Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Modular Image Manipulation Localization Framework using a Dual-Stream Classifier and Conditional Diffusion Models

Zohaib Hamdule

24cse2014@nitgoa.ac.in

National Institute of Technology Goa https://orcid.org/0009-0007-7490-225X

Venkatanareshbabu Kuppili

National Institute of Technology Goa

Research Article

Keywords: Image Manipulation Localization, Conditional Diffusion, Dual-Stream Classifier, Image Forensics

Posted Date: July 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10474507/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Version of Record: A version of this preprint was published at Multimedia Tools and Applications on September 24th, 2026. See the published version at https://doi.org/10.1007/s11042-026-21940-0.

A Modular Image Manipulation Localization Framework using a Dual-Stream Classifier and Conditional Diffusion Models

Mohammad Zohaib Hamdule[1*†] and Venkatanareshbabu Kuppili[1†]

[1*]Department of Computer Science and Engineering, National Institute of Technology Goa, Cuncolim, 403703, Goa, India.

*Corresponding author(s). E-mail(s): 24CSE2014@nitgoa.ac.in;
Contributing authors: venkatanaresh@nitgoa.ac.in;
†These authors contributed equally to this work.

Abstract

Image manipulation localization (IML), the task of accurately identifying the manipulated regions within an image, remains a significant challenge. Conventional deep learning methods often struggle with generalization and adapting to new manipulation schemes. To address these limitations, a novel two-stage IML system is proposed. The first stage employs a Dual-Stream Manipulation Classifier, fusing features from the standard RGB domain with low-level forensic noise artifacts extracted via Steganalysis Rich Model (SRM) filters using a ResNet like 4-stage backbone, to accurately classify the input image into one of four common manipulation types: Copy-Move, Splicing, Removal (Inpainting), or Enhancement. In the second stage, the classified image is directed to a specialized Conditional Diffusion Model (CDM), which treats the localization task as an image-to-mask generation problem. These CDMs are conditioned on the input image, and are trained with a modified loss function incorporating both Mean Squared Error (MSE) and Intersection over Union (IoU) to better handle smaller manipulation masks and encourage better spatial structure. Evaluated on the DF2023 dataset, the manipulation classifier achieves an accuracy of 89% and localization CDMs achieve average IoU of 0.70 and F1-Score of 0.77. The system demonstrates competitive performance on benchmark datasets, confirming the efficacy of a modular, generative approach for tackling the complexity of image forgeries.

Keywords: Image Manipulation Localization, Conditional Diffusion, Dual-Stream Classifier, Image Forensics

1 Introduction

In the digital age today, image manipulation has become increasingly prevalent due to accessible and easy to use editing tools and ubiquity of social media. This results in a large volume of manipulated images shared and circulated online daily. Humans often struggle to identify fake images. In fact, they fail to identify two thirds of the manipulated images [1]. Even when they are aware an image is fake, they fail to locate the specific alteration 45% of the times [1]. People surfing the internet, especially the elderly are victims of misinformation spread using these manipulated images. Fake and inciting news is 70% more likely to be retweeted and spreads six times faster than factual information [2]. Fake images also impact the online shopping and ecommerce industry, where according to a study, 75% of online shoppers completely rely on product images to make purchasing decisions [3]. It hinders scientific research which relies on accurate image data and plots to confirm research findings. The Journal of Cell Biology reported that one in every ten accepted articles had inappropriate manipulations of image data [4]. Integrity of image data is of paramount importance in legal cases where digital evidence, including images, plays a role in 90% of criminal cases [5]. Image manipulation or image forgery includes subtle enhancements, adding, duplicating, removing, or moving parts of the image, to make it more appealing or to mislead the audience. Altered images have been used to defame entities, incite violence, and spread misinformation and hatred.

Image manipulation detection is the task of identifying whether an image has been manipulated or not. Whereas, Image manipulation localization is the process of locating / highlighting the exact manipulations made. The following sections describe the common types of image manipulations found, metrics used to measure the performance of image manipulation detection and localization (IMDL) and a detailed review of recent work done in these tasks.

2 Types of Image Manipulation

Though there are many complex image manipulation schemes possible, four types of manipulations are common due to the minimal effort required to perform. These include Copy-Move, Splicing, Removal or Inpainting and Enhancement [6]. These four types are not exhaustive and an image may contain a combination of these or other manipulations.

Copy-Move forgery involves copying a part of an image and pasting it elsewhere within the same image to duplicate or conceal an object. Whereas, splicing creates a composite image by cutting an object from one image and pasting it into a different image. Removal or Inpainting involves deleting an object from an image and then artificially filling the resulting hole with a plausible, computer-generated background. Finally, enhancement improves an image's visual quality by adjusting properties like brightness, sharpness or contrast, without altering its core content. Figure 1 shows examples of these manipulation types.



Fig. 1 Examples of the discussed manipulation types from common datasets.

3 Performance Metrics

Image manipulation detection is a binary classification problem, therefore, to measure the performance of manipulation detection, standard classification metrics such as Accuracy, Precision, Recall and F1 Score are used. These are defined mathematically as follows:

$$\begin{aligned} \text{Accuracy} &= \frac{TP + TN}{TP + TN + FP + FN} \\ \text{Precision} &= \frac{TP}{TP + FP} \\ \text{Recall} &= \frac{TP}{TP + FN} \\ F_1 &= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \end{aligned} \qquad \begin{aligned} &\text{where:} \\ &TP = \text{True Positives} \\ &TN = \text{True Negatives} \\ &FP = \text{False Positives} \\ &FN = \text{False Negatives} \end{aligned} \tag{1}$$

Manipulation Localization is an image segmentation problem and involves pixel wise comparison between the predicted mask highlighting the manipulations and the ground truth mask. Many such metrics with their advantages/drawbacks are discussed in [7]. Metrics used to evaluate manipulation localization in this study are discussed below.

3.1 Pixel Wise Accuracy

Pixel-wise accuracy measures the proportion of pixels in an image that are correctly classified as either belonging to the manipulated region or the original (unmanipulated)

region. It is calculated by dividing the number of correctly classified ($TP+TN$) pixels by the total number of pixels in the image. This metric's primary drawback is that in cases of sparse manipulations, where only a few pixels are altered, it tends to yield a high value even for an empty predicted mask [7].

3.2 Intersection over Union (IoU) / Jaccard Distance

IoU measures the overlap between predicted segmentation mask and ground-truth mask. It is the ratio of their intersection to their combined area. IoU is a strict metric because, unlike accuracy, it completely ignores true negatives (background correctly classified as background), which often dominate in segmentation tasks. Mathematically, IoU is defined as follows:

$$IoU_{\text{soft}} = \frac{\sum_{i\in\Omega}(y_i \cdot \hat{y}_i)}{\sum_{i\in\Omega}(y_i + \hat{y}_i - y_i \cdot \hat{y}_i)} \quad \textbf{\textit{or}} \quad IoU_{\text{soft}} = \frac{\sum_{i\in\Omega}\min(y_i, \hat{y}_i)}{\sum_{i\in\Omega}\max(y_i, \hat{y}_i)} \tag{2}$$

where:

y_i	is the ground-truth value at pixel i,
$\hat{y}_i$	is the predicted value at pixel i,
$i \in \Omega$	represents all pixels in the union of both masks.

3.3 F1-Score

F1 Score is the ratio of precision (out of predicted pixels, how many are correctly predicted), and recall (out of ground-truth pixels, how many are correctly detected). It is closely related to IoU but gives higher weight to correctly predicted regions when overlap is small. Figure 2. illustrates values of IoU and F1 Score and the extent of overlap between the masks. F1 is expressed mathematically as:

$$F_{1,\text{soft}} = \frac{2\sum_{i\in\Omega} y_i\hat{y}_i}{\sum_{i\in\Omega} y_i + \sum_{i\in\Omega}\hat{y}_i} \quad \textbf{\textit{or}} \quad F_1 = \frac{2 \times \text{IoU}}{1 + \text{IoU}} \tag{3}$$

where:

y_i	is the ground truth pixel value (0 or 1),
$\hat{y}_i$	is the predicted pixel probability (0 to 1),
Ω	is the set of pixels in the union of both masks.

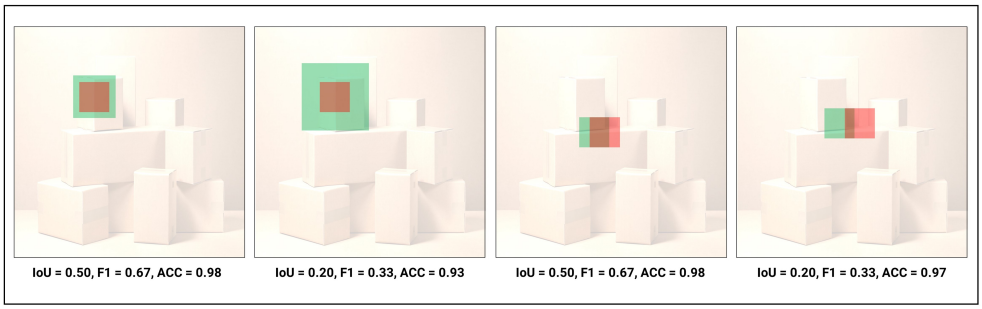


Fig. 2 Intersection over Union (Jaccard Distance) and F1 Score visualizations.

4 Related Work

The following literature review is structured into three main categories based on the methodologies employed: Conventional Techniques, Deep Learning Techniques (like CNNs, LSTMs, and Attention-based models), and Generative Models (such as GANs, VAEs, ViTs, Diffusion models, etc) used for image manipulation detection and localization.

4.1 Conventional Techniques

The foundational era of image manipulation detection was led by conventional techniques, which operate by identifying inconsistencies in statistical or structural properties of an image. These methods rely on the assumption that the manipulation process inevitably alters the inherent artifacts in the image left during capture or encoding of the image. This category of methods can be further broadly grouped based on the type of artifact they exploit. One of the most common avenues of research has been the analysis of JPEG artifacts. The method proposed by Lakovidou et al. [8] targets inconsistencies in the 8x8 pixel grid alignment, a common side effect of splicing a region from a different image. Its primary innovation is its "content aware" nature. The algorithm first segments the image to understand the natural object boundaries, which allows it to intelligently suppress false-positive grid inconsistency alerts that might be triggered by these natural object boundaries. Similarly, Dua et al. [9] also explicit JPEG properties by analyzing the statistical distribution of the block wise Discrete Cosine Transform (DCT) coefficients. This method extracts a feature vector composed of the standard deviation and the count of non-zero AC coefficients for each frequency component. Then it uses a Support Vector Machine (SVM) for final classification of both splicing and copy-move forgeries. Some techniques also exploit the artifacts attributed to the camera sensor. Liu et al. [10] propose a localization method based on Color Filter Array (CFA) pattern. It relies on the fact that a spliced region will lack or have an inconsistent pattern of artifacts left by the interpolation algorithm. It employs a statistical analysis based on pixel value probabilities and local curvature to detect CFA inconsistencies, making it effective against the non linear,

content aware interpolation used in modern cameras. Niu et al. [11] developed a keypoint based method. First, it performs dense keypoint extraction to ensure coverage across the entire image. Second, it describes these keypoints using "complex-valued moment invariants" descriptors that are resilient to geometric transformation like rotation and scaling. Finally, it uses a hierarchical strategy for matching and finding duplicate regions among the massive set of keypoints. The reliance of these conventional techniques on a specific, assumed artifact makes them brittle and unable to generalize to new, or artifact free manipulation.

4.2 Deep Learning Techniques

Deep learning methods learn to identify more general and subtle manipulation traces directly from data, leading to significantly improved robustness and generalizability. A foundational work that applies deep learning to IMDL is the ManTra-Net by Wu et al. [12]. It views the task of forgery localization from the perspective of anomaly detection. It has a two stage design. A CNN feature extractor is pretrained on a self-supervised image manipulation classification task with 385 types of manipulations. A "Local Anomaly Detection Network" (LADN) processes these feature maps to assess how much a local feature deviates from its neighborhood using a Z-score style standardization and a Conv-LSTM [13] based far-to-near sequential analysis. This coarse to local anomaly detection strategy lets ManTra-Net generalize to unseen forgeries. A dominant trend across these deep learning techniques is Multi-Stream fusion. Fusing features extracted through multiple streams (RGB, SRM, YCrCb, DCT, etc) provide a more robust and complete signal. Another common aspect among these techniques is their use of multi-scale processing, where the image is analyzed from coarse to fine levels of detail. The MVSS-Net by Dong et al. [14] fuses features from the RGB image stream with a noise stream derived from Steganalysis Rich Model (SRM) filter. It then uses "multi-scale supervision" and "dual-attention" mechanisms to effectively merge these two views. The MSF-Net by Dou et al. [15] utilizes Multi-Stream Fusion that employs four parallel streams, to capture artifacts in color (YCrCb), content (RGB), compression (DCT) and noise (SRM) fusing them at multiple stages. A similar multi-stream (RGB + SRM) and multi-scale approach is presented by Bai in MSCL-Net [16]. Its primary innovation lies in incorporating Contrastive Learning, the network is explicitly trained to maximize the distance in the feature space between representations of authentic and manipulated regions, thereby enhancing its discriminative power. Integrating sophisticated attention mechanisms to capture global dependencies have proved beneficial in enhancing manipulation localization accuracy. The "Spatial Pyramid Attention Network" (SPAN) by Hu et al. [17] builds a spatial pyramid of feature maps at various scales, allowing the model to capture both coarse and fine-grained cues. Within each scale, local self-attention blocks are applied to model relationships among nearby patches, while a position projection module encodes their spatial layout. This hierarchical attention structure efficiently combines local detail and global context, producing sharper and accurate localization masks across diverse manipulation types and image scales. Further advancing this, the PSCC-Net by Liu et al. [18] introduces two key innovations. First is its progressive bottom-up architecture, starting from a coarse-scale mask and sequentially refining it at finer scales, each stage

conditioned on the previous prediction. Second is its novel "Spatio-Channel Correlation Module" (SCCM), an attention mechanism that captures correlations not only spatially (like SPAN) but also across feature channels. Because the backbone is deliberately lightweight and the progressive mechanism allows early termination, PSCC-Net achieves high-speed performance (processing 1080p images at 50+ frames per second).

4.3 Generative Models in IMDL

The current frontier in IMDL is driven by the use of generative models. This paradigm moves away from just detecting low level artifacts and focuses on understanding the content, semantics, and underlying probability distribution of images. These models are better equipped to handle generative fakes. The ViT-VAE by Chen et al. [19] is an unsupervised method that operates by "run-time learning" from the suspicious image itself. It trains a Variational Autoencoder (VAE) [20] to reconstruct small, authentic patches of the image. Forged regions, being anomalous, are identified by their significantly higher reconstruction error. The model's power is enhanced by using a Vision Transformer (ViT) [21] as the VAE's encoder for a more robust feature representation. The DH-GAN by Liu et al. [22] uses the Generative Adversarial Network's (GAN) [23] generator as a feature extractor to map the image into a more discriminative latent space, where a subsequent model performs localization. The model is trained to detect inconsistencies at multiple levels, such as pixel-level artifacts and patch-level structural relationships. The WSMD-Net by Wang et al. [24] tackles the annotation bottleneck by operating under weak supervision, requiring only image-level labels. It fuses a ViT stream, which captures long-range dependencies, with an "SR-CA stream", which uses a conventional SRM filter and Convolutional Block Attention Modules (CBAM) [25] to focus on weak features in forgery regions allowing it to achieve localization from weak labels. The CLIP-IML by Hou et al. [26] leverages the vast multi-modal (image-text) knowledge of the CLIP [27] model. By using CLIP token features within a multi-scale fusion decoder, this method can detect semantic or stylistic inconsistencies. The InpDiffusion model by Wang et al. [28] introduced diffusion for inpainting (algorithmic removal) localization. It generates the binary mask by progressively refining a random noise map. The generation process is conditioned on semantic and edge feature maps extracted from the image by a "Dual-stream Multi-scale Feature Extractor" (DMFE) into each step of the denoising process to guide the mask's formation. This generative, conditional diffusion framework fundamentally changes the problem from one of pixel-wise discrimination to one of holistic, conditional probability, demonstrating a path forward for the field. Table 1. summarizes the techniques discussed in this review.

5 Conditional Diffusion Models

Diffusion models are a class of generative models that learn data distributions by gradually destroying structure in an image with a forward noising process and then learning to reverse this corruption through a neural network denoiser that predicts the noise added at timestep t. The generation process starts with pure noise, iteratively denoising it to generate a clean sample. Usually, the denoising process is probabilistic and results in unique samples each time. The core assumptions are: (1) complex data

Table 1 Summary of recent Image Manipulation Detection and Localization techniques, arranged chronologically in descending order of year.

Year	Method	Type	Core Idea	Task	Forgery Type
2025	**InpDiffusion** [28]	GenNN	Diffusion + DMFE	Localization	InPainting
2025	**WSMD-Net** [24]	GenNN	Multi Stream(RGB, SRM, Bayer) + Vision Transformer, CBAM	Localization	Mixed
2025	**CLIP-IML** [26]	GenNN	CLIP token, feature pyramid, multi-scale fusion decoder	Localization	Mixed
2025	**MSF-Net** [15]	DNN	Multi Stream (YrCrB, DCT, SRM) Fusion, Multi Stage Attention	Detection + Localization	Mixed
2024	**MSCL-Net** [16]	DNN	Dual stream (RGB + SRM), Self Attention, Contrast learning	Detection + Localization	Mixed
2024	**CFA** [10]	Trad	Statistical analysis of CFA patterns	Localization	Splicing
2024	**DH - GAN** [22]	GenNN	Pixel & Patch level consistency analysis	Localization	Mixed
2023	**ViT - VAE** [19]	GenNN	Anomaly Detection (VAE)	Localization	Mixed
2022	**MVSS-Net** [14]	DNN	Noise view + Edge view + Dual attention	Detection + Localization	Mixed
2022	**PSCC-Net** [18]	DNN	Progressive mechanism + SCCM	Detection + Localization	Mixed
2021	**Keypoint** [11]	Trad	Complex-valued moment invariants keypoint descriptors	Detection + Localization	Copymove
2020	**SPAN** [17]	DNN	Pyramid Self Attention	Localization	Mixed
2020	**DCT** [9]	Trad	SVM on statistical features of DCT blocks	Detection	Copymove, Splicing
2019	**ManTra-Net** [12]	DNN	CNN feature extractor + Anomaly Detection	Detection + Localization	Mixed
2018	**CAGI** [8]	Trad	Detecting misalignment in JPEG compression grids.	Localization	Mixed

Trad : Traditional/Conventional Techniques

DNN : Deep Learning Techniques

GenNN : Generative Models for IMDL

can be mapped to pure Gaussian noise using a fixed Markovian forward diffusion process, and (2) adding Gaussian noise to a Gaussian remains Gaussian, which makes the forward transitions analytically tractable and allows closed-form sampling at any timestep. The forward diffusion process is defined as:

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}\, x_{t-1},\, \beta_t I\right) \tag{4}$$

and by repeatedly applying this, x_t becomes a Gaussian with known parameters. Using the closed-form reparameterization,

$$x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I) \tag{5}$$

The reverse model $p_\theta\,(x_t - 1|x_t)$ is trained to predict the noise ϵ added at each timestep t. The standard denoising objective becomes:

$$\mathcal{L}_{\text{simple}} = \mathbb{E}_{x_0,\,\epsilon,\,t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2\right] \tag{6}$$

Denoising Diffusion Probabilistic Models (DDPM) [29] introduced by Ho et al. established the modern framework, providing high sample quality through a learned reverse diffusion process. Denoising Diffusion Implicit Models (DDIM) [30] by Song et. al. generalized the sampling process to be non-Markovian and deterministic, enabling faster generation.

Conditional Diffusion Models (CDMs) extend this framework by generating outputs guided by an external input (conditioning) [31]. The denoising model takes conditioning along with noisy sample and timestep to predict the noise added. CDMs have become the backbone of many applications including text-to-image generation [32–35], image-to-image translation [36–38], and image segmentation [39, 40]. In image segmentation, CDMs model the conditional distribution of segmentation masks given an image, treating segmentation as an image-to-image generation problem.

6 Proposed Image Manipulation Localization System

In this study, a modular two-stage pipeline for image manipulation localization is proposed. Initially, a manipulation classifier processes the input image to determine the type of forgery present. Based on this classification, the image is then directed to a specialized CDM which generates the precise manipulation mask. This modular design addresses the issue of adaptability in these systems [6], as introducing a new forgery type only requires adding a new CDM and retraining only the classifier, avoiding complete system retraining. It also simplifies training by enabling smaller sub-models to be trained on subsets of the data, while reducing the scope targeted by each model. This modularity is particularly well suited to CDMs because each diffusion model can be specialized to learn the distinct noise and structural characteristics of a specific manipulation type. Figure 3 illustrates the overall system.

A CDM learns the conditional distribution of manipulation masks through an iterative denoising process, enabling probabilistic predictions and improved uncertainty handling in visually ambiguous forensic regions. Moreover, the diffusion process itself acts as a strong form of regularization by decomposing mask prediction into a sequence of simpler denoising steps, improving generalization and training stability.

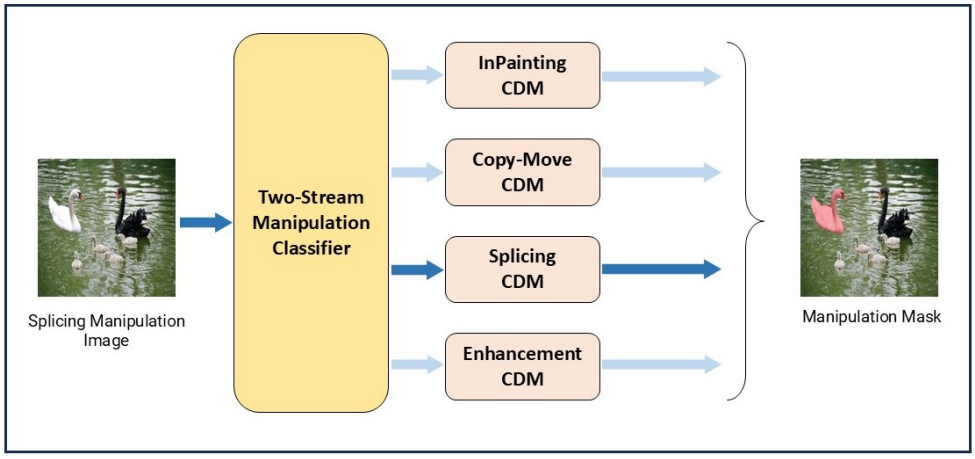


Fig. 3 Overview of 2-Stage pipeline for Image Manipulation Localization.

6.1 Dual-Stream Manipulation Classifier

The proposed classifier architecture utilizes a dual-stream framework to process an input image. The RGB stream extracts high level visual content, while the SRM stream applies 30 distinct SRM filters [41] to capture low level forensic noise details. Feature maps from both streams are concatenated and passed through a four stage ResNet like backbone composed of Residual Blocks [42] equipped with Squeeze and Excitation modules [43]. This stablizes the gradient flow while adaptively emphasizing the more informative feature channels from the fused RGB and forensic representations. This is followed by a Spatial Self Attention layer to capture long range dependencies. Finally, a classifier head classifies the image into one of the four manipulation types. Figure 4 shows the detailed architecture of the Manipulation Classifier.

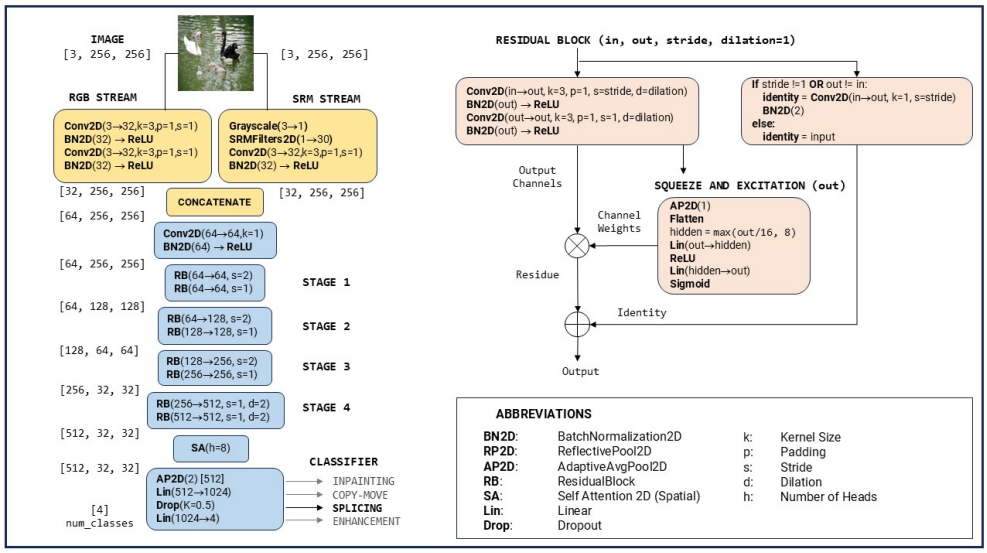


Fig. 4 Architecture overview of Dual-Stream Manipulation Classifier.

6.2 Manipulation Localization via Conditional Diffusion Models

Upon the successful classification of the input image, the system activates a specialized CDM for generating the localization mask for that manipulation type. It utilizes a U-Net [44] conditioned on the manipulated image to predict the noise residuals to reconstruct the localization mask. The model is trained using the DDPM framework over 1000 timesteps (T=1000) with a fixed linear noise schedule, where the variance β_t ranges from 10^{-4} to 0.02. To handle the imbalance where manipulated regions are much smaller than the background and encourage accurate spatial structure, the standard objective function is modified. The total loss is calculated as the sum of the Mean Squared Error (MSE) between predicted and actual noise and an Intersection over Union (IoU) loss term.

$$\mathcal{L}_{\text{total}} = \mathbb{E}_{x_0,t,\epsilon}\left[\underbrace{\|\epsilon - \epsilon_\theta(x_t,t,c)\|_2^2}_{\text{Noise Loss (MSE)}} + \lambda \underbrace{\mathcal{L}_{\text{IoU}}(\tilde{x}_0,\tilde{y})}_{\text{Mask Loss (IoU)}}\right] \tag{7}$$

$$\hat{x}_0 = \frac{x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t,t,c)}{\sqrt{\bar{\alpha}_t}}, \qquad \tilde{x}_0 = \frac{\text{clip}(\hat{x}_0,-1,1)+1}{2}, \qquad \tilde{y} = \frac{x_0+1}{2} \tag{8}$$

To capture dependencies within the image, a Spatial Self Attention module was incorporated in the bottleneck layer. Since CDMs are sensitive to the image size, a resolution of 256x256 was chosen for image size. This can be changed with minimal

effort by modifying the corresponding convolutions in the *Initial Conv* and *Final Conv* blocks, while keeping the remaining U-Net unchanged. Figure 5. shows the generalized architecture of the U-Net for the specialized CDMs

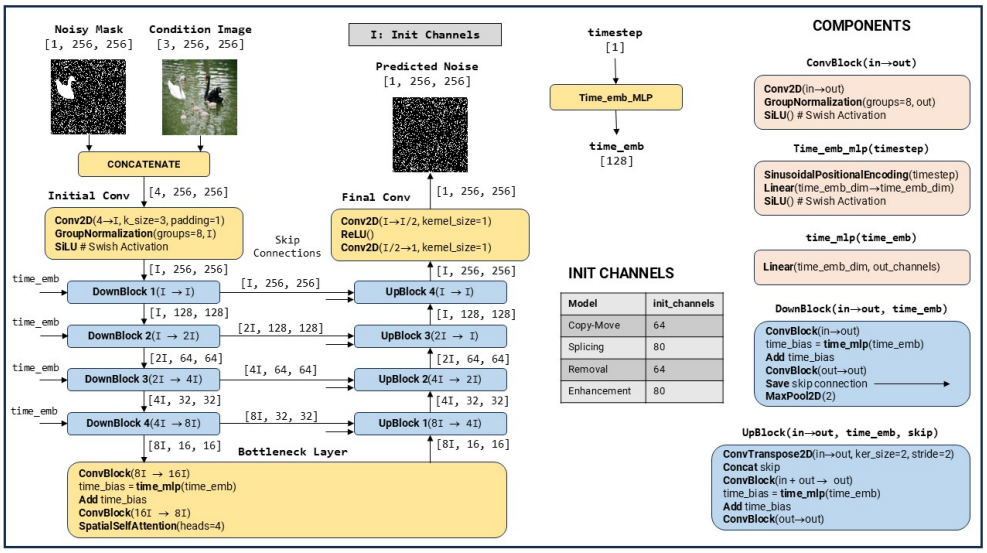


Fig. 5 Manipulation Localization CDM U-Net Architecture.

Although forensic filters such as SRM were explored as an additional conditioning stream applied to the input image, their inclusion consistently degraded localization performance. The deep U-Net backbone within the CDM provides sufficient capacity to implicitly learn manipulation-relevant features directly from RGB inputs. Furthermore, because the U-Net is invoked repeatedly across diffusion timesteps, a lightweight architecture was preferred, and incorporating additional forensic feature streams would substantially increase feature dimensionality and inference time without substantial performance gains.

For faster inference, the DDIM sampling strategy was also employed, reducing the generation process to 200 timesteps (T=200).

7 Datasets Used and Training Methodology

Primarily, the DF2023 dataset [45] was used for the purpose of training, due to the large number of samples present, labelled manipulation types and fixed resolution of images (256x256). Few other datasets namely: IMD2020 [46], CASIAv1 [47], CoMoFoD [48] and COVERAGE [49] were acquired to test the system. Images were resized appropriately to match the input size. Table 2. Shows the summary of the datasets used.

Table 2 Summary of Datasets used and type of manipulations present

Dataset	Total Images	Training Samples	Testing Samples
DF2023	Copy-move : 300,000	Copy-move : 50,000	Copy-move : 2,000
	Splicing : 400,000	Splicing : 67,000	Splicing : 2,000
	Removal : 100,000	Removal : 50,000	Removal : 2,000
	Enhancement : 200,000	Enhancement : 50,000	Enhancement : 2,000
IMD2020	Mixed : 2010	–	Mixed : 2010
CASIAv1	Mixed : 867	–	Mixed : 867
CoMoFoD	Copy-move : 200	Copy-move : 100[1]	Copy-move : 100
COVERAGE	Copy-move : 100	Copy-move : 50[1]	Copy-move : 50

[1]Samples used only for fine-tuning.

Each of the four CDMs were trained for up to 4 sessions, each with 5 epochs, on the images of corresponding manipulation type from the DF2023 dataset. The AdamW optimizer [50] was used, configured with a low learning rate and weight decay (10^{-4}). The training was conducted on Nvidia A100 GPUs. More details on training times needed can be referred from Table 10.

The Manipulation Classifier was trained on 25,000 images of each type from the DF2023 dataset for over 5 epochs (3910 steps, batch size = 32). The Adam optimizer was used with Cross Entropy Loss as the training objective.

8 Results and Discussion

The manipulation classifier achieved an accuracy of 89% when tested on the DF2023 dataset with labeled manipulation types. Figure 6. shows the confusion matrix for classification and other metrics. Table 3 shows the performance of the classifier with and without the SRM stream.

Table 3 Performance comparison of Manipulation Classifier with and without SRM stream

Metric	Single Stream (RGB)	Dual-Stream (RGB+SRM)
Accuracy	84.90	89.00
Precision	86.26	89.11
Recall	83.21	89.00
F1-Score	84.70	89.01

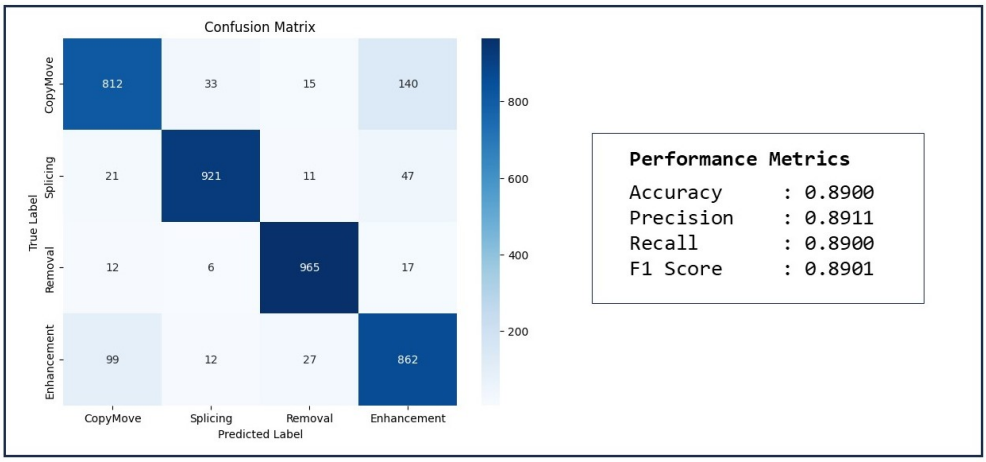


Fig. 6 Confusion Matrix and Performance Metrics for Manipulation Classifier.

Each of the localization CDMs were tested individually on test images of corresponding manipulation type from the DF2023 dataset. Both DDPM (T=1000) and DDIM ($T = 200, \eta = 1$) sampling was used to generate masks. Since generation is probabilistic and produces slightly different masks each run, masks from multiple runs can be averaged over to get more stable, and soft manipulation masks / heatmaps [39]. Tables 4, 5 and Figure 7 illustrate the performance of these CDMs on DF2023 dataset.

Table 4 Results of manipulation localization on DF2023 dataset using DDPM sampling.

Manipulation Type	Number of Masks used to Generate Heatmap											
	1			2			3			Average		
	IoU	F1	Acc	IoU	F1	Acc	IoU	F1	Acc	IoU	F1	Acc
Inpainting / Removal	88	92	98	88	92	98	88	92	98	88	92	98
Copy-Move	80	85	97	79	84	96	79	84	97	79	84	97
Splicing	52	62	86	49	59	86	52	63	87	51	61	86
Enhancement	63	70	94	62	69	93	63	70	94	63	70	94
Average	71	77	94	70	76	93	71	77	94	70	77	94

Table 5 Results manipulation localization on DF2023 dataset using DDIM sampling.

Manipulation Type	Number of Masks used to Generate Heatmap											
	1			2			3			Average		
	IoU	F1	Acc	IoU	F1	Acc	IoU	F1	Acc	IoU	F1	Acc
Inpainting / Removal	89	93	99	89	94	99	89	94	99	89	94	99
Copy-Move	79	84	96	78	84	96	78	84	97	78	84	96
Splicing	50	60	86	50	61	86	49	61	87	50	61	86
Enhancement	64	70	94	63	70	93	62	69	94	63	70	94
Average	71	77	94	70	77	94	70	77	94	70	77	94

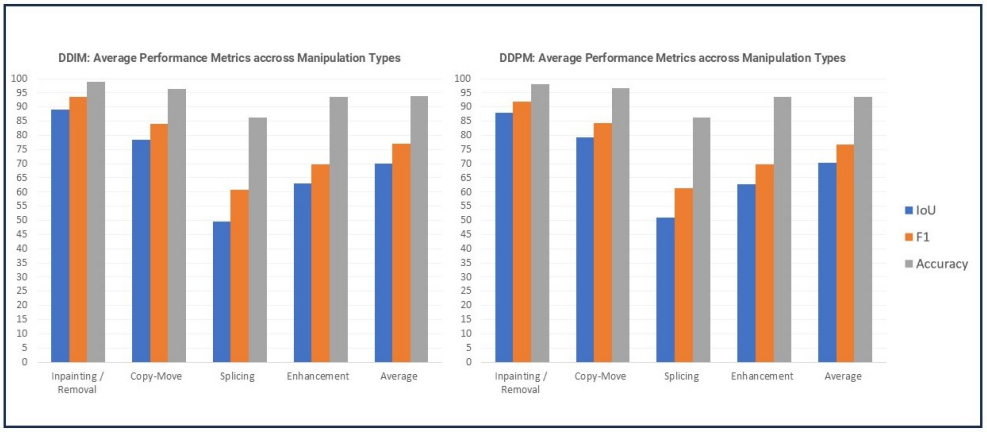


Fig. 7 Comparison of performance on different manipulation types in DF2023 using DDPM and DDIM sampling.

As expected, there is no impact on performance when using DDIM sampling despite it being significantly faster. Table 6. shows the performance of the complete end-to-end system (Classifier + CDMs) on DF2023 dataset. It also illustrates influence of classifiers performance on the overall system. Table 7 quantifies the effect of bottleneck self-attention added in the UNet.

Table 6 Performance of the end-to-end Manipulation Localization System.

Metric	Correctly Classified Images	Incorrectly Classified Images	All Images (Overall)	Incorrectly Classified Images *(using Correct CDM)*
F1	0.80	0.25	0.74	0.53
IoU	0.73	0.19	0.67	0.46
Accuracy	0.94	0.82	0.93	0.92

Table 7 Comparison of performance with/without Spatial Self Attention Module measured using DDIM sampling.

Manipulation Type	Without Bottleneck Self Attention			With Bottleneck Self Attention		
	IoU	F1	Acc	IoU	F1	Acc
Inpainting / Removal	87	93	99	89	94	99
Copy-Move	73	81	93	78	84	96
Splicing	49	60	86	50	61	86
Enhancement	60	69	93	63	70	94
Average	67	76	93	70	77	94

The Copy-Move CDM was fine-tuned on few samples from CoMoFoD and COVERAGE, and the entire system was tested on the benchmark datasets. Table 8. shows the performance on these datasets.

Table 8 Results of the entire end-to-end Manipulation Localization system on Benchmark datasets.

Metric	Dataset					
	DF2023	IMD2020	CoMoFoD	CASIAv1	COVERAGE	Average
Pixel Wise Accuracy	0.93	0.83	0.97	0.84	0.92	0.90
IoU	0.60	0.22	0.40	0.26	0.28	0.35
F1	0.68	0.27	0.50	0.32	0.40	0.43

Sparsity of manipulations adversely affects the performance, when the manipulated region is less than 5% of the image. Since the images are downsized to a resolution of 256x256 before generating manipulation masks, this drastically reduces the pixel count (proportional to the square of downscaling factor). This can be mitigated by

increasing the input resolution of the U-Net, as discussed earlier. Figure 8 illustrates the impact of manipulation region size on the system's performance. The second plot presents the distribution of images across different manipulation area ratios.

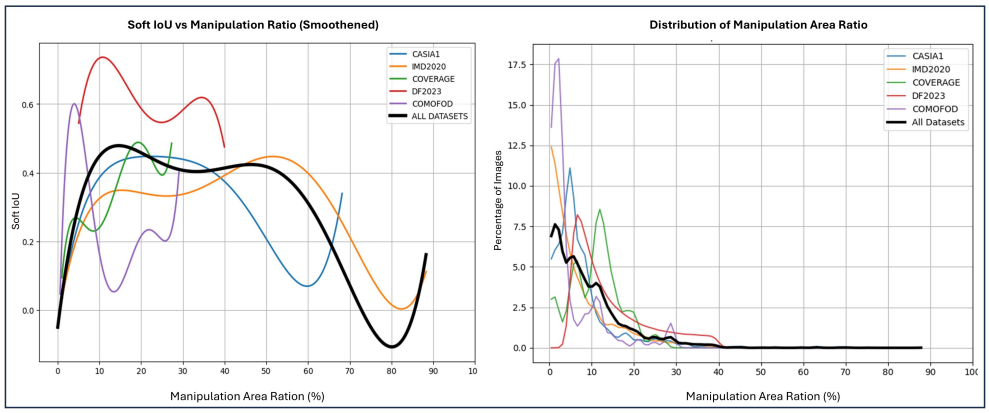


Fig. 8 Effect of manipulation area ratio on performance, and the distribution of images across different manipulation area ratios plotted for various datasets.

An alternative approach was also evaluated, in which the final manipulation mask is obtained as a weighted average of the masks produced by all CDMs, using the classifier's softmax probabilities as weights. However, this strategy did not result in a significant performance improvement. It is observed empirically that high top-class classifier confidence is a requirement for good performance.

$$P(\text{good IoU}\,(>0.3)\mid \text{low top-class confidence}\,(<0.5)) = 0.33$$
$$P(\text{good IoU}\,(>0.3)\mid \text{high top-class confidence}\,(>0.5)) = 0.84$$

In such cases the weighted mask will be nearly indentical to the one generated using the original approach. Nevertheless, this weighting scheme may offer advantages in scenarios involving multiple manipulation types within a single image, where contributions from multiple CDMs could become relevant.

For baseline comparisons, performance metrics reported in [51] are used. Table 9 shows the proposed system achieves competitive performance with other similar systems. Figures 9 and 10 show some sample outputs of the proposed system.

DDIM sampling is significantly faster than DDPM. This can be seen from Table 10 listing training and inference times for various configurations. Figure 11 shows the effect of DDIM timesteps on localization performance, encouraging smaller number of timesteps without noticeable drop in performance.

Table 9 Comparison with other recent, similar systems with performance metrics as reported in [51]

Method	**CASIAv1**		**IMD2020**	
	IoU	F1	IoU	F1
SPAN [17]	0.11	0.14	0.09	0.14
ManTraNet [12]	0.09	0.13	0.10	0.16
HiFi-Net [52]	0.13	0.18	0.09	0.14
PSCC-Net [18]	0.36	0.46	0.22	0.32
CAT-Net [53]	0.44	0.51	0.14	0.19
MVSS-Net [14]	0.40	0.48	0.23	0.31
CDM-Net	0.26	0.32	0.22	0.27

Table 10 Training and Testing time required for CDMs with different configurations.

Task	Device	Sampling	Model	Manipulation Type	Time (avg)
Training	Nvidia A100	DDPM (T=1000)	CDM (I=64)	Copy-Move, Removal	30 mins / session
		DDPM (T=1000)	CDM (I=80)	Splicing, Enhancement	50 mins / session
Inference	Nvidia L4	DDPM (T=1000)	CDM (I=64)	Copy-Move, Removal	7734 ms / mask 129 timesteps/sec
		DDPM (T=1000)	CDM (I=80)	Splicing, Enhancement	11482 ms / mask 87 timesteps/sec
		DDIM (T=200)	CDM (I=64)	Copy-Move, Removal	2418 ms / mask 83 timesteps/sec
		DDIM (T=200)	CDM (I=80)	Splicing, Enhancement	1660 ms / mask 120 timesteps/sec

One training session corresponds to 5 epochs.

With batch size of 32, for Splicing CDM, each epoch corresponds to 2094 steps.

For other CDMs, each epoch corresponds to 1563 steps

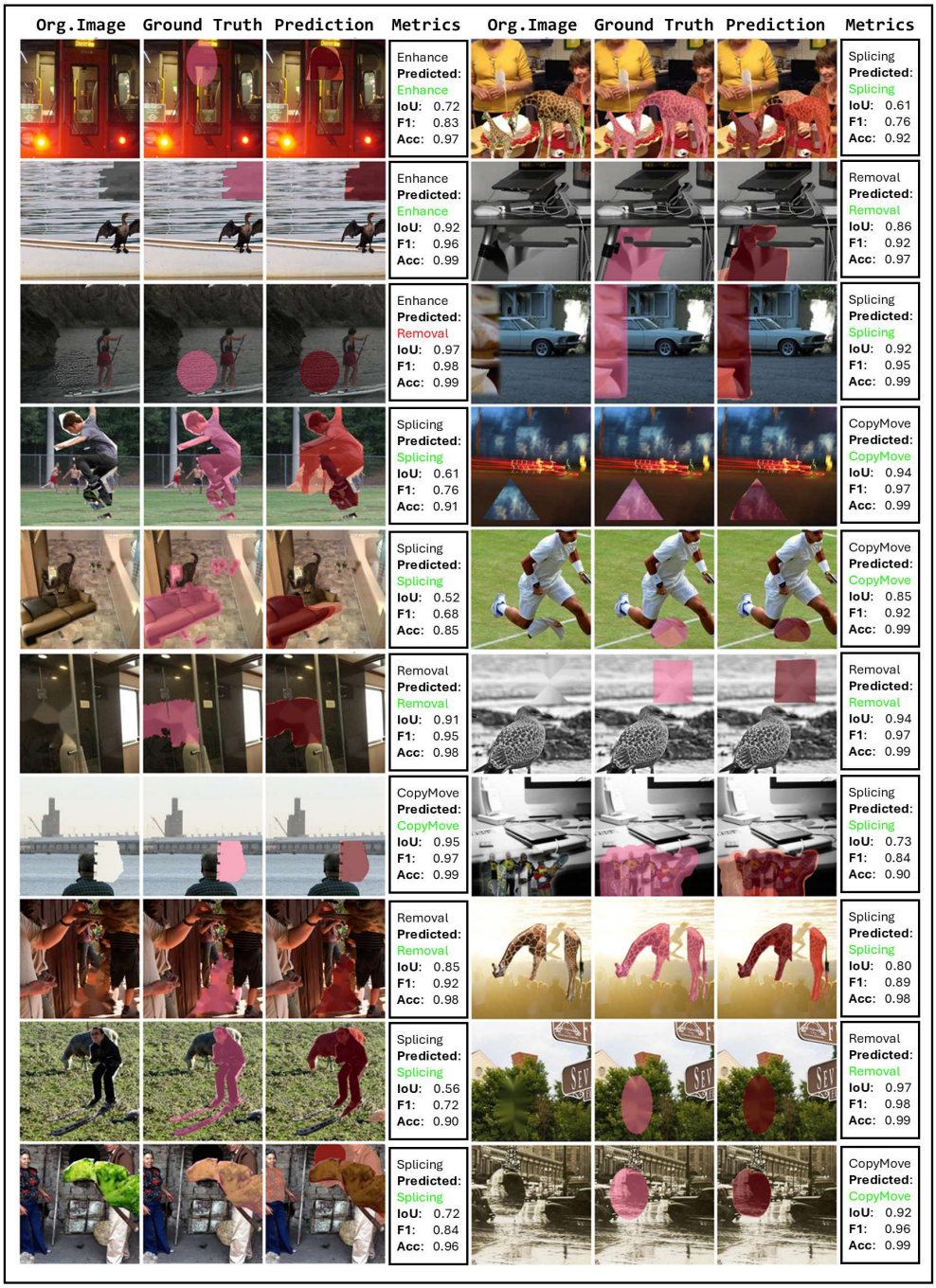


Fig. 9 Sample end-to-end outputs with manipulation classification and localization.

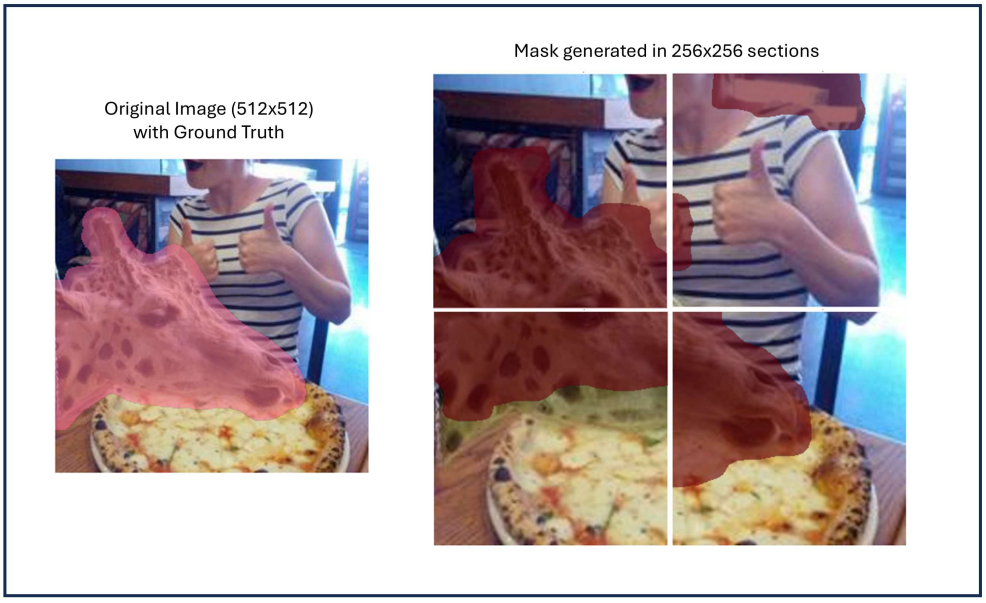


Fig. 10 Demonstration of Tiling where a large image is processed in parts of 256x256 sections.

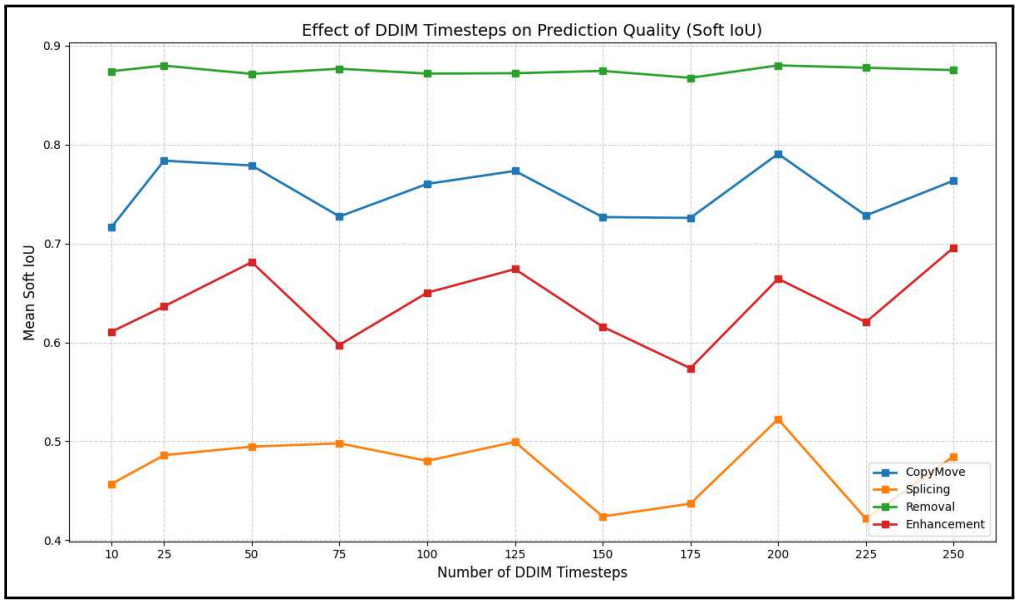


Fig. 11 Effect of DDIM timesteps used to sample a mask on localization performance measured on DF2023 dataset.

9 Conclusions

This paper successfully introduced a novel two-stage approach for Image Manipulation Localization (IML), leveraging a Dual-Stream Manipulation Classifier to categorize input images and routing them to specialized CDMs for precise mask generation. The system demonstrated competitive performance across several benchmark datasets, confirming the efficacy of decoupling manipulation type identification from the localization task and employing CDMs for pixel-wise segmentation. The decoupling also provides an added benefit of not requiring complete retraining to adapt to new manipulation types and improving ease of training. Future work can focus on expanding the system's capability to handle images containing multiple, co-occurring forgery types. This would involve training the classifier to output a multi-hot encoded vector representing all detected manipulations, followed by running the corresponding specialized CDMs to localize each. Furthermore, research into more effective fine-tuning strategies, such as Elastic Weight Consolidation (EWC) [54], can be explored to improve performance on diverse, smaller benchmark datasets while mitigating the risk of catastrophic forgetting of the general knowledge learned from large-scale training data.

In real-world deployments, the proposed framework can be integrated as a modular IML suite within digital forensics pipelines used by news agencies, law enforcement, or social media platforms. New manipulation types can be introduced as plugin-style CDMs without affecting existing components, enabling continuous system evolution as new forgery techniques emerge.

Declarations

Conflict of Interest

The authors declare that they have no conflict of interest.

Data Availability

Data produced / analyzed will be provided on request. The following publicly available datasets are used in this research.

Primary Dataset

- DF2023 [45]: available online on request at: https://zenodo.org/records/7326540

Benchmark Datasets

- IMD2020 [46]: available online at: https://staff.utia.cas.cz/novozada/db
- CoMoFoD [48]: available online at: https://www.vcl.fer.hr/comofod/
- CASIA [47]: available online at: https://github.com/namtpham/casia1groundtruth
- COVERAGE [49]: available online at: https://github.com/wenbihan/coverage

Author Contribution

All authors contributed equally to the conceptualization, implementation, evaluation, and documentation of this work.

Consent to Publish

All authors consent to the publication of this manuscript.

Funding

This research did not receive any grants/funding.

Ethics Approval

Not Applicable

References

[1] Nightingale, S.J., Wade, K.A., Watson, D.G.: Can people identify original and manipulated photos of real-world scenes? Cogn. Res. Princ. Implic. **2**(1), 30 (2017) https://doi.org/10.1186/s41235-017-0067-2

[2] Vosoughi, S., Roy, D., Aral, S.: The spread of true and false news online. Science **359**(6380), 1146–1151 (2018) https://doi.org/10.1126/science.aap9559

[3] Weebly Guides: How Important Is Product Photography? Weebly Guides Online. Accessed: September 2025. https://www.weebly.com/guides/product-photography-guide/how-important-is-product-photography (2025)

[4] Martin, C., Blatt, M.: Manipulation and misconduct in the handling of image data. Plant Cell **25**(9), 3147–3148 (2013) https://doi.org/10.1105/tpc.113.250980

[5] Miller, C.M.: A survey of prosecutors and investigators using digital evidence: A starting point. Forensic Sci. Int. Synergy **6**, 100296 (2022) https://doi.org/10.1016/j.fsisyn.2022.100296

[6] Kadha, V.K., Bakshi, S., Das, S.K.: Unravelling digital forgeries: A systematic survey on image manipulation detection and localization. ACM Comput. Surv. **57**(12), 1–36 (2025) https://doi.org/10.1145/3731243

[7] Müller, D., Soto-Rey, I., Kramer, F.: Towards a guideline for evaluation metrics in medical image segmentation. BMC Res. Notes **15**(1), 210 (2022) https://doi.org/10.1186/s13104-022-06096-y

[8] Iakovidou, C., Zampoglou, M., Papadopoulos, S., Kompatsiaris, Y.: Content-aware detection of JPEG grid inconsistencies for intuitive image forensics. J. Vis.

Commun. Image Represent. **54**, 155–170 (2018) https://doi.org/10.5281/zenodo.1246437

[9] Dua, S., Singh, J., Parthasarathy, H.: Image forgery detection based on statistical features of block DCT coefficients. Procedia Comput. Sci. **171**, 369–378 (2020) https://doi.org/10.1016/j.procs.2020.04.038

[10] Liu, L., Sun, P., Lang, Y., Li, J.: CFA-based splicing forgery localization method via statistical analysis. IET Signal Process. **18**(1), 1–11 (2024) https://doi.org/10.1049/2024/9929900

[11] Niu, P., Wang, C., Chen, W., Yang, H., Wang, X.: Fast and effective keypoint-based image copy-move forgery detection using complex-valued moment invariants. J. Vis. Commun. Image Represent. **77**, 103068 (2021) https://doi.org/10.1016/j.jvcir.2021.103068

[12] Wu, Y., AbdAlmageed, W., Natarajan, P.: ManTra-Net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9543–9552 (2019). https://doi.org/10.1109/CVPR.2019.00977

[13] Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.C.: Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Preprint at https://arxiv.org/abs/1506.04214 (2015)

[14] Dong, C., Chen, X., Hu, R., Cao, J., Li, X.: MVSS-Net: Multi-view multi-scale supervised networks for image manipulation detection. IEEE Trans. Pattern Anal. Mach. Intell. **45**(3), 3539–3553 (2022) https://doi.org/10.1109/TPAMI.2022.3180556

[15] Dou, L., Chen, M., Qiu, J., Wang, J.: MSF-Net: Multi-stream fusion network for image manipulation detection and localization. Digit. Signal Process. **161**, 105114 (2025) https://doi.org/10.1016/j.dsp.2025.105114

[16] Bai, R.: Image manipulation detection and localization using multi-scale contrastive learning. Appl. Soft Comput. **163**, 111914 (2024) https://doi.org/10.1016/j.asoc.2024.111914

[17] Hu, X., Zhang, Z., Jiang, Z., Chaudhuri, S., Yang, Z., Nevatia, R.: SPAN: Spatial pyramid attention network for image manipulation localization. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 312–328. Springer, Glasgow, UK (2020). https://doi.org/10.1007/978-3-030-58589-1_19

[18] Liu, X., Liu, Y., Chen, J., Liu, X.: PSCC-Net: Progressive spatio-channel correlation network for image manipulation detection and localization. IEEE Trans. Circuits Syst. Video Technol. **32**(11), 7505–7517 (2022) https://doi.org/10.1109/

TCSVT.2022.3189545

[19] Chen, T., Li, B., Zeng, J.: Learning traces by yourself: Blind image forgery localization via anomaly detection with ViT-VAE. IEEE Signal Process. Lett. **30**, 150–154 (2023) https://doi.org/10.1109/LSP.2023.3245947

[20] Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: International Conference on Learning Representations (ICLR) (2014)

[21] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021). Preprint at https://arxiv.org/abs/2010.11929

[22] Liu, W., Cun, X., Pun, C.-M.: DH-GAN: Image manipulation localization via a dual homology-aware generative adversarial network. Pattern Recognit. **155**, 110658 (2024) https://doi.org/10.1016/j.patcog.2024.110658

[23] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, vol. 27. Curran Associates, Inc., Montreal, QC, Canada (2014)

[24] Wang, G., Yang, M., Chen, R.: A weakly supervised multi-stream feature enhancement network for image manipulation detection and localization. IEEE Access **13**, 157111–157125 (2025) https://doi.org/10.1109/ACCESS.2025.3607317

[25] Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: Convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19. Springer, Munich, Germany (2018). https://doi.org/10.1007/978-3-030-01234-2_1

[26] Hou, X.-Y., Yaermaimaiti, Y., Cheng, S.-Q.: CLIP-IML: A novel approach for CLIP-based image manipulation localization. J. Electron. Sci. Technol. **23**(3), 56–70 (2025)

[27] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML), vol. 139, pp. 8748–8763 (2021)

[28] Wang, K., Niu, S., Hao, Q., Zhang, J.: InpDiffusion: Image inpainting localization via conditional diffusion models. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 7771–7779 (2025). https://doi.org/10.1609/aaai.v39i7.32837

[29] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 34, pp. 6840–6851. Curran Associates Inc., Red Hook, NY, USA (2020)

[30] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (ICLR) (2021)

[31] Panagiotakopoulos, T., Kotsiantis, S., Gkillas, A., Lalos, A.S.: Conditional diffusion models: A survey of techniques, applications, and challenges. IEEE Access **13**, 183617–183643 (2025) https://doi.org/10.1109/ACCESS.2025.3625094

[32] Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In: Proceedings of the 39th International Conference on Machine Learning (ICML). PMLR, vol. 162, pp. 16784–16804 (2022)

[33] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical Text-Conditional Image Generation with CLIP Latents. Preprint at https://arxiv.org/abs/2204.06125 (2022)

[34] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S., Ayan, B., Mahdavi, S., Gontijo-Lopes, R., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 35, pp. 36479–36494 (2022)

[35] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10674–10685 (2022). https://doi.org/10.1109/CVPR52688.2022.01042

[36] Saharia, C., Chan, W., Chang, H., Lee, C.A., Ho, J., Salimans, T., Fleet, D.J., Norouzi, M.: Palette: Image-to-image diffusion models. In: Proceedings of ACM SIGGRAPH, pp. 1–10 (2022). https://doi.org/10.1145/3528233.3530757

[37] Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D.J., Norouzi, M.: Image super-resolution via iterative refinement. IEEE Trans. Pattern Anal. Mach. Intell. **45**(4), 4713–4726 (2023) https://doi.org/10.1109/TPAMI.2022.3204461

[38] Tian, J., Wu, J., Chen, H., Ma, M.: MapGen-Diff: An end-to-end remote sensing image to map generator via denoising diffusion bridge model. Remote Sens. **16**(19), 3716 (2024) https://doi.org/10.3390/rs16193716

[39] Amit, T., Shaharbany, T., Nachmani, E., Wolf, L.: SegDiff: Image Segmentation with Diffusion Probabilistic Models. Preprint at https://arxiv.org/abs/2112.00390 (2022)

[40] Wu, J., Fu, R., Fang, H., Zhang, Y., Yang, Y., Xiong, H., Liu, H., Xu, Y.: MedSegDiff: Medical image segmentation with diffusion probabilistic model. In: Proceedings of Medical Imaging with Deep Learning (MIDL). PMLR, vol. 227, pp. 1623–1639 (2024)

[41] Rubio, A., Grisales, J., Tabares Soto, R., Ramos-Pollán, R., Ortíz, M., Arteaga Arteaga, H., Isaza, G., Arias, D., Orozco Arias, S., Burbano Jàcome, A.: GBRAS-Net: A convolutional neural network architecture for spatial image steganalysis. IEEE Access **9**, 14340–14350 (2021) https://doi.org/10.1109/ACCESS.2021.3052494

[42] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016). https://doi.org/10.1109/CVPR.2016.90

[43] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141 (2018). https://doi.org/10.1109/CVPR.2018.00745

[44] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28

[45] Fischinger, D., Boyer, M.: DF2023: The digital forensics 2023 dataset for image forgery detection. In: Proceedings of the 25th Irish Machine Vision and Image Processing Conference (IMVIP), pp. 128–135 (2023)

[46] Novozamsky, A., Mahdian, B., Saic, S.: IMD2020: A large-scale annotated dataset for image manipulation detection. In: 2020 IEEE Winter Applications of Computer Vision Workshops (WACVW), pp. 71–80 (2020). https://doi.org/10.1109/WACVW50321.2020.9096940

[47] Dong, J., Wang, W., Tan, T.: CASIA image tampering detection evaluation database. In: Proceedings of the IEEE China Summit International Conference on Signal and Information Processing, pp. 422–426 (2013). https://doi.org/10.1109/ChinaSIP.2013.6625374

[48] Tralic, D., Zupancic, I., Grgic, S., Grgic, M.: CoMoFoD — new database for copy-move forgery detection. In: Proceedings of the 55th International Symposium ELMAR, pp. 49–54 (2013)

[49] Wen, B., Zhu, Y., Subramanian, R., Ng, T.-T., Shen, X., Winkler, S.: COVERAGE — a novel database for copy-move forgery detection. In: Proceedings of the IEEE International Conference on Image Processing (ICIP), pp. 161–165 (2016). https://doi.org/10.1109/ICIP.2016.7532339

[50] Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR) (2019)

[51] Xu, Z., Zhang, X., Li, R., Tang, Z., Huang, Q., Zhang, J.: FAKESHIELD: Explainable image forgery detection and localization via multi-modal large language models. In: International Conference on Learning Representations (ICLR) (2025)

[52] Guo, X., Liu, X., Ren, Z., Grosz, S., Masi, I., Liu, X.: Hierarchical fine-grained image forgery detection and localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3155–3165 (2023). https://doi.org/10.1109/CVPR52729.2023.00308

[53] Kwon, M.J., Yu, I.J., Nam, S.H., Lee, H.K.: CAT-Net: Compression artifact tracing network for detection and localization of image splicing. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 375–384 (2021). https://doi.org/10.1109/WACV48630.2021.00042

[54] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., Hadsell, R.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences **114**(13), 3521–3526 (2017) https://doi.org/10.1073/pnas.1611835114