Amplifying Discrepancies: Exploiting Macro and Micro Inconsistencies for Image Manipulation Localization

Shenghao Chen[1], Yibo Zhao[1], Tianyi Wang[2], Chunjie Ma[3*], Weili Guan[4], Ming Li[5], Zan Gao[1,3]

[1]School of Computer Science and Engineering, Tianjin University of Technology
[2]School of Computing, National University of Singapore
[3]Shandong Artificial Intelligence Institute, Qilu University of Technology (Shandong Academy of Sciences)
[4]Harbin Institute of Technology (Shenzhen)
[5]Shandong Inspur Database Technology Co., Ltd.

Abstract

The rapid development of image manipulation technologies poses significant challenges to multimedia forensics, especially in accurate localization of manipulated regions. Existing methods often fail to fully explore the intrinsic discrepancies between manipulated and authentic regions, resulting in sub-optimal performance. To address this limitation, we propose the Focus Region Discrepancy Network (FRD-Net), a novel and efficient framework that significantly enhances manipulation localization by amplifying discrepancies at both macro- and micro-levels. Specifically, our proposed Iterative Clustering Module (ICM) groups features into two discriminative clusters and refines representations via backward propagation from cluster centers, improving the distinction between tampered and authentic regions at the macro level. Thereafter, our Differential Progressive Module (DPM) is constructed to capture fine-grained structural inconsistencies within local neighborhoods and integrate them into a Central Difference Convolution, increasing sensitivity to subtle manipulation details at the micro level. Finally, these complementary modules are seamlessly integrated into a compact architecture that achieves a favorable balance between accuracy and efficiency. Extensive experiments on multiple benchmarks demonstrate that FRD-Net consistently surpasses state-of-the-art methods in terms of manipulation localization performance while maintaining a lower computational cost.

Introduction

With the rapid development of image editing tools and deep generative models (Nichol et al. 2022; Verdoliva 2020; Sun et al. 2024), it has become increasingly easy for people to produce realistic images for design, creativity, and entertainment, even for novice users. However, misuse of these tools online erodes public trust and distorts information. In response to these challenges, multimedia forensics has attracted growing attention. Among various manipulation approaches, partially altering an image can alter its semantic information, thus impacting social trust and information integrity. Therefore, in this work, we focus on image manipulation localization to trace and correct misrepresented information.

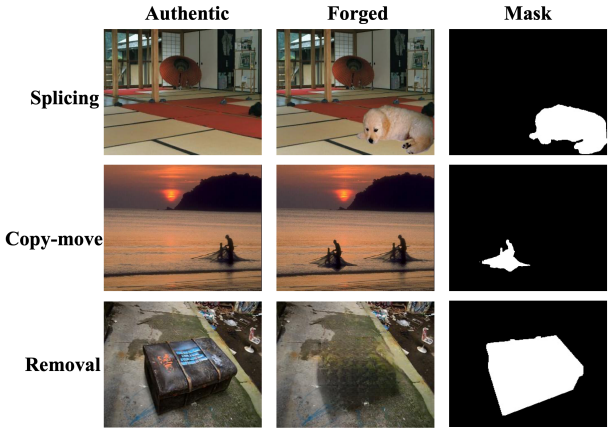

Figure 1: Examples of different types of image manipulations. From left to right: authentic images, manipulated images, and their corresponding masks.

Image manipulation techniques can be classified into three primary categories: splicing, copy-move, and removal. Splicing involves transferring a region from one image to another, potentially disrupting semantic coherence. Copy-move refers to duplicating and pasting a region within the same image, while removal techniques erase specific content and synthesize plausible fillers to maintain visual continuity. As illustrated in Figure 1, increasing public concern over image forgery has driven the development of a wide range of detection methods. Most recent approaches (Dong et al. 2022; Guillaro et al. 2023; Kwon et al. 2022; Wang et al. 2022; Li et al. 2024a) utilize noise and frequency domain information combined with multi-view feature fusion to enhance localization performance. Despite these advances, the insufficient exploration of intrinsic discrepancies between manipulated and authentic regions at multiple levels significantly constrains the accuracy and robustness when confronted with modern manipulation techniques.

At the macro level, we focus on global or contextual inconsistencies. Existing methods often fail to fully cap-

*Corresponding author: Chunjie Ma machunjie@qlu.edu.cn

ture discrepancies introduced by object-level manipulation. For example, Zhu et al. (2025) found that over 80 percent of tampered images in the CASIA dataset involve object modifications, with varying semantic impacts depending on the manipulated object's role in the scene. Although some approaches adopt Transformer-based models (Dosovitskiy et al. 2020; Liu et al. 2021) to model token-level relationships, the grid-based tokenization process commonly used in these models disrupts object integrity. When a manipulated region is divided into multiple tokens, its internal coherence is lost, making it difficult to extract meaningful object-level semantics from RGB images.

At the micro level, we focus on pixel-level artifact differences. Existing methods struggle to characterize discrepancies in local detail. While CNN-based methods (Long, Shelhamer, and Darrell 2015; Chen et al. 2017; Dong et al. 2022) aggregate local neighborhood information through receptive fields, they mainly focus on feature aggregation and weighting, which limits their ability to detect fine-grained structural differences. As a result, subtle manipulation cues, particularly those affecting local texture or edge consistency, are often overlooked.

Inspired by the insights above, we observe that existing methods fail to deeply model the intrinsic discrepancies between manipulated and authentic regions, leading to suboptimal performance. As noted in (Xu et al. 2022), amplifying discrepancies is critical for distinguishing anomalous regions. What matters is not merely what we see, but how we interpret inconsistencies. We argue that effectively understanding and amplifying such discrepancies is critical for precise manipulation localization. Based on this insight, we propose the Focus Region Discrepancy Network (FRD-Net), a novel framework designed to progressively enhance the discrepancies between manipulated and authentic regions from both macro and micro perspectives. For macro-level, we introduce an iterative clustering module that groups features based on similarity, revealing semantic correlations and generating two distinct cluster centers. Each pixel feature is then reassigned according to its similarity to these centers, and the feature representations are optimized to amplify discrepancies in the latent feature space. For micro-level, we design a differential progressive module based on difference convolution. Specifically, angular differences are computed to identify subtle structural discrepancies within local neighborhoods, this information is then embedded into a central difference convolution as prior knowledge, improving the model's ability to detect discrepancies between the center pixel and its surrounding detail information. Importantly, the macro-level module guides the micro-level module by providing global semantic cues, while the micro-level information helps refine regional segmentation and enhance representation quality. This mutual reinforcement leads to more accurate and robust localization of manipulated regions. The main contributions of this paper are summarized as follows:

- We propose FRD-Net, a novel Focus Region Discrepancy Network that utilizes and amplifies the intrinsic discrepancies between manipulated and authentic regions as key cues for accurate image manipulation localization.
- We propose two novel modules to amplify discrepancies between manipulated and authentic regions: an Iterative Clustering Module (ICM) that learns macro-level inconsistencies via similarity-based clustering with interactions between cluster centers and features, and a Differential Progressive Module (DPM) that captures micro-level inconsistencies through angular-aware perception and enhancement via central difference convolution. Their mutual reinforcement enables accurate and robust manipulation localization.
- Extensive experiments on five public image manipulation localization datasets demonstrate that FRD-Net achieves superior performance and robustness compared to state-of-the-art methods, while maintaining low computational complexity.

Related Work

Image Manipulation Localization

Image manipulation localization is a pivotal task in digital image forensics. With the rapid development of deep learning, recent methods (Dong et al. 2022; Kwon et al. 2022; Wang et al. 2022; Guillaro et al. 2023) have sought to enhance generalization by integrating deep features with handcrafted cues. For example, CAT-Net (Kwon et al. 2022) combines frequency-domain and RGB features, while MVSS-Net (Dong et al. 2022) fuses edge, multi-scale, and noise features to provide multi-perspective information for forgery detection and localization. Although frequency and noise features are effective in detecting splicing forgeries, their performance on copy-move manipulation remains suboptimal. To improve robustness, TruFor (Guillaro et al. 2023) applies contrastive learning to retrain image filters, capturing intrinsic noise discrepancies across camera models. While it achieves strong localization performance, its heavy dependence on camera diversity limits its generalization. Mesoscopic Insights (Zhu et al. 2025) advances localization by jointly analyzing high- and low-frequency components with a dynamic weighting mechanism. Meanwhile, SparseViT (Su et al. 2025) employs an interleaved attention design to extract non-semantic cues for more effective forgery detection. Despite these advances, most existing methods still struggle to capture the internal structure of manipulated objects and their contextual relationships. To address this issue, UnionFormer (Li et al. 2024a) and ObjectFormer (Wang et al. 2022) introduce learnable tokens to explicitly model interactions among manipulated regions.

Despite great progress, their reliance on global attention mechanisms can impair the structural coherence within tampered areas, limiting their ability to represent intrinsic correlations accurately.

Cluster in Image Processing

Clustering algorithms have a long-standing role in image processing, with early approaches grouping pixels based on visual similarity. For example, SuperPixel (Wang et al. 2017) aggregates similar pixels to segment an image into coherent regions through clustering. However, such methods

often suffer from high computational costs. SLIC (Achanta et al. 2012) addresses this issue by restricting clustering operations to local neighborhoods and initializing K-means centers uniformly, thereby accelerating both convergence and computation. Traditional clustering techniques rely heavily on raw data representations and often perform poorly when dealing with high-dimensional or complex feature spaces. With the advent of deep learning, clustering methods have seen notable progress. CLUSTSEG (Liang et al. 2023) iteratively updates cluster centers and feature tokens through cross-attention mechanisms, while CoC (Ma et al. 2023b) eliminates both convolution and self-attention entirely, instead using clustering for deep feature extraction. FEC (Chen et al. 2024) alternates between grouping pixels into clusters to obtain abstract representations and refining deep features based on these representations, achieving impressive performance.

Recently, Li et al. (2024b) show that clustering can amplify fine-grained image discrepancies, yet this strategy is still rarely applied to manipulation localization. Early attempts such as Focal (Wu, Chen, and Zhou 2023) combine clustering with contrastive learning, but the full potential of clustering-based discrepancy modeling for localization remains largely untapped.

The Proposed Approach

Despite the remarkable performance of existing image manipulation localization methods, these methods often lack a deeper modeling of the intrinsic differences between manipulated and authentic regions, most prior works detect forgeries by analyzing artifacts from multiple perspectives, such as noise residuals or frequency-domain patterns. Although effective in certain cases, these methods often fail to explicitly model semantic and structural inconsistencies, leading to suboptimal localization performance. To overcome these limitations, we propose the Focus Region Discrepancy Network (FRD-Net), an efficient framework that amplifies both macro-level and micro-level discrepancies to achieve more accurate manipulation localization.

In this work, we utilize clustering to amplify global semantic discrepancies and introduce differential analysis to emphasize subtle, localized manipulations. We design a novel architecture that explicitly captures and amplifies these discrepancies through a specialized block, as illustrated in Figure 2. Given an input RGB image $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$, a backbone feature extractor G first generates a high-dimensional feature map $\mathbf{F} \in \mathbb{R}^{2C\times h\times w}$, where C is 256. This feature map is then split along the channel dimension into two components, $\mathbf{F}_1$ and $\mathbf{F}_2 \in \mathbb{R}^{C\times h\times w}$, which are used to learn macroscopic and microscopic inconsistencies, respectively.

The features $\mathbf{F}_1$ and $\mathbf{F}_2$ are further refined through our proposed Focus Region Discrepancy (FRD) block, which integrates two key modules: the Iterative Clustering Module (ICM) and the Differential Progressive Module (DPM). The ICM captures macro-level inconsistencies by clustering features into semantically coherent groups, identifying the most divergent cluster centers, and feeding their signals back into the feature space to enhance regional contrast. In parallel, the DPM focuses on micro-level inconsistencies by applying differential convolution operations to progressively enhance local structural variations, thereby maintaining sensitivity to subtle manipulation traces. Together, these components enable FRD-Net to accurately model and localize diverse types of image manipulations.

Iterative Clustering Module (ICM)

To effectively model macro-level semantic inconsistencies, we design an Iterative Clustering Module (ICM) that performs cluster-based forward and backward operations. ICM groups correlated features, identifies representative clusters, and propagates this cluster-aware information back to the feature space for enhanced discrimination.

Forward Cluster Process. To capture the macro semantic discrepancies in images more effectively, we design an iterative clustering process consisting of T rounds in the forward pass. At the beginning of each iteration $t \in \{1, 2, \ldots, T\}$, given the reshaped input feature map $\mathbf{F}_1^t \in \mathbb{R}^{N\times C}$, where N represents the number of feature elements, we have $N = h \times w$ when $t = 1$, we first use a 1×1 convolution to project $\mathbf{F}$ into the key and value spaces: $\mathbf{K}^t, \mathbf{V}^t \in \mathbb{R}^{N\times C}$, it is noteworthy that $\mathbf{K}^t$ is employed to construct the similarity matrix between features and cluster centers, while $\mathbf{V}^t$ performs aggregation based on the computed similarity measures. Specifically, we initialize the cluster centers by applying adaptive average pooling to the current projected features $\mathbf{K}^t$ and $\mathbf{V}^t$:

$$\begin{aligned}\hat{\mathbf{K}}^t &= [C_1^k; \ldots; C_O^k]^t = (\text{ada-pool}(\mathbf{K}^t)) \in \mathbb{R}^{O\times C},\\ \hat{\mathbf{V}}^t &= [C_1^v; \ldots; C_O^v]^t = (\text{ada-pool}(\mathbf{V}^t)) \in \mathbb{R}^{O\times C},\end{aligned} \tag{1}$$

where O is the number of cluster centers,and ada-pool$(\cdot)$ is adaptive average pooling, this strategy enables each round of iteration to keep the center positions consistently aligned with the latest feature distribution, thereby ensuring that the initialized cluster centers are more balanced and representative of the underlying features. After the center initialization, we compute the cosine similarity matrix $\mathbf{M}^t \in \mathbb{R}^{N\times O}$ based on $\mathbf{K}^t$ and $\hat{\mathbf{K}}^t$, and then generate the assignment one-hot matrix $\mathbf{A}^t$ using the cluster center o with the highest similarity:

$$\mathbf{A}_{ij}^t = \begin{cases}1, & \text{if } j = \arg\max_o \mathbf{M}_{io}^t,\\ 0, & \text{otherwise.}\end{cases} \tag{2}$$

Based on $\mathbf{A}^t$, we update the cluster centers $\mathbf{R}^t$ by aggregating the value features:

$$\begin{aligned}\mathbf{R}_o^t &= \frac{\left[C_o^v\right]^t + \sum_{n=1}^N \mathbf{A}_{no}^t\, \mathbf{V}_n^t}{1 + \sum_{n=1}^N \mathbf{A}_{no}^t},\\ \mathbf{F}_1^{t+1} &= \mathbf{R}^t \in \mathbb{R}^{O\times C}\end{aligned} \tag{3}$$

where $\left[C_o^v\right]^t$ denotes the o-th cluster center in $\hat{\mathbf{V}}^t$, and $\mathbf{V}_n^t$ denotes the value feature vector at the n-th location. Repeating this procedure for T rounds enables the model to progressively approach a globally salient semantic partition; ultimately, we only retain the two centers with the largest discrepancy, each corresponding to one of the two most seman-

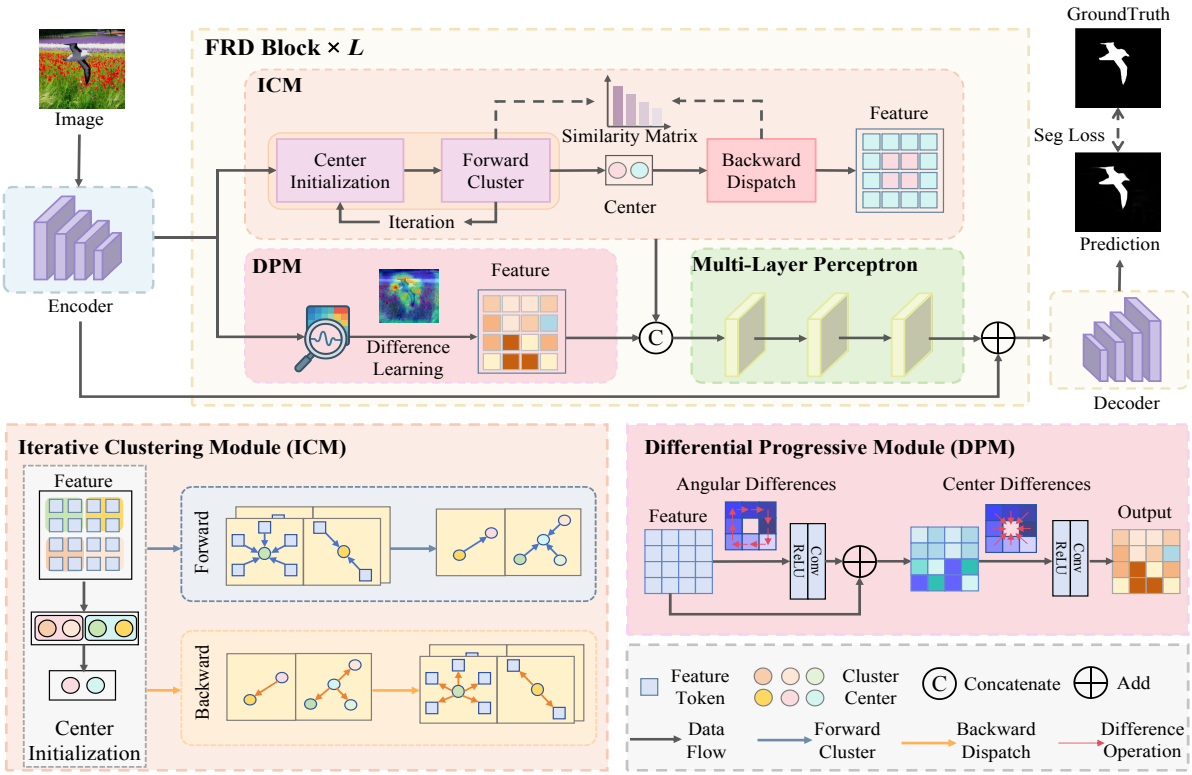

Figure 2: Overview of the proposed framework. It consists of an Encoder, FRD Blocks with Iterative Clustering Module (ICM) and Differential Progressive Module (DPM) for feature refinement, continue L times (where we set to 4), and a Decoder for final prediction. ICM iteratively updates clusters, while DPM enhances features with dynamic difference weights.

tically inconsistent regions in the image. This iterative process drives a progressive shift from localized feature grouping to global semantic understanding, enhancing discrepancy perception and yielding richer features for decoding.

Backward Cluster Process. We have already obtained the representative information of the image feature. Next, we will enhance the understanding of each element regarding this representative information through the method of backpropagation:

$$\mathbf{F}_{\text{macro}} = \sigma(\alpha(\mathbf{M}^1 \cdots \mathbf{M}^{(T)}) + \beta)\boldsymbol{R}^{(T)}, \tag{4}$$

where α and β denote learnable parameters. Through the above reverse clustering process, the ICM achieves a closed loop from local feature aggregation to global semantic perception: the forward stage is responsible for capturing the major conflict centers, and the backward stage is responsible for returning these center information to each pixel position, thereby improving the overall recognition ability of manipulation.

Differential Progressive Module (DPM)

While the iterative clustering process effectively amplifies the perception of macro-level semantic inconsistencies, it lacks explicit modeling of fine-grained local variations, which are often crucial in detecting subtle forgeries. To address this limitation, inspired by (Su et al. 2021) we propose a Differential Progressive Module (DPM) that focuses on capturing micro-level inconsistencies. Specifically, DPM leverages difference convolution to measure the discrepancy between each pixel and its local neighborhood, thereby enhancing the model's sensitivity to local inconsistency.

Specifically, we employ difference convolution with a 3×3 kernel to perceive local information. Initially, neighborhood variations are captured using Angular Pixel Difference Convolution(APDC), whose operation can be formulated as follows:

$$\mathbf{F}_{\text{APDC}}(p) = \sum_{(x_i,\, x_{i+1})\in\Omega^A} w_i(x_i - x_{i+1}), \mathbf{F}_{\text{APDC}}(p) \in \mathbb{R}^C, \tag{5}$$

where p denotes the spatial location of elements in feature $\mathbf{F}_2$, and Ω^A represents the eight token pairs formed by arranging the 3×3 neighborhood feature vectors around position p, in a clockwise manner. This approach effectively captures neighborhood variations, and angular difference operations explicitly detect subtle edge discontinuities along circular orientations, thus better exploring artifacts resulting from rotation, tilt, or other local deformations. Subsequently, we add these resulting features back to the original features, aiming to retain sensitivity to artifacts such as rotations and tilts. Afterwards, we utilize a Central Pixel Difference Convolution(CPDC) to perceive local discrepancies along horizontal, vertical, or diagonal orientations. The CPDC opera-

tion can be expressed as follows:

$$\begin{aligned}
&\tilde{\mathbf{F}} = \mathbf{F}_2 + conv(ReLU(\mathbf{F}_{\text{APDC}})),\\
&\tilde{\mathbf{F}}_{\text{CPDC}}(p) = \sum_{x_i \in \Omega^C} w_i(x_i - x_p), \tilde{\mathbf{F}}_{\text{CPDC}}(p) \in \mathbb{R}^C,\\
&\mathbf{F}_{\text{micro}} = conv(ReLU(\tilde{\mathbf{F}}_{\text{CPDC}})),
\end{aligned} \tag{6}$$

Here, $\tilde{\mathbf{F}}$ represents the features integrated with surrounding angular discrepancies, Ω^C denotes all feature vectors within the 3×3 neighborhood around position p, and x_p denotes the center within the neighborhood, and x_i represents all adjacent pixel tokens. Through this mechanism, the network effectively learns discriminative differential features while maintaining robustness against perturbations from rotations and similar distortions. Subsequently, we concatenate features $\mathbf{F}_{\text{micro}}$ and $\mathbf{F}_{\text{macro}}$ and apply a simple MLP operation with residual connections to obtain refined representations.

$$\mathbf{F}_{\text{final}} = \mathbf{F} + \text{MLP}\left(\text{Concat}(\mathbf{F}_{\text{micro}}, \mathbf{F}_{\text{macro}})\right). \tag{7}$$

We have thus completed the full operation of the FRD block. The optimized representation $\mathbf{F}_{\text{final}}$ is then fed into a deconvolution-based decoder (Long, Shelhamer, and Darrell 2015) layer to generate the final predicted mask for image manipulation localization.

Loss Function

Manipulated areas tend to have irregular shapes, jagged edges, and inconsistent textures, making localization difficult. We mitigate this by introducing a joint region and boundary loss that leverages morphological edge generation from (Ma et al. 2023a), adding no extra branches and keeping the model compact and general.

Specifically, given the binary ground-truth mask M_{gt}, we apply morphological operations such as dilation and erosion to extract a corresponding edge map $M^{\star}_{gt}$. This edge map serves as an auxiliary supervision signal to encourage the model to pay more attention to the spatial structure and boundaries of the tampered regions. Instead of the commonly used binary cross-entropy loss, we employ dice loss (Dong et al. 2022) for both the mask and edge supervision, which is particularly effective in handling class imbalance between manipulated and pristine regions. The total loss is formulated as:

$$\mathcal{L} = \mathcal{L}^{\text{seg}}_{\text{Dice}}(M_{gt}, M_{pred}) + \lambda \mathcal{L}^{\text{edge}}_{\text{Dice}}(M^{\star}_{gt}, M^{\star}_{pred}), \tag{8}$$

where M_{pred} and $M^{\star}_{pred}$ denote the predicted segmentation mask and its corresponding edge map, respectively. The dice loss is defined as:

$$\mathcal{L}_{\text{Dice}}(G, P) = 1 - \frac{2\sum_i p_i g_i}{\sum_i p_i + \sum_i g_i}, \tag{9}$$

where $G = \{g_i\}$ and $P = \{p_i\}$ represent the ground-truth and predicted binary maps. The weighting factor λ balances the contributions of the segmentation and edge terms, and is empirically set to 1 in our experiments.

Experiments and Discussion

We conduct extensive experiments under fair comparative settings. This section covers: (1) datasets, (2) experimental competitors, (3) experimental settings, and (4) performance evaluations and comparisons across the five public datasets.

Datasets

The training data comprises CASIAv2 (Dong, Wang, and Tan 2013), IMD2020 (Novozamsky, Mahdian, and Saic 2020), FantasticReality (Kniaz, Knyaz, and Remondino 2019), and a custom dataset by (Kwon et al. 2022) that features splicing and copy-move manipulations using COCO (Lin et al. 2014), RAISE (Dang-Nguyen et al. 2015), and COCO masks. For evaluation, we use five widely adopted benchmarks: CASIAv1 (Dong et al. 2022), COVER (Wen et al. 2016), Columbia (Hsu and Chang 2006), COCOGlide (Guillaro et al. 2023), and NIST16 (Guan et al. 2019).

Competitors and Evaluation Metrics

We compare our FRD-Net with recent state-of-the-art methods, including ManTra-Net (Wu, AbdAlmageed, and Natarajan 2019), SPAN (Hu et al. 2020), IF-OSN (Wu et al. 2022), MVSS-Net (Dong et al. 2022), PSCC-Net (Liu et al. 2022), CAT-Net (Kwon et al. 2022), TruFor (Guillaro et al. 2023), UnionFormer (Li et al. 2024a), Mesorch (Zhu et al. 2025), and SparseViT (Su et al. 2025). Following CAT-Net (Kwon et al. 2022; Guillaro et al. 2023), we adopt its evaluation protocol for fair comparison, reporting the best F1 scores from the literature. Pixel-wise F1 scores are computed via a permuted comparison strategy, using both a fixed threshold (0.5) and an optimal threshold. During training, balanced sampling is applied across datasets. Additionally, we performed a comparison using the IMDLBenCo method (Ma et al. 2024), further details can be found in the Appendix.

Implementation Details

We employ the ImageNet-pretrained Segformer-b2 (Xie et al. 2021) as the baseline, set the number of iterations T to 3, and configure the cluster centers with the number of 64, 16, and 2. All images are resized to 512 × 512, and the experiments run on two A100 GPUs. We use the Adam optimizer with an initial learning rate of 0.0005. The model is trained for 100 epochs using a batch size of 16. Following (Dong et al. 2022), we apply common data augmentations, such as flipping, blurring, compression, and rudimentary manipulations achieved by cropping and pasting square regions or using OpenCV's built-in inpainting functions.

Performance Evaluations and Comparisons

Localization Performance Table 1 reports localization results, highlights the best model in bold and the runner-up with underline across all datasets. FRD-Net, evaluated with the optimal and a fixed threshold, consistently secures either first- or second-place F1 scores, markedly surpassing UnionFormer on Columbia, CASIAV1, NIST16, and COCOGlide while matching it on COVER. These gains confirm FRD-Net's ability to localize manipulations accurately without

Method	Optimal Threshold F1						Fixed Threshold F1					
	Columbia	COVER	CASIAv1	NIST16	COCOGlide	avg	Columbia	COVER	CASIAv1	NIST16	COCOGlide	avg
Mantra-Net	0.650	0.486	0.320	0.225	0.673	0.471	0.508	0.317	0.180	0.172	0.516	0.339
SPAN	0.873	0.428	0.169	0.363	0.350	0.437	0.759	0.235	0.112	0.228	0.298	0.326
IF-OSN	0.836	0.472	0.676	0.449	0.589	0.604	0.753	0.304	0.553	0.330	0.428	0.473
MVSS-Net	0.781	0.659	0.650	0.372	0.642	0.621	0.729	0.514	0.528	0.320	0.486	0.515
PSCC-Net	0.760	0.615	0.670	0.210	0.685	0.588	0.604	0.473	0.520	0.113	0.515	0.445
CAT-Net2	0.923	0.582	0.852	0.417	0.603	0.675	0.859	0.381	0.752	0.308	0.434	0.547
TruFor	0.914	**0.735**	0.822	0.470	0.720	0.732	0.859	0.600	0.737	0.399	0.523	0.623
UnionFormer	0.925	0.720	0.863	0.489	0.742	0.747	0.861	0.592	0.760	**0.413**	0.536	0.632
SparseViT*	0.980	0.703	0.839	0.514	0.667	0.740	0.955	0.531	0.747	0.318	0.497	0.609
Mesorch*	0.991	0.729	0.894	**0.559**	0.724	0.779	0.936	0.481	0.744	0.357	0.482	0.600
FRD-Net	**0.994**	0.717	**0.903**	0.536	**0.762**	**0.782**	**0.973**	**0.601**	**0.829**	0.412	**0.626**	**0.688**

Table 1: Permute-F1 performance comparison using both optimal and fixed thresholds under CAT-based protocol, * denotes methods retrained under the same setting for fair comparison.

heavy threshold tuning. Figure 4 further shows that FRD-Net captures both global structure and fine-grained traces, producing cleaner and more precise manipulation masks than competing methods.

Computational Efficiency As shown in Table 2, we also evaluate the efficiency of FRD-Net, all experiments are conducted on a hardware platform equipped with an Intel(R) Xeon(R) Gold 6226R CPU @ 2.90 GHz, an NVIDIA Tesla A100 GPU. In our experiments, the number of parameters, GFlops is employed as the evaluative criteria. We find that the number of parameters and GFlops is the least among all SOTA methods; thus, our proposed FRD-Net is simple yet efficient, while maintaining excellent performance.

Robustness Evaluation To comprehensively assess robustness, we conduct experiments on the CASIAv1 dataset under various common perturbations. Specifically, we introduce four types of distortions: (1) Gaussian noise, (2) Gaussian blur, (3) JPEG compression, and (4) Gamma. We simulate real-world degradations that often occur during image acquisition, editing, or transmission. As summarized in Figure 3, our model consistently achieves the best or second-best performance across all perturbation settings and retains strong detection capability even under severe distortion. It ranks second only to Mesorch under Gaussian blurring and Gaussian noise while surpassing every other competitor. Compared with Mesorch, it uses far fewer parameters and reaches faster inference speed, highlighting a lightweight and robust design. These results demonstrate the method's effectiveness and generalizability for noisy or degraded images.

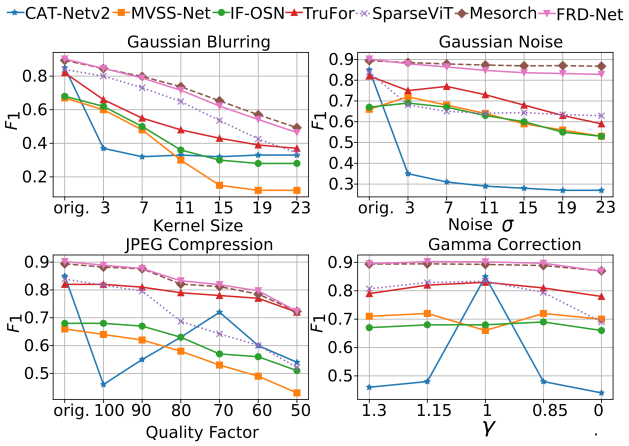

Figure 3: Comparison of optimal Permute-F1 scores for model robustness under various perturbations.

Ablation Study

To verify the effectiveness of each proposed component in our framework, we conduct ablation experiments on three benchmark datasets: CASIAv1, COVER, and COCOGlide. We perform ablations on CASIAv1, COVER, and COCOGlide: beginning with a Segformer-b2 baseline, we toggle ICM and DPM and report Permute-F1 (threshold 0.5) in Table 3.

Effectiveness of ICM To assess ICM's effectiveness, we set the clustering iterations to t = 1, reducing ICM to single-step semantic clustering. Even so, F1 scores rise by an average of 1.8 % across the three benchmarks, showing that semantic clustering alone already improves the separation between manipulated and pristine regions. We then enable full iterative clustering (+ICM), which further boosts F1 by 3.4 %, 4.2 %, and 2.4 % on CASIAv1, COVER, and COCOGlide, respectively, yielding an average F1 of 0.668 and the best accuracy on COCOGlide. These results demonstrate that the iterative mechanism progressively amplifies semantic inconsistencies and markedly enhances manipulation localization across diverse scenarios.

Effectiveness of DPM To assess the progressive refinement strategy, we remove it from DPM, using only single-stage central-difference convolution. This variant slightly improves CASIAv1 but fails on COVER and COCOGlide, giving an average F1 of 0.6004—below the Baseline (0.6176). Without multi-step refinement, the module cannot capture diverse, subtle manipulations. Restoring the progressive strategy lifts F1 to 0.7631, 0.5701, and 0.5826 on CASIAv1, COVER, and COCOGlide, respectively, raising the average to 0.6386 (+2.1 % over Baseline) with the largest gain on COVER (+5.2 %). These results show that progres-

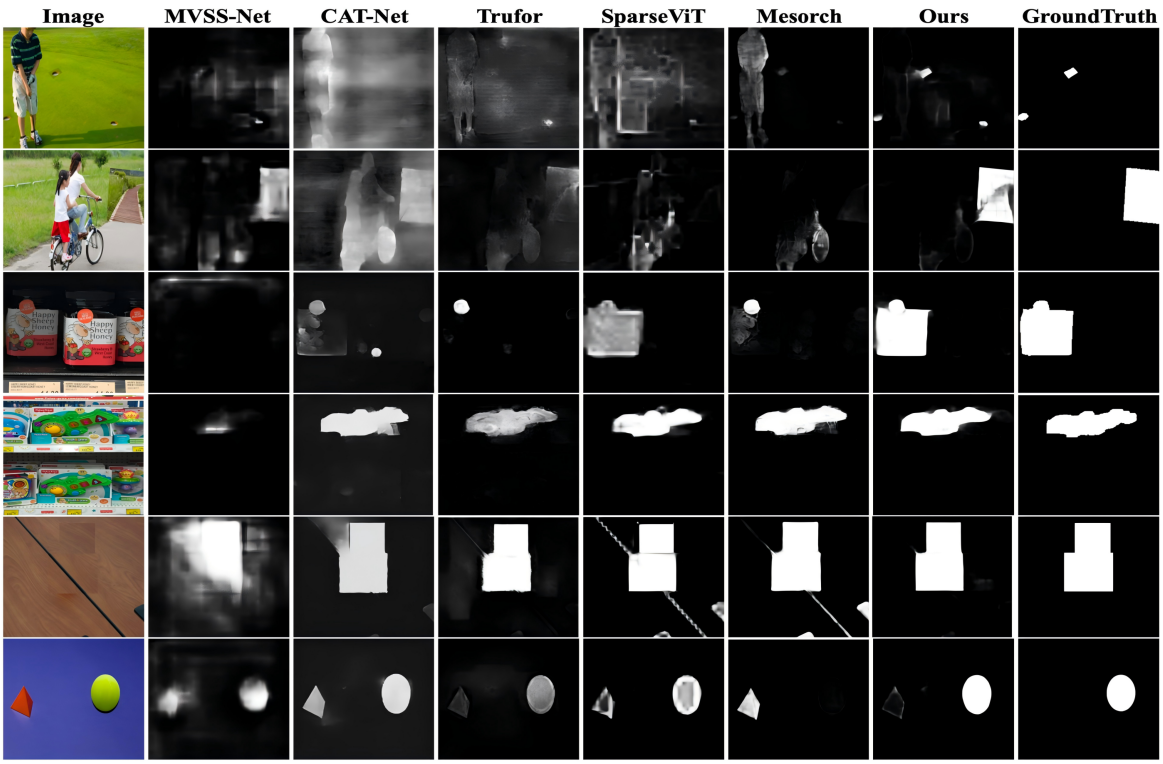

Figure 4: Qualitative results of different image manipulation localization algorithms. The first column shows manipulation images on different datasets. The second to seventh columns indicate the final segmentation prediction results of MVSS-Net, CAT-Net, Trufor, SparseViT, Mesorch, and FRD-Net (Ours), respectively, the last column is the GroundTruth.

Method	Size	Parameters	FLOPs
MVSS-Net	512 × 512	142.7M	163.5G
TruFor	512 × 512	68.7M	221.8G
UnionFormer	512 × 512	210M	-
SparseViT	512 × 512	50.3M	41.5G
Mesorch	512 × 512	82.7M	145.9G
Ours	512 × 512	**39M**	**25.6G**

Table 2: Computational efficiency statistics where the number of parameters and GFlops are assessed.

sive difference modeling better tracks multi-scale manipulations and generalizes across complex scenarios.

Combined Effectiveness of ICM and DPM Combining ICM with DPM yields the strongest ablation result—an average F1 of 0.6851 and the top scores on CASIAv1 and COVER—by uniting enhanced low-level traces with semantically aware clustering. COCOGlide shows a small dip because its diffusion-generated artifacts differ from manual edits, yet semantic cues still lift the combo to second place. Overall, ablation confirms every module's value: feature clustering offers the largest solo boost, inverse guidance amplifies it, and difference enhancement supplies complementary local cues; together they deliver substantial, consistent gains across all datasets.

Method	CASIAv1	COVER	COCOGlide	AVG
Segformer-b2	0.7532	0.5181	0.5789	0.6176
+ ICM(w/o iteration)	0.7690	0.5277	0.6082	0.6350
+ ICM	<u>0.8030</u>	0.5690	**0.6320**	<u>0.6680</u>
+ DPM(w/o progress)	0.7644	0.5047	0.5322	0.6004
+ DPM	0.7631	<u>0.5701</u>	0.5826	0.6386
FRD-Net	**0.8287**	**0.6007**	<u>0.6259</u>	**0.6851**

Table 3: Ablation study of each component on CASIAv1, COVER, and COCOGlide datasets. We report Permute-F1 scores, along with their average, the best results are shown in bold, and the second best results are underlined.

Conclusion

This paper presents FRD-Net, a novel framework for image manipulation localization that jointly amplifies inconsistencies at macro and micro scales. The proposed ICM captures semantic-level differences via iterative clustering, while the DPM enhances local structural cues using difference convolution. Extensive experiments demonstrate that FRD-Net achieves superior performance. In the future, we plan to explore frequency and noise domain cues to further enhance the modeling of both macro- and micro-level inconsistencies for more accurate tampering localization.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No.62372325, No.62402255, No.62502344, No.U25A20444), Natural Science Foundation of Tianjin Municipality (No.23JCZDJC00280), Shandong Provincial Natural Science Foundation (No.ZR2024QF020), Shandong Province National Talents Supporting Program (No.2023GJJLJRC-070), Shandong project towards the integration of education and industry (No.801822020100000024), Young Talent of Lifting engineering for Science and Technology in Shandong (No. SDAST2024QTB001), Shandong Project towards the Integration of Education and Industry (No.2024ZDZX11), Wenzhou major science and technology innovation project(Grant No. ZG2022011).

References

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2012. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11): 2274–2282.

Chen, G.; Li, X.; Yang, Y.; and Wang, W. 2024. Neural Clustering Based Visual Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5714–5725.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4): 834–848.

Dang-Nguyen, D.-T.; Pasquini, C.; Conotter, V.; and Boato, G. 2015. Raise: A Raw Images Dataset for Digital Image Forensics. In *Proceedings of the 6th ACM multimedia systems conference*, 219–224.

Dong, C.; Chen, X.; Hu, R.; Cao, J.; and Li, X. 2022. MVSS-Net: Multi-View Multi-Scale Supervised Networks for Image Manipulation Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3): 3539–3553.

Dong, J.; Wang, W.; and Tan, T. 2013. CASIA Image Tampering Detection Evaluation Database. In *2013 IEEE China Summit and International Conference on Signal and Information Processing*, 422–426. IEEE.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, 1–21.

Guan, H.; Kozak, M.; Robertson, E.; Lee, Y.; Yates, A. N.; Delgado, A.; Zhou, D.; Kheyrkhah, T.; Smith, J.; and Fiscus, J. 2019. MFC Datasets: Large-Scale Benchmark Datasets for Media Forensic Challenge Evaluation. In *2019 IEEE Winter Applications of Computer Vision Workshops*, 63–72. IEEE.

Guillaro, F.; Cozzolino, D.; Sud, A.; Dufour, N.; and Verdoliva, L. 2023. TruFor: Leveraging All-Round Clues for Trustworthy Image Forgery Detection and Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20606–20615.

Hsu, J.; and Chang, S. 2006. Columbia Uncompressed Image Splicing Detection Evaluation Dataset. *Columbia DVMM Research Lab*.

Hu, X.; Zhang, Z.; Jiang, Z.; Chaudhuri, S.; Yang, Z.; and Nevatia, R. 2020. SPAN: Spatial Pyramid Attention Network for Image Manipulation Localization. In *European Conference on Computer Vision*, 312–328. Springer.

Kniaz, V. V.; Knyaz, V.; and Remondino, F. 2019. The Point Where Reality Meets Fantasy: Mixed Adversarial Generators for Image Splice Detection. *Advances in Neural Information Processing Systems*, 32: 215–226.

Kwon, M.-J.; Nam, S.-H.; Yu, I.-J.; Lee, H.-K.; and Kim, C. 2022. Learning JPEG Compression Artifacts for Image Manipulation Detection and Localization. *International Journal of Computer Vision*, 130(8): 1875–1895.

Li, S.; Ma, W.; Guo, J.; Xu, S.; Li, B.; and Zhang, X. 2024a. UnionFormer: Unified-Learning Transformer with Multi-View Representation for Image Manipulation Detection and Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12523–12533.

Li, X.; Guo, Z.; Zhu, R.; Ma, Z.; Guo, J.; and Xue, J.-H. 2024b. A simple scheme to amplify inter-class discrepancy for improving few-shot fine-grained image classification. *Pattern Recognition*, 156: 110736.

Liang, J. C.; Zhou, T.; Liu, D.; and Wang, W. 2023. CLUSTSEG: Clustering for Universal Segmentation. In *International Conference on Machine Learning*, 20787–20809. PMLR.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, 740–755. Springer.

Liu, X.; Liu, Y.; Chen, J.; and Liu, X. 2022. PSCC-Net: Progressive Spatio-Channel Correlation Network for Image Manipulation Detection and Localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 7505–7517.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10012–10022.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully Convolutional Networks for Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Ma, X.; Du, B.; Jiang, Z.; Hammadi, A. Y. A.; and Zhou, J. 2023a. IML-ViT: Benchmarking Image Manipulation Localization by Vision Transformer. *arXiv preprint arXiv:2307.14863*.

Ma, X.; Zhou, Y.; Wang, H.; Qin, C.; Sun, B.; Liu, C.; and Fu, Y. 2023b. Image as Set of Points. In *International Conference on Learning Representations*, 1–18.

Ma, X.; Zhu, X.; Su, L.; Du, B.; Jiang, Z.; Tong, B.; Lei, Z.; Yang, X.; Pun, C.-M.; Lv, J.; and Zhou, J. 2024. IMDL-BenCo: A Comprehensive Benchmark and Codebase for Image Manipulation Detection & Localization. In *Advances in Neural Information Processing Systems*, volume 37, 134591–134613.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*, 16784–16804. PMLR.

Novozamsky, A.; Mahdian, B.; and Saic, S. 2020. IMD2020: A Large-Scale Annotated Dataset Tailored for Detecting Manipulated Images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, 71–80.

Su, L.; Ma, X.; Zhu, X.; Niu, C.; Lei, Z.; and Zhou, J.-Z. 2025. Can We Get Rid of Handcrafted Feature Extractors? SparseViT: Nonsemantics-Centered, Parameter-Efficient Image Manipulation Localization Through Spare-Coding Transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 7024–7032.

Su, Z.; Liu, W.; Yu, Z.; Hu, D.; Liao, Q.; Tian, Q.; Pietikäinen, M.; and Liu, L. 2021. Pixel Difference Networks for Efficient Edge Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5117–5127.

Sun, Z.; Fang, H.; Cao, J.; Zhao, X.; and Wang, D. 2024. Rethinking Image Editing Detection in the Era of Generative AI Revolution. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 3538–3547.

Verdoliva, L. 2020. Media Forensics and DeepFakes: An Overview. *IEEE Journal of Selected Topics in Signal Processing*, 14(5): 910–932.

Wang, J.; Wu, Z.; Chen, J.; Han, X.; Shrivastava, A.; Lim, S.-N.; and Jiang, Y.-G. 2022. ObjectFormer for Image Manipulation Detection and Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2364–2373.

Wang, M.; Liu, X.; Gao, Y.; Ma, X.; and Soomro, N. Q. 2017. Superpixel segmentation: A benchmark. *Signal Processing: Image Communication*, 56: 28–39.

Wen, B.; Zhu, Y.; Subramanian, R.; Ng, T.-T.; Shen, X.; and Winkler, S. 2016. COVERAGE—A novel database for copy-move forgery detection. In *2016 IEEE International Conference on Image Processing (ICIP)*, 161–165. IEEE.

Wu, H.; Chen, Y.; and Zhou, J. 2023. Rethinking Image Forgery Detection via Contrastive Learning and Unsupervised Clustering. *arXiv preprint arXiv:2308.09307*.

Wu, H.; Zhou, J.; Tian, J.; and Liu, J. 2022. Robust Image Forgery Detection over Online Social Network Shared Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13440–13449.

Wu, Y.; AbdAlmageed, W.; and Natarajan, P. 2019. ManTra-Net: Manipulation Tracing Network for Detection and Localization of Image Forgeries With Anomalous Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9543–9552.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090.

Xu, J.; Wu, H.; Wang, J.; and Long, M. 2022. Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy. In *International Conference on Learning Representations*, 1–20.

Zhu, X.; Ma, X.; Su, L.; Jiang, Z.; Du, B.; Wang, X.; Lei, Z.; Feng, W.; Pun, C.-M.; and Zhou, J.-Z. 2025. Mesoscopic Insights: Orchestrating Multi-Scale & Hybrid Architecture for Image Manipulation Localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 11022–11030.