What if Retrieval Could Work Before Decoding? The case of JPEG AI Latents for Deepfake Source Attribution

Claudio Vittorio Ragaglia
Department of Mathematics and Computer Science,
University of Catania
Catania, Italy
claudio.ragaglia@phd.unict.it

Lorenzo Catania
Department of Mathematics and Computer Science,
University of Catania
Catania, Italy
lorenzo.catania@phd.unict.it

Francesco Guarnera
Department of Mathematics and Computer Science,
University of Catania
Catania, Italy
francesco.guarnera@unict.it

Dario Allegra
Department of Mathematics and Computer Science,
University of Catania
Catania, Italy
dario.allegra@unict.it

Sebastiano Battiato
Department of Mathematics and Computer Science,
University of Catania
Catania, Italy
sebastiano.battiato@unict.it

Abstract

We explore whether the latent space of the recent JPEG AI compression standard can be employed for high-level semantic tasks. Specifically, we propose a decoding-free approach to image-to-image retrieval and deepfake generator attribution that operates directly on JPEG AI latents, using simple global average pooling and cosine similarity, without any training or learned parameters. Our experiments show that these latent representations retain both semantic content and generator-specific signatures. Using only two latents per image, we achieve consistent mean Top-1 accuracy across eight retrieval classes and high attribution performance in a multi-generator deepfake setting. Compared to traditional RGB-based pipelines, our method eliminates synthesis transformation and color post-processing, yielding substantial efficiency gains. We argue that compressed-domain semantic indexing may play a central role in large-scale generative content monitoring, assuming that appropriate transparency and user consent mechanisms are implemented.

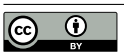
CCS Concepts

• **Computing methodologies** → *Image compression*; **Image representations**; **Visual content-based indexing and retrieval**; • **Information systems** → **Image search**.

Keywords

JPEG AI; latents; learned image compression; deepfake attribution; generative model; image retrieval; content-based image retrieval

ACM Reference Format:
Claudio Vittorio Ragaglia, Lorenzo Catania, Francesco Guarnera, Dario Allegra, and Sebastiano Battiato. 2025. What if Retrieval Could Work Before Decoding? The case of JPEG AI Latents for Deepfake Source Attribution. In *Proceedings of the 1st Deepfake Forensics Workshop: Detection, Attribution, Recognition, and Adversarial Challenges in the Era of AI-Generated Media (DFF '25), October 27–28, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3746265.3759672



DFF '25, Dublin, Ireland

ACM ISBN 979-8-4007-2047-5/2025/10
https://doi.org/10.1145/3746265.3759672

1 Introduction

In domains such as social media moderation, visual search, and digital forensics, modern image analysis pipelines typically begin by decompressing each image into a full-resolution RGB array. This step, although often considered inevitable, imposes substantial computational and storage costs, especially in high-volume media processing scenario. Moreover, it is semantically redundant: the decompressed image itself does not offer immediate insight into its content. What many visual analysis tasks really require is not the pixel grid, but the semantic content embedded in the image. This distinction between data representation and its semantic description raises a fundamental question: is full image reconstruction always necessary for semantic-level tasks? With the increasing volume of media generated and transmitted daily, and the rising need for both real-time processing and privacy-aware systems, reconsidering this assumption is essential for designing more efficient and scalable visual processing systems. The emergence of JPEG AI [2], a learned image compression standard recently introduced by the JPEG committee, presents the opportunity to rethink how we structure visual pipelines. Unlike traditional codecs such as JPEG, which rely on hand-crafted transforms like DCT, JPEG AI uses neural networks to encode images into latent tensors. These latents are then decoded by a learned synthesis network to reconstruct the original RGB image. However, the expressive power of the latent space suggests it may already contain rich, structured information, embedding both high-level semantic features and low-level artifacts indicative of the image's generative origin. This observation motivates a deeper investigation into the role of JPEG AI latents as first-class representations for analysis. If these compressed features retain high-level cues about the image content, then it may be possible to perform meaningful inference directly in the compressed domain. By operating entirely in the latent domain, the pipeline

eliminates any need to recreate pixel level content, thereby mitigating the exposure of sensitive visual information and offering an inherent privacy advantage. This could limit unnecessary decoding and demanding pre-processing, while enabling lighter, faster, and potentially more privacy-preserving systems.

In this work, we introduce a retrieval-based pipeline for deepfake attribution that operates entirely within the compressed latent space of JPEG AI, in contrast to traditional RGB-based approaches that require full image decoding. Starting from the latent tensors produced by the encoder, we apply global average pooling to generate compact descriptors, which are then indexed using Facebook AI Similarity Search (FAISS) [11], an open source library optimized for vector similarity search. Given a query, attribution is performed by retrieving the most similar latent representations and inferring the source model through a majority voting system based on k-NN. This design eliminates the need for decoding, training, or pixel-level feature extraction, resulting in an efficient and scalable solution. Beyond attribution, we also investigate whether JPEG AI latents encode semantic information that supports broader visual understanding. Our experiments provide preliminary evidence in this direction, showing that the same latent representations can be effectively used for content-based image retrieval—suggesting that JPEG AI may serve not only as a compression standard, but also as a compact semantic embedding space for a wide range of visual tasks. The paper is structured as follows: Section 2 presents an overview of the state-of-the-art of topics related to our work; in Section 3 we provide background concepts and formalisms used in our discussion; Section 4 details our proposed feature extraction method that works directly in compressed domain; in Section 5 we describe the experimental settings, while the results are presented and discussed in Section 6; in Section 7 we analyse implications and possibilities for future works, and Section 8 ends the paper.

2 Related Work

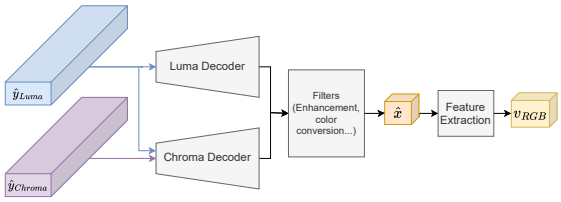
Many image analysis tasks such as classification and retrieval typically operate on raw pixel data. However, most images are stored in compressed formats like JPEG, and decompressing them is a computationally expensive step. This overhead has motivated a growing body of research aimed at performing inference directly within the compressed domain, either by adapting existing codecs or designing new encoders optimized for analysis. Early work by Shen et al. [27] showed that DCT coefficients could be used to extract low-level features for processing tasks, and subsequent studies further demonstrated their effectiveness for high-level vision problems such as semantic segmentation [26]. These approaches share a common motivation with efficient scene representation techniques developed for real-time context classification on platforms [14] with resource constraints, where compact and discriminative descriptors are essential. Subsequent research extended these ideas, adapting neural architectures to operate directly on compressed data for various tasks, including video understanding [31], action recognition [32], image classification [13, 33], and image retrieval [15, 16]. These methods often match or outperform methods that rely on fully decoded images. For example, Edmundson et al. [12] designed a content-based image retrieval system that bypasses full decoding by extracting features directly from the JPEG stream. Gueguen et al. [21] demonstrated that a modified ResNet-50 operating on blockwise DCT coefficients could improve both speed and accuracy. Building on this, Ehrlich and Davis [13] reformulated residual network operations for the JPEG domain while maintaining mathematical equivalence. Similar gains have been reported in video settings [5, 30] by using motion vectors and residuals from compressed streams. The JPEG committee has recently introduced JPEG AI [2], a learned image compression standard that has attracted growing attention. Unlike traditional codecs, JPEG AI relies on neural networks and latent representations. Initial work has demonstrated its practical potential. Alkhateeb et al. [1] proposed a face detection method operating on JPEG AI latents, achieving performance on par with pixel-domain approaches but at lower computational cost. Kovalev et al. [23] presented a methodology for evaluating the adversarial robustness of JPEG AI, comparing its resilience to attacks against other neural codecs. Recently, the implications of JPEG AI are being actively examined in the domain of multimedia forensics. Bergmann et al. [7] introduced three forensic cues to detect JPEG AI compression, identify recompression, and differentiate AI-compressed images from synthetic ones. Cannas et al. [10] studied counter-forensic effects, showing that JPEG AI artifacts can mislead deepfake and splicing detectors. JPEG AI was also employed in the construction of the WILD dataset [8], for one of the post-processing step in the assessment phase. Notably, JPEG AI was designed not only for efficient image compression and transmission, but also with the goal of facilitating downstream computer vision tasks. Its predecessor, JPEG, has been widely explored for such purposes [12, 13, 15], and research has also addressed the psychovisual and statistical optimization of its quantization tables to improve coding efficiency in a way that is consistent with semantic content [6], partly due to the interpretability and structure of its DCT-based compressed domain. In contrast, the learned latent space of JPEG AI is more abstract and less interpretable. Although early work [1] has hinted at its potential for vision tasks, the understanding of how effectively these learned latents support downstream applications is still very limited. In this context, demonstrating that JPEG AI latents are not merely compressed representations of the original images, but also compact carriers of semantic information, could unlock new opportunities for large-scale image processing. This is particularly relevant for scenarios involving massive image datasets, where minimizing storage and computational overhead is critical. A prominent example is Content-Based Image Retrieval (CBIR) [9, 20, 24], which enables retrieval based on visual content rather than textual metadata. A common query modality in CBIR is Image-to-Image (I2I) retrieval, where the goal is to find the most semantically similar images to a given query image. In light of these considerations, we posit that the semantic observed in JPEG AI latent representations could also support complex forensic tasks. Actually, as generative models for multimedia contents proliferate, safeguarding intellectual property rights has become a pressing concern. This has led to significant research in the area of ownership verification, which aims to protect the creators of generative content from unauthorized use or replication. This led researchers to explore the deepfake attribution task, or deepfake model recognition, which focuses on identifying the specific generative system behind a manipulated media [19]. Unlike classic deepfake detection (real/fake), attribution aims to

recognize the type of generator (e.g., GAN or diffusion model) and, where possible, the exact model instance [3]. SOTA methods have demonstrated strong capabilities in associating generated content with its source architecture, highlighting the ability to recognize not just the general structure, but also the subtle artifacts and patterns introduced by the generative process. A notable example is the work by Marra et al. [25], the first to show that GANs leave distinctive, source-specific artifacts (GAN fingerprints) that can be revealed through correlation analysis and exploited for reliable source attribution. Yu et al. [34], subsequently proposed an artificial fingerprinting scheme that actively embeds binary coded marks into training data, causing the generator itself to learn a robust controllable signature that greatly simplifies downstream deepfake attribution. More recently, Guarnera et al. [18] introduced a unified frequency domain framework that distinguishes images produced by GANs and diffusion models, achieving state-of-the-art detection accuracy across a wide spectrum of generators and underscoring the persistence of model specific artifacts in synthetic imagery. Sun et al. [28, 29] tackled the more challenging setting of open-world attribution with the OW-DFA++ benchmark. In this scenario, the attribution system must handle not only known generators, but also unseen ones that were not present during training. OW-DFA++ is designed to evaluate method robustness across diverse forgery types and unknown model instances. To address the complexities of this open-world setup, the authors proposed the Multi-Perspective Sensory Learning framework, which aggregates multiple feature perspectives to improve generalization across source models.

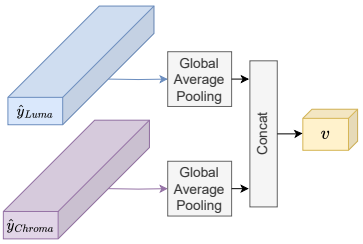
In this work, we investigate the unexplored potential of JPEG AI's compressed latent space for two distinct semantic tasks: Image-to-Image (I2I) retrieval and deepfake generator attribution. Unlike prior studies on JPEG AI that focus on low-level tasks such as face detection [1] or adversarial robustness [23], we aim to assess whether its learned latent representations encode sufficient high-level semantics to support tasks traditionally carried out in the pixel domain. Our approach operates entirely within the compressed domain: we extract the latent tensors used for image reconstruction in JPEG AI, apply global average pooling to obtain fixed-size embeddings, and use them to build a similarity index via FAISS. This pipeline requires no image decoding, no pixel-based feature extraction, and no fine-tuning or training on the downstream tasks. To the best of our knowledge, this is the first attempt to apply JPEG AI latent features directly to both semantic retrieval and forensic attribution. We show that these compressed-domain embeddings enable efficient similarity search across large datasets and can distinguish between synthetic images generated by different deepfake models, all while maintaining high performance and reducing computational overhead.

3 Background

This section provides the technical foundations required to contextualize our study. We first outline the learned compression pipeline of JPEG AI and discuss the structure of its latent space. We then introduce the task of deepfake attribution and motivate why working directly in latent space is advantageous for forensic analysis.



(a) Traditional embedding generation process, based on decoding the compressed latents and performing feature extraction on the decoded RGB image.



(b) Our proposed embedding generation process from compressed JPEG AI latents, avoiding decoding, post-processing and feature extraction steps.

Figure 1: Embedding generation process following the traditional approach in the decoded RGB domain (Figure 1a) and our proposed approach, using the compressed JPEG AI latents with minimal processing (Figure 1b).

3.1 Exploring JPEG AI

The recent JPEG AI standard departs from traditional block-based encoding by proposing a novel end-to-end neural network for learned image compression, which is based on the most performant architectures proposed in the field. The input RGB image $x \in [0,1]^{H\times W\times 3}$ is converted to the YUV BT.709 [22] colorspace before being fed to the network. Luma and chroma channels are encoded by two pipelines sharing the same structure. The luma pipeline takes the Y channel at full resolution, while the chroma pipeline takes both with the Y and UV components downsampled at half resolution. The resulting features are processed through a network g_a, of which the output latents y are quantized to form $\hat{y}$. This quantized $\hat{y}$ tensor is then entropy-coded, estimating a Gaussian distribution and learning a hyper-prior h_a [4]. Decoding is performed with specular pipelines that adopt a *synthesis* network g_s instead, with enhancement filters applied to the intermediate YUV values to reduce visual artifacts. Notably, latents are predicted and only the residual is entropy-coded and transmitted to exploit the spatial redundancy in the transformed domain itself. The result of the decoding process is the reconstructed RGB image $\hat{x}$.

Despite the purpose of JPEG AI is to optimize the rate-distortion ratio of compressed images, the domain of compressed latents $\hat{y}$ exhibits other interesting characteristics, as one of the declared objectives of this standard is to support various tasks other than straightforward picture transmission. In this respect, the channels of $\hat{y}$ exhibit a coarse-to-fine ordering: early channels code global color and luminance, while later ones refine high-frequency texture.

(a) Adobe Firefly (b) Deep AI (c) Hotpot AI (d) NVIDIA SANA (e) StableDiffusion (f) StyleGAN2 (g) StyleGAN3

Figure 2: Sample of images used in the deepfake attribution task. Images from (a) to (e) are generated by diffusion models, while images (f) and (g) by GANs.

For instance, the JPEG AI codec is designed to enable progressive decoding of $\hat{y}$, starting from its first channels to achieve a rough reconstruction of the image. This suggests that different channels capture information at varying levels of abstraction and frequency content. Therefore, we hypothesize that characteristic artifacts or statistical traces left by different generative models could manifest distinctly across these different feature scales. For instance, structural regularities or low-frequency noise might be encoded in early channels, while high-frequency textures or specific noise patterns could reside in later channels. Analyzing specific subsets of channels of $\hat{y}$ could thus isolate these generator-specific signatures, potentially offering robustness against variations captured by other channels or improving efficiency by focusing on the most informative features. Operating directly on the quantized latents $\hat{y}$ offers several advantages: it avoids the g_s decoding cost, yielding a dense, low-dimensional representation (compared to pixels) amenable to fast similarity search. The core intuition of this work is that while optimized for compression, the latent tensors $\hat{y}$ capture rich visual features and potentially residual statistical patterns that can serve as fingerprints for the generating model. In this study we evaluate three configurations of $\hat{y}$ channels: (i) the two most energetic channels, (ii) the top six channels, and (iii) the total number of channels.

3.2 Deepfake Attribution

Deepfake attribution is the task of identifying which generative model $\mathcal{G}_k$ produced a synthetic image x'. The goal of the task is to produce a multi-class label among K candidate generators (for instance StyleGAN2, Stable Diffusion, Adobe Firefly, etc.). Accurate attribution is essential for provenance tracking, regulatory compliance, and assigning responsibility in investigations. State-of-the-art methods typically operate on decoded pixels (spatial domain) or exploit frequency artifacts, camera-row dependencies, or learned classifier fingerprints. These approaches often rely on complex networks, often with millions of parameters, operating on the high-dimensional pixel data. Our hypothesis is that JPEG AI latents already encode generative model specific signatures as inherent statistical traces or micro-patterns. We therefore reformulate attribution as nearest-neighbor retrieval in the $\hat{y}$ space: given query latents $\hat{y}^{(q)}$, we search a FAISS index of reference latents and assign x' the majority generator label among its top-k neighbors. Section 4 details the extraction of used features and Section 6 demonstrates that this retrieval-based strategy achieves consistently high accuracy across a wide range of generative models.

4 Lightweight feature extraction from the compressed JPEG AI domain

Our investigation is based on the hypothesis that the latent feature tensors produced by modern backbones already encode sufficient semantic structure to enable accurate retrieval without any specific finetuning. We therefore pose two research questions:

- Can purely tensor-level representations support effective retrieval across heterogeneous image domains?
- Can the same representations be leveraged for the forensic task of deepfake attribution?

Answering the first question offers a controlled way to test how well the chosen features represent the data. If successful, this motivates the second step, a more challenging evaluation to see if those same features can capture subtle artifacts that distinguish different types of generated content. The starting point of our study is the observation that modern codecs such as JPEG AI expose intermediate tensors that remain largely unexplored by the computer vision community. The conventional workflow, reproduced in Figure 1a, shows the pipeline to decode these tensors back to the RGB domain, applies color conversion and other post–processing operations. We suppose that the original latents already encode discriminative information that can be exploited directly. Our proposed pipeline, shown in Figure 1b, therefore bypasses the expensive synthesis stage: the entropy-decoded Y and UV latents are aggregated with a global average pooling operator and transformed into a compact descriptor without any further supervision. The idea is to operate directly on compressed images, relying on the latent representations produced during the decoding stage of a reference JPEG AI codec. Rather than using reconstructed pixel data, it leverages internal latent features generated by the synthesis network g_s, including predicted latents with associated information, plus the entropy-decoded latent residuals. These features serve as the basis for subsequent tasks. In the initial retrieval experiment, three distinct subsets of these latent features are evaluated:

(1) **`2lat`**: the pair $\{\hat{y}_{Luma}, \hat{y}_{Chroma}\}$, i.e. the latents of the luminance and chrominance branches;
(2) **`selected`**: the previous two plus the residual and quantised–residual latents for both branches, yielding six tensors in total;
(3) **`all`**: every tensor exported by the reference JPEG AI decoder.

For each chosen tensor $T \in \mathbb{R}^{C\times H\times W}$ a single vector $v = \mathrm{GAP}(T) \in \mathbb{R}^C$ has been computed via global average pooling over the spatial dimensions. Vectors are concatenated in a fixed, deterministic order

and cast to 32-bit float, producing an embedding $f(x) \in \mathbb{R}^d$ whose dimensionality d depends on the configuration and on the codec hyper-parameters. No spatial information is retained, reflecting our focus on instance retrieval rather than localisation.

5 Experimental settings

To prepare the datasets, composed of lossless compressed PNG images, for our experiments, we compress each picture with the reference JPEG AI encoder [1]. We have adopted a quality factor of 65, which produces bitstreams in a range of $0.6 - 0.7$bpp, plausible in practical settings. After encoding an image, we extract the features given to the synthesis network as described in Section 4. This set of tensors is then serialized in `.pt` files and used in our experiments. The result of this process for each image is the compressed bitstream together with a directory of PyTorch tensors that capture the complete latent state of the encoder; no decoded pixels are ever used downstream.

Experimental design for image retrieval. To probe the discriminative power of the latent descriptors, we construct three scenarios of increasing difficulty. For these experiments, the images are collected from Caltech Dataset [17]; in particular, 16 classes are sampled: ak47, american-flag, .backpack, baseball-bat, baseball-glove, basketball-hoop, bat, bathtub, bear, beer-mug, billiards, binoculars, birdbath, bonsai-101, bowling-ball, and cake. Embeddings are z-scored using a *StandardScaler*, subsequently ℓ_2-normalised. Similarity between two images is then expressed as the inner product of their unit-norm vectors (i.e. cosine similarity). We used FAISS for large scale nearest-neighbour retrieval. Unless otherwise stated, the experiments employ the exact index IndexFlatIP, which stores the entire database in RAM and performs a scan optimised with SIMD (Single Instruction, Multiple Data). The index does not require any training and guarantees reproducible results across platforms. For every subset we index all images belonging to the selected classes in a single FAISS structure and then issue queries using the same pool, discarding the trivial self-match that FAISS returns. The procedure is repeated for each of the three feature configurations (`2lat`, `selected` and `all`) described in Section 4, yielding a total of 1080 retrieval runs.

From retrieval to deepfake attribution. The second experiment assesses whether the same latent descriptors can support provenance analysis of deepfake images. Here we restrict our attention to the feature `2lat` (two latents $\hat{y}_{Luma}$ and $\hat{y}_{Chroma}$ concatenated into a single vector per image) due to the results of first retrieval experiments demonstrating how `2lat` reach same results with less latent (described in Section 6.1). A FAISS index is constructed containing every compressed image, regardless of its origin. The dataset used in this experiment contains random sampled images from WILD dataset [8], in particular based on seven generators: five diffusion models (Adobe Firefly, Deep AI, Hotpot AI, NVIDIA SANA, and Stable Diffusion 3.5) and two GANs (StyleGAN2 and StyleGAN3). Each generator contributes 450 samples. Figure 2 presents one image for generator; it is indicative how visually are the images of different model. Retrieval is then performed under two complementary attribution granularities: **Engine level**, where generators are grouped into GANs and diffusion models (DM) classes and the task is to discover the dominant macro categories among the nearest neighbours of a query; **Model Generator level**, where each individual model constitutes its own class, enabling a finer attribution when the retrieval set contains all GAN or diffusion variants. Thus, at the engine level, the index therefore holds 2 macro classes (diffusion models and GANs). At the generator level, it preserves the original 7 classes (one per generative model). These class labels are used only for direct evaluation; the retrieval itself remains fully unsupervised. For both settings every image serves in turn as a query against the complete index; the evaluation reflects a pure nearest-neighbour scenario based on the latent space.

[1] https://gitlab.com/wg1/jpeg-ai/jpeg-ai-reference-software

6 Results

Throughout this section we report results obtained from our experiments. They are based on four retrieval metrics that offer complementary perspectives on performance. **Top-1 Accuracy** measures the accuracy that the first neighbour retrieved by the index belongs to the same class as the query. **Precision@10** quantifies the fraction of correct items within the ten retrieved, thus rewarding ranks that contain multiple true matches. Finally, **mean Average Precision** (mAP) computes the area under the precision–recall curve on a per-query basis before averaging, integrating both rank and abundance of true positives. For the deepfake attribution experiment we provide an additional **Top-1 Type Accuracy**, which integrates the criterion to the macro-class level (GAN vs. diffusion model) and is therefore insensitive to intra-family confusions.

6.1 Image retrieval based on JPEG AI latent

Figure 3 illustrates a semantic retrieval case (4 classes scenario), starting from a backpack query image, the FAISS index returns 7/10 semantic correct images, belonging to the same class. Table 1 presents retrieval performance metrics aggregated by feature configuration and number of classes. The minimalist `2lat` configuration, based solely on the two latent tensors $\hat{y}_{Luma}$ and $\hat{y}_{Chroma}$, achieves a mean Top-1 accuracy of 0.743 in the simplest scenario with two classes, which decreases to 0.542 and 0.384 for scenarios with four and eight classes, respectively. Precision@10 similarly decreases as the class count grows, from 0.713 (2 classes) to 0.498 (8 classes). The mAP metric shows the same trend, ranging from 0.793 to 0.649, indicating consistently good retrieval performance despite increased complexity. Comparing feature configurations, the `all` configuration yields slightly better performance than `2lat` when dealing with more classes (8-class Top-1 accuracy of 0.416 vs. 0.384), suggesting that additional latent tensors can provide beneficial discriminative information in more challenging scenarios. The intermediate `selected` feature set generally underperforms compared to the `2lat` and `all` configurations across all class numbers, confirming that a selective combination of tensors does not necessarily enhance retrieval quality and may introduce noise. Overall, these findings highlight JPEG AI latent tensors robustness and effectiveness for semantic-level instance retrieval, supporting their utility in subsequent applications such as deepfake attribution. Based on this hint, in the next section we present a case study where our retrieval pipeline based on JPEG AI latents achieved promising results in deepfake attribution tasks.

(a) Query

(b) ✓ (c) ✗ (d) ✗ (e) ✓ (f) ✓

(g) ✗ (h) ✓ (i) ✓ (j) ✓ (k) ✓

Figure 3: Example of latent space image retrieval based on semantic. The query depicts a backpack (top). Among the ten nearest neighbours returned by the latent search, green ✓ denote the same semantic class, while red ✗ identify mismatches. Here 7 / 10 neighbours belong to the correct class.

Table 1: Instance retrieval - Performance metrics aggregated by feature set and number of classes.

Feature	Classes	Top-1 Acc.	Prec.@10	mAP
2lat	2	0.743	0.713	0.793
2lat	4	0.542	0.498	0.646
2lat	8	0.384	0.341	0.549
all	2	0.763	0.714	0.803
all	4	0.566	0.500	0.662
all	8	0.416	0.347	0.577
selected	2	0.711	0.676	0.765
selected	4	0.482	0.441	0.600
selected	8	0.307	0.276	0.492

Each entry corresponds to 120 experimental runs queries.

6.2 Retrieval for attribution of deepfake images

The main study of this work evaluates the implementation of `2lat` latent descriptors feature configuration, seen in the previous experiments, applied in a challenging deepfake source attribution task. As an illustrative example, Figure 4 displays a Deep AI query image alongside the ten nearest neighbours returned by our FAISS search, ranked left-to-right and top-to-bottom: Deep AI, Deep AI, NVIDIA SANA, Stable Diffusion, Deep AI, NVIDIA SANA, NVIDIA SANA, NVIDIA SANA, Stable Diffusion and Deep AI. Thus, in this example all ten neighbours share the correct generative engine (diffusion model), and four of them (including the first and second, in order the closest ones) also match the exact generator Deep AI. Table 2 summarises the results at the macro-class level, comparing diffusion models (DM) and GAN-based methods. Diffusion models demonstrate slightly superior Top-1 accuracy (0.769) compared to GANs (0.692). Both categories achieve comparable Precision@10 (0.682 for DM and 0.659 for GAN) and mean average precision (mAP) scores (0.779 DM, 0.744 GAN). Notably, GAN models achieve an exceptional Top-1 Type accuracy of 0.991, suggesting that while individual GAN generators may occasionally be misclassified at the model level, they are consistently correctly identified at the broader class level. Detailed per model results in Table 3 reveal specific strengths and limitations. NVIDIA SANA obtains the highest Top-1 accuracy among diffusion models (0.858), closely followed by Adobe Firefly (0.808) and Hotpot AI (0.804). Among GAN-based models, StyleGAN2 exhibits strong performance with a Top-1 accuracy of 0.818, whereas StyleGAN3 notably underperforms at 0.438, despite maintaining a high Top-1 Type accuracy (0.982). This discrepancy suggests substantial latent feature overlap between GAN models, particularly evident between StyleGAN2 and StyleGAN3. Despite not being explicitly optimised for attribution, the retrieval framework robustly captures generator-specific fingerprints, underscoring the effectiveness of JPEG AI latent tensors for deepfake provenance analysis. In summary, both experiments corroborate the central hypothesis of the paper: compressed JPEG AI latents, when subjected to minimal processing, encode semantic and forensic cues that can be harnessed for image retrieval and deepfake attribution without any additional learning.

Table 2: Deepfake Attribution - Performance metrics aggregated by generative model type.

Type	Top-1 Acc.	Prec.@10	mAP	Top-1 Type Acc.
dm	0.769	0.682	0.779	0.927
gan	0.692	0.659	0.744	0.991

Table 3: Deepfake Attribution - Performance metrics per generative model.

Model	Top-1 Acc.	Prec.@10	mAP	Top-1 Type Acc.
Adobe Firefly	0.808	0.735	0.814	0.890
Deep AI	0.587	0.473	0.622	0.907
Hotpot AI	0.804	0.727	0.816	0.971
NVIDIA SANA	0.858	0.782	0.857	0.991
Stable Diffusion 3.5	0.749	0.638	0.748	0.911
StyleGAN2	0.818	0.812	0.857	0.996
StyleGAN3	0.438	0.349	0.515	0.982

7 Discussion

Our experiments show that performing complete decoding of an image is not a prerequisite for either instance retrieval or deepfake forensic attribution. Entropy decoding a JPEG AI bitstream

Figure 4: Example of latent space image retrieval. The top image is the Deep AI query; below, the ten nearest neighbours are shown in rank order. Each neighbour is annotated with two markers: the left-hand marker refers to the *generation family* (GAN vs. diffusion), whereas the right-hand marker tests the exact generator instance. For both markers a green ✓ indicates a correct prediction and a red ✗ denotes an error.

already yields latent tensors that, after a trivial global average pooling, retain enough discriminatory power to achieve a consistent mean Top-1 accuracy in eight–class retrieval using only two latents, while reach high Top-1 and macro–type accuracy in deepfake generator attribution. These results are obtained *without any learned parameters*, relying only on cosine similarity in FAISS, confirming our hypothesis that the learned analysis transform g_a embeds generator-specific traces alongside semantic content. Compared to pixel-domain pipelines, our approach removes two costly blocks: synthesis transform g_s and color space post-processing. From a forensic point of view, this opens a lot of useful implementations; just for instance, newsrooms and social platforms could flag latent neighbours of a suspicious upload before deeper verification, or since latents exist prior to decryption, a client-side software could query a public index without ever revealing the reconstructed image. In general, compressed-domain indexing could enhance privacy by avoiding pixel disclosure, and it also lowers the barrier for large-scale tracking of generative usage.

7.1 Future Work

The present study establishes a robust proof-of-concept, demonstrating that latent representations are rich in both forensic and semantic information. Based on these initial findings, we outline a structured roadmap for future work, progressing from immediate validation to more ambitious research directions. The next step could be to go deeper into the understanding of the current results. From a forensic perspective, this involves conducting targeted attribution experiments on generative models from the same architectural family to test the sensitivity of latent fingerprints to subtle variations. Another fundamental step is to discover these fingerprints and provide interpretable evidence of what constitutes a generative signature. From a semantic point of view, a rigorous quantitative analysis is needed to confirm the obtained preliminar results. We plan to extend our evaluation to include a larger and more diverse set of deepfake generators, increasing the number of classes significantly beyond those considered in this study. This will help assess the scalability and robustness of our latent space retrieval approach in more challenging multiclass attribution scenarios. Additionally, we intend to evaluate the proposed methodology on alternative deepfake datasets, such as newly emerging generators not present in the current WILD dataset. This will allow us to investigate the generalization capabilities of the retrieval system when faced with unseen forgery sources. Subsequently, the focus can shift to enhance the representation's robustness and utility. For forensic applications, we must evaluate the resilience of the latent fingerprints against a standardized suite of common post-processing operations, such as recompression and filtering. Some alternative to global average pooling should be explored to capture localized forensic artifacts that could be currently averaged out. In the semantic domain, the application of metric learning objectives could refine the feature space, improving the separation of closely related classes. A critical contribution would be to analyze a novel "Rate-Semantics", conceptualizing the trade-off between compression efficiency and the preservation of semantic info. For semantic analysis, future work should investigate the latent space's ability to capture hierarchical relationships between concepts. A final, transformative direction is to explore cross-modal retrieval, examining whether the semantic space can be aligned with other modalities like text, bridging the gap between image compression and large-scale, multimodal search.

8 Conclusion

We have demonstrated that the latent space generated by JPEG AI codec could preserve important info for semantic analysis and pattern for deepfake attribution. These insights elevate compressed domain forensics from a niche optimisation to a promising research frontier. Treating latents as important forensic evidence opens multiple directions: cross-codec generalisation, adversarially robust latent descriptors, web-scale indexing, and privacy-preserving on-device provenance services. We could imagine a future in which methods based on latent study complement traditional pixel-level analysis, in a trustworthy and efficient media ecosystems.

Acknowledgments

The work of Claudio Vittorio Ragaglia has been supported by the Spoke 1 "Future HPC BigData" of the Italian Research Center on High-Performance Computing, Big Data and Quantum Computing (ICSC) funded by MUR Missione 4 Componente 2 Investimento 1.4: Potenziamento strutture di ricerca e creazione di "campioni nazionali di R&S (M4C2-19)" - Next Generation EU (NGEU).

The work of Francesco Guarnera has been supported by MUR in the framework of PNRR PE0000013, under project "Future Artificial Intelligence Research – FAIR".

References

[1] Ayman Alkhateeb, Alessandro Gnutti, Fabrizio Guerrini, Riccardo Leonardi, João Ascenso, and Fernando Pereira. 2024. JPEG AI Compressed Domain Face Detection. In *2024 IEEE 26th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 1–6.

[2] João Ascenso, Elena Alshina, and Touradj Ebrahimi. 2023. The jpeg ai standard: Providing efficient human and machine visual data consumption. *Ieee Multimedia* 30, 1 (2023), 100–111.

[3] Vishal Asnani, Xi Yin, Tal Hassner, and Xiaoming Liu. 2023. Reverse engineering of generative models: Inferring model hyperparameters from generated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 12 (2023), 15477–15493.

[4] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. 2018. Variational image compression with a scale hyperprior. *arXiv preprint arXiv:1802.01436* (2018).

[5] Barak Battash, Haim Barad, Amit Bleiweiss, and Hanlin Tang. 2020. Mimic The Raw Domain: Accelerating Action Recognition in the Compressed Domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 744–753.

[6] Sebastiano Battiato, Massimo Mancuso, Angelo Bosco, and Mirko Guarnera. 2001. Psychovisual and statistical optimization of quantization tables for DCT compression engines. In *Proceedings 11th International Conference on Image Analysis and Processing*. IEEE, 602–606.

[7] Sandra Bergmann, Fabian Brand, and Christian Riess. 2025. Three Forensic Cues for JPEG AI Images. *arXiv preprint arXiv:2504.03191* (2025).

[8] Pietro Bongini, Sara Mandelli, Andrea Montibeller, Mirko Casu, Orazio Pontorno, Claudio Vittorio Ragaglia, Luca Zanchetta, Mattia Aquilina, Taiba Majid Wani, Luca Guarnera, et al. 2025. WILD: a new in-the-Wild Image Linkage Dataset for synthetic image attribution. *arXiv preprint arXiv:2504.19595* (2025).

[9] Akshara Preethy Byju, Begüm Demir, and Lorenzo Bruzzone. 2020. A progressive content-based image retrieval in JPEG 2000 compressed remote sensing archives. *IEEE Transactions on Geoscience and Remote Sensing* 58, 8 (2020), 5739–5751.

[10] Edoardo Daniele Cannas, Sara Mandelli, Nataša Popović, Ayman Alkhateeb, Alessandro Gnutti, Paolo Bestagini, and Stefano Tubaro. 2025. Is JPEG AI going to change image forensics? *arXiv preprint arXiv:2412.03261* (2025).

[11] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The Faiss library. (2024). arXiv:2401.08281 [cs.LG]

[12] David Edmundson and Gerald Schaefer. 2012. An Overview and Evaluation of JPEG Compressed Domain Retrieval Techniques. In *54th International Symposium ELMAR-2012*. IEEE, 75–78.

[13] Max Ehrlich and Larry S Davis. 2019. Deep residual learning in the jpeg transform domain. In *Proceedings of the IEEE/CVF international conference on computer vision*. 3484–3493.

[14] Giovanni Maria Farinella, Daniele Ravì, Valeria Tomaselli, Mirko Guarnera, and Sebastiano Battiato. 2015. Representing scenes for real-time context classification on mobile devices. *Pattern Recognition* 48, 4 (2015), 1086–1100.

[15] Guocan Feng and Jianmin Jiang. 2002. JPEG image retrieval based on features from DCT domain. In *International conference on image and video retrieval*. Springer, 120–128.

[16] Wei-bin Fu, Jingbing Li, Mengxing Huang, and Yicheng Li. 2015. The image retrieval based on transform domain. In *2015 International Conference on Mechatronics, Electronic, Industrial and Control Engineering (MEIC-15)*. Atlantis Press, 431–435.

[17] Gregory Griffin, Alex Holub, Pietro Perona, et al. 2007. *Caltech-256 object category dataset*. Technical Report. Technical Report 7694, California Institute of Technology Pasadena.

[18] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. 2024. Mastering deepfake detection: A cutting-edge approach to distinguish gan and diffusion-model images. *ACM Transactions on Multimedia Computing, Communications and Applications* 20, 11 (2024), 1–24.

[19] Luca Guarnera, Oliver Giudice, Matthias Nießner, and Sebastiano Battiato. 2022. On the exploitation of deepfake model recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 61–70.

[20] Venkat N Gudivada and Vijay V Raghavan. 1995. Content based image retrieval systems. *Computer* 28, 9 (1995), 18–22.

[21] Lionel Gueguen, Alex Sergeev, Ben Kadlec, Rosanne Liu, and Jason Yosinski. 2018. Faster neural networks straight from jpeg. *Advances in Neural Information Processing Systems* 31 (2018).

[22] ITU-T. 2015. *Parameter values for the HDTV standards for production and international programme exchange*. Technical Report BT.709-6. Recommendation.

[23] Egor Kovalev, Georgii Bychkov, Khaled Abud, Aleksandr Gushchin, Anna Chistyakova, Sergey Lavrushkin, Dmitriy Vatolin, and Anastasia Antsiferova. 2024. Exploring adversarial robustness of JPEG AI: methodology, comparison and new methods. *arXiv preprint arXiv:2411.11795* (2024).

[24] Zhe-Ming Lu, Su-Zhi Li, and Hans Burkhardt. 2006. A content-based image retrieval scheme in JPEG compressed domain. *International Journal of Innovative Computing, Information and Control* 2, 4 (2006), 831–839.

[25] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. 2019. Do gans leave artificial fingerprints?. In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*. IEEE, 506–511.

[26] Daniele Ravì, Miroslaw Bober, Giovanni Maria Farinella, Mirko Guarnera, and Sebastiano Battiato. 2016. Semantic segmentation of images exploiting DCT based features and random forest. *Pattern Recognition* 52 (2016), 260–273.

[27] Bo Shen and Ishwar K Sethi. 1996. Direct feature extraction from compressed images. In *Storage and Retrieval for Image and Video Databases IV*, Vol. 2670. SPIE, 404–414.

[28] Zhimin Sun, Shen Chen, Taiping Yao, Ran Yi, Shouhong Ding, and Lizhuang Ma. 2025. Rethinking open-world deepfake attribution with multi-perspective sensory learning. *International Journal of Computer Vision* 133, 2 (2025), 628–651.

[29] Zhimin Sun, Shen Chen, Taiping Yao, Bangjie Yin, Ran Yi, Shouhong Ding, and Lizhuang Ma. 2023. Contrastive pseudo learning for open-world deepfake attribution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 20882–20892.

[30] Chengyu Wang, Jing Li, Saurabh Kumar, Seok-Jun Lee, and Hamid R. Sheikh. 2022. Compressed Domain Multiframe Processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 1954–1963.

[31] Olivia Wiles, Joao Carreira, Iain Barr, Andrew Zisserman, and Mateusz Malinowski. 2022. Compressed Vision for Efficient Video Understanding. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*. 4581–4597.

[32] Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R Manmatha, Alexander J Smola, and Philipp Krähenbühl. 2018. Compressed video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6026–6035.

[33] Yibo Xu, Weidi Liu, and Kevin F Kelly. 2020. Compressed domain image classification using a dynamic-rate neural network. *IEEE Access* 8 (2020), 217711–217722.

[34] Ning Yu, Vladislav Skripniuk, Sahar Abdelnabi, and Mario Fritz. 2021. Artificial fingerprinting for generative models: Rooting deepfake attribution in training data. In *Proceedings of the IEEE/CVF International conference on computer vision*. 14448–14457.