Preprint – IJCAI 2025: This is the accepted version made available for conference attendees.
Do not cite. The final version will appear in the IJCAI 2025 proceedings.

ForgDiffuser: General Image Forgery Localization with Diffusion Models

Mengxi Wang[1] , **Shaozhang Niu**[1,2] , **Jiwei Zhang**[1,3]*
[1]Beijing Key Lab of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, China
[2]Southeast Digital Economy Development Institute, China
[3]Key Laboratory of Interactive Technology and Experience System, Ministry of Culture and Tourism (BUPT), China
{wangmengxi, szniu, jwzhang666}@bupt.edu.cn

Abstract

Current general image forgery localization (GIFL) methods confront two main challenges: decoder overconfidence causing misidentification of the authentic regions or incomplete predicted masks, and limited accuracy in localizing forgery details. Recently, diffusion models have excelled as dominant approach for generative models, particularly effective in capturing complex scene details. However, their potential for GIFL remains underexplored. Therefore, we propose a GIFL framework named ForgDiffuser with diffusion models. The core of ForgDiffuser lies in leveraging diffusion models conditioned on the forgery image to efficiently generate the segmentation mask for tampered regions. Specifically, we introduce the attention-guided module (AGM) to aggregate and enhance image feature representations. Meanwhile, we design the boundary-driven module (BDM) with edge supervision to improve the localization accuracy of boundary details. Additionally, the probabilistic modeling and stochastic sampling mechanisms of diffusion models effectively alleviate the overconfidence issue commonly observed in traditional decoders. Experiments on six benchmark datasets demonstrate that ForgDiffuser outperforms existing mainstream GIFL methods in both localization accuracy and robustness, especially under challenging manipulation conditions.

1 Introduction

With the rapid development of AI and image generation techniques, the public can easily and inexpensively fake high-quality images. These images are almost indistinguishable from real ones, and have greatly blurred the boundaries between reality and fiction, bringing unprecedented challenges and crises to social order, information security, and even public perception. Examples span fake news dissemination, judicial evidence falsification, insurance fraud, and academic cheating. It makes the development of general image forgery localization (GIFL) techniques an important issue in the field of computer vision and security, which aims at precisely locating the tampered areas in the forgery image. Generally, image forgery techniques can be categorized into: traditional image forgery techniques (TIF) and AI-generated image forgery techniques (AIGIF). TIF include: splicing [He *et al.*, 2012; Xiao *et al.*, 2020], copy-move [Wu *et al.*, 2018; Chen *et al.*, 2020] and removal [Chen *et al.*, 2024; Feng *et al.*, 2022]. Splicing is copying and pasting specific content from one image to another; copy-move is moving specific content from one area to another area of the image; removal is deleting specific content from the image; AIGIF is redrawing specific areas of an image with diffusion models, GAN, or other generative techniques.

*Corresponding Author

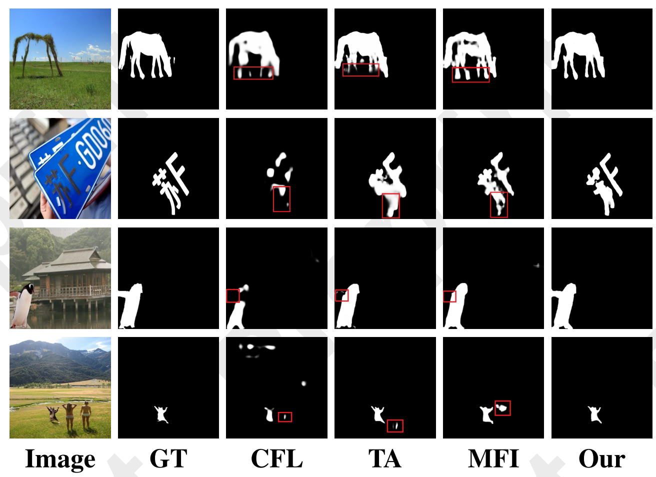

Figure 1: Current GIFL methods suffer from low segmentation accuracy in edge details, as well as overconfident mispredictions and incomplete segmentation masks. We utilize diffusion models to generate predicted masks and incorporate attention-guided feature representation enhancement along with boundary supervision, significantly improving the accuracy of predicted masks.

GIFL methods are usually achieved by capturing specific forgery features to achieve accurate localization of the tampered region. The diverse tampering types impose higher requirements on the model's ability to balance the differences and commonalities of various forgery features, which makes challenging for GIFL algorithms. Currently, numerous GIFL methods have been proposed. Examples include manipulation tracing network (ManTra-Net) [Wu *et al.*, 2019], multi-view multi-scale supervision network (MVSS-Net) [Chen *et al.*, 2021], contrastive learning image forgery localization network (CFL-Net) [Shi *et al.*, 2023], transformer-auxiliary neural network (TA-Net) [Niloy *et al.*, 2023], multi-feature fusion identification network (MFI-Net) [Ren *et al.*, 2023], and edge distribution guidance and contrastive learning network (EC-Net) [Hao *et al.*, 2024], etc. These methods usually learn the forgery clues left by the forgery manipulation to identify the tampered region, which has made significant progress in GIFL. However, current traditional frameworks still suffer from the following problems: 1) The end-to-end network design often leads to decoder overconfidence, resulting in inaccurate and incomplete predictions. 2) The high degree of blending between forgery regions and the background causes blurred and imprecise boundary localization. Examples of these issues are demonstrated in Figure 1.

To address the above challenges, we consider GIFL as a mask generation task using diffusion models. The diffusion model models the probability distribution of data through progressive denoising and incrementally refining the predicted mask, effectively mitigates the detail localization ambiguity problem. Meanwhile, random sampling generates multiple predictions and evaluates the uncertainty of the predictions, thus effectively mitigating the overconfidence of the decoder. Therefore, we propose a diffusion model-based framework ForgDiffuser, which aims to efficiently generate the tampered region mask by leveraging the faked image as conditional input. Specifically, in the training phase, first, we design the attention-guided module (AGM) to aggregate multi-layer image features efficiently to enhance the richness and contextual expression of image features. Second, we devise the boundary-driven module (BDM) to enhance the detail processing capability of ForgDiffuser by incorporating edge supervision. In the inference stage, We propose the global-local consistency fusion (GLCF) strategy to enhance prediction stability and reliability by fusing predicted masks from multiple sampling steps.

Our main contributions are as follows:

1) We propose a diffusion model-based GIFL method called ForgDiffuser. To improve tampered region prediction, we design the attention-guided module within the conditional network to extract more reliable image features.

2) We design the edge-driven module to further enhance the detail perception capability of the ForgDiffuser.

3) We conduct extensive experiments on six benchmark datasets, demonstrating that ForgDiffuser outperforms existing GIFL methods, especially in localization accuracy and robustness.

2 Related Work

2.1 General Image Forgery Localization

The core issue of GIFL is commonly formulated as a binary segmentation task, and it requires the methods capable of accurately dividing the forgery image into two categories: the untampered region and the tampered region. Earlier, GIFL methods mainly depended on handcrafted features or specific artifacts, such as JPEG artifacts [Amerini *et al.*, 2017], noise patterns [Zhou *et al.*, 2018], and edge inconsistencies [Salloum *et al.*, 2018], etc. However, handcrafted features are redundant and costly, limiting their prevalence in practical applications.

In response to the above, deep neural networks (DNN) can automatically extract deep robust features due to their strong feature learning capabilities [He and Xiao, 2023], which improves the accuracy and generalization of the algorithms. This effectively addresses the limitations of traditional methods and has been widely used in GIFL tasks. For example, ManTra-Net [Wu *et al.*, 2019] proposes an end-to-end network and formulates the task as a local anomaly detection problem. However, it struggles to effectively model global contextual information and accurately capture tampering details. Therefore, recent research has focused on developing more sophisticated and powerful feature extraction mechanisms. Several approaches enhance feature learning by incorporating auxiliary information such as boundaries, texture, and frequency cues. MVSS-Net [Chen *et al.*, 2021] exploits noise distribution and boundary artifacts around tampered regions to facilitate more generalizable feature learning. TA-Net [Shi *et al.*, 2023] introduces the edge-assisted strategy to further refine the boundary details of the predicted mask. CFL-Net [Niloy *et al.*, 2023] leverages noise information extracted by steganalysis rich model (SRM) filters and contrast learning approach to improve the separability between authentic and tampered regions. EC-Net [Hao *et al.*, 2024] employs a two-stage localization strategy from coarse to fine that significantly improves the localization accuracy. Despite the significant progress achieved in GIFL, current methods still suffer from decoder overconfidence that results in incorrect and incomplete predictions, as well as imprecise boundary localization in complex forgery scenarios.

2.2 Diffusion Models

Diffusion models are generative methods grounded in probabilistic modeling[Ho *et al.*, 2020], leveraging the Markov chain to iteratively denoise random noise into high-quality data. Diffusion models have recently achieved a wide range of successful applications in computer vision, demonstrating excellent scalability, stability, and strong capabilities in addressing complex visual tasks. For example, diffusion models have shown remarkable performance in tasks such as semantic segmentation [Wu *et al.*, 2023], image super-resolution [Gao *et al.*, 2023], anomaly detection [Zhang *et al.*, 2023], object detection [Chen *et al.*, 2023], and monocular depth estimation [Saxena *et al.*, 2024]. Compared to the traditional GIFL framework, the iterative denoising mechanism of diffusion models offers significant advantages in handling complex scenes and diverse objects, while also enabling more precise control over the generative process. In this research, we use the conditional diffusion model framework for the GIFL task, which significantly improves the localization accuracy of tampered regions.

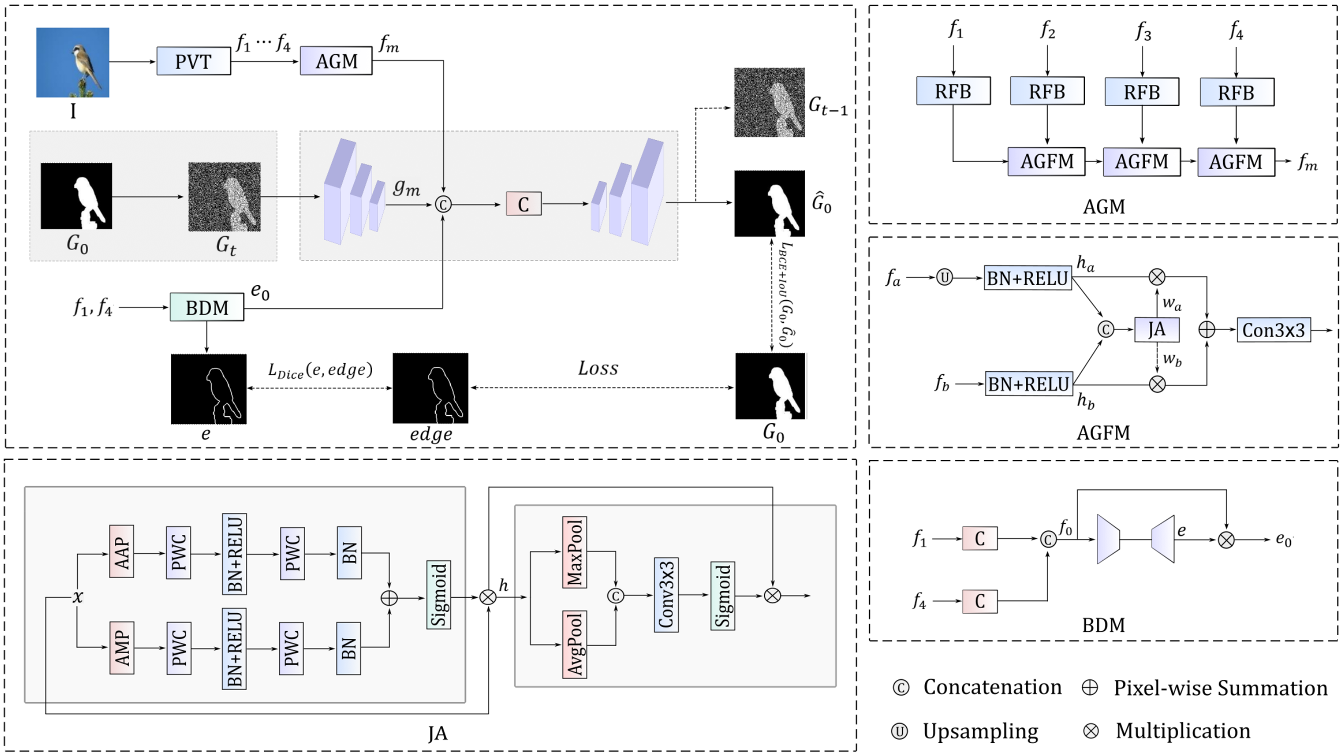

Figure 2: The architecture of ForgDiffuser, the core modules include the attention-guided module (AGM) and the edge-driven module (BDM). AGM is composed of the joint attention mechanism (JA) and the attention-guided feature fusion module (AGFM). In the training process, the ground truth G_0 is transformed into the noisy version G_t through the diffusion process. The conditional image features f_m, edge information e_0, and the noisy mask G_t are then fed into the denoising network to generate the predicted mask $\hat{G}_0$. The model is trained by minimizing the combined loss of the predicted mask and the predicted edge.

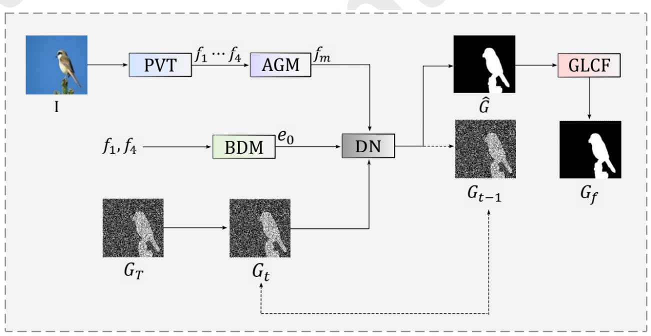

Figure 3: The inference architecture of ForgDiffuser, mainly composed of the denoising network (DN) and the global-local consistency fusion (GLCF) strategy.

3 Proposed Method

3.1 Overview

The overall architecture of ForgDiffuser is illustrated in Figure 2. In the training process (shown in the top-left of Figure 2), we first input the RGB image into the pre-trained PVTv2 backbone [Wang *et al.*, 2022] to extract the multi-level features, denoted as f_1, f_2, f_3, and f_4. These features are then processed by the proposed AGM to produce the multi-level fused representation f_m. Specifically, the features f_1, f_2, f_3, and f_4 are passed through four receptive field blocks (RFB), composed of inflated convolutions with various kernel sizes, and then fused via the attention-guided feature fusion module (AGFM) to obtain the fusion feature f_m. Next, we input the features f_1 and f_4 into the BDM to generate the predicted edge map e and the edge feature representation e_0. Simultaneously, we add noise to the ground truth mask G_0 via the forward diffusion process to produce the noisy mask G_t, which is then fed into the denoising network to obtain the predicted mask $\hat{G}_0$. In ForgDiffuser, a lightweight UNet-based architecture is adopted as the denoising network. Specifically, The noisy mask G_t is first encoded through a series of convolutional layers combined with downsampling operations. Following the encoding stage, the f_m output from the AGM is concatenated with the e_0 from the BDM to form the fusion feature representation g_m. Subsequently, g_m is passed through the decoder, which consists of convolutional layers and upsampling operations, to produce the predicted mask.

In the inference process, as illustrated in Figure 3, ForgDiffuser starts from a random noise image G_T and progressively generates predicted masks over T time steps, guided by forgery features and edge information. To improve the stability and reliability of the result, we introduce the GLCF strategy to fuse T predictions and obtain the final mask G_f.

3.2 Diffusion Model Process

The core idea of diffusion models is to progressively add noise to the data in the forward process, driving it toward randomness, and to learn the reverse denoising process that gradually reconstructs the original data.

Forward process: The forward process is modeled as a Markov chain that gradually adds Gaussian noise to the original data G_0 over T time steps, resulting in a noisy sample G_T. At each step, Gaussian noise is added according to the following formulation:

$$q(G_t \mid G_{t-1}) = \mathcal{N}(G_t; \sqrt{1-\beta_t} G_{t-1}, \beta_t I), \tag{1}$$

where β_t denotes the noise variance at time step t, typically increasing with t. The forward process from step 1 to t can be equivalently described by the distribution:

$$q(G_t \mid G_0) = \mathcal{N}(G_t; \sqrt{\bar{\alpha}_t} G_0, (1-\bar{\alpha}_t) I), \tag{2}$$

where $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$, and $\alpha_t = 1 - \beta_t$.

Reverse process: The reverse process aims to recover the original data G_0 from the pure noise sample G_T by iteratively denoising. Assuming the forward noise schedule β_t is known, the reverse process is formulated as a conditional distribution:

$$p_\theta(G_{t-1} \mid G_t) = \mathcal{N}(G_{t-1}; \mu_\theta(G_t, t), \Sigma_\theta(G_t, t)), \tag{3}$$

where $\Sigma_\theta(G_t, t) = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t} \beta_t$, and $\mu_\theta(G_t, t)$ is parameterized by ForgDiffuser as:

$$\mu_\theta(G_t, t) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t} G_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t} \hat{G}_0, \tag{4}$$

where $\hat{G}_0$ is the predicted mask generated by ForgDiffuser.

3.3 Attention-guided Module

With the continued development of image forgery and generation techniques, forgery images have become increasingly sophisticated and diverse. Effectively leveraging local artifacts and global semantic consistency is therefore critical for GIFL. Shallow Transformer layers tend to retain more local details, whereas deeper layers capture richer contextual representations. To effectively combine these hierarchical representations, the AGM is introduced. It consists of the joint attention mechanism and the attention-guided feature fusion module. The fused feature obtained from the AGM is utilized as the additional condition to support the subsequent mask prediction process.

Joint attention mechanism (JA): Inspired by [Woo *et al.*, 2018], we propose the JA that combines channel and spatial attention, aiming to improve the ForgDiffuser's ability to extract key features. For the channel attention component, as shown in the bottom-left of Figure 2, the input feature x is processed in parallel by two branches: one applies adaptive average pooling (AAP), and the other applies adaptive max pooling (AMP), both along the spatial dimensions. Then, each obtained feature is independently passed through the pointwise convolution (PWC), batch normalization (BN), and ReLU. The two resulting feature maps are summed element-wise and passed through the sigmoid function. The obtained map is applied to the input via element-wise multiplication to generate the channel-attention feature h.

For the spatial attention component, the channel-attended feature h is taken as input to further emphasize salient spatial information. Specifically, we compute the maximum (Max) and average (Avg) values along the channel dimension. These two resulting maps are concatenated and processed by a 3×3 convolution followed by the sigmoid function. The obtained map is then multiplied element-wise with the feature h to produce the final output. The formulation of JA is given below:

$$\begin{aligned} h = {} & Sigmoid(BN(PWC(BNR(PWC(AMP(x))))) \\ & + BN(PWC(BNR(PWC(AAP(x))))))x, \end{aligned} \tag{5}$$

$$JA(x) = Sigmoid(C(Cat([Avg(h), Max(h)])))h, \tag{6}$$

where BNR denotes BN and ReLU, and C is the convolution operation.

Attention-guided feature fusion module (AGFM): The AGFM is designed to aggregate the multi-layer feature as the comprehensive guide for the subsequent mask prediction. As illustrated on the right side of Figure 2, the inputs f_a and f_b are first aligned in spatial dimensions by upsampling (Up) f_a via bilinear interpolation. Subsequently, both features are then normalized with BN and activated by ReLU, resulting in feature maps h_a and h_b. To enhance salient features, h_a and h_b are concatenated and fed into the JA, which generates the attention weight w_a and its complementary weight w_b. These weights are then applied to h_a and h_b, respectively, through element-wise multiplication. Finally, the weighted features are summed and processed by a 3×3 convolution. The formula for AGFM is as follows:

$$h_a = R(BN(Up(f_a))), h_b = R(BN(f_b)), \tag{7}$$

$$w_a = JA(Cat([h_a, h_b]), w_b = 1 - w_a \tag{8}$$

$$AGFM(f_a, f_b) = C(Cat([w_a h_a, w_b h_b])), \tag{9}$$

where R stands for ReLu.

3.4 Boundary-driven Module

Currently, the transitions between tampered regions and their backgrounds have become more visually consistent. Therefore, accurately detecting the boundaries between forgery and authentic regions is crucial for improving the performance of GIFL methods. To address this, we propose the BDM designed to enhance boundary representation and thus improve localization accuracy.

As illustrated in the bottom-right of Figure 2, the input features to BDM consist of two features: f_1 and f_4. Here, f_1 is the low-level feature map that preserves rich spatial details, while f_4 is the high-level semantic feature map capturing abstract contextual information. In the BDM, f_1 is upsampled and then concatenated with f_4 to form the fused feature map f_0, which integrates both fine-grained details and high-level semantics. This fused feature is subsequently fed into a lightweight encoder-decoder network to predict the edge map e, which serves as a supervision signal to enhance boundary localization during training. Then, we perform an element-wise multiplication between the predicted edge map e and the fused feature f_0, yielding the edge-enhanced feature representation e_0. This representation is then used to guide the subsequent tampering mask prediction. The operations are formally defined as:

$$f_0 = Cat([C(f_1), C(f_4)]), \tag{10}$$

$$BDM(f_1, f_4) = ED(f_0)f_0, \tag{11}$$

where ED is the lightweight encoder-decoder network.

3.5 Loss Function

ForgDiffuser is designed to predict the forgery localization mask directly. To ensure that the predicted mask generated through the reverse diffusion process progressively approximates the ground truth, we adopt the Weighted Binary Cross-Entropy (BCE) and Weighted Intersection over Union (IoU) losses for mask supervision [Wei *et al.*, 2020]. Additionally, the Dice loss is employed to supervise the edge prediction [Xie *et al.*, 2020]. The overall training objective of ForgDiffuser is defined as follows:

$$Loss = \lambda_1 L_{BCE+IoU}(G_0, \hat{G_0}) + \lambda_2 L_{Dice}(e, edge). \tag{12}$$

3.6 Sampling Strategy

To mitigate overconfident incorrect segmentations in GIFL, inspired by [Zhang *et al.*, 2021], we employ time ensemble to integrate predicted masks from T sampling steps. Then, we design the global-local consistency fusion (GLCF) strategy to enhance the stability and reliability of the predicted mask.

Specifically, let the predicted mask at time t be $\hat{G}_t(x, y)$ and all predicted masks from sampling phase be $\left\{\hat{G}_t(x, y)\right\}_{t=1}^{T}$. First, for each step, calculate the global variance of the sample to quantify the predictive stability.

$$\sigma_{\text{g}}^2(x, y) = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{G}_t(x, y) - \bar{G}(x, y)\right)^2, \tag{13}$$

where $\bar{G}(x, y)$ denotes the mean value of predicted masks across all time steps. The global weights $W_{\text{global}}(x, y) = e^{-\sigma_{\text{g}}^2(x,y)}$ are constructed based on global variance, which suppress low-quality sampling steps with large global variance and focus on stable predictions. Next, for the sampling result at each time step, calculate the local variance in 5×5 neighborhood of each pixel, measuring the uncertainty of the local prediction.

$$\sigma_{\text{l},t}^2(x, y) = \frac{1}{N}\sum_{(u,v)\in\mathcal{N}_{x,y}}\left(\hat{G}_t(u, v) - \bar{G}_{\text{l},t}(x, y)\right)^2, \tag{14}$$

where $\bar{G}_{\text{l},t}(x, y)$ is the neighborhood mean, $\mathcal{N}_{x,y}$ represents the neighborhood window centered at (x, y), and $N = |\mathcal{N}_{x,y}| = 25$. The local weights $W_{\text{local},t}(x, y) = e^{-\sigma_{\text{l},t}^2(x,y)}$ are based on local variance, which can suppress the high-frequency noise and clear the boundary of predicted mask.

Finally, multiply and normalize the global and local weights, and perform weighted fusion on predictions at each time step to obtain the final predicted mask $G_f(x, y)$.

$$W_t(x, y) = \frac{W_{\text{global}}(x, y) \cdot W_{\text{local},t}(x, y)}{\sum_{k=1}^{T} W_{\text{global}}(x, y) \cdot W_{\text{local},k}(x, y) + \epsilon}, \tag{15}$$

$$G_f(x, y) = \sum_{t=1}^{T} W_t(x, y) \cdot \hat{G}_t(x, y), \tag{16}$$

where $W_t(x, y)$ denote the integration weights, and $\epsilon = 10^{-8}$ prevents division-by-zero errors.

4 Experiments

4.1 Datasets and Evaluation Metrics

ForgDiffuser is evaluated on widely used forgery image datasets: CASIA1 [Dong *et al.*, 2013], DID [Wu and Zhou, 2021], IMD [Novozamsky *et al.*, 2020], Auto [Jia *et al.*, 2023], BSN and RLS26K [Hao *et al.*, 2024]. These datasets encompass main types of current image forgery techniques. CASIA1 contains 921 images, which includes both splicing and copy-move, and uses image enhancement for data post-processing. DID contains 10 different image inpainting methodologies, including deep learning-based and traditional-based methods, each method contributing 1,000 images, for a total of 10,000 images. IMD is the real manipulation dataset with 2,010 forgery images. Auto is the AIGIF dataset generated by the DALL-E2 model [Ramesh *et al.*, 2022]. BSN is an AIGIF dataset constructed with Brushnet method [Ju *et al.*, 2024] and contains 2500 images. RLS26K is a large-scale TIF dataset containing splicing, copy-move, and removal, which includes 26,000 images. We divide the above datasets into train and test sets in the ratio of 9:1 for experimentation.

In order to evaluate the performance of ForgDiffuser comprehensively, we adopt two evaluation metrics: F1-score (F1), and Intersection over Union (IoU).

4.2 Implementation Details

We implemented ForgDiffuser based on the PyTorch with one NVIDIA L20 with 48 GB memory for training and inference. We trained 100 epochs with batch sizes of 16. AGM is initialized using PVTv2-B4, and the input images are resized to 352×352. The AdamW optimizer is employed, and the initial learning rate is set to 0.001. λ_1 and λ_2 in the loss function of Equation 12 are set to 0.8 and 0.2, respectively. A higher value of λ_1 encourages the model to prioritize mask prediction, while still maintaining a balanced emphasis on edge information. The time step T is set to 10 for sampling.

4.3 Comparison with State-of-the-arts

Quantitative comparisons: Table 1 demonstrates the quantitative results of ForgDiffuser with five baseline methods on six benchmark datasets. It is obvious from experimental data that ForgDiffuser achieves optimal results on five datasets, which proves that ForgDiffuser can effectively detect splicing, copy-move, real-world forgery, and AI forgery. Especially on IMD dataset, F1 of ForgDiffuser increases by 0.05 over the suboptimal baseline model EC-Net, and IOU increases by 0.04. On the CASIA1 dataset, F1 of ForgDiffuser only decreases by 0.008 compared to EC-Net, and the quantitative results outperformed EC-Net on the other five benchmark datasets. ForgDiffuser has superior performance in real-world and complex forgery scenarios, primarily due to its integration of the diffusion models' iterative denoising mechanism with edge-enhanced supervision. In conclusion, the experimental results can demonstrate the superior performance of ForgDiffuse in GIFL.

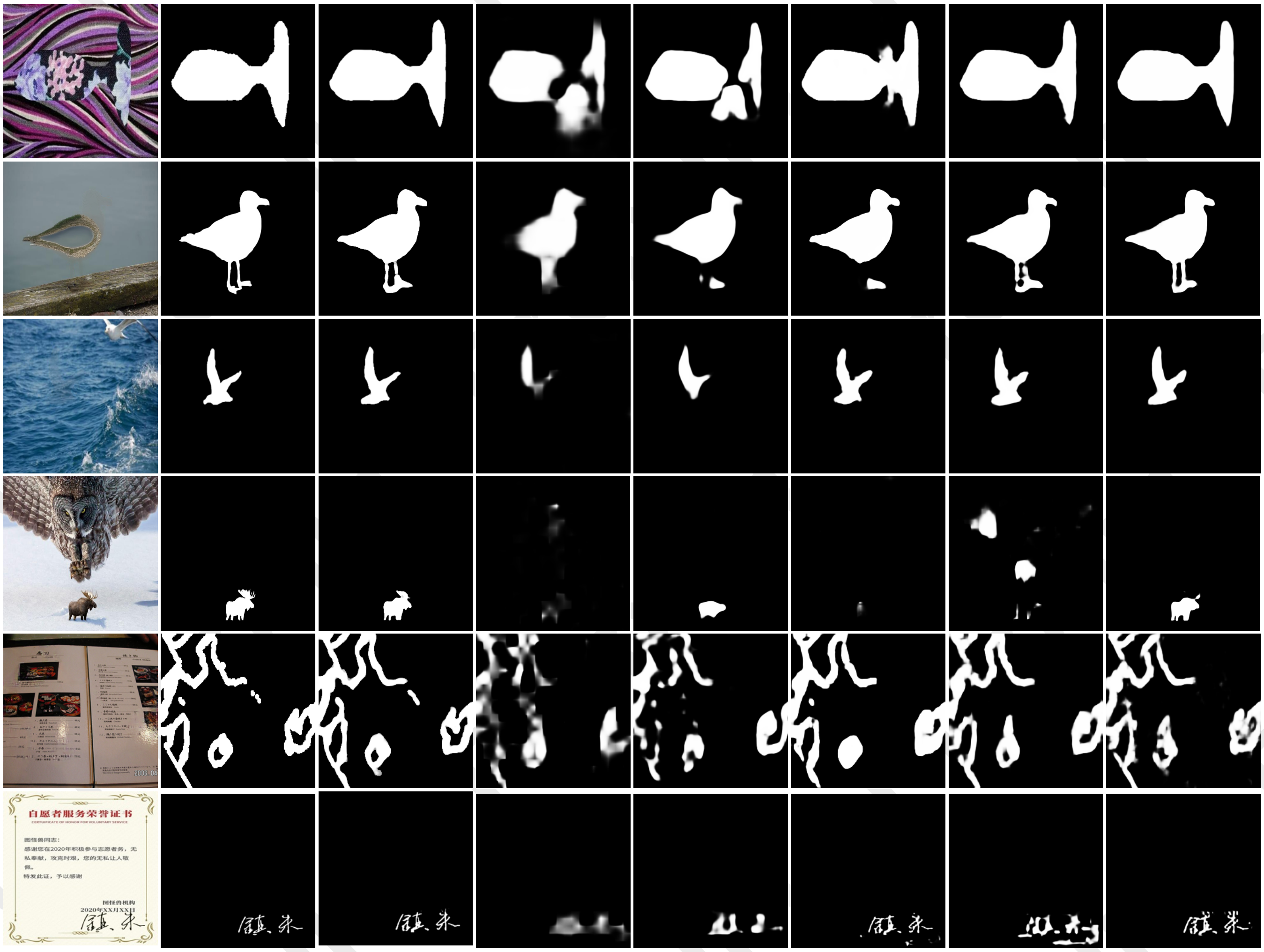

Figure 4: Visual comparison of localization results with different methods.

Qualitative comparisons: In order to compare the results of different methods more intuitively, we show the visualization of the predicted masks of ForgDiffuse and other baseline models on six benchmark datasets in Figure 4. From the visualization results, it is obvious that ForgDiffuser is able to avoid overconfident incorrect segmentation and provide more complete localization results (e.g., lines 1-2). In addition, ForgDiffuser achieved greater accuracy in localizing edge details (e.g., lines 4-6), which proves the effectiveness of BDM.

4.4 Ablation Study

In this subsection, we conduct ablation experiments on the proposed ForgDiffuser to verify the effectiveness of each designed module. The experiments were conducted on CASIA1, DID, IMD, and BSN datasets with the default settings described in Section 4.2. Table 2 presents the results of ablation experiments.

As shown in rows 2-3 of Table 2, the introduction of AGM significantly improved the performance of ForgDiffuser. AGM adaptively extracts features from the conditioned image through the attention mechanism and better combines local information and global context. Specifically, compared to the baseline, the F1 on CASIA1, DID, IMD, and BSN datasets increase by 0.06, 0.02, 0.06, and 0.03, respectively. Meanwhile, the experimental results in lines 3-4 indicate that F1 and IoU show improvements on all datasets with the introduction of BDM, further validating its effectiveness. In addition, the results in lines 1-2 demonstrate the effectiveness of the sampling strategy in the proposed method.

4.5 Robustness Evaluation

To evaluate the robustness of ForgDiffuser, we conducted experiments on CASIA1 and IMD datasets using common image attack methods, including Gaussian noise with standard deviation of 0.02, 0.04, 0.06, 0.08 and 0.1; salt & pepper noise with noise intensity of 0.02, 0.04, 0.06, 0.08 and 0.1. Gaussian noise simulates continuous perturbations such as sensor noise or transmission interference. In contrast, salt & pepper noise represents discrete distortions like pixel-level corruption or abrupt intensity changes. These attacks degrade

Methods	Datasets											
	CASIA1		DID		IMD		Auto		BSN		RLS26K	
	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU
MVSS-Net	0.6114	0.5323	0.9141	0.8615	0.3405	0.2533	0.9669	0.9382	0.5456	0.4466	0.3145	0.2523
CFL-Net	0.6148	0.5365	0.9313	0.8844	0.2842	0.2014	0.9682	0.9403	0.5531	0.4478	0.3531	0.2875
TA-Net	0.6325	0.5754	0.9706	0.9469	0.3961	0.3961	0.9761	0.9544	0.7881	0.7002	0.4594	0.3981
MFI-Net	0.7126	0.7126	0.9590	0.9268	0.4532	0.3634	0.9736	0.9498	0.8157	0.7220	0.5037	0.4697
EC-Net	**0.8194**	**0.7676**	0.9641	0.9350	0.5561	0.4765	0.9757	0.9537	0.8344	0.7506	0.5693	0.4997
ForgDiffuser	0.8113	0.7612	**0.9645**	**0.9357**	**0.6076**	**0.5166**	**0.9768**	**0.9577**	**0.8358**	**0.7527**	**0.5708**	**0.5003**

Table 1: Quantitative comparison of F1 and IoU on six benchmark datasets. The best results are in bold.

Methods	Datasets							
	CASIA1		DID		IMD		BSN	
	F1	IoU	F1	IoU	F1	IoU	F1	IoU
base w/o sampling strategy	0.7447	0.6641	0.9340	0.8863	0.5272	0.4298	0.7672	0.6547
base	0.7479	0.6663	0.9340	0.8861	0.5312	0.4336	0.7703	0.6589
base+AGM	0.8048	0.7482	0.9530	0.9230	0.5963	0.5084	0.8044	0.7052
base+AGM+BDM	**0.8113**	**0.7612**	**0.9645**	**0.9357**	**0.6076**	**0.5166**	**0.8358**	**0.7527**

Table 2: Ablation study of module contributions in ForgDiffuser, evaluated based on F1 and IoU.

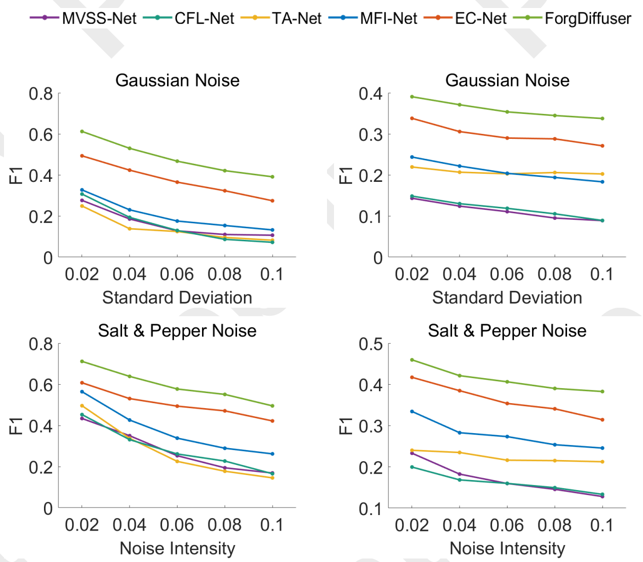

Figure 5: Experimental results of different methods on Gaussian noise and salt & pepper noise. The experiments were performed on CASIA1 and IMD datasets, using F1 as evaluation indicators. ForgDiffuser has significant advantages in robustness.

edge and structural cues in tampered regions, making the tampered areas harder to distinguish from authentic areas and challenging the GIFL task in both localization accuracy and robustness. The experimental results are shown in Figure 5. The left column shows the results on CASIA1 dataset and the right column presents the ones on IMD dataset. From the experimental results, it can be concluded that as the intensity of Gaussian and salt & pepper noise increases, the tampering localization accuracy of all models exhibits a decreasing trend. ForgDiffuser achieves the best performance under both image attack types, significantly outperforming other methods and demonstrating strong robustness.

5 Conclusion

In this paper, we propose ForgDiffuser, a GIFL framework based on conditional diffusion models. The core of ForgDiffuser lies in predicting the tampered region mask through the iterative generation mechanism of diffusion models. It effectively alleviates decoder overconfidence through the iterative sampling strategy. To further improve detection accuracy, we design the AGM to deeply fuse the global semantic features with the low-level detail features of the conditioned image, providing more precise guidance for subsequent mask prediction. In addition, the BDM is introduced to precisely capture edge details between tampered regions and the background, effectively enhancing the accuracy of boundary localization. Experimental results on multiple benchmark datasets show that ForgDiffuser achieves superior performance in detection accuracy and robustness compared to existing mainstream methods, demonstrating its strong potential in GIFL.

Acknowledgments

This work is supported by National Key Research and Development Program of China [grant number 2024YFF0907404].

References

[Amerini *et al.*, 2017] Irene Amerini, Tiberio Uricchio, Lamberto Ballan, and Roberto Caldelli. Localization of jpeg double compression through multi-domain convolutional neural networks. In *2017 IEEE Conference on computer vision and pattern recognition workshops (CVPRW)*, pages 1865–1871. IEEE, 2017.

[Chen *et al.*, 2020] Beijing Chen, Weijin Tan, Gouenou Coatrieux, Yuhui Zheng, and Yun-Qing Shi. A serial image copy-move forgery localization scheme with source/target distinguishment. *IEEE Transactions on Multimedia*, 23:3506–3517, 2020.

[Chen *et al.*, 2021] Xinru Chen, Chengbo Dong, Jiaqi Ji, Juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14185–14193, 2021.

[Chen *et al.*, 2023] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 19830–19843, 2023.

[Chen *et al.*, 2024] Shuang Chen, Amir Atapour-Abarghouei, and Hubert PH Shum. Hint: High-quality inpainting transformer with mask-aware encoding and enhanced attention. *IEEE Transactions on Multimedia*, 2024.

[Dong *et al.*, 2013] Jing Dong, Wei Wang, and Tieniu Tan. Casia image tampering detection evaluation database. In *2013 IEEE China summit and international conference on signal and information processing*, pages 422–426. IEEE, 2013.

[Feng *et al.*, 2022] Xin Feng, Wenjie Pei, Fengjun Li, Fanglin Chen, David Zhang, and Guangming Lu. Generative memory-guided semantic reasoning model for image inpainting. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7432–7447, 2022.

[Gao *et al.*, 2023] Sicheng Gao, Xuhui Liu, Bohan Zeng, Sheng Xu, Yanjing Li, Xiaoyan Luo, Jianzhuang Liu, Xiantong Zhen, and Baochang Zhang. Implicit diffusion models for continuous super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10021–10030, 2023.

[Hao *et al.*, 2024] Qixian Hao, Ruyong Ren, Kai Wang, Shaozhang Niu, Jiwei Zhang, and Maosen Wang. Ecnet: General image tampering localization network based on edge distribution guidance and contrastive learning. *Knowledge-Based Systems*, 293:111656, 2024.

[He and Xiao, 2023] Yang He and Lingao Xiao. Structured pruning for deep convolutional neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 2023.

[He *et al.*, 2012] Zhongwei He, Wei Lu, Wei Sun, and Jiwu Huang. Digital image splicing detection based on markov features in dct and dwt domain. *Pattern recognition*, 45(12):4292–4299, 2012.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[Jia *et al.*, 2023] Shan Jia, Mingzhen Huang, Zhou Zhou, Yan Ju, Jialing Cai, and Siwei Lyu. Autosplice: A text-prompt manipulated image dataset for media forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 893–903, 2023.

[Ju *et al.*, 2024] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision*, pages 150–168. Springer, 2024.

[Niloy *et al.*, 2023] Fahim Faisal Niloy, Kishor Kumar Bhaumik, and Simon S Woo. Cfl-net: image forgery localization using contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4642–4651, 2023.

[Novozamsky *et al.*, 2020] Adam Novozamsky, Babak Mahdian, and Stanislav Saic. 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*, pages 71–80, 2020.

[Ramesh *et al.*, 2022] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[Ren *et al.*, 2023] Ruyong Ren, Qixian Hao, Shaozhang Niu, Keyang Xiong, Jiwei Zhang, and Maosen Wang. Mfi-net: Multi-feature fusion identification networks for artificial intelligence manipulation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(2):1266–1280, 2023.

[Salloum *et al.*, 2018] Ronald Salloum, Yuzhuo Ren, and C-C Jay Kuo. Image splicing localization using a multi-task fully convolutional network (mfcn). *Journal of Visual Communication and Image Representation*, 51:201–209, 2018.

[Saxena *et al.*, 2024] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. *Advances in Neural Information Processing Systems*, 36, 2024.

[Shi *et al.*, 2023] Zenan Shi, Haipeng Chen, and Dong Zhang. Transformer-auxiliary neural networks for image manipulation localization by operator inductions. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4907–4920, 2023.

[Wang *et al.*, 2022] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022.

[Wei *et al.*, 2020] Jun Wei, Shuhui Wang, and Qingming Huang. F^3net: fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12321–12328, 2020.

[Woo *et al.*, 2018] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[Wu and Zhou, 2021] Haiwei Wu and Jiantao Zhou. Iid-net: Image inpainting detection network via neural architecture search and attention. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3):1172–1185, 2021.

[Wu *et al.*, 2018] Yue Wu, Wael Abd-Almageed, and Prem Natarajan. Busternet: Detecting copy-move image forgery with source/target localization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 168–184, 2018.

[Wu *et al.*, 2019] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. 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*, pages 9543–9552, 2019.

[Wu *et al.*, 2023] Weijia Wu, Yuzhong Zhao, Mike Zheng Shou, Hong Zhou, and Chunhua Shen. Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1206–1217, 2023.

[Xiao *et al.*, 2020] Bin Xiao, Yang Wei, Xiuli Bi, Weisheng Li, and Jianfeng Ma. Image splicing forgery detection combining coarse to refined convolutional neural network and adaptive clustering. *Information Sciences*, 511:172–191, 2020.

[Xie *et al.*, 2020] Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. Segmenting transparent objects in the wild. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*, pages 696–711. Springer, 2020.

[Zhang *et al.*, 2021] Jing Zhang, Deng Ping Fan, Yuchao Dai, Saeed Anwar, and Nick Barnes. Uncertainty inspired rgb-d saliency detection. *IEEE Transactions on Software Engineering*, PP(99), 2021.

[Zhang *et al.*, 2023] Xinyi Zhang, Naiqi Li, Jiawei Li, Tao Dai, Yong Jiang, and Shu-Tao Xia. Unsupervised surface anomaly detection with diffusion probabilistic model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6782–6791, 2023.

[Zhou *et al.*, 2018] Peng Zhou, Xintong Han, Vlad I. Morariu, and Larry S. Davis. Two-stream neural networks for tampered face detection. 2018.