Learned Image Compression for Deepfake Detection

Alessandro Gnutti*, Ayman Alkhateeb*, Chia-Hao Kao*, Edoardo Daniele Cannas†,
Sara Mandelli†, Wen-Hsiao Peng‡, Riccardo Leonardi*§ and Pierangelo Migliorati*
*Department of Information Engineering, University of Brescia, Italy
†Department of Electronics, Information, and Bioengineering, Politecnico di Milano, Italy
‡Department of Computer Science, National Yang Ming Chiao Tung University, Taiwan
§Department of Electronics Engineering, University of Rome Tor Vergata, Italy

***Abstract*—Learned image compression introduces new challenges for image forensics, as deepfake detectors may mistakenly classify authentic images as artificially generated due to compression artifacts or misidentify synthetic images that have undergone neural compression as real. In this paper, the first neural image compression system specifically designed for deepfake detection is presented. To achieve this, spatial-frequency modulation adapters are integrated into an existing image compression architecture, eliminating the need to retrain the underlying codec. Images decoded with the proposed method achieve detection performance comparable to uncompressed images while also exhibiting superior perceptual quality than images reconstructed from conventional image compression.**

***Index Terms*—Learned image compression, coding for machines, deepfake detection.**

I. Introduction

Image compression has been an active research field for decades, driven by the need for efficient storage and transmission across a wide range of applications that support modern digital life. The recent rise of deep learning has sparked a new wave of advancements in this domain, with end-to-end learned compression systems gaining significant attention. Unlike traditional approaches, these methods optimize the entire compression pipeline holistically. Notably, some of the latest works [1]–[5] have surpassed VVC, which is the most advanced conventional image and video coding standard, on key metrics such as Peak Signal-to-Noise Ratio (PSNR). These results highlight the fact that learned image compression will impact the next generation coding techniques [6].

However, the emergence of Learned Image Compression (LIC) presents new challenges in image forensics [7], [8]. Image forensics refers to the field of study dedicated to analyzing digital images to determine their authenticity, integrity, and origin. Given the increasing prevalence of AI-generated media, it plays a crucial role in combating misinformation, digital fraud, and other malicious uses of manipulated imagery. Recently, it has been proved that LIC methods create specific footprints that require dedicated analysis to understand and leverage [9]. Furthermore, studies suggest that LIC techniques leave characteristic upsampling artifacts, resembling those observed with neural network generative models [8], [10]. These factors pose a critical concern: the risk of forensic detectors mistakenly identifying authentic content as artificially generated due to LIC traces [8]. More broadly, there is also the potential for misclassification of synthetic images that have undergone neural image compression, further complicating forensic analysis (see Fig. 1).

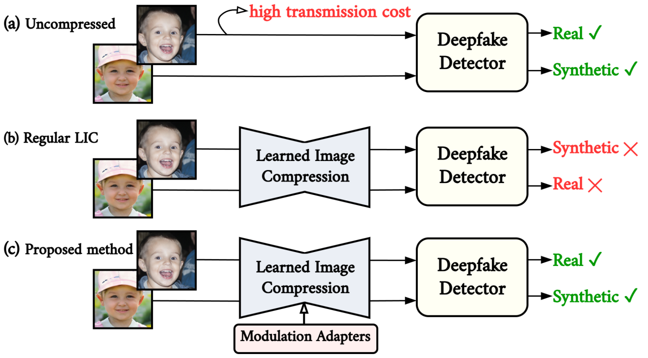
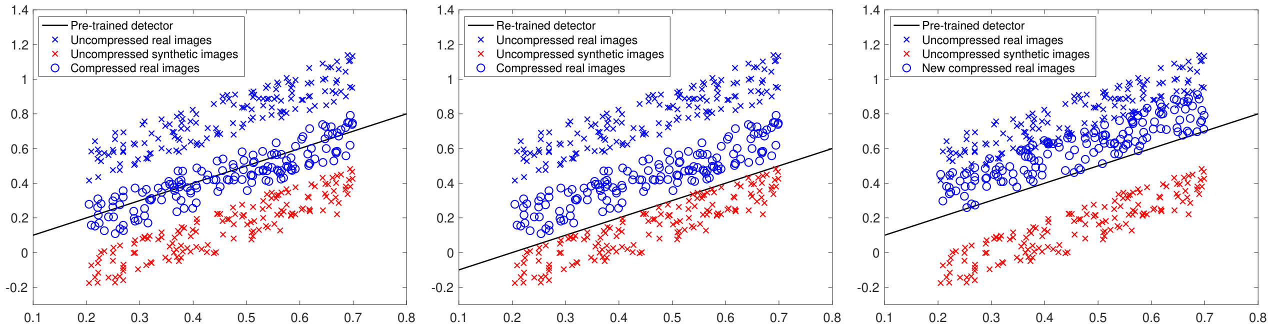


Fig. 1. High-level framework comparison between no compression, using original LIC, and the proposed method.

The problem can be illustrated through the simplified visualization in Fig. 2(a). In this representation, real and synthetic images are depicted as two distinct distributions, with blue crosses denoting real images and red crosses representing synthetic ones. Ideally, a well-trained real vs. synthetic image detector, which is represented by a black decision boundary, should effectively separate these two clusters. However, when images are compressed using a specific LIC method, their pixel distribution shifts (blue circles). As a consequence, the detector may struggle to distinguish between them accurately, leading to potential misclassification.

A possible approach to solve this issue is to retrain the detector to correctly identify compressed real images as real rather than misclassifying them as synthetic. This involves adjusting the decision boundary so that real images, even after compression, remain within the real image distribution, as depicted in Fig. 2(b). However, retraining an entire detector, or more generally a recognition model, to incorporate new data may not always be the most optimal solution. This approach can be computationally expensive due to the complexity of the training process, requiring significant time and resources. Additionally, retraining a model each time new data is intro-

This work was partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on “Telecommunications of the Future” (PE00000001 - program “RESTART”), Projects “S7 FUN-Media” and “FRAME”.



(a) Problem formulation: a pre-trained detector may misclassify the compressed real images.

(b) First approach: re-train the detector.

(c) Second approach: guide the compression process, preserving deepfake detection capabilities (ICM).

Fig. 2. A simplified visualization of the clustering process for a trained real vs. synthetic image detector, represented by a black decision boundary. While uncompressed real images (blue crosses) and synthetic images (red crosses) are well-separated, compressed real images (blue circles) may be misclassified. For clarity, only the distribution of compressed real images are shown. The same analysis applies to compressed synthetic images as well.

duced is inefficient and impractical. Furthermore, modifying the model in this way carries the risk of altering its overall performance, potentially degrading its accuracy on previously learned data or introducing unintended biases.

An alternative recent approach to addressing this issue is to modify the compression model rather than retrain the detector. Instead of adjusting the decision boundary of the recognition model, we can guide the training of the compression algorithm to ensure that the compressed images remain within the distribution of real images, as shown in Fig. 2(c). This strategy, known as Image Coding for Machines (ICM), focuses on optimizing the compression process to produce images that are more suitable for specific downstream tasks. In the proposed scenario, the goal would be to tailor the compression model to preserve the forensic characteristics necessary for accurate deepfake detection. However, while this approach avoids modifying the detector, it still requires retraining the entire compression model, which is still computationally inefficient.

To address these challenges, in this paper, we propose to integrate the Spatial-Frequency Modulation Adapter (SFMA) proposed in [11] into an existing neural image compression system, tailoring it specifically for the task of deepfake detection. SFMA removes non-semantic redundancy using a spatial modulation adapter while enhancing task-relevant frequency components and suppressing task-irrelevant ones through a frequency modulation adapter. The goal is to generate compressed images that are correctly classified by a deepfake detection system without modifying or retraining either the base codec or the detector. Moreover, it will experimentally be shown that it benefits the perceptual quality of reconstruction for real content.

The main contributions of this paper are:

- The first neural image compression system specifically designed for deepfake detection is introduced.
- The proposed model integrates the SFMA, training it exclusively while keeping the main image codec unchanged.
- Experimental results indicate that images decoded using the proposed method achieve performance levels close to those of uncompressed images and comparable to the fully fine-tuned approach.
- Experimental results also demonstrate that decoded images including SFMA achieve superior perceptual quality compared to reconstructed images of the original LIC optimized for PSNR.

II. Preliminaries

A. Learned image compression

An end-to-end learned image compression system typically consists of two primary components: the main autoencoder and the hyperprior autoencoder. The main autoencoder includes an analysis transform g_a and a synthesis transform g_s. The analysis transform g_a encodes an RGB image $x \in \mathbb{R}^{H \times W \times 3}$, with height H and width W, into a smaller latent representation $y \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times C_y}$ using an encoding distribution $q_{g_a}(y|x)$. The latent y is then uniformly quantized as $\hat{y}$ and entropy encoded into a bitstream using a learned prior distribution $p(\hat{y})$. On the decoder side, $\hat{y}$ is entropy decoded and reconstructed as $\hat{x} \in \mathbb{R}^{H \times W \times 3}$ through a decoding distribution $q_{g_s}(\hat{x}|\hat{y})$, implemented by the synthesis transform g_s. During this process, the prior distribution $p(\hat{y})$ significantly influences the number of bits required to signal the quantized latent $\hat{y}$. To address this, it is typically modeled in a content-adaptive manner by a hyperprior autoencoder [12], which consists of a hyperprior analysis transform h_a and a hyperprior synthesis transform h_s. Specifically, h_a transforms the image latent y into an even smaller side information $z \in \mathbb{R}^{\frac{H}{64} \times \frac{W}{64} \times C_z}$, which statistically models a co-located portion of the compressed bitstream. The quantized version of z is decoded from the bitstream through h_s, resulting in $p(\hat{y})$. In this work, we adopt the image compression model proposed in [13] as a reference.

B. Synthetic image detection

The rapid advancement of AI has made it possible to generate highly realistic synthetic images and videos, commonly known as deepfakes [14]. In particular, deep learning techniques can create synthetic images with unprecedented

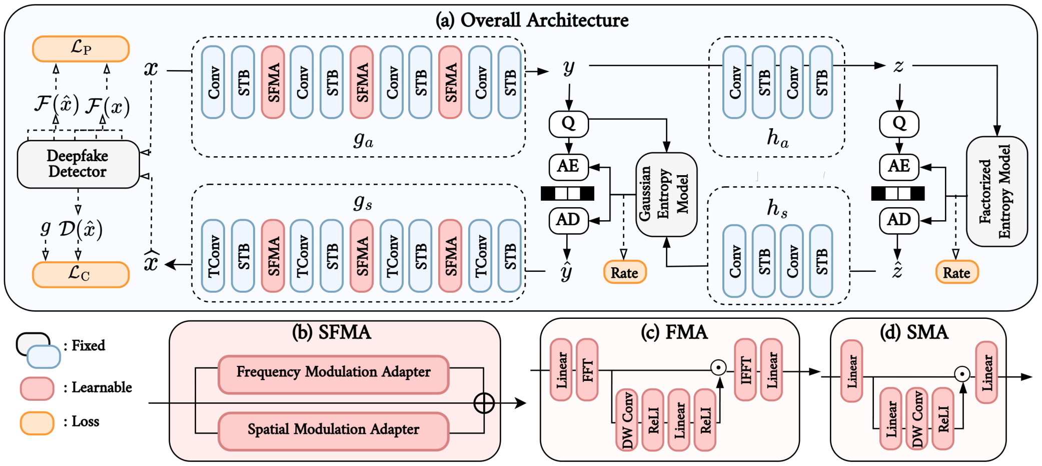


Fig. 3. The architecture of the proposed compression system for deepfake detection and the introduced adapters.

realism, posing a serious threat to the trustworthiness and integrity of multimedia content. To counter the spread of such manipulated content, the multimedia forensics community has placed significant focus on developing effective detection tools and techniques [15]–[17].

In this work, the forensic detector recently introduced in [16] is considered, which has demonstrated outstanding performance in distinguishing real from synthetic images depicting different semantic contents. This detector operates by extracting small square patches from the input image and aggregating their features to produce a single detection score per image. In specific, the detector randomly extracts $N = 800$ color patches $\{P_i\}_1^N$ with size 96×96 pixels, regardless of the input image size. Every patch is processed through a backbone Convolutional Neural Network (CNN) [18], which assigns a detection score s_i, where $s_i > 0$ indicates that the patch is classified as synthetic and real otherwise. Finally, the scores are aggregated by selecting the M highest values, corresponding to the uppermost 75%, and computing their arithmetic mean. The final image score is defined as $s_{\mathbf{I}} = \sum_i^M s_i/M$.

III. Proposed Method

A. System Overview

In this paper, a novel learned image compression system specifically designed for deepfake detection is proposed. Conventional pre-trained deepfake detectors are prone to suffer significant drop in performance when encountering compressed images, particularly at lower bit-rates. To address the issue, the Spatial-Frequency Modulation Adapter [11] is integrated into an off-the-shelf learned image compression system to modulate the reconstructed image signals better suited for the downstream detector instead of reconstruction fidelity. This approach ensures that the reconstructed images remain within the decision bounds for improved classification accuracy. Moreover, the flexibility of the base codec is preserved, as the SFMA modules can be selectively removed to restore pristine reconstruction quality suited for minimal distortion upon reconstruction.

B. Adaptation for Deepfake Detection

Fig. 3 illustrates the overall system architecture, where the base codec is derived from [13]. Unlike the approaches that require re-training an entire codec, lightweight modulation adapters are included into both the encoder g_a and decoder g_s. Specifically, three separate SFMAs are inserted between each Swin Transformer Block (STB) and convolutional layer. The SFMA has two main components: Spatial Modulation Adapter (SMA) and Frequency Modulation Adapter (FMA). These components adjust the image representation in the spatial and frequency domains, respectively in parallel. In each SFMA, the input features from the base codec are processed separately by SMA and FMA, and their outputs are combined via element-wise summation with the original features (Fig. 3 (b)).

The Frequency Modulation Adapter modifies the feature representations in the frequency domain. This is achieved by first applying Fast Fourier Transform (FFT) to extract frequency components. The transformed features are then modulated though element-wise multiplication with the matrix obtained through a lightweight processing pipeline comprising a convolutional layer, a linear layer, and ReLU activations. Finally, an inverse FFT is applied to obtain resulting features (Fig. 3 (c)). On the other hand, the Spatial Modulation Adapter operates directly in the spatial domain without performing FFT/IFFT. It applies a similar transformation mechanism as FMA, using convolutional layers. (Fig. 3 (d)). By jointly modulating both spatial and frequency information, the proposed approach effectively aligns reconstructed images with the expectations of the target deepfake detector.

C. Training Objective

The proposed method follows the standard training paradigm for learned image compression, utilizing a rate-distortion (RD) optimization objective. In symbols, we have

$$\mathcal{L} = \underbrace{-\log p(\hat{z}) - \log p(\hat{y}|\hat{z})}_{\text{R}} + \lambda \cdot \underbrace{(\mathcal{L}_{\text{C}}(\mathcal{D}(\hat{x}), g) + \mathcal{L}_{\text{P}}(x, \hat{x}; \mathcal{D}))}_{\text{D}}, \tag{1}$$

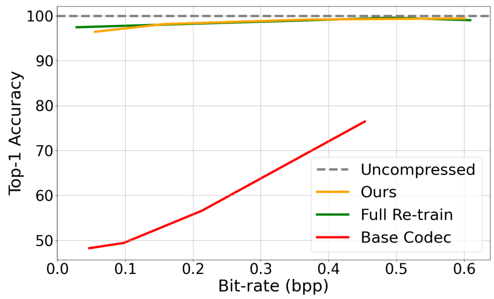


Fig. 4. Rate-accuracy comparison.

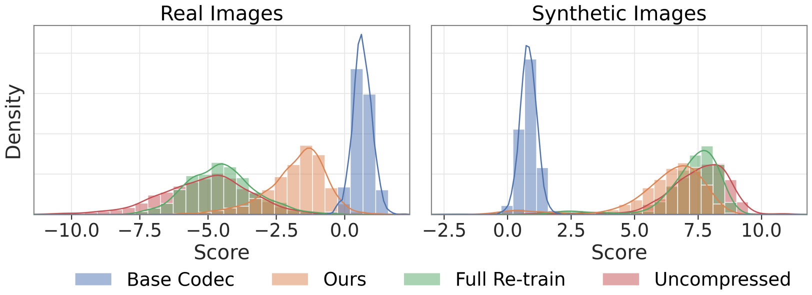


Fig. 5. Detection score distribution on real and synthetic images across different settings. Positive (negative) detection scores indicate the images are classified as synthetic (real).

where R represents the total estimated bit rate, and D quantifies the effect of the distortion in the reconstructed image $\hat{x}$ on the performance of the deepfake detector $\mathcal{D}$. The hyperparameter λ controls the trade-off between D and R. Specifically, the distortion D is measured using two loss functions: (i) $\mathcal{L}_{\mathrm{C}}$, which represents the cross-entropy loss between the ground truth label g of x and the detector's output $\mathcal{D}(\hat{x})$, and (ii) $\mathcal{L}_{\mathrm{P}}$, a perceptual loss computed as the Mean Squared Error (MSE) between the intermediate features $\mathcal{F}$ extracted by the detector $\mathcal{D}$ when processing the original uncompressed image and the reconstructed image, that is $\mathcal{L}_{\mathrm{P}} = \mathrm{MSE}\left(\mathcal{F}(x), \mathcal{F}(\hat{x})\right)$. This second term enforces alignment between the two feature representations, reducing discrepancies introduced by compression that may affect deepfake detection. Throughout the training process, the base codec parameters remain frozen, while only the SFMA modules are learned.

IV. Experimental Results

A. *Experimental Settings*

Deepfake Detector and Dataset. In this work, the real versus synthetic detector from [16] has been used as target deepfake detector, with no modifications or fine-tuning. The considered dataset contains images that the pre-trained detector can correctly classify in their uncompressed form. To this end, images from the original training dataset [16] that includes real and synthetic human faces have been chosen. The real images are sourced from FFHQ [19] and CelebAHQ [20], while the synthetic images are generated using a wide variety of state-of-the-art generative models ranging from GANs to diffusion models. The proposed system has been trained with roughly 11,000 images, with additional 4,000 images for evaluation. Both training and evaluation sets maintain a balanced ratio of real and synthetic images.

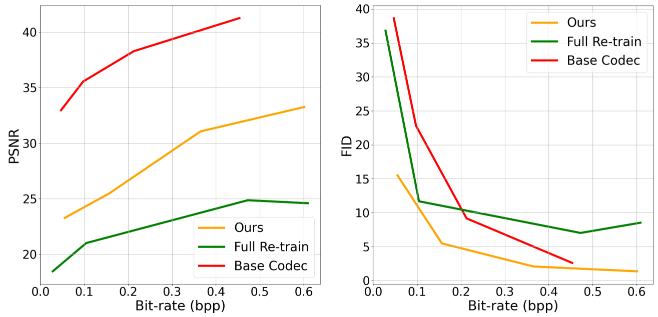
Training and Evaluation. The pre-trained TIC [13] is the base codec for all experiments. To expedite the training process, input images are randomly cropped to 256 × 256 before encoding, and the reconstructed images are processed by the detector using 50 patches to compute the cross-entropy loss and perceptual loss. Separate models have been trained using four different rate parameters $\lambda \in \{0.02, 0.03, 0.045, 0.07\}$, corresponding to different bit rates. Training is conducted on 50 epochs using an Adam optimizer and a learning rate of 0.001, which is reduced by half after 30 epochs.

For evaluation, the proposed method has been compared against two baselines: (1) the base codec optimized for PSNR and (2) a fully re-trained codec optimized for deepfake detection with the same proposed training objective. The deepfake detection process in evaluation follows exactly the original approach in [16], with $M = 600$ and top-1 accuracy and bits per pixel (bpp) as main evaluation metric.

B. *Performance Comparison*

Fig. 4 shows the rate-accuracy performance comparison between the three methods. The detection accuracy using uncompressed images serves as the upper bound, achieving 100% accuracy. Three key observations are drawn from Fig. 4: (1) The detection accuracy for images compressed with the base codec shows a drastic drop, nearing random guessing at the lowest rate point. (2) The proposed method, which integrates SFMA into the fixed base codec, achieves a considerable accuracy improvement with up to a 45% gain, reaching an overall accuracy of 96% even at extremely low bit rates ($\approx$ 0.05 bpp). (3) Compared to full re-training of the base codec, the proposed method achieves comparable performance while requiring significantly fewer learnable parameters, making it a more efficient solution.

Detection scores obtained from reconstructed images of the different settings are used to investigate the effect of the proposed method (see Fig. 5). We recall that a positive detection score indicates the image is classified as synthetic, while a negative score signifies it is classified as real (see Sec. II-B). The score distribution of the base codec has a mean close to 0 for both real and synthetic images, indicating that the detector fails to differentiate between them. This suggests that compression artifacts disrupt the features necessary for an accurate classification. In contrast, the proposed method, as well as the fully re-trained model and the detector applied to uncompressed images, result in distinct score distributions, with most real images leading to negative scores while synthetic images leading to positive ones.



(a) PSNR↑ (b) FID↓

Fig. 6. Rate-distortion performance comparison on PSNR and FID.

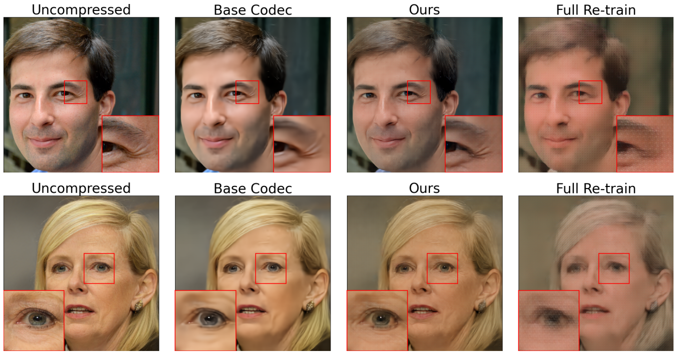


Fig. 7. Reconstructed images of different settings.

C. Reconstruction Quality Comparison

Fig. 6 compares RD performance using PSNR and FID [21], where FID measures perceptual similarity to uncompressed images, capturing realism-related distributional differences [22]. While the base codec achieves higher PSNR, the proposed method yields lower (better) FID scores, indicating greater realism and a distribution closer to uncompressed images. This explains why the proposed approach induces a better preservation of the information needed for accurate deepfake detection.

A qualitative comparison of reconstructed images is also shown in Fig. 7. The base codec removes fine details and produces smoother, slightly blurred images. In contrast, the proposed method preserves more textural details while maintaining good perceptual quality. On the other hand, full retraining introduces heavy artifacts, leading to visually unappealing reconstructions.

V. Conclusion

LIC poses challenges for image forensics, as it can lead to misclassification of both authentic and synthetic images. In this work, the first neural image compression system tailored for deepfake detection has been proposed by integrating an SFMA into an existing codec, avoiding the need for full retraining. Experimental results demonstrate that the proposed method achieves detection performance close to uncompressed images while also enhancing perceptual quality with respect to images reconstructed from the original LIC scheme.

References

[1] R. Zou, C. Song, and Z. Zhang, “The devil is in the details: Window-based attention for image compression,” *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[2] Y. Zhu, Y. Yang, and T. Cohen, “Transformer-based transform coding,” in *International Conference on Learning Representations*, 2022.

[3] D. He, Z. Yang, W. Peng, R. Ma, H. Qin, and Y. Wang, “Elic: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[4] D. Wang, W. Yang, Y. Hu, and J. Liu, “Neural data-dependent transform for learned image compression,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[5] J. Liu, H. Sun, and J. Katto, “Learned image compression with mixed transformer-cnn architectures,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 14 388–14 397.

[6] E. Alshina, J. Ascenso, and T. Ebrahimi, “Jpeg ai: The first international standard for image coding based on an end-to-end learning-based approach,” *IEEE MultiMedia*, vol. 31, no. 4, pp. 60–69, 2024.

[7] N. Hofer and R. Böhme, “A taxonomy of miscompressions: Preparing image forensics for neural compression,” 2024. [Online]. Available: https://arxiv.org/abs/2409.05490

[8] E. D. Cannas, S. Mandelli, N. Popovic, A. Alkhateeb, A. Gnutti, P. Bestagini, and S. Tubaro, “Is jpeg ai going to change image forensics?” *arXiv preprint arXiv:2412.03261*, 2024.

[9] S. Bergmann, D. Moussa, F. Brand, A. Kaup, and C. Riess, “Frequency-domain analysis of traces for the detection of ai-based compression,” in *International Workshop on Biometrics and Forensics (IWBF)*, 2023.

[10] ——, “Forensic analysis of ai-compression traces in spatial and frequency domain,” *Pattern Recognition Letters*, vol. 180, pp. 41–47, 2024.

[11] H. Li, S. Li, S. Ding, W. Dai, M. Cao, C. Li, J. Zou, and H. Xiong, “Image compression for machine and human vision with spatial-frequency adaptation,” in *Computer Vision – ECCV 2024*. Springer Nature Switzerland, 2025, pp. 382–399.

[12] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, “Variational image compression with a scale hyperprior,” 2018. [Online]. Available: https://arxiv.org/abs/1802.01436

[13] M. Lu, P. Guo, H. Shi, C. Cao, and Z. Ma, “Transformer-based image compression,” in *2022 Data Compression Conference (DCC)*, 2022.

[14] L. Verdoliva, “Media forensics and deepfakes: An overview,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.

[15] U. Ojha, Y. Li, and Y. J. Lee, “Towards universal fake image detectors that generalize across generative models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 24 480–24 489.

[16] S. Mandelli, P. Bestagini, and S. Tubaro, “When synthetic traces hide real content: Analysis of stable diffusion image laundering,” in *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2024, pp. 1–6.

[17] D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, and L. Verdoliva, “Raising the bar of ai-generated image detection with clip,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 4356–4366.

[18] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[19] “Flickr-Faces-HQ dataset (FFHQ),” https://github.com/NVlabs/ffhq-dataset.

[20] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of gans for improved quality, stability, and variation,” in *6th International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, April 30 - May 3, 2018*, 2018.

[21] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” *Advances in neural information processing systems*, vol. 30, 2017.

[22] E. Agustsson, D. Minnen, G. Toderici, and F. Mentzer, “Multi-realism image compression with a conditional generator,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 324–22 333.