * mean_n$
$mean_a$ = MFConv(a)
$mean_b$ = MFConv(b)
 # Output
q = $mean_a$ * G_n + $mean_b$
return q

JS divergence to compose the loss function for noise representation learning, which can be written as:

$$\mathbf{L_n} = \lambda\left(1 - JSD\right) + \left(1 - \lambda\right)\mathcal{L}\left(Y, G_c\right), \tag{6}$$

where $\mathcal{L}$ denotes the Dice loss (Chen et al. 2021), $Y \in \mathbb{R}^{H\times W\times 1}$ is the ground-truth, and λ is the hyperparameter to balance the two terms which is set as 0.80.

Noise Guided Network

As shown in Figure 2b, after the NRLNet converges, we embed the well-trained discriminative noise extractor (shown in the box: the denoiser and BayarConv) into the noise guided network (NGNet). Different from previous work, we apply noise in the form of explicit guidance. The network architecture of NGNet, cross-attention-based guided filter (CAGF) and optimization are detailed as follows.

Network architecture. We utilize two branches to handle RGB and noise information. The well-trained discriminative noise extractor is used to obtain the input of the noise branch to better extract forged traces. We use ResNet-50 pretrained on ImageNet (Deng et al. 2009) as the backbone network. Then to guarantee the guiding effect of noise inconsistency on RGB, we design CAGF and place it alternately with the ResNet block, as shown in Figure 2b. Guided by the noise, the RGB branch can extract features highly correlated with tampering artifacts. Finally, we transform the extracted features with plain convolutional layers and bilinear upsampling into the final predicted mask $G_{out} \in \mathbb{R}^{H\times W\times 1}$.

Cross-attention-based guided filter. Existing IFDL methods directly use fusion strategies, which cannot explicitly guarantee that the tampering artifacts in the noisy domain are fully exploited. The guided filter can guarantee the transfer of structural information from the guided image to the target image and has the edge-preserving property (He, Sun, and Tang 2012). Inspired by this, we explore the fusion of noise and RGB information from a guidance perspective thus proposing CAGF, as shown in Algorithm 1. And we use three CAGF blocks in practice. The traditional guided filter is derived from a local linear model. It generates the filtering output by considering the guidance.

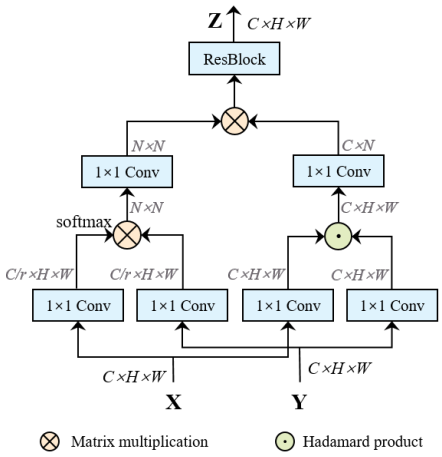


Figure 3: Cross-Modal Attention. It can flexibly calculate the variance and covariance of noise and RGB (Equation 7 to 8). If X and Y are the same, then Z is the variance, and if they are not, then Z is the covariance.

However, the traditional guided filter is a non-trainable algorithm without considering the mutual dependency between the guidance and the target, which is inappropriate for IFDL. Because of the large information gap between noise and RGB, simply transferring structural information from noise to RGB would result in various artifacts. Therefore, based on the traditional algorithm, we use the attention mechanism to calculate the variance and covariance and use the convolutional layer instead of the mean filter. Specifically, denoting input features derived from the noise stream and the RGB stream as $G_n \in \mathbb{R}^{H_s\times W_s\times C_s}$ and $G_r \in \mathbb{R}^{H_s\times W_s\times C_s}$. We take G_n as the guide image and G_r as the input image. First, we design novel cross-modal attention (CMA) to obtain covariance and variance. Taking the calculation of covariance as an example, the input of CMA is G_n and G_r. CMA leverages the computation block described in Figure 3 to convert them to $cov_{nr} \in \mathbb{R}^{H_s\times W_s\times C_s}$:

$$\begin{aligned} cov_{nr} &= \mathrm{CMA}\left(G_n, G_r\right) \\ &= \mathrm{Res}(\mathcal{C}(\mathcal{C}(G_n)^T \otimes \mathcal{C}(G_r)) \otimes \mathcal{C}(\mathcal{C}(G_n) \odot \mathcal{C}(G_r))), \end{aligned} \tag{7}$$

where $\otimes$ is the matrix multiplication, $\odot$ is the Hadamard product, Res denotes the res-block containing two 3×3 convolution and ReLU function, and $\mathcal{C}$ denotes the 1×1 convolution. We perform matrix multiplication on G_n and G_r to obtain $C \in \mathbb{R}^{N\times N}$ ($N = H_s \times W_s$) instead of the correlation coefficient $corr_{nr}$ of traditional guided filtering, and calculate the Hadamard product of the two to replace $mean_n * mean_r$ of the traditional algorithm (He, Sun, and Tang 2012). Before the matrix multiplication of G_n and G_r, the two are converted to $C_s/r\times N$, where r is a scalar that reduces channel dimension for computation efficiency. In the same way, when both inputs of CMA are the noise features G_n, we can obtain the variance $var_n \in \mathbb{R}^{H_s\times W_s\times C_s}$:

$$var_n = \mathrm{CMA}\left(G_n, G_n\right). \tag{8}$$

The res-block is used to obtain coefficient $a \in \mathbb{R}^{H_s\times W_s\times C_s}$:

$$a = \mathrm{Res}\left(\mathrm{Concat}\left(cov_{nr}, var_n\right)\right). \tag{9}$$

Next, we follow the equation in Algorithm 1 to calculate $b \in \mathbb{R}^{H_s \times W_s \times C_s}$. Inspired by (Wu et al. 2018), we use convolution operation (MFConv) instead of the mean filter:

$$mean_M = \mathrm{MFConv}\,(M) = \frac{\mathrm{Conv}\,(M)}{\mathrm{Conv}\,(J)}, \qquad (10)$$

where M is the input of MFConv, J is the all-ones matrix with the same size as M, and Conv is the 3×3 convolution. Finally, we obtain the output of CAGF $q \in \mathbb{R}^{H_s \times W_s \times C_s}$ according to the local linear relationship:

$$q = mean_a \odot G_r + mean_b, \qquad (11)$$

where $mean_a$ and $mean_b$ are the results of a and b after MFConv, the size is $H_s \times W_s \times C_s$.

Detector. For the detector, we apply the ConvGeM proposed by MVSS-Net++ (Dong et al. 2023), which can convert localization results G_{out} into detection prediction D_{out}. ConvGeM strikes a good balance between detection and localization through a decayed skip connection. Thus, we use ConvGem to obtain a more accurate detection result:

$$D_{out} = ConvGeM(G_{out}) \qquad (12)$$

Optimization. Following most studies (Chen et al. 2021; Wang et al. 2022c), we also utilize the edge supervision. However, this is not the focus of this work, so we have used some common methods. Following (Chen et al. 2021), we use the Sobel layer and the residual block to obtain the edge prediction $G_e \in \mathbb{R}^{H_e \times W_e \times 1}$ in a shallow-to-deep manner. For edge loss, the ground-truth edges $E \in \mathbb{R}^{H \times W \times 1}$ is downsampled to a smaller size $E' \in \mathbb{R}^{H_e \times W_e \times 1}$ to match G_e. This strategy outperforms upsampling G_e in terms of computational cost and performance. The loss of NGNet can be written as:

$$\mathbf{L_N} = \alpha\mathcal{L}_1\left(Y, G_{out}\right) + \beta\mathcal{L}_2\left(y, D_{out}\right) + (1-\alpha-\beta)\,\mathcal{L}_3\left(E', G_e\right), \qquad (13)$$

where $\mathcal{L}_1$ and $\mathcal{L}_3$ denote the Dice loss (Chen et al. 2021), $\mathcal{L}_2$ is BCE loss, y is a label that represents the authenticity of the image and α, β are the hyperparameters to balance the loss function. In practice, α is set as 0.60 and β is set as 0.2. Note that authentic images are only used to compute $\mathcal{L}_2$.

Experiments

Experimental Setup

Pre-training Data. We construct a substantial image tampering dataset and employ it for pre-training our model. This dataset comprises three categories: 1) splicing, 2) copy-move, and 3) removal.

Testing Datasets. Following (Liu et al. 2022), we evaluate our model on CASIA (Dong, Wang, and Tan 2013), Coverage (Wen et al. 2016), Columbia (Hsu and Chang 2006), Nist Nimble 2016 (NIST16) (Guan et al. 2019) and IMD20 (Novozamsky, Mahdian, and Saic 2020). We apply the same training/testing splits as (Hu et al. 2020; Wang et al. 2022b) to fine-tune our model for fair comparisons.

Evaluation Metrics. To quantify the localization performance, following previous works (Hu et al. 2020; Wang et al. 2022b), we use pixel-level Area Under Curve (AUC) and F1 score on manipulation masks. To evaluate detection performance, we use image-level AUC and F1 score. Since binary masks are required to compute F1 scores, we adopt the Equal Error Rate (EER) threshold to binarize them.

Image Forgery Localization

Following SPAN (Hu et al. 2020) and ObjectFormer (Wang et al. 2022b), our model is compared with other state-of-the-art tampering localization methods under two settings: 1) training on the synthetic dataset and evaluating on the full test datasets, and 2) fine-tuning the pre-trained model on the training split of test datasets and evaluating on their test split.

Pre-trained Model. Table 1a shows the localization performance of pre-trained models for different methods on five datasets under pixel-level AUC. We compare our model NGNet with MantraNet (Wu, AbdAlmageed, and Natarajan 2019), SPAN (Hu et al. 2020), PSCCNet (Liu et al. 2022), ObjectFormer (Wang et al. 2022b), TANet (Shi, Chen, and Zhang 2023) and HiFi-Net (Guo et al. 2023) when evaluating pre-trained models. The pre-trained NGNet achieves the best localization performance on Coverage, CASIA, NIST16 and IMD20 and ranks second on Columbia. Especially, NGNet achieves 94.1 % on the copy-move dataset Coverage, whose image forgery regions are indistinguishable from the background. This validates our model owns the superior ability to capture tampering traces in the noise domain. We fail to achieve the best performance on Columbia, falling behind TANet 0.2 % under AUC. We contend that the explanation may be that the distribution of their synthesized training data closely resembles that of the Columbia dataset. This is further supported by the results in Table 1b, which show that NGNet performs better than TANet in terms of both AUC and F1 scores. Furthermore, it is worth pointing out NGNet achieves decent results with less pre-training data.

Fine-tuned Model. The network weights of the pretrained model are used to initiate the fine-tuned models that will be trained on the training split of Coverage, CASIA, and NIST16 datasets, respectively. We evaluate the fine-tuned models of different methods in Table 1b. As for AUC and F1, our model achieves significant performance gains. This validates that our method could precisely capture subtle tampering traces by discriminative noise representation learning and cross-attention-based guided filtering.

Image Forgery Detection

To avoid false alarms, we also consider the forgery detection task. Following ObjectFormer (Wang et al. 2022b), we conduct experimental comparisons on the CASIA-D dataset introduced by (Liu et al. 2022). As shown in Table 1c, our method has excellent detection performance, *i.e.*, 99.81% in terms of AUC and 98.72% in F1. Our method explicitly models and exploits noise inconsistencies, thus accurately distinguishing forged images from authentic ones.

Robustness Evaluation

To analyze the robustness of our model for localization, we follow the distortion settings in (Wang et al. 2022b) to degrade the forged images from NIST16. These distortion types include resizing images to different scales, applying Gaussian blur with a kernel size k, adding Gaussian noise with a standard deviation σ, and performing JPEG compression with a quality factor q. We compare the forgery local-

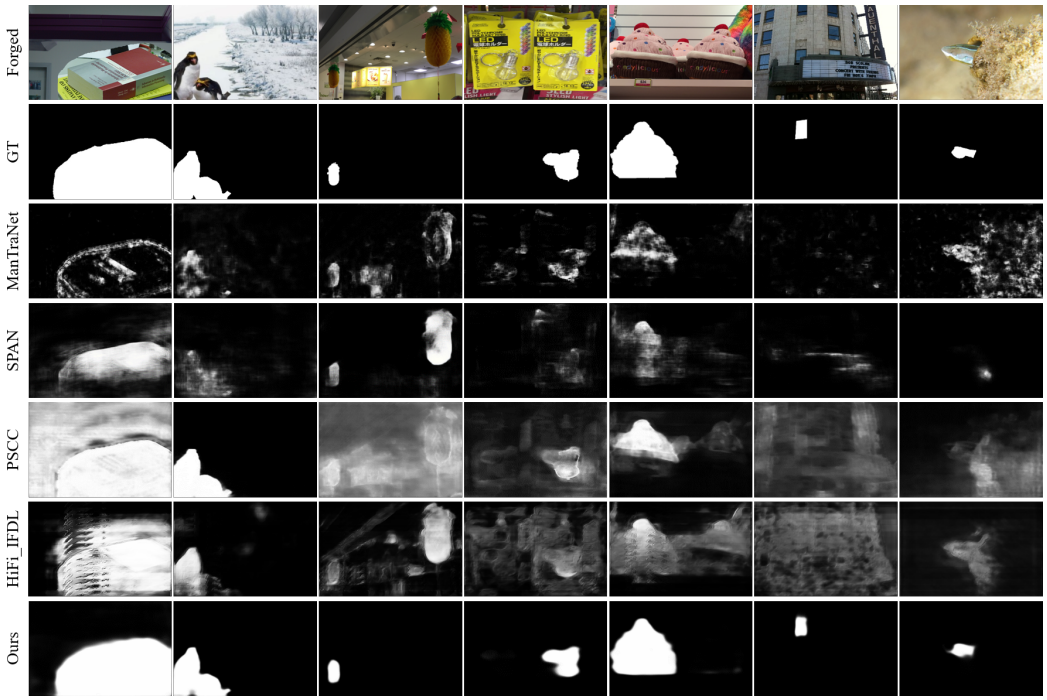


Figure 4: Visualization of the predicted manipulation mask by different methods. From top to bottom, we show forged images, GT masks, predictions of ManTraNet, SPAN, PSCC-Net, HiFi-Net and ours.

Loc.	Data	Col.	Cov.	CAS.	NI.16	IM.20
		Metric: AUC(%) – Pre-trained				
ManTra	64K	82.4	81.9	81.7	79.5	74.8
SPAN	96k	93.6	92.2	79.7	84.0	75.0
PSCC	100k	98.2	84.7	82.9	85.5	80.6
Ob.Fo.	62K	95.5	92.8	84.3	87.2	82.1
TANet	60K	**98.7**	91.4	85.3	89.8	84.9
HiFi	100k	98.3	93.2	85.8	87.0	82.9
Ours	60K	98.5	**94.1**	**87.2**	**90.0**	**85.2**

(a)

Loc.	Cov.	CAS.	NI.16
	Metric: AUC(%) / F1(%) – Fine-tuned		
RGB-N	81.7/43.7	79.5/40.8	93.7/72.2
SPAN	93.7/55.8	83.8/38.2	96.1/58.2
PSCC	94.1/72.3	87.5/55.4	99.6/81.9
Ob.Fo.	95.7/75.8	88.2/57.9	99.6/82.4
TANet	97.8/78.2	89.3/ 61.4	99.7/86.5
HiFi	96.1/80.1	88.5/61.6	98.9/85.0
Ours	**98.1/81.2**	**91.3/62.1**	**99.8/86.8**

(b)

Det.	AUC(%)	F1(%)
ManTra	59.94	56.69
SPAN	67.33	63.48
PSCC	99.65	97.12
Ob.Fo.	99.70	97.34
HiFi	99.50	97.40
Ours	**99.81**	**98.72**

(c)

Table 1: Image forgery detection and localization results. (a) Localization performance of the pre-train model. (b) Localization performance of the fine-tuned model. (c) Detection performance on *CASIA-D* dataset. (Bold means best, underline means second best).

ization performance (AUC scores) of our pretrained models with SPAN and ObjectFormer on these corrupted data, and report the results in Table 2. Our model demonstrates better robustness against various distortion techniques. It is worth noting that JPEG compression is commonly performed when uploading images to social media. And our model performs significantly better on compressed images.

Ablation Study

In this section, we conduct experiments to demonstrate the effectiveness of our method. The noise representation learning (NRL) is designed to explicitly enlarge the difference in the noise distribution between the two regions (authentic and forged). The cross-attention-based guided filter contains the cross-modal attention (CMA) and guided filtering mechanism (GF). CMA fully integrates the complementary information contained in the RGB and noise branches, while GF guarantees the transfer of structural information from the noise to the RGB and has the edge-preserving property. To evaluate the effectiveness of NRL, CMA and GF, we remove them separately from our method and evaluate the forgery localization performance on CASIA and NIST16.

Table 3 presents the quantitative outcomes. The baseline denotes that we just use ResNet-50. It can be seen that without GF, the AUC scores decrease by 4.5 % on CASIA and 5.9 % on NIST16, while without CMA, the AUC scores decrease by 7.6 % on CASIA and 10.9 % on NIST16. Furthermore, when NRL is discarded, serious performance degradation in Table 3, i.e., 9.5% in terms of AUC and 20.9% in terms of F1 on CASIA can be observed.

Since different denoiser may derive different performance, we perform ablation study on the choice of de-

Distortion	SPAN	Ob.Fo.	HiFi	Ours	
no distortion	83.95	87.18	87.0	**90.04**	
Resize(0.78×)	83.24	87.17	86.9	**89.85**	↓0.19
Resize(0.25×)	80.32	86.33	86.5	**88.83**	↓1.21
Blur($k = 3$)	83.10	85.97	86.1	**89.24**	↓0.80
Blur($k = 15$)	79.15	80.26	81.0	**88.77**	↓1.27
Noise($\sigma = 3$)	75.17	79.58	81.9	**89.16**	↓0.88
Noise($\sigma = 15$)	67.28	78.15	79.5	**86.28**	↓3.76
Compress($q = 100$)	83.59	86.37	86.5	**89.19**	↓0.85
Compress($q = 50$)	80.68	86.24	86.0	**88.98**	↓1.06

Table 2: The performance on NIST16 dataset under various distortions. AUC scores are reported (in %), (Blur: GaussianBlur, Noise: GaussianNoise, Compress: JPEG-Compress.)

Variants	CASIA		NIST16	
	AUC	F1	AUC	F1
baseline	75.6	43.0	79.9	69.9
w/o NRL	82.6	49.1	83.2	73.1
w/o CMA	84.4	50.3	88.7	77.2
w/o GF	87.2	55.1	93.9	81.2
Ours	**91.3**	**62.1**	**99.8**	**86.8**

Table 3: Ablation results on CASIA and NIST16 dataset using different variants of our proposed scheme.

noiser. We choose blind denoising networks such as DnCNN (Zhang et al. 2017), FFDNet (Zhang, Zuo, and Zhang 2018), RIDNet (Anwar and Barnes 2019) and DRUNet (Zhang et al. 2021) for comparison. As shown in Table 4, CBDNet trades off performance and computational complexity.

Denoiser	FLOPs(G)	Col.	Cov.	CAS.	NI.16	IM.20
FFDNet	31.80	88.4	87.3	85.2	82.1	75.6
DnCNN	145.52	87.0	89.4	86.1	83.9	78.4
CBDNet*	161.13	**98.5**	**94.1**	**87.2**	90.0	**85.2**
RIDNet	391.82	92.2	93.5	84.6	87.1	82.9
DRUNet	411.65	98.1	93.3	86.8	**90.2**	83.2

Table 4: The ablation study of different denoising architectures. FLOPs(G) is calculated on images of size 512 × 512. AUC scores are reported (in %).

Visualization Results

Qualitative results. As shown in Figure 4, we provide predicted masks of various methods. The results demonstrate that our method can not only locate the tampering regions accurately but also develop sharp boundaries. It benefits from the ability of our model to explicitly enlarge the noise difference between the two regions and preserve the edges.

Visualization of noise representation learning. We show the change of features with and without the NRL in Figure 5. It is clear that NRL facilitates the learning of forgery features and obtains more accurate contours of forged regions. This is because NRL helps the network capture tampering traces in the noise domain.

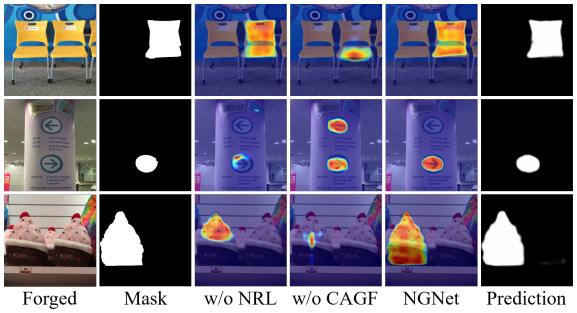


Figure 5: Visualization of noise representation learning and cross-attention-based guided filter. From left to right, we display the forged images, masks, GradCAM (Selvaraju et al. 2017) of the feature map without (w/o) NRL and without CAGF and with both (NGNet), and predictions.

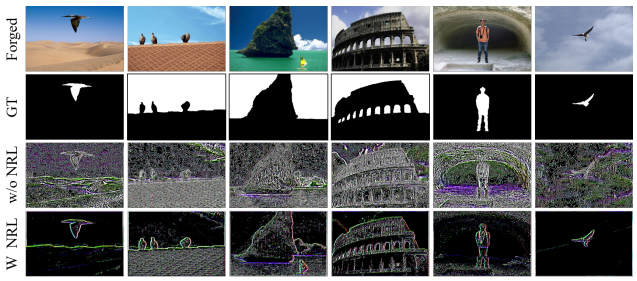


Figure 6: Noise obtained without (w/o) and with (w) NRL. NRL can model more explicit noise inconsistencies.

Visualization of cross-attention-based guided filter. To verify the effect of CAGF, we show the change of features before and after the filter in Figure 5. It can be seen that CAGF can improve the accuracy of forgery localization. The network without CAGF will make false judgments about objects that are similar to the forgery.

Visualization of discriminative noise representation. To further validate the motivation and effectiveness of our method, we show the extracted noise without and with the noise representation learning (NRL) in Figure 6, respectively. It can be seen that NRL obtains a more discriminative noise representation, which is forgery-informed.

Conclusion

In this paper, we propose a two-step noise-guided scheme containing noise representation learning and noise guided network. The first step is to explicitly highlight the discriminability in noise distribution between authentic and forged regions. In the second step, a customized cross-attention-based guided filter that combines model-driven and data-driven technologies is devised to enhance the guiding effect of noise inconsistencies on the RGB branch, fully utilizing the forgery-informed noise representations. Our work provides a new research strategy to solve the problem of difficult extraction of subtle forged traces. Extensive experimental results on several benchmarks demonstrate the effectiveness of the proposed scheme.

Acknowledgements

This work was supported by the National Key R&D Program of China under Grant 2020AAA0105702, the National Natural Science Foundation of China (NSFC) under Grants 62225207, U19B2038 and 62276243.

References

Aloraini, M.; Sharifzadeh, M.; and Schonfeld, D. 2020. Sequential and patch analyses for object removal video forgery detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(3): 917–930.

Anwar, S.; and Barnes, N. 2019. Real image denoising with feature attention. In *Proceedings of the IEEE/CVF international conference on computer vision*, 3155–3164.

Chen, X.; Dong, C.; Ji, J.; Cao, J.; and Li, X. 2021. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14185–14193.

Cong, W.; Tao, X.; Niu, L.; Liang, J.; Gao, X.; Sun, Q.; and Zhang, L. 2022. High-Resolution Image Harmonization via Collaborative Dual Transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18470–18479.

Cozzolino, D.; Poggi, G.; and Verdoliva, L. 2015. Efficient dense-field copy–move forgery detection. *IEEE Transactions on Information Forensics and Security*, 10(11): 2284–2297.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Dong, C.; Chen, X.; Hu, R.; Cao, J.; and Li, X. 2023. 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.

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.

Guo, S.; Yan, Z.; Zhang, K.; Zuo, W.; and Zhang, L. 2019. Toward convolutional blind denoising of real photographs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1712–1722.

Guo, X.; Liu, X.; Ren, Z.; Grosz, S.; Masi, I.; and Liu, X. 2023. Hierarchical Fine-Grained Image Forgery Detection and Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3155–3165.

He, K.; Sun, J.; and Tang, X. 2012. Guided image filtering. *IEEE transactions on pattern analysis and machine intelligence*, 35(6): 1397–1409.

Hsu, J.; and Chang, S. 2006. Columbia uncompressed image splicing detection evaluation dataset. *Columbia DVMM Research Lab*.

Hu, X.; Zhang, Z.; Jiang, Z.; Chaudhuri, S.; Yang, Z.; and Nevatia, R. 2020. SPAN: Spatial pyramid attention network for image manipulation localization. In *European conference on computer vision*, 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.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

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.

Lin, X.; Wang, S.; Deng, J.; Fu, Y.; Bai, X.; Chen, X.; Qu, X.; and Tang, W. 2023. Image manipulation detection by multiple tampering traces and edge artifact enhancement. *Pattern Recognition*, 133: 109026.

Ling, J.; Xue, H.; Song, L.; Xie, R.; and Gu, X. 2021. Region-aware adaptive instance normalization for image harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9361–9370.

Liu, J.; Xie, J.; Wang, Y.; and Zha, Z.-J. 2024. Adaptive Texture and Spectrum Clue Mining for Generalizable Face Forgery Detection. *IEEE Transactions on Information Forensics and Security*, 19: 1922–1934.

Liu, X.; Liu, Y.; Chen, J.; and Liu, X. 2022. PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*.

Lyu, S.; Pan, X.; and Zhang, X. 2014. Exposing region splicing forgeries with blind local noise estimation. *International journal of computer vision*, 110(2): 202–221.

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

Novozamsky, A.; Mahdian, B.; and Saic, S. 2020. IMD2020: A large-scale annotated dataset tailored for detecting manipulated images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, 71–80.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Shi, Z.; Chen, H.; and Zhang, D. 2023. Transformer-Auxiliary Neural Networks for Image Manipulation Localization by Operator Inductions. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–1.

Van Den Oord, A.; Vinyals, O.; et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Wang, J.; Li, Z.; Zhang, C.; Chen, J.; Wu, Z.; Davis, L. S.; and Jiang, Y.-G. 2022a. Fighting Malicious Media Data: A Survey on Tampering Detection and Deepfake Detection. *arXiv preprint arXiv:2212.05667*.

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

Wang, M.; Fu, X.; Liu, J.; and Zha, Z.-J. 2022c. JPEG Compression-aware Image Forgery Localization. In *Proceedings of the 30th ACM International Conference on Multimedia*, 5871–5879.

Wang, T.; and Chow, K. P. 2023. Noise Based Deepfake Detection via Multi-Head Relative-Interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 14548–14556.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803.

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.; Zheng, S.; Zhang, J.; and Huang, K. 2018. Fast end-to-end trainable guided filter. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1838–1847.

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/CVF Conference on Computer Vision and Pattern Recognition*, 9543–9552.

Zhang, F.; Liu, J.; Zhang, Q.; Sun, E.; Xie, J.; and Zha, Z.-J. 2023a. ECENet: Explainable and Context-Enhanced Network for Muti-modal Fact verification. In *Proceedings of the 31st ACM International Conference on Multimedia*, 1231–1240.

Zhang, K.; Li, Y.; Zuo, W.; Zhang, L.; Van Gool, L.; and Timofte, R. 2021. Plug-and-play image restoration with deep denoiser prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6360–6376.

Zhang, K.; Zuo, W.; Chen, Y.; Meng, D.; and Zhang, L. 2017. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7): 3142–3155.

Zhang, K.; Zuo, W.; and Zhang, L. 2018. FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *IEEE Transactions on Image Processing*, 27(9): 4608–4622.

Zhang, Q.; Liu, J.; Zhang, F.; Xie, J.; and Zha, Z.-J. 2023b. Hierarchical Semantic Enhancement Network for Multimodal Fake News Detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3424–3433.

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.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.