**DOI 10.70286/isu-06.08.2025.1.002**

# METHODS FOR DETECTING ARTIFICIALLY GENERATED IMAGES: A REVIEW OF CURRENT APPROACHES

**Vorobel Oksana**
Student
Department of Computer Science
Lviv Polytechnic National University, Ukraine

Today, the rapid development of generative models has enabled us to create photorealistic images with just a few clicks. On the one hand, this improves creative and production processes, but on the other hand, it creates numerous threats, including the widespread dissemination of misinformation and manipulation of public opinion through fake photos, copyright issues (because it is unclear whether machine-generated images can have their own copyright [1]), as well as the possibility of fraud and political manipulation [2]. Therefore, it is very important to be able to identify such artificially generated images. As most researchers note, detection tools are becoming a ‘critical means’ of protecting the information space and user rights [3].

The purpose of this review is to systematically summarise and classify modern methods for detecting artificially generated images. We will consider the main categories of approaches, analyse their advantages and limitations, and formulate recommendations for further research. As noted by the authors of recent reviews, such works aim to create a ‘structured analysis of detection methods’ and to distinguish between ‘unimodal and multimodal approaches’ [3].

Methods for detecting artificially generated images can be divided according to the modal principle and feature domain. Unimodal approaches work exclusively with the image itself, analysing its pixel content, while multimodal approaches combine image processing with text descriptions or prompts. Unimodal methods are often classified by feature domain into spatial (operating in the original pixel space), frequency (analysing spectral properties) and hybrid (combining spatial and frequency information). Multimodal methods typically use language models to match images with text [3].

Spatial methods are based on the analysis of geometric, colour, and texture patterns in an image. For example, Marra et al. (2018) compared several detectors of generated images and showed that deep neural networks are capable of maintaining high accuracy even when images are compressed [3]. Many works from 2018 to 2022 used CNNs or autoencoders to remove residual noise from the generator: Hsu et al. (2018) trained a contrastive model to learn the salient features of synthetic images, Yu et al. (2019) restored images using an autoencoder and compared its ‘fingerprint’ with known fingerprints of the GAN model [3]. In modern approaches, CNNs are complemented by self-attention and ensemble mechanisms, and pre-trained models

such as ResNet are used for classification [3][4]. However, such methods work well on known model generations but may not generalise well to new high-quality generators.

Frequency-based methods are based on the analysis of spectral artefacts that often appear in synthetic images. For example, Zhang et al. (2019) discovered unique frequency artifacts caused by scaling operations in GANs [3]. Frank et al. (2020) applied discrete cosine transform (DCT) and optimized the regressor to detect synthetic images [3]. These approaches are useful for detecting the transition between generative and real spectra, especially at high frequencies. However, modern generative models are beginning to eliminate obvious spectral differences. New methods (e.g., FreqNet) are trained separately to use the phase and amplitude components of the spectrum to focus on high-frequency artifacts [3].

Hybrid methods combine information from multiple domains to improve recognition. For example, Wang et al. (2023) obtained a medium-high frequency image through discrete wavelet transform and combined it with the original image in CNN [3]. Hybrid methods typically improve detection accuracy, but their complexity and computational requirements increase.

Multimodal methods use the relationship between images and text. Modern large vision and language models (e.g., CLIP) can generate compatible features for images and captions. Research shows that such models are easily adaptable to the detection task: for example, Roy et al. (2024) fine-tuned CLIP on a dataset of various generated images and outperformed specialised detectors in terms of accuracy without any architecture-specific model [1]. Other works directly use text prompts: Wu et al. (2023) aligned image vectors with ‘fake photo’ text vectors, and Sha et al. (2024) studied the impact of prompts on detection [3]. Liu et al. (2024) developed the FatFormer transformer with a ‘fake’ adapter that compares images and corresponding text [3]. Multimodal approaches demonstrate high generalisability on unseen data but require costly training with large text-visual datasets and are sensitive to the quality of image-text pairs [3].

Active methods are based on embedded tags or watermarks that can be generated by the model (e.g., hidden patterns in the source data). In theory, active tags provide reliable identification, but in practice, most generators do not add such tags. Therefore, passive methods that analyse the image content prevail. According to Guo et al. (2023), active approaches are currently ‘very limited’ because there are no markers[3], and the main focus in detection remains on analysing visible differences in image structure and statistics.

Table 1. Comparison of the main approaches to recognising AI-generated images

| Approach | Description | Advantages | Disadvantages |
|---|---|---|---|
| Spatial CNN | Training CNN/neural networks on pixel-level or residual features. | High accuracy on known generators; does not require metadata. | Often poorly generalize to unseen models; require large datasets. |
| Color/Texture | Use of statistical color and texture features. | Simple and intuitive. | Generators improve color modeling; textures can become more realistic. |

Continuation of table 1

| Approach | Description | Advantages | Disadvantages |
|---|---|---|---|
| Frequency-based | Analysis of DCT/FFT coefficients, high-frequency components. | Effectively detect spectral inconsistencies, especially in high frequencies. | Advanced generators minimize spectral artifacts; sensitive to compression/filtering. |
| Cross-domain (Hybrid) | Combination of spatial and frequency-based features. | Complementary features enhance detection; capable of recognizing new patterns. | High computational complexity; possible information redundancy and overfitting. |
| VLM/CLIP (Multimodal) | Use of pre-trained image/text models. Methods include prompt tuning and contrastive learning. | High generalization and robustness to various generators; no manual feature design needed. | Requires large multilingual datasets and significant computational resources; sensitive to text-image alignment accuracy. |
| Active | Embedded marks or watermarks. | Instant verification when marks are present. | Rarely used (lack of embedded marks). |
| Passive | Content analysis. | Independent of metadata. | Limited by algorithms ability to detect complex artifacts. |

Artificially created images are becoming increasingly realistic, which increases the relevance of their reliable detection. As shown in Table 1, each category of methods has its strengths and weaknesses. Spatial detectors often provide high accuracy on familiar data, but their effectiveness decreases with the emergence of new generators with higher quality. Frequency-based methods complement spatial methods by detecting artefacts invisible to the human eye, but they also lose their recognition ability on improved models. Multimodal approaches demonstrate good generalisability because they take into account the semantics of images, but they require complex architectures and large datasets. Overall, no method is universal. Further research should focus on developing more ‘universal’ methods that are resistant to unseen generation models. In summary, this comparison shows that the development of artificial image detection technologies will continue to require the integration of different approaches, balanced between accuracy and generalisability.

## References

1. Mahara, A., & Rishe, N. (2025). Methods and Trends in Detecting Generated Images: A Comprehensive Review (No. arXiv:2502.15176). arXiv. https://doi.org/10.48550/arXiv.2502.15176
2. Moskowitz, A. G., Gaona, T., & Peterson, J. (2024). Detecting AI-Generated Images via CLIP (No. arXiv:2404.08788). arXiv. https://doi.org/10.48550/arXiv.2404.08788

3. Wang, H. (2025). Vision Transformer-Based Framework for AI-Generated Image Detection in Interior Design. Informatica, 49(16). https://doi.org/10.31449/inf.v49i16.7979
4. Zhang, Y., Pang, Z., Huang, S., Wang, C., & Zhou, X. (2025). Unmasking AI-created visual content: A review of generated images and deepfake detection technologies. Journal of King Saud University Computer and Information Sciences, 37(6), 148. https://doi.org/10.1007/s44443-025-00154-8