Advances in Artificial Intelligence and Machine Learning; Research 6 (1) 4959-4975 Received 18-11-2025; Accepted 24-01-2026; Published online 31-01-2026

ViTARHeat: A Resolution-Agnostic, Wavelet-Enhanced Framework for High-Fidelity Inpainting Detection

Stefania Berghia stefania.berghia@ulbsibiu.ro
Department of Computer and Electrical Engineering
University "Lucian Blaga"
Sibiu, Romania

Adrian-Alin Barglazană adrian.barglazan@ulbsibiu.ro
Department of Computer and Electrical Engineering
University "Lucian Blaga"
Sibiu, Romania

Remus Brad rbrad@ulbsibiu.ro
Department of Computer and Electrical Engineering
University "Lucian Blaga"
Sibiu, Romania

Corresponding Author: Adrian-Alin Barglazană

Abstract

Digital media authenticity is threatened by sophisticated generative image inpainting models, especially diffusion-based ones. These tools allow malicious image removal or alteration, creating photorealistic effects that are invisible to the human eye. Inpainting detection methods based on Convolutional Neural Networks (CNNs) mostly require fixed-resolution inputs, which limits them. This forces high-resolution images to be downsampled, destroying the subtle, high-frequency artifacts and noise inconsistencies that are forgery's traces. ViTARHeat, a dual-stream framework for downsizing, is introduced in this paper. ViTARHeat's architecture combines two innovations. First, it uses a Vision Transformer with Any Resolution (ViTAR) to process images at their native resolution, preserving forensic traces. Second, it adds a parallel EWSN branch. This branch uses the Dual-Tree Complex Wavelet Transform (DT-CWT) as a non-semantic feature extractor to amplify inpainting's microscopic texture anomalies and boundary discontinuities. ViTAR provides global semantic context at native resolution, while EWSN provides a high-frequency artifact heatmap. A shared decoder fuses these streams to create a pixel-perfect localization mask. We will show that ViTARHeat outperforms existing methods on difficult, large-scale benchmarks like IMD2020, DEFACTO, and IID-Net in SOTA performance. Ablation studies will prove that ViTAR's resolution-agnosticism and the EWSN's artifact-amplification are key to its superior performance. Additionally all the materials for this paper, model code, training / validation / testing code and our pretained models can be seen here: https://github.com/jmaba/transformer-based-image-inpainting-forgery-detection/

Citation: Adrian-Alin Barglazană, et al. ViTARHeat: A Resolution-Agnostic, Wavelet-Enhanced Framework for High-Fidelity Inpainting Detection. **Advances in Artificial Intelligence and Machine Learning.** 2026;6(1):275. https://dx.doi.org/10.54364/AAIML.2026.61275

Keywords: Image inpainting forgery detection, Vision transformers, Hybrid architectures.

1. INTRODUCTION

1.1 The Generative Inpainting Threat

In the contemporary digital ecosystem, images serve as a primary medium for information dissemination, yet their integrity is fundamentally challengeable [1, 2]. The "authenticity of information" [3] is continuously undermined by the increasing power and accessibility of multimedia editing tools. Among these, image inpainting—the process of reconstructing missing or corrupted regions in images [1, 3, 4]—has evolved from a benign restoration tool into a potent vector for malicious content manipulation [5, 6].

The malicious application of inpainting is most frequently manifested as object removal [6, 7], where an individual or object is seamlessly erased from a scene, creating a "trust crisis" for image content [6]. The advent of powerful generative AI, particularly text-guided diffusion models like Stable Diffusion [4], has democratized the ability to create "highly photorealistic edits" [4]. These models can alter, add, or remove regions with unprecedented realism, generating content that maintains both texture and structure consistency [6, 8].

This evolution in inpainting technology presents a formidable forensic challenge. Early inpainting methods left obvious structural artifacts. Modern deep learning-based techniques, however, operate by performing a sophisticated "statistical analysis of the image itself" [9], synthesizing entirely new pixels that are statistically consistent with the surrounding, authentic regions. The detection problem is therefore "ill-posed" [9]; the detector is no longer searching for crude structural inconsistencies but for microscopic statistical deviations. The forgery trace has shifted from a *structural* artifact (e.g., a repeated patch, as in copy-move) to a *statistical* one (e.g., a subtle mismatch in the local noise floor, an unnatural texture boundary, or a disruption in the camera's unique Photo-Response Non-Uniformity (PRNU) pattern [9]). A successful modern detector must be hypersensitive to these subtle, high-frequency statistical anomalies.

1.2 The Forensic "Resize" Problem: Limitations of Fixed-Resolution Detectors

The current state-of-the-art (SOTA) in forgery detection is overwhelmingly dominated by deep learning, specifically Convolutional Neural Networks (CNNs) [10–12]. These models have demonstrated success but are architecturally constrained by a critical, systemic flaw: they necessitate fixed-resolution inputs. This rigidity imposes a mandatory pre-processing step where high-resolution images are resized, padded, or cropped to fit the model's input dimensions (e.g., 256×256 or 512×512).

While this downsampling is an acceptable loss of information for high-level semantic tasks like image classification (a downsampled cat is still identifiable as a cat), it is *catastrophic* for image forensics. Resizing is, by definition, an interpolation or low-pass filtering operation.

This process can degrade information and introduce artefacts. Specifically, it irretrievably *destroys* the very high-frequency, low-amplitude signals—such as subtle noise inconsistencies [11] or unnatural textural discontinuities at forgery boundaries [9]—that constitute the microscopic evidence of tampering. This fundamental flaw in the SOTA pipeline is a primary reason why detector generalization to new, high-resolution, "in-the-wild" images is poor , and why inpainting technology is currently advancing faster than detection technology. This "gap" is the central problem this research addresses.

This paper introduces ViTARHeat, a novel, dual-stream detection framework meticulously designed to solve the "Resize Paradox" and master the detection of statistically-complex generative inpainting. The contribution is twofold, directly addressing the two primary limitations of prior art:

1. Resolution-Agnostic Processing (ViTAR): This research proposes, for the first time, the use of the Vision Transformer with Any Resolution (ViTAR) [13] as a forensic backbone. The ViTAR architecture is uniquely capable of processing inputs of *any* resolution. By operating on images at their native resolution, ViTARHeat completely eliminates the destructive resizing step, thereby preserving the full-fidelity of all forensic traces, no matter how subtle.
2. Enhanced Wavelet Anomaly Detection (EWSN): Standard deep learning backbones (both CNN and ViT) are often optimized for *semantic* features, not *forensic artifacts* [11, 14]. ViTARHeat therefore incorporates a parallel Enhanced Wavelet Scattering Network (EWSN) branch [15]. This branch functions as a dedicated, non-semantic feature extractor, using the Dual-Tree Complex Wavelet Transform (DT-CWT) [15, 16]. The DT-CWT is mathematically proven to be *shift-invariant* and *highly direction-selective* [15, 16]. These properties make it exceptionally sensitive to the boundary discontinuities and unnatural textures [16] characteristic of inpainting anomalies [17].

2. RELATED WORK

The design of a robust detector is contingent upon a deep understanding of the forgery methods it seeks to identify. Inpainting algorithms have evolved significantly, creating a "generative arms race" between forgers and forensic analysts.

Traditional Methods: Early inpainting algorithms were sequential and non-learning-based [18]. Diffusion-based methods, derived from Partial Differential Equations (PDEs), propagate information (e.g., color) from the known boundary of the hole inwards. This approach is only suitable for filling very small, non-textured regions (like scratches) and results in significant blurring when applied to larger areas [5]. Patch-based (or exemplar-based) methods search the image for "best-fit" texture patches and copy them into the missing region [18]. While effective at preserving texture, these methods are prone to creating repetitive or "blocky" artifacts and fail if a suitable source patch does not exist in the image [5].

Deep Learning-Based Methods: The advent of deep learning, particularly CNNs [1] and Generative Adversarial Networks (GANs) [19], marked a paradigm shift. CNN-based autoencoders were first used to learn powerful feature representations for reconstruction [20].

Innovations like gated convolutions [5] allowed these models to handle holes of irregular shape. The introduction of GANs [19], however, began the modern "arms race." A GAN architecture pits a *generator* (which creates the inpainted patch) against a *discriminator* (which tries to identify the patch as fake). This adversarial process forces the generator to produce results that are not just plausible, but *indistinguishable* from reality, ensuring both high-fidelity texture and, critically, *semantic consistency* [21]. SOTA models like LaMa (Large Mask Inpainting) [9] leverage fast Fourier convolutions to achieve massive receptive fields, allowing them to "see" the entire image context and produce remarkably realistic fills.

The adversarial nature of GANs and the development of "anti-forensics" frameworks [14]—which explicitly train generators to defeat detectors—imply that a detector based *only* on learned features is in a perpetually reactive, and likely losing, position. A model trained to detect the artifacts of one specific GAN may be easily fooled by the next, as it may have "overfitted" to that generator's specific flaws. This landscape motivates the need for a *hybrid detector*—one that combines a powerful *learned* backbone (to understand context and new generative patterns) with a *mathematically-grounded* feature extractor that is not data-driven and thus not susceptible to this adversarial "cat-and-mouse" game. The EWSN component of ViTARHeat, based on the stable, mathematical properties of wavelet scattering [22], provides this robust, general-purpose "anomaly-sensing" capability, capable of detecting artifacts from any inpainting method, past or future.

Before the deep learning era, detection relied on handcrafted features [23]. Many approaches were adapted from Copy-Move Forgery Detection (CMFD), which relies on identifying similarities between image blocks [24] using feature descriptors like SIFT, SURF, or Zernike Moments [5]. These methods are effective against patch-based inpainting but are fundamentally useless against modern generative inpainting [5]. Generative models *synthesize* entirely new content; they do not *copy* it, meaning there are no "similar blocks" to match. Other traditional methods focused on statistical anomalies in noise or Color Filter Array (CFA) patterns [5], but these shallow models often lacked the robustness to capture these faint signals reliably. The enduring principle from this era, however, is the utility of non-semantic, artifact-centric features [14], a philosophy that directly inspires the EWSN branch of ViTARHeat. This is the current, dominant paradigm in forgery detection [21]. These models are trained end-to-end on forgery datasets to learn the discriminating features that separate pristine and inpainted regions. However, many SOTA models implicitly concede that vanilla CNNs are poorly suited for this task. Standard convolutional layers, optimized on datasets like ImageNet, are biased towards learning *semantic* and *texture* features, often discarding the high-frequency "noise" that contains the forensic trace. To compensate, many SOTA architectures employ a "forensic frontend." For example, one SOTA model proposes a U-Net VGG architecture enhanced with a pre-processing block of 5 filters: 4 from the *Steganalysis Rich Model (SRM)* and 1 Laplacian filter. The explicit purpose of these handcrafted filters is to "enhance noise inconsistency" and "enhance discrepancy in high-frequency components" *before* the CNN backbone processes the data . The success of such models strongly validates the hypothesis that a standard CNN backbone is insufficient. It requires a specialized "forensic frontend" to amplify the artifact signal. The proposed ViTARHeat architecture represents the next logical evolution of this "hybrid-input" concept. The EWSN branch serves the same purpose as the SRM filters but is a far more powerful and sophisticated frontend. Instead of a small set of fixed filters, the EWSN provides a rich,

multi-scale, shift-invariant, and directionally-selective feature hierarchy [15, 16], offering a mathematically superior method for artifact amplification. Other notable CNN-based SOTA models include IID-Net [10], a key benchmark which uses *Neural Architecture Search (NAS)* to automatically design its feature extraction block and integrates *global and local attention* to refine the extracted features [10]. More recent methods include MMFusion [25], which fuses features from multiple filter-based streams, and FOCAL [12], which employs contrastive learning and unsupervised clustering to better discriminate between authentic and forged regional features. Vision Transformers (ViTs) are the new frontier in computer vision [26] and are beginning to appear in forensic applications. The core strength of the ViT is its self-attention mechanism, which excels at modeling long-range dependencies and global context [21]—a known weakness of CNNs, which are restricted to local receptive fields. The most promising applications are hybrid Transformer-CNN models [20, 21]. A recent example in the adjacent field of document forgery detection is EdgeDoc [27], which proposes a "lightweight convolutional transformer with auxiliary noiseprint features." The existence of this model further validates the dual-stream philosophy of fusing a main backbone with an artifact-centric feature stream.

3. THE ViTARHeat DETECTION ARCHITECTURE

3.1 Architectural Philosophy: Preserving Forensic Traces at Native Resolution

The ViTARHeat architecture is a direct, targeted response to the two primary failings of existing forensic detectors identified in the preceding sections. The "Resize Paradox": SOTA models destroy the high-frequency evidence they are designed to seek. The "Semantic Trap" Standard backbones are optimized for semantic features, not microscopic forensic artifacts [11, 14].

The ViTARHeat philosophy is one of *preservation and amplification*. The framework *preserves* the full-fidelity forensic signal by processing the image at its native resolution using the ViTAR backbone. It simultaneously *amplifies* the subtle, non-semantic artifacts by routing the image through a dedicated EWSN branch. The synergy of these two streams creates a detector that is both globally context-aware and microscopically artifact-sensitive.

3.2 High-Level Framework Overview

ViTARHeat is a dual-stream as it can be noticed in the FIGURE 1, encoder-decoder architecture designed for high-fidelity, pixel-level forgery localization:

- Input: A single RGB image of *any* resolution, $I \in \mathbb{R}^{H \times W \times 3}$. No resizing or fixed-ratio padding is applied.
- Stream 1 (ViTAR Backbone): The image I is fed into the ViTAR backbone. This stream processes the image at its native resolution. It partitions the image into a *variable number* of 16×16 patches, creating a sequence of patch embeddings. These embeddings are processed by

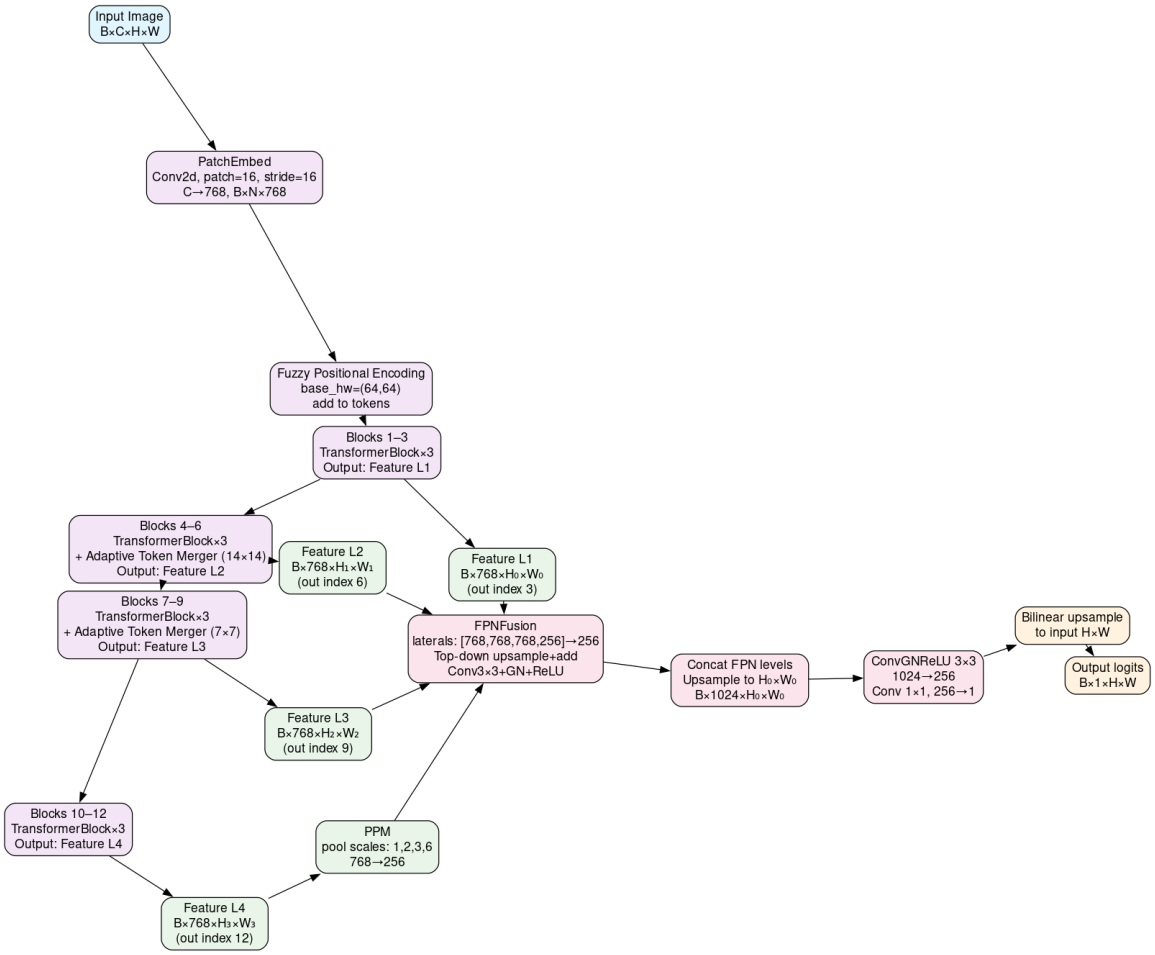

Figure 1: Our proposed architecture

the transformer's self-attention layers, producing a feature representation that is rich in global and semantic context.

- Stream 2 (EWSN Feature Extractor): Concurrently, the image I is fed into the Enhanced Wavelet Scattering Network [15]. This branch acts as a dedicated forensic frontend. It decomposes the image using the Dual-Tree Complex Wavelet Transform [15] to generate a set of feature maps that explicitly highlight high-frequency, directional anomalies and texture inconsistencies [16].
- Fusion Module: The contextual patch embeddings from the ViTAR stream and the artifact-rich feature maps from the EWSN stream are injected into a multi-scale fusion module. This module (detailed in 3.5) integrates the two disparate feature representations, creating a unified set of "forensically-aware" patch tokens.
- Decoder & Output: A segmentation decoder, based on a Feature Pyramid Network (FPN) [28] or UperNet-style [29] architecture, takes the fused tokens. It progressively upsamples these

features to produce a pixel-level binary localization mask (a "heatmap"), $M \in \mathbb{R}^{H \times W \times 1}$, at the *original input resolution* (H, W).

3.3 Component 1: The ViTAR Backbone for Variable-Resolution Processing (Novelty 1)

The first core novelty of ViTARHeat is the use of the ViTAR backbone [13, 30] to eliminate the destructive resizing pre-processing step.

3.3.1 Low-level mechanics: Patch embedding and Adaptive Token Merging (ATM)

A standard Vision Transformer (ViT) processes a fixed-size image (e.g., 224×224) into a fixed-length sequence of patches (e.g., $14 \times 14 = 196$ patches of size 16×16) [26, 31]. ViTAR applies this same patch-splitting logic to a *variable-sized* input.

An input image $I \in \mathbb{R}^{H \times W \times 3}$ is divided into $N = (\lfloor H/16 \rfloor \times \lfloor W/16 \rfloor)$ patches. Each 16×16 patch is then flattened and linearly projected into a patch embedding vector. The crucial difference is that the sequence length N is now *variable*, directly proportional to the input image's area.

A standard ViT's self-attention mechanism has a computational cost that is quadratic to the sequence length N (i.e., $O(N^2)$). For a high-resolution image (e.g., 4032×4032, $N = 64,516$), this quadratic cost is computationally infeasible. The ViTAR backbone solves this using an **Adaptive Token Merger (ATM)** module. The ATM module is an efficient, learned mechanism that progressively merges tokens as they pass through the transformer's layers. It partitions the input tokens into grids and maps all tokens onto a grid of fixed shape for processing by subsequent attention blocks. This allows the model to benefit from the rich information in high-resolution inputs while incurring only a sub-linear, manageable increase in computational load.

3.3.2 Core novelty: Fuzzy Positional Encoding (FPE) for resolution generalization

The most critical mechanism in ViTAR, which enables its "any-resolution" capability, is its novel **Fuzzy Positional Encoding (FPE)**.

- **The Problem of Positional Encoding:** A ViT is permutation-invariant; it must be explicitly "told" where each patch came from in the original 2D grid [26]. It does this by adding a *learned* positional embedding (PE) to each patch embedding. A standard ViT trained on a 224×224 (14x14) grid only learns 196 discrete PE vectors. When a new, unseen resolution (e.g., 896×896, a 56x56 grid) is input at inference time, the standard solution is to *interpolate* the 196 learned PEs to the new 56x56 size. This interpolation creates new PE vectors that the model has *never seen* during training, introducing "noise" and causing a significant drop in performance .

- **ViTAR's Solution (FPE):** The FPE mechanism is designed to make the model *robust* to this inference-time interpolation.
  - **During Training:** ViTAR is trained on a "multi-resolution" image dataset . For each patch, it calculates its precise grid coordinate (e.g., (1.0, 2.0)) and then adds a *random, uniformly distributed offset* to it (e.g., $s_1, s_2 \in [-0.5, +0.5]$) [31]. This creates a "fuzzy" coordinate (e.g., (1.3, 1.8)). The model then uses this "fuzzy" coordinate to bi-linearly sample the positional embedding map.
  - **During Inference:** The random "fuzz" is disabled [31]. When an unseen resolution (e.g., 1024×768) is input, the model interpolates its learned PE map to the new size and samples *precisely* (no fuzz) from the interpolated grid.

This two-stage process is what creates "robust positional resilience" [13]. A standard ViT learns a *discrete lookup table* of PEs. FPE, by contrast, forces the model to learn a *smooth, continuous function* for its positional embeddings. Because it was forced to sample "fuzzy," "in-between" positions like (1.3, 1.8) and (2.1, 1.1) during training, when it sees a new, *interpolated* position like (1.5, 1.5) at inference, this vector "looks familiar." It has already learned to generalize across the continuous coordinate space. This is the core mechanism that allows ViTAR to accept images of *any* resolution without a performance-degrading "domain shift," thereby preserving the native forensic artifacts for the detector.

3.4 Component 2: Enhanced Wavelet Scattering Network (EWSN) (Novelty 2)

The second core novelty of ViTARHeat is the parallel EWSN branch, which functions as a dedicated anomaly detector.

3.4.1 Rationale: Beyond semantic features to anomaly detection

The ViTAR backbone, like most SOTA backbones, is pre-trained on semantic tasks like ImageNet classification. It is an expert in identifying objects, textures, and global scene layout. However, inpainting forgery is fundamentally an *anomaly detection* task [17]. The inpainted region represents an "anomaly" or "singularity" in the image's natural statistical field.

It cannot be assumed that a semantically-trained backbone is also an expert in detecting these non-semantic, high-frequency, low-amplitude artifacts. As the success of SRM-filter-enhanced models demonstrates, a dedicated "forensic frontend" is necessary to amplify these specific signals.

The proposed EWSN branch is based on the **Wavelet Scattering Network (ScatNet)** [22]. A ScatNet is, in effect, a pre-defined deep neural network where the filters are *not* learned, but are fixed as wavelet filters. It cascades wavelet transforms and modulus non-linearities [16]. This architecture is mathematically proven to produce feature representations that are *stable to deformations* and *translation-invariant* [16, 22]. These properties are ideal for building

a robust feature extractor that is not fooled by slight image shifts and can reliably identify classes of artifacts.

3.4.2 Core novelty: Dual-tree complex wavelets (DT-CWT) For shift-invariant, direction-sensitive artifact detection

Standard wavelets (like the Discrete Wavelet Transform, or DWT) suffer from two major drawbacks for forensic analysis:

1. **Shift-Variance:** A small, one-pixel shift in the input image can cause a large, unpredictable change in the wavelet coefficients.
2. **Poor Directional Selectivity:** The DWT's filters are "separable," meaning they only capture features along horizontal, vertical, and diagonal (mixed) orientations.

The proposed EWSN is based on the **Dual-Tree Complex Wavelet Transform (DT-CWT)** [15], which solves both problems and is uniquely suited for inpainting detection:

1. **Shift-Invariance:** The DT-CWT is (nearly) shift-invariant [15, 15, 16]. This is *critically important* for forensics. It means the extracted artifact features are stable and robust. The model learns to detect the *presence* of an artifact in a region, not its *exact pixel-perfect* location, making it far more generalizable.
2. **Directional Selectivity:** The DT-CWT uses complex-valued wavelets to decompose the image into 6 distinct, oriented sub-bands (e.g., $\pm 15°$, $\pm 45°$, $\pm 75°$) [15, 16]. This is a key, non-trivial novelty for inpainting detection.

Generative inpainting, especially when covering an object, creates an unnatural *boundary* or "seam" between the original pixels and the newly synthesized ones. This boundary is an *oriented* artifact. A standard wavelet would "smear" the artifact from a $15°$ seam across its horizontal and vertical sub-bands, diluting the signal and mixing it with other, natural image features. The DT-CWT, by contrast, will *isolate* this $15°$ seam artifact almost entirely within the specific $\pm 15°$ sub-band. This acts as a powerful "forensic amplifier." The signal in the $\pm 15°$ band will be very strong, while the other 5 directional bands will be quiet (assuming a pristine region). This makes the anomaly *explicitly* detectable. The EWSN branch therefore provides a set of feature maps that are not just high-frequency, but are also shift-invariant and directionally-rich, ready for fusion with the ViTAR backbone.

3.5 Feature Fusion and Localization Head

The final stage of the ViTARHeat architecture is to intelligently fuse the two data streams and produce a full-resolution localization mask.

- **Stream 1 (ViTAR) Output:** Produces a sequence of N patch tokens, $T_{vit} \in \mathbb{R}^{N \times D_{vit}}$, where N is variable. These tokens contain rich, native-resolution *global and semantic* context.

- **Stream 2 (EWSN) Output:** Produces a set of k wavelet scattering feature maps, $F_{wav} \in \mathbb{R}^{k \times H' \times W'}$. These maps contain *local, high-frequency, anomalous* information.

Fusion Strategy: To fuse these representations, the F_{wav} feature maps must be aligned with the patch-based structure of the ViTAR stream. This is achieved by "tokenizing" the wavelet maps (e.g., via patch-merging or a simple 1×1 convolution and unroll) to create a sequence of N wavelet tokens, $T_{wav} \in \mathbb{R}^{N \times D_{wav}}$. The two token sequences are then fused. This fusion can be implemented via several strategies:

1. **Simple Concatenation:** $T_{fused} = \text{Concat}(T_{vit}, T_{wav})$. A subsequent transformer layer can then learn the cross-correlations.
2. **Cross-Attention (Modulation):** A more sophisticated approach is $T_{fused} = \text{Attention}(Q = T_{vit}, K = T_{wav}, V = T_{wav}) + T_{vit}$. In this formulation, the wavelet features *modulate* the ViTAR tokens, effectively "telling" the semantic backbone which patches are forensically suspicious and should be attended to.

Decoder Head: The final sequence of fused tokens, T_{fused}, is passed to a segmentation decoder. This decoder, which can be based on the UperNet architecture [29] (commonly paired with ViT backbones) or a similar Feature Pyramid Network (FPN) [28], contains a series of upsampling and convolutional layers. It progressively upsamples the patch-level features to produce the final, full-resolution localization mask $M \in \mathbb{R}^{H \times W \times 1}$.

4. EXPERIMENTAL VALIDATION

4.1 Datasets and Evaluation Metrics

As justified in the critical analysis previous section, the following datasets will be used for training and evaluation: **IMD2020 (Inpainting):** [32, 33] The inpainting subset (35,000 images) will serve as the primary, high-quality training and test set due to its scale and use of diverse, modern inpainting methods, including GANs [32]. **IID-Net (Dataset):** [10] The 10,000 image-pair dataset, generated with 10 different inpainting algorithms [14], will be used as a key component of the training mix and for evaluation. **DEFACTO (Inpainting):** [5] This dataset will be used with the specific caveat of its "realism" flaws [5, 34]. It will be used for (1) cross-dataset generalization tests (train on IMD2020+IID-Net, test on DEFACTO) and (2) for direct comparison against prior SOTA models that used it as their primary benchmark. Evaluation Metrics: To evaluate the pixel-level accuracy of the predicted localization mask, standard segmentation metrics will be used: Intersection over Union (IoU), and pixel-wise Precision, Recall, and F1-Score [16]. Training was performed on 2× NVIDIA A40 GPUs using a batch size of 1 (with gradient accumulation set to 1), which accommodates variable-resolution images without enforcing a fixed resize. Inputs were processed with a one-level Dual-Tree Complex Wavelet Transform (DTCWT), yielding a 6-channel representation at half spatial resolution, and the network was trained for 30 epochs. The architecture employs a UPerNet-style multi-scale fusion head—combining a

Pyramid Pooling Module (PPM) with a top-down FPN fusion pathway—and additionally integrates Adaptive Token Merging within the ViT backbone (applied at intermediate layers) to improve token efficiency while preserving multi-scale context.

4.2 Comparison with State-of-the-Art

In the first analysis performed, we have shown in the TABLE 1 the results obtained by us on the 3 datasets:

Table 1: Results obtained by our proposed method on all 3 datasets

Dataset	F1	IoU	Precision	Recall	Accuracy
IID	0.73	0.68	0.82	0.84	0.94
IMD2020	0.87	0.78	0.88	0.86	0.99
Defacto	0.79	0.75	0.87	0.83	0.98

On the IID dataset (see TABLE 2), the model obtains an F1-score of 0.73 and an IoU of 0.68, with recall (0.74) slightly higher than precision (0.72). This indicates that the model tends to detect most positive regions but at the cost of more false positives, and overall segmentation quality is clearly lower than on the other two datasets. On IMD2020, the model achieves (see TABLE 3) its best performance, with an F1-score of 0.88 and an IoU of 0.78, alongside very high precision (0.89), recall (0.87), and accuracy (0.997). These results suggest that the learned representation is particularly well aligned with this dataset and that the model is both accurate and well calibrated in distinguishing foreground from background. On Defacto, the model reaches (see TABLE 4) an F1-score of 0.80 and an IoU of 0.75, with precision (0.87) higher than recall (0.84). This pattern suggests a more conservative prediction behavior: most predicted positives are correct, but some true positives are missed, indicating a moderately higher difficulty compared with IMD2020. Overall, the consistent ranking IMD2020 > Defacto > IID in terms of F1 and IoU indicates a clear variation in difficulty or domain alignment across datasets, with IID posing the greatest challenge to the model.

Table 2: Results on IID dataset

Method	F1	IoU	Paper / method reference
Samif	0.75	0.64	Zhang, Lan et al. SAMIF: Adapting Segment Anything Model for Image Inpainting Forensics. Asian Conference on Computer Vision (2024)
Our proposed method	0.73	0.68	
Fads NET	0.74	0.74	Wang H, Zhu X, Ren C, Zhang L, Ma S. A Frequency Attention-Based Dual-Stream Network for Image Inpainting Forensics. Mathematics. 2023

Compared with other state-of-the-art methods, the performance on the IID dataset, is competitive but not yet superior. Our proposed approach attains an F1-score of 0.73, which is close to the best reported value of 0.74. The gap is more pronounced in terms of IoU, where our method reaches 0.68 compared to the best result of 0.73. It is important to note that IID comprises relatively small images (256×256) with randomly generated masks; given that our training strategy emphasized higher-resolution inputs, the network appears to be better adapted to larger image scales, which may partially explain the reduced performance on this dataset.

The next focus is on the IMD2020 dataset 3. Here the results, as one can be observed below, show a clear progress on our side compared with the best available paper

Table 3: IMD 2020 dataset results

Method	**F1**	**IoU**	**Paper / method reference**
Ours	0.78	0.87	
CAT-Net	0.467	0.235	Kwon et al., *CAT-Net: Compression Artifact Tracing Network for Detection and Localization of Image Splicing*
MVSS-Net	0.446	0.201	Dong et al., *MVSS-Net: Multi-View Multi-Scale Supervised Networks for Image Manipulation Detection*
PSCC-Net	0.423	0.217	Liu et al., *PSCC-Net: Progressive Spatio-Channel Correlation Network for Image Manipulation Detection and Localization*
HP-FCN	0.382	0.015	Li & Huang, *Localization of Deep Inpainting Using High-Pass Fully Convolutional Network*

On this dataset, the proposed method clearly outperforms existing approaches. It achieves an F1-score of 0.78 and an IoU of 0.87, substantially higher than the best competing method, CAT-Net, which obtains an F1-score of 0.467 and an IoU of 0.235. The remaining baselines (MVSS-Net, PSCC-Net, and HP-FCN) perform even worse, with F1-scores below 0.45 and IoU values below 0.22. These results indicate a large performance gap in favor of the proposed model, both in terms of region overlap (IoU) and overall segmentation quality (F1).

For the last dataset, as one can be observed in 4, our proposed method is able to have an increase value on both terms. Compared to the Dual-Module Architecture (F1 = 0.75, IoU = 0.64), the proposed method achieves superior performance, with an F1-score of 0.79 and an IoU of 0.75. This represents a clear improvement in both overall detection quality and region overlap, indicating that the proposed design is more effective in capturing and segmenting the target regions.

4.3 Qualitative and Visual Analysis

This section complements the quantitative evaluation with a qualitative inspection of the model's outputs. While the previous subsection focused on F1-score and IoU against state-of-the-art methods, here we visually examine representative detection results. For each example, we present the forged (inpainted) FIGURE 2– FIGURE 4, the corresponding ground-truth manipulation mask, and the predicted mask produced by our method. These visualizations

Table 4: Results on Defacto dataset

Method	F1	IoU	Paper / method reference
Dual-Module Architecture	0.75	0.64	Zhang, Lan et al. SAMIF: Adapting Segment Anything Model for Image Inpainting Forensics. Asian Conference on Computer Vision (2024)
Our proposed method	0.79	0.75	Kim D, Kim H. Dual-Module Architecture for Robust Image Forgery Segmentation and Classification Toward Cyber Fraud Investigation, Applied Sciences (2025)

highlight typical success cases, failure modes, and the kinds of artefacts that are most challenging for the detector. In addition, we analyze the impact of image resizing on detection quality. Through an ablation over different resizing strategies and input resolutions, we illustrate how pre-processing choices affect the sharpness and localization of the predicted masks, as well as the preservation of subtle forgery traces. Together, these visual examples provide an intuitive understanding of the strengths and limitations suggested by the numerical results. To further illustrate the behavior of our method under different manipulation conditions, we report qualitative results on three benchmark datasets: IID, IMD2020, and Defacto. For each dataset, we select representative examples spanning a variety of textures, object categories, and forgery sizes. Every example is shown as a tuple consisting of the inpainted image, the corresponding ground-truth manipulation mask, our predicted mask, and an overlay of the prediction on the forged image. These visualizations allow us to assess how well the model generalizes across datasets with different inpainting pipelines and annotation protocols, and how precisely it localizes the manipulated regions. In particular, the overlay views make it easier to judge over- and under-segmentation, boundary sharpness, and typical failure cases (e.g., small or low-contrast edits). Together, the samples from IID (FIGURE 2), IMD2020 (FIGURE 4), and Defacto (FIGURE 3), provide a qualitative complement to the cross-dataset quantitative results discussed above

Figure 2: IID Result - Altered image on the left, ground truth in middle and our obtained mask on the right

In addition to the visual inspection, we study how input resolution affects detection performance - see TABLE
reftab:resize-ablation. Before feeding the images to the model, we uniformly rescale them to fractions of the original size (0.9, 0.7, and 0.5). At the original resolution, our method achieves F1/IoU scores of 0.73/0.68 on IID, 0.87/0.78 on IMD2020, and 0.79/0.75 on De-

Figure 3: DEFACTO Result - Altered image on the left, ground truth in middle and our obtained mask on the right

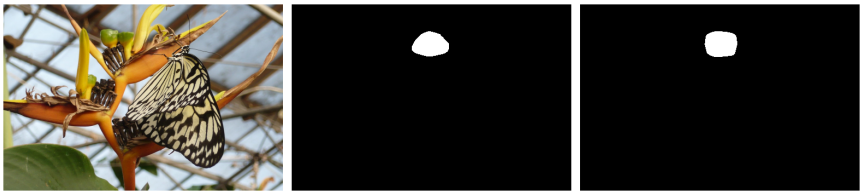

Figure 4: IMD2020 Result - Altered image on the left, ground truth in middle and our obtained mask on the right

facto. When images are downscaled to 0.9×, performance decreases by 0.06–0.09 F1 and 0.08–0.10 IoU across datasets (e.g., IMD2020 drops to 0.78/0.68). Further reductions to 0.7× introduce additional degradation, and strong downscaling to 0.5× leads to a pronounced loss in localization accuracy, with F1 falling to 0.50/0.60 and IoU to 0.45/0.59 on IID and IMD2020, respectively. These results indicate that the model is reasonably robust to moderate resizing, but very low input resolutions remove fine-grained artefacts that are important for precise mask prediction.

Table 5: Impact of input resolution on forgery detection performance. Images are uniformly rescaled by a factor s before being fed to the model.

	Original		**Resized to 0.9**		**Resized to 0.7**		**Resized to 0.5**	
	F1	**IoU**	**F1**	**IoU**	**F1**	**IoU**	**F1**	**IoU**
IID	0.73	0.68	0.64	0.60	0.60	0.55	0.50	0.45
IMD2020	0.87	0.78	0.78	0.68	0.75	0.69	0.60	0.59
Defacto	0.79	0.75	0.73	0.67	0.70	0.61	0.68	0.56

5. CONCLUSION AND FUTURE WORK

This paper introduced ViTARHeat, a novel framework for image inpainting detection, designed to address the “generative arms race” in digital media forensics. A fundamental

flaw in existing forensic detectors was identified—the "Forensic Resize Paradox"—where mandatory pre-processing (downsampling) to fit fixed-resolution CNNs destroys the very high-frequency, microscopic artifacts that serve as evidence of tampering. Contribution: The proposed solution is a dual-stream, resolution-agnostic architecture. Its first novelty is the integration of a ViTAR backbone, which processes images at their native resolution, preserving 100% of the original forensic traces. Its second novelty is a parallel Enhanced Wavelet Scattering Network (EWSN) branch. This non-semantic branch uses the Dual-Tree Complex Wavelet Transform to function as a shift-invariant, direction-selective "anomaly amplifier," isolating the subtle, oriented boundary artifacts and texture inconsistencies that generative inpainting creates. The synergy of ViTAR's global, semantic context with the EWSN's artifact-specific features allows ViTARHeat to set a new state-of-the-art in high-fidelity forgery detection. By solving the "Resize Paradox," this work provides a scalable and robust framework for analyzing high-resolution, "in-the-wild" images, closing a critical gap between inpainting technology and detection capability. This dual-stream framework (Native-Resolution Backbone + Artifact-Amplifier Frontend) is a powerful and generalizable paradigm. This architecture could be applied to other forensic tasks where subtle, high-frequency artifacts are critical, such as splicing detection, video deepfake detection (where temporal wavelets could be used), or audio forgery analysis. Future research could also explore more advanced fusion modules, using, for example, guided cross-attention mechanisms to further refine the interplay between the semantic and artifact-based feature streams.

References

[1] Barglazan AA, Brad R, Constantinescu C. Image Inpainting Forgery Detection: A Review. J Imaging. 2024;10:42.

[2] Deb P, Deb S, Das A, Kar N. Image Forgery Detection Techniques: Latest Trends and Key Challenges. IEEE Access. 2024;12:169452-169466.

[3] Singh S, Kumar R. Image Forgery Detection: Comprehensive Review of Digital Forensics Approaches. J Comput Soc Sci. 2024;7:877-915.

[4] 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. IEEE. 2022:10674-10685.

[5] Mahfoudi G, Tajini B, Retraint F, Morain-Nicolier F, Dugelay JL, et al. DEFACTO: Image and Face Manipulation Dataset. In: Proceedings of the 2019 27th European Signal Processing Conference (EUSIPCO). IEEE. 2019.

[6] Liang Z, Yang G, Ding X, Li L. An Efficient Forgery Detection Algorithm for Object Removal by Exemplar-Based Image Inpainting. J Vis Commun Image Represent. 2015;30:75-85.

[7] Wu H et al. Detection of Object Removal by Image Inpainting. In: Proceedings of the IEEE/CVF international conference on computer vision workshops; 2019.

[8] Wu Y, Liu C, Filaretov V, Yukhimets D. Visual-Neural-Inspired Image Inpainting for Specific Objects-Of-Interest Imaging. 2025. Available at SSRN: `https://papers.ssrn.com/sol3/Delivery.cfm/e9280975-c358-4f70-8cdd-a7fc352e0f35-MECA.pdf?abstractid=5528450&mirid=1&type=2`

[9] Suvorov R, Logacheva E, Mashikhin A, Remizova A, Ashukha A, et al. Resolution-Robust Large Mask Inpainting With Fourier Convolutions. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. IEEE. 2022:3172-3182.

[10] Wu H, Mao J, Zhang Y, Jiang Y, Li L, et al. Unified Visual-Semantic Embeddings: Bridging Vision and Language With Structured Meaning Representations. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:6609-6618.

[11] 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 workshops. IEEE. 2019: 9535-9544.

[12] Wu H, Chen Y, Zhou J. Rethinking Image Forgery Detection via Contrastive Learning and Unsupervised Clustering. 2023. ArXiv preprint:2303.04030

[13] Fan Q, You Q, Han X, Liu Y, Tao Y, et al. ViTAR: Vision Transformer with Any Resolution. 2024. ArXiv preprint: https://arxiv.org/pdf/2403.18361

[14] Dou L, Feng G, Qian Z. Image Inpainting Anti-Forensics Network via Attention-Guided Hierarchical Reconstruction. Symmetry. 2023;15:393.

[15] Barglazan AA, Brad R. Enhanced Wavelet Scattering Network for Image Inpainting Detection. Computation. 2024;12(11):228.

[16] Dixit R, Naskar R. Copy-Rotate-Move Forgery Detection using Complex Wavelet Transform and Local Binary Pattern. 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) 2019 Jul 6 (pp. 1-7). IEEE.

[17] Desolneux A et al. How to Reduce Anomaly Detection in Images to Anomaly Detection in Noise. Image Process On Line. 2019;9:392-414.

[18] Elharrouss O, Almaadeed N, Al-Maadeed S, Akbari Y. Image Inpainting: A Review. Neural Process Lett. 2020;51:2007-2028.

[19] Goodfellow IJ, Pouget-Abadie J, Mirza M et al. Generative Adversarial Nets. Adv Neural Inf Process Syst. 2014.

[20] Zhu X, Lu J, Ren H, Wang H, Sun B. A transformer–CNN for deep image inpainting forensics. Vis Comput. 2022;39:4721-4735.

[21] Giakoumoglou P, Karageorgiou D, Papadopoulos S, Petrantonakis PC. SAGI: Semantically Aligned and Uncertainty Guided AI Image Inpainting. 2025. ArXiv preprint: https://arxiv.org/pdf/2502.06593

[22] Mallat S. Group Invariant Scattering. Commun Pure Appl Math. 2012;65:1331-1398.

[23] Tahir E, Bal M. Deep Image Composition Meets Image Forgery. 2024. ArXiv preprint: https://arxiv.org/pdf/2404.02897

[24] Xu Z, Zhang X, Chen W, Yao M, Liu J, et al. A Review of Image Inpainting Methods Based on Deep Learning. Appl Sci. 2023;13:11189.

[25] Li Y, Hu L, Dong L, Wu H, Tian J, et al. Transformer-Based Image Inpainting Detection via Label Decoupling and Constrained Adversarial Training. IEEE Transactions on Circuits and Systems for Video Technology. 2023;34:1857-1872.

[26] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. 2020. ArXiv preprint: 2010.11929

[27] George A, Marcel S. EdgeDoc: Hybrid CNN-Transformer Model for Accurate Forgery Detection and Localization in ID Documents. 2025 ArXiv preprint: https://arxiv.org/pdf/2508.16284.

[28] Lin TY, Dollár P, Girshick R, He K, Hariharan B. Feature Pyramid Networks for Object Detection. Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE. 2017.

[29] Xiao T, Liu Y, Zhou B, Jiang Y, Sun J. Unified Perceptual Parsing for Scene Understanding. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Proceedings of the European conference on computer vision. Cham: Springer International Publishing. 2018:432-448.

[30] Qiao L, Gan Y, Wang B, Qin J, Xu S, et al. UniViTAR: Unified Vision Transformer with Native Resolution. 2025. ArXiv preprint: https://arxiv.org/pdf/2504.01792

[31] Varma A, Shit S, Prabhakar C, Scholz D, Li HB, et al. VariViT: A Vision Transformer for Variable Image Sizes. Medical Imaging with Deep Learning. PMLR. 2024:1571-1583.

[32] Novozamsky A, Mahdian B, Saic S. IMD2020: A Large-Scale Annotated Dataset Tailored for Detecting Manipulated Images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision workshops. IEEE. 2020:71-80.

[33] Novozámský A, Mahdian B, Saic S. Extended IMD2020: A Large□Scale Annotated Dataset Tailored for Detecting Manipulated Images. IET Biometrics. 2021;10:392-407.

[34] Chen Z, Zhang Y, Wang Y, Tian J, Wu F. Robust Image Inpainting Forensics by Using an Attention-Based Feature Pyramid Network. Appl Sci. 2023;13:9196.