Journal of Future Artificial Intelligence and Technologies

E-ISSN: 3048-3719

Research Article

Tuned Attention-Guided Fusion of Deep and Handcrafted Features for Distinguishing AI-Generated and Human-Created Artworks

Vedaan Kumar and Apoorva Purohit *

Independent Researcher, Gurugram, Haryana 122003, India; e-mail : kvedaan@gmail.com; apoorva.pur@gmail.com

* Corresponding Author : Apoorva Purohit

Abstract: Generative AI models can now produce artwork that is virtually indistinguishable from human-created art, posing critical verification challenges for art galleries, educational institutions, and digital content platforms. Existing detection approaches face notable limitations: handcrafted feature–based methods offer interpretability but achieve limited accuracy, while deep learning approaches typically require substantial computational resources and provide minimal explanation for their predictions. We propose an attention-guided fusion framework that integrates Discrete Cosine Transform (DCT)–based frequency-domain features with deep learned representations through a learned attention mechanism, enabling both improved detection performance and interpretable decision-making grounded in signal processing theory. The attention module dynamically weights each feature modality based on input-specific reliability patterns, allowing adaptive fusion across diverse artistic styles and generation methods. We evaluate the proposed framework using two convolutional backbones: the lightweight MobileNetV2 for efficiency-critical deployment scenarios and the higher-capacity ResNet50 for settings where computational resources permit stronger feature extraction. Experiments are conducted on 18,288 artwork images spanning traditional paintings, digital illustrations, and outputs from multiple generative models, using stratified train–validation–test splits and five random seeds for statistical robustness. Under frozen-backbone settings, MobileNetV2-based attention fusion achieves an F1-score of 90.1%, while ResNet50-based attention fusion reaches 90.5%. When backbones are fine-tuned end-to-end, incorporating handcrafted features via attention fusion yields consistent performance gains: MobileNetV2 improves from 93.9% to 94.5% F1-score (+0.6%), and ResNet50 improves from 94.2% to 95.1% F1-score (+0.9%). Feature importance analysis further reveals that low-frequency DCT energy is the most discriminative handcrafted feature, confirming that frequency-domain characteristics effectively distinguish AI-generated from human-created artwork across diverse artistic styles. These results demonstrate that attention-guided fusion of signal processing features and deep learned representations provides consistent accuracy–efficiency benefits across both lightweight and heavyweight architectures, offering a practical and interpretable solution to the emerging challenge of AI-generated artwork detection.

Keywords: AI-generated art detection; Attention mechanism; Discrete cosine transform; Feature fusion; Frequency-domain analysis; Interpretability; MobileNetV2; ResNet50.

Received: November, 6th 2025
Revised: December, 18th 2025
Accepted: December, 25th 2025
Published: January, 4th 2026
Curr. Ver.: January, 4th 2026

1. Introduction

AI-based generative models can produce art that mimics human work so closely that distinguishing between the two had become a significant verification challenge with substantial real-world implications. Art galleries require reliable authentication systems to maintain buyer confidence and pricing integrity, as the provenance and authorship of artworks directly affect market value. Educational institutions must detect AI-generated submissions in art and

design programs to enforce academic integrity policies and ensure that students develop genuine creative skills rather than relying on automated generation tools. Digital content platforms face unprecedented scaling challenges as the volume of synthetic artwork grows exponentially, making manual review economically infeasible for verifying millions of daily uploads [1].

These applications demand detection systems that not only achieve high accuracy but also provide transparent decision-making grounded in interpretable features. Deep learning approaches based on convolutional neural networks achieve strong performance but often function as black boxes, offering limited insight into which image characteristics distinguish AI-generated from human-created artworks. This opacity reduces stakeholder trust and limits practitioners' ability to understand classification decisions. Conversely, handcrafted features grounded in signal processing theory offer clear interpretability but typically achieve limited accuracy when used in isolation. The central research question is whether combining these complementary approaches through learned adaptive fusion can simultaneously improve detection performance while preserving interpretability advantages. To address this question, we propose an attention-guided fusion framework that integrates frequency-domain handcrafted features with deep learned representations extracted from pretrained convolutional networks.

Our approach extends Discrete Cosine Transform (DCT)–based frequency analysis, traditionally applied to face-centric deepfake detection [2], to the substantially more heterogeneous domain of artwork encompassing diverse styles, media, and generative models. Frequency-domain features capture systematic spectral characteristics that neural generators struggle to reproduce perfectly, as demonstrated by Frank et al. [2]. However, such handcrafted features may fail to capture complex compositional and semantic patterns that vary widely across artistic styles. Deep learned representations, by contrast, model hierarchical visual structures and high-level semantic content through convolutional filters but lack explicit encoding of signal-processing principles. The proposed attention mechanism explicitly addresses this complementarity by dynamically weighting each modality according to input-specific reliability patterns.

The proposed framework is architecture-agnostic and is validated across two representative convolutional backbones with substantially different capacities: MobileNetV2 (3.4 million parameters), which emphasizes lightweight and efficient feature extraction, and ResNet50 (25.6 million parameters), which provides more expressive but computationally intensive representations [3], [4]. Evaluating across architectures is essential to determine whether performance gains arise from the attention-based fusion mechanism itself rather than from backbone-specific properties. Each architecture is evaluated under two training paradigms: a frozen-backbone setting where only the fusion and classification layers are trained, and an end-to-end fine-tuning setting where all network parameters are optimized.

Fine-tuning pretrained convolutional networks for domain adaptation constitutes a critical component of the proposed framework. While ImageNet-pretrained models provide strong general-purpose visual representations, the artwork domain exhibits fundamentally different characteristics, including intentional stylization, diverse artistic media, abstract compositions, and varying degrees of photorealism. Fine-tuning enables backbone networks to adapt their learned filters to artwork-specific patterns while retaining hierarchical feature extraction capabilities acquired during pretraining. By systematically evaluating both frozen and fine-tuned training paradigms, we isolate the individual contributions of domain adaptation, handcrafted features, and attention-based fusion.

The core contribution of this work is to demonstrate, through controlled experiments, that handcrafted features integrated via attention-guided fusion yield consistent, measurable performance improvements across both architectures and training paradigms. Using stratified data splits and evaluation over five random seeds, we isolate the effect of handcrafted features by comparing each backbone with and without fusion under identical training protocols. In addition to achieving high classification accuracy, the proposed framework preserves interpretability properties that are important for practical deployment. The remainder of this paper is organized as follows. Section 2 reviews related work and identifies research gaps motivating the proposed approach. Section 3 describes the methodology, including feature engineering, fusion architecture, and experimental design. Section 4 presents the experimental results. Section 5 discusses architectural insights, interpretability, and limitations. Section 6 concludes the paper and outlines directions for future research.

2. Literature Review

Traditional forensic techniques were not originally designed to address the challenges posed by modern generative models, while many recent deep learning–based approaches prioritize accuracy at the expense of interpretability and computational efficiency. Understanding where existing detection methods fall short for artistically diverse and style-rich artworks requires examining their evolution across several key research directions. This section reviews representative approaches, highlights their limitations, and identifies the research gaps the proposed framework addresses.

2.1. Classical Image Forensics and Early GAN Detection

Before the widespread adoption of AI-based image generation, classical image forensics primarily focused on detecting human-made manipulations by analyzing compression artifacts, sensor pattern noise, and resampling traces. Farid et al. [5]provided a comprehensive overview of traditional forgery detection techniques. While effective for conventional photo editing, these methods show limited effectiveness for AI-generated content, as generative models introduce fundamentally different artifact patterns originating from neural network architectures rather than explicit editing operations.

As generative models became more accessible, research attention shifted toward GAN detection. Early studies identified low-level artifacts in generated images. McCloskey and Albright [6]reported anomalies in the color-channel correlations in synthetic faces, while Marra et al. [7] identified architecture-specific fingerprints introduced by convolutional operations. More recent work has extended these analyses beyond GANs to diffusion-based generators [8], demonstrating that different generative architectures leave distinct and exploitable forensic signatures. However, such approaches often focus on specific artifact types associated with particular generation mechanisms, limiting their robustness and generalization across the rapidly evolving landscape of generative models.

2.2. Frequency-Domain Analysis Methods

Beyond spatial-domain artifacts, researchers explored the spectral characteristics of synthetic images. As the field progressed, frequency-domain analysis emerged as a particularly effective direction for AI-generated image detection [5], [9]. Frank et al. [2] demonstrated that DCT-based spectral analysis can reliably distinguish GAN-generated faces from real photographs by showing that neural generators distribute energy across frequency bands differently from natural imaging processes. Similarly, Zhang et al. analyzed artifacts in both spatial and frequency domains, revealing that even state-of-the-art GANs leave detectable spectral fingerprints.

Frequency-domain approaches exploit systematic differences between natural and synthetic images. The DCT provides near-optimal energy compaction for natural images by concentrating most of the information in the low-frequency coefficients. In contrast, neural generators trained with perceptual or adversarial losses primarily optimize spatial-domain similarity, potentially neglecting frequency-domain fidelity. Nevertheless, most existing frequency-based studies focus on facial imagery rather than artistically diverse content [10], [11]. Comprehensive evaluations of frequency-domain detection methods on heterogeneous artwork datasets spanning multiple styles and generative models remain limited in the literature [10]–[13].

2.3. Deep Learning Approaches

Deep learning–based methods remain the dominant paradigm for AI-image detection due to their ability to learn highly discriminative representations that often surpass traditional approaches in classification accuracy [14], [15]. CNN-based detectors commonly employ transfer learning from ImageNet-pretrained weights [4], [16], with architectures such as ResNet50 widely adopted due to their strong representational capacity and the availability of robust pretrained models.

Wang et al. [17] demonstrated that pretrained CNNs can transfer effectively to synthetic image detection despite the domain shift between natural photographs and generated content. Similarly, Tolosana et al. [18]surveyed deepfake and face manipulation detection methods, though their analysis primarily focused on facial imagery . Martin-Rodriguez et al. [15] proposed pixel-wise feature extraction combined with CNNs for detecting AI-created images,

achieving strong performance at the cost of high computational complexity and limited interpretability. However, this apparent transferability has important limitations. Yu et al. [19] showed that detectors trained on outputs from a single generative model often struggle to generalize to images produced by unseen architectures, suggesting that many detectors learn generator-specific artifacts rather than model-agnostic synthesis characteristics. In response, Ojha et al. [20] explored universal detection strategies to improve cross-generator robustness, concluding that generalization across diverse, rapidly evolving generative models remains an open and pressing challenge.

Despite their strong performance, CNN-based approaches typically operate as black boxes, offering limited interpretability and reducing stakeholder trust in automated decisions. This lack of transparency complicates deployment in real-world applications where explainability is often required [17], [21], [22].

2.3. Feature Fusion Techniques

As limitations of single-modality detectors became evident, recent studies began exploring feature fusion strategies that combine heterogeneous representations to improve robustness [23], [24]. Traditional ensemble methods typically rely on majority voting or simple feature concatenation, applying fixed fusion rules regardless of input characteristics. However, such static strategies fail to account for varying modality reliability across samples.

Zuama et al. [23] demonstrated that combining FaceNet and DenseNet201 features through learned fusion outperforms individual models for face spoofing detection, particularly when modalities contribute unevenly across inputs. Similarly, Yu and Xu [24] proposed a multi-modal texture fusion network for detecting AI-generated images, showing that integrating complementary feature representations improves detection robustness. Despite their effectiveness, these approaches focus primarily on texture-based features and do not explicitly incorporate frequency-domain information or provide detailed analysis of computational efficiency.

Attention-based mechanisms have emerged as powerful tools for adaptive multimodal fusion by enabling input-dependent weighting of feature sources [24], [25]. Vaswani et al. [25] introduced self-attention mechanisms within the transformer architecture, revolutionizing representation learning across multiple domains. Earlier work by Bahdanau et al. [26] demonstrated that learned attention for aligning and weighting heterogeneous information sources outperforms fixed combination strategies in neural machine translation. These principles naturally extend to multimodal fusion scenarios, where different feature modalities exhibit varying reliability depending on the input. However, attention-based fusion remains relatively underexplored in forensic detection tasks, particularly for artistically diverse content.

2.5. Research Gaps

Despite extensive prior research across a range of detection techniques (as summarized in Table 1), several critical gaps remain, particularly in the context of AI-generated artwork detection. These gaps directly motivate the methodological choices and contributions of this study. First, although frequency-domain analysis has proven effective for detecting synthetic face images[2], [12], its systematic evaluation on artistically diverse images spanning multiple styles and generation pipelines remains limited [1], [18], [27]. Artwork poses fundamentally different challenges compared to facial imagery due to intentional stylization, abstraction, and fewer structural constraints than photographic content. Consequently, findings derived from face-centric datasets cannot be assumed to generalize reliably to fine art. A comprehensive evaluation of frequency-based detection across diverse artistic genres and generative models is therefore still lacking.

Second, recent studies on artwork detection increasingly rely on deep learning approaches, revealing both their strengths and limitations. Agrawal et al. [1] investigated synthetic art generation and deepfake detection using Jamini Roy–inspired datasets, highlighting the difficulty of distinguishing AI-generated art from human artistic styles. Tinago et al. [27] employed CNN-based methods to differentiate AI-generated visual art from human-created art, achieving competitive accuracy but without addressing computational efficiency or interpretability—two factors essential for real-world deployment. While deep learning methods demonstrate strong classification performance, practical considerations such as inference cost, deployment feasibility, and explainability remain underexplored despite their critical

importance [28], [29]. Many existing works continue to optimize primarily for accuracy, overlooking training time, inference latency, and transparency requirements that determine operational viability [30].

Table 1. Summary of representative studies on detecting AI-generated and related synthetic images.

Reference	Method Category	Core Technique	Dataset Domain	Key Limitations
Frank et al. [2]	Frequency-Domain Detection	DCT-based spectral analysis	Face images	Face-centric evaluation; limited generalization beyond photographic facial content.
Wang et al. [17]	Deep Learning Detection	ResNet50 (ImageNet pretrained)	Synthetic images	High training cost (25.6M parameters); low interpretability due to black-box nature.
Tinago et al. [27]	Deep Learning Detection	Custom CNN classifier	Artwork	Limited analysis of computational efficiency and interpretability.
Yu et al. [24]	Feature Fusion	Multi-modal texture fusion network	AI-generated images	Focuses on texture features; does not incorporate frequency-domain descriptors.
Zuama et al. [23]	Feature Fusion	FaceNet + DenseNet201 with learned fusion	Face spoofing detection	Task-specific to face spoofing; limited generality for broader AI-image detection.
Our work	Attention-Guided Feature Fusion	DCT + MobileNetV2 / ResNet50 with attention-based fusion	Human- and AI-generated artwork	Single-dataset evaluation; focus on CNN-based backbones.

Third, although attention mechanisms have transformed numerous domains by enabling dynamic weighting of heterogeneous information sources, they remain insufficiently explored in forensic feature fusion contexts. In detection tasks, different feature modalities often exhibit varying reliability depending on input characteristics. However, most existing ensemble approaches rely on fixed feature concatenation or majority voting strategies that cannot adapt to such input-dependent variability [31], [32].

Fourth, existing fusion-based detection approaches are typically evaluated using a single backbone architecture. This practice makes it difficult to determine whether observed performance gains stem from the fusion strategy itself or from properties of the chosen feature extractor. Without cross-architecture validation, fusion benefits may reflect architecture-specific effects rather than genuine methodological contributions.

These gaps are addressed in this work through an attention-based fusion framework that integrates frequency-domain handcrafted features with deep learned representations. The proposed approach enables input-dependent feature weighting, systematic evaluation across artistically diverse content, controlled experimental comparisons, and explicit reporting of computational cost alongside detection performance. To assess the robustness of handcrafted feature fusion, the framework is evaluated across two backbone architectures—MobileNetV2 and ResNet50—with substantially different capacities and computational profiles, using identical training protocols and direct comparisons with and without fusion.

This combination of spectral analysis, adaptive attention-based fusion, controlled experimental validation, and cross-backbone evaluation constitutes the core novelty of this study. To the best of our knowledge, this is also the first work to extend frequency-domain detection techniques beyond face-centric applications to the broader and more heterogeneous domain of fine art, demonstrating their discriminative value across diverse artistic styles and generative models through rigorous controlled comparisons.

3. Proposed Method

We formulate AI artwork detection as a supervised binary classification problem. The proposed framework operates through a structured multi-stage pipeline that combines domain-informed handcrafted features with data-driven deep representations. Specifically, the method consists of four sequential stages:

1. image preprocessing and parallel feature extraction,
2. projection of heterogeneous features into a shared representation space,
3. attention-based adaptive fusion of handcrafted and deep features, and
4. training and final classification.

Unlike purely end-to-end learning approaches, our design explicitly incorporates signal–processing–based handcrafted features to encode interpretable domain knowledge, while allowing the model to adaptively weight these features based on input characteristics. The following subsections describe each stage in detail and justify the architectural choices based on established principles from signal processing and machine learning.

3.1. Image Preprocessing and Feature Extraction

3.1.1. Input Preprocessing

All input images $\mathrm{I} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times 3}$ are resized to 128×128 pixels using bilinear interpolation, yielding standardized inputs $I_{std} \in \mathbb{R}^{128\times128\times3}$. This resolution represents a trade-off between computational efficiency and preservation of sufficient spatial detail for frequency-domain analysis. Lower resolutions would degrade the quality of DCT-based spectral features, while higher resolutions would increase computational cost without proportional benefit, given the coarse spectral partitioning employed in this work. Pixel intensities are retained in the $[0,255]$ range prior to subsequent normalization steps.

3.1.2. Handcrafted Feature Engineering

We extract a total of 23 handcrafted features, denoted $f_{\text{hand}} \in \mathbb{R}^{23}$, designed to capture statistical, textural, frequency-domain, and structural characteristics shown to be discriminative in prior work [10] [33]–[35]. These features are grouped into four complementary categories, each grounded in established principles of signal processing and computer vision.

- **Color statistics (15 features):**

For each RGB channel, five statistical moments are computed from the flattened pixel intensity distribution: mean, standard deviation, median, skewness, and excess kurtosis. While generative models often match first- and second-order statistics during training, higher-order moments can expose subtle distributional biases introduced by synthetic generation processes [33], [34]. The resulting feature set comprises 15 values (5 statistics × 3 channels).

- **Texture descriptors (3 features).**

Images are first converted to grayscale using standard luminance weighting:

$$I_{gray} = 0.299 \cdot I_R + 0.587 \cdot I_G + 0.114 \cdot I_B \tag{1}$$

From the grayscale image, three texture-related descriptors are computed: Shannon entropy, contrast, and Laplacian variance. Shannon entropy, $H = -\sum_{i=0}^{255} p(i) \log_2 p(i)$ measures intensity randomness and complexity, where $p(i)$ denotes the normalized grayscale histogram. Contrast is defined as the standard deviation of grayscale intensities, capturing global variation, while Laplacian variance, $L = Var(\nabla^2 I_{gray})$ quantifies edge sharpness and local spatial structure [33], [34].

- **Frequency-domain analysis (3 features).**

Neural generators trained with perceptual losses emphasize spatial-domain similarity and do not explicitly enforce frequency-domain fidelity [13]. As a result, visually plausible synthetic images may still exhibit systematic spectral biases. We apply a two-dimensional DCT to grayscale images and follow established GAN-detection practices [2], [10]–[12]. The DCT coefficient matrix is partitioned into low- and high-frequency regions. For a 128×128 DCT, the low-frequency region corresponds to coefficients $D(i,j)$ with $i < 64$ and $j < 64$. The normalized low-frequency energy is computed as $E_{low} = \frac{1}{N_{low}} \sum_{i<64, j<64} |D(i,j)|$, where N_{low} denotes the number of low-frequency coefficients. This formulation leverages the near-optimal energy compaction property of the DCT for natural images, which is often imperfectly reproduced by generative models.

- **Edge characterization (2 features).**

Edges produced by adversarially trained generators may differ statistically from those in natural scenes or in deliberate artistic strokes. We apply Canny edge detection with thresholds of 50 and 150 to generate a binary edge map. From this map, two features are extracted: edge

density (fraction of edge pixels) and edge smoothness, measured as the mean contour perimeter of connected edge components [34].

Overall, this 23-dimensional feature set balances descriptive richness and computational efficiency. Rather than performing exhaustive feature engineering, we select a compact set of theoretically motivated descriptors that capture complementary aspects of color, texture, frequency, and structure.

3.1.3. Deep Feature Extraction

While handcrafted features encode interpretable, theoretically grounded information, they cannot capture all discriminative patterns in complex visual data. Deep convolutional networks complement this limitation by learning hierarchical representations directly from data. Accordingly, we extract deep features from pretrained convolutional backbones and integrate them into the fusion framework.

The proposed framework is architecture-agnostic and accepts features from arbitrary CNN backbones. To demonstrate this generality and evaluate fusion behavior across different points on the accuracy–efficiency spectrum, we consider two representative architectures: MobileNetV2 and ResNet50 [4].

- **MobileNetV2 (lightweight backbone).**

MobileNetV2 is selected for its favorable trade-off between accuracy and computational efficiency [3], [36], [37]. Its use of depthwise separable convolutions significantly reduces parameter count and runtime while maintaining competitive performance. With approximately 3.4 million parameters, MobileNetV2 is well suited for deployment in resource-constrained environments such as mobile or edge devices [38], [39]. Prior studies have validated its robustness across diverse image classification tasks and real-time applications [3], [39].

- **ResNet50 (heavyweight backbone).**

To assess whether fusion benefits extend to stronger feature extractors, we also evaluate ResNet50, which contains approximately 25.6 million parameters and represents a widely used high-capacity architecture. Residual connections enable stable training of deep networks and have been shown to yield strong performance across a broad range of visual recognition tasks [4], [40]. From ResNet50, we extract 2048-dimensional feature vectors from the global average pooling layer preceding the classification head. For both backbones, input images are normalized using ImageNet statistics:

$$I_{norm} = \frac{I_{std} - \mu_{ImageNet}}{\sigma_{ImageNet}} \tag{2}$$

Where $\mu_{ImageNet} = [0.485, 0.456, 0.406]$ and $\sigma_{ImageNet} = [0.229, 0.224, 0.225]$. Deep feature vectors are extracted as

$$f_{deep} = GlobalAvgPool\big(MobileNetV2(I_{norm})\big) \in \mathbb{R}^{1280} \tag{3}$$

$$f_{deep} = GlobalAvgPool\big(ResNet50(I_{norm})\big) \in \mathbb{R}^{2048} \tag{4}$$

Consistent with prior findings, ImageNet-pretrained CNN features transfer effectively to synthetic image detection, likely because low- and mid-level visual patterns such as edges, textures, and color distributions remain relevant across domains [17], [38].

3.2. Feature Projection into a Shared Representation Space

Handcrafted features with 23 dimensions and deep features with 1280 dimensions for MobileNetV2 (or 2048 dimensions for ResNet50) reside in fundamentally different geometric spaces and exhibit substantially different scales and semantic meanings. Direct concatenation or comparison of these features would be problematic for several reasons. First, higher-dimensional deep features would numerically dominate distance-based or weighted combinations simply due to their larger dimensionality. Second, the two feature types encode different forms of information: handcrafted features represent explicit signal-processing measurements, while deep features capture learned abstract patterns without inherent interpretability.

To address these issues, both feature modalities are projected into a shared 64-dimensional representation space using learnable linear transformations followed by ReLU nonlinearity:

$$h_{hand} = \mathrm{ReLU}(W_{\mathrm{hand}} f_{\mathrm{hand}} + b_{hand}) \tag{5}$$

$$h_{\mathrm{deep}} = \mathrm{ReLU}\left(W_{\mathrm{deep}} f_{\mathrm{deep}} + b_{\mathrm{deep}}\right) \tag{6}$$

where, f_{hand} and f_{deep} denote the handcrafted and deep feature vectors, respectively. The projection matrices satisfy $W_{\mathrm{hand}} \in \mathbb{R}^{64\times 23}$ and $W_{\mathrm{deep}} \in \mathbb{R}^{64\times 1280}$ for MobileNetV2 (or $W_{\mathrm{deep}} \in \mathbb{R}^{64\times 2048}$ for ResNet50), and b_{hand}, $b_{\mathrm{deep}} \in \mathbb{R}^{64}$ are learnable bias terms.

This projection serves multiple purposes. First, it enables meaningful fusion by mapping both modalities into a common representation space where their relative contributions can be compared directly. Second, it significantly reduces dimensionality for computational efficiency; for example, the 1280-dimensional MobileNetV2 feature vector is compressed by 20-fold to 64 dimensions. Third, enforcing equal dimensionality prevents either modality from dominating the fusion process purely due to scale, providing the attention mechanism with balanced inputs for learning adaptive weighting. The choice of a 64-dimensional projection space reflects a deliberate trade-off between representational expressiveness and computational efficiency. This dimensionality preserves sufficient capacity for discriminative information while maintaining a compact representation, resulting in approximately 2.8:1 compression for handcrafted features and 20:1 compression for deep features in the MobileNetV2-based configuration.

3.3. Attention-Based Adaptive Fusion

The core methodological contribution of this work lies in the use of learned adaptive fusion through an attention mechanism that dynamically determines the relative importance of handcrafted and deep features for each input image [26]. This design directly addresses the limitations of fixed fusion strategies, which apply identical weighting regardless of input characteristics and thus fail to adapt to heterogeneous visual content.

3.3.1. Concatenation and Attention Network

After projection into a shared embedding space, the handcrafted and deep feature representations are concatenated to provide the attention mechanism with joint access to both modalities:

$$h_{\mathrm{concat}} = \left[h_{hand}^{T}, h_{deep}^{T}\right]^{T} \in \mathbb{R}^{128} \tag{7}$$

This concatenated representation is processed by a lightweight two-layer attention network that computes input-dependent fusion weights:

$$a = \mathrm{softmax}(W_2 \tanh(W_1\, h_{\mathrm{concat}} + b_1) + b_2) \tag{8}$$

where $W_1 \in \mathbb{R}^{32\times 128}$, $b_1 \in \mathbb{R}^{32}$, $W_2 \in \mathbb{R}^{2\times 32}$, and $b_2 \in \mathbb{R}^{2}$ are learnable parameters. The softmax operation ensures that the resulting attention weights $a = \left[a_{hand}, a_{deep}\right]^{T}$ are non-negative and sum to one.

The use of a nonlinear hidden layer with the tanh activation function enables the attention network to capture complex interactions among feature modalities. This allows the model to learn input-specific reliability patterns, where certain combinations of handcrafted and deep features indicate that one modality should be emphasized over the other.

Traditional ensemble methods typically rely on simple concatenation, averaging, or majority voting, applying fixed fusion rules irrespective of input content [24]. In contrast, the proposed attention mechanism dynamically adapts the fusion strategy. For example, photorealistic AI-generated images may exhibit subtle frequency-domain artifacts that are better captured by handcrafted features, while highly stylized or abstract artworks may benefit more from deep representations that encode higher-level compositional structure.

3.3.2. Weighted Feature Integration and Classification

The final fused representation is computed as a weighted combination of the projected features

$$h_{fused} = a_{hand} \cdot h_{hand} + a_{deep} \cdot h_{deep} \in \mathbb{R}^{64} \tag{9}$$

This adaptive integration emphasizes the more reliable feature source for each input while maintaining a compact representation. Keeping the fused embedding at 64 dimensions ensures modest computational overhead while preserving complementary information from both modalities. The fused representation is then passed to a lightweight classification head. This consists of a fully connected layer with 32 hidden units and ReLU activation, followed by dropout with probability 0.3 to reduce overfitting. A final linear layer with softmax activation produces class probabilities for human-created (class 0) and AI-generated (class 1) artwork. By integrating classification directly into the attention fusion pipeline, the model remains end-to-end trainable while preserving interpretability and efficiency.

3.4. Training and Experimental Configuration

The proposed model is trained using categorical cross-entropy loss with L2 regularization, defined as:

$$\mathcal{L} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=0}^{1} y_i^j \log\left(\hat{y}_i^j\right) + \lambda \sum_{\theta} \theta^2 \tag{10}$$

Where B denotes the batch size, y_i^j is the one-hot encoded ground-truth label of sample i for class j, and $\hat{y}_i^j$ is the corresponding predicted probability. The regularization coefficient is set to $\lambda = 10^{-4}$ and applied to all learnable parameters θ.

During training, the optimization objective includes the projection layers, attention fusion module, and classification network. For fine-tuned variants, the entire model—including the backbone convolutional layers—is optimized end-to-end using backpropagation with the AdamW optimizer. For frozen-backbone variants, the backbone parameters remain fixed, while only the projection, attention, and classification layers are trainable.

Figure 1 provides a high-level overview of the attention-guided fusion architecture. Handcrafted features (23-dimensional) and deep features extracted from a pretrained backbone are projected into a shared 64-dimensional embedding space, concatenated, and fused through an attention mechanism that learns input-dependent weighting coefficients. The resulting fused representation is then passed to a lightweight classification head for binary prediction.

Although Figure 1 illustrates the MobileNetV2-based configuration, the same architecture is applied to the ResNet50-based variant by replacing the 1280-dimensional deep feature vector with a 2048-dimensional one. All remaining components—including the projection layers, attention network, and classifier—remain unchanged, highlighting the framework's modular design.

Algorithm 1 formally describes the complete attention fusion pipeline, detailing each computational stage from input preprocessing and feature extraction to attention-based fusion and final classification, with explicit dimensionality tracking. The algorithm corresponds directly to the architecture shown in Figure 1. It is used consistently during both training and inference, with parameter updates applied according to the selected training paradigm (frozen or fine-tuned).

To rigorously evaluate the contribution of handcrafted feature fusion, we adopt a controlled experimental design that minimizes confounding factors. Performance differences may arise from backbone capacity, training strategy (frozen vs. fine-tuned), or fusion mechanism. Our evaluation protocol explicitly controls these variables to isolate the incremental benefit of attention-guided handcrafted feature fusion.

Frozen-backbone baselines are first used to assess the discriminative value of individual feature types. A Handcrafted + Random Forest baseline evaluates signal-processing descriptors alone, while a Deep + Random Forest baseline assesses the transferability of frozen deep features. A Concatenation + Random Forest baseline further examines whether performance gains arise simply from combining feature vectors without adaptive weighting. Next, attention fusion models with frozen backbones are evaluated to determine whether learned weighting improves over fixed concatenation. Both MobileNetV2 and ResNet50 are tested under this setting to assess consistency across backbones with different representational capacities.

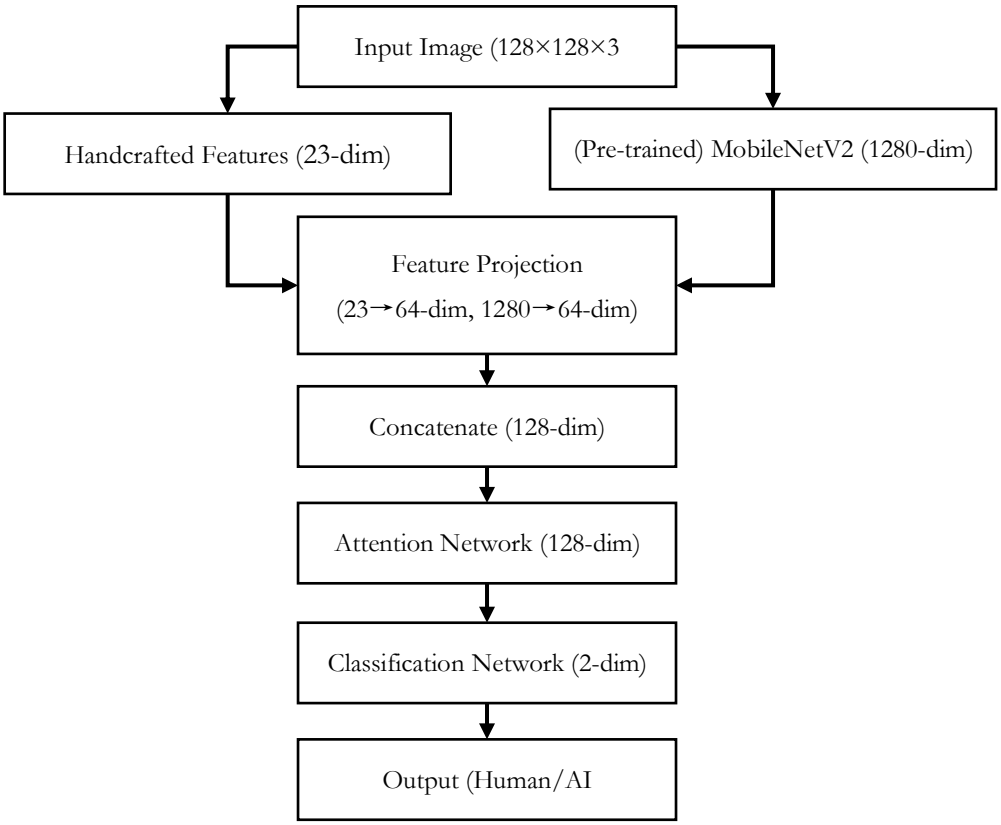

Figure 1. Attention-guided fusion architecture combining handcrafted and deep features for AI-generated artwork detection.

Algorithm 1. Attention Fusion-Based Art Authentication

INPUT: Image I, trained model parameters θ
OUTPUT: Prediction $\hat{y}$, attention weights a
// Stage 1: Feature Extraction
$I_{std} \leftarrow$ Resize(I, 128×128)
$f_{\text{hand}} \leftarrow$ ExtractHandcraftedFeatures(I_{std}) // 23-dim
$I_{norm} \leftarrow$ Normalize(I_{std}, $\mu_{ImageNet}$, $\sigma_{ImageNet}$)
$f_{\text{deep}} \leftarrow$ MobileNetV2(I_{norm}) // 1280-dim

// Stage 2: Projection
$h_{\text{hand}} \leftarrow$ ReLU($W_{\text{hand}} \cdot f_{\text{hand}} + b_{\text{hand}}$) // 64-dim
$h_{\text{deep}} \leftarrow$ ReLU($W_{\text{deep}} \cdot f_{\text{deep}} + b_{\text{deep}}$) // 64-dim

// Stage 3: Attention Fusion
$h_{\text{concat}} \leftarrow$ Concatenate(h_{hand}, h_{deep}) // 128-dim
$a \leftarrow$ softmax($W_2 \cdot \tanh(W_1 \cdot h_{\text{concat}} + b_1) + b_2$) // 2-dim weights
$h_{fused} \leftarrow a_{hand} \cdot h_{hand} + a_{\text{deep}} \cdot h_{\text{deep}}$ // 64-dim

Finally, fine-tuned experiments are conducted to disentangle the effects of backbone optimization and handcrafted feature fusion. For each backbone, a fine-tuned model without handcrafted features is directly compared to its counterpart trained with attention-guided fusion under identical training schedules, data splits, augmentation strategies, and hyperparameters. This design isolates whether observed improvements stem from handcrafted features rather than from fine-tuning alone.

Across all experiments, identical data splits, training protocols, and evaluation procedures are used to ensure fairness. This controlled setup enables a clear and interpretable assessment of the conditions under which handcrafted features and attention-based fusion contribute to improved detection performance.

3.5. Experimental Setup

We evaluate the proposed framework on a publicly available dataset from Kaggle, selected for its substantial artistic diversity and suitability for assessing generalization. The dataset consists of two balanced subsets: 9,144 human-created images and 9,144 AI-generated images. The human-created subset spans a wide range of artistic forms, including traditional oil paintings, watercolors, acrylics, digital illustrations, photography, and mixed media, and encompasses styles from photorealism and impressionism to cubism, surrealism, abstract expressionism, and contemporary art. The AI-generated subset contains images produced by two diffusion-based models—Standard Diffusion and Latent Diffusion—which represent state-of-the-art generative approaches capable of producing visually convincing synthetic artwork.

We acknowledge that exhaustive manual verification of all labels was not performed and that some degree of label noise may exist, particularly for edge cases such as heavily post-processed human artworks or procedurally generated content. Nevertheless, the dataset provides sufficient scale, stylistic diversity, and realism to enable meaningful evaluation of AI artwork detection methods.

To ensure fair and unbiased evaluation, we employ stratified random splitting to maintain class balance across all subsets: 70% for training (12,801 images), 15% for validation (2,743 images), and 15% for testing (2,744 images). The validation set is used exclusively for model selection, selecting the checkpoint with the highest validation F1-score, while the test set is held out until final evaluation. To assess robustness and statistical reliability, all experiments are repeated across five independent random seeds. For each seed, data are re-split, model parameters are re-initialized, and training is performed from scratch, controlling for randomness arising from both data partitioning and weight initialization. All methods operate on images resized to a uniform resolution of 128 × 128 pixels.

Data augmentation strategies are applied selectively based on training paradigm and architectural constraints to ensure methodological fairness. End-to-end trained models—including fine-tuned MobileNetV2, ResNet50, and their handcrafted attention fusion variants—employ standard image-level augmentation consisting of random resized crops and horizontal flips. This follows established best practices for convolutional neural networks trained directly on pixel data. In contrast, frozen-backbone methods operate on pre-extracted deep features that are computed once from unaugmented images and cached in memory throughout training. Image-level augmentation is not applied in this setting for three reasons: (1) frozen feature extractors produce identical feature vectors for a given input regardless of repetition, (2) fusion training operates solely on cached features without accessing raw images, and (3) re-extracting features from augmented images at every epoch would negate the computational efficiency advantages of frozen-backbone approaches. This design ensures that each method is evaluated under its optimal and practically realistic training protocol rather than under artificially constrained conditions.

Table 2. Summarizes the training hyperparameter settings used across all experiments.

Parameter	Value
Optimizer	AdamW
Batch Size	64
Loss Function	CrossEntropyLoss
Learning Rate	1e-4
Weight Decay (L2)	1e-4
Epochs	5
Precision	Mixed (FP16)

Training hyperparameters are kept consistent across all methods to enable fair comparison. We do not employ automated hyperparameter optimization techniques such as grid search or Optuna. All models are trained using the AdamW optimizer with a learning rate of $1e{-}4$ and weight decay of $1e{-}4$ for L2 regularization. The batch size is fixed at 64, and training proceeds for exactly five epochs, with the model checkpoint achieving the highest validation F1-score selected for final testing. Mixed-precision (FP16) training is used to reduce memory consumption and accelerate computation without sacrificing accuracy. All

experiments are implemented in PyTorch 1.12 with CUDA 11.3 and conducted on NVIDIA Tesla T4 GPUs.

This controlled experimental setup ensures that observed performance differences arise from methodological choices rather than variations in training configuration or computational environment. To support reproducibility, all implementation details—including data loading, preprocessing, model architectures, training procedures, and evaluation protocols—are documented in publicly accessible code, as described in the Data Availability Statement. The complete set of training hyperparameters used consistently across all experimental configurations is summarized in Table 2.

4. Implementation and Results

In this section, we evaluate the proposed framework against baseline methods across multiple performance dimensions. We adopt F1-score as the primary evaluation metric because it provides a balanced assessment of precision and recall, which is particularly important for fair decision-making in binary classification tasks. Although the dataset is perfectly balanced (9,144 images per class), reducing the risk of misleading accuracy scores, the F1-score remains a conservative and widely accepted metric in detection research.

Precision measures the proportion of predicted AI-generated artworks that are truly synthetic, which is critical for minimizing false accusations that could unfairly disadvantage human artists. Recall reflects the proportion of AI-generated artworks correctly identified, ensuring detection completeness and preventing synthetic content from evading scrutiny. Accuracy provides an overall correctness measure but does not distinguish between error types, while ROC-AUC offers a threshold-independent evaluation of class separability. Together, these metrics provide a comprehensive performance characterization, with F1-score serving as the primary basis for comparison.

4.1. Comprehensive Performance Comparison

Our evaluation follows a controlled experimental design to isolate the contribution of handcrafted features while maintaining consistency across all other methodological factors. The comparison is structured along two complementary axes. First, we analyze frozen-backbone methods to establish baseline performance for handcrafted features alone, deep features alone, simple feature concatenation, and the proposed attention-based fusion. Second, we compare identical architectures trained end-to-end with and without handcrafted feature fusion, while holding constant training schedules, data augmentation, optimization procedures, and hyperparameters. This design ensures that any observed performance differences can be explicitly attributed to the inclusion of handcrafted features via attention mechanisms. Table 3 reports the performance of all evaluated methods, with results averaged over five random seeds. The methods are grouped to enable fair and direct comparisons between frozen-backbone approaches, fine-tuned baselines without handcrafted features, and fine-tuned variants incorporating handcrafted attention fusion.

4.1.3. Frozen Backbone Results

Using handcrafted features alone achieves an F1-score of 80.9% with a pipeline time of 182 seconds, indicating that signal-processing descriptors capture meaningful discriminative cues but are insufficient on their own. Frozen deep features from MobileNetV2 achieve an 87.2% F1-score in 129 seconds, confirming effective transfer learning from ImageNet pretraining. Simple concatenation of handcrafted and deep features yields only a modest gain (87.9% F1-score) while incurring substantial computational overhead (318 seconds), suggesting that naive fusion does not effectively exploit feature complementarity.

In contrast, the proposed attention-based fusion with frozen MobileNetV2 achieves 90.1% F1-score in 262 seconds, representing a 2.9 percentage point improvement over deep features alone. The learned attention mechanism dynamically weights feature modalities per input, outperforming fixed concatenation while maintaining a reasonable computational cost. Small standard deviations across five seeds indicate stable and reproducible performance.

Applying the same fusion strategy to ResNet50 further improves performance to 90.5% F1-score in 310 seconds, demonstrating that handcrafted features also benefit stronger backbones. Although ResNet50 yields slightly higher accuracy, both architectures show consistent gains from attention-guided fusion.

Table 3. Performance comparison across detection methods (averaged over five random seeds).

Method	F1-score (%)	Precision (%)	Recall (%)	Accuracy (%)	ROC-AUC	Total Pipeline time (second)
Handcrafted	80.9±0.4	79.3±0.5	82.6±0.6	79.0±0.4	0.870±0.003	181.6±3.3
Deep features (MobileNetV2)	87.2±0.1	86.8±0.2	87.6±0.2	85.5±0.1	0.932±0.001	129.1±1.2
Concatenation (MobileNetV2 + Handcrafted)	87.9±0.1	87.3±0.2	88.5±0.2	86.3±0.1	0.936±0.001	318.3±2.4
Attention Fusion (MobileNetV2 + Handcrafted)	90.1±0.1	89.8±0.2	90.4±0.2	88.9±0.2	0.959±0.001	261.7±4.8
Attention Fusion (ResNet50 + Handcrafted)	90.5±0.2	89.8±0.2	91.3±0.2	89.5±0.2	0.958±0.001	310.3±5.1
MobileNetV2 (Fine-tuned)	93.9±0.2	93.1±0.2	94.6±0.2	93.3±0.2	0.982±0.001	365.4±5.1
Attention Fusion (MobileNetV2 Fine-tuned + Handcrafted)	94.5±0.2	94.1±0.2	94.9±0.3	93.6±0.2	0.983±0.001	382.8±6.2
ResNet50 (Fine-tuned)	94.2±0.2	93.9±0.2	94.5±0.2	93.6±0.2	0.983±0.001	394.2±6.0
Attention Fusion (ResNet50 Fine-tuned + Handcrafted)	95.1±0.2	94.7±0.2	95.4±0.2	94.3±0.2	0.985±0.002	409.6±7.4

4.1.2. Fine-Tuning Comparisons

End-to-end fine-tuning of MobileNetV2 without handcrafted features achieves 93.9% F1-score with a pipeline time of 365 seconds. Introducing handcrafted features via attention fusion increases the F1-score to 94.5% with a modest runtime increase to 383 seconds. This controlled comparison demonstrates that handcrafted features provide a measurable benefit even when the backbone is already optimized through fine-tuning.

A similar trend is observed for the heavier ResNet50 architecture. Fine-tuned ResNet50 achieves 94.2% F1-score in 394 seconds, while adding handcrafted attention fusion improves performance to 95.1% F1-score with a pipeline time of 410 seconds. The consistent gains across both architectures indicate that attention-based integration of handcrafted and deep features offers complementary information beyond what is learned through end-to-end optimization alone.

4.1.3. Architecture-Dependent Fusion Benefits and Trade-offs

The observed fusion gains are architecture-dependent but consistently positive. For MobileNetV2, a 0.6 percentage point improvement is notable given the already high baseline performance, where further gains become increasingly difficult. For ResNet50, the larger 0.9 percentage point improvement suggests that even high-capacity models benefit from explicit encoding of frequency-domain characteristics that may not be fully captured by deep representations alone.

Comparing frozen and fine-tuned variants highlights practical trade-offs between accuracy and computational cost. MobileNetV2 attention fusion with a frozen backbone achieves 90.1% F1-score in 262 seconds, while its fine-tuned counterpart reaches 94.5% F1-score at the cost of 383 seconds. A similar pattern is observed for ResNet50. These results suggest that model selection should be guided by deployment requirements, balancing detection accuracy against available computational resources.

Overall, the results in Table 3 demonstrate that the proposed attention-guided fusion achieves its strongest performance when combined with end-to-end fine-tuning. While both attention-based feature fusion and backbone fine-tuning individually improve detection

accuracy, their combination consistently yields the highest F1-scores across both lightweight (MobileNetV2) and heavyweight (ResNet50) architectures. This indicates that handcrafted features provide complementary information that is most effectively exploited when backbone representations are adapted to the artwork domain via fine-tuning.

4.2. Qualitative Error Analysis

To better understand the behavior of the proposed model beyond aggregate metrics, we conduct a qualitative error analysis on representative test samples. Figure 2 presents 20 randomly selected artworks from the test set, illustrating both correct and incorrect predictions produced by the MobileNetV2-based Attention Fusion model.

The first two rows show correctly classified examples, including five human-created artworks and five AI-generated artworks. The correctly identified human artworks span diverse styles and media, including traditional oil paintings with visible brushstrokes, impressionist landscapes characterized by natural color blending, abstract geometric compositions, digital illustrations, and watercolor paintings exhibiting organic texture. Correctly classified AI-generated samples include photorealistic landscapes, stylized portraits, abstract digital compositions, architectural renderings, and fantasy scenes, indicating that the model can effectively handle a wide range of generative styles.

The bottom two rows highlight misclassified cases, revealing recurring failure modes. False positives—human-created artworks incorrectly classified as AI-generated—often correspond to heavily post-processed digital art, geometric abstract compositions with procedurally generated appearance, artworks with extensive filter application, minimalist designs with unnaturally clean edges, and photographs subjected to strong color grading. In these cases, deliberate artistic choices or digital editing introduce visual patterns that resemble synthetic artifacts targeted by the model.

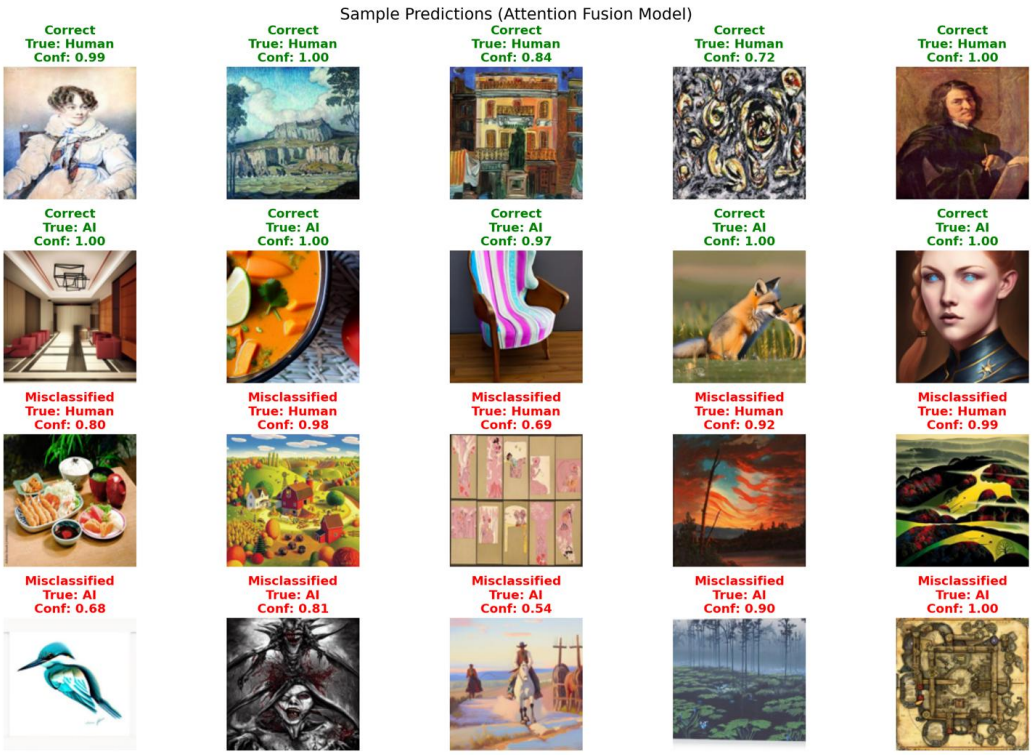

Figure 2. Representative test samples illustrating correct predictions and misclassifications produced by the MobileNetV2-based attention fusion model across diverse artistic styles.

Conversely, false negatives—AI-generated artworks misclassified as human-created—include photorealistic landscapes that closely replicate natural frequency characteristics, portraits with convincing texture detail, AI-generated images emulating traditional painting styles, architectural scenes with realistic material rendering, and organic abstract compositions lacking obvious synthetic cues. These cases arise when generative models successfully reproduce

the statistical properties that the handcrafted and deep features are designed to capture. These error patterns highlight an inherent challenge in AI artwork detection: the boundary between human and AI-created art is not always sharply defined. Human artists may intentionally employ digital techniques that mimic synthetic patterns, while advanced generative models increasingly replicate natural and artistic image statistics with high fidelity. These observations reinforce the importance of combining complementary cues and avoiding overconfident claims of perfect separability.

4.3. Feature Importance and Attention Mechanism Analysis

To further interpret the behavior of the proposed framework, we analyze both the discriminative power of handcrafted features and the learned behavior of the attention fusion mechanism. Figure 3 provides complementary quantitative insights into these aspects.

Figure 3(a) reports the top-ranked handcrafted features based on mean permutation importance derived from the Random Forest baseline. The mean low-frequency DCT energy emerges as the most informative feature, with an importance score of 0.070 ± 0.001 across random seeds. This result directly supports our hypothesis that frequency-domain characteristics capture systematic differences between AI-generated and human-created artworks. Importantly, this finding extends prior frequency-based observations from face-centric detection studies to a much broader and stylistically diverse art domain.

Beyond frequency features, several texture- and color-related descriptors also exhibit high importance. Edge density and edge smoothness contribute substantially, indicating that structural boundary characteristics remain informative cues. Color statistics, such as the green channel standard deviation (0.065 ± 0.002), red channel mean (0.055 ± 0.001), and blue channel standard deviation (0.053 ± 0.003), further reinforce the idea that subtle distributional differences across color channels play a meaningful role. Notably, no single handcrafted feature overwhelmingly dominates the ranking, suggesting genuine complementarity across multiple signal characteristics rather than reliance on a single discriminative cue.

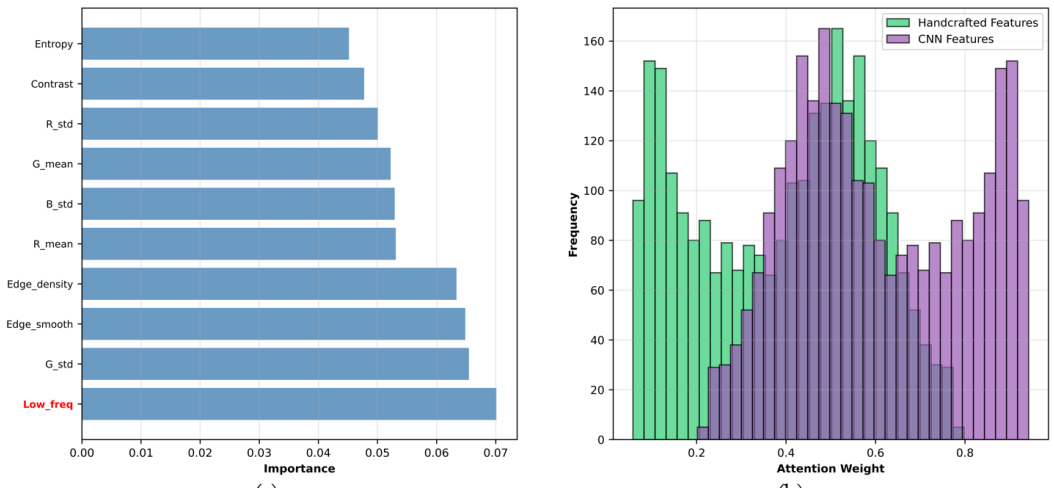

Figure 3. Analysis of feature importance and learned attention behavior in the MobileNetV2-based attention fusion model: (**a**) top-ranked handcrafted features based on permutation importance; (**b**) distribution of learned attention weights for handcrafted and deep features across test samples.

Figure 3(b) visualizes the learned attention weight distributions for handcrafted and deep features in the MobileNetV2-based (frozen) Attention Fusion model. Across all test samples and random seeds, handcrafted features receive an average attention weight of 0.435 ± 0.012, while deep CNN features receive 0.565 ± 0.014. This indicates a modest overall preference toward deep representations, consistent with the stronger baseline performance of deep features alone (as shown in Table 3). However, the widespread and overlapping distributions demonstrate that the model does not converge to a fixed weighting. Instead, it adapts dynamically per sample.

This adaptive behavior explains why attention-based fusion consistently improves performance across architectures and training paradigms. For inputs where frequency-domain or

low-level structural cues are highly informative, the attention mechanism assigns greater weight to handcrafted features. Conversely, for images where high-level semantic or compositional patterns dominate, the model relies more heavily on deep learned representations. This dynamic balancing behavior is observed consistently despite substantial differences in backbone capacity between MobileNetV2 and ResNet50.

Taken together, these analyses validate that the proposed attention mechanism performs meaningful feature integration rather than merely increasing model capacity. The consistent performance gains observed across metrics and random seeds, combined with interpretable feature importance and adaptive attention behavior, support the claim that integrating handcrafted signal-processing features with deep representations yields tangible and explainable benefits. For AI artwork detection, this hybrid approach offers a practical alternative to purely end-to-end models, particularly in scenarios where transparency and interpretability are valued alongside accuracy.

5. Discussion and Implications

This section interprets the experimental findings and discusses their implications for model design, interpretability, and practical deployment, while also outlining limitations that motivate future research.

5.1. Cross-Architecture Generalization and Fusion Benefits

Our results demonstrate that handcrafted features provide consistent performance gains across both lightweight and heavyweight backbones, although the magnitude of improvement depends on the base model capacity. For MobileNetV2 fine-tuned end-to-end, integrating handcrafted features via attention fusion increases the F1-score from 93.9% to 94.5%. For the higher-capacity ResNet50, the improvement is more pronounced, rising from 94.2% to 95.1%.

This pattern suggests that stronger backbones may be better positioned to exploit complementary information introduced by handcrafted features. While deep networks learn powerful hierarchical representations, the observed gains indicate that frequency-domain characteristics are not fully captured through standard end-to-end optimization alone. Neural generators trained with perceptual losses primarily optimize spatial similarity, potentially leaving systematic spectral artifacts insufficiently modeled by convolutional filters. Explicit DCT-based descriptors, therefore, continue to provide complementary discriminative signals even when combined with fine-tuned, high-capacity networks.

The attention mechanism itself plays a critical role in realizing these gains. In frozen-backbone settings, attention-based fusion clearly outperforms fixed concatenation (90.1% vs. 87.9% F1 for MobileNetV2), confirming that adaptive, input-dependent weighting is more effective than static fusion rules. The learned attention selectively emphasizes handcrafted features when frequency cues are reliable, while favoring deep representations for images where high-level compositional patterns are more informative.

5.2. Feature Interpretability and Attention Behavior

Beyond accuracy improvements, the interpretability offered by handcrafted features is a central strength of the proposed framework. Feature importance analysis highlights low-frequency DCT energy as the most discriminative handcrafted descriptor, validating prior observations from face-centric detection studies and extending them to the far more diverse domain of artwork. This finding confirms that generative models introduce systematic spectral biases that persist across styles and generation pipelines.

Additional contributions from edge density, edge smoothness, and color statistics further indicate that no single descriptor dominates the decision process. Instead, discrimination emerges from complementary cues spanning frequency, texture, and color characteristics. These insights provide actionable explanations unavailable in purely black-box deep learning approaches, allowing practitioners to reason about why particular images are classified as AI-generated or human-created based on measurable signal properties.

Attention weight analysis reinforces this interpretation. While deep features receive slightly higher average weights, substantial variance across samples confirms genuinely dynamic fusion behavior rather than fixed preference. The model adapts its reliance on

handcrafted versus learned representations depending on image characteristics, explaining why fusion benefits both frozen and fine-tuned architectures.

5.3. Practical Deployment Considerations

The proposed framework supports multiple deployment configurations tailored to different operational constraints. Frozen-backbone attention fusion offers a favorable balance between efficiency, interpretability, and performance. Such configurations are well-suited for resource-constrained environments, including mobile applications, edge computing platforms, and large-scale moderation systems where per-image computational cost accumulates rapidly.

In contrast, fine-tuned fusion variants are preferable in high-stakes scenarios where maximum accuracy is paramount. Applications such as art authentication, museum acquisition, and forensic analysis justify the additional computational overhead, as even modest performance gains translate into fewer costly errors. The observed improvements for fine-tuned ResNet50 with handcrafted fusion demonstrate that accuracy gains remain achievable even at high baseline performance levels [40], [41].

5.4. Limitations and Future Directions

Several limitations warrant discussion. First, dataset labeling relies on contributor-provided annotations rather than verified provenance, which can introduce label noise. This reflects a broader challenge in generative AI research, where reliable documentation of authorship is often unavailable. Future work should prioritize curated datasets with verified metadata, multi-dataset evaluation, and more nuanced labeling schemes that capture human–AI collaboration.

Second, while our handcrafted features are theoretically motivated, more expressive spectral representations—such as wavelets or learned frequency embeddings—may uncover subtler artifacts. Automated feature discovery methods also offer promising directions for extending beyond manually designed descriptors. Finally, although attention weights indicate the importance of each modality, they do not fully explain which internal deep features drive decisions. Integrating explainability techniques such as concept activation vectors or prototype-based models could further improve transparency, particularly for deep representations.

6. Conclusion

This work addresses a central challenge in AI-generated artwork detection: achieving high classification performance while maintaining interpretability and practical deployability. Through controlled experiments across multiple architectures and training paradigms, we demonstrate that attention-guided fusion of handcrafted signal-processing features with deep learned representations yields consistent and measurable improvements. Our results show that handcrafted frequency-domain descriptors remain valuable even when combined with fine-tuned deep networks, indicating that standard end-to-end training does not fully capture spectral characteristics introduced by generative models. The proposed attention mechanism enables dynamic, input-dependent fusion, allowing the model to adaptively leverage complementary information sources rather than relying on fixed combination strategies.

The framework supports flexible deployment. Lightweight frozen-backbone configurations provide efficient and interpretable solutions for large-scale or resource-constrained settings, while fine-tuned fusion variants deliver maximal accuracy for high-stakes applications. Notably, the inclusion of handcrafted features enables explanations grounded in signal processing theory, offering transparency that purely deep learning approaches often lack. Beyond artwork detection, the underlying principle of adaptive fusion extends naturally to other domains involving heterogeneous data sources with variable reliability. By demonstrating the benefits of combining domain knowledge with learned representations in a principled and interpretable manner, this work contributes to a broader understanding of how hybrid models can support trustworthy, informed decision-making at the boundary between human creativity and machine-generated content.

Author Contributions: Conceptualization: A.P. and V.K.; Methodology: A.P.; Software: A.P. and V.K.; Validation: A.P. and V.K.; Formal analysis: A.P. and V.K.; Investigation: A.P. and V.K.; Resources: A.P. and V.K.; Data curation: A.P. and V.K.; Writing—original draft preparation: A.P.; Writing—review and editing: A.P.; Visualization: V.K.; Supervision: A.P.; Project administration: A.P.; Funding acquisition: NA. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The dataset used in this study is publicly available at www.kaggle.com/datasets/kausthubkannan/ai-and-human-art-classification. To facilitate reproducibility, the Python code is available at https://github.com/apoorvapu/data_science/blob/main/AI_art_human.ipynb. We commit to maintaining this repository and to responding to inquiries about reproducing results to support open science principles.

Acknowledgments: The authors thank the Kaggle community for providing the dataset used in this study. ChatGPT was used for language improvement after the authors completed all research, analysis, and original writing. No AI tools were used for experimental design, data analysis, result generation, figure creation, or substantive content development. All scientific contributions, methodological decisions, interpretations, and intellectual content are solely the work of the authors, who take full responsibility for the accuracy and integrity of the entire manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] K. Agrawal and R. Banerjee, "Synthetic Art Generation and Deepfake Detection: A Study on Jamini Roy Inspired Dataset," *SSRN*. Mar. 29, 2025. doi: 10.2139/ssrn.5358869.

[2] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *Proceedings of the 37th International Conference on Machine Learning*, 2020. [Online]. Available: https://dl.acm.org/doi/abs/10.5555/3524938.3525242

[3] K. Dong, C. Zhou, Y. Ruan, and Y. Li, "MobileNetV2 Model for Image Classification," in *2020 2nd International Conference on Information Technology and Computer Application (ITCA)*, Dec. 2020, pp. 476–480. doi: 10.1109/ITCA52113.2020.00106.

[4] B. Koonce, "ResNet 50," in *Convolutional Neural Networks with Swift for Tensorflow*, Berkeley, CA: Apress, 2021, pp. 63–72. doi: 10.1007/978-1-4842-6168-2_6.

[5] H. Farid, "Image forgery detection," *IEEE Signal Process. Mag.*, vol. 26, no. 2, pp. 16–25, Mar. 2009, doi: 10.1109/MSP.2008.931079.

[6] S. McCloskey and M. Albright, "Detecting GAN-Generated Imagery Using Saturation Cues," in *2019 IEEE International Conference on Image Processing (ICIP)*, Sep. 2019, pp. 4584–4588. doi: 10.1109/ICIP.2019.8803661.

[7] F. Marra, D. Gragnaniello, D. Cozzolino, and L. Verdoliva, "Detection of GAN-Generated Fake Images over Social Networks," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, Apr. 2018, pp. 384–389. doi: 10.1109/MIPR.2018.00084.

[8] R. Corvi, D. Cozzolino, G. Poggi, K. Nagano, and L. Verdoliva, "Intriguing properties of synthetic images: from generative adversarial networks to diffusion models," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Jun. 2023, pp. 973–982. doi: 10.1109/CVPRW59228.2023.00104.

[9] Y. Zhang, S. Huang, L. Huangfu, and D. Dajun Zeng, "Learning Feature Exploration and Selection With Handcrafted Features for Few-Shot Learning," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 55, no. 4, pp. 2599–2610, Apr. 2025, doi: 10.1109/TSMC.2024.3524390.

[10] D. V. Fevralev, N. N. Ponomarenko, V. V. Lukin, S. K. Abramov, K. O. Egiazarian, and J. T. Astola, "Efficiency analysis of DCT-based filters for color image database," in *Proceedings Volume 7870, Image Processing: Algorithms and Systems IX*, Feb. 2011, p. 78700R. doi: 10.1117/12.871944.

[11] F. Franzen, "Image Classification in the Frequency Domain with Neural Networks and Absolute Value DCT," in *Image and Signal Processing*, 2018, pp. 301–309. doi: 10.1007/978-3-319-94211-7_33.

[12] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and Simulating Artifacts in GAN Fake Images," in *2019 IEEE International Workshop on Information Forensics and Security (WIFS)*, Dec. 2019, pp. 1–6. doi: 10.1109/WIFS47025.2019.9035107.

[13] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution," in *Computer Vision – ECCV 2016*, 2016, pp. 694–711. doi: 10.1007/978-3-319-46475-6_43.

[14] S. Kan, Y. Cen, Z. He, Z. Zhang, L. Zhang, and Y. Wang, "Supervised Deep Feature Embedding With Handcrafted Feature," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5809–5823, Dec. 2019, doi: 10.1109/TIP.2019.2901407.

[15] F. Martin-Rodriguez, R. Garcia-Mojon, and M. Fernandez-Barciela, "Detection of AI-Created Images Using Pixel-Wise Feature Extraction and Convolutional Neural Networks," *Sensors*, vol. 23, no. 22, p. 9037, Nov. 2023, doi: 10.3390/s23229037.

[16] A. Khan *et al.*, "A survey of the vision transformers and their CNN-transformer based variants," *Artif. Intell. Rev.*, vol. 56, no. S3, pp. 2917–2970, Dec. 2023, doi: 10.1007/s10462-023-10595-0.

[17] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-Generated Images Are Surprisingly Easy to Spot… for Now," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 8692–8701. doi: 10.1109/CVPR42600.2020.00872.

[18] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A Survey of face manipulation and fake detection," *Inf. Fusion*, vol. 64, pp. 131–148, Dec. 2020, doi: 10.1016/j.inffus.2020.06.014.

[19] N. Yu, L. Davis, and M. Fritz, "Attributing Fake Images to GANs: Learning and Analyzing GAN Fingerprints," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2019, pp. 7555–7565. doi: 10.1109/ICCV.2019.00765.

[20] U. Ojha, Y. Li, and Y. J. Lee, "Towards Universal Fake Image Detectors that Generalize Across Generative Models," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 24480–24489. doi: 10.1109/CVPR52729.2023.02345.

[21] M. Groh, Z. Epstein, C. Firestone, and R. Picard, "Deepfake detection by human crowds, machines, and machine-informed crowds," *Proc. Natl. Acad. Sci.*, vol. 119, no. 1, Jan. 2022, doi: 10.1073/pnas.2110013119.

[22] W. Fang, F. Zhang, V. S. Sheng, and Y. Ding, "A Method for Improving CNN-Based Image Recognition Using DCGAN," *Comput. Mater. Contin.*, vol. 57, no. 1, pp. 167–178, 2018, doi: 10.32604/cmc.2018.02356.

[23] L. R. Zuama, D. R. I. M. Setiadi, A. Susanto, S. Santosa, H.-S. Gan, and A. A. Ojugo, "High-Performance Face Spoofing Detection using Feature Fusion of FaceNet and Tuned DenseNet201," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 4, pp. 385–400, Feb. 2025, doi: 10.62411/faith.3048-3719-62.

[24] H. Yu and B. Xu, "Multi-modal texture fusion network for detecting AI-generated images," *Front. Artif. Intell.*, vol. 8, Oct. 2025, doi: 10.3389/frai.2025.1663292.

[25] A. Vaswani *et al.*, "Attention Is All You Need," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Jun. 2017, vol. 30. [Online]. Available: http://arxiv.org/abs/1706.03762

[26] D. Bahdanau, K. Cho, and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, Sep. 2014, pp. 1–15. [Online]. Available: http://arxiv.org/abs/1409.0473

[27] N. Tinago, S. F. Verkijika, and K. Eva Mamabolo, "Deepfakes in Visual Art: Differentiating AI-Generated Art From Human Art Using Convolutional Neural Networks (CNN)," *IEEE Access*, vol. 13, pp. 141484–141495, 2025, doi: 10.1109/ACCESS.2025.3596882.

[28] A. Mahara and N. Rishe, "Methods and Trends in Detecting AI-Generated Images: A Comprehensive Review," *arXiv*. Oct. 17, 2025. [Online]. Available: http://arxiv.org/abs/2502.15176

[29] S. Mavali, J. Ricker, D. Pape, A. Fischer, and L. Schönherr, "Adversarial Robustness of AI-Generated Image Detectors in the Real World," *arXiv*. Jun. 03, 2025. [Online]. Available: http://arxiv.org/abs/2410.01574

[30] L. Nataraj *et al.*, "Detecting GAN generated Fake Images using Co-occurrence Matrices," *arXiv*. Oct. 03, 2019. [Online]. Available: http://arxiv.org/abs/1903.06836

[31] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020, pp. 3204–3213. doi: 10.1109/CVPR42600.2020.00327.

[32] H. H. Nguyen, J. Yamagishi, and I. Echizen, "Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 2307–2311. doi: 10.1109/ICASSP.2019.8682602.

[33] R. Bello-Cerezo, F. Bianconi, F. Di Maria, P. Napoletano, and F. Smeraldi, "Comparative Evaluation of Hand-Crafted Image Descriptors vs. Off-the-Shelf CNN-Based Features for Colour Texture Classification under Ideal and Realistic Conditions," *Appl. Sci.*, vol. 9, no. 4, p. 738, Feb. 2019, doi: 10.3390/app9040738.

[34] L. K. Pavithra and T. S. Sharmila, "An efficient framework for image retrieval using color, texture and edge features," *Comput. Electr. Eng.*, vol. 70, pp. 580–593, Aug. 2018, doi: 10.1016/j.compeleceng.2017.08.030.

[35] J. Ma, X. Jiang, A. Fan, J. Jiang, and J. Yan, "Image Matching from Handcrafted to Deep Features: A Survey," *Int. J. Comput. Vis.*, vol. 129, no. 1, pp. 23–79, Jan. 2021, doi: 10.1007/s11263-020-01359-2.

[36] L. Fei-Fei, J. Deng, and K. Li, "ImageNet: Constructing a large-scale image database," *J. Vis.*, vol. 9, no. 8, pp. 1037–1037, Mar. 2010, doi: 10.1167/9.8.1037.

[37] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015, doi: 10.1007/s11263-015-0816-y.

[38] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 4510–4520. doi: 10.1109/CVPR.2018.00474.

[39] D. Hussain, M. Ismail, I. Hussain, R. Alroobaea, S. Hussain, and S. S. Ullah, "Face Mask Detection Using Deep Convolutional Neural Network and MobileNetV2-Based Transfer Learning," *Wirel. Commun. Mob. Comput.*, vol. 2022, no. 1, Jan. 2022, doi: 10.1155/2022/1536318.

[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, vol. 2016-Decem, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[41] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv*, Jun. 2021, [Online]. Available: http://arxiv.org/abs/2010.11929