Contrastive Residual Network with Sliding-Window Fusion for Precise Deepfake Detection and Localization

Yilin Wang[1] , **Zunlei Feng**[2] , **Yijun Bei**[3]
[123]Zhejiang University
{ylwang2024, zunleifeng, beiyj}@zju.edu.cn

Abstract

Most current deepfake image detection methods focus on distinguishing real from fake images but lack the ability to accurately localize forged regions, limiting their interpretability and applicability in scenarios such as content moderation and forensic analysis. Existing approaches primarily rely on image features like deep or frequency-domain representations, often ignoring the intrinsic patterns introduced by generator architectures. However, these features may retain semantic content from the original image, potentially hindering detection accuracy and generalization. We observe that upsampling in generators induces strong correlations among neighboring pixels, which are further amplified by residual connections. Leveraging this insight, we propose a **C**ontrastive **R**esidual **F**orgery **D**etection Network (**CRFD-Net**). It employs a tailored residual structure and a local contrastive enhancement module to highlight abnormal pixel correlations in forged areas. A U-Net-inspired decoder enables spatial localization, while a sliding window-based fusion strategy further refines the prediction of forged regions.

1 Introduction

With the rapid advancement of AIGC technologies [Cao *et al.*, 2023], particularly diffusion-based generative models, applications in image and video synthesis have grown significantly. While these innovations bring convenience, they also pose serious threats. AIGC enables the easy creation of highly realistic fake content, facilitating disinformation, malicious tampering, and identity fraud—raising concerns over economic losses and societal trust [Zhao *et al.*, 2023].

Current forged data detection methods mainly fall into two categories: feature-based and reconstruction error-based. Feature-based methods leverage generative model characteristics, such as [Ma *et al.*, 2023], which exploits the reverse process and denoising error in diffusion models, and [Sha *et al.*, 2023], which performs model attribution based on unique image features. Reconstruction-based approaches, like [Wang *et al.*, 2023], compare input and reconstructed images to detect forgeries, showing robustness even under unknown models. However, such methods are sensitive to the nature of reconstruction targets; for example, a model trained on animals may misclassify real plant images as fake due to domain mismatch.

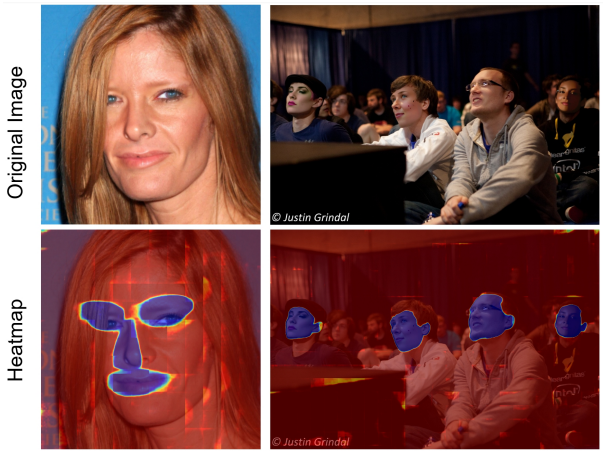

Figure 1: Heatmaps generated by our model.

Recent deep learning-based methods have made notable progress in localizing forged regions [Hu *et al.*, 2020; Park *et al.*, 2018; Wu *et al.*, 2019a; Yang *et al.*, 2020; Zhou *et al.*, 2018]. However, they often struggle in real-world scenarios due to the wide variety of forgery techniques, including generation, insertion, deletion, cloning, and replacement. Existing approaches typically use shared architectures and project diverse forgeries into a unified feature space, neglecting forgery pattern diversity and thus limiting generalization.

Although current methods have made encouraging progress in forged image detection and localization, they still face challenges such as limited generalization and low localization accuracy under complex and diverse forgery patterns. In particular, the upsampling operations commonly found in generative models introduce abnormal local correlations—phenomena that have not been fully modeled or leveraged by existing methods.

To address the limitations of existing methods, we propose a **C**ontrastive **R**esidual **F**orgery **D**etection Network (**CRFD-Net**), designed to capture abnormal local correlations in generated images. A residual structure and a local contrastive en-

hancement module are introduced to amplify pixel similarity differences in forged regions, improving sensitivity to forgery traces. We adopt a U-Net-inspired decoder and a sliding window fusion strategy to enhance localization across varying scales and resolutions. Our main contributions are:

- A local contrastive mechanism is proposed to capture subtle yet consistent pixel correlations in forged regions, enhancing detection robustness.
- A sliding window fusion strategy is designed to reduce boundary artifacts and improve localization of small or dispersed forgeries.
- Our method achieves top performance in the IJCAI Forgery Detection Challenge, demonstrating strong generalization and practical value.

2 Related Work

2.1 Image Forgery Detection

Early research primarily extracted spatial clues such as edge blending [Li *et al.*, 2020], color [McCloskey and Albright, 2018], and saturation [McCloskey and Albright, 2019]. However, as image generation technologies evolve, traditional spatial features have become less effective. Compression during image transmission further degrades quality, making forgery traces harder to detect. To address this, researchers turned to frequency-domain methods, introducing approaches based on scale [Wang *et al.*, 2022b], frequency bands [Li *et al.*, 2021], and adaptive feature extraction [Qian *et al.*, 2020]. While effective on compressed images, these methods still struggle with unseen forgery techniques.

Most existing methods target GAN-generated images. Although Corvi et al. [Corvi *et al.*, 2023a; Corvi *et al.*, 2023b; Ricker *et al.*, 2022] observed spectral artifacts in diffusion-generated images, detection methods based on such artifacts remain limited in effectiveness. To address this, recent approaches shift focus to features from the diffusion process itself. For example, DIRE [Sha *et al.*, 2023] and SeDIE [Wang *et al.*, 2023] leverage reconstruction errors, while LaRE2 [Luo *et al.*, 2024] and AEROBLADE [Ricker *et al.*, 2024] explore errors in latent space. DRCT [Chen *et al.*, 2024] employs four reconstruction types and uses contrastive loss to train a classifier.

However, these methods heavily rely on pre-trained reconstruction models, making them vulnerable to the limitations of the training data. For instance, if the reconstruction model is trained on a specific category (e.g., cats) but is used to detect forgeries from different categories (e.g., dogs or plants), misclassifications may occur. The core issue lies in using reconstruction error as the primary detection signal, without deeply leveraging the characteristics of the diffusion generation mechanism itself, thereby limiting the understanding of the generative patterns behind forged images.

2.2 Image Forgery Region Localization

Current image forgery localization methods typically detect and localize forgeries by identifying inconsistencies or differences between forged and authentic regions. These methods often rely on feature extractors to capture forgery-related cues, such as RGB noise [Cozzolino and Verdoliva, 2019], [Guillaro *et al.*, 2023], high-frequency features [Kwon *et al.*, 2022; Wang *et al.*, 2022a], or edge artifacts [Dong *et al.*, 2022; Zhou *et al.*, 2020]. For example, [Cozzolino and Verdoliva, 2019] and [Guillaro *et al.*, 2023] extract low-level forgery traces from camera model fingerprints; ManTra-Net [Wu *et al.*, 2019b] uses both BayarConv and SRM as noise extractors to obtain rich features; CAT-Net [Kwon *et al.*, 2022] leverages Discrete Cosine Transform (DCT) coefficients to localize tampered regions; ObjectFormer [Wang *et al.*, 2022a] captures subtle forgery traces by combining high-frequency and RGB features; TruFor [Guillaro *et al.*, 2023] integrates high-level semantic features from RGB images with noise-sensitive fingerprint features for forgery localization.

In addition, many methods utilize edge artifacts for forgery region detection. For instance, GSR-Net [Zhou *et al.*, 2020] introduces edge detection and refinement branches to better recognize boundary artifacts; MVSS-Net [Dong *et al.*, 2022] designs an edge supervision branch that progressively extracts fine-grained boundary information from shallow to deep layers.

However, most existing methods adopt a unified feature extractor to handle all types of forged images, without fully considering the significant visual differences across various forgery patterns. This limits their generalization ability in diverse forgery scenarios. Moreover, these methods often over-rely on global or single-scale features, making it difficult to accurately capture subtle structural anomalies around the boundaries of forged regions, which negatively impacts localization precision.

In light of these challenges, this paper proposes a Contrastive Residual Forgery Detection Network (CRFD-Net), designed to more sensitively perceive and localize forgery traces in images. Starting from the observation that generative models introduce structural defects during the upsampling process — notably, significant pixel correlations in neighborhood regions — we design a residual enhancement module and a local contrast mechanism to effectively amplify and capture these micro-level forgery features.

3 Method

This study proposes a forgery localization method called the **C**ontrastive **R**esidual **F**orgery **D**etection Network (**CRFD-Net**). As shown in the figure, the overall architecture consists of two main functional modules:

- Contrastive Residual Feature Encoder: To address the neighborhood correlation introduced by upsampling operations, a contrastive enhancement module is designed. This module amplifies the anomalous relationships between pixels in forged regions through a local contrastive mechanism and residual connections, enabling the extraction of low-level forgery features from images.
- Symmetric U-Net Decoder: Inspired by the skip connection design in the U-Net architecture, this decoder fuses low-level detailed information with high-level semantic information. It effectively reconstructs the spa-

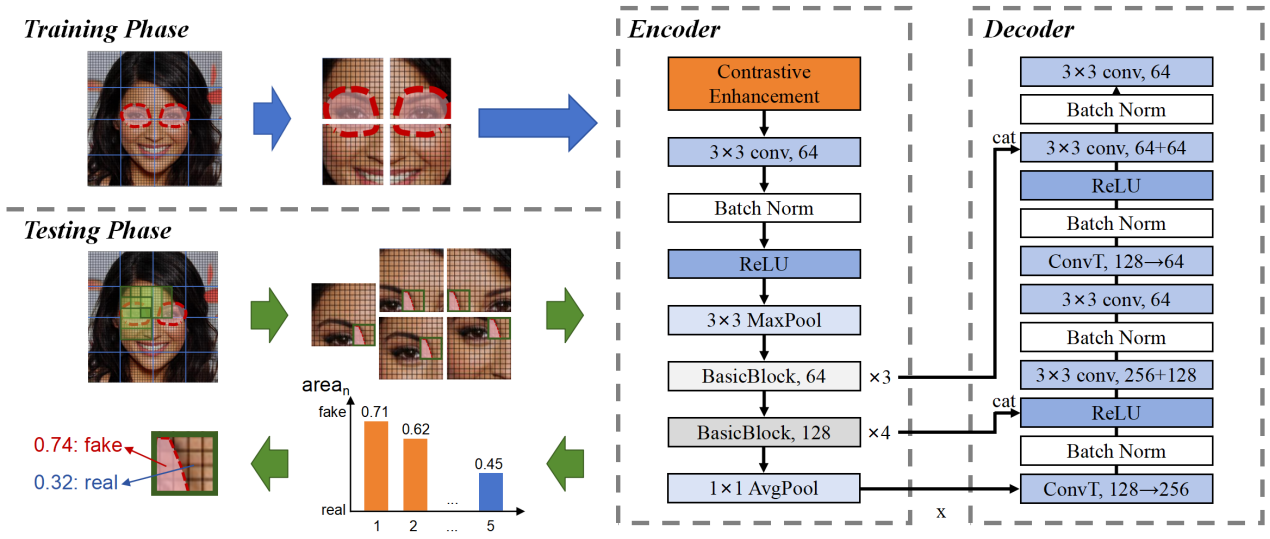

Figure 2: Architecture of the proposed CRFD-Net. The model integrates a feature encoder, a contrastive residual module to enhance anomalous correlations in upsampled regions, and a U-Net-like decoder with sliding window fusion for precise localization of forged areas.

tial distribution of forged regions and generates pixel-level forgery masks.

3.1 Contrastive Residual Feature Encoder

To improve the model's sensitivity to subtle and localized tampering traces in forged images, we design a Contrastive Enhancement Module (CEM) as a preprocessing step. This module extracts high-frequency residual information from the input image through dual interpolation operations:

$$\text{CEM}(x) = x - \text{Upsample}(\text{Downsample}(x)) \quad (1)$$

Both the upsampling and downsampling operations are implemented using nearest-neighbor interpolation, aiming to preserve edge details and local perturbations. Ultimately, the residual signals are scaled to enhance the influence of subtle differences on the downstream network.

Residual Feature Extraction Module The core network adopts the standard ResNet framework, consisting of two groups of residual structure modules—BasicBlock$_1$ and BasicBlock$_2$—each formed by stacking multiple residual units. Within each residual unit, skip connections are employed to implement identity mappings of features, effectively alleviating the vanishing gradient problem and enhancing the model's generalization ability. The number of convolution channels increases progressively across layers to extract higher-order semantic features from images.

Global Semantic Aggregation and Prediction After deep feature extraction, the final output feature map is compressed into a 1×1 global vector through adaptive average pooling. This operation performs spatial averaging of features while preserving deep semantic information across channels.

3.2 Symmetric U-Net Decoder

We designed an improved decoder module aimed at generating high-quality semantic segmentation results through multi-scale feature fusion and progressive upsampling. Inspired by the U-Net architecture, this module incorporates skip connections to retain low-level spatial information, and fuses deep semantic features with shallow spatial details to improve segmentation accuracy. It combines transposed convolutions with bilinear interpolation to gradually restore the spatial resolution of feature maps and minimize information loss.

The decoder takes three feature maps of different scales as input: A high-resolution feature map x_1 from the shallow layer, a medium-resolution feature map x_2 from the middle layer, and a low-resolution feature map x_3 from the deep layer.

First, the deep feature map x_3 is upsampled to match the spatial size of the intermediate feature map x_2, followed by feature fusion:

$$x_3^{'} = \text{ReLU}(\text{BN}(\text{ConvT}(x_3))) \quad (2)$$

$$x_3^{''} = \text{Interpolate}(x_3^{'}, \text{size} = x_2.\text{size}) \quad (3)$$

$$x_{fusion_1} = \text{ReLU}(\text{BN}(\text{Conv}(\text{Concat}(x_3^{''}, x_2)))) \quad (4)$$

Next, the fused feature map x_{fusion_1} is upsampled to match the spatial resolution of the shallow feature map x_1, followed by feature fusion:

$$x_{fusion_1}^{'} = \text{ReLU}(\text{BN}(\text{ConvT}(x_{fusion_1}))) \quad (5)$$

$$x_{fusion_1}^{''} = \text{Interpolate}(x_{fusion_1}^{'}, \text{size} = x_1.\text{size}) \quad (6)$$

$$x_{fusion_2} = \text{ReLU}(\text{BN}(\text{Conv}(\text{Concat}(x_{fusion_1}^{''}, x_1)))) \quad (7)$$

The fused feature map x_{fusion_2} is projected to the category space through a 1×1 convolution, and then upsampled to the target resolution:

$$x_{out} = \text{Conv}_{1\times1}(x_{fusion_2}) \quad (8)$$

$$x_{seg} = \text{Interpolate}(x_{out}, \text{size} = \text{target.size}) \quad (9)$$

3.3 Data Preprocessing

Training Phase During the training phase, the window size is set to $P \times P$, and each slide moves P pixels. Given an input image with height H and width W, the sliding starts from the top-left corner and proceeds with a stride of S, traversing all possible window positions. For each position (x, y), an image patch of size $P \times P$ is extracted. For each extracted image patch $patch(x, y)$, the corresponding binarized mask patch is $mask(x, y)$. The image patch is retained only if the mask patch contains foreground pixels (i.e., pixels with a value of 255). Through the above steps, multiple local regions containing forgery annotations are extracted from the original image, providing sample data for subsequent model training.

Testing Phase In the testing phase, to reconstruct a complete mask image from the predicted image patches, we adopt a sliding window–based overlapping region fusion method. Let the size of the reconstructed image be $H \times W$. We initialize two matrices of the same size: Accumulation map $A \in \mathbb{R}^{H \times W}$: used to store the sum of prediction values at each pixel position; Count map $C \in \mathbb{R}^{H \times W}$: used to record how many times each pixel position has been predicted.

For each predicted patch P_k and its starting coordinates (x_k, y_k) in the original image, we add the predicted values to the corresponding positions in the accumulation map, and increment the values in the count map at the same positions by one:

$$A_{i,j} += P_k(i - y_k, j - x_k), \forall i \in [y_k, y_k + p) \quad (10)$$

$$C_{i,j} += 1, \forall i \in [y_k, y_k + p) \quad (11)$$

where p denotes the size of the image patch. To obtain the average predicted value at each pixel position, compute $M_{i,j} = \frac{A_{i,j}}{C_{i,j}}$

Then, apply thresholding to the values in the average prediction map M to generate the binary mask image B:

$$B_{i,j} = \begin{cases} 255, & \text{if } M_{i,j} > 0.5 \\ 0, & \text{otherwise} \end{cases}$$

Finally, crop B to match the original image size of $H \times W$.

3.4 Real vs. Fake Image Classification

The Binary Cross-Entropy (BCE) Loss is adopted, which is suitable for binary classification tasks. In forgery region detection, it effectively measures the difference between predicted probabilities and ground truth labels. For an image with N pixels, the BCE loss function is defined as:

$$\text{BCE} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (12)$$

where $y_i \in \{0, 1\}$ is the ground truth label for the i-th pixel (0 for background, 1 for foreground), and $p_i \in (0, 1)$ is the predicted probability that the i-th pixel belongs to the foreground. N is the total number of pixels in the image.

To effectively distinguish between real and forged images, we propose a detection strategy based on the proportion of forged regions. This method calculates the proportion of the image detected as forged and compares it with a predefined threshold to determine whether the image is forged.

For each input image, a corresponding binary mask image $M \in \{0, 255\}^{H \times W}$ is first generated by the forgery localization model, where H and W represent the height and width of the image, respectively. In the mask image, a pixel value of 255 indicates that the position is detected as part of a forged region, while 0 indicates it is not.

The proportion p of the forged region in the entire image is calculated as follows:

$$p = \frac{\sum_{i=1}^{H} \mathbb{I}[M_{i,j} = 255]}{H \times W}$$

Here, $\mathbb{I}$ is an indicator function, which takes the value 1 when the condition is true, and 0 otherwise.

To set a reasonable decision threshold θ, we analyze the distribution of the forgery region proportions on the validation set for both real and forged images.

Let $p_{\text{real}}^{\max}$ be the maximum forgery region proportion among all real images in the training set; $p_{\text{fake}}^{\min}$ be the minimum forgery region proportion among all forged images in the validation set. Then, the decision threshold θ is set as $\theta = \frac{p_{\text{real}}^{\max} + p_{\text{fake}}^{\min}}{2}$. This approach ensures that the threshold lies between the distribution ranges of real and forged images, which helps improve classification accuracy. For a test image, if its forgery region proportion p satisfies $p > \theta$, the image is classified as forged; otherwise, it is classified as real.

Let the sliding window be of size $p \times p$ and stride s. Then the number of windows in the horizontal and vertical directions is $N_h = \left\lfloor \frac{H-p}{s} \right\rfloor + 1$, $N_w = \left\lfloor \frac{W-p}{s} \right\rfloor + 1$. Therefore, the total number of windows is $N = N_h \times N_w \approx \frac{H \cdot W}{s^2}$, and the resulting time complexity of the sliding-window procedure is $O\left(\frac{H \cdot W}{s^2}\right)$.

4 Experiment

In the first part of this section, we present our detailed experimental setup. In the second part, we evaluate and compare CRFD-Net against other forgery detection methods, including state-of-the-art approaches. Finally, we conduct an ablation study on our proposed model.

4.1 Experimental Setup

Dataset and Evaluation Metrics All experiments are conducted on the public DDL-I dataset from the IJCAI challenge, designed for both deepfake detection and forged-region localization. It includes 1.2 million face images across Real, Fake, and Mask subsets, supporting joint classification and segmentation tasks. Each fake image has a pixel-level annotated mask, enabling precise evaluation of boundary and small-region localization. DDL-I covers single and multi-face "in-the-wild" scenes and incorporates 61 forgery generators, including face swapping, reenactment, and editing, to ensure broad generalization and robustness. Real and fake images are labeled 0 and 1 respectively, with 400K masks matching 400K fake images.

During training, we supervise the localization branch with both fake images and their masks. At evaluation time, we

Method	Year	AUC	F1
AIDE	2024	91.84	74.91
DRCT	2024	92.03	72.73
NPR	2024	94.27	75.98
CRFD-Net (Ours)	-	**97.61**	**78.05**

Table 1: Classification Performance Evaluation (values in %) on different methods.

Method	Year	IoU	AUC	F1
AdaIFL	2024	68.02	95.73	76.42
FLTNet	2024	70.39	94.52	**79.71**
CRFD-Net (Ours)	-	**72.8**	**97.61**	78.05

Table 2: Localization Performance Evaluation (values in %) on different methods.

use classification accuracy (ACC) and area under the ROC curve (AUC) for the Real vs. Fake task, and intersection over union (IoU) plus boundary-F1 score to assess pixel-level localization quality by comparing predicted masks against the ground truth.

Implementation Details All training was conducted on an Ubuntu 22.04 workstation equipped with an NVIDIA RTX A6000 GPU. The software environment comprises Python 3.10 and PyTorch 2.7.0 (CUDA 12.6/cu126). We employ the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of 4, training for 100 epochs. The loss function is binary cross-entropy with class-balancing weights to further sharpen forged-region detection precision. The inference time per 126×126 image is approximately 2.88 ms.

4.2 Comparison with State-of-the-Art Methods

Classification Performance Evaluation In our comparative experiments, CRFD-Net surpasses prior state-of-the-art methods in both AUC and F1. The results are showed in Table 1. Specifically, it raises AUC from NPR's 94.27% to 97.61%, showing better overall discrimination and robustness to diverse forgery artifacts. For F1, CRFD-Net improves from 75.98% to 78.05%, highlighting stronger pixel-level localization with fewer misses and false alarms.

These gains result from two key innovations: a contrastive residual module that amplifies upsampling-related anomalies, and a sliding-window fusion strategy that enhances boundary precision and small-region detection. Together, they make CRFD-Net highly effective for deepfake detection and localization.

Localization Performance Evaluation In our study, CRFD-Net achieves the highest IoU at 72.85%, surpassing FLTNet (70.39%) and AdaIFL (68.02%). The results are showed in Table 2. This 2.46-point gain over FLTNet confirms that our contrastive enhancement and sliding-window strategies effectively improve localization, particularly in small and boundary regions.

For classification, AdaIFL and FLTNet reach AUCs of 95.73% and 94.52%, respectively, while CRFD-Net sets a new benchmark at 97.61%, maintaining high recall with fewer false positives across diverse forgery types.

Threshold Strategy	AUC	F1	Precision	Recall
$p_{\max}^{\text{real}}$	94.83	74.62	78.26	71.31
$p_{\min}^{\text{fake}}$	95.54	75.69	71.39	**80.54**
$(p_{\max}^{\text{real}} + p_{\min}^{\text{fake}})/2$	**97.61**	**78.05**	**77.76**	78.34

Table 3: Comparison of threshold strategies on model performance (values in %).

In terms of F1, CRFD-Net scores 78.05%, outperforming AdaIFL (76.42%) and slightly trailing FLTNet (79.71%), as our method favors fine-grained localization, potentially trading off recall in larger forgeries. Overall, CRFD-Net demonstrates strong performance in both detection and localization tasks.

4.3 Ablation Study

Forgery Detection Threshold

To validate the proposed threshold $\theta = \frac{p_{\max}^{\text{real}} + p_{\min}^{\text{fake}}}{2}$, Table 3 compares the performance of three threshold strategies in terms of IoU, AUC, Precision, and Recall. $p_{\max}^{\text{real}}$ is the maximum proportion of falsely detected forgery regions in genuine training images; $p_{\min}^{\text{fake}}$ is the minimum proportion of correctly detected forgery regions in forged training images; θ represents the midpoint between these two extremes.

Adopting $p_{\max}^{\text{real}}$ as the threshold yields high precision but relatively low recall, indicating a conservative decision rule that misses many forgery regions. Conversely, using $p_{\min}^{\text{fake}}$ increases recall at the expense of precision, reflecting a more aggressive threshold that produces more false positives. In contrast, the midpoint threshold θ achieves the highest AUC and F1 scores and balances precision and recall. This aligns with the general precision–recall trade-off observed across thresholds. Overall, these results confirm that $\theta = \frac{p_{\max}^{\text{real}} + p_{\min}^{\text{fake}}}{2}$ provides an optimal compromise between detection accuracy and coverage.

Generalization Across Generator Families

To evaluate the generalization ability of the proposed method across different forgery techniques, we tested CRFD-Net on various types of forged datasets. These include the GAN-based FFHQ [Karras *et al.*, 2020], the diffusion-based DIRe-CelebA-HQ [Wang *et al.*, 2023], and the face-swap-based DeepFakes [Li *et al.*, 2019]. We computed both classification and localization metrics, including AUC, F1, Precision, and Recall. As shown in Table X, for the GAN-based dataset FFHQ, CRFD-Net achieved the highest AUC and F1 scores, indicating excellent accuracy and stability in detecting traditional GAN-generated images. For the diffusion-based DIRe-CelebA-HQ, the model's performance was slightly lower than on the FFHQ dataset but still remained at a high level. In the case of the face-swap-based DeepFakes dataset, although there was a minor performance drop, the results were still significantly better than random, demonstrating that the proposed forgery cues remain effective in real-world face-swapping scenarios. Despite the differences in generation

Generator Type	AUC	F1	Precision	Recall
FFHQ	**98.46**	**81.71**	**80.76**	**82.70**
DIRe-CelebA-HQ	97.80	80.03	78.45	81.67
DeepFakes	95.98	73.62	72.81	74.45

Table 4: Performance comparison across generator types (values in %).

Configuration	**AUC**	**F1**	**IoU**	**Time (s)**	**GPU (GB)**
CRFD-Net	97.61	78.05	72.80	28.8 ± 1.1	28.2
w/o CEM	90.83	68.53	64.53	28.8 ± 1.0	28.2
w/o Residual	96.17	74.84	68.06	27.5 ± 1.2	26.6
w/o SWF	95.26	73.21	68.21	22.3 ± 0.9	25.3

Table 5: Comparison of performance and resource consumption under different component configurations. Each input image size is 126×126.

mechanisms and forgery traces among various generators, our method consistently achieves high detection performance across all types.

Component Effectiveness and Efficiency

Table 5 presents a comparison of CRFD-Net and its ablated variants in terms of performance and resource consumption. The complete model achieves the best results across all metrics, demonstrating the effectiveness of the proposed architecture. Removing the CEM leads to a significant performance drop, with AUC and F1 scores decreasing by approximately 7%, highlighting the crucial role of the context enhancement module in improving the discrimination of forged regions. Although removing the residual connections or the sliding window fusion (SWF) slightly reduces resource consumption, it also causes noticeable declines in F1 and IoU scores, indicating their importance in facilitating information flow and refining spatial boundaries. Overall, the full model integrates these components effectively to achieve an optimal balance between performance and efficiency.

5 Conclusion

This paper proposes a novel Contrastive Residual Forgery Detection Network (CRFD-Net) designed to enhance both the interpretability and localization accuracy in deepfake image detection. Unlike conventional methods that primarily focus on binary authenticity classification, CRFD-Net emphasizes fine-grained localization of forged regions, addressing the core limitation of poor explainability in current forgery detection systems. By introducing a dedicated residual structure and integrating a local contrastive enhancement mechanism, the model effectively amplifies abnormal correlations among neighboring pixels in forged areas, thereby improving its sensitivity to forgery traces. Combined with a U-Net-like decoder and a sliding window-based prediction fusion strategy, CRFD-Net achieves dual improvements in both localization and classification tasks. Experimental results demonstrate that the proposed method consistently outperforms state-of-the-art approaches across multiple metrics, exhibiting strong robustness and generalization capabilities. Looking forward, CRFD-Net holds significant potential for practical applications that demand high-precision localization, such as content moderation and forensic analysis.

References

[Cao *et al.*, 2023] Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip S Yu, and Lichao Sun. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*, 2023.

[Chen *et al.*, 2024] Baoying Chen, Jishen Zeng, Jianquan Yang, and Rui Yang. Drct: Diffusion reconstruction contrastive training towards universal detection of diffusion generated images. In *Forty-first International Conference on Machine Learning*, 2024.

[Corvi *et al.*, 2023a] Riccardo Corvi, Davide Cozzolino, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. Intriguing properties of synthetic images: from generative adversarial networks to diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 973–982, 2023.

[Corvi *et al.*, 2023b] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Cozzolino and Verdoliva, 2019] Davide Cozzolino and Luisa Verdoliva. Noiseprint: A cnn-based camera model fingerprint. *IEEE Transactions on Information Forensics and Security*, 15:144–159, 2019.

[Dong *et al.*, 2022] Chengbo Dong, Xinru Chen, Ruohan Hu, Juan Cao, and Xirong Li. Mvss-net: Multi-view multi-scale supervised networks for image manipulation detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3539–3553, 2022.

[Guillaro *et al.*, 2023] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. 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*, pages 20606–20615, 2023.

[Hu *et al.*, 2020] Xuefeng Hu, Zhihan Zhang, Zhenye Jiang, Syomantak Chaudhuri, Zhenheng Yang, and Ram Nevatia. Span: Spatial pyramid attention network for image manipulation localization. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 312–328. Springer, 2020.

[Karras *et al.*, 2020] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020.

[Kwon *et al.*, 2022] Myung-Joon Kwon, Seung-Hun Nam, In-Jae Yu, Heung-Kyu Lee, and Changick Kim. Learning jpeg compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, 130(8):1875–1895, 2022.

[Li *et al.*, 2019] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019.

[Li *et al.*, 2020] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5001–5010, 2020.

[Li *et al.*, 2021] Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6458–6467, 2021.

[Luo *et al.*, 2024] Yunpeng Luo, Junlong Du, Ke Yan, and Shouhong Ding. Lareˆ 2: Latent reconstruction error based method for diffusion-generated image detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17006–17015, 2024.

[Ma *et al.*, 2023] Ruipeng Ma, Jinhao Duan, Fei Kong, Xiaoshuang Shi, and Kaidi Xu. Exposing the fake: Effective diffusion-generated images detection. *arXiv preprint arXiv:2307.06272*, 2023.

[McCloskey and Albright, 2018] Scott McCloskey and Michael Albright. Detecting gan-generated imagery using color cues. *arXiv preprint arXiv:1812.08247*, 2018.

[McCloskey and Albright, 2019] Scott McCloskey and Michael Albright. Detecting gan-generated imagery using saturation cues. In *2019 IEEE international conference on image processing (ICIP)*, pages 4584–4588. IEEE, 2019.

[Park *et al.*, 2018] Jinseok Park, Donghyeon Cho, Wonhyuk Ahn, and Heung-Kyu Lee. Double jpeg detection in mixed jpeg quality factors using deep convolutional neural network. In *Proceedings of the European conference on computer vision (ECCV)*, pages 636–652, 2018.

[Qian *et al.*, 2020] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European conference on computer vision*, pages 86–103. Springer, 2020.

[Ricker *et al.*, 2022] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. Towards the detection of diffusion model deepfakes. *arXiv preprint arXiv:2210.14571*, 2022.

[Ricker *et al.*, 2024] Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9130–9140, 2024.

[Sha *et al.*, 2023] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. 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*, pages 3418–3432, 2023.

[Wang *et al.*, 2022a] Junke Wang, Zuxuan Wu, Jingjing Chen, Xintong Han, Abhinav Shrivastava, Ser-Nam Lim, and Yu-Gang Jiang. Objectformer for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2364–2373, 2022.

[Wang *et al.*, 2022b] Junke Wang, Zuxuan Wu, Wenhao Ouyang, Xintong Han, Jingjing Chen, Yu-Gang Jiang, and Ser-Nam Li. M2tr: Multi-modal multi-scale transformers for deepfake detection. In *Proceedings of the 2022 international conference on multimedia retrieval*, pages 615–623, 2022.

[Wang *et al.*, 2023] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li. Dire for diffusion-generated image detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22445–22455, 2023.

[Wu *et al.*, 2019a] 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.*, 2019b] 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.

[Yang *et al.*, 2020] Chao Yang, Huizhou Li, Fangting Lin, Bin Jiang, and Hao Zhao. Constrained r-cnn: A general image manipulation detection model. In *2020 IEEE International conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2020.

[Zhao *et al.*, 2023] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2), 2023.

[Zhou *et al.*, 2018] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Learning rich features for image manipulation detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1053–1061, 2018.

[Zhou *et al.*, 2020] Peng Zhou, Bor-Chun Chen, Xintong Han, Mahyar Najibi, Abhinav Shrivastava, Ser-Nam Lim, and Larry Davis. Generate, segment, and refine: Towards generic manipulation segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13058–13065, 2020.