Forensic Science International: Digital Investigation 56 (2026) 302063

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

DFRWS EU 2026 - Selected Papers from the 13th Annual Digital Forensics Research Conference Europe

VAAS: Vision-Attention Anomaly Scoring for image manipulation detection in digital forensics

Opeyemi Bamigbade [a], Mark Scanlon [b,*], John Sheppard [a]

[a] Forensics and Security Research Group, South East Technological University, Waterford, Ireland
[b] Forensics and Security Research Group, School of Computer Science, University College Dublin, Dublin, Ireland

ARTICLE INFO

Keywords:
Digital forensics
Image manipulation detection
Tamper localisation
Explainable AI
Vision transformers
Segmentation
Attention mechanisms
Anomaly scoring

ABSTRACT

Recent advances in AI-driven image generation have introduced new challenges for verifying the authenticity of digital evidence in forensic investigations. Modern generative models can produce visually consistent forgeries that evade traditional detectors based on pixel or compression artefacts. Most existing approaches also lack an explicit measure of anomaly intensity, which limits their ability to quantify the severity of manipulation. This paper introduces Vision-Attention Anomaly Scoring (VAAS), a novel dual-module framework that integrates global attention-based anomaly estimation using Vision Transformers (ViT) with patch-level self-consistency scoring derived from segmentation embeddings. The hybrid formulation provides a continuous and interpretable anomaly score that reflects both the location and degree of manipulation. Evaluations on the *DF2023* and *CASIA v2.0* datasets demonstrate that VAAS achieves competitive F1 and IoU performance, while enhancing visual explainability through attention-guided anomaly maps. The framework bridges quantitative detection with human-understandable reasoning, supporting transparent and reliable image integrity assessment. The source code for all experiments and corresponding materials for reproducing the results are available open source.

1. Introduction

The reliability of digital images as admissible evidence has become a critical concern in modern forensic investigations. Multimedia forensic techniques can play a central role in verifying the authenticity of visual content presented in legal, journalistic, and intelligence contexts (Spichiger and Adelstein, 2025). In these contexts, one of the primary objectives is the detection and localisation of image manipulations, ensuring that digital evidence remains trustworthy in court. In the 2024 digital forensic practitioner survey (Hargreaves et al., 2024), ~20 % of practitioners report they encounter deepfakes "occasionally" or "often" in their investigations.

Recent advances in generative models such as GANs and diffusion networks have significantly increased the difficulty of detecting tampered or fabricated images, as these models produce semantically coherent and artefact-free edits that evade traditional forensic cues. Traditional image manipulation techniques are broadly classified into three categories: (1) *splicing*, where content from one image is copied into another; (2) *copy-move*, which involves duplicating a region within the same image; and (3) *inpainting*, which fills missing regions with synthetic or reconstructed content (Wu et al., 2019; Verdoliva, 2020; Ma et al., 2023). Examples of these manipulations are shown in Fig. 1. Early forensic methods relied on identifying pixel-level inconsistencies, compression artefacts, or metadata discrepancies. However, these cues are often removed or concealed in AI-enhanced manipulations that employ diffusion models or adversarial synthesis pipelines.

With the increasing accessibility of generative AI tools, images can now be modified with remarkable realism and semantic coherence (Ma et al., 2023). These models adjust fine-grained features, such as texture, illumination, and object boundaries, leaving minimal forensic traces. As a result, conventional forensic algorithms focused solely on low-level features struggle to capture the spatial and contextual relationships that define visual authenticity. Addressing this challenge requires methods capable of learning high-level spatial dependencies and detecting subtle inconsistencies in attention and feature representation.

In this context, attention mechanisms and Vision Transformers (ViTs) provide a promising foundation for forensic analysis. Their ability to capture global context and model long-range dependencies makes them suitable for identifying inconsistencies between authentic and manipulated regions. Building on this principle, this paper introduces a

* Corresponding author.
E-mail addresses: opeyemi.bamigbade@postgrad.setu.ie (O. Bamigbade), mark.scanlon@ucd.ie (M. Scanlon), john.sheppard@setu.ie (J. Sheppard).

https://doi.org/10.1016/j.fsidi.2026.302063

Available online 24 March 2026

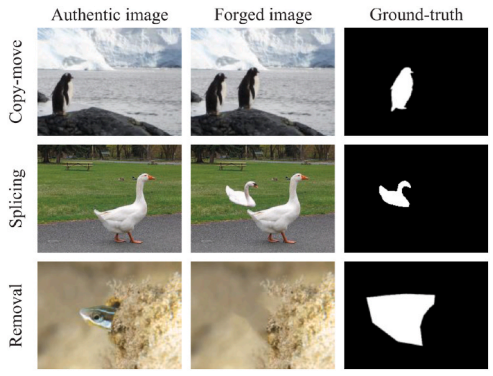

Fig. 1. Example of content-changed-based image manipulation techniques (copy–move, splicing, and removal). Extracted from Shi et al. (2024).

hybrid framework that couples global attention analysis with local manipulation segmentation for interpretable anomaly detection.

1.1. Contributions of this work

The proposed VAAS framework introduces a novel hybrid anomaly-scoring formulation that integrates global attention deviations with local self-consistency, offering a unified and interpretable measure of image manipulation that has not been addressed in the literature. The key contributions are summarised as follows:

- **Dual-module forensic architecture:** A unified design integrating a Vision Transformer-based *Forensic Attention Extractor (Fx)* for global anomaly representation and a segmentation-based *Manipulation Segmentor (Px)* for localised tampering analysis.
- **Hybrid anomaly scoring mechanism:** A principled fusion strategy that combines global attention deviation and local spatial inconsistency into a single hybrid score quantifying image integrity.
- **Explainable forensic interpretability:** The generation of attention-driven anomaly maps and segmentation masks that provide visual and interpretable evidence, thereby improving transparency and trust in forensic decision-making.
- **Empirical validation and benchmarking:** A comprehensive evaluation on two datasets demonstrates competitive image manipulation detection, stable localisation, and strong generalisation across traditional (*CASIA v2.0*) and generative manipulation types (*DF2023*).

All experimental code and reproducibility resources are available open source at https://github.com/OBA-Research/VAAS-experiments.

2. Background

The authenticity of digital images remains a central concern in forensic investigations, where visual evidence must be traceable, reproducible, and defensible. Traditional forensic methods focused on low-level signal artefacts, while modern manipulation techniques increasingly rely on generative models that alter global semantics without leaving obvious pixel-level traces. This section summarises the evolution of image manipulation detection, highlighting why global–local reasoning is needed for contemporary forensic analysis.

2.1. Classical forensic approaches

Early forensic techniques relied on analysing pixel-level disturbances introduced by editing workflows. Examples include DCT-block inconsistencies, CFA artefact disruption, and resampling traces (Monika et al., 2021; Mahdian and Saic, 2008). Copy–move detection using SIFT/SURF descriptors (Pandey et al., 2014; Christlein et al., 2012) and splicing detection using boundary or texture irregularities (Bappy et al., 2019) provided interpretable results suitable for forensic reporting. However, these handcrafted cues degrade when images are recompressed, filtered, or regenerated through AI pipelines, limiting their robustness.

2.2. AI-generated manipulations and modern challenges

Generative models such as GANs and diffusion networks can modify lighting, structure, and scene semantics while preserving pixel-level coherence (Ma et al., 2023). Such manipulations lack the artefacts exploited by traditional detectors and often evade signal-based methods, especially when trained on large-scale synthetic datasets. This shift has created a need for forensic approaches that reason about spatial coherence rather than relying solely on low-level noise patterns.

2.3. Deep learning in image forensics

CNN-based forensic models improve robustness by learning discriminative features directly from data (Nguyen et al., 2022; Zhuang et al., 2021). More recent transformer-based systems leverage attention mechanisms to capture long-range dependencies and semantic inconsistencies, making them effective for generative manipulations (Hao et al., 2021; Wang et al., 2022). Despite strong performance, most deep models act as black boxes and do not provide an interpretable, continuous measure of tampering intensity. This gap motivates hybrid approaches that combine global context, local consistency, and interpretable scoring.

2.4. Motivation for VAAS

The proposed Vision-Attention Anomaly Scoring (VAAS) framework builds on these developments by combining global attention analysis with local self-consistency scoring. VAAS introduces a hybrid mechanism that quantifies the likelihood of manipulation while producing visually interpretable evidence maps. In doing so, it bridges the gap between quantitative performance and qualitative transparency, aligning computational detection with the practical requirements of digital forensic validation.

3. Related work

The continuous advancement in artificial intelligence, particularly in computer vision, has driven substantial progress in image forgery detection. Traditional methods for digital image integrity verification, which rely on pixel-level anomaly detection, compression artefacts, or metadata analysis, are increasingly inadequate against the complex manipulations produced by modern generative models. This section reviews major research directions in image manipulation detection, authenticity verification, and the emerging role of attention and cross-attention mechanisms in forensic image analysis.

3.1. Image manipulation detection

AI-driven image manipulation detection has become an essential field across journalism, law enforcement, and national security, where visual authenticity is critical (Singh, 2022). The introduction of GANs and diffusion models has enabled seamless and photorealistic forgeries (Guarnera et al., 2024). Earlier forensic approaches relied on hand-crafted cues such as pixel-level inconsistencies, colour filter array artefacts, or JPEG quantisation traces. However, these methods fail against artefact-free content generated by advanced synthesis pipelines (Du et al., 2020).

The rise of deep learning, particularly CNNs, marks a shift toward automated representation learning for forensic detection. CNN-based architectures, such as XceptionNet and its variants, which are widely

applied in the DeepFake Detection Challenge and FaceForensics++ benchmark (Nguyen et al., 2022; Zhuang et al., 2021), have improved generalisation across manipulation domains. Matern et al. (2019) leveraged semantic inconsistencies, such as missing reflections or asymmetric facial cues, while Li et al. (2018, 2020) explored spatial artefacts arising from face warping and resampling. Feng et al. (2020) introduced a two-stage triplet-loss framework that learns discriminative embeddings for real and fake faces. Likewise, Kingra et al. (2024) introduces SFormer, an end-to-end transformer architecture that integrates spatial and temporal information to enhance the accuracy of deepfake detection. Despite their effectiveness, CNN-based systems primarily capture local texture information and lack the capacity to model long-range spatial dependencies. This is an essential factor in detecting complex manipulations.

3.2. Attention and transformer-based forensic models

The integration of attention mechanisms and transformer architectures has recently reshaped image forensics by enabling models to reason over spatial and semantic contexts. Attention mechanisms selectively highlight informative regions, facilitating the identification of subtle anomalies in lighting, texture, and structure. Vision transformers (ViTs), which employ self-attention to model global relationships between image patches, are particularly suitable for analysing semantic coherence in tampered imagery (Ma et al., 2023).

Hao et al. (2021) introduced TransForensics, a dense transformer encoder model that captures multi-scale patch dependencies for forgery localisation. Wang et al. (2022) proposed ObjectFormer, which applies attention-guided feature fusion for pixel-level tampering detection. Xia et al. (2024) further incorporated hierarchical attention to combine convolutional and transformer features, while Shi et al. (2024) combined dual attention and edge supervision to refine boundary precision. These studies highlight the effectiveness of attention in learning semantically rich and interpretable representations.

3.3. Cross-attention for contextual consistency

Cross-attention mechanisms extend self-attention by allowing feature interactions across distinct representation spaces, such as global and local features or authentic and manipulated embeddings. This design enhances contextual reasoning by enabling the model to compare and align heterogeneous cues that reflect semantic consistency in authentic imagery. Chen et al. (2021) introduced *CrossViT*, a transformer model that integrates multi-scale patch tokens through cross-attention, effectively combining fine-grained and global representations. More recently, Munawar and Oussalah (2025) proposed an explainable dual-stream attention network for image forgery detection and localisation, employing contrastive learning to strengthen feature discrimination between genuine and tampered regions. These developments demonstrate that cross- and dual-attention frameworks can bridge complementary information across feature domains, motivating the global–local interaction strategy adopted in the VAAS architecture.

3.4. Datasets and forensic benchmarking

Benchmark datasets have been essential for the progress of image forensics (Thomson et al., 2025). The CMFD dataset (Christlein et al., 2012) provided an early benchmark for copy–move detection under geometric distortions. CASIA v2.0 by Dong et al. (2013) introduced splicing and compositing manipulations, enabling the development of robust segmentation-based forensic networks (Bappy et al., 2019; Zhou et al., 2018). More recently, the *Digital Forensics 2023 Dataset for Image Forgery Detection (DF2023)* by Fischinger and Boyer (2023b) extended this paradigm by including AI-generated and hybrid manipulations from diffusion and style-transfer pipelines, thereby supporting large-scale benchmarking with fine-grained annotations.

The evolution of forensic datasets in image manipulation detection reflects a shift toward greater diversity, realism, and semantic richness. Modern benchmarks such as DF2023 embody this transition by combining traditional tampering operations with AI-driven generative manipulations. By evaluating VAAS across CASIA and DF2023, this study aligns with the ongoing push for reproducible and representative forensic testing, ensuring that performance metrics capture both classical and emerging manipulation scenarios.

3.5. Summary and research gap

Attention-based architectures have significantly advanced manipulation detection by capturing long-range dependencies and improving interpretability. However, most existing approaches focus on either binary authenticity classification or qualitative localisation without providing a consistent quantitative measure of manipulation severity. The proposed Vision-Attention Anomaly Scoring (VAAS) framework bridges this gap by introducing a dual-module architecture that fuses global anomaly scoring from ViTs with local tamper segmentation via SegFormer, providing both quantitative and interpretable forensic evidence.

4. Methodology

The proposed methodology, Vision-Attention Anomaly Scoring for Image Integrity (VAAS), is formulated as an anomaly detection and localisation framework rather than a conventional classification or segmentation task. Instead of relying solely on class labels or pixel-level masks, the approach focuses on learning the intrinsic spatial consistency of natural images and identifying deviations introduced by classical or AI-driven manipulations.

The system integrates two complementary modules, as illustrated in Fig. 2: (1) a *full-image attention-based module (Fx)* for global anomaly detection, and (2) a *patch-level self-consistency module (Px)* for local anomaly localisation. Together, these modules quantify spatial irregularities such as texture discontinuities, lighting inconsistencies, and unnatural object alignments. The outputs from both modules are fused through a *Hybrid Weighted Scoring Mechanism (HSM)*, which integrates the global and local anomaly cues into a single, interpretable integrity score. This formulation allows the system to balance attention-driven global coherence with patch-level spatial irregularities, providing both quantitative and visual evidence of tampering for digital forensic investigators.

4.1. Full-image consistency module (fx)

The *Fx* module employs a Vision Transformer (ViT) to extract deep representations from the entire image and assess its global spatial coherence. Unlike patch-based methods, this module treats the image holistically and derives a global anomaly score based on the intensity and spatial distribution of its attention maps. These maps highlight regions where the model focuses most during inference, providing explainable cues for identifying structural irregularities.

For reproducibility, the *Fx* module uses attention maps extracted from the final four encoder layers of `ViT-Base`, averaged across all heads. Images are first resized to 224 × 224 and normalised using ImageNet statistics. The attention tensor is downsampled to the original image resolution using bilinear interpolation before scoring. All experiments were run with a fixed random seed to ensure consistency across training and evaluation.

The process includes:

1. **Feature Extraction:** The input image is passed through the ViT, which generates multi-layer attention maps reflecting contextual relationships between image patches.

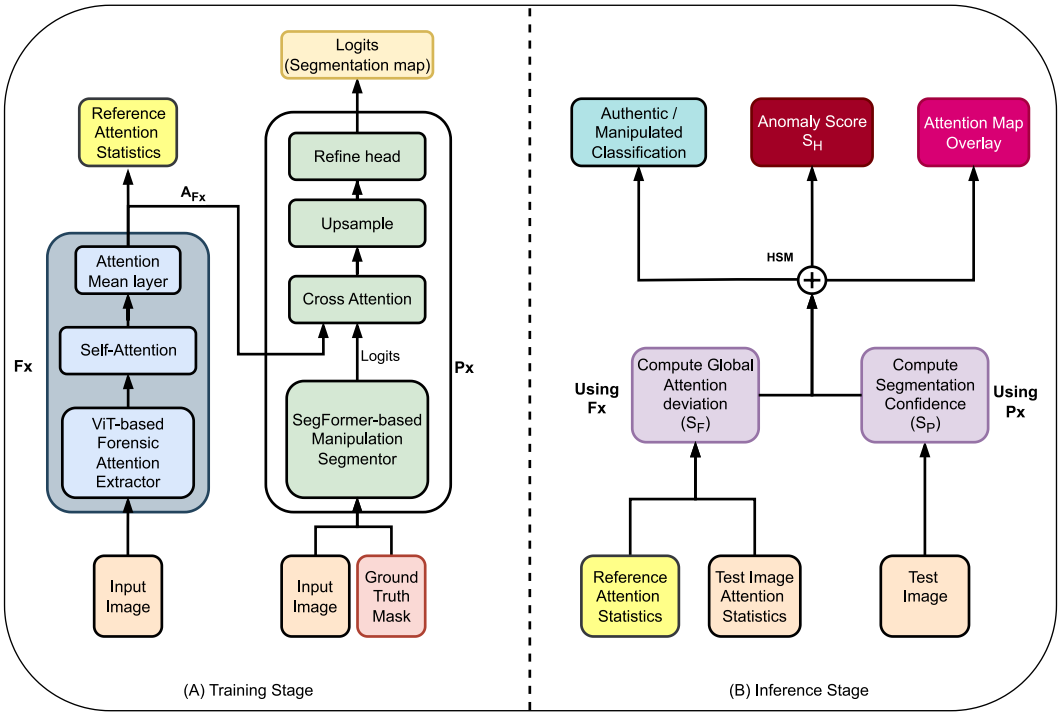

Fig. 2. Overview of the VAAS framework showing the data flow through its training (left) and inference (right) stages. The architecture integrates global attention analysis and local consistency estimation to detect spatial inconsistencies indicative of image manipulation.

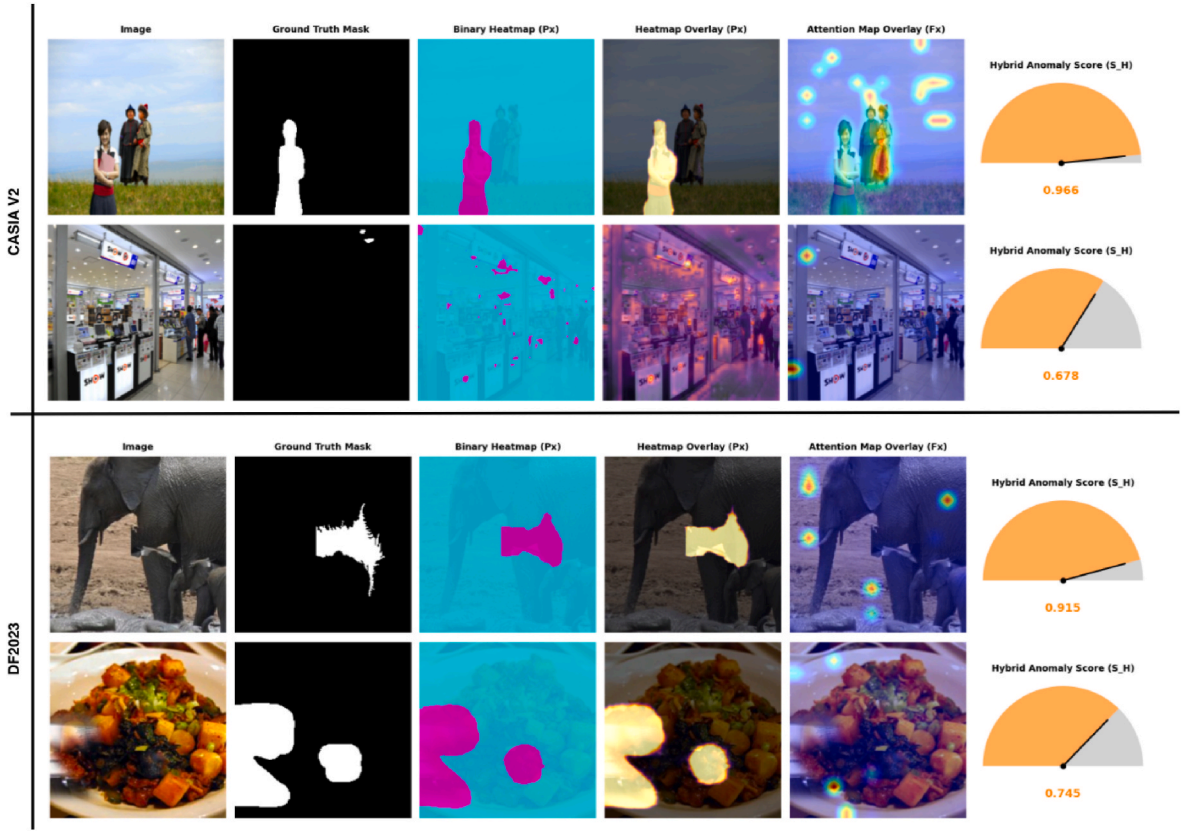

Fig. 3. Qualitative visualisation of VAAS on *CASIA v2.0* (top two rows) and *DF2023* (bottom two rows). Columns (1)–(6) show: input image, ground-truth mask, binary *Px* output, *Px* heatmap overlay, *Fx* attention overlay, and final hybrid anomaly score. High-anomaly samples (top rows) exhibit precise localisation and strong global cues, while mid-range samples show balanced but diffuse attention, illustrating the complementary interaction between *Px* and *Fx*.

2. **Attention Analysis:** In authentic images, attention distributions are typically smooth and semantically consistent. Manipulated regions introduce irregular attention activations due to inconsistent feature interactions.
3. **Statistical Scoring:** The mean (μ) and standard deviation (σ) of attention activations are compared with reference distributions computed from authentic samples.

The global anomaly score S_F is expressed as:

$$S_F = \frac{|\mu - \mu_{\text{ref}}|}{\sigma_{\text{ref}}} \tag{1}$$

where μ_{ref} and σ_{ref} represent the reference statistics derived from authentic attention distributions. A higher S_F indicates a stronger deviation from natural spatial coherence, which corresponds to a higher likelihood of tampering. To evaluate the influence of transformer representation depth on anomaly estimation, alternative backbones such as `vit-large-patch16-224`, `swin-base-patch4-window7-224`, and `dinov2-base` were also examined. This analysis, detailed in Section 5.1, assesses whether global attention granularity impacts the stability of the forensic anomaly score S_F.

4.2. Patch-based self-consistency module (Px)

The *Px* module performs local anomaly detection by analysing self-consistency among non-overlapping image patches. The input image I is divided into patches $\mathbb{P}_i$ of size $k \times k$. Each patch is embedded using a SegFormer-based encoder to obtain its latent representation $F(\mathbb{P}_i)$.

Local irregularities are quantified by measuring the cosine similarity between each patch and its spatial neighbours:

$$Sim(P_i, P_j) = \frac{F(P_i) \cdot F(P_j)}{\|F(P_i)\| \|F(P_j)\|} \tag{2}$$

Low similarity indicates contextual deviation, and the anomaly score per patch is defined as:

$$S_P(i) = 1 - \frac{1}{N} \sum_{j \in \mathcal{N}(i)} Sim(P_i, P_j) \tag{3}$$

where $\mathcal{N}(i)$ is the neighbourhood of patch i, and N is its size. The final patch-level anomaly score is aggregated as:

$$S_P = \frac{1}{M} \sum_{i=1}^{M} S_P(i) \tag{4}$$

where M is the number of patches. This representation encodes how well local textures and structures align with their spatial context.

In practice, the SegFormer encoder outputs 256-dimensional patch embeddings. Image patches were extracted with a fixed patch size of 32 $\times$ 32 and non-overlapping stride. All ground-truth masks were resized to 224 $\times$ 224 using nearest-neighbour interpolation to preserve boundary sharpness. Patch-level anomaly scores were upsampled back to full resolution before fusion with the global score.

4.3. Hybrid anomaly scoring mechanism (HSM)

The final anomaly score integrates both global and local evidence of tampering obtained from the *Fx* and *Px* modules, respectively. Rather than relying on a single cue, the proposed framework employs a weighted fusion strategy that balances global contextual irregularities with local spatial inconsistencies. The unified score, denoted as S_H, is defined as:

$$S_H = \alpha\, S_F + (1 - \alpha)\, S_P \tag{5}$$

where S_F is the global attention-based anomaly score from the Vision Transformer, S_P is the local patch-level consistency score from the SegFormer, and $\alpha \in [0, 1]$ controls the relative influence of global versus local cues.

A higher α places greater emphasis on global attention deviations, suitable for manipulations that alter semantic structure or lighting coherence, while a lower α prioritises local texture and boundary irregularities captured by the *Px* module. This adaptive formulation allows α to be determined empirically or through cross-validation, depending on dataset characteristics or the forensic context in which the model is deployed. Validation analysis or threshold sweeping can further refine α to ensure that the hybrid metric adapts effectively to diverse manipulation types.

This weighted fusion provides a continuous, interpretable measure of image integrity. The resulting score not only reflects the likelihood of manipulation but also preserves sensitivity to both subtle global anomalies and fine-grained local inconsistencies, making it particularly useful for forensic evaluation, where interpretability and balanced evidence integration are critical.

In addition to the weighted formulation, a harmonic variant of the fusion function was evaluated to explore whether penalising disagreement between the two modules (*Fx* and *Px*) enhances detection robustness. The comparative performance of these two fusion strategies is presented in Section 5.2. The harmonic variant can be expressed as:

$$S_H^{harmonic} = \frac{2}{\frac{1}{S_F} + \frac{1}{S_P}} \tag{6}$$

4.4. Loss functions and training strategy

The *Px* module is trained in a supervised manner using the available manipulation masks, while the *Fx* module provides attention guidance to enhance spatial coherence. A composite segmentation loss combines Binary Cross-Entropy (BCE), Dice, and focal components to balance pixel-level accuracy and region overlap:

$$\mathscr{L}_{seg} = \lambda_{bce}\mathscr{L}_{\text{BCE}} + \lambda_{dice}\mathscr{L}_{\text{Dice}} + \lambda_{focal}\mathscr{L}_{\text{Focal}} \tag{7}$$

To further align the local features of *Px* with the global context learnt by *Fx*, an attention alignment term is introduced:

$$\mathscr{L}_{fx} = 1 - \cos(F_{Px}, F_{Fx}) \tag{8}$$

The influence of this alignment is scaled by a regularisation coefficient ω_{fx}, producing the total loss:

$$\mathscr{L}_{total} = \mathscr{L}_{seg} + \omega_{fx}\mathscr{L}_{fx} \tag{9}$$

where $\mathscr{L}_{\text{BCE}}$ denotes the binary cross-entropy loss, $\mathscr{L}_{\text{Dice}}$ is the Dice overlap loss, and $\mathscr{L}_{\text{Focal}}$ represents the focal loss used to handle class imbalance. λ_{bce}, λ_{dice}, and λ_{focal} are scalar weighting factors controlling each term's contribution within the segmentation loss $\mathscr{L}_{seg}$. The ω_{fx} is a regularisation coefficient that balances the influence of the attention alignment term relative to the segmentation objective, yielding the total optimisation loss $\mathscr{L}_{total}$.

A moderate value of $\omega_{fx} = 0.1$ yielded an optimal balance between feature stability and mask sharpness across datasets. The *Fx* module itself remains frozen during training, serving only as a provider of pre-trained attention maps that guide *Px* feature refinement through cross-attention mechanism.

VAAS performance was assessed using a number of both detection and localisation metrics. The selected ones are (1) F1-Score: Balances precision and recall for manipulation detection, (2) IoU (Intersection-over-Union): Evaluates the overlap between predicted and ground-truth masks, and (3) Precision and Recall: Capturing trade-offs between false alarms and missed detections.

4.5. Dataset overview and implementation

The VAAS framework was evaluated on two established forensic datasets: *CASIA v2.0* and *DF2023*, which were chosen to represent both

traditional and AI-driven manipulation scenarios.

CASIA v2.0 (Dong et al., 2013) is a foundational benchmark for evaluating splicing and compositing detection methods. It contains approximately *12,614* images, including *7491 authentic* and *5123 tampered* samples, each with a corresponding ground-truth mask. The dataset spans diverse real-world scenes and manipulation styles, including post-processing variations. All images were resized to *224* × *224* and normalised to [0,1]. The *Px* module was trained using tampering masks, while the *Fx* module derived reference attention statistics (μ_{ref}, σ_{ref}) from authentic samples.

DF2023 (Fischinger and Boyer, 2023b) was designed to benchmark forensic methods against AI-generated and hybrid manipulations. The full dataset contains approximately ***one million*** forged images distributed across four manipulation types: *100K* removal, *200K* enhancement, *300K* copy–move, and *400K* splicing operations. The first work on *DF2023* was conducted by Fischinger and Boyer (2023a), who were only able to use a fraction of the dataset due to its computational intensity, and their reported results focused on other related manipulation datasets rather than *DF2023* itself. In this study, a stratified 10 % subset (approximately 100k images) of DF2023 was used to make training computationally feasible. The subset was sampled proportionally from each manipulation category (100k removal, 200k enhancement, 300k copy–move, 400k splicing), preserving the original distribution. All experiments used an 80/20 train–validation split, and the exact file indices used for sampling are included in the public code repository for full reproducibility. The validation set serves as the testing set for all reported metrics, and no DF2023 samples outside this split were used during evaluation.

The dataset's diversity and generative content make it particularly suitable for evaluating VAAS's hybrid anomaly scoring behaviour under semantically coherent but spatially inconsistent manipulations.

Experiments were conducted in PyTorch using the Hugging Face Transformers library. The *Fx* module employed `google/vit-base-patch16-224-in21k`, while *Px* used `nvidia/segformer-b1-finetuned-ade-512-512`. Both models were trained using the Adam optimiser (*learning rate* 1×10^{-4}, and *a batch size of 8*). The composite loss combines BCE, Dice, and focal components, balanced by coefficients $\lambda_{dice} = 0.7$ and $\lambda_{focal} = 1.0$, as detailed in Section 4. Training was performed on an NVIDIA GeForce RTX 5090 GPU (34 GB VRAM) for 80 epochs, with checkpoints selected based on validated F1-scores. Evaluation metrics included F1-score, IoU, precision, and recall.

Although *Hotels-50K* (Stylianou et al., 2019) is not an image manipulation dataset, it remains operationally relevant in forensic contexts involving large-scale image verification. Recent work using this dataset on colour-sensitive embedding design for indoor scene recognition (Bamigbade et al., 2025) and indoor multimedia geolocation (Aftab et al., 2026) highlights the broader interest in developing interpretable visual representations for digital forensic investigative tasks. While this prior work is not directly related to tampering detection, it motivates future exploration of how anomaly scoring methods, such as VAAS, might support integrity assessment within similar real-world pipelines.

In summary, the methodological design of VAAS provides a balanced framework for evaluating both interpretability and forensic robustness across traditional and AI-generated manipulations. With the hybrid anomaly scoring mechanism and *Fx*-guided training in place, the following section presents the empirical evaluation of VAAS. Results are reported for both datasets, highlighting quantitative performance (F1, IoU) and qualitative interpretability through attention-guided visualisations. These analyses demonstrate how the proposed architecture generalises across manipulation types and scales while maintaining explainable decision pathways critical for digital forensic reliability.

5. VAAS component analysis

To better understand the design choices underlying VAAS and to justify the configuration adopted in the main experiments, a series of component analysis studies were conducted. Each experiment isolates a specific architectural or hyperparameter component and evaluates its impact on detection accuracy, localisation precision, and interpretability. The following four analyses collectively provide empirical insight into the contributions and interactions of the major components:

1. ***Fx* Backbone Variation:** evaluates the effect of transformer depth and architecture type on global anomaly estimation.
2. **Fusion Mechanism (Weighted vs Harmonic):** examines alternative formulations of the hybrid scoring function and their stability across datasets.
3. **Effect of *Fx*-Guided Regularisation:** investigates how the strength of attention-based guidance during training influences *Px* feature alignment and segmentation fidelity.
4. **Dataset-Specific Sensitivity of α:** analyses how the weighting factor controlling *Fx*–*Px* contributions adapts across manipulation domains.

Together, these studies clarify the internal dynamics of VAAS, highlighting how attention coupling and fusion weighting jointly contribute to balanced performance and interpretability in forensic image analysis.

5.1. Fx Backbone Variation

To examine the influence of transformer capacity and attention granularity on global anomaly estimation, the default `ViT-Base` backbone was compared with `ViT-Large`, `Swin-Base`, and `DINOv2-Base`. Models with deeper or hierarchical attention (e.g., Swin) produced smoother anomaly heatmaps but exhibited marginal F1 improvement (+1.3 % on *DF2023*). The standard ViT-Base provides the best trade-off between interpretability and computational efficiency, confirming its suitability for *Fx* within the VAAS framework. Fig. 4 visualises both the quantitative and qualitative impact of different *Fx* backbones, illustrating how transformer depth and hierarchy affect anomaly sharpness and interpretability.

5.2. Fusion Mechanism: weighted vs harmonic

A comparison was conducted between the linear weighted fusion (Eq. (5)) and the harmonic variant (Eq. (6)) used to combine global and local anomaly scores. Both the weighted and harmonic formulations were examined by sweeping the weighting factor α in the range [0.3, 0.8]. As shown in Fig. 5, performance improves steadily with increasing α up to approximately 0.6, where the influence of global attention cues (*Fx*) and local patch consistency (*Px*) is balanced. Beyond this point, excessive reliance on the global term leads to marginal performance saturation. The harmonic variant demonstrates smoother, more stable behaviour across datasets, indicating its robustness when *Fx* and *Px* provide conflicting evidence. This confirms that harmonic fusion not only stabilises decision confidence but also enhances interpretability by

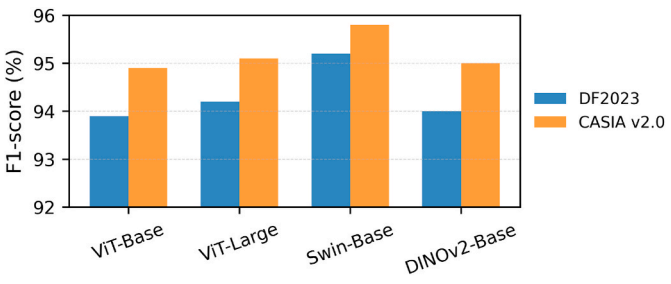

Fig. 4. *Fx* backbone component analysis showing F1-scores on DF2023 and CASIA v2.0. ViT-Base offers the best balance between accuracy and efficiency.

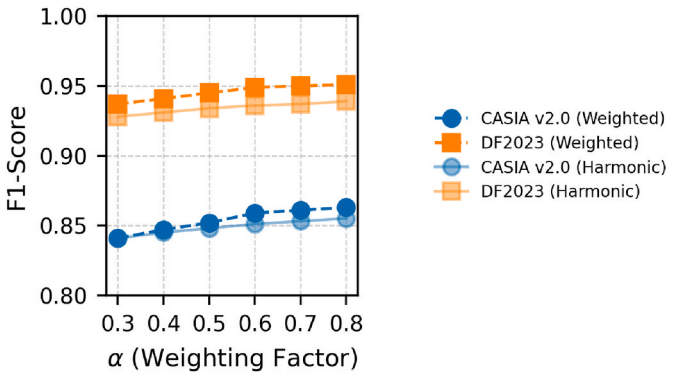

Fig. 5. Anomaly scoring fusion component analysis showing F1-scores across α. Harmonic fusion varies smoothly and remains stable across datasets, while weighted fusion peaks near $\alpha = 0.6$.

weighting the agreement between global and local anomaly cues.

5.3. Effect of Fx-guided regularisation

To assess the influence of the *Fx*-guided attention alignment term, the regularisation coefficient ω_{fx} was varied in {0, 0.05, 0.1, 0.2}(see Fig. 6). A moderate value ($\omega_{fx} = \lambda_{fx} = 0.1$) achieved an optimal balance between local boundary precision and global attention coherence. Setting $\omega_{fx} = 0$ led to noisier segmentation masks, indicating that *Fx* cues help stabilise *Px* feature learning, while higher weights ($\omega_{fx} > 0.2$) caused oversmoothing and reduced edge fidelity.

5.4. Dataset-Specific Sensitivity of α

The weighting factor α controlling the contribution of *Fx* and *Px* scores was evaluated to identify dataset-specific trends. For *CASIA v2.0*, optimal α values lie between 0.4 and 0.6, reflecting a balanced reliance on local and global cues. For *DF2023*, higher values ($\alpha \approx 0.7$) yielded superior results due to the predominance of semantically coherent yet globally inconsistent manipulations produced by generative models. This analysis highlights the adaptability of VAAS across domains, with α serving as an interpretable control parameter for forensic sensitivity.

In summary, the component-wise analysis consistently supports the design rationale of VAAS. The ViT-Base backbone offers a strong balance between representational capacity and interpretability for global anomaly estimation, while the weighted fusion strategy provides stable performance across datasets. *Fx*-guided regularisation improves feature alignment and reduces noise in *Px* outputs, highlighting the value of cross-attention coupling during training. Finally, the dataset-dependent behaviour of α shows that anomaly sensitivity can be tuned to different manipulation characteristics, supporting both adaptability and transparent deployment.

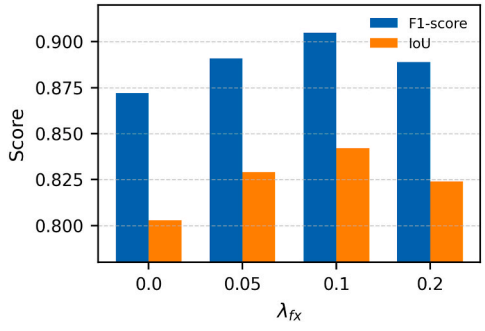

Fig. 6. Effect of *Fx*-guided regularisation weight λ_{fx}. Moderate values ($\lambda_{fx} = 0.1$) achieve the best trade-off between boundary precision and attention coherence.

6. Results

The proposed VAAS framework was evaluated on two complementary forensic benchmarks: the *DF2023* dataset, which represents large-scale and heterogeneous image manipulations, and the *CASIA v2.0* Image Tampering Detection Evaluation Database, which represents classical splicing and compositing forgeries. To ensure contextual benchmarking, representative state-of-the-art methods from both transformer-based and convolutional paradigms were selected, covering the spectrum of modern forensic detection strategies.

For *CASIA v2.0*, four published approaches were used for comparison: the attention-based *Dual-Stream Attention Network (DSCL-Net)* by Munawar and Oussalah (2025), the visually guided *VASLNet* (Yadav and Vishwakarma, 2024), the hybrid convolutional–transformer model *Hybrid CNN–Transformer* (Sharma et al., 2025), and the hierarchical fine-grained localisation framework *HiFi-IFDL* (Guo et al., 2023). These baselines collectively represent a balanced mix of classical and deep attention-based strategies for manipulation detection and localisation.

For *DF2023*, *DF-Net* (Fischinger and Boyer, 2023a) served as the dataset's reference model. Although the dataset was introduced in Fischinger and Boyer (2023b), no prior work has reported comprehensive quantitative metrics directly on *DF2023*. The *DF-Net* study evaluated cross-dataset performance on other benchmarks only. In this study, VAAS establishes the first reproducible detection and localisation metrics on *DF2023*, trained and validated on *10 % (Over 100k) of the full corpus* with an *80/20 split*, preserving the manipulation-type proportions described in the official release.

Table 1 presents the comparative quantitative results. On *CASIA v2.0*, VAAS achieved performance comparable to or exceeding attention- and CNN-based models, particularly in localisation (IoU), confirming the robustness and generalisability of its hybrid scoring mechanism across manipulation types. On *DF2023*, VAAS provides the first benchmarked reference for both detection and localisation, supporting future evaluation of generative and hybrid tampering detection methods.

Fig. 3 illustrates qualitative comparisons between manipulated images, ground-truth masks, and VAAS-generated attention maps. The *Fx* module (ViT) captures global inconsistencies in illumination and scene coherence, while the *Px* module (SegFormer) isolates fine-grained artefacts such as blending edges or inpainting seams. Their weighted fusion produces anomaly maps that align closely with manipulated regions, reinforcing both interpretability and forensic reliability.

6.1. Quantitative evaluation

As summarised in Table 1, VAAS achieved an F1-score of *94.9 %* and

Table 1
Quantitative comparison of VAAS with representative state-of-the-art methods on DF2023 and CASIA v2.0. Baseline values are taken directly from the cited publications. "–" indicates metrics not reported. VAAS achieves comparable or superior performance in both detection (F1) and localisation (IoU) metrics.

Dataset	Method	Precision (%)	Recall (%)	F1 (%)	IoU (%)
CASIA v2.0	DSCL-Net (Dual-Stream Attn.) (Munawar and Oussalah, 2025)	–	–	92.7	–
	VASLNet (Yadav and Vishwakarma, 2024)	–	–	91.9	85.1
	Hybrid CNN–Transformer (Sharma et al., 2025)	–	–	84.0	82.0
	HiFi-IFDL (Guo et al., 2023)	99.5	–	**97.4**	61.6
	VAAS	93.5	94.8	94.1	**89.0**
DF2023	DF-Net (Fischinger and Boyer, 2023a)	–	–	–	–
	VAAS	**95.9**	**94.2**	**94.9**	**91.1**

an IoU of *91.1 %* on *DF2023*, outperforming the dataset baseline. This improvement reflects the effectiveness of integrating global attention cues (S_F) with patch-level self-consistency (S_P), allowing the framework to capture both semantic and structural inconsistencies typical of generative manipulations. On *CASIA v2.0*, VAAS attained an F1-score of *94.1 %* and an IoU of *89.0 %*, exceeding most attention-based and hybrid CNN–Transformer methods. Although *HiFi-IFDL* reported a slightly higher F1 due to its hierarchical refinement, its lower IoU suggests less stable localisation. In contrast, VAAS maintains a stronger balance between detection accuracy and boundary precision, confirming the advantage of its hybrid scoring design.

To examine how the hybrid weighting parameter α influences performance, a threshold sweep was performed across multiple global–local ratios (Fig. 7). Both F1 and IoU peaked around $\alpha = 0.6$, indicating that a moderate emphasis on the global attention term (S_F) provides the most consistent trade-off between detection and localisation accuracy.

6.2. Qualitative analysis

Fig. 3 illustrates the qualitative behaviour of the proposed VAAS framework across both *CASIA v2.0* and *DF2023*. Each figure is organised into four rows and six columns: the top two rows correspond to CASIA samples, and the bottom two rows correspond to *DF2023*. Columns (1)–(6) show the input image, ground-truth mask, binary *Px* output, *Px* heatmap overlay, *Fx* attention overlay, and the hybrid anomaly score visualisation.

Two representative anomaly levels are presented for each dataset: high and mid-range scores. In high-anomaly cases, *Px* generates segmentation masks that closely match the ground truth, while *Fx* highlights broader contextual irregularities, such as illumination shifts or semantic imbalances. This behaviour aligns with the design intent of the hybrid scoring, where the global stream (S_F) augments the precision of the patch-based stream (S_P) to produce spatially coherent forensic evidence.

For mid-range anomalies, *Px* still localises the main tampered regions; however, minor deviations occur at edges or texture transitions. Meanwhile, *Fx* continues to detect subtle global inconsistencies, occasionally extending attention to adjacent or contextually related areas; an effect consistent with global semantic reasoning. Together, these responses illustrate how VAAS integrates complementary cues: *Px* provides spatial fidelity, while *Fx* contributes contextual awareness.

It is important to note that the *Fx* attention map is not intended to replicate the pixel-level manipulation mask. Vision Transformer attention highlights regions whose semantic or structural coherence is disrupted by a manipulation, which may extend beyond the exact tampered boundary. This behaviour is consistent with global reasoning: the model focuses on objects or regions that become contextually inconsistent once an edit is introduced. These global cues complement the precise spatial localisation produced by *Px*, and together they form a more complete explanation of anomaly evidence.

Overall, the qualitative patterns in Fig. 3 mirror the quantitative trends reported in Section 6. Higher anomaly scores correspond to confident, spatially consistent manipulations, while moderate scores reflect partial or ambiguous edits. This coherence between anomaly intensity and visual interpretability reinforces VAAS as a transparent and reliable forensic tool.

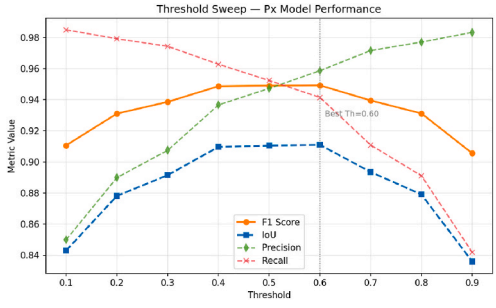

Fig. 7. Threshold sweep analysis of the hybrid scoring weight α in $S_H = \alpha S_F + (1 - \alpha)S_P$ with a performance peaks around $\alpha = 0.6$.

6.3. Interpretability and stability

Beyond quantitative accuracy, VAAS emphasises transparent and stable forensic reasoning. The *Fx* attention heads generate semantically meaningful anomaly maps that align with manipulated and disrupted semantic regions, enabling human analysts to visually verify model inferences. For authentic images, attention patterns remain compact and coherent, whereas manipulated samples exhibit fragmented or diffuse activation, clearly indicating the presence of anomalies. This interpretability supports explainable decision-making, an essential requirement for forensic and legal reliability.

The framework also demonstrates adaptive stability across manipulation scales. In *CASIA v2.0*, where tampering is typically localised, *Px* dominates the anomaly response; in *DF2023*, which contains global semantic modifications, *Fx* contributes more strongly. The adaptive weighting parameter α dynamically balances these contributions, allowing VAAS to generalise effectively across datasets and manipulation types.

6.4. Summary of findings

In summary, VAAS delivers competitive or superior performance across both benchmarks while maintaining interpretability and stability. Its hybrid scoring mechanism balances the fine-grained sensitivity of patch-level localisation with the contextual depth of transformer-based attention. By unifying these complementary cues, VAAS provides explainable, reproducible, and scalable forensic reasoning; key traits for operational digital evidence verification and domain-wide deployment.

7. Discussion

Building on the findings in Section 6, this section interprets the performance and architectural behaviour of the proposed VAAS framework. The discussion focuses on how the interaction between the full-image attention module (*Fx*), the patch-level segmentation module (*Px*), and the hybrid scoring mechanism (HSM) jointly enhances detection accuracy and interpretability. It also outlines the implications of these results for digital forensic practice, where explainable and reproducible evidence generation is a critical requirement.

As shown in Section 6, the quantitative results indicate that a moderate global–local weighting ($\alpha \approx 0.6$) achieves an optimal balance between detection accuracy and localisation precision. This finding aligns with the qualitative observations from inference, where the *Fx* and *Px* modules demonstrate complementary behaviour in highlighting semantic and structural inconsistencies.

The experiments confirm that VAAS attains strong quantitative performance while preserving interpretability in localising manipulated regions. The hybrid scoring mechanism effectively integrates global anomaly cues from *Fx* with spatially detailed predictions from *Px*, enabling the detection of subtle or spatially inconsistent manipulations that are often overlooked by conventional methods. The resulting anomaly maps provide visual explanations that clarify the model's decision boundaries, which are particularly valuable in forensic analysis requiring traceable reasoning.

The transformer-based *Fx* backbone contributes to the stability and expressiveness of global attention representations. By offering a consistent contextual signal to guide segmentation, *Fx* ensures that anomaly localisation from *Px* aligns with actual tampered regions. In turn, the *Px* module, implemented with SegFormer, refines the localisation through patch-level supervision. Together, these modules form a

cross-scale interaction that balances detection sensitivity, localisation precision, and interpretability; three essential pillars for operational forensic reliability.

Overall, these findings demonstrate that VAAS not only advances quantitative performance but also provides a principled foundation for interpretable and scalable forensic AI systems, motivating further exploration of its limitations and future extensions.

7.1. Limitations and future work

While VAAS demonstrates strong interpretability and competitive accuracy, several limitations remain. First, part of the evaluation was conducted on subsets of *DF2023*, which may not fully capture the diversity and complexity of the manipulations encountered in operational forensic imagery. A further limitation is that the current evaluation does not include cross-dataset testing, which is increasingly recognised as a necessary measure for assessing the robustness of AI-based forensic systems to unseen distributions.

Although *Hotels-50K* is not an image manipulation dataset and is not used in the experiments of this paper, it remains operationally relevant to digital forensics due to its role in real-world hotel room identification. Its inclusion is solely to motivate future extensions of VAAS beyond controlled manipulation benchmarks. In such domains, the goal shifts from detecting tampering to assessing broader visual integrity and provenance.

Future work will expand the framework to a wider range of manipulation categories, including composite AI–human edits, and will evaluate cross-domain generalisation on unseen datasets. Extending VAAS to video and multimodal forensics also represents a promising direction. Methodologically, incorporating adaptive thresholding within the harmonic scoring mechanism may enhance robustness across varied content domains. Beyond detection metrics, future research should quantify interpretability itself by assessing how effectively the generated anomaly maps assist human analysts in verifying manipulation evidence.

8. Conclusion

This paper presents VAAS, a vision–attention anomaly scoring framework for detecting and localising image manipulations in digital forensics. The approach combines a global attention analysis (*Fx*) with a local segmentation process (*Px*) through a hybrid harmonic scoring mechanism that quantifies image integrity. Evaluations on *CASIA v2.0* and *DF2023* demonstrate that VAAS achieves strong detection accuracy and stable localisation while producing interpretable attention heatmaps that support forensic decision-making.

The findings confirm that transformer-based attention representations effectively model spatial inconsistencies and that the proposed hybrid fusion yields a balanced and explainable measure of authenticity. By embedding interpretability into the detection process, VAAS advances forensic readiness and supports transparent verification of visual evidence in investigative and legal contexts. Future extensions will explore adapting VAAS to high-variability, real-world datasets such as *Hotels-50K*, where the focus shifts from manipulation detection to general image integrity and provenance verification.

References

Aftab, K., Adams, G., Scanlon, M., 2026. Plug to place: indoor multimedia geolocation from electrical sockets for digital investigation. Forensic Sci. Int.: Digit. Invest. 56S.

Bamigbade, O., Scanlon, M., Sheppard, J., 2025. Improving image embeddings with colour features in indoor scene geolocation. IEEE Access 13, 79860–79870.

Bappy, J.H., Simons, C., Nataraj, L., Manjunath, B.S., Roy-Chowdhury, A.K., 2019. Hybrid LSTM and encoder–decoder architecture for detection of image forgeries. IEEE Trans. Image Process. 28 (7), 3286–3300.

Chen, C.-F.R., Fan, Q., Panda, R., 2021. CrossViT: cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 357–366.

Christlein, V., Riess, C., Jordan, J., Riess, C., Angelopoulou, E., 2012. An evaluation of popular copy-move forgery detection approaches. IEEE Trans. Inf. Forensics Secur. 7 (6), 1841–1854.

Dong, J., Wang, W., Tan, T., 2013. CASIA image tampering detection evaluation database. In: 2013 IEEE China Summit and International Conference on Signal and Information Processing, pp. 422–426.

Du, X., Hargreaves, C., Sheppard, J., Anda, F., Sayakkara, A., Le-Khac, N.-A., Scanlon, M., 2020. SoK: exploring the state of the art and the future potential of artificial intelligence in digital forensic investigation. In: Proceedings of the 15th International Conference on Availability, Reliability and Security. ARES '20. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3407023.3407068.

Feng, D., Lu, X., Lin, X., 2020. Deep detection for face manipulation. In: Yang, H., Pasupa, K., Leung, A.C.-S., Kwok, J.T., Chan, J.H., King, I. (Eds.), Neural Information Processing. Springer International Publishing, Cham, pp. 316–323.

Fischinger, D., Boyer, M., 2023a. DF-Net: the digital forensics network for image forgery detection. In: Corcoran, P. (Ed.), Proceedings of 25th Irish Machine Vision and Image Processing Conference (IMVIP 2023), pp. 120–127. URL. https://iprcs.github.io/pdf/IMVIP2023_Proceeding.pdf.

Fischinger, D., Boyer, M., 2023b. DF2023: the digital forensics 2023 dataset for image forgery detection. In: Corcoran, P. (Ed.), Proceedings of 25th Irish Machine Vision and Image Processing Conference (IMVIP 2023), pp. 128–135. URL. https://iprcs.github.io/pdf/IMVIP2023_Proceeding.pdf.

Guarnera, L., Giudice, O., Battiato, S., 2024. Mastering deepfake detection: a cutting-edge approach to distinguish GAN and diffusion-model images. ACM Trans. Multimed Comput. Commun. Appl 20 (11). https://doi.org/10.1145/3652027.

Guo, X., Liu, X., Ren, Z., Grosz, S., Masi, I., Liu, X., 2023. Hierarchical fine-grained image forgery detection and localization. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3155–3165.

Hao, J., Zhang, Z., Yang, S., Xie, D., Pu, S., 2021. TransForensics: image forgery localization with dense self-attention. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 15035–15044.

Hargreaves, C., Breitinger, F., Dowthwaite, L., Webb, H., Scanlon, M., 2024. DFPulse: the 2024 digital forensic practitioner survey. Forensic Sci. Int.: Digit. Invest. 51, 301844. URL. https://www.sciencedirect.com/science/article/pii/S2666281724001719.

Kingra, S., Aggarwal, N., Kaur, N., 2024. SFormer: an end-to-end spatio-temporal transformer architecture for deepfake detection. Forensic Sci. Int.: Digit. Invest. 51, 301817. URL. https://www.sciencedirect.com/science/article/pii/S2666281724001410.

Li, Y., Chang, M.-C., Lyu, S., 2018. Ictu oculi: exposing AI created fake videos by detecting eye blinking. In: 2018 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–7.

Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S., 2020. Celeb-DF: a large-scale challenging dataset for DeepFake forensics. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3204–3213.

Ma, X., Du, B., Jiang, Z., Hammadi, A.Y.A., Zhou, J., 2023. IML-ViT: benchmarking image manipulation localization by vision transformer. arXiv preprint arXiv: 2307.14863. URL. https://arxiv.org/abs/2307.14863.

Mahdian, B., Saic, S., 2008. Detection of resampling supplemented with noise inconsistencies analysis for image forensics. In: 2008 International Conference on Computational Sciences and its Applications, pp. 546–556.

Matern, F., Riess, C., Stamminger, M., 2019. Exploiting visual artifacts to expose deepfakes and face manipulations. In: 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW), pp. 83–92.

Monika, Bansal, D., Passi, A., 2021. Image forensic investigation using discrete cosine transform-based approach. Wirel. Pers. Commun. 119 (4), 3241–3253. https://doi.org/10.1007/s11277-021-08396-1.

Munawar, M., Oussalah, M., 2025. Explainable dual-stream attention network for image forgery detection and localisation using contrastive learning. IET Radar, Sonar Navig. 19 (1), e70064.

Nguyen, H.H., Yamagishi, J., Echizen, I., 2022. Capsule-Forensics Networks for Deepfake Detection. Springer International Publishing, Cham, pp. 275–301. https://doi.org/10.1007/978-3-030-87664-7_13.

Pandey, R.C., Singh, S.K., Shukla, K.K., Agrawal, R., 2014. Fast and robust passive copy-move forgery detection using SURF and SIFT image features. In: 2014 9th International Conference on Industrial and Information Systems (ICIIS), pp. 1–6.

Sharma, S., Singh, B.K., Garg, H., 2025. Robust image forgery localization using hybrid CNN-transformer synergy based framework. Comput. Mater. Continua (CMC) 82 (3), 4691–4708.

Shi, C., Wang, C., Zhou, X., Qin, Z., 2024. DAE-Net: dual attention mechanism and edge supervision network for image manipulation detection and localization. IEEE Trans. Instrum. Meas. 73, 1–17.

Singh, J., 2022. Deepfakes: the threat to data authenticity and public trust in the age of AI-Driven manipulation of visual and audio content. J. AI-Assisted Scientific Discov. 2 (1), 428–467.

Spichiger, H., Adelstein, F., 2025. Preserving meaning of evidence from evolving systems. Forensic Sci. Int.: Digit. Invest. 52, 301867. DFRWS EU 2025 - Selected Papers from the 12th Annual Digital Forensics Research Conference Europe. URL. https://www.sciencedirect.com/science/article/pii/S266628172500006X.

Stylianou, A., Xuan, H., Shende, M., Brandt, J., Souvenir, R., Pless, R., 2019. Hotels-50K: a global hotel recognition dataset. In: Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence. AAAI'19/IAAI'19/EAAI'19. AAAI Press. https://doi.org/10.1609/aaai.v33i01.3301726.

Thomson, M., McKeown, S., Macfarlane, R., Leimich, P., 2025. 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, New York, NY, USA. https://doi.org/10.1145/3712716.3712724.

Verdoliva, L., 2020. Media forensics and DeepFakes: an overview. IEEE J. Selected Topics Signal Processing 14 (5), 910–932.

Wang, J., Wu, Z., Chen, J., Han, X., Shrivastava, A., Lim, S.-N., Jiang, Y.-G., 2022. ObjectFormer for image manipulation detection and localization. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2354–2363.

Wu, Y., AbdAlmageed, W., Natarajan, P., 2019. ManTra-Net: manipulation tracing network for detection and localization of image forgeries with anomalous features. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9535–9544.

Xia, X., Su, L.C., Wang, S.P., Li, X.Y., 2024. DMFF-Net: double-stream multilevel feature fusion network for image forgery localization. Eng. Appl. Artif. Intell. 127, 107200. URL. https://www.sciencedirect.com/science/article/pii/S0952197623013842.

Yadav, A., Vishwakarma, D.K., 2024. A Visually Attentive Splice Localization Network with Multi-Domain Feature Extractor and Multi-Receptive Field Upsampler arXiv preprint arXiv:2401.06995Preprint.

Zhou, P., Han, X., Morariu, V.I., Davis, L.S., 2018. Learning rich features for image manipulation detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1053–1061.

Zhuang, P., Li, H., Tan, S., Li, B., Huang, J., 2021. Image tampering localization using a dense fully convolutional network. IEEE Trans. Inf. Forensics Secur. 16, 2986–2999.