Synthetic Image Detection under Distribution Shift: CLIP-FFT Fusion and ForensicGAT

Berkay Bayramoglu[1,*,†], Ismail Erol[1,†], Rukiye Savran Kiziltepe[1,†] and Murat Karakus[1,†]

[1]*Department of Software Engineering, Ankara University, Ankara, 06830, Türkiye*

Abstract

We present the ANLAM-NET-M team submission for Subtask A: Real vs. Synthetic Images of the MediaEval 2026 task Synthetic Images: Advancing Detection and Localization of Generative AI Used in Real-World Online Images, addressing the binary classification of real versus synthetic images collected from social media. The central challenge is distribution shift: training data consist of controlled GAN and diffusion outputs, while test images reflect real-world compression, resizing, and modern generators. For the constrained run, we combine a frozen Contrastive Language-Image Pre-training (CLIP) ViT-L/14 encoder with 16 deterministic Fast Fourier Transform (FFT) spectral statistics in a compact fusion head, reaching F1-Score = 0.6294. For the open run, ForensicGAT pairs a trainable Swin Transformer (Swin-Tiny) with four ForensicCNN branches fused by a Graph Attention Network v2 (GATv2); admitting the validation set into training raises F1-Score to 0.8320. Aligning the training and test distributions affects detection more than architectural complexity alone.

1. Introduction

Photorealistic image synthesis is now within reach of any web user, and generated content circulates on social platforms before any verification. By the time a detector sees an image, it has typically passed through compression and re-uploads that erode the low-level traces detectors rely on. This distribution shift between controlled training data and real-world posts is the core difficulty of Subtask A of the MediaEval 2026 *Synthetic Images* task [1]: standard corpora of GAN-based and diffusion-based outputs [2, 3] differ systematically from in-the-wild social-media images. In early experiments we observed training F1-Score near 0.97 but validation F1-Score below 0.15, showing that a model can memorize training statistics without learning generalizable signals.

Our contributions are threefold: (i) CLIP-FFT Fusion, a constrained detector that combines a frozen CLIP ViT-L/14 encoder with 16 deterministic FFT statistics and a compact head; (ii) ForensicGAT, an open-run dual-stream model that treats four forensic channels (gradient, frequency, residual, correlation) as graph nodes processed by a Graph Attention Network v2 (GATv2) and fused with a trainable Swin Transformer (Swin-Tiny) backbone; (iii) an analysis showing that validation-augmented training mitigates distribution shift more than added architectural complexity alone.

MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online

*Corresponding author.

†These authors contributed equally.

24290604@ogrenci.ankara.edu.tr (B. Bayramoglu); 24290049@ogrenci.ankara.edu.tr (I. Erol); rukiyekiziltepe@ankara.edu.tr (R. S. Kiziltepe); mrtkarakus@ankara.edu.tr (M. Karakus)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

2. Related Work

Early synthetic image detectors exploit spectral artifacts produced by GAN upsampling [2] or diffusion-model denoising [3], but these low-level cues degrade after the compression and resizing common on social media.

Large pretrained encoders offer representations that transfer more reliably across generator types. Li et al. [4] show that a frozen CLIP ViT-L/14 encoder with a linear probe generalizes across generators when fine-tuned on in-distribution data. Huda et al. [5] compare Transformer-based and CNN-based architectures, finding that the two architecture families capture complementary aspects of synthetic image artifacts.

Augmenting RGB input with forensic channels is a complementary direction. Avantikaa et al. [6] and Le et al. [7] show that frequency, edge, and residual maps reveal manipulation traces invisible in RGB. Khan et al. [8] combine CNN confidence with handcrafted texture features (Gabor, LBP), showing that unsupervised descriptors complement learned ones. We benchmark against BFree [9] and RINE+TWIGMA [10], the strongest in-the-wild detectors from the prior MediaEval benchmark.

3. Task and Dataset

Subtask A of the MediaEval 2026 *Synthetic Images* task [1] asks whether each image is real or synthetic, under a *Constrained Run* (only the official training dataset) or an *Open Run* (any data). The official training corpus consists of Wang et al. [2] and Corvi et al. [3] (≈681,000 images in total): Wang et al. provides GAN-based synthetic images (StyleGAN2, BigGAN, ProGAN), while Corvi et al. contributes diffusion-generated synthetic images. The validation and test sets each contain 5,000 real and 5,000 synthetic images collected in the wild from social media; real images come from LAION and RAISE, while synthetic images span recent generators (GigaGAN, Stable Diffusion, MidJourney, DALL·E 3, and others), then post-processed to mimic online distribution conditions. This mismatch between training and test generators constitutes the core distribution shift. Following SIDBench [11], F1-Score is the primary metric, as it balances false alarms and missed detections.

4. Approach

We submitted one system per run: a frozen-encoder detector with deterministic spectral features (constrained), and a trainable dual-branch network that admits the validation set into training (open).

4.1. Constrained Run: CLIP-FFT Fusion

The constrained run must confront distribution shift with no extra data, since the official training set (Wang et al. [2], Corvi et al. [3]) is dominated by controlled GAN and diffusion outputs unlike the in-the-wild test set. Following Li et al. [4], whose open-run results show that a frozen CLIP encoder transfers well across generators when a linear probe is trained on in-distribution data, we use a frozen CLIP ViT-L/14 backbone, but replace their linear probe with 16 deterministic spectral statistics and a deeper head. The feature vector is $\mathbf{z} = [\mathbf{f}_{\text{CLIP}}; \mathbf{s}_{\text{FFT}}] \in \mathbb{R}^{784}$, where $\mathbf{f}_{\text{CLIP}} \in \mathbb{R}^{768}$ is the CLIP embedding and $\mathbf{s}_{\text{FFT}} \in \mathbb{R}^{16}$ the spectral vector; Figure 1 illustrates this pipeline.

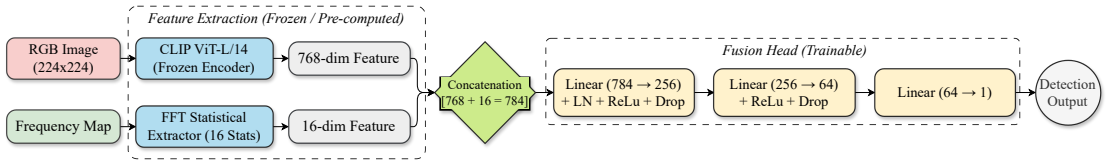

Figure 1: CLIP-FFT Fusion. A frozen CLIP ViT-L/14 embedding (768-dim) is concatenated with 16 deterministic FFT spectral statistics and classified by the FusionHead.

GAN and diffusion models leave systematic high-frequency traces [3]. We capture these without training via 16 statistics from each magnitude spectrum: energy ratio (1), directional sector energies (8), spectral entropy (1), radial-ring energies (4), and peak-frequency coordinate (2). Combining learned representations with handcrafted descriptors of this kind follows Khan et al. [8], who show that unsupervised texture features complement CNN-based detectors.

Because the backbone is frozen, all features are precomputed once over the roughly 681,000 training samples and cached in GPU memory; only the FusionHead, $\text{Linear}(784{\to}256) \to \text{LN} \to \text{ReLU} \to \text{Dropout}(0.3) \to \text{Linear}(256{\to}64) \to \text{Linear}(64{\to}1)$ (about 218,000 parameters), is updated, giving a $10\times$ to $20\times$ reduction in per-epoch time over end-to-end fine-tuning. Training uses Focal Loss ($\alpha{=}0.25$, $\gamma{=}2.0$, label smoothing 0.05) [12] with AdamW and cosine annealing (batch 512), and Temperature Scaling [13] calibrates the output probabilities.

4.2. Open Run: ForensicGAT

With additional data allowed, we attack distribution shift directly by training on $\mathcal{D}_{\text{Wang+Corvi}} \cup \mathcal{D}_{\text{val},\,90\%}$, holding out the remaining 10% of the validation set for monitoring. This single decision drives most of the gap between the two runs.

The architecture has two branches. Motivated by the comparative analysis of Huda et al. [5], which shows Transformer-based and CNN-based architectures to capture complementary artifact signals, we pair Swin-Tiny [14] as a global semantic branch with a forensic branch, ForensicGAT, over four low-level maps: gradient (Sobel), frequency (FFT log-magnitude), residual (non-local means), and correlation (local Pearson). This multi-channel choice follows Avantikaa et al. [6] and Le et al. [7], who show such maps encode traces invisible in RGB; we add a graph stage to model inter-channel dependencies; Figure 2 shows the complete architecture.

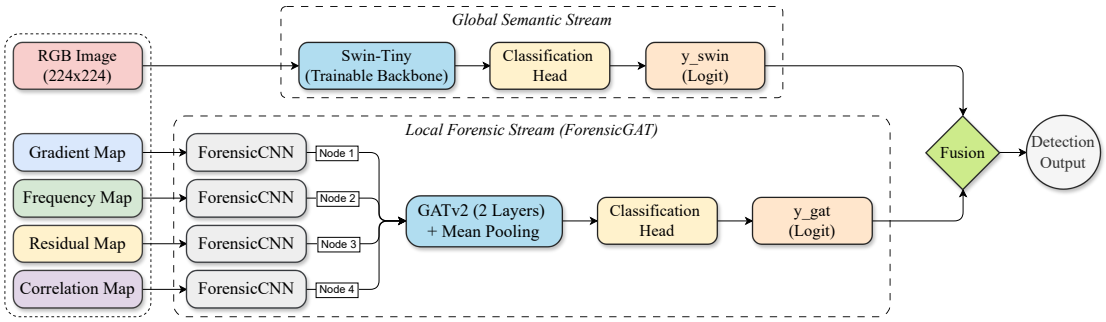

Figure 2: ForensicGAT. The Swin-Tiny stream encodes global features from RGB; four ForensicCNNs encode gradient, frequency, residual, and correlation maps as graph nodes processed by two GATv2 layers. A learnable scalar w weights the two streams.

Each map passes through an independent ForensicCNN ($\text{Conv2d}(1{\to}32{\to}64{\to}128) \to \text{pool} \to \text{Linear}(128{\to}64)$), giving four 64-dim embeddings that form the nodes of a fully connected 4-node graph. Two GATv2 [15] layers (4 heads, hidden 128) with mean pooling yield

a logit $\hat{y}_{\text{gat}}$, while the Swin-Tiny head yields $\hat{y}_{\text{swin}}$. The two combine through a learnable scalar w, initialized to 0:

$$\hat{y} = \sigma(w)\,\hat{y}_{\text{swin}} + \big(1 - \sigma(w)\big)\,\hat{y}_{\text{gat}}. \tag{1}$$

Training applies the multi-task loss to the fused output of (1): $\mathcal{L} = \mathcal{L}_{\text{focal}}(\hat{y}) + 0.15\,\mathcal{L}_{\text{focal}}(\hat{y}_{\text{swin}}) + 0.15\,\mathcal{L}_{\text{focal}}(\hat{y}_{\text{gat}})$, using AdamW ($\eta$=$10^{-4}$), cosine annealing, batch 16, and early stopping (patience 8); Temperature Scaling [13] calibrates the final probabilities.

5. Results and Analysis

Table 1 reports the full test-set metrics for both submitted systems. The Open Run (ForensicGAT) reaches F1-Score = 0.8320 and AUC = 0.9163. Benchmarking against MediaEval 2025 Synthetic Image Detection participants as a reference baseline, our Open Run places above RINE+TWIGMA [10] (0.806) and the closest open-track participant, Mariappan et al. [16] (0.8315). Li et al. [4] achieve a higher F1-Score = 0.8945; although their open-run training similarly incorporates the validation set, their single linear probe on frozen CLIP features yields superior generalization relative to our fully trainable Swin-Tiny and ForensicCNN branches, which introduce additional parameters and a heightened risk of overfitting.

Table 1
ANLAM-NET-M system performance on the Task A test set.

System	Run	Acc.	Prec.	Rec.	F1	AUC	AP
CLIP-FFT Fusion	Constrained	0.5574	0.5413	0.7516	0.6294	0.5645	0.5294
ForensicGAT	Open	0.8450	0.9082	0.7676	0.8320	0.9163	0.9322

The Constrained Run (CLIP-FFT Fusion) yields F1-Score = 0.6294; among 2025-edition constrained-run systems, this exceeds Avantikaa et al. [6] (0.5225) but falls below Huda et al. [5] (0.859) and Mariappan et al. [16] (0.8485). Its 3,184 false positives reflect the domain gap between legacy training data and in-the-wild social media images. The gap between runs, ΔF1-Score $= 0.2026$, is the clearest signal: once validation images enter training, precision rises from 0.5413 to 0.9082, reducing false positives more than eightfold (from 3,184 to $\approx$388 on the 5,000-image real test split), confirming that data strategy, not architecture, drives the gain.

6. Discussion and Outlook

Several limitations should be considered when interpreting these results. The 16 FFT statistics were not ablated against a plain CLIP probe [4]. The submitted threshold of 0.10 was suboptimal (0.045 would have given F1-Score = 0.6701). ForensicGAT was monitored on a held-out validation subset, introducing leakage that may inflate its score. Future work should address threshold search and test alternative frozen backbones such as DINOv2 or SigLIP to reduce *distribution shift* without extra labels.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT in order to: Improve writing style, Grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] O. Papadopoulou, D. Karageorgiou, C. Koutlis, E. Gavves, H. Mareen, S. Papadopoulos, Synthetic images at mediaeval 2026: Advancing detection of generative ai in real-world online images, in: Proceedings of MediaEval'26: Multimedia Evaluation Workshop, Amsterdam, Netherlands and Online, 2026.

[2] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, A. A. Efros, CNN-generated images are surprisingly easy to spot...for now, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8692–8701. doi:10.1109/CVPR42600.2020.00872.

[3] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, L. Verdoliva, On the detection of synthetic images generated by diffusion models, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023, pp. 1–5.

[4] Q. Li, A. Ciamarra, R. Caldelli, S. Berretti, A CLIP-based approach for synthetic image detection under distribution shift, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025. URL: https://2025.multimediaeval.com/paper6.pdf, cEUR Workshop Proceedings.

[5] N. u. Huda, A. Fayyaz, U. Asad, Y. S. Afridi, Synthetic vs real image detection using vision transformers and CNN-based architectures, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025. URL: https://2025.multimediaeval.com/paper18.pdf, cEUR Workshop Proceedings.

[6] R. Avantikaa, S. K. Sangeetha, B. Madhuri, A. VijayaLakshmi, Six-channel deep learning approach for detecting AI-generated images, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025. URL: https://2025.multimediaeval.com/paper23.pdf, cEUR Workshop Proceedings.

[7] M.-H. Le, M.-K. Le-Phan, K.-N. Vu-Nguyen, M.-T. Tran, T.-L. Do, A resolution-agnostic three-stage framework for image forgery detection and localization, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025. URL: https://2025.multimediaeval.com/paper25.pdf, cEUR Workshop Proceedings.

[8] Z. Khan, Z. Zainab, A three-way decision-based synthetic image detection, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025. URL: https://2025.multimediaeval.com/paper9.pdf, cEUR Workshop Proceedings.

[9] F. Guillaro, G. Zingarini, B. Usman, A. Sud, D. Cozzolino, L. Verdoliva, A bias-free training paradigm for more general AI-generated image detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 18685–18694.

[10] C. Koutlis, S. Papadopoulos, Leveraging representations from intermediate encoder-blocks for synthetic image detection, in: European Conference on Computer Vision (ECCV), Springer, 2024, pp. 394–411.

[11] M. Schinas, S. Papadopoulos, SIDBench: A python framework for reliably assessing synthetic image detection methods, in: Proceedings of the 3rd ACM International Workshop on Multimedia AI Against Disinformation, 2024, pp. 55–64.

[12] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988. doi:10.1109/ICCV.2017.324.

[13] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, in: Proceedings of the 34th International Conference on Machine Learning (ICML), volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 1321–1330.

[14] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10012–10022. doi:10.1109/ICCV48922.2021.00986.

[15] S. Brody, U. Alon, E. Yahav, How attentive are graph attention networks?, in: International Conference on Learning Representations (ICLR), 2022. URL: https://openreview.net/forum?id=F72ximsx7C1.

[16] S. M. T. Mariappan, M. Ramasamy, B. Arul, An EfficientNet framework: Methods and results for synthetic image detection and manipulation localization, in: Proceedings of the MediaEval 2025 Workshop, Dublin, Ireland and Online, 2025. URL: https://2025.multimediaeval.com/paper8.pdf, cEUR Workshop Proceedings.