Morphology-optimized Multi-Scale Fusion: Combining Local Artifacts and Mesoscopic Semantics for Deepfake Detection and Localization

Chao Shuai[1,*] , **Gaojian Wang**[1,*] , **Kun Pan**[1,*] , **Tong Wu**[1,*] , **Fanli Jin**[1,*] , **Haohan Tan**[1,*] , **Mengxiang Li**[1,*] , **Zhenguang Liu**[1,2] , **Feng Lin**[1,2] and **Kui Ren**[1,2]

[1]The State Key Laboratory of Blockchain and Data Security, Zhejiang University
[2]Hangzhou High-Tech Zone (Binjiang) Institute of Blockchain and Data Security
{chaoshui, 12321142, pankun, cocotwu, 12421197, tanhh2023, limengxiang, flin, kuiren}@zju.edu.cn, liuzhenguang2008@gmail.com

Abstract

While the pursuit of higher accuracy in deepfake detection remains a central goal, there is an increasing demand for precise localization of manipulated regions. Despite the remarkable progress made in classification-based detection, accurately localizing forged areas remains a significant challenge. A common strategy is to incorporate forged region annotations during model training alongside manipulated images. However, such approaches often neglect the complementary nature of local detail and global semantic context, resulting in suboptimal localization performance. Moreover, an often-overlooked aspect is the fusion strategy between local and global predictions. Naively combining the outputs from both branches can amplify noise and errors, thereby undermining the effectiveness of the localization.

To address these issues, we propose a novel approach that independently predicts manipulated regions using both local and global perspectives. We employ morphological operations to fuse the outputs, effectively suppressing noise while enhancing spatial coherence. Extensive experiments reveal the effectiveness of each module in improving the accuracy and robustness of forgery localization.

1 Introduction

The rapid advancement of deep generative models [Goodfellow *et al.*, 2014; Ho *et al.*, 2020], particularly deepfakes, has led to an unprecedented level of realism and sophistication in synthetic media. When such technologies are misused, the consequences are evident: the proliferation of misinformation, reputational damage, and the erosion of public trust. While existing deepfake detection models [Hu *et al.*, 2022; Shiohara and Yamasaki, 2022] have achieved notable progress in classification, their ability to pinpoint tampered regions remains limited. This deficiency hinders practical applications such as forensic analysis and content moderation, where identifying manipulated pixels is critical for accountability and trustworthiness.

Current deepfake localization methods exhibit three primary issues. First, *local artifact detectors*, exemplified by CNN-based noise residual analysis [Hu *et al.*, 2020] and edge inconsistency modeling [Wu and Natarajan, 2019], suffer from semantic blindness, failing to detect contextually implausible manipulations. Second, *global approaches* that leverage frequency spectrum analysis [Qian *et al.*, 2020a] or transformer-based contextual reasoning [Guillaro *et al.*, 2023] lack the granularity to pinpoint fine-grained spatial anomalies. Third, hybrid frameworks [Wang *et al.*, 2022] that mechanically combine local/global features employ static fusion strategies, neglecting dynamic adaptation to manipulation characteristics across scales. Crucially, existing methods universally overlook *mesoscopic artifacts*—manipulation traces manifesting at intermediate scales between pixel-level distortions and semantic contradictions [Zhu *et al.*, 2025].

To tackle these challenges, we propose a novel morphology-optimized multi-scale fusion framework for deepfake detection and localization, which synergizes local forgery artifacts with mesoscopic semantic information. First, we present a two-stream Local Facial Forgery Detection and Location (**LFDL**) network that combines RGB and SRM (Steganalysis Rich Model) features through cross-modality consistency enhancement, effectively capturing fine-grained forensic traces like noise and edge artifacts. Second, we propose a Mesoscopic Image Tampering Localization (**MITL**) network that integrates frequency-enhanced representations with adaptive multi-scale weighting to encode both object-level and scene-consistency information. Third, we introduce a Morphology-Driven Mask Fusion (**MDMF**) strategy that intelligently merges LFDL and MITL localization results through differential dilation/erosion operations, generating reconciled global predictions.

Our key contributions are summarized as below:

- We propose a novel hybrid deepfake detection and localization framework that synergistically combines local and mesoscopic forgery cues.
- We introduce a morphology-driven mask fusion strategy that adaptively refines localization masks by dilating local predictions and eroding mesoscopic predictions.
- We evaluate the effectiveness and robustness of the proposed framework through extensive evaluation.

[*] Equal contribution.

Table 1: Supplementary Data Registration for Deepfake Detection Model Training: Model Types, Methods, and Forgery Details

Model Type	Method	Forgery Types	Fake/Mask Image	Reference
Image Edit	SBIs	FaceSwap	18135	[Shiohara and Yamasaki, 2022]
	Random combination	FaceSwap	17728	-
GAN	Simswap	FaceSwap	14999	[Chen *et al.*, 2020]
	MaskFaceGAN	Face Attribute Editing	14999	[Pernuš *et al.*, 2023]
	Facedancer	FaceSwap	20000	[Rosberg *et al.*, 2023]
Diffusion Model	BELM	Diffusion Inversion	14674	[Wang *et al.*, 2024]
	SD-inpanting	Inpanting	18347	[Podell *et al.*, 2023]

2 Related Work

Deepfake Detection Deepfake detection techniques have evolved significantly to counter the increasing sophistication of forgery paradigms. Current methods can be broadly categorized into four main approaches based on the cues they exploit: spatial domain [Liu *et al.*, 2020; Nirkin *et al.*, 2021; Cao *et al.*, 2022; Wang and Chow, 2023; Tan *et al.*, 2023], temporal domain [Yang *et al.*, 2019; Gu *et al.*, 2022; Yang *et al.*, 2023; Choi *et al.*, 2024; Peng *et al.*, 2024], frequency domain [Qian *et al.*, 2020b; Li *et al.*, 2021; Miao *et al.*, 2022; Guo *et al.*, 2023b; Tan *et al.*, 2024], and data-driven methods [Dang *et al.*, 2020; Zhao *et al.*, 2021; Hu *et al.*, 2022; Huang *et al.*, 2023; Guo *et al.*, 2023a; Zhang *et al.*, 2024a]. However, most existing methods focus solely on binary classification (real vs. fake), neglecting the localization of manipulated regions. This limitation not only restricts the practical utility of detection systems in forensic scenarios but also weakens model interpretability, as the lack of localization prevents both decision validation and forgery patterns analysis.

Deepfake Location Deepfake localization [Miao *et al.*, 2024; Miao *et al.*, 2023; Zhang *et al.*, 2024b] aims to identify manipulated regions through pixel-level analysis. Existing methods for deepfake localization primarily focus on either local or global approaches to identifying manipulated regions. Local location methods, such as ManTra-Net [Wu *et al.*, 2019] and TruFor [Guillaro *et al.*, 2023], detect anomalies within smaller regions of an image, focusing on fine-grained inconsistencies like noise and texture irregularities. Global location methods, including F3-Net [Qian *et al.*, 2020b], ObjectFormer [Wang *et al.*, 2022], and MVSS-Net [Chen *et al.*, 2021], consider the entire image or video frame, identifying larger-scale inconsistencies and manipulation artifacts that span broader areas.

In contrast to existing methods, our approach integrates both local, mesoscopic, and global location information, applying morphological fusion techniques to enhance the robustness and precision of deepfake localization.

3 Method

3.1 Overview

Our proposed framework is designed for robust deepfake detection and localization by synergizing local forgery artifacts with mesoscopic semantic information, to formulate a comprehensive and precise global prediction. The overall framework, as depicted in Figure 1, comprises three main components: the Local Facial Forgery Detection and Localization (LFDL) network, the Mesoscopic Image Tampering Localization (MITL) network, and the Morphology-Driven Mask Fusion (MDMF) strategy.

3.2 Data Preparation

1) Generating Diverse Forgery Images and Masks: To improve the model's robustness to diverse manipulation patterns, we construct a supplementary dataset by applying various forgery techniques to authentic images. Instead of relying on external datasets, we utilize the original real images to generate manipulated samples using representative methods spanning traditional image editing, GAN-based synthesis, and diffusion models. This process not only increases the diversity of forgery types but also provides paired masks for supervised learning. The details of the constructed dataset are shown in Table 1. In total, 118,882 manipulated images were generated using various forgery techniques, each accompanied by a corresponding mask.

2) Data Preprocessing Strategy: To tackle the challenging task of facial forgery detection, we employ two distinct models, each is designed to analyze different forgery scales (e.g., facial attribute and background context), respectively. Correspondingly, two preprocessing pipelines are adopted to prepare the input for these models, ensuring alignment with their respective detection strategies. The LFDL module focuses on analyzing local facial regions, and its preprocessing pipeline emphasizes face localization and adaptive cropping to highlight detailed facial features. In contrast, the MITL module leverages global image context, and its preprocessing approach uses the entire original image and corresponding forgery mask without any facial cropping. By pairing these models with tailored preprocessing techniques, we simultaneously utilize localized and global cues, enhancing the robustness and comprehensiveness of the overall detection system.

In the LFDL module, face detection is applied to localize the face region in the image. If the detected face occupies a moderate portion of the entire image, the face region is cropped; otherwise, the original image is fed directly into the model. This adaptive cropping strategy highlights local facial details while preventing excessive cropping of small face regions, thereby preserving key information.

In contrast, the MITL module avoids face detection or cropping altogether. Instead, the entire original image and its corresponding forgery mask are directly input into the model. By preserving both global structure and background, this method leverages the overall context of the image in

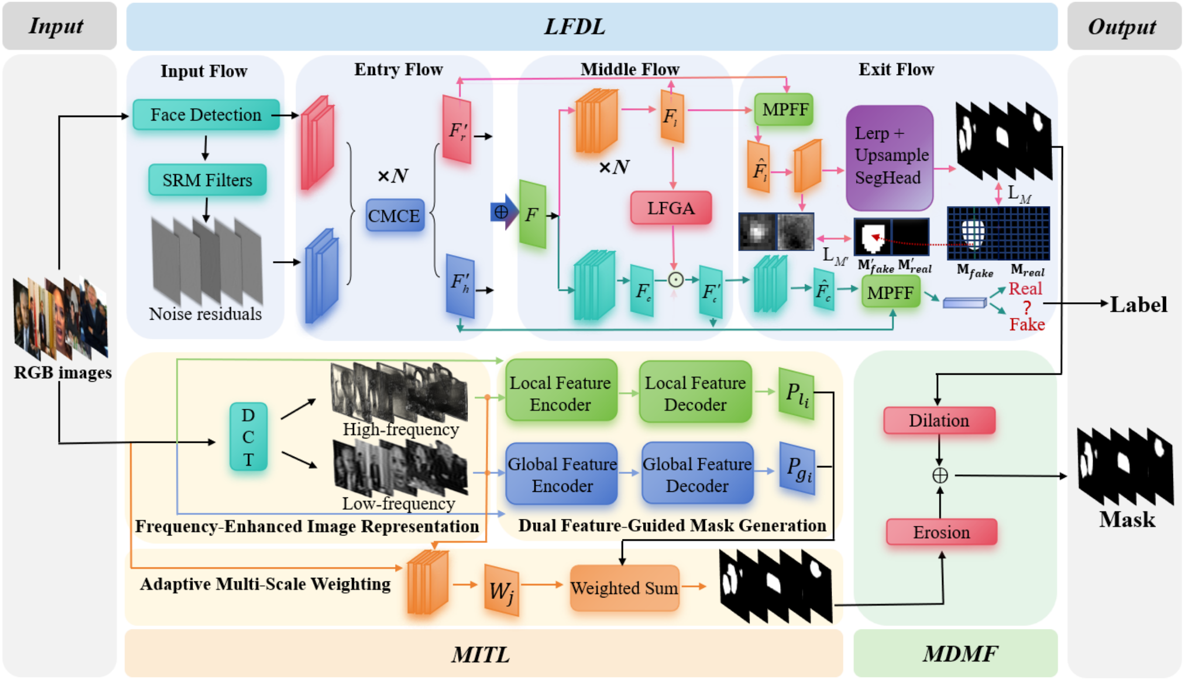

Figure 1: Overview of our proposed framework. (1) The **LFDL** module employs a two-stream architecture (RGB and SRM streams) and utilizes cross-modality consistency mechanisms for manipulation detection. (2) The **MITL** module processes frequency-enhanced features through dual encoders to achieve semantic-level tampering identification. (3) The **MDMF** module combines the dilated mask from **LFDL** (enhancing edge coherence) and the eroded mask from **MITL** (suppressing over-prediction) through a union operation to achieve comprehensive localization.

addition to facial cues. This preprocessing pipeline emphasizes global information, complementing the localized focus of LFDL, and together they enhance the completeness and robustness of facial forgery detection.

3.3 LFDL: Local Facial Forgery Detection and Location with Two-Stream Architecture

Deepfake manipulations often introduce subtle and spatially confined artifacts in key facial regions, such as the eyes, mouth, and contours. These fine-grained inconsistencies are difficult to capture using coarse global features alone. To address this, we employ a patch-based strategy wherein facial regions are first detected and cropped, then processed by a dedicated Two-Stream Network [Shuai *et al.*, 2023] designed to capture both RGB image and SRM noise residuals forgery cues at a local level.

The architecture comprises two parallel branches, classification and localization, each augmented by four key components that jointly exploit cross-modal consistency and patch-level context, as described below.

1) Cross-Modality Consistency Enhancement (CMCE): Let $F_{\text{rgb}}^l \in \mathbb{R}^{C_l \times H_l \times W_l}$ and $F_{\text{srm}}^l \in \mathbb{R}^{C_l \times H_l \times W_l}$ denote the features extracted from the l-th layer of the RGB and SRM streams, respectively. CMCE computes a cross-modal consistency map by measuring the cosine similarity at each spatial location:

$$\text{Corr}(f_i^r, f_i^h) = \frac{f_i^r \cdot f_i^h}{\|f_i^r\|_2 \|f_i^h\|_2}, \quad i \in \{1, \dots, H_l W_l\} \tag{1}$$

where $f_i^r, f_i^h \in \mathbb{R}^{C_l \times 1 \times 1}$ are the corresponding spatial vectors from the RGB and SRM streams. This correlation is used to refine both modalities:

$$F_r' = \text{ReLU}(F_r + \text{Corr} \odot F_h), \quad F_h' = \text{ReLU}(F_h + \text{Corr} \odot F_r) \tag{2}$$

Finally, the enhanced representation is obtained by summing the two refined features:

$$F_{\text{cmce}}^l = F_r' + F_h' \tag{3}$$

This mechanism reinforces cross-modal interactions, allowing the network to exploit complementary forgery clues present in both spatial and frequency domains.

2) Local Forgery Guided Attention (LFGA): To mitigate the common failure case where networks overly rely on unmanipulated regions, LFGA explicitly guides attention toward potentially tampered areas. It first constructs a self-attention map from the localization branch's feature F_l as:

$$\text{Att}_{ij} = \text{Softmax}(g(F_l)_i \cdot g(F_l)_j) \tag{4}$$

where $g(\cdot)$ is a learnable linear projection and i, j index spatial locations. This attention map highlights patch-wise similarity and forgery salience. The attention is then used to recalibrate the classification feature F_c:

$$F_c^* = \text{ReLU}(\text{Reshape}(h(F_c) \otimes \text{Att}) + F_c) \tag{5}$$

where $h(\cdot)$ is a projection function and $\otimes$ denotes matrix multiplication. LFGA is applied at multiple scales to progressively refine spatial focus and promote forgery-aware classification.

3) Multi-Scale Patch Feature Fusion (MPFF): Forgery traces often vary in spatial granularity. MPFF fuses features across multiple resolution levels to capture both coarse and fine-level inconsistencies. For the localization stream, high-resolution intermediate features $F_{ml} \in \mathbb{R}^{c_2 \times h_2 \times w_2}$ and low-resolution features $F_l \in \mathbb{R}^{c_1 \times h_1 \times w_1}$ are partitioned into corresponding non-overlapping patches. Within each patch P_k, an intra-patch consistency is computed as:

$$f_{ml}^{(k,j)} = \text{Tanh}\left(\frac{\theta(p_k^j) \cdot \theta(f_l^k)}{c}\right) \tag{6}$$

where $\theta(\cdot)$ is a shared projection and c is a normalization constant. The same strategy is applied in the classification stream to generate $f_{mc}^{(k,j)}$, allowing joint multi-scale fusion and consistency modeling. This patch-wise alignment helps bridge semantic gaps across scales while preserving spatial fidelity.

4) Semi-Supervised Patch Similarity Learning (SSPSL): Most public deepfake datasets lack pixel-level annotations of manipulated regions, which hinders effective supervision of the localization branch. To overcome this, SSPSL adopts a semi-supervised approach to generate pseudo ground-truth masks based on patch similarity analysis.

First, facial landmarks are used to detect the nose position, around which a rectangular area is heuristically designated as the manipulated region. Patches from this area are treated as pseudo-forged samples, while patches from authentic images serve as real references.

Given the feature map $F_f \in \mathbb{R}^{C \times H \times W}$, we compute cosine similarity between each patch feature f_{ij}^f and the average features of real (f_r) and forged (f_a) patches:

$$S_{ij}^{fr} = \frac{f_{ij}^f \cdot f_r}{\|f_{ij}^f\|_2 \|f_r\|_2}, \quad S_{ij}^{ff} = \frac{f_{ij}^f \cdot f_a}{\|f_{ij}^f\|_2 \|f_a\|_2} \tag{7}$$

Each patch is then assigned a binary label indicating whether it is more similar to real or forged examples:

$$M_{ij} = \begin{cases} 0, & S_{ij}^{fr} - S_{ij}^{ff} \geq 0 \\ 1, & S_{ij}^{fr} - S_{ij}^{ff} < 0 \end{cases} \tag{8}$$

This results in a pseudo mask M that approximates manipulated regions, which is further converted into patch-level labels. For a patch P_k, its label M_k is defined by the average value of its corresponding pixels:

$$M_k = \begin{cases} 0, & \text{avg}(M_{P_k}) = 0 \\ 1, & \text{avg}(M_{P_k}) > 0 \end{cases} \tag{9}$$

The localization branch is supervised using a binary cross-entropy loss between predicted mask logits $\hat{M}_k$ and pseudo labels M_k:

$$\mathcal{L}_{\text{loc}} = -\frac{1}{h_1 w_1} \sum_{k=1}^{h_1 w_1} \left[M_k \log \hat{M}_k + (1 - M_k) \log(1 - \hat{M}_k) \right] \tag{10}$$

Meanwhile, the classification branch is optimized with a standard binary cross-entropy loss:

$$\mathcal{L}_{\text{cls}} = -\left[y \log \hat{y} + (1 - y) \log(1 - \hat{y})\right] \tag{11}$$

Overall Training Objective: The total loss integrates both classification and localization terms for joint optimization:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{loc}} \tag{12}$$

This formulation allows the network to learn fine-grained forgery localization without explicit ground-truth annotations, enabling more scalable and flexible training under semi-supervised settings.

3.4 MITL: Mesoscopic Image Tampering Localization with Mesorch

Existing image tampering localization methods primarily rely on microscopic features, such as image RGB noise, edge signals, or high-frequency features, to detect tampering traces. However, these methods fail to effectively capture the macroscopic semantic information of tampering, leading to insufficient localization accuracy in complex scenes and difficulty in handling tampering related to semantics. To address this challenge, we apply the Mesorch [Zhu *et al.*, 2025] architecture to extract both microscopic details and macroscopic semantic information, and dynamically adjust the importance of different features through an adaptive weighting module.

The framework mainly consists of three parts: Frequency-Enhanced Image Representation, Dual Feature-Guided Mask Generation, and Adaptive Multi-Scale Weighting, as described below.

1) Frequency-Enhanced Image Representation (FEIR): The FEIR module initiates spectral decomposition using Discrete Cosine Transform (DCT) to amplify mesoscopic-level artifacts. Given an input RGB image $x \in \mathbb{R}^{H \times W \times 3}$, we decompose it into high-frequency (x_h) and low-frequency (x_l) components through DCT filtering, retaining spatial dimensions $H \times W \times 3$ for both components. These frequency-enhanced features are concatenated with the original image along the channel axis to form enhanced representations:

$$\begin{aligned} I_h &= \text{Concat}(x, x_h) \in \mathbb{R}^{H \times W \times 6}, \\ I_l &= \text{Concat}(x, x_l) \in \mathbb{R}^{H \times W \times 6} \end{aligned} \tag{13}$$

I_h and I_l are enhanced representations used for further processing. This hybrid representation bridges pixel-level anomalies (microscopic) with semantic contradictions (macroscopic), enabling the detection of subtle manipulation traces that span hierarchical levels.

2) Dual Feature-Guided Mask Generation (DFGMG): This module is a key component of the whole architecture. It processes high-frequency and low-frequency enhanced images separately to capture fine-grained details and macroscopic semantics, respectively. The outputs from these encoders are then decoded to generate initial predictions, which will be combined in the next module to produce the final mask.

The high-frequency enhanced image I_h is processed by the Local Feature Encoder, a CNN-based architecture specifically engineered to capture fine-grained local details. These details play a critical role in the detection of microscopic artifacts within the image. Concurrently, the low-frequency enhanced image I_l is processed by the Global Feature Encoder, a Transformer-based framework designed to extract macroscopic semantics and global contextual information. Such information is essential for comprehending object-level manipulations in the image.

Each encoder outputs feature maps at four distinct scales:

$$\begin{aligned}\{L_{s_1}, L_{s_2}, L_{s_3}, L_{s_4}\} &= \text{LocalFeatureEncoder}(I_h),\\ \{G_{s_1}, G_{s_2}, G_{s_3}, G_{s_4}\} &= \text{GlobalFeatureEncoder}(I_l)\end{aligned} \tag{14}$$

where C_{local} and C_{global} denote the total number of output channels at each scale i for the local and global encoders, respectively. These feature maps are then processed by the corresponding decoders to generate initial predictions of manipulated regions:

$$\begin{aligned}M_{l_i} &= \text{LocalFeatureDecoder}(L_{s_i}),\\ M_{g_i} &= \text{GlobalFeatureDecoder}(G_{s_i})\end{aligned} \tag{15}$$

where M_{l_i} and M_{g_i} are the local and global prediction masks, respectively, with a shape of $H/4 \times W/4 \times 1$ for each scale i.

3) Adaptive Multi-Scale Weighting (AMW): This Module dynamically optimizes the fusion of multi-scale representations by adjusting the importance of features across different scales. Unlike traditional methods that assume equal weighting, this module identifies and emphasizes critical scales while suppressing redundant or noisy information, thereby enhancing localization precision and robustness.

The module takes three inputs: the original RGB image $x \in \mathbb{R}^{H\times W\times 3}$, high-frequency components $x_h \in \mathbb{R}^{H\times W\times 3}$, and low-frequency components $x_l \in \mathbb{R}^{H\times W\times 3}$, both derived via DCT. These inputs are concatenated to form a composite representation $I_{\text{concat}} = \{x, x_h, x_l\} \in \mathbb{R}^{H\times W\times 9}$.

Then the weighting network will produce a tensor $W \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 8}$, where each element reflects the relative importance of predictions from local and global features across four scales (totaling eight branches). Formally:

$$W = \text{WeightingModule}(I_{\text{concat}}) \tag{16}$$

The final prediction mask M is computed through pixel-wise weighted fusion of concatenated multi-scale predictions $M_{\text{merge}} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times 8}$, followed by upsampling to the original resolution:

$$M_{\text{merge}} = \text{Concat}(M_{l_1}, M_{g_1}, M_{l_2}, M_{g_2}, M_{l_3}, M_{g_3}, M_{l_4}, M_{g_4}) \tag{17}$$

$$M = \text{Resize}\left(\sum_{j=1}^{8} W_j \odot M_{\text{merge}_j}, H, W\right) \tag{18}$$

where $\odot$ denotes element-wise multiplication. This adaptive fusion ensures spatial coherence while focusing on regions critical for mesoscopic artifact detection.

3.5 MDMF: Morphology-Driven Mask Fusion for Comprehensive Forgery Localization

To leverage the complementary strengths of LFDL and MITL, we introduce a Morphology-Driven Mask Fusion strategy that combines their outputs for more accurate localization.

The LFDL module first detects and crops the facial region, then performs fine-grained forgery localization. While effective in identifying facial manipulations, the cropping process may discard edge information, making the model sensitive to local artifacts. As a result, the output masks may have irregular boundaries or fragmented areas. Based on our observation that ground truth masks are generally smooth and coherent, we apply a dilation operation to the LFDL mask M_{LFDL} to smooth edges and connect nearby manipulated regions:

$$M_{\text{LFDL}} \oplus B = \{z \in \mathbb{Z}^2 \mid (B)_z \cap M_{\text{LFDL}} \neq \emptyset\} \tag{19}$$

$B \subseteq \mathbb{Z}^2$ is a structuring element, here we set it as a 5×5 matrix of ones. $(B)_z = \{b + z \mid b \in B\}$ represents the translation of B by displacement z.

In contrast, the MITL module directly detects manipulated regions on the full image, effectively capturing non-facial forgeries such as hair alterations. However, the image resizing process during training may lead to loss of fine details, potentially resulting in ambiguous and over-extended predictions. To reduce this effect, we apply an erosion operation to the MITL mask M_{MITL}, which helps suppress over-prediction while retaining mesoscopic perception capabilities:

$$M_{\text{MITL}} \ominus B = \{z \in \mathbb{Z}^2 \mid (B)_z \subseteq M_{\text{LFDL}}\} \tag{20}$$

Finally, we take the union of the two processed masks to obtain the final localization result M_{final}:

$$M_{\text{final}} = (M_{\text{LFDL}} \oplus B) \cup (M_{\text{MITL}} \ominus B) \tag{21}$$

This morphology-based fusion compensates for the limitations of both modules: dilation recovers LFDL's lost edge coherence, while erosion suppresses MITL's over-prediction tendencies. Their union combines fine-grained localization with consistency analysis, yielding more accurate manipulation detection.

4 Experiments

4.1 Experimental setup

Metrics. The performance of our model is evaluated using several key metrics. For detection, we report Area Under the ROC Curve (AUC). For spatial localization, we employ the F1 Score and Intersection over Union (IoU) to assess the accuracy of the localized tampered regions.

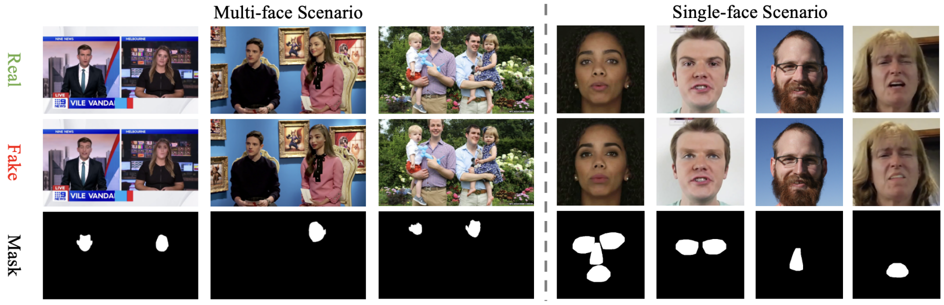

Figure 2: Examples of the DDL-I dataset.Multi-Face Scenario refers to an image with multiple faces where one or more faces have been altered. Single-Face Scenario involves altering a local region within an image that contains only one face.

The F1 Score balances precision and recall, where precision measures the accuracy of predicted tampered pixels, and recall reflects the completeness of detection. It is computed as follows:

$$F1\text{-}Score = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{22}$$

Where:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN} \tag{23}$$

Here, TP denotes the correctly predicted tampered pixels, FP represents incorrectly predicted tampered pixels, and FN refers to missed tampered pixels.

IoU measures the overlap between the predicted and ground truth manipulation regions. It is calculated as:

$$IoU = \frac{\text{Area of Intersection}}{\text{Area of Union}} \tag{24}$$

or equivalently

$$IoU = \frac{TP}{TP + FP + FN} \tag{25}$$

Finally, the overall performance is summarized by a Final Score, which is the average of the three metrics:

$$\text{Final score} = (\text{AUC} + \text{IoU} + \text{F1})/3 \tag{26}$$

Final score provides a comprehensive evaluation by combining detection and localization performance into a single metric.

Implementation Details. Our approach consists of two primary models: LFDL and MITL, which are trained separately. The LFDL model utilizes an Xception backbone and is trained on 8 × A100 GPUs with a learning rate of 5×10^{-4}, a batch size of 76, and for 30 epochs. The MITL model, which combines a Segformer-B3 and ConvNeXt-Tiny hybrid backbone, is trained on 8 × RTX 4090 GPUs. It is trained with a learning rate of 10^{-4}, a batch size of 10, and for 30 epochs. Following training, the models are fused during the inference stage to produce the final prediction. This modular training strategy allows each model to optimize performance by leveraging appropriate hardware for its own training, resulting in improved performance and efficiency.

Dataset. We evaluate our model on the Deepfake Detection and Localization Image (DDL-I) dataset [Miao *et al.*, 2025], which contains over 1.5 million samples with pixel-level annotations. DDL-I covers 61 latest deepfake methods across four forgery types: face swapping, face reenactment, full-face synthesis, and face editing. It encompasses both single-face and multi-face scenarios, providing a diverse range of forgery contexts. Additionally, pixel-level forgery region masks are provided, enabling precise localization of tampered areas.

4.2 Experimental results

Table 2 presents the results of our ablation study, aimed at evaluating the contribution of different modules and fusion strategies in the proposed framework. We report Detection AUC, F1-score, IoU for localization accuracy, and a Final Score that aggregates the classification and localization quality.

LFDL is a patch-based forgery detection and localization module, which performs both binary classification and spatially localizes forgery artifacts. It focuses on fine-grained and local inconsistencies, especially in manipulated facial regions. Despite its localized input, it achieves strong performance with an AUC of 0.9790 and a respectable IoU of 0.5981, indicating its ability to precisely detect local manipulations.

MITL represents a standalone end-to-end deepfake detection model that takes the whole image as input and produces a full-resolution forgery mask. This model provides coarse but global localization information. When used alone, it yields a significantly lower Final Score of 0.2349, possibly due to lack of fine-grained localization capacity.

LFDL + MITL corresponds to a naive combination strategy where the detection confidence is derived from the LFDL module, while the global mask from the MITL module is directly used for localization. This combination improves the Final Score to 0.3200, suggesting that LFDL's strong classification signal can benefit global prediction. However, without careful integration, the spatial coherence remains limited.

LFDL + MITL + Mask Naive Fusion applies a simple yet effective fusion strategy: the final localization mask is obtained by a pixel-wise logical OR operation between the LFDL module's patch-based local mask and the MITL module's global mask. This ensures that any region predicted as fake by either model is retained. This strategy significantly

Table 2: Ablation Study. LFDL is a two-stream architecture focusing on local feature-based detection. MITL leverages the Mesorch backbone to extract both local artifacts and global semantic features. Mask Naive Fusion directly adds the output masks from LFDL and MITL, treating a pixel as fake if either branch flags it. MDMF further refines this fusion using morphological operations to enhance coherence and suppress noise. The final score is the average of AUC, F1-score, and IoU.

Method	Detection AUC	F1-score	IoU	Final Score
LFDL	0.9790	0.6840	0.5981	0.7497
MITL	-	-	-	0.2349
LFDL + MITL	-	-	-	0.3200
LFDL + MITL + Mask Naive Fusion	0.9790	0.7598	0.6657	0.8015
LFDL + MITL + MDMF	0.9790	0.7759	0.6902	0.8150

boosts both the F1-score (to 0.7598) and IoU (to 0.6657), highlighting the complementary strengths of the two models—local detail sensitivity and global coverage.

LFDL + MITL + MDMF further applies post-processing to refine the fused mask using morphological operations (e.g., dilation, erosion). This enhances mask smoothness and removes small noisy regions, achieving the highest performance with a Final Score of 0.8150, F1-score of 0.7759, and IoU of 0.6902.

In summary, the LFDL module offers precise local forgery detection, while the MDMF module complements it by providing global contextual information. Their fusion—especially through naive logical integration and morphological refinement—yields substantial gains in both detection and localization, underscoring the value of combining local and global cues.

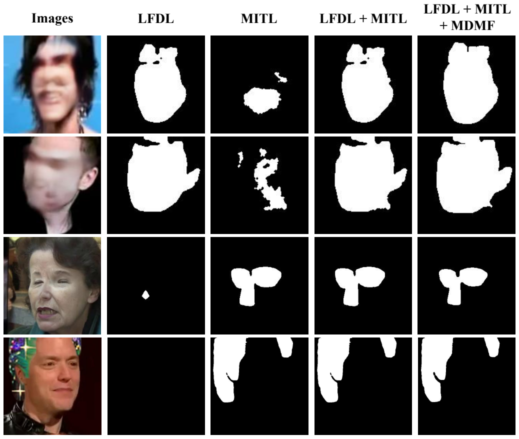

Figure 3: Visualization results of our methods. We apply dilation to LFDL masks and erosion to MITL masks, then combine them to achieve precise and coherent forgery localization.

4.3 Visualization

As shown in Figure 3, our method achieves precise forgery localization by combining the strengths of LFDL and MITL. The LFDL module accurately detects facial manipulations but suffers from irregular edges and fragmented regions, while the MITL module effectively identifies peripheral artifacts (e.g., hair region anomalies) yet tends to over-expand boundaries. To address these issues, we innovatively adopt MDMF, which performs dilation to LFDL masks to smooth fragmented areas and erosion to MITL masks to suppress boundary over-extension, followed by an intersection operation for final localization. Experimental results demonstrate that this strategy significantly improves boundary precision and structural coherence of the detection masks.

5 Conclusion

In this work, we address the challenging task of deepfake localization by proposing a dual-branch framework that separately models local and global manipulation cues. While existing methods often overlook the importance of semantic-aware localization and suffer from ineffective fusion strategies, our approach explicitly leverages both fine-grained local details and global semantic context. By applying morphological fusion to the independently predicted masks, we suppress noisy activations and improve spatial coherence. Extensive experiments demonstrate the effectiveness of our method, showing that it achieves more accurate and robust localization results compared to existing approaches. This work highlights the importance of multi-perspective modeling and adaptive fusion in advancing the state-of-the-art in deepfake forensics.

References

[Cao *et al.*, 2022] Junyi Cao, Chao Ma, Taiping Yao, Shen Chen, Shouhong Ding, and Xiaokang Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4113–4122, 2022.

[Chen *et al.*, 2020] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2003–2011, 2020.

[Chen *et al.*, 2021] Xinru Chen, Chengbo Dong, Jiaqi Ji, juan Cao, and Xirong Li. Image manipulation detection by multi-view multi-scale supervision. In *The IEEE International Conference on Computer Vision (ICCV)*, 2021.

[Choi *et al.*, 2024] Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. Exploiting style latent flows for generalizing deepfake detection video detection. *arXiv e-prints*, pages arXiv–2403, 2024.

[Dang *et al.*, 2020] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, pages 5781–5790, 2020.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[Gu *et al.*, 2022] Zhihao Gu, Yang Chen, Taiping Yao, Shouhong Ding, Jilin Li, and Lizhuang Ma. Delving into the local: Dynamic inconsistency learning for deepfake video detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 744–752, 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 (CVPR)*, pages 20606–20615, June 2023.

[Guo *et al.*, 2023a] Xiao Guo, Xiaohong Liu, Zhiyuan Ren, Steven Grosz, Iacopo Masi, and Xiaoming Liu. Hierarchical fine-grained image forgery detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3165, 2023.

[Guo *et al.*, 2023b] Zhiqing Guo, Zhenhong Jia, Liejun Wang, Dewang Wang, Gaobo Yang, and Nikola Kasabov. Constructing new backbone networks via space-frequency interactive convolution for deepfake detection. *IEEE Transactions on Information Forensics and Security*, 19:401–413, 2023.

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

[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. page 312–328, Berlin, Heidelberg, 2020. Springer-Verlag.

[Hu *et al.*, 2022] Juan Hu, Xin Liao, Jinwen Liang, Wenbo Zhou, and Zheng Qin. Finfer: Frame inference-based deepfake detection for high-visual-quality videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 951–959, 2022.

[Huang *et al.*, 2023] Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4490–4499, 2023.

[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.

[Liu *et al.*, 2020] Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8060–8069, 2020.

[Miao *et al.*, 2022] Changtao Miao, Zichang Tan, Qi Chu, Nenghai Yu, and Guodong Guo. Hierarchical frequency-assisted interactive networks for face manipulation detection. *IEEE Transactions on Information Forensics and Security*, 17:3008–3021, 2022.

[Miao *et al.*, 2023] Changtao Miao, Qi Chu, Zhentao Tan, Zhenchao Jin, Wanyi Zhuang, Yue Wu, Bin Liu, Honggang Hu, and Nenghai Yu. Multi-spectral class center network for face manipulation detection and localization. *arXiv preprint arXiv:2305.10794*, 2023.

[Miao *et al.*, 2024] Changtao Miao, Qi Chu, Tao Gong, Zhentao Tan, Zhenchao Jin, Wanyi Zhuang, Man Luo, Honggang Hu, and Nenghai Yu. Mixture-of-noises enhanced forgery-aware predictor for multi-face manipulation detection and localization. *arXiv preprint arXiv:2408.02306*, 2024.

[Miao *et al.*, 2025] Changtao Miao, Yi Zhang, Weize Gao, Man Luo, Weiwei Feng, Zhiya Tan, Jianshu Li, Ajian Liu, Yunfeng Diao, Qi Chu, Tao Gong, Li Zhe, Weibin Yao, and Joey Tianyi Zhou. Ddl: A dataset for interpretable deepfake detection and localization in real-world scenarios. *arXiv preprint arXiv:2506.23292*, 2025.

[Nirkin *et al.*, 2021] Yuval Nirkin, Lior Wolf, Yosi Keller, and Tal Hassner. Deepfake detection based on discrepancies between faces and their context. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6111–6121, 2021.

[Peng *et al.*, 2024] Chunlei Peng, Zimin Miao, Decheng Liu, Nannan Wang, Ruimin Hu, and Xinbo Gao. Where deepfakes gaze at? spatial-temporal gaze inconsistency analysis for video face forgery detection. *IEEE Transactions on Information Forensics and Security*, 2024.

[Pernuš *et al.*, 2023] Martin Pernuš, Vitomir Štruc, and Simon Dobrišek. Maskfacegan: High-resolution face editing with masked gan latent code optimization. *IEEE Transactions on Image Processing*, 32:5893–5908, 2023.

[Podell *et al.*, 2023] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

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

[Qian *et al.*, 2020b] 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.

[Rosberg *et al.*, 2023] Felix Rosberg, Eren Erdal Aksoy, Fernando Alonso-Fernandez, and Cristofer Englund. Facedancer: Pose-and occlusion-aware high fidelity face swapping. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3454–3463, 2023.

[Shiohara and Yamasaki, 2022] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18720–18729, 2022.

[Shuai *et al.*, 2023] Chao Shuai, Jieming Zhong, Shuang Wu, Feng Lin, Zhibo Wang, Zhongjie Ba, Zhenguang Liu, Lorenzo Cavallaro, and Kui Ren. Locate and verify: A two-stream network for improved deepfake detection. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7131–7142, 2023.

[Tan *et al.*, 2023] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12105–12114, 2023.

[Tan *et al.*, 2024] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5052–5060, 2024.

[Wang and Chow, 2023] Tianyi Wang and Kam Pui Chow. Noise based deepfake detection via multi-head relative-interaction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14548–14556, 2023.

[Wang *et al.*, 2022] 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.*, 2024] Fangyikang Wang, Hubery Yin, Yue-Jiang Dong, Huminhao Zhu, Chao Zhang, Hanbin Zhao, Hui Qian, and Chen Li. BELM: Bidirectional explicit linear multi-step sampler for exact inversion in diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[Wu and Natarajan, 2019] Yue Wu and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *CVPR Workshops*, 2019.

[Wu *et al.*, 2019] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgerieswith anomalous features. 2019.

[Yang *et al.*, 2019] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 8261–8265. IEEE, 2019.

[Yang *et al.*, 2023] Ziming Yang, Jian Liang, Yuting Xu, Xiao-Yu Zhang, and Ran He. Masked relation learning for deepfake detection. *IEEE Transactions on Information Forensics and Security*, 18:1696–1708, 2023.

[Zhang *et al.*, 2024a] Yi Zhang, Weize Gao, Changtao Miao, Man Luo, Jianshu Li, Wenzhong Deng, Zhe Li, Bingyu Hu, Weibin Yao, Wenbo Zhou, et al. Inclusion 2024 global multimedia deepfake detection: Towards multi-dimensional facial forgery detection. *arXiv preprint arXiv:2412.20833*, 2024.

[Zhang *et al.*, 2024b] Yi Zhang, Changtao Miao, Man Luo, Jianshu Li, Wenzhong Deng, Weibin Yao, Zhe Li, Bingyu Hu, Weiwei Feng, Tao Gong, and Qi Chu. Mfms: Learning modality-fused and modality-specific features for deepfake detection and localization tasks. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11365–11369, 2024.

[Zhao *et al.*, 2021] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15023–15033, 2021.

[Zhu *et al.*, 2025] Xuekang Zhu, Xiaochen Ma, Lei Su, Zhuohang Jiang, Bo Du, Xiwen Wang, Zeyu Lei, Wentao Feng, Chi-Man Pun, and Ji-Zhe Zhou. Mesoscopic insights: orchestrating multi-scale & hybrid architecture for image manipulation localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 11022–11030, 2025.