big data and cognitive computing

MDPI

Article

Hierarchical Progressive Image Forgery Detection and Localization Method Based on UNet

Yang Liu, Xiaofei Li, Jun Zhang, Shuohao Li, Shengze Hu and Jun Lei *

Laboratory for Big Data and Decision, National University of Defense Technology, Changsha 410000, China; liuyang.nudt@nudt.edu.cn (Y.L.); xf@nudt.edu.cn (X.L.); zhangjun1975@nudt.edu.cn (J.Z); lishuohao@nudt.edu.cn (S.L); springsun@nudt.edu.cn (S.H.)
* Correspondence: leijun1987@nudt.edu.cn

Abstract: The rapid development of generative technologies has made the production of forged products easier, and AI-generated forged images are increasingly difficult to accurately detect, posing serious privacy risks and cognitive obstacles to individuals and society. Therefore, constructing an effective method that can accurately detect and locate forged regions has become an important task. This paper proposes a hierarchical and progressive forged image detection and localization method called HPUNet. This method assigns more reasonable hierarchical multi-level labels to the dataset as supervisory information at different levels, following cognitive laws. Secondly, multiple types of features are extracted from AI-generated images for detection and localization, and the detection and localization results are combined to enhance the task-relevant features. Subsequently, HPUNet expands the obtained image features into four different resolutions and performs detection and localization at different levels in a coarse-to-fine cognitive order. To address the limited feature field of view caused by inconsistent forgery sizes, we employ three sets of densely cross-connected hierarchical networks for sufficient interaction between feature images at different resolutions. Finally, a UNet network with a soft-threshold-constrained feature enhancement module is used to achieve detection and localization at different scales, and the reliance on a progressive mechanism establishes relationships between different branches. We use ACC and F1 as evaluation metrics, and extensive experiments on our method and the baseline methods demonstrate the effectiveness of our approach.

Keywords: computer vision; fake images; detection and localization; hierarchical network; UNet

Citation: Liu, Y.; Li, X.; Zhang, J.; Li, S.; Hu, S.; Lei, J. Hierarchical Progressive Image Forgery Detection and Localization Method Based on UNet. *Big Data Cogn. Comput.* **2024**, *8*, 119. https://doi.org/10.3390/bdcc8090119

Academic Editor: Moulay A. Akhloufi

Received: 5 July 2024
Revised: 8 August 2024
Accepted: 4 September 2024
Published: 10 September 2024

1. Introduction

Is seeing really believing? The answer is no. With the rapid advancements in AI-generated content (AIGC) technology within the realm of image processing, individuals can now produce highly realistic images using generative models. However, a generated model is akin to a double-edged sword. While it offers numerous conveniences in areas such as image editing [1], image repair [2], and image fusion [3], when AI-tampered images are used with malicious intent, people are unable to discern their authenticity through visual inspection alone. Furthermore, the traditional detection models struggle to accurately assess their authenticity attributes, let alone pinpoint the manipulated areas. As illustrated in Figure 1a, the traditional detectors frequently misclassify real and forged images, leading humans to face unprecedented security risks in the domains of information and cognitive security.

The most prevalent image generation methods can be categorized into three main types: Variational Auto-Encoder (VAE) [4], Generative Adversarial Networks (GANs) [5], and Diffusion Models (DMs) [6]. These generative models fundamentally learn the distribution of images from extensive training datasets to produce similar images. Initially, the generative models, such as DDPM [7] and GDDIM [8], focused on generating complete images by learning the distribution of the images, where all the pixels were considered

to be synthetic. However, with advancements in the generative models, the recent methods have evolved to facilitate partial image editing, known as "inpainting." Examples include Inpaint Anything [3] and HD-painter [2]. As depicted in Figure 1b, tampered images now contain both real and synthetic pixels, with the attacker's primary focus being on the manipulated region. Contrary to the complex and intricate requirements of the traditional manual image tampering methods, the generative-model-based partial editing methods offer significant advantages by simplifying the tampering process based on the input instructions. This approach is gradually becoming the mainstream method for image tampering, posing greater challenges for the identification of synthetic images.

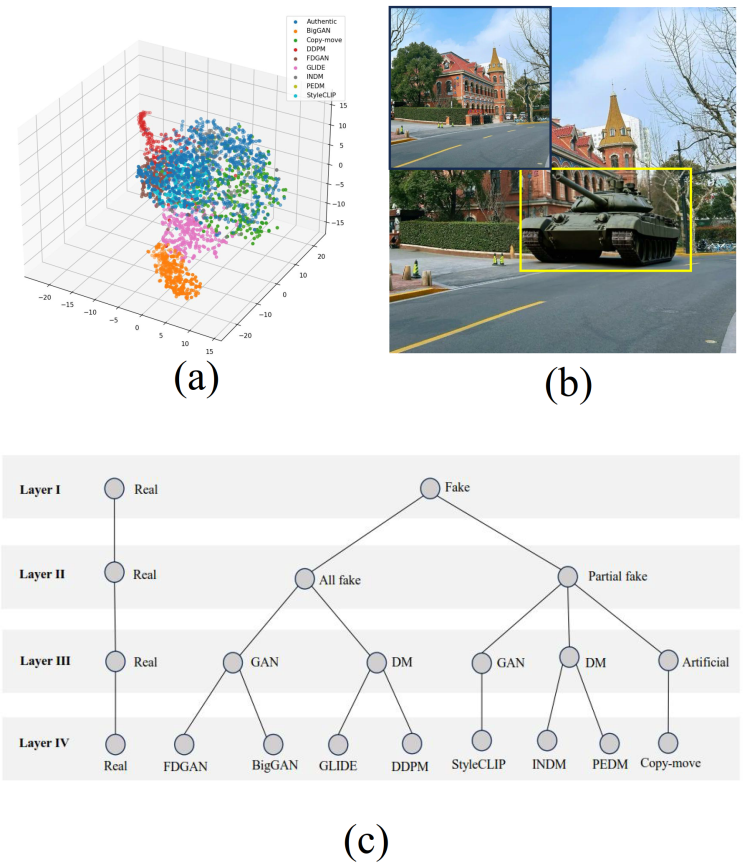

Figure 1. Description of forged image detection. (**a**) Classified t-SNE images of the dataset in the ResNet50 network. (**b**) Examples of AI tampering with images. (**c**) Schematic diagram of image multi-level label division.

Indeed, the detection and localization of forged images have always constituted a pivotal research area within the domain of artificial intelligence security. When an image is found to be forged, individuals not only aspire to identify its forgery but also aim to further refine the process by pinpointing the specific forged region within the image. This enables the comprehension of the attacker's intentions, ultimately facilitating the mitigation of the adverse effects stemming from such forged information.

Previous studies [9–13] have indeed attained notable accomplishments in the domain of forged image detection and localization. However, a majority of these methodologies primarily concentrate on the traditional artificial tampering techniques, such as replication and splicing, or are restricted to categorizing image authenticity. Consequently, their efficacy in detecting and locating the latest AI-generated forged images is significantly diminished. Hence, apart from the legal constraints imposed on the generation models, it becomes imperative to enhance the detection and localization capabilities of forged images at the technical level.

This paper addresses the challenges posed by AI-generated forged images. We propose a progressive layered network, based on UNet, for the refined detection and localization of forged images. Initially, to facilitate progressive detection and localization, we re-classified the forged images from our previously established AI-generated forged image dataset and assigned more reasonable multi-level labels. The specific hierarchical classification and multi-level labels are depicted in Figure 1c. Specifically, from the first to the fourth layer, we gradually subdivide the image attributes, transitioning from the coarse-grained category of authenticity to fine-grained forgery methods. For instance, the forgery image in Figure 1b is assigned a multi-level label of 'Forgery -> partial tampering -> DM -> INDM'.

Furthermore, it has been observed in [11,14,15] that images generated by different forgery methods exhibit distinct frequency domain deviations, which can serve as forgery fingerprints for detection tasks. At the same time, we have also noticed that the generation process of fake images is closely related to noise. Moreover, the quality of the generated models varies, often leading to color fluctuations in the forgery regions. Therefore, we draw inspiration from the previous methods [10,12,16] to jointly utilize multi-type image features from both the spatial and frequency domains. In our method, the spatial domain features specifically include RGB features that capture abnormal fluctuations in the color space and noise features that mine the noise level of the image. The frequency domain features are captured using multiple Laplacian operators to capture the image frequency fluctuations. In addition, to learn richer feature representations, we use a dual-branch attention fusion module to fuse the spatial features. In the dual-branch attention fusion module, we introduce external attention [17] to handle the positional relationships of the spatial features. At the same time, we utilize detection and localization results as thresholds to filter those channel features that are strongly related to the detection performance. We apply these channel thresholds to the spatial features to enhance those features that are strongly related to the task and suppress those features that are weakly related to the task.

Subsequently, due to variations in the size of the tampering region within the generated model, utilizing a fixed-resolution feature map poses the challenge of restricting the feature field of view for forged regions of different scales. Therefore, to mitigate the adverse effects of the scale variations in the forged region on detection and localization, we employ three sets of densely interconnected hierarchical networks. These networks facilitate comprehensive information exchange between the features of varying resolutions through multiple upsampling and downsampling operations, enabling the preservation of both the local and global features and addressing the issue of a limited feature field of view.

Finally, we adopt a hierarchical and progressive approach for detection and localization. To establish dependencies between feature maps of different resolutions, we integrate a multi-scale feature interaction module into the UNet [18] network structure. Using the decoder, we fuse low-resolution feature maps with high-resolution feature maps from bottom to top. Additionally, we leverage the detection and localization results from the low-resolution feature maps as priors to guide the detection and localization process at higher resolutions. The experimental results demonstrate the effectiveness of this approach as the coarse detection and localization outcomes from the low-resolution feature maps prove beneficial when used as priors for guiding the detection and localization at higher resolutions. In the skip connections of our UNet structure, we introduce a convolutional block attention module with soft-threshold constraints (t-CBAM) to capture rich contextual dependencies. The threshold selection is achieved by multiplying the channel and spatial average pooling with the channel attention weights and spatial attention weights, respectively. With our proposed model, we have achieved significant improvements in the detection and localization accuracy of AI-tampered images, surpassing the baseline methods.

The contributions of this article are the following:

1. This article combines external attention and channel attention as a dual-branch attention feature enhancement module, using the feedback results of the detection and localization as dynamic thresholds to enhance the strongly related features and suppress the weakly related features.

2. This article proposes a combination of a hierarchical network and UNet network structure with soft-threshold attention, and it establishes hierarchical dependency relations.
3. This article proposes a hierarchical and progressive forged image detection method called HPUNet, which successfully achieves the accurate detection and localization of AI-generated forged images and further improves the accuracy of detection and localization compared to the baseline methods.

This work extends from our previous research [19] in several key aspects. Firstly, instead of merely discussing whether an image is generated from text, we have assigned more reasonable multi-level labels to our AI-generated forged image dataset. This approach ensures that the hierarchical detection results are more aligned with human cognitive laws. Secondly, we have introduced an external attention mechanism to optimize the spatial attention process of the features. Additionally, we utilize detection results as dynamic thresholds to constrain the dual-branch feature fusion within the context feature enhancement module. This enhancement strategy aims to amplify the task-relevant features while suppressing the weakly relevant ones. Thirdly, we have incorporated the UNet network structure, leveraging the decoder to establish connections between feature maps of different resolutions. Furthermore, we have introduced a soft-threshold dual-attention mechanism in the skip connections to retain the main semantic features and eliminate irrelevant ones.

2. Related Works

2.1. Image Forgery Generation

Generative models and detection models can be viewed as a paired sword and shield. The traditional methods for creating forged images often involve manual tampering techniques, such as copying, splicing, and moving, which demand intricate operations and consume significant effort. In contrast, the latest generative methods utilize image generative models to produce forged images. These models are based on the principle of learning the distribution of a large number of real images and subsequently generating similar ones. Currently, the mainstream image generative models are primarily based on VAEs, GANs, and DMs, with DMs being the most widely used. Apart from the classification based on the fundamental models, different generative methods can also be categorized into two types according to the generation approach. One type generates a complete image in a whole-image generation manner, such as CycleGAN [20], DDPM [7], StyleGAN [21], GLIDE [22], Stable Diffusion [23], etc. The images generated by these methods can be considered as having all fake pixels. The other type partially edits real images by locally tampering with certain parts, for instance, DragGAN [24], HD-painter [2], Paint-by-example [25], StyleCLIP [26], Imagic [1], etc. The fake images produced by the generative models are more realistic and have more concealed forgery boundaries than the traditional manually tampered images. Users can complete image tampering with simple instructions, making it difficult for detection models to accurately discern, and thus such images are harder for models to successfully detect and locate.

2.2. Image Forgery Detection

The objective of fake image detection is to accurately distinguish between fake and real images. This type of detection is regarded as binary classification of images, and several effective detection methods have been proposed in the early research works. Li et al. [27] propose a framework that enhances the performance of forgery localization by integrating tampering possibility maps. This framework selects and improves detectors based on statistical features and copy-move forgery detectors, adjusting their results to generate tampering possibility maps. Both Arshed et al. [28] and Ojha et al. [29] utilize ViT as a powerful feature extractor to capture the global features of images for detecting fake images. AISMSNet [30] introduces a fake image detection method based on Siamese networks, which learns the unique features of specific image regions and detects tampering attributes through feature comparison. Wan et al. [31] incorporate a two-branch data

augmentation and attention mechanism into the task of fake image detection. Guo et al. [15] propose the use of a hierarchical mechanism to achieve the layer-by-layer detection of fake images. To address the generalization issue of detectors towards unknown forgery methods, Epstein et al. [32] adopt an online learning approach. They train N models based on the historical release dates of known forgery methods and test them on the next (N + k) model.

Apart from the efforts regarding detection frameworks, many researchers adopt methods that jointly utilize multiple image features for fake image detection and achieve promising results. Verdoliva [33] investigates data-driven forensic methods to detect deepfake images. Xi et al. [13] and Huang et al. [12] both combine the noise features and RGB features of images to design a dual-stream detection network for image detection. Niloy et al. [16] utilize images processed with the SRM filter and combine them with RGB features as the network input. Guo et al. [15] extract the frequency features and color features of images as input signals for their hierarchical detection network in their research. Some researchers have also introduced image descriptions for fake detection. For instance, Sha [14] and Wu [34] et al. leverage the correlation between text prompts and images for fake detection. Additionally, there are researchers exploring new representations of artifacts. For example, Wang et al. [35] determine that generated images exhibit smaller reconstruction errors than real images when reprocessed by the generative model. Zhong et al. [36] discover that AI-generated images leave varying degrees of artifacts in texture-rich and texture-poor regions. Guillaro et al. [10] divide images into blocks and use a denoising network to find that fake images produce more severe noise deviations. However, most of the aforementioned works only consider fake detection at the image level, neglecting the localization of specific forgery regions.

2.3. Image Forgery Localization

The generation methods of fake images are diverse, and, in practical applications, it is necessary to not only identify that an image is fake but also to further locate the forgery regions. Some researchers have conducted corresponding research work in fake image localization and achieved progress. Cozzolino et al. [9] and Guillaro et al. [10] both use a noise-sensitive fingerprint to learn the relevant noise deviations caused by external camera processing to achieve pixel-level localization. Wu et al. [37] employ a self-supervised learning approach to learn the features from 385 types of manipulations and treat the manipulation localization problem as a local outlier detection problem, using Z-score features to capture the local outliers. Dong et al. [38] utilize semantic-irrelevant image noise distribution features and boundary features to achieve the accurate localization of manipulation regions. Liu et al. [39] and Guo et al. [15] both use hierarchical networks to detect the fake attributes of images and employ self-attention mechanisms to locate the forgery regions. Zhang et al. [40] analyze both the original image and the noise image to locate the forgery regions in the image. Zhou et al. [41] propose a class activation map for manipulation edges and use this map in a weakly supervised framework to locate the manipulation regions in manipulated images. Liu et al. [42] leverage noise features to fully expose subtle changes in images caused by manipulation operations. Although the above methods have achieved good results in the task of fake image localization, there is still room for improvement regarding localization accuracy.

3. Methods

Our goal is to extract artifacts hidden in forged images and leverage a hierarchical and progressive network to enhance the model's ability to detect and locate forged images. In this section, we will introduce HPUNet, as shown in Figure 2, which consists of three modules: the contextual feature enhancement module (Section 3.1), the multi-scale feature interaction module (Section 3.2), and the progressive detection and localization module (Section 3.3). In Section 3.4, the loss function used in the experiments will be introduced.

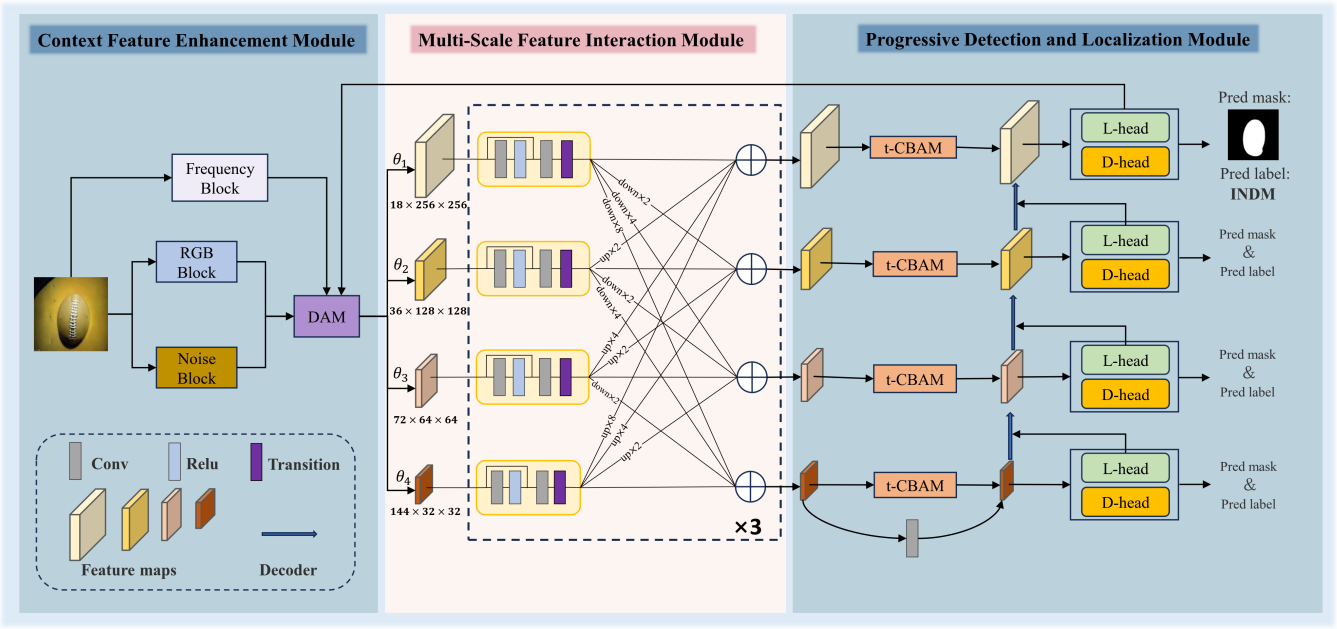

Figure 2. General structure of the HPUNet network. It combines multiple types of image features for detection and localization, and the dual-branch attention mechanism amplifies strongly relevant features while suppressing weakly relevant features. Combined with UNet to construct a hierarchical network, it achieves accurate detection and localization of forged images in a coarse-to-fine cognitive order.

3.1. Context Feature Enhancement Module

We leverage the spatial domain features and frequency domain features of images jointly for the detection and localization of forged images. Specifically, in the spatial domain, we capture abnormal color fluctuations and noise characteristics in the forged areas by extracting the RGB and noise features of images, and employ a dual-branch attention fusion module to integrate these features. In the frequency domain, we use multiple Laplacian operators to capture different responses of images in the frequency domain. Finally, the image features in the spatial and frequency domains are processed jointly in a cascaded manner as the overall feature input.

During the aforementioned process, we noticed that different channels of the feature map possess varying degrees of expressive power, and suppressing weakly task-related features under resource constraints is a crucial task. Therefore, we utilize a dual-branch attention fusion module to fuse and enhance the RGB and noise features in the spatial domain, as shown in Figure 3. In this module, we introduce an external attention [17] mechanism to capture the contextual positional relationships contained in the image features. At the same time, we directly use the input features as key and query to obtain the similarity between each feature channel through multiplication, and derive the feature channel weights after excluding self-similarity. To enhance task-related features and suppress weakly task-related ones, we utilize the detection and localization results as dynamic thresholds to filter the relationships captured between feature channels in channel attention according to Equation (1). The new channel weights are then applied to the rich contextual positional features output by the external attention. This mechanism helps the model to focus more on task-related features while suppressing weakly related ones.

$$C_w = \begin{cases} C_w + C_w * \alpha & \text{if } C_w > threshold \\ C_w - C_w * \beta & \text{if } C_w < 1 - threshold \\ C_w & \text{else} \end{cases} \quad (1)$$

where C_w represents the channel weight, threshold represents the threshold value, and the threshold is obtained through the feedback of detection and positioning results. α and β are hyperparameters that control the variation in the channel weight. Through this weight transformation, the model focuses more attention on the strongly task-related features.

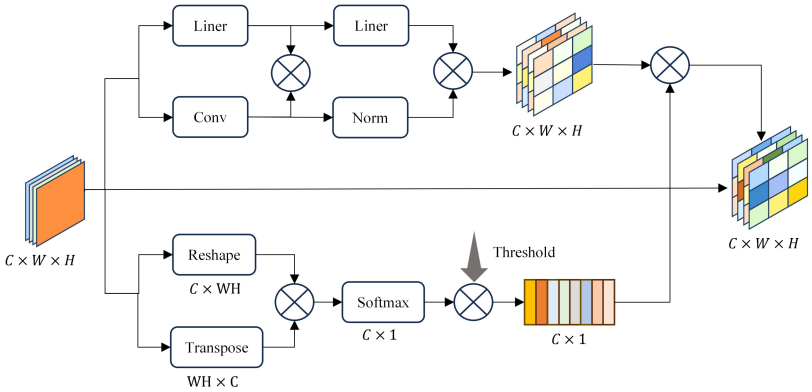

Figure 3. Two-branch attention fusion module.

3.2. Multi-Scale Feature Interaction Module

Due to the varying sizes of the tampered regions in images caused by the generative model, and the potential information loss when using a feature map of a fixed resolution, we aim to overcome the limited visibility of features across different forgery scales. Therefore, we leverage a hierarchical network composed of three sets of densely cross-connected interaction modules to fully interact with feature maps of different resolutions, capturing richer local and global features. The structure of this module is illustrated in the middle part of Figure 2. For any feature map of a specific resolution, we fuse all feature maps of different resolutions with the current resolution's feature map through sampling techniques, and this serves as the output of the current branch.

First, we extract features from the given input image through a context feature enhancement module. Then, we set up four branches to obtain feature maps of four different resolutions, denoted as θ_b, where b belongs to (1...4). Each branch can extract feature maps of a specific resolution, and full connectivity is established between different branches to enable sufficient feature interaction. The output of each scale branch is obtained by fusing the outputs from all branches through upsampling or downsampling. Specifically, taking the output process of branch θ_2 as an example, as shown in Figure 4, we downsample the feature map from branch θ_1 by a factor of 2, upsample the feature map from branch θ_3 by a factor of 2, and upsample the feature map from branch θ_4 by a factor of 4. The sampled outputs from these three branches are then fused with the feature map of θ_2 itself to obtain the output feature map of θ_2.

3.3. Progressive Detection and Localization Module

Inspired by previous works [15,39], we employ a hierarchical network to achieve coarse-to-fine forgery detection and localization tasks. Specifically, we first obtain four feature maps of different resolutions from the aforementioned process and fully interact between these feature maps of varying resolutions. Then, starting from the lowest-resolution feature map in a bottom-up manner, we perform class prediction and region prediction in a detection module and a localization module for each hierarchical level. It is worth mentioning that, considering the possible information loss in lower-resolution features that may hinder precise classification, we adopt a multi-level label structure, as shown in Figure 1c, for predictions at different levels. Finally, we utilize a decoder to concatenate low-resolution and high-resolution feature maps and constrain the detection and localization results of the previous layer with the current layer's detections and localizations as prior knowledge until the final predicted labels and prediction masks are obtained. During

the progressive detection process, we express the output of the detection head for branch θ_b and the predicted probability as $D_b(X)$ and $p(y_b|X)$, and we can calculate

$$p(y_b|X) = softmax(D_b(X) \odot (1 + p(y_{b-1}|X))) \tag{2}$$

Concurrently, we express the output mask and predicted probability map of the localization head for branch θ_b as $mask_b(X)$ and $p(mask_b|X)$, and we can compute

$$p(mask_b|X) = L_b(mask_{b-1}(X) \odot F_b) \tag{3}$$

where L_b represents the localization head of branch θ_b, and F_b represents the feature input for branch θ_b. By relating the output of branch θ_{b-1} to branch θ_b according to the above two equations, we establish a progressive detection and localization path.

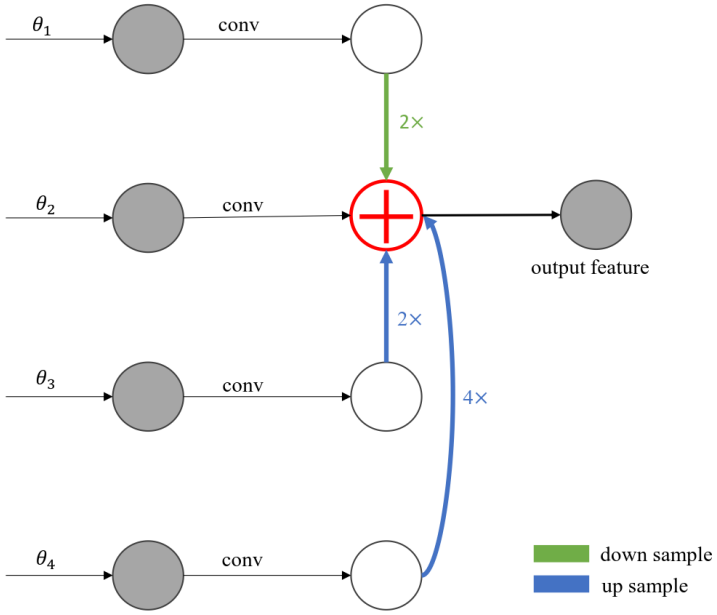

Figure 4. Diagram of feature fusion for branch θ_2.

In addition, to preserve the semantic features of the main subject and eliminate irrelevant features, while considering computational resources and better adaptability, we propose a soft-threshold dual-attention module within the skip connection structure of HPUNet to process the features, as shown in Figure 5. To ensure flexibility between each branch, the threshold selection remains relatively independent between different branches. First, we calculate the absolute value of the intermediate feature map. Then, we multiply the channel feature's average pooling and the spatial feature's average calculation along the channel dimension by the channel attention weight coefficient and the spatial attention weight coefficient, respectively, to obtain the target thresholds. Finally, the feature map is filtered using these two thresholds. The mathematical description of this process is as follows:

$$F_{\text{fused}} = (F_{\text{input}} \otimes (C_a \otimes C_b)) \otimes (P_a \otimes P_b) \tag{4}$$

where C_a represents the feature description after average pooling of the channel features, P_a represents the feature description of spatial features averaged across channels, and C_b and P_b represent the weight coefficients for the channel features and spatial features, respectively.

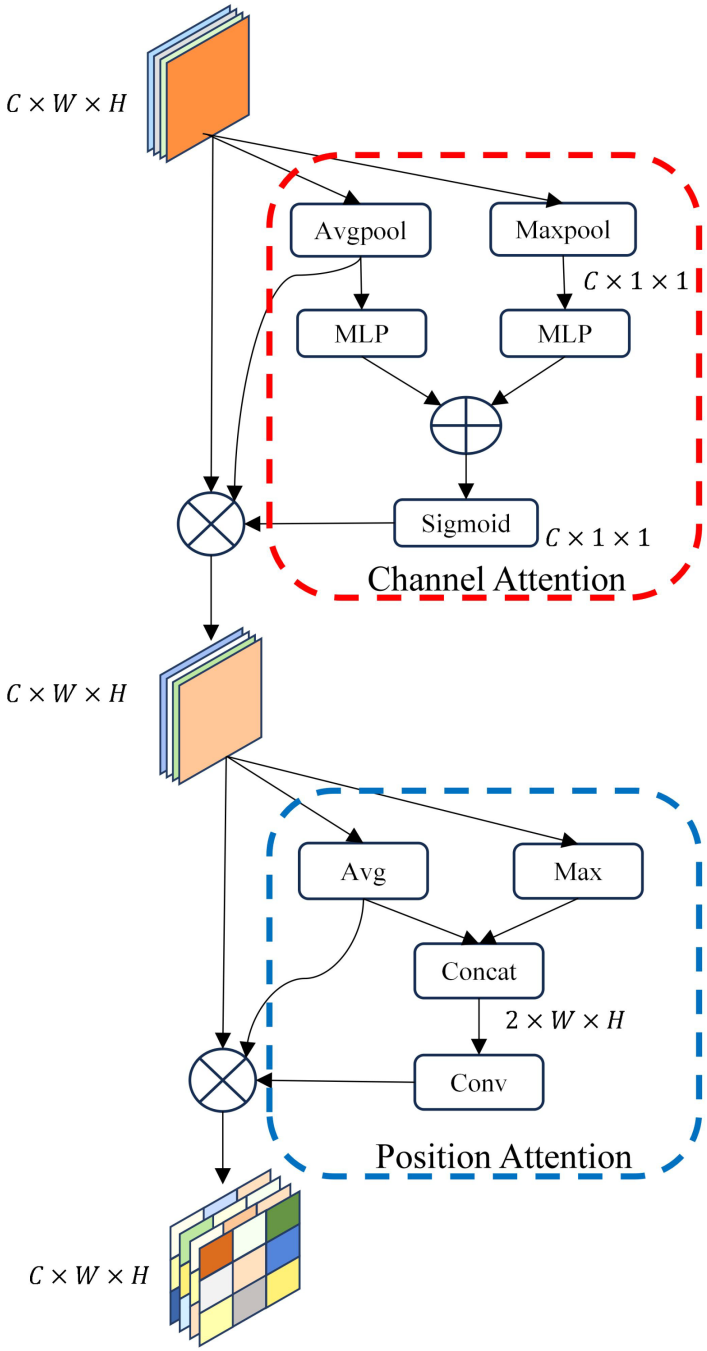

Figure 5. Soft-threshold dual-attention module.

3.4. Loss Function

In the use of loss functions in this paper, three main factors are considered. Firstly, for the task of fake image detection, we perform the detection of the forgery attributes of the image at each level; therefore, we express the optimization goal of each branch as

$$\mathcal{L}_{det}^{b}(X) = -\frac{1}{N}\sum y_i \cdot log(p(y_b|X)) \tag{5}$$

where N represents the number of samples, y_i represents the true class label, and $p(y_b|X)$ represents the predicted probability of the class on branch b.

Secondly, for the task of fake image localization, we also perform localization of the forged regions at each level. Therefore, we use the binary cross-entropy loss function to represent the optimization goal of each branch as

$$\mathcal{L}_{loc}^{b}(X) = -\frac{1}{H^b W^b} \sum_{i=1}^{H^b} \sum_{j=1}^{W^b} (y_{i,j}^{b} \cdot log(p_{i,j}^{b}(X))) \tag{6}$$

where H^b and W^b represent the length and width of the image on branch θ_b, $y_{i,j}^{b}$ represents the true label of the mask position (i, j) on branch θ_b, and $p_{i,j}^{b}(X)$ represents the predicted probability for position (i, j) of the input image on branch θ_b. To accommodate the hierarchical network structure of HPUNet, classification loss and localization loss are both calculated for branches at different resolutions.

In addition, considering that segmentation tasks often tend to misjudge false pixels and true pixels at boundary positions, we also introduce an edge loss $\mathfrak{L}_{edge}$ at the highest resolution layer to constrain the segmentation. The edge loss $\mathfrak{L}_{edge}$ is expressed as

$$\mathcal{L}_{edge} = mean(S_x(M) - S_x(m)) + mean(S_y(M) - S_y(m)) \tag{7}$$

Here, S_x and S_y represent the Sobel convolution kernels in the x and y directions, respectively. (M) stands for the ground truth mask image, while (m) represents the predicted mask image.

In summary, combining Equations (5)–(7), we use a combination of these three losses as the total loss:

$$\mathcal{L}_{\text{total}} = \sum_{b=1}^{4} \mathcal{L}_{det}^{b}(X) + \sum_{b=1}^{4} \mathcal{L}_{loc}^{b}(X) + \mathcal{L}_{edge} \tag{8}$$

4. Experiments

4.1. Dataset and Experimental Settings

Dataset. To verify the advanced performance of our method, we expand the AI-generated fake image dataset, DA-HFNet dataset, created in our previous work. We incorporate a method of manually edited images. This portion of the fake image dataset consists of the MICC-F2000 dataset, the CoMoFod dataset, and approximately 2500 hand-manipulated images produced by us, totaling about 3400 fake images. For ease of distinction, we name this dataset the AITfake dataset in this paper. Its relevant composition structure is shown in Table 1, encompassing a total of 8 types of fake images, with each category containing 3K images. For real images, they are randomly selected from COCO2017 and ImageNet. We validated our method on four datasets: the CoCoGLIDE dataset, the HIFI-IFDL dataset, the GenImage dataset, and the Casia dataset. Specifically, CoCoGLIDE is a small image editing dataset created using the GLIDE model, which contains a total of 512 manipulated images. The HIFI-IFDL dataset is a comprehensive dataset that includes multiple image manipulation methods, and we selected only 500 images for each method. We randomly selected 2000 images from the GenImage dataset. The Casia dataset contains 900 manually tampered manipulated images.

Evaluation Metrics. Drawing on the experiment of previous work, we select accuracy score (ACC) and F1 score (F1) as the evaluation metrics for the experiments.

Experimental Basic Setup. HPUNet is implemented on PyTorch and trained on four NVIDIA 3090 GPUs. The input image size is set to 256 × 256. The initial learning rate is set to 0.0002 and decays periodically to 1×10^{-8}. The batch size is set to 16, and the number of training epochs is set to 50. The initial threshold value in the DAM module was set to 0.5 and is updated iteratively as the training progresses. We apply common data augmentation methods to the training process, including rotation and flipping. The split ratio between the training set and the test set is 9:1.

Table 1. Composition of the AITfake dataset. ✓and × represent whether the item is involved. Copy-Move is a manual forgery method.

Method	Model		Forgery Region		Guidance	Num
	GAN	Diffusion	Full	Partial		
BigGAN [43]	✓	×	✓	×	image	3k
DDPM [7]	×	✓	✓	×	image	3k
FuseDream [44]	✓	×	✓	×	text	3k
GLIDE [22]	×	✓	✓	×	text	3k
Inpaint Anything [45]	×	✓	×	✓	text	3k
Paint by Example [25]	×	✓	×	✓	image	3k
StyleCLIP [26]	✓	×	×	✓	text	3k
Copy-Move	-	-	-	-	-	3k

4.2. Fake Image Detection

Firstly, HPUNet is subjected to comparative experiments with the baseline methods on the task of fake image detection, and the results of the different methods on fake image detection are reported in Table 2. Specifically, from the first to the third row in Table 2, it can be observed that the detection results using pre-trained detectors on the AITfake dataset are generally poor, which is directly related to the feature expression ability learned by the detectors from the traditional fake images. Among them, Trufor is an excellent fake image detector, while PSCC-Net and HIFI-IFDL are representative methods based on hierarchical networks. After unifying the training datasets, from the fourth row to the last row, it can be observed that, compared with the baseline methods, HPUNet exhibits the best ACC and F1 scores on both GAN-based and DM-based fake images. On the manually manipulated images, the ACC score is 0.87% higher than the second-best, and the F1 score is also highly competitive. Overall, our method outperforms the second-best in both the ACC and F1 scores, with average improvements of 1.4% and 0.43%, respectively. We believe this is because HPUNet is able to learn image features with stronger representation capabilities. We present a comparison diagram in Figure 6, which shows the dimensionality reduction regarding the validation set data using the t-SNE approach for different methods. It can be observed that HPUNet is better able to distinguish different categories of fake images compared to the other methods. Meanwhile, in Figure 7, we compare HPUNet with our previous work, DA-HFNet, and it can be seen that HPUNet has improved in both ACC and F1 metrics, demonstrating that our improvement measures are effective in enhancing the detection performance of the model.

Table 2. Experimental results of fake image detection. "*" indicates that we applied the pre-trained model released by the authors. Models without "*" were trained on the same training set. [**Bold**: best result; underline: second-best result].

Method	GANs		DMs		Artificial		AVG	
	ACC (%)	F1 (%)	ACC (%)	F1 (%)	ACC (%)	F1 (%)	ACC (%)	F1 (%)
Trufor * [10]	65.42	62.47	62.15	61.27	59.75	52.84	62.44	58.86
PSCC-Net * [39]	47.24	51.36	46.31	53.28	47.61	52.16	47.05	52.27
HIFI-IFDL * [15]	65.37	60.19	62.31	56.18	70.24	64.29	65.97	60.22
CNN-det. [46]	81.29	77.42	79.38	65.77	82.56	71.48	81.08	71.56
ResNet50 [47]	84.50	70.19	77.69	65.82	78.49	70.93	80.23	68.98
Mantra-Net [37]	82.95	77.27	84.52	78.94	81.39	77.38	82.95	77.86
DIRE [35]	71.25	60.08	90.59	89.64	62.18	58.39	74.67	69.37
PSCC-Net [39]	94.16	97.26	92.38	97.41	92.67	96.42	93.07	97.03
HIFI-IFDL [15]	95.67	97.13	95.19	96.28	96.37	97.19	95.74	96.87
DA-HFNet [19]	98.14	97.09	97.92	97.61	98.42	**98.10**	98.16	97.60
HPUNet (ours)	**99.69**	**98.54**	**99.70**	**98.13**	**99.29**	97.43	**99.56**	**98.03**

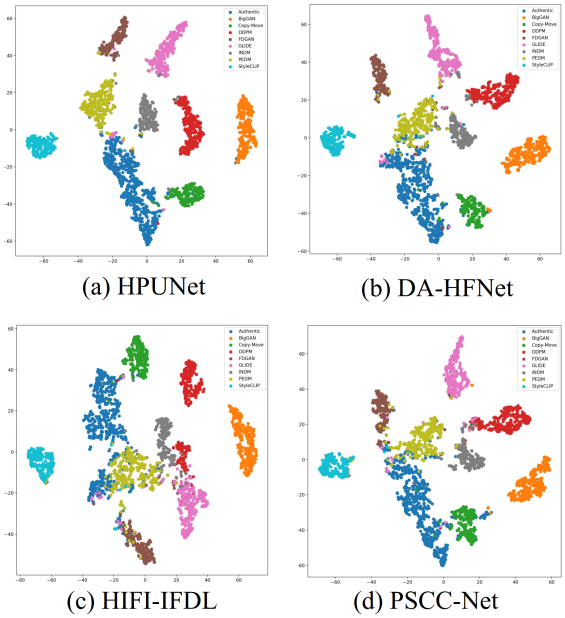

Figure 6. t-SNE visual comparison.

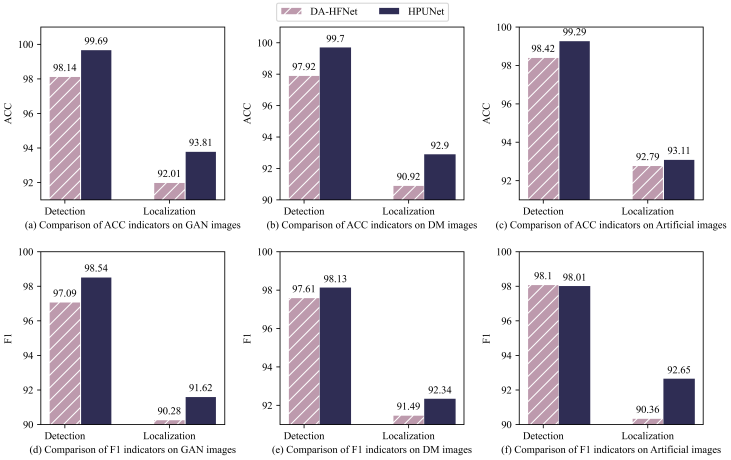

Figure 7. Comparison picture of HPUNet and DA-HFNet.

4.3. Fake Image Localization

Subsequently, we conducted comparative experiments with the baseline methods on the task of fake image localization and report the performance of the different methods on fake image localization in Table 3. Our baseline model follows the methods used in fake image detection. As can be observed from the first to the third row in Table 3, the pre-trained methods do not work well on our dataset. This is because the segmentation objects executed by the pre-trained PSCC-Net and HIFI-IFDL mainly come from manually edited fake images, which exhibit significant artifact differences from AI-edited fake images. Then, we used our AITfake dataset to train the PSCC and HIFI-IFDL detectors. As can be observed from the sixth and seventh rows, they both showed significant improvements in their ACC and F1 scores. From the fourth row to the last row, it can be observed that, under unified training data conditions, HPUNet achieved the best performance in both the ACC and F1 scores. The average ACC score is 1.36% higher than the second-best, and the average F1 score is 2.03% higher. We present some localization results on large-scale

fake images in Figure 8 and on small-scale fake images in Figure 9. It can be observed that HPUNet exhibits excellent localization capabilities in fake images of different scales.

Table 3. Fake image localization experimental results. "*" indicates that we applied the pre-trained model released by the authors. Models without "*" were trained on the same training set. [**Bold**: best result; underline: second-best result].

Method	GANs		DMs		Artificial		AVG	
	ACC (%)	F1 (%)	ACC (%)	F1 (%)	ACC (%)	F1 (%)	ACC (%)	F1 (%)
Trufor * [10]	77.18	78.49	74.27	76.34	68.23	61.71	73.23	72.18
PSCC-Net * [39]	55.39	51.67	58.24	54.89	57.42	61.58	57.02	56.05
HIFI-IFDL * [15]	65.27	42.88	54.36	51.94	57.81	49.82	59.15	48.21
Unet [18]	65.49	59.91	68.35	58.32	59.48	50.16	64.44	56.13
Mantra-Net [37]	84.34	79.61	81.92	74.38	86.94	80.11	84.40	78.03
PSCC-Net [39]	86.15	62.37	87.04	59.67	88.45	67.89	87.21	63.31
HIFI-IFDL [15]	89.94	88.26	87.22	86.39	90.18	91.06	88.57	88.57
DA-HFNet [19]	92.01	90.28	90.92	91.49	92.79	90.36	91.91	90.71
HPUNet (ours)	**93.81**	**91.62**	**92.90**	**92.34**	**93.11**	**92.65**	**93.27**	**92.20**

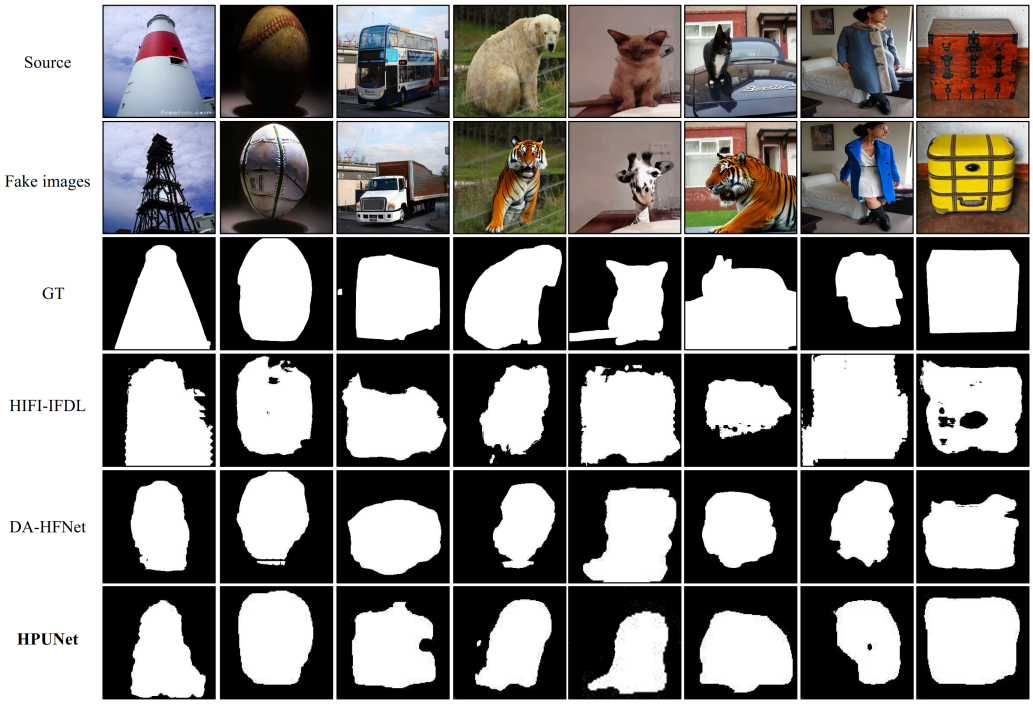

Figure 8. Comparison of large-scale fake image localization results.

4.4. Cross-Dataset Validation

To validate the effectiveness of HPUNet, we conducted cross-dataset validation experiments. The training dataset was composed of the AITfake dataset, while the validation datasets were derived from four public datasets: CoCoGLIDE [10], HIFI-IFDL [15], GenImage, and Casia. We report the experimental results of the cross-dataset validation in Table 4. As can be seen from Table 4, in the cross-dataset validation tasks, HPUNet performed at the first or second level on all the datasets, still demonstrating a clear advantage over the baseline methods.

Table 4. Cross-dataset validation results. [**Bold**: best result; underline: second-best result].

Method	CoCoGLIDE				HIFI-IFDL Dataset				GenImage		Casia	
	Detection		Localization		Detection		Localization		Detection		Detection	
	ACC (%)	F1 (%)	ACC (%)	F1 (%)	ACC (%)	F1 (%)	ACC (%)	F1 (%)	ACC (%)	F1 (%)	ACC (%)	F1 (%)
PSCC-Net	44.26	39.57	58.76	43.59	74.21	68.55	68.19	71.26	62.84	57.23	75.34	70.29
HIFI-IFDL	48.75	40.33	66.89	50.17	**83.84**	79.91	73.59	79.48	72.58	69.81	74.38	**79.54**
DA-HFNet	52.18	**52.36**	72.15	55.47	81.67	**82.36**	71.10	77.68	73.05	69.54	72.82	69.93
HPUNet (ours)	**54.19**	48.97	**78.57**	**61.09**	82.97	81.85	**74.99**	**81.53**	**80.28**	**74.62**	**77.16**	75.80

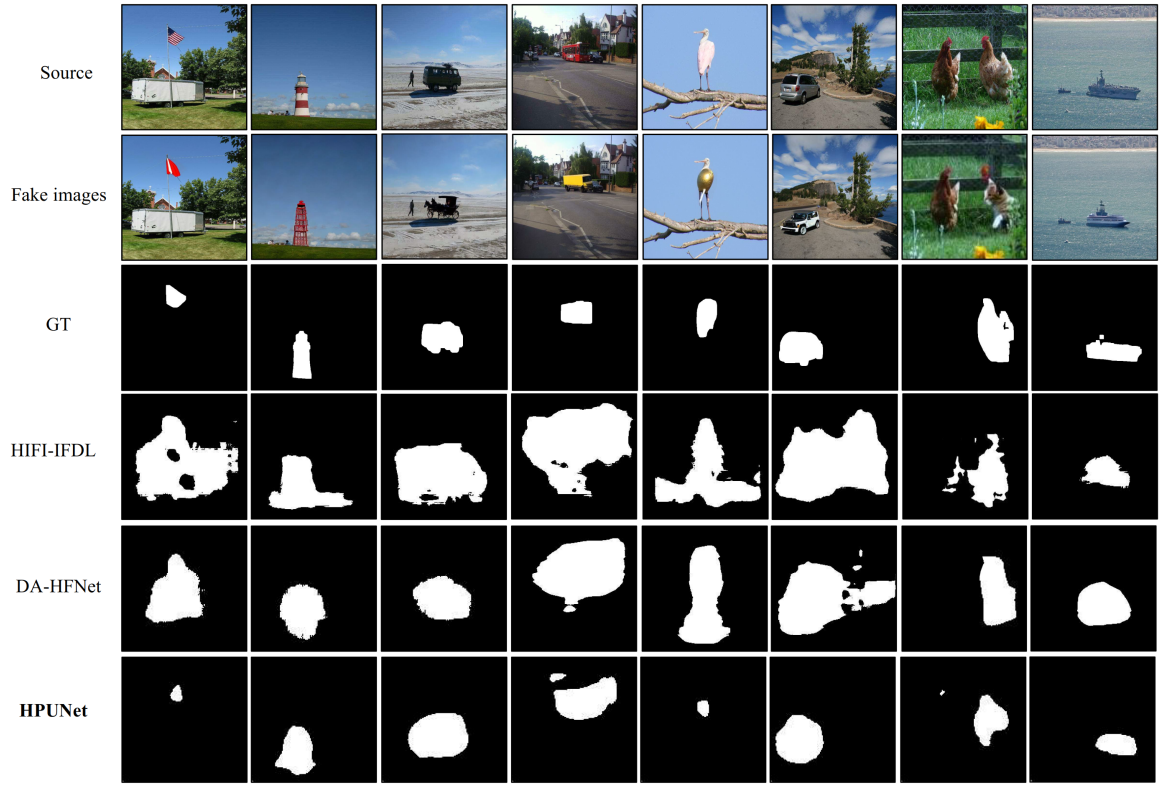

Figure 9. Comparison of small-scale fake image localization results.

4.5. Ablation Experiment

To validate the effectiveness of the context feature enhancement module, we conducted an ablation study. First, we performed an ablation on the extracted image forgery features and report the experimental results in Table 5. We individually ablated the features used in the method, and, as observed in rows 1–3 of Table 5, eliminating the image features from any one branch results in varying degrees of performance degradation. Therefore, we believe that the features from each branch contribute positively to the overall performance of the method. In addition, we conducted experiments using the ResNet model directly as our feature extraction network. As can be observed from the fourth and fifth rows, using a deep network directly for feature extraction results in a significant decrease in the model's detection and localization performance. Therefore, we believe that using a deep network directly for feature extraction is not beneficial for detection and localization tasks.

In the subsequent experiments, we investigated the impact of the edge loss on the experimental results and reported the corresponding findings in Table 6. As shown in Table 6, when the edge loss is removed, the model's ACC and F1 scores in the forgery category detection task drop by 4.98% and 4.31%, respectively, while the ACC and F1 scores in the forgery region localization task decrease by 2.99% and 2.71%, respectively. This demonstrates the positive influence of the edge loss we implemented in the experiments.

Table 5. Ablation experiment results on image feature extraction. [**Bold**: best result; ✓and × represent whether to use the item]

Image Features			Detection		Localization	
RGB	**Noise**	**Frequency**	**ACC (%)**	**F1 (%)**	**ACC (%)**	**F1 (%)**
✓	✓		93.51	92.18	87.64	85.27
✓		✓	92.67	91.28	83.95	80.39
	✓	✓	95.81	93.47	89.16	82.83
×	×	×	94.59	93.18	82.91	81.49
✓	✓	✓	**99.56**	**98.03**	**93.27**	**92.20**

Table 6. Ablation experiment results on the impact of edge loss on the model's detection performance. [**Bold**: best result].

Method	**Detection**		**Localization**	
	ACC (%)	**F1 (%)**	**ACC (%)**	**F1 (%)**
No edgeloss	94.58	93.72	90.28	89.49
HPUNet	**99.56**	**98.03**	**93.27**	**92.20**

We also conducted another set of ablation experiments. We investigated the impact of the proposed dual-branch feature fusion module and soft-threshold attention module on the model's performance and report the experimental results in Table 7. As observed in Table 7, the experiments without the dual-branch feature fusion module and soft-threshold attention module exhibited varying degrees of performance degradation in both the forgery image detection and forgery image localization tasks. This validates that the attention modules we proposed are beneficial for the experimental tasks.

Table 7. Ablation experiment results on the impact of attention mechanisms on the model's detection performance. [**Bold**: best result].

Attention Modules		Detection		Localization	
DAM	**t-CBAM**	**ACC (%)**	**F1 (%)**	**ACC (%)**	**F1 (%)**
✓		84.76	82.19	88.63	82.41
	✓	92.94	89.96	90.47	87.52
✓	✓	**99.56**	**98.03**	**93.27**	**92.20**

Finally, we conducted experiments to verify the interrelationship between multi-level detection and localization. The relevant experimental results are reported in Table 8. From the first and second rows, it can be observed that, when we only use detection or localization as the training objective, there is a significant decrease in the performance for that task. Among them, the model performance decreases the most when only localization is used as the training objective. Subsequently, we treated both tasks as training objectives and conducted ablation experiments on the hierarchical structure. We recorded the corresponding results for the detection task. From the third row to the last row, it can be noted that, as we reduce the number of branches in the model, the model performance continues to decline. This result confirms that our hierarchical detection and localization indeed improve the model's detection results at each level.

Table 8. Interaction between multi-level detection and localization. When the "training task" is detection, the testing task is localization. When the "training task" is localization, the testing task is detection. When the "training task" is all tasks, the testing task is detection.

Training Task	Branches	GANs		DMs		Artificial	
		ACC (%)	F1 (%)	ACC (%)	F1 (%)	ACC (%)	F1 (%)
Detection	4 branches	94.51	93.47	95.39	94.67	96.28	94.19
Localization	4 branches	91.83	90.27	89.67	90.15	92.18	91.12
All tasks	1 branch	95.74	95.13	94.55	94.31	94.28	93.94
	2 branches	97.92	96.75	96.86	96.48	97.12	95.24
	3 branches	98.59	97.46	97.21	96.84	98.25	96.18
	4 branches	99.69	98.54	99.70	98.13	99.29	97.43

5. Conclusions

In this paper, we propose a progressive UNet-based network for the detection and localization of AI-generated forged images, achieving further improvements compared to the baseline methods. This approach utilizes spatial domain image noise features and RGB features, combined with frequency domain features, to capture richer forgery traces. Firstly, the image features are extracted through a context feature enhancement module and fused in a dual-branch attention fusion module, incorporating multiple types of image features. Then, the detection results are used as dynamic thresholds to enhance the task-relevant features and suppress the task-irrelevant features. Subsequently, a multi-branch feature interaction network is employed to enable information exchange between features of different resolutions, addressing the limited field of view of the feature maps caused by inconsistent forgery image sizes. Finally, in a hierarchical network structure, the decoder correlates feature maps of different resolutions, and a soft-threshold dual-attention feature enhancement mechanism is introduced in the skip connections to preserve the main features. The model progressively improves the higher-level results under the guidance of the lower-level results in a hierarchical manner.

We have conducted extensive experiments to demonstrate the effectiveness of our method for detecting and locating AI-generated fake images. We uniformly evaluated HPUNet and the baseline model on the AITfake dataset, and our method achieved the best scores for both the fake image detection and localization tasks. Subsequently, we conducted cross-dataset validation experiments on four open datasets: CoCoGLIDE, HIFI-IFDL, GenImage, and Casia. The experimental results show that HPUNet has better stability than the baseline model. We also conducted sufficient ablation experiments to verify the effectiveness of HPUNet from multiple perspectives.

Author Contributions: Conceptualization and Methodology, Y.L. and X.L.; Data Collection, Y.L.; Model Building, Y.L. and S.H.; Experiment, Data Analysis, and Writing—Original Draft Preparation, Y.L.; Writing—Review and Editing, J.Z., S.L. and J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Laboratory of Big Data and Decision Making of National University of Defense Technology.

Data Availability Statement: The code and dataset will be finalized and made publicly available online upon acceptance of the paper.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; Irani, M. Imagic: Text-based real image editing with diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 6007–6017.

2. Manukyan, H.; Sargsyan, A.; Atanyan, B.; Wang, Z.; Navasardyan, S.; Shi, H. Hd-painter: High-resolution and prompt-faithful text-guided image inpainting with diffusion models. *arXiv* **2023**, arXiv:2312.14091.
3. Bar-Tal, O.; Yariv, L.; Lipman, Y.; Dekel, T. Multidiffusion: Fusing diffusion paths for controlled image generation. In Proceedings of the ICML'23: International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023.
4. Kingma, D.P.; Welling, M. Auto-encoding variational bayes. *arXiv* **2013**, arXiv:1312.6114.
5. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*.
6. Sohl-Dickstein, J.; Weiss, E.; Maheswaranathan, N.; Ganguli, S. Deep unsupervised learning using nonequilibrium thermodynamics. In Proceedings of the International Conference on Machine Learning, PMLR, Lille, France, 6–11 July 2015; pp. 2256–2265.
7. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
8. Zhang, Q.; Tao, M.; Chen, Y. gddim: Generalized denoising diffusion implicit models. *arXiv* **2022**, arXiv:2206.05564.
9. Cozzolino, D.; Verdoliva, L. Noiseprint: A cnn-based camera model fingerprint. *IEEE Trans. Inf. Forensics Secur.* **2019**, *15*, 144–159. [CrossRef]
10. Guillaro, F.; Cozzolino, D.; Sud, A.; Dufour, N.; Verdoliva, L. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 20606–20615.
11. Corvi, R.; Cozzolino, D.; Zingarini, G.; Poggi, G.; Nagano, K.; Verdoliva, L. On the detection of synthetic images generated by diffusion models. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: New York, NY, USA, 2023; pp. 1–5.
12. Huang, Y.; Bian, S.; Li, H.; Wang, C.; Li, K. Ds-unet: A dual streams unet for refined image forgery localization. *Inf. Sci.* **2022**, *610*, 73–89. [CrossRef]
13. Xi, Z.; Huang, W.; Wei, K.; Luo, W.; Zheng, P. Ai-generated image detection using a cross-attention enhanced dual-stream network. *arXiv* **2023**, arXiv:2306.07005.
14. Sha, Z.; Li, Z.; Yu, N.; Zhang, Y. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, Copenhagen, Denmark, 26–30 November 2023; pp. 3418–3432.
15. Guo, X.; Liu, X.; Ren, Z.; Grosz, S.; Masi, I.; Liu, X. Hierarchical fine-grained image forgery detection and localization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 14–24 June 2023; pp. 3155–3165.
16. Niloy, F.F.; Bhaumik, K.K.; Woo, S.S. Cfl-net: Image forgery localization using contrastive learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 4642–4651.
17. Guo, M.-H.; Liu, Z.-N.; Mu, T.-J.; Hu, S.-M. Beyond self-attention: External attention using two linear layers for visual tasks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 5436–5447. [CrossRef]
18. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, Germany, 5–9 October 2015; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241.
19. Liu, Y.; Li, X.; Zhang, J.; Hu, S.; Lei, J. Da-hfnet: Progressive Fine-Grained Forgery Image Detection and Localization Based on Dual Attention. 2024. Available online: https://api.semanticscholar.org/CorpusID:270214687 (accessed on 5 June 2024).
20. Zhu, J.-Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
21. Sauer, A.; Karras, T.; Laine, S.; Geiger, A.; Aila, T. Stylegan-t: Unlocking the power of gans for fast large-scale text-to-image synthesis. *arXiv* **2023**, arXiv:2301.09515.
22. Nichol, A.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; McGrew, B.; Sutskever, I.; Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv* **2021**, arXiv:2112.10741.
23. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695.
24. Pan, X.; Tewari, A.; Leimkühler, T.; Liu, L.; Meka, A.; Theobalt, C. Drag your gan: Interactive point-based manipulation on the generative image manifold. In Proceedings of the ACM SIGGRAPH 2023 Conference Proceedings, Los Angeles, CA, USA, 6–10 August 2023; pp. 1–11.
25. Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; Wen, F. Paint by example: Exemplar-based image editing with diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 18381–18391.
26. Patashnik, O.; Wu, Z.; Shechtman, E.; Cohen-Or, D.; Lischinski, D. Styleclip: Text-driven manipulation of stylegan imagery. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2085–2094.
27. Li, H.; Luo, W.; Qiu, X.; Huang, J. Image forgery localization via integrating tampering possibility maps. *IEEE Trans. Inf. Forensics Secur.* **2017**, *12*, 1240–1252. [CrossRef]

28. Arshed, M.A.; Alwadain, A.; Ali, R.F.; Mumtaz, S.; Ibrahim, M.; Muneer, A. Unmasking deception: Empowering deepfake detection with vision transformer network. *Mathematics* **2023**, 11, 3710. [CrossRef]
29. Ojha, U.; Li, Y.; Lee, Y.J. Towards universal fake image detectors that generalize across generative models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 24480–24489.
30. Ramirez-Rodriguez, A.E.; Arevalo-Ancona, R.E.; Perez-Meana, H.; Cedillo-Hernandez, M.; Nakano-Miyatake, M. Aismsnet: Advanced image splicing manipulation identification based on siamese networks. *Appl. Sci.* **2024**, *14*, 5545. [CrossRef]
31. Wan, D.; Cai, M.; Peng, S.; Qin, W.; Li, L. Deepfake detection algorithm based on dual-branch data augmentation and modified attention mechanism. *Appl. Sci.* **2023**, *13*, 8313. [CrossRef]
32. Epstein, D.C.; Jain, I.; Wang, O.; Zhang, R. Online detection of ai-generated images. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Paris, France, 2–6 October 2023; pp. 382–392.
33. Verdoliva, L. Media forensics and deepfakes: An overview. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 910–932. [CrossRef]
34. Wu, H.; Zhou, J.; Zhang, S. Generalizable synthetic image detection via language-guided contrastive learning. *arXiv* **2023**, arXiv:2305.13800.
35. Wang, Z.; Bao, J.; Zhou, W.; Wang, W.; Hu, H.; Chen, H.; Li, H. Dire for diffusion-generated image detection. *arXiv* **2023**, arXiv:2303.09295.
36. Zhong, N.; Xu, Y.; Qian, Z.; Zhang, X. Rich and poor texture contrast: A simple yet effective approach for ai-generated image detection. *arXiv* **2023**, arXiv:2311.12397.
37. Wu, Y.; AbdAlmageed, W.; Natarajan, P. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9543–9552.
38. 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.* **2022**, *45*, 3539–3553. [CrossRef]
39. 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.* **2022**, *32*, 7505–7517. [CrossRef]
40. Zhang, J.; Tohidypour, H.; Wang, Y.; Nasiopoulos, P. Shallow-and deep-fake image manipulation localization using deep learning. In Proceedings of the 2023 International Conference on Computing, Networking and Communications (ICNC), Honolulu, HI, USA, 20–22 February 2023; pp. 468–472.
41. Zhou, Y.; Wang, H.; Zeng, Q.; Zhang, R.; Meng, S. Exploring weakly-supervised image manipulation localization with tampering edge-based class activation map. *Expert Syst. Appl.* **2024**, *249*, 123501. [CrossRef]
42. Liu, Q.; Li, H.; Liu, Z. Image forgery localization based on fully convolutional network with noise feature. *Multimed. Tools Appl.* **2022**, *81*, 17919–17935. [CrossRef]
43. Brock, A.; Donahue, J.; Simonyan, K. Large scale GAN training for high fidelity natural image synthesis. In Proceedings of the 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, 6–9 May 2019. Available online: https://openreview.net/forum?id=B1xsqj09Fm (accessed on 5 June 2024).
44. Liu, X.; Gong, C.; Wu, L.; Zhang, S.; Su, H.; Liu, Q. Fusedream: Training-free text-to-image generation with improved clip+ gan space optimization. *arXiv* **2021**, arXiv:2112.01573.
45. Yu, T.; Feng, R.; Feng, R.; Liu, J.; Jin, X.; Zeng, W.; Chen, Z. Inpaint anything: Segment anything meets image inpainting. *arXiv* **2023**, arXiv:2304.06790.
46. Wang, S.-Y.; Wang, O.; Zhang, R.; Owens, A.; Efros, A.A. Cnn-generated images are surprisingly easy to spot... for now. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8695–8704.
47. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.