Exploring Swin Transformers, Luminance Input, and Hybrid Architectures for Synthetic Image Detection

Amrit Gopinath[1,*,†], Raghul[1,†] and P Mirunalini[1,†]

[1]Department of Computer Science and Engineering, Sri Sivasubramaniya Nadar College of Engineering, Chennai, India

Abstract

We participated in the MediaEval 2026 Synthetic Image Detection task, addressing Task A: binary classification of real versus synthetic images. Our submissions were designed as a controlled progression across model families and input variants. Starting with pretrained Swin-based systems, we evaluated standard, global, MIL-style, and luminance-only variants, and further explored compact CNN- and ResNet-based GRU+ViT hybrid architectures under RGB and luminance-only settings. The results show that open runs achieved stronger performance than constrained runs, with the best F1 score reaching 0.6997. In the constrained setting, Swin-based systems behaved more conservatively, whereas CNN- and ResNet-based hybrid architectures were more synthetic-recall oriented. These findings suggest that constrained runs are particularly affected by architectural design and input representation.

1. Introduction

We took part in the MediaEval 2026 Synthetic Image Detection task, focusing on Task A: binary classification of real versus synthetic images [1]. Our overall goal was not only to submit a strong final detector, but also to study how different architectures and input representations affect synthetic image detection under constrained and open settings. In particular, we explored a progression from pretrained Swin-based models to luminance-only variants and compact hybrid systems combining convolutional based architectures with GRU and transformer-based feature modeling. The task is challenging because synthetic image generators have improved substantially, often producing highly realistic visual content while still leaving subtle forensic traces. Therefore, we compare strong pretrained reference systems with more compact task-specific hybrid models to examine whether detection relies more on high-level semantic features, local artifact cues, or a combination of both.

2. Related Work

Recent synthetic image detection systems have largely been built around pretrained Vision Transformers [2], multimodal representations such as CLIP [3], and self-supervised foundation models such as DINOv2 [4]. These systems are strong starting points because large-scale pretraining provides robust semantic and texture priors, but this can also make it difficult to determine whether performance is driven by task-specific forensic cues or by more general visual knowledge. From a forensic perspective, prior work has studied CNN-based discriminators, diffusion-oriented detection settings, and color-component disparities between real

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

*Corresponding author.

†All authors contributed equally.

amrit2410182@ssn.edu.in (A. Gopinath); raghul2510435@ssn.edu.in (Raghul); miruna@ssn.edu.in (P. Mirunalini)

and generated images [5, 6, 7]. The latter is particularly relevant to our use of luminance-only variants, since it highlights that the choice of input representation can affect detectability.

Prior work has also explored patch-level reasoning and localization-oriented aggregation, motivated by the observation that synthetic artifacts are often spatially uneven rather than uniformly distributed across an image. Our submissions build on this line of work by comparing pretrained Swin-based reference systems, luminance-only variants, and compact hybrid architectures under both constrained and open training settings. This design allows us to study not only final performance, but also how backbone strength, input representation, and aggregation strategy influence the precision-recall trade-off in synthetic image detection.

3. Approach

All systems were trained for binary classification with the labels *real* and *synthetic*. Constrained runs used: Wang real images for the authentic class, and Wang fake plus Corvi latent diffusion images[6, 5] for the synthetic class. Open runs followed the same task formulation, but additionally incorporated SID_Set as external training data [8], with its labels mapped to the required binary real/synthetic labels. In both settings, we used fixed random seeds, controlled source sampling, and held-out validation splits to support consistent comparison across model families. Images were resized or cropped to 224 pixels and trained using standard augmentation, including random resized crops, horizontal flips, Gaussian blur, JPEG compression, and color perturbation where appropriate. Constrained training used 200,000 Wang real images together with 100,000 Wang fake and 100,000 Corvi latent diffusion images, while open training used 70,000 SID_Set real images together with 50,000 full synthetic and 50,000 tampered images.

Runs 1–4 are pretrained Swin-based systems intended as strong reference models, while Runs 5–8 are from-scratch hybrid architectures designed to test how luminance input and convolutional feature extraction interact with GRU- and transformer-based token-level aggregation.

3.1. Swin-Based Systems

We began with Swin Transformer because it is a strong hierarchical vision backbone that combines transformer reasoning with locality through shifted-window self-attention [9, 10]. This makes it a useful reference architecture for synthetic image detection, where both fine artifact patterns and broader image-level structure may matter. Run 1 uses a standard RGB Swin classifier as the main baseline. Run 2 keeps the same backbone but changes the final summarization by using a stronger global aggregation head, allowing us to test whether image-level pooling changes the decision boundary.

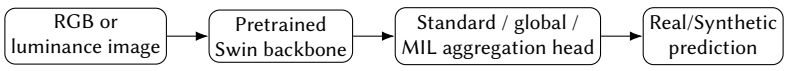

Figure 1: Overview of Runs 1–4: Swin-based systems with different input and aggregation variants.

3.2. MIL and Luminance Variants

Run 3 adds multiple-instance style reasoning to the Swin backbone by treating the image representation as a collection of local instances arranged on a 14×14 grid rather than reducing it immediately to a single global vector. The final decision is then based on aggregated regional evidence, which is useful when synthetic artifacts are spatially localized or unevenly distributed.

Run 4 keeps the Swin family but changes the input representation by converting RGB images to a single luminance channel before feature extraction. This luminance-only setting reduces dependence on color semantics and tests whether structural, brightness, and texture cues alone remain discriminative.

3.3. CNN and ResNet Hybrid Systems

Runs 5–8 replace the pretrained Swin backbone with compact from-scratch hybrids. In these models, a convolutional frontend first extracts local spatial features, which are then pooled into a token grid. A bidirectional GRU mixes the token sequence, after which a lightweight transformer encoder performs global reasoning over the token set. The resulting token-level representation is fused with a global pooled embedding for final classification. Runs 5 and 6 use a compact CNN frontend under luminance and RGB input respectively, providing a direct representation-controlled comparison. Runs 7 and 8 replace the simple CNN with a ResNet-style residual frontend, allowing us to test whether deeper staged convolution and residual propagation preserve forensic cues differently from the smaller CNN stem.

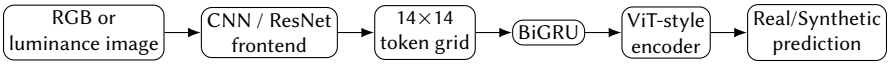

Figure 2: Overview of Runs 5–8: compact CNN/ResNet hybrid systems with token-level sequence and attention modeling.

All submitted systems were trained with a largely shared configuration. Images were resized or cropped to 224 pixels and trained for 3 epochs with a batch size of 16 and 2 gradient accumulation steps. Optimization used AdamW with a learning rate of 10^{-4} and weight decay of 10^{-4}. For the hybrid architectures, dropout was set to 0.2, the token grid size was 14, the GRU hidden dimension was 128, the transformer dimension was 192, the number of transformer heads was 3, and the number of transformer layers was 2.

4. Results and Analysis

Table 1
Constrained validation results. Precision, recall, and F1 are computed for the synthetic class.

Run	System	Acc.	Prec.	Rec.	F1	AUC	AP
1	Swin Base	0.4712	0.4770	0.5978	0.5306	0.4486	0.4492
2	Swin + Global	0.4495	0.4501	0.4560	0.4531	0.4270	0.4421
3	Swin + MIL 14	0.4434	0.4379	0.3992	0.4177	0.4183	0.4424
4	Swin + Luma	**0.5272**	**0.5222**	0.6406	0.5754	**0.5235**	**0.4981**
5	CNN + GRU + ViT + Luma	0.4999	0.4999	0.9906	0.6645	0.4207	0.4426
6	CNN + GRU + ViT + RGB	0.5000	0.5000	**1.0000**	**0.6667**	0.4610	0.4784
7	ResNet + GRU + ViT + Luma	0.5003	0.5002	0.9374	0.6523	0.4943	0.4835
8	ResNet + GRU + ViT + RGB	0.4952	0.4976	0.9752	0.6589	0.4639	0.4780

Across experiments, open runs were substantially more stable than constrained runs, and most of the strongest overall performance was achieved in the open setting. This suggests that access to broader external training data contributed more to final generalization than architectural refinements alone. In contrast, constrained runs showed noticeably higher variance across

model families, indicating that when data diversity is limited, model inductive bias and input representation play a larger role in determining decision behavior.

Table 2
Open validation results with SID_Set available as external training data.

Run	System	Acc.	Prec.	Rec.	F1	AUC	AP
1	Swin Base	0.6690	0.6403	0.7712	**0.6997**	**0.7471**	0.7623
2	Swin + Global	0.6287	0.5979	0.7858	0.6791	0.7228	0.7578
3	Swin + MIL 14	**0.6693**	**0.6570**	0.7086	0.6818	0.7435	**0.7853**
4	Swin + Luma	0.6389	0.6150	0.7430	0.6729	0.7155	0.7454
5	CNN + GRU + ViT + Luma	0.5259	0.5140	0.9496	0.6670	0.5802	0.5820
6	CNN + GRU + ViT + RGB	0.4994	0.4997	**0.9984**	0.6660	0.5591	0.5971
7	ResNet + GRU + ViT + Luma	0.5610	0.5355	0.9208	0.6772	0.6048	0.5789
8	ResNet + GRU + ViT + RGB	0.5012	0.5006	0.9976	0.6667	0.5424	0.5669

Table 3
Confusion matrices for Run 4 (Swin + Luma) under constrained and open validation.

	Constrained			Open		
Actual	Pred. Real	Pred. Synth.	Total	Pred. Real	Pred. Synth.	Total
Real	2069	2931	5000	2674	2326	5000
Synthetic	1797	3203	5000	1285	3715	5000
Total	3866	6134	10000	3959	6041	10000

An important pattern emerged in the precision-recall trade-off. The stronger pretrained Swin-based systems tended to behave more conservatively, preserving higher precision and lower false-positive rates. In contrast, the compact CNN+GRU+ViT hybrids were more recall-oriented, often identifying more synthetic examples at the cost of increased false positives on real images. The ResNet-based hybrids occupied an intermediate position, suggesting that deeper residual structure can change how low- and mid-level forensic cues are preserved before token-level reasoning. Relative to the other runs, Run 4 (Swin + Luma) showed the most balanced behavior across the two settings. As shown in Table 3, it avoided the extreme synthetic overprediction seen in several hybrid runs and maintained a more even trade-off between retaining real images and recovering synthetic ones, especially after moving from constrained to open training.

5. Discussion and Outlook

The full model family highlights three main points. First, pretrained transformer baselines remain strong reference systems for synthetic image detection. Second, luminance-only input is not merely an auxiliary curiosity, but a meaningful representation that can remain competitive and, in some settings, improve robustness. Third, compact hybrid models trained from scratch can still recover useful forensic structure, although their behavior may shift toward more aggressive synthetic detection boundaries.

Future work should combine luminance-aware preprocessing, patch-level reasoning, and hybrid architectures to balance pretrained backbone precision with forensic model recall. We also plan to explore alternative input channels and chromatic systems, such as RGB, HSV, YCbCr, Lab, edge, noise, and frequency representations.

Declaration on Generative AI

ChatGPT 5 was used for grammar, spelling, and wording suggestions in this paper. The authors reviewed and edited the resulting 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] H. Wang, Vision transformer-based framework for ai-generated image detection in interior design, Informatica 49 (2025). URL: https://www.informatica.si/index.php/informatica/article/view/7979. doi:10.31449/inf.v49i16.7979.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning (ICML), 2021, pp. 8748–8763.

[4] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. HAZIZA, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, P. Bojanowski, DINOv2: Learning robust visual features without supervision, Transactions on Machine Learning Research (2024). URL: https://openreview.net/forum?id=a68SUt6zFt, featured Certification.

[5] 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.

[6] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, L. Verdoliva, On the detection of synthetic images generated by diffusion models, CoRR abs/2211.00680 (2022). doi:10.48550/arXiv.2211.00680.

[7] H. Li, B. Li, S. Tan, J. Huang, Identification of deep network generated images using disparities in color components, Signal Processing 174 (2020) 107616. doi:10.1016/j.sigpro.2020.107616.

[8] Z. Huang, J. Hu, X. Li, Y. He, X. Zhao, B. Peng, B. Wu, X. Huang, G. Cheng, Sida: Social media image deepfake detection, localization and explanation with large multimodal model, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025. doi:10.1109/CVPR52734.2025.02685.

[9] 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.

[10] P. Mehta, A. Sagar, S. Kumari, Enhancing image authenticity detection: Swin transformers and color frame analysis for cgi vs. real images, 2024. URL: https://arxiv.org/abs/2409.04742. arXiv:2409.04742.