Detecting Localized Deepfakes: How Well Do Synthetic Image Detectors Handle Inpainting?

Serafino Pandolfini[1,*], Lorenzo Pellegrini[1], Matteo Ferrara[1] and Davide Maltoni[1]

[1]*Department of Computer Science and Engineering (DISI), University of Bologna, Italy*

Abstract

The rapid progress of generative AI has enabled highly realistic image manipulations, including inpainting and region-level editing. These approaches preserve most of the original visual context and are increasingly exploited in cybersecurity-relevant threat scenarios. While numerous detectors have been proposed for identifying fully synthetic images, their ability to generalize to localized manipulations remains insufficiently characterized. This work presents a systematic evaluation of state-of-the-art detectors, originally trained for the deepfake detection on fully synthetic images, when applied to a distinct challenge: localized inpainting detection. The study leverages multiple datasets spanning diverse generators, mask sizes, and inpainting techniques. Our experiments show that models trained on a large set of generators exhibit partial transferability to inpainting-based edits and can reliably detect medium- and large-area manipulations or regeneration-style inpainting, outperforming many existing ad hoc detection approaches.

Keywords

Generative AI, Image Inpainting, Deepfake Detection

1. Introduction

The rapid advancement of generative AI has fundamentally reshaped the digital media landscape, enabling the creation of highly realistic synthetic imagery at an unprecedented scale. While these technologies offer powerful creative and professional capabilities, they have also amplified the spread of manipulated content intended to mislead, deceive, or imitate reality. This threat is further amplified by the proliferation of image inpainting models, which can seamlessly remove, modify, or insert visual elements into authentic images with minimal perceptual artifacts. Originally developed for image restoration, inpainting now facilitates subtle yet highly effective manipulations that pose significant risks in cybersecurity-critical domains, including coordinated disinformation campaigns, political influence operations, social engineering, and digital fraud. Beyond classical inpainting, recent advances in image editing and image-to-image generative pipelines, such as prompt-based editing, object replacement, background regeneration, and region-guided diffusion, enable far more sophisticated manipulations that merge characteristics of both inpainting and full image generation. These systems can locally regenerate large image regions, rewrite semantic content, or even produce fully edited images while preserving the global structure and altering key elements in ways that remain almost imperceptible to human observers. From a forensic perspective, such manipulations represent an evolution of inpainting: they still rely on regenerating selected regions while maintaining coherence with the original image, but they operate at much larger scales and introduce transformations that blur the boundary between "edited" and "synthetic." This emerging spectrum of localized-to-global modifications defines a new frontier in media manipulation, expanding the capabilities available to malicious actors and significantly complicating traditional detection strategies. In recent years, synthetic content has been increasingly exploited to fabricate evidence, distort public narratives, and undermine trust in authentic digital media. Unlike fully synthetic images, inpainted or selectively regenerated images preserve most of the original visual context, making them particularly difficult to detect and highly effective for deceptive purposes.

1st Workshop on Digital Forensics (WDF 2026), co-located with the 2nd Joint National Conference on Cybersecurity (ITASEC & SERICS 2026), February 09–13, 2026, Cagliari, Italy

*Corresponding author.

serafino.pandolfini2@unibo.it (S. Pandolfini); l.pellegrini@unibo.it (L. Pellegrini); matteo.ferrara@unibo.it (M. Ferrara); davide.maltoni@unibo.it (D. Maltoni)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-14

CEUR-WS.org/Vol-4329/paper1.pdf

This emerging class of "localized deepfakes" expands the attack surface for adversaries and raises urgent concerns for platforms, institutions, and investigators that rely on trustworthy visual information.

In response, the research community has introduced a wide range of detection strategies, from handcrafted forensic analyses and frequency-domain artifact detection to patch-based classifiers and vision models fine-tuned to distinguish authentic from manipulated content. However, as with deepfake and synthetic image detection more broadly, reported performance often varies significantly depending on factors such as the choice of training generators, data preprocessing, augmentation strategies, and architectural biases. A recent work has shown that the performance of inpainting localization methods is highly sensitive to dataset composition, including image content and the specific generative tools used, raising concerns about evaluation bias and real-world generalization [1]. This lack of standardization complicates method comparison and obscures which design principles truly enhance generalization, particularly for inpainting detection, where manipulated regions may be small, context-dependent, and visually indistinguishable from their surroundings.

In this work, we present a comprehensive study of AI-generated image detection, with a particular focus on image inpainting, image editing, and localized manipulations, framed within the context of cybersecurity and media integrity. Following the principles established in AI-GenBench [2], we design an evaluation protocol tailored to inpainting and localized-editing forensics. Our protocol spans a variety of generators, mask dimensions, and manipulation types, providing a controlled environment for assessing the robustness of current detection approaches. This setting allows us to explore how well detectors trained for conventional real–fake classification on fully synthetic images transfer to more subtle, localized manipulations such as inpainting.

Rather than introducing a new detection model, our study evaluates state-of-the-art detectors originally trained on AI-GenBench [2] for detecting fully synthetic images and analyzes their robustness when applied to localized inpainted content. We investigate the factors that most strongly influence generalization performance, including the size of the inpainted region, the underlying manipulation algorithm, and the characteristics of different generative models. By comparing detection behavior across these dimensions and comparing it to previous results on fully synthetic images, we identify patterns that reveal how inpainting-based manipulations differ from fully synthetic content in the eyes of current detectors. Through this targeted analysis, our work provides actionable insights for designing future forensic systems, highlighting failure modes and opportunities to improve robustness against localized attacks.

This paper is organized as follows. Section 2 introduces the background and relevant related work. Section 3 details the experimental setup, including the inpainting datasets used, the training procedure, and the input pipeline. Section 4 presents the experimental results from multiple dimensions of analysis. Section 5 summarizes our findings and highlights open research questions and directions for future work.

2. Related Work

Detecting AI-assisted image manipulations increasingly extends beyond traditional deepfakes to include inpainting, localized edits, and partial image regeneration. These techniques preserve most of the original visual context while subtly altering specific regions, making them significantly harder to detect than fully synthetic images. In this section, we review the state-of-the-art in two closely related areas: (i) detection methods, focusing on architectural choices, forensic cues, and algorithmic strategies used to identify manipulations, and (ii) benchmarks and datasets, which define evaluation protocols, manipulation types, and generator diversity for assessing the robustness and generalization of detection systems.

2.1. Inpainting Techniques

Modern inpainting and localized editing methods can be broadly grouped into two methodological categories based on their usage in the generation process: (i) masked inpainting, or(ii) full regeneration.

Masked (pure) inpainting - One of the earliest approaches to inpainting relies on a binary mask to constrain the region to be modified. The objective is to reconstruct or hallucinate plausible content within the masked area while preserving global visual coherence. This category includes:

- classical inpainting algorithms, such as Telea's method [3],
- feed-forward convolutional architectures, such as LaMa [4] and MAT [5],
- diffusion-based inpainting models, such as Stable Diffusion 2 [6] and BrushNet [7].

Full regeneration - Fully regenerative methods are becoming increasingly prevalent, as they are often the default inpainting approach in simplified AI-systems, such as chat-based interfaces. These methods replace large masked regions or entire semantic components of an image using powerful generative priors. In doing so, they may substantially alter scene semantics, lighting conditions, or texture statistics while maintaining a high degree of perceptual realism. Examples of fully regenerative methods include:

- diffusion-based region regeneration models, such as Stable Diffusion 2 (SD2) [6] and Stable Diffusion XL (SDXL) [6],
- text-guided editing pipelines, including Adobe Firefly [8] and PowerPaint [9],
- large generative models, such as FLUX.1 schnell, FLUX.1 dev, and FLUX.1 filldev [10].

Image regeneration often introduces unwanted modifications, including identity alterations, changes in lighting conditions, the introduction of noticeable graininess, and a tendency toward cartoonization, among others.

These two categories pose different levels of difficulty for forensic analysis. Masked inpainting typically introduces subtle inconsistencies confined to the manipulated region, often resulting in anomalous noise patterns relative to the rest of the image. In contrast, semantic regeneration may produce globally coherent outputs that exhibit distributional shifts, challenging detectors that rely on local artifacts while potentially introducing image-wide signatures.

2.2. Detection Methods

Early forensic techniques primarily relied on handcrafted cues and statistical inconsistencies. Representative examples include Noiseprint [11], which models camera-specific noise residuals to localize anomalous regions, and frequency-domain or wavelet-based analyses that have been recently adapted to detect artifacts introduced by diffusion-based inpainting.

More recent approaches leverage deep learning for segmentation or classification. CTNet targets local tampering through patch-wise transformer reasoning, while DiffSeg [12] introduces multi-feature segmentation specifically designed for diffusion-based inpainting detection. FakeShield [13] further advances the field by integrating multi-modal forensic cues and robust post-processing strategies to improve localized manipulation detection.

2.3. Benchmarks and Datasets

Benchmarks for inpainting and localized manipulation forensics remain considerably less mature than those developed for full-image deepfake or synthetic image detection. Existing datasets vary in scale, manipulation diversity, mask design, and realism, and often cover only a limited subset of inpainting algorithms.

The Inpainting Localization Evaluation Dataset [14] provides pixel-level annotations for manipulations generated using a combination of classical and deep-learning-based inpainting algorithms. The dataset includes results from foundational deep inpainting methods such as Contextual Attention [15], Gated Convolution [16], ShiftNet [17], EdgeConnect [18], and Region Normalization [19], alongside classical algorithms including Telea's fast marching method [20], Navier–Stokes inpainting [21], and several patch-based or structure-guided approaches [22, 23, 24]. While these methods were representative of

the state of the art between 2017–2020, they predate modern diffusion and transformer-based inpainting models and therefore exhibit less realistic regeneration capabilities compared to contemporary systems. Although smaller in scale compared to more recent datasets, it remains a widely used benchmark for assessing segmentation-based approaches for inpainting localization.

DEFACTO [25] is a large-scale dataset designed for generic image manipulation detection, including splicing, copy-move, and inpainting. Its inpainting subset provides a variety of object-removal scenarios; however, most manipulations rely on classical or non-generative restoration techniques. Consequently, DEFACTO is less suitable for evaluating robustness against contemporary generative inpainting pipelines.

DeepfakeArt [26] is primarily oriented toward manipulation detection in artistic and stylized imagery. Although it includes localized alterations and generative transformations, its non-photographic domain limits its applicability to inpainting forensics on natural-image.

InpaintCOCO [27] contains inpainted versions of COCO images generated using neural inpainting models, along with ground-truth masks for the removed regions. The dataset primarily emphasizes object-removal scenarios and masked inpainting pipelines, but includes fewer large-scale semantic regenerations compared to more recent benchmarks.

Beyond the Brush [28] targets modern generative editing workflows, including region-level editing, semantic modifications, and object-guided content generation. Although not exclusively focused on inpainting, portions of the dataset involve localized content replacement and regeneration, making it highly relevant to inpainting detection research.

TGIF [29] and TGIF-2 [30] provide standardized resources for evaluating tampering localization and object-removal detection under modern generative editing pipelines. Unlike earlier inpainting benchmarks, both datasets incorporate manipulations produced by state of the art diffusion-based and text-guided inpainting systems, including models such as Adobe Firefly [8] and various Flux1 models [10]. The datasets further includes cases where the editing model regenerates the entire image rather than only the masked region, an emerging challenge for localized forensics, as traditional inpainting localization methods often fail under full-image regeneration.

BR-Gen [31] is a recent large-scale dataset designed for localized generative forgeries, emphasizing both object-level manipulations and broader scene edits such as sky, ground, or background regions. Although the dataset relies on slightly older but still widely used generative models—including LaMa [4], BrushNet [7], PowerPaint [9], SDXL [6], and MAT [5], it offers a highly diverse set of masks and scene-aware annotations produced through an automated perception–creation–evaluation pipeline. This process ensures semantic coherence and high visual realism across the dataset.

3. Experimental setup

Our objective is to establish a baseline for detecting AI-generated inpainting and localized manipulations using models originally trained for binary real-versus-fake classification on fully synthetic images. Specifically, we assess whether state-of-the-art deepfake detectors trained on AI-GenBench can generalize to the inpainting and regeneration scenarios.

We adopt a standard image-level detection formulation. Given an input image $I \in \mathbb{R}^{H \times W \times 3}$, the detector outputs a scalar probability $p_\theta(I)$ representing the likelihood that the image contains any AI-generated content. No spatial supervision is provided, and the detector is not informed about the location or size of manipulated regions. The task is therefore strictly binary: the model must decide whether the image contains any form of inpainting or regeneration.

To investigate architectural effects and to cover both convolutional and transformer-based paradigms, we evaluate two representative backbones previously considered in AI-GenBench: ResNet-50 CLIP [32] by OpenAI and DINOv2 [33] and DINOv3 [34], both in their ViT-L variants.

3.1. Training Procedure

The two models are trained on AI-GenBench [2] following the optimized designed choices outlined in [35], including optimization parameters, augmentation strategies, and model-specific hyperparameters. Key components include:

- *Resize-based preprocessing* - Input images are resized to the native resolution of the model without cropping.
- *Augmentation pipeline* - Training uses the evaluation-based augmentation suite from [35], which includes multi-pass JPEG compression and limited perturbations such as Gaussian noise addition, grayscale conversion, rotations, and flipping.
- *Supervision* - Only binary real-versus-fake supervision is used. No inpainting-specific data is included during training, nor any labels related to the model used to generate the inpainted region.

This protocol ensures that any performance on inpainting detection arises from generalization rather than from exposure to localized manipulations during training.

3.2. Testing Datasets

To construct our evaluation corpus, we aggregate real and manipulated images from established inpainting and localized-editing datasets, ensuring broad coverage across manipulation types, mask configurations, and generative pipelines.

Real images are taken from the unaltered subsets of BR-Gen [31], which draws from COCO2017 [36], ImageNet [37], and Places [38], as well as from TGIF [29] and TGIF-2 [30], both of which also originate from COCO2017. Manipulated samples are collected from BR-Gen, TGIF, and TGIF-2, capturing a heterogeneous range of modern generative inpainting techniques. These datasets collectively include forgeries produced by a diverse set of contemporary models. Figure 1 provides representative example images from the dataset.



(a) LaMa (b) MAT (c) PowerPaint (d) BrushNet (e) SD2



(f) SDXL (g) Firefly (h) Flux1 dev (i) Flux1 filldev (j) Flux1 schnell

Figure 1: Inpainted images from different generators.

The combined dataset inherits the rich diversity of mask shapes and scales present across its sources, spanning small occlusions, object-level removals, background and scene-region edits, and free-form semantic inpainting. As a result, the assembled dataset provides a broad and representative foundation for evaluating detection methods under realistic modern inpainting scenarios. An overview of the dataset composition is provided in Table 1, which reports the number of inpainted and real images per source dataset. Figure 2 illustrates the mask size distributions for each dataset, while Figure 3 shows the distribution of masked area percentages across generative models. Consistent with these distributions, Table 2 summarizes the average mask size and standard deviation for each generator, revealing substantial variability across models. It is worth noting that the source datasets are unbalanced

with respect to mask size (expressed as a percentage of of image area), and images manipulated with modern generative models tend to exhibit smaller masks.

Table 1
Number of images in the datasets.

Dataset	# of inpainted images	# of real images
BR-Gen	150000	15000
TGIF	74808	9372
TGIF2	112464	6248
Total	**337272**	**30620**

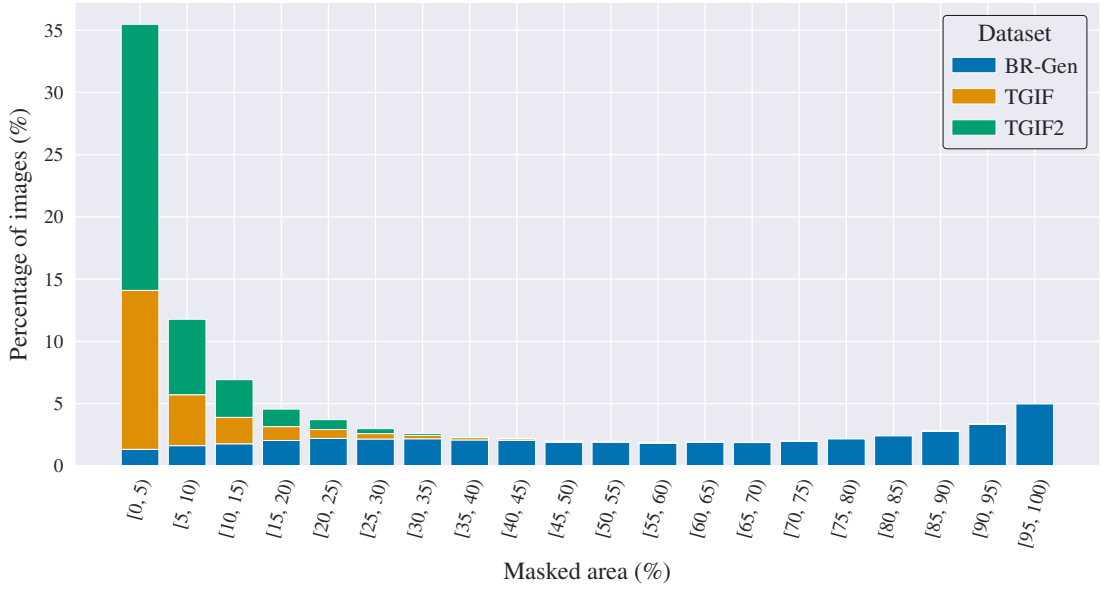


Figure 2: Mask size distribution across the datasets.

Table 2
Average mask size and standard deviation for positive samples by manipulation generator.

Generator	Avg Mask Size (%)	Std (%)
LaMa	56.4	30.1
BrushNet	56.4	30.1
PowerPaint	56.4	30.1
MAT	56.4	30.1
SD2	9.8	12.9
SDXL	36.9	34.4
Firefly	5.5	6.4
Flux.1 schnell	5.5	6.4
Flux.1 dev	5.5	6.4
Flux.1 filldev	5.5	6.4

3.3. Input pipeline

The input pipeline closely follows the design validated in [35]. Previous work has shown that classification performance and cross-generator generalization are maximized when detectors are trained and

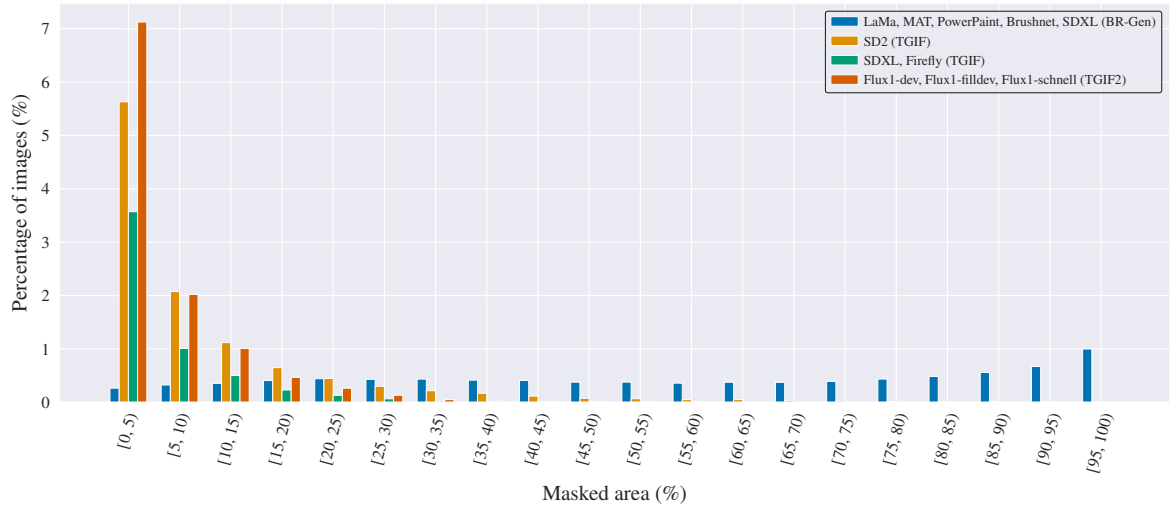


Figure 3: Distribution of masked area percentages across representative generative models. For clarity and compactness, models with identical distributions in terms of number of images per bin have been grouped together, and only one representative model from each group is shown.

evaluated on entire images rather than on cropped or patch-based inputs. Accordingly:

- *Training* - Images are resized to the model's native input resolution and normalized according to the backbone specifications.
- *Evaluation* - Each image is processed in its entirety using a single forward pass, with no test-time augmentation.
- *Inference strategy* - No sliding-window or multi-crop inference is used, allowing us to measure the inherent sensitivity of the model to localized manipulations under realistic constraints.

4. Results

For all experiments, we report the Area Under the Receiver Operating Characteristic curve (AUROC) as the primary evaluation metric. To analyze performance under different conditions, results are disaggregated by:

- *Testing dataset* - BR-Gen, TGIF, and TGIF2.
- *Mask size* - expressed as a percentage of the image area.
- *Inpainting type* - masked inpainting and semantic regeneration.
- *Generator Identity* - SD2, SDXL, Firefly, Flux.1 dev, Flux.1 filldev, Flux.1 schnell, LaMa, BrushNet, PowerPaint, and MAT.
- *Image compression* - uncompressed, JPEG Q80, and JPEG Q60.

4.1. Evaluation on single datasets

In this section, we compare the performance of our models (ResNet-50 CLIP, DINOv2 ViT-L/14 and DINOv3 ViT-L/16) against a range of baseline and state-of-the-art detection approaches reported in the literature across three datasets: BR-Gen, TGIF, and TGIF2.

4.1.1. BR-Gen

Table 3 reports AUROC results for detecting inpainted images within the BR-Gen dataset. Our ResNet-50 CLIP achieves performance comparable to other methods reported in the literature. In contrast,

transformer-based self-supervised features substantially improve detection performance: DINOv2 ViT-L/14 significantly outperforms all prior baselines, achieving an AUROC above 0.83, while DINOv3 ViT-L/16 slightly improves upon its predecessor, reaching an AUROC of 0.86. These findings demonstrate the strong capability of self-supervised vision transformers in distinguishing inpainted content from authentic images.

Table 3
AUROC scores for the BR-Gen dataset. Values for other methods are reported from [31], while results marked as (ours) are obtained from our own evaluation under the same protocol. The highest and second-highest scores are highlighted in **bold** and underline, respectively.

Method	AUROC
LGrad [39]	0.64
Dire [40]	0.48
FreqNet [41]	0.47
NPR [42]	0.50
FatFormer [43]	0.60
ResNet-50 CLIP (ours)	0.64
DINOv2 ViT-L/14 (ours)	0.83
DINOv3 ViT-L/16 (ours)	**0.86**

4.1.2. TGIF

Following the evaluation protocol in [29], Table 4 presents AUROC results for detecting fully regenerated images within the TGIF dataset under different compression settings (uncompressed, JPEG Q80, and JPEG Q60). Our models maintain robust performance across all compression levels. Notably, DINOv2 ViT-L/14 and especially DINOv3 ViT-L/16 consistently achieve high AUROC values, demonstrating strong generalization to varying compression artifacts. The ResNet-50 CLIP model also exhibits competitive performance, particularly in moderately compressed images, indicating resilience to quality degradation.

Table 4
AUROC scores for the TGIF dataset using only fully regenerated images. Results for baseline methods are taken from [29], while methods marked as (ours) are evaluated under the same protocol. The highest and second-highest scores in each column are highlighted in **bold** and underline, respectively.

	Uncompressed		JPEG Q80		JPEG Q60		Average	
Method	SD2	SDXL	SD2	SDXL	SD2	SDXL	SD2	SDXL
CNNDetect [44]	0.57	0.61	0.57	0.62	0.56	0.61	0.57	0.61
DIMD [45]	**1.00**	0.94	**1.00**	**0.95**	**0.99**	**0.92**	**1.00**	**0.94**
Dire [40]	0.49	0.62	0.49	0.61	0.49	0.60	0.49	0.61
FreqDetect [46]	0.50	0.40	0.72	0.71	0.71	0.72	0.64	0.61
UnivFD [47]	0.82	0.80	0.73	0.78	0.70	0.73	0.75	0.77
Fusing [48]	0.55	0.47	0.52	0.59	0.53	0.61	0.53	0.56
GramNet [49]	0.82	0.55	0.48	0.56	0.47	0.54	0.59	0.55
LGrad [39]	0.85	0.83	0.51	0.57	0.51	0.64	0.62	0.68
NPR [42]	0.51	0.67	0.50	0.30	0.51	0.68	0.51	0.55
RINE [50]	0.89	0.89	0.79	0.88	0.75	0.85	0.81	0.87
DeFake [51]	0.65	0.61	0.54	0.45	0.54	0.43	0.58	0.50
PatchCraft [52]	0.98	**0.95**	0.66	0.67	0.60	0.58	0.75	0.73
ResNet-50 CLIP (ours)	0.55	0.65	0.65	0.72	0.64	0.70	0.61	0.69
DINOv2 ViT-L/14 (ours)	0.91	0.86	0.90	0.85	0.88	0.83	0.90	0.85
DINOv3 ViT-L/16 (ours)	0.96	0.92	0.95	0.91	0.93	0.90	0.95	0.91

Table 5 reports AUROC scores on the TGIF dataset when considering all manipulated images jointly,

including both fully regenerated and spliced/inpainted samples. Overall performance is notably lower than in the fully regenerated setting (Table 4), reflecting the increased difficulty of detecting localized manipulations, that often affect only a small portion of the image. As illustrated in Figure 2, inpainted regions in TGIF frequently cover less than 20% of the image area, resulting in weaker global cues.

Among the evaluated methods, DINOv3 ViT-L/16 achieves the highest AUROC, followed closely by DINOv2 ViT-L/14, indicating that large-scale self-supervised vision transformers retain a clear advantage even when multiple manipulation types are aggregated. In contrast, the ResNet-50 CLIP model performs closer to chance level, further highlighting the benefit of transformer-based representations for detecting subtle and heterogeneous image manipulations.

Table 5
AUROC scores for the TGIF dataset.

Model	AUROC
ResNet-50 CLIP	0.54
DINOv2 ViT-L/14	0.66
DINOv3 ViT-L/16	**0.69**

4.1.3. TGIF2

For TGIF2 (Table 6), all evaluated models achieve AUROC scores close to random chance. This behavior can be partly attributed to the use of Flux1 generator variants, which produce images with significantly higher visual quality and fewer detectable artifacts compared to the generative models included in TGIF. As a result, the global image statistics captured by current feature representations provide very limited discriminative power for distinguishing real from synthetic content.

In addition, TGIF2 exhibits a manipulation mask area distribution similar to TGIF, where the majority of spliced or inpainted regions cover less than 20% of the image area 2. The prevalence of such localized and subtle modifications further increases detection difficulty, as global features are weakly affected by small manipulated regions. Compared to TGIF, where fully regenerated images can often be detected with high confidence, TGIF2 therefore poses a substantially harder detection problem due to the combination of stronger generative models and limited spatial extent of the manipulations.

Overall, these results highlight the limitations of existing feature-based detection approaches and underscore the need for representations that are both more sensitive to localized artifacts and more robust to advances in generative image quality.

4.2. Effect of Mask Size

To quantify the impact of manipulation scale, we evaluate detection performance across mask-size bins defined by the percentage of the image area that is replaced. Figure 4 reports the AUROC for each bin.

Detection accuracy correlates positively with mask size: for regions smaller than 5% of the image, all detectors perform near chance. For regions greater than 20% performance improves substantially, with DINOv3 consistently outperforming DINOv2 and ResNet-50, achieving higher AUROC scores, although both remain below their performance on fully generated images in AI-GenBench. For regions covering most of the image, DINOv3 reaches performance comparable to that observed on fully generated images at approximately 65% mask size, whereas DINOv2 requires larger manipulated regions ($> 80\%$) to achieve comparable similar results [35].

These findings suggest that current fully synthetic image detectors rely heavily on global artifacts, which are attenuated or absent in small localized edits.

4.3. Effect of Inpainting Type

We consider two categories of manipulations: masked (pure) inpainting and semantic full regeneration. Table 7 reports AUROC performance stratified by manipulation type.

Table 6
AUROC scores for the TGIF2 dataset.

Model	AUROC
ResNet-50 CLIP	**0.53**
DINOv2 ViT-L/14	0.52
DINOv3 ViT-L/16	0.52

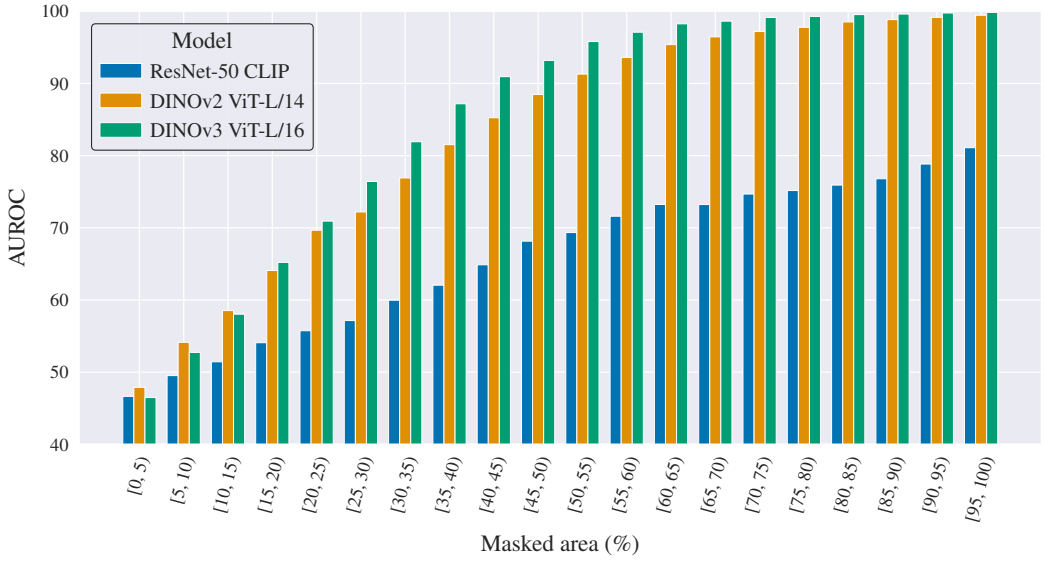


Figure 4: AUROC as a function of mask size (percentage of image area modified). Larger regions yield more detectable artifacts.

Table 7
Overall AUROC for detecting masked inpainting and full regeneration across all generators and mask sizes.

Model	Masked Inpainting	Full Regeneration
ResNet-50 CLIP	0.59	0.53
DINOv2 ViT-L/14	0.70	**0.60**
DINOv3 ViT-L/16	**0.71**	**0.60**

Results indicate that masked inpainting is generally more detectable, particularly for classical or feed-forward models that introduce localized texture inconsistencies. In contrast, full regeneration often yields globally coherent outputs that challenge detectors trained on full-image classification, sometimes producing near-chance performance for small masks (see Section 4.2).

4.4. Effect of Generator Identity

We further analyze cross-generator generalization by evaluating detectors separately on manipulations produced by different generative models. Table 8 reports AUROC performance stratified by generator.

A consistent observation is the difficulty of detecting manipulations produced by models of the Flux and Firefly families, particularly when masks are small. These models generate high-quality texture statistics that are distributionally closer to real content, thereby reducing detectable artifacts. Diffusion-based regenerators are often more difficult to distinguish than masked inpainting networks such as LaMa or MAT, which tend to leave structural or frequency-domain irregularities.

Table 8
AUROC by manipulation generator. Detection performance varies widely across generators, with diffusion-based models often proving the most challenging to identify.

Generator	ResNet-50	DINOv2	DINOv3
LaMa	0.63	0.84	**0.88**
BrushNet	0.67	0.87	**0.89**
PowerPaint	0.69	0.88	**0.90**
MAT	0.70	0.86	**0.89**
SD2	0.55	**0.67**	**0.67**
SDXL	0.64	0.79	**0.83**
Firefly	0.45	**0.46**	0.43
Flux.1 schnell	**0.44**	0.43	0.41
Flux.1 dev	**0.46**	**0.46**	0.43
Flux.1 filldev	**0.49**	0.48	0.45

4.5. Robustness to JPEG Compression

Since a substantial fraction of online images undergo lossy recompression, we evaluate whether JPEG compression affects inpainting detectability. Following the compression settings used in prior work, we assess performance at multiple JPEG quality levels.

Results in Table 9 indicate that JPEG compression does not substantially affect detection accuracy. This finding is consistent with the training strategy proposed in [35], which incorporates aggressive multi-pass JPEG augmentation.

Table 9
AUROC by JPEG compression.

Model	Uncompressed	JPEG Q80	JPEG Q60
ResNet-50	0.57	0.56	0.56
DINOv2	0.67	**0.67**	**0.67**
DINOv3	**0.68**	**0.67**	0.66

4.6. Qualitative Analysis

To better understand the limitations of current detectors, Figure 5 presents qualitative examples of images with low and high predicted manipulation scores.

Detectors tend to succeed on:

- manipulations that introduce clear boundary inconsistencies,
- classical inpainting outputs with texture discontinuities,
- large semantic changes not fully aligned with global context.

They struggle with:

- small inpainting masks,
- high-quality diffusion-based regeneration (especially Flux models),
- semantically coherent edits that preserve lighting and texture statistics.

5. Conclusions

This study examined the extent to which state-of-the-art pre-trained detectors for AI-generated content can generalize from fully synthetic images to localized inpainting. Our results show that detectors



Figure 5: Qualitative examples of inpainted images with corresponding masks, illustrating successful detections (left) and failure cases (right) for the DINOv2 model.

trained on a wide range of generators, such as those included in the AI-GenBench [2], exhibit competitive transferability to the inpainting detection task: even without explicit localization mechanisms (e.g., segmentation or bounding-box prediction), they can often identify whether an image contains inpainted regions.

Performance is particularly strong when the manipulated area is relatively large ($> 50\%$) or when the inpainting approach involves full image regeneration. In these scenarios, our models can surpass state-of-the-art synthetic image detection mechanisms. However, we observe consistent performance degradation for fully regenerated images or when the edited region occupies only a small fraction of the image ($< 20\%$), demonstrating that these cases remain challenging for models trained on the binary classification task (real vs fake).

Overall, the results indicate that, while broad generator exposure and robust augmentation during training enhance generalization, current classification-based detectors still struggle to detect small localized manipulations without specific detection or segmentation mechanisms. Future work should explore hybrid approaches that combine global classification with fine-grained localization to improve robustness against subtle inpainting attacks.

6. Acknowledgments

We acknowledge:

- the support of the European funds from the Emilia-Romagna Region under the Fse+ 2021-2027 program, for Lorenzo Pellegrini.
- ISCRA for awarding this project access to the LEONARDO supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CINECA (Italy).
- IdentifAI for insightful discussions and constructive feedback on the topics addressed in this paper.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Gemini and Copilot in order to: Grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. Thomson, S. McKeown, R. Macfarlane, P. Leimich, Exploring dataset diversity for genai image inpainting localisation in digital forensics, in: Proceedings of the Digital Forensics Doctoral Symposium, DFDS '25, Association for Computing Machinery, 2025. doi:10.1145/3712716.3712724.

[2] L. Pellegrini, D. Cozzolino, S. Pandolfini, D. Maltoni, M. Ferrara, L. Verdoliva, M. Prati, M. Ramilli, Ai-genbench: A new ongoing benchmark for ai-generated image detection, in: 2025 International Joint Conference on Neural Networks (IJCNN), 2025, pp. 1–9. doi:10.1109/IJCNN64981.2025.11228377.

[3] A. Telea, An image inpainting technique based on the fast marching method, Journal of Graphics Tools 9 (2004) 23–34. doi:10.1080/10867651.2004.10487596.

[4] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, V. Lempitsky, Resolution-robust large mask inpainting with fourier convolutions, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2149–2159.

[5] W. Li, Z. Lin, K. Zhou, L. Qi, Y. Wang, J. Jia, Mat: Mask-aware transformer for large hole image inpainting, 2022. URL: https://arxiv.org/abs/2203.15270. arXiv:2203.15270.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 10684–10695.

[7] X. Ju, X. Liu, X. Wang, Y. Bian, Y. Shan, Q. Xu, Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion, 2024. URL: https://arxiv.org/abs/2403.06976. arXiv:2403.06976.

[8] Adobe firefly, https://www.adobe.com/it/products/firefly/landpa.html, ???? Access: 2025-12-03.

[9] J. Zhuang, Y. Zeng, W. Liu, C. Yuan, K. Chen, A task is worth one word: Learning with task prompts for high-quality versatile image inpainting, 2024. URL: https://arxiv.org/abs/2312.03594. arXiv:2312.03594.

[10] Black Forest Labs, Flux, https://github.com/black-forest-labs/flux, 2024. Access: 2025-12-03.

[11] D. Cozzolino, L. Verdoliva, Noiseprint: a cnn-based camera model fingerprint, 2018. URL: https://arxiv.org/abs/1808.08396. arXiv:1808.08396.

[12] R. A. Frick, M. Steinebach, Diffseg: Towards detecting diffusion-based inpainting attacks using multi-feature segmentation, in: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2024, pp. 3802–3808. doi:10.1109/CVPRW63382.2024.00384.

[13] Z. Xu, X. Zhang, R. Li, Z. Tang, Q. Huang, J. Zhang, Fakeshield: Explainable image forgery detection and localization via multi-modal large language models, 2025. URL: https://arxiv.org/abs/2410.02761. arXiv:2410.02761.

[14] H. Wu, J. Zhou, Iid-net: Image inpainting detection network via neural architecture search and attention, IEEE Transactions on Circuits and Systems for Video Technology (2021) 1–1. doi:10.1109/TCSVT.2021.3075039.

[15] J. H. Yu, Z. Lin, J. M. Yang, X. H. Shen, X. Lu, T. S. Huang, Generative image inpainting with contextual attention, in: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., 2018, pp. 5505–5514.

[16] J. H. Yu, Z. Lin, J. M. Yang, X. H. Shen, X. Lu, T. S. Huang, Free-form image inpainting with gated convolution, in: Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 4471–4480.

[17] Z. Y. Yan, X. M. Li, M. Li, W. M. Zuo, S. G. Shan, Shift-net: image inpainting via deep feature rearrangement, in: Proc. Eur. Conf. Comput. Vis., 2018, pp. 1–17.

[18] K. Nazeri, E. Ng, T. Joseph, F. Z. Qureshi, M. Ebrahimi, Edgeconnect: generative image inpainting with adversarial edge learning, in: Proc. IEEE Int. Conf. Comput. Vis. Workshop, 2019.

[19] T. Yu, Z. Guo, X. Jin, S. Wu, Z. Chen, W. Li, Z. Zhang, S. Liu, Region normalization for image inpainting, in: Proc. AAAI Conf. Arti. Intell., 2020, pp. 12733–12740.

[20] A. Telea, An image inpainting technique based on the fast marching method, J. of Graph. Tools 9 (2004) 23–34.

[21] M. Bertalmio, A. L. Bertozzi, G. Sapiro, Navier-stokes, fluid dynamics, and image and video inpainting, in: Proc. IEEE Conf. Comput. Vis. Pattern Recogn., 2001, pp. 355–362.
[22] J. Herling, W. Broll, High-quality real-time video inpaintingwith pixmix, IEEE Trans. Vis. Comput. Graph. 20 (2014) 866–879.
[23] J. Huang, S. Kang, N. Ahuja, J. Kopf, Image completion using planar structure guidance, ACM Trans. Graph. 33 (2014) 1–10.
[24] Q. Guo, S. Gao, X. Zhang, Y. Yin, C. Zhang, Patch-based image inpainting via two-stage low rank approximation, IEEE Trans. Vis. Comput. Graph. 24 (2017) 2023–2036.
[25] G. Mahfoudi, B. Tajini, F. Retraint, F. Morain-Nicolier, J. L. Dugelay, M. Pic, Defacto: Image and face manipulation dataset, in: 2019 27th European Signal Processing Conference (EUSIPCO), 2019, pp. 1–5. doi:10.23919/EUSIPCO.2019.8903181.
[26] H. Aboutalebi, D. Mao, R. Fan, C. Xu, C. He, A. Wong, Deepfakeart challenge: A benchmark dataset for generative ai art forgery and data poisoning detection, 2024. URL: https://arxiv.org/abs/2306.01272. arXiv:2306.01272.
[27] P. J. Rösch, N. Oswald, M. Geierhos, J. Libovický, Enhancing conceptual understanding in multimodal contrastive learning through hard negative samples, 2024. arXiv:2403.02875.
[28] G. Bertazzini, C. Albisani, D. Baracchi, D. Shullani, A. Piva, Beyond the brush: Fully-automated crafting of realistic inpainted images, in: 2024 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2024, pp. 1–6. doi:10.1109/WIFS61860.2024.10810722.
[29] H. Mareen, D. Karageorgiou, G. Van Wallendael, P. Lambert, S. Papadopoulos, Tgif: Text-guided inpainting forgery dataset, in: Proc. Int. Workshop on Information Forensics and Security (WIFS) 2024, 2024.
[30] IDLabMedia, Tgif: Text-guided inpainting forgery dataset, https://github.com/IDLabMedia/tgif-dataset, 2024. Access: 2025-12-03.
[31] L. Cai, H. Wang, J. Ji, Y. ZhouMen, Y. Ma, X. Sun, L. Cao, R. Ji, Zooming in on fakes: A novel dataset for localized ai-generated image detection with forgery amplification approach, 2025. URL: https://arxiv.org/abs/2504.11922. arXiv:2504.11922.
[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, 2021. URL: https://arxiv.org/abs/2103.00020. arXiv:2103.00020.
[33] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, P. Bojanowski, Dinov2: Learning robust visual features without supervision, 2024. URL: https://arxiv.org/abs/2304.07193. arXiv:2304.07193.
[34] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, F. Massa, D. Haziza, L. Wehrstedt, J. Wang, T. Darcet, T. Moutakanni, L. Sentana, C. Roberts, A. Vedaldi, J. Tolan, J. Brandt, C. Couprie, J. Mairal, H. Jégou, P. Labatut, P. Bojanowski, Dinov3, 2025. URL: https://arxiv.org/abs/2508.10104. arXiv:2508.10104.
[35] L. Pellegrini, S. Pandolfini, D. Maltoni, M. Ferrara, M. Prati, M. Ramilli, Generalized design choices for deepfake detectors, 2025. URL: https://arxiv.org/abs/2511.21507. arXiv:2511.21507.
[36] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, P. Dollár, Microsoft coco: Common objects in context, 2015. URL: https://arxiv.org/abs/1405.0312. arXiv:1405.0312.
[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.
[38] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: A 10 million image database for scene recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (2018) 1452–1464. doi:10.1109/TPAMI.2017.2723009.
[39] C. Tan, Y. Zhao, S. Wei, G. Gu, Y. Wei, Learning on gradients: Generalized artifacts representation for gan-generated images detection, in: Proceedings of the IEEE/CVF Conference on Computer

Vision and Pattern Recognition, 2023, pp. 12105–12114.
[40] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, H. Li, Dire for diffusion-generated image detection, 2023. URL: https://arxiv.org/abs/2303.09295. arXiv:2303.09295.
[41] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, Y. Wei, Frequency-aware deepfake detection: Improving generalizability through frequency space learning, 2024. URL: https://arxiv.org/abs/2403.07240. arXiv:2403.07240.
[42] C. Tan, H. Liu, Y. Zhao, S. Wei, G. Gu, P. Liu, Y. Wei, Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection, 2023. URL: https://arxiv.org/abs/2312.10461. arXiv:2312.10461.
[43] H. Liu, Z. Tan, C. Tan, Y. Wei, Y. Zhao, J. Wang, Forgery-aware adaptive transformer for generalizable synthetic image detection, 2023. URL: https://arxiv.org/abs/2312.16649. arXiv:2312.16649.
[44] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, A. A. Efros, Cnn-generated images are surprisingly easy to spot... for now, 2020. URL: https://arxiv.org/abs/1912.11035. arXiv:1912.11035.
[45] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, L. Verdoliva, On the detection of synthetic images generated by diffusion models, 2022. URL: https://arxiv.org/abs/2211.00680. arXiv:2211.00680.
[46] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, T. Holz, Leveraging frequency analysis for deep fake image recognition, 2020. URL: https://arxiv.org/abs/2003.08685. arXiv:2003.08685.
[47] U. Ojha, Y. Li, Y. J. Lee, Towards universal fake image detectors that generalize across generative models, 2024. URL: https://arxiv.org/abs/2302.10174. arXiv:2302.10174.
[48] Y. Ju, S. Jia, L. Ke, H. Xue, K. Nagano, S. Lyu, Fusing global and local features for generalized ai-synthesized image detection, 2022. URL: https://arxiv.org/abs/2203.13964. arXiv:2203.13964.
[49] Z. Liu, X. Qi, P. Torr, Global texture enhancement for fake face detection in the wild, 2020. URL: https://arxiv.org/abs/2002.00133. arXiv:2002.00133.
[50] C. Koutlis, S. Papadopoulos, Leveraging representations from intermediate encoder-blocks for synthetic image detection, 2024. URL: https://arxiv.org/abs/2402.19091. arXiv:2402.19091.
[51] Z. Sha, Z. Li, N. Yu, Y. Zhang, De-fake: Detection and attribution of fake images generated by text-to-image generation models, 2023. URL: https://arxiv.org/abs/2210.06998. arXiv:2210.06998.
[52] N. Zhong, Y. Xu, S. Li, Z. Qian, X. Zhang, Patchcraft: Exploring texture patch for efficient ai-generated image detection, 2024. URL: https://arxiv.org/abs/2311.12397. arXiv:2311.12397.