Swin Meets EfficientNet: Lightweight Architectures for GAN-Based Face Forensics

Sejuti Basu[1], Ashima Sood[2], Vijay Kumar[3], Sahil Sharma[2]
{basusejuti@gmail.com, sood-a1@ulster.ac.uk, vijaykumarchahar@gmail.com, s.sharma@ulster.ac.uk}
[1]International Institute of Information Technology, Bangalore, India
[2]School of Computing, Engineering, and Intelligent Systems, Ulster University, Londonderry, United Kingdom
[3]Department of Information Technology, Dr B R Ambedkar National Institute of Technology Jalandhar, Punjab, India

Abstract. Modern generative models, such as GANs, diffusion architectures, and autoregressive systems, now produce facial images that are nearly indistinguishable from authentic photographs. This capability makes detecting forged images increasingly difficult, raising serious concerns about identity theft, fraud, and misinformation campaigns. Our research focuses specifically on GAN-generated synthetic faces, which underpin many face-centric deepfakes, and investigates efficient detection approaches using image analysis alone. Existing detection systems rely heavily on either convolutional neural networks (CNNs) or global vision transformers. While CNNs excel at identifying texture-based local features, they struggle with broader contextual understanding. Traditional Vision Transformer (ViT) models can capture long-range structures effectively, but demand substantial computational resources. Our work explores Swin-Transformer-based architectures across three implementations: a compact Swin Transformer trained from the ground up, ImageNet-1K pre-trained Swin-Tiny and Swin-Small models adapted for binary classification, and a novel hybrid combining EfficientNet-B0's convolutional processing with a Swin Transformer backend. We evaluated all models using the 140K Real and Fake Faces dataset, which includes StyleGAN-generated fake faces alongside authentic images from Flickr and DFDC, with balanced splits for training, validation, and testing. The EfficientNetB0+Swin hybrid achieved 99% accuracy and a 99.44% recall on 5,000 test images, outperforming both pure Swin variants and a previous CNN-only baseline on this dataset. Our results suggest that combining hierarchical CNN features with shifted-window self-attention provides an efficient and computationally lightweight method for detecting GAN-generated synthetic faces.

Keywords: Deepfake, Transformers, EfficientNet, Swin-Transformer

1 Introduction

Facial images are commonly utilised as biometric identifiers. Mobile devices, border control systems, financial institutions, and social media platforms collect such data, which is increasingly

IICTDS 2025, December 19-20, Chandigarh, India
Copyright © 2026 EAI
DOI 10.4108/eai.19-12-2025.2362519

repurposed for training large-scale recognition and generative models. At the same time, concerns about biometric leakage and unauthorised reuse of facial data without proper consent continue to grow. These tensions have made data synthesis an important research direction in biometrics. Synthetic faces could help facilitate model development and evaluation while reducing dependence on identifiable subjects and raw operational logs. Regulators and international organisations stress the need for stronger protections against harmful uses of synthetic media, including identity theft, financial fraud, political misinformation, and reputational damage.[1]

In this environment, high-fidelity face synthesis has become widespread. Beyond classical Generative Adversarial Networks (GANs) [1], modern diffusion and autoregressive models can generate photorealistic human faces and perform effective face editing [2, 3]. Our study examines synthetic faces produced by GANs, specifically architectures similar to StyleGAN [4]. These models play a central role in current biometric practices, being used to construct large-scale synthetic face datasets, generate identities with controllable attributes, and support attack pipelines requiring rapid and cost-effective identity fabrication. From a forensic standpoint, GAN-generated faces display unique artifact patterns and distributional biases distinct from those in diffusion models. Understanding how detector architectures respond to GAN-specific signatures is crucial for improving biometric system security and for clarifying the limitations of synthetic face datasets used in training and evaluation.

Deepfake detection is typically framed as a supervised binary classification problem for images or videos, though effective modeling requires more than simple label assignment. Detectors must respond to detailed local evidence like texture inconsistencies, blending artifacts, and abnormal noise statistics, while also accounting for global facial structure and cross-region consistency [5, 6].Convolutional neural networks (CNNs) are frequently used here because they leverage local correlations and perform well across diverse benchmarks. Backbones like VGG, ResNet, Xception, and EfficientNet focus on local patterns. Although their effective receptive field expands with depth, they model global context only indirectly [7]. Vision Transformers (ViTs) [8] take a different approach by applying global self-attention across all patches. This enables robust long-range modeling but comes with higher memory and computational costs, plus greater reliance on extensive pretraining. The Swin Transformer [9] offers a middle ground by combining shifted window self-attention with a hierarchical architecture, balancing local detail, global structure, and computational efficiency. Our paper investigates how architectural choices affect the detection of GAN-generated synthetic faces using Swin-based architectures and a hybrid CNN-Swin design. This approach computes self-attention within non-overlapping windows, alternating with shifted windows, yielding a hierarchical representation with three useful properties for deepfake detection: (i) a multi-scale feature pyramid capturing both fine-grained and semantic information, (ii) self-attention with computational cost scaling linearly with image size, and (iii) cross-window interactions approximating global context without requiring full global attention. Swin is lighter and simpler to implement than ViT-style global transformers, while offering better contextual modeling than pure CNNs.

The primary concern is *how far can Swin-based architectures be advanced for GAN-based synthetic face detection, and how do they perform in comparison to robust CNN baselines?* We analyse Swin both independently and in conjunction with an EfficientNet backbone, utilising StyleGAN-

[1] https://www.unesco.org/en/articles/deepfakes-and-crisis-knowing

generated faces from the 140K Real and Fake Faces dataset as a controlled testbed. This research emphasises GAN-based images, yet the architectural insights are applicable to other generative models.

Our contributions are as follows:

- This study systematically investigates Swin-Transformer architectures for deepfake image detection, analysing the impact of design choices, including patch size, window size, and training regime (from scratch versus transfer learning), on performance with GAN-generated synthetic faces.
- This study presents a hybrid architecture that merges EfficientNet-B0 with a Swin-Transformer head. This integration facilitates convolutional feature extraction alongside shifted-window self-attention, enabling the joint modelling of local artefacts and global facial context, while maintaining a significantly lighter profile compared to a standard ViT.
- This study assesses three Swin-based variants: a task-specific Swin trained from scratch, ImageNet-pretrained Swin-Tiny and Swin-Small, and the proposed hybrid EfficientNet+Swin, using the 140K Real and Fake Faces dataset. These variants are compared to an EfficientNet-based baseline. The hybrid model demonstrates 99% accuracy and a 99.44% recall on a test set of 5,000 images, significantly surpassing the performance of both solely Swin-based and CNN-based detectors on this GAN-centric benchmark.

The remainder of the paper is organised as follows. Section 2 reviews related work on deepfake detection with CNNs and transformers. Section 3 introduces our Swin-based and hybrid architectures. Section 4 presents the experimental setup and results. Section 5 concludes with a discussion of limitations and directions for extending this work beyond GAN-based synthetic faces.

2 Related Work

Initial research focused on manually designed cues and traditional image forensics while contemporary approaches are primarily characterised by deep learning models that function directly on pixel data. This section outlines three primary categories: CNN-based detectors, transformer and hybrid architectures, and Swin-Transformer-based approaches. Our work is situated at the convergence of robustness and computational efficiency with a clear focus on resource-aware model design.

CNN-based detectors: CNNs remain the dominant framework for deepfake detection because of their strong locality bias and favorable accuracy-efficiency tradeoff. Architectures like VGG, ResNet, and Xception have been adapted to emphasize spatial artifacts and abnormal noise statistics in manipulated facial images. Chang et al. use SRM filtering to highlight noise residuals before feeding them to a VGG-16 classifier [10]. Bondi et al. show that carefully designed augmentations and loss functions substantially improve CNN performance on DFDC [11]. EfficientNet variants have been explored as more parameter-efficient backbones for both curated and in-the-wild datasets [12]. Beyond standard CNNs, several studies incorporate frequency-domain or spectral information [13]. Earlier work [14] combines Error Level Analysis (ELA) with a CNN backbone on the 140K Real and Fake Faces dataset, creating a competitive all-convolutional baseline.

Transformer and hybrid detectors: Vision transformers have been implemented in deepfake detection to effectively address the deficiency of global context. ViT-style models utilise self-attention across all image patches, enabling robust long-range modelling that effectively captures inconsistencies in global facial layout, illumination, and geometry. Coccomini et al. integrate an EfficientNet backbone with a ViT branch through cross-attention, effectively merging local and global cues for the purpose of video forgery detection [15]. A range of hybrid CNN-transformer architectures has been developed, wherein a convolutional stage extracts low-level features while a transformer stage analyzes higher-level context. Research on deepfake detection consistently indicates that hybrid models represent a promising avenue, as they combine the inductive biases and efficiency of convolutional neural networks (CNNs) with the modelling flexibility of transformers. Global transformers, such as ViT-like models, exhibit a quadratic cost relative to the number of patches, rendering them less attractive for practitioners constrained by practical computational and energy limitations.

Swin-Transformer-based approaches: Swin-Transformer models [9] mitigate scalability challenges by confining self-attention to non-overlapping windows and employing shifted windows to facilitate interactions across windows. This results in a hierarchical feature pyramid characterised by linear complexity relative to image size. Swin has been utilised in various deepfake-related applications, such as serving as a backbone for thumbnail-based video detectors [16] or as a component in more intricate multi-branch architectures [17]. Recent studies have investigated Swin-based detectors aimed at enhancing cross-manipulation robustness and cross-dataset generalisation, demonstrating that local-global reasoning within a hierarchical transformer can improve synthetic face detection [18]. Many current Swin-based methods either operate at high input resolutions with substantial backbone sizes or are integrated within larger frameworks that incur a high overall computational cost.

Positioning and scope of this work: This study concentrates on GAN-based synthetic faces and architectures that can be trained and deployed with limited computational resources. Swin-Transformer variants represent an intermediary between CNNs and global ViTs, providing enhanced contextual reasoning compared to CNN-only baselines while maintaining a significantly lower computational cost than traditional ViTs. Swin-based configurations (Swin Base, Swin Transfer, and Hybrid EfficientNet+Swin) aim to provide a compact, reproducible framework for practitioners who require practical, resource-efficient detectors rather than a singular optimally large model.

Prior research indicates that deep and complex architectures achieve high performance on various deepfake benchmarks. Our contribution demonstrates that well-designed Swin-based and hybrid models can be comparable or exceed robust CNN baselines on a GAN-based dataset without the need for large-scale training or advanced hardware. This establishes our method as a robust, resource-efficient advancement in scalable synthetic face forensics.

3 Method

The objective is to develop an image-only detector capable of differentiating between authentic faces and GAN-generated synthetic faces while adhering to practical computational limitations. This section first formalises the problem and the design principles that inform our choices, followed

by a description of the instantiation of Swin-Transformer backbones and their integration with EfficientNet in our hybrid architecture.

3.1 Problem formulation and design principles

Let $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ represent a cropped RGB face image, and let $y \in \{0,1\}$ denote its label, where $y = 1$ indicates a GAN-generated (synthetic) face and $y = 0$ indicates a real face. A detector with parameters θ assigns a score $p_\theta(\mathbf{I}) \in [0,1]$ to input $\mathbf{I}$, interpreted as the probability that $\mathbf{I}$ is synthetic.

We train all models as binary classifiers using standard binary cross-entropy loss on a training set $\mathscr{D}$:

$$\mathscr{L}(\theta) = -\frac{1}{|\mathscr{D}|} \sum_{(\mathbf{I},y)\in\mathscr{D}} \big[y \log p_\theta(\mathbf{I}) + (1-y) \log\big(1 - p_\theta(\mathbf{I})\big)\big]. \tag{1}$$

The complete 140K Real and Fake Faces dataset contains 140,000 images, but we work with a balanced subset: 25,000 training images, 5,000 validation images, and 5,000 test images for this binary real-fake classification task (see Section 4). The image quality and diversity in this subset are sufficient for building a robust discriminator. Using all 140,000 images would yield only marginal improvements while substantially increasing computational and environmental costs. This choice reflects our core design philosophy.

We establish three guiding principles for our design. First, we focus on a GAN framework within an image-only setting: the detector trains on StyleGAN-generated synthetic faces and operates exclusively with still images, not using temporal information or extra metadata. Second, we aim to build architectures that are both context-aware and efficient, going beyond local CNN filters by incorporating long-range interactions while keeping memory footprint and FLOP count much lower than global vision transformers. Third, rather than proposing a single large model optimized for one benchmark, we present a compact set of Swin-based configurations that can be trained on standard hardware and adapted to other synthetic face datasets with minimal tuning, providing a resource-efficient framework for practical deployment.

The choice of Swin-Transformer backbones and their integration with EfficientNet-B0 follows from these principles, which we discuss next.

3.2 Swin-Transformer backbone for face forensics

The Swin Transformer [9] offers a hierarchical vision backbone that substitutes global self-attention with local, window-based attention utilising shifted windows. This paper summarises the components utilised and their specific applications in face forgery detection.

Given an input image $\mathbf{I}$, Swin initially divides it into non-overlapping patches of size $P \times P$ and subsequently linearly projects each patch into a C-dimensional token. Tokens undergo a series of processing stages. The stages comprise:

1. *Window Multi-Head Self-Attention (W-MSA)*, which computes self-attention independently within fixed-size windows (e.g., $M \times M$ patches).

2. *Shifted Window Multi-Head Self-Attention (SW-MSA)*, which applies the same operation after cyclically shifting the window partition, enabling cross-window interactions.

3. *Feed-forward MLP blocks* with GELU activations and residual connections.

4. *Patch merging* layers between stages, which halve the spatial resolution and increase the channel dimension to build a multi-scale representation.

The above design produces a feature pyramid that effectively captures fine-grained local structures in the initial stages and more semantic, global information in the deeper stages. In the context of face forensics, this approach is advantageous: early layers can effectively model skin texture, boundary artefacts, and local inconsistencies, whereas deeper layers can assess global facial coherence (e.g., consistency among the eyes, mouth, and background) at a cost that scales approximately linearly with image size. In all Swin-based models, the final token sequence undergoes global average pooling and is subsequently mapped to a scalar logit, which is processed through a sigmoid function to yield $p_\theta(\mathbf{I})$.

3.3 Model variants

We instantiate the above backbone in three ways, shown schematically in Figure 1. The variants differ in how much prior knowledge they exploit (from scratch vs. pretraining) and how they combine convolutional and transformer-based reasoning.

3.3.1 Swin Base: task-specific Swin from scratch

The Swin Base model is a compact Swin-Transformer designed specifically for GAN-based face detection and trained from scratch on a subset of our dataset. We employ a low input resolution (e.g., 128×128) and small window sizes to maintain a reduced parameter count and FLOPs, while still leveraging hierarchical attention. This variant fulfils two functions. Initially, it serves as a baseline for Swin that does not utilise ImageNet pretraining, enabling an assessment of Swin's capability to develop effective artefact detectors from GAN data. Additionally, it serves as a robust benchmark for assessing the comparative advantages of transfer learning and hybridisation.

3.3.2 Swin Transfer: ImageNet-pretrained Swin-Tiny and Swin-Small

The Swin Transfer configuration employs pretraining on ImageNet-1K. We utilise two standard backbones, Swin-Tiny[2] and Swin-Small,[3], resizing all input images to 224×224. The original ImageNet classification head is substituted with a single sigmoid unit, and all layers, including the Swin backbone, undergo fine-tuning on the deepfake dataset.

Analysing Swin-Tiny and Swin-Small facilitates an examination of the accuracy-efficiency trade-off present in the Swin family. Swin-Small possesses an increased number of layers and

[2]https://huggingface.co/microsoft/swin-tiny-patch4-window7-224

[3]https://huggingface.co/microsoft/swin-small-patch4-window7-224

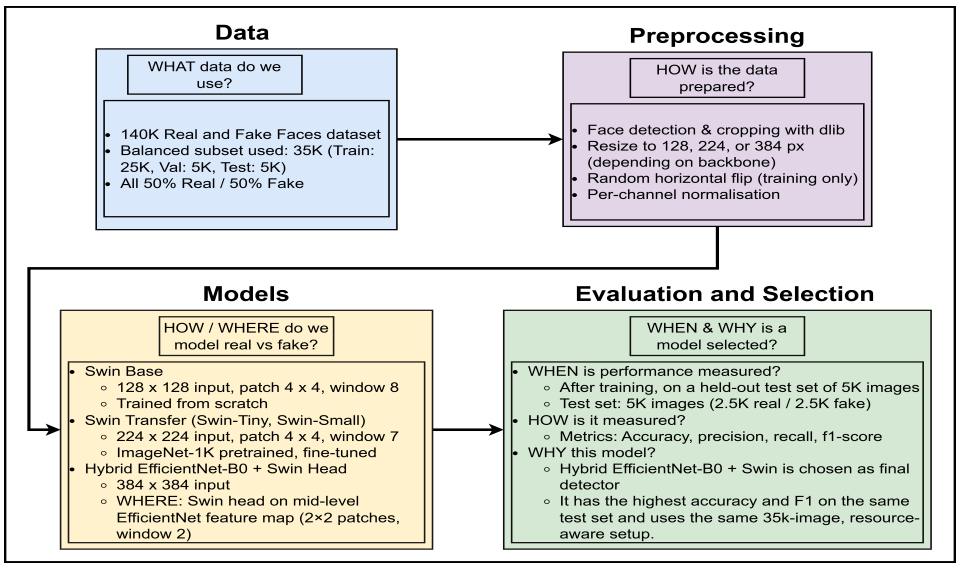


Fig. 1. Proposed deepfake detection pipeline, organised according to the *what/how/where/when/why* framework.

broader channels, resulting in elevated computational costs. Our experiments in Section 4 demonstrate that, for a binary GAN-based task, the smaller Swin-Tiny outperforms its larger counterpart, thereby supporting the use of compact backbones in this context.

3.3.3 Hybrid EfficientNet and Swin integration

The primary architecture employed is a hybrid model that combines EfficientNet-B0 with a Swin-Transformer head. The design is inspired by the observation that convolutional networks excel at capturing fine-grained local artefacts, while Swin-Transformers are more adept at modelling broader facial context. The aim is to utilise both types of cues while maintaining a lightweight overall model. Each incoming face is processed at a resolution of 384×384 using an EfficientNet-B0 backbone pre-trained with ImageNet-1K weights. We extract two complementary representations from this backbone: a mid-level feature map from an intermediate block[4], which retains spatial detail and rich texture information, and a final feature map from the last convolutional layer, encoding higher-level semantics in a more compressed spatial form. The mid-level feature map is divided into small patches (e.g., 2×2), linearly embedded, and processed through a shallow Swin stack with a limited window size and number of heads. The Swin head enhances convolutional features by modelling interactions within and between facial regions at a lower spatial resolution. The final EfficientNet feature map undergoes global average pooling to generate a compact CNN descriptor.

[4] `block6a_expand_activation`

The outputs from these two branches are subsequently integrated in a straightforward fusion stage. Global average pooling is applied to the Swin output to generate a single vector, which is subsequently concatenated with the pooled EfficientNet descriptor. The joint representation is subsequently processed by a fully connected layer utilising sigmoid activation to yield $p_\theta(\mathbf{I})$. This hybrid design offers three practical advantages for GAN-based face forensics. Initially, it integrates complementary information: EfficientNet-B0 focuses on local artefacts and noise patterns, whereas the Swin head incorporates global context and cross-region consistency. Secondly, Swin processes a mid-level feature map instead of the raw input image, resulting in a reduced number of tokens and maintaining a low cost of self-attention. Third, we initialise EfficientNet-B0 using ImageNet-1K weights and train the Swin head on its features, enhancing data efficiency and aiding the model in generalising from a relatively small training set.

Section 4 demonstrates that the EfficientNet+Swin hybrid significantly surpasses both the pure Swin variants and the prior CNN-only baseline on the same dataset, while maintaining a compact form suitable for deployment within standard computational constraints.

4 Experiments and Discussion

4.1 Dataset

All models are evaluated using the 140K Real and Fake Faces dataset[5], comprising 70,000 authentic faces and 70,000 faces generated by StyleGAN. The real images are obtained from Flickr and DeepFake Detection Challenge (DFDC) [19], encompassing a diverse array of ages, demographics, and accessories. In accordance with previous research [14], faces are detected and cropped utilising `dlib` [20], subsequently resized to dimensions of 128, 224, or 384 pixels based on the selected backbone. Random horizontal flipping and per-channel normalisation are applied during training. In accordance with our resource-aware design, we utilise a balanced subset of the dataset for a binary real-fake classification task. We utilise 25,000 images for training, 5,000 for validation, and 5,000 for testing. Each subset comprises 50% real and 50% synthetic faces, resulting in 12,500 real and 12,500 synthetic images for training, and 2,500 real and 2,500 synthetic images for both validation and testing. The image quality and diversity within this subset are adequate for developing a robust discriminator, while utilising the complete set of 140k images would yield diminishing returns at a significantly increased computational expense.

4.2 Implementation details

All models utilise the same balanced splits and are trained with the Adam optimiser. The Swin Base model functions with 128×128 inputs, utilising a patch size of 4×4 and a window size of 8. It is trained from scratch over 200 epochs, employing a learning rate of 2×10^{-5} and a batch size of 32. Swin-Tiny and Swin-Small utilise 224×224 inputs, employing a patch size of 4 and a window size of 7. Both models are fine-tuned from ImageNet-1K pretraining, with a learning rate of 5×10^{-4}, batch sizes of 128 and 64, and training durations of 100 and 150 epochs, respectively.

[5] https://www.kaggle.com/datasets/xhlulu/140k-real-and-fake-faces

The hybrid EfficientNet+Swin model handles inputs of size 384×384. The Swin head is utilised on mid-level EfficientNet feature maps through the application of 2×2 patches and a window size of 2, employing 8 attention heads. The model undergoes training for 20 epochs, utilising a learning rate of 0.5 and a batch size of 8. Figure 2(a,b) illustrates the training dynamics. The model converges quickly, with the training loss decreasing from 0.97 to 0.20 in the first ten epochs, while the validation loss remains around 0.24 after nine epochs. Additionally, we can see that accuracy reaches 97% by 11 epochs, compared to 99.8% peak validation accuracy at epoch 17. Results are reported on the held-out 5,000-image test set unless stated otherwise.

4.3 Quantitative results

Table 1 reports precision, recall, F1-score and accuracy on the test set, and compares our Swin-based architectures with the EfficientNet-based ELA+CNN baseline of Kumar et al. [14]. The Swin Base model, trained from scratch, demonstrates performance comparable to the baseline, suggesting that a compact Swin configuration can rival a meticulously designed CNN approach on this GAN-based dataset without the need for external pretraining. The introduction of ImageNet-1K pretraining using Swin-Tiny results in a notable enhancement, with accuracy increasing from 88.50% (Swin Base) to 91.13%, and the F1-score increasing from 88.00% to 91.14%.

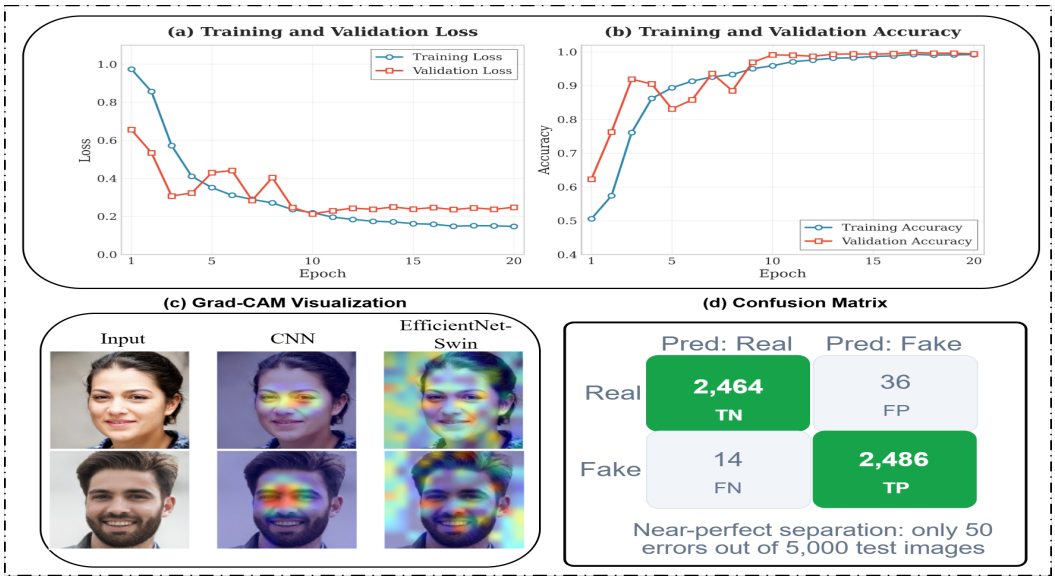


Fig. 2. Performance over proposed EfficientNet-Swin hybrid Transformer model: (a) loss plots, (b) accuracy plots (peak: 99.8%), (c) Grad-CAM visualizations comparing CNN and hybrid model attention regions, and (d) confusion matrix on 5,000 test images (recall: 99.44%).

Swin-Small exhibits lower performance than Swin-Tiny, despite its greater capacity and ex-

Table 1: Evaluation results on the 140K Real and Fake Faces test set (5,000 images).

Model	Epochs	Precision	Recall	F1-score	Accuracy
ELA + CNN [14]	20	86.76	88.40	87.57	87.46
Swin Base (from scratch)	200	88.00	87.00	88.00	88.50
Swin-Tiny (transfer learning)	100	88.00	94.00	91.14	91.13
Swin-Small (transfer learning)	150	87.00	91.00	89.13	89.14
Hybrid EfficientNet+Swin	20	98.57	99.44	99.00	99.00

tended training duration, achieving an accuracy of 89.14% and an F1-score of 89.13%. This indicates that merely increasing model size does not inherently lead to improved performance in this binary GAN-based task, thereby supporting the argument for compact backbones.

The hybrid EfficientNet+Swin architecture achieves superior performance, with 98.57% precision, 99.44% recall, a 99% F1-score, and 99% accuracy (see Figure 2(d)). Compared to the ELA+CNN baseline, this represents an increase of over 11 percentage points in both accuracy and F1 Score, achieved with a relatively small model trained on a subset of the complete dataset. The findings substantiate our primary assertion that integrating convolutional features with shifted-window self-attention is notably effective for GAN-based synthetic face detection within realistic computational constraints.

4.4 Qualitative analysis and limitations

Grad-CAM visualisations for the hybrid model indicate attention to both local artefacts, including skin texture irregularities and boundary inconsistencies, as well as broader facial regions. Conversely, a baseline utilising only EfficientNet tends to focus on more localised patches. The qualitative evidence aligns with our architectural motivation, indicating that the Swin head contributes authentic contextual reasoning rather than simply re-weighting features from the CNN backbone (see Figure 2(c)).

This evaluation is limited to one dataset and one generator family, specifically StyleGAN. Recent studies demonstrate that both CNN and transformer-based detectors may encounter significant performance declines when utilised on unseen datasets or alternative manipulation pipelines [21]. Our results indicate the potential of Swin-based and hybrid architectures within controlled GAN-based conditions, rather than serving as evidence of complete real-world robustness. Subsequent steps should involve cross-dataset and cross-generator evaluations, as well as the extension of the proposed hybrid design to video deepfakes, which present additional challenges due to temporal consistency and compression artefacts.

5 Conclusion and Future Scope

This study examines Swin-Transformer architectures for GAN-based synthetic face detection within defined computational constraints, utilising StyleGAN-generated faces from the 140K Real and Fake Faces dataset. A compact Swin model trained from scratch achieved performance comparable to a robust ELA+CNN baseline. Swin-Tiny, pretrained on ImageNet-1K, yielded a moderate improvement, while the larger Swin-Small variant did not provide additional benefits. The EfficientNet-B0+Swin hybrid demonstrated 99% accuracy and a 99.44% recall on a balanced test set of 5,000 images, while being trainable on standard hardware. This suggests that integrating mid-level convolutional features with shifted-window self-attention represents a viable design approach for this binary GAN configuration.

This study concentrates on a specific context: static images, a singular category of GAN-based generators, and a single benchmark dataset. The results should be interpreted as controlled evidence concerning architectural choices, rather than as a comprehensive solution for detecting synthetic media in real-world scenarios. It is necessary to replicate the study across various datasets and generator families, encompassing both diffusion and autoregressive face models. One alternative method involves shifting from images to video, utilising temporal and compression-related cues. A third direction involves clarifying the relationship between model selection, data scale, and energy consumption, with the aim of establishing resource-aware benchmarks for synthetic media forensics.

References

[1] Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. Advances in neural information processing systems. 2014;27.

[2] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Advances in neural information processing systems. 2020;33:6840-51.

[3] Tian K, Jiang Y, Yuan Z, Peng B, Wang L. Visual autoregressive modeling: Scalable image generation via next-scale prediction. Advances in neural information processing systems. 2024;37:84839-65.

[4] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2019. p. 4401-10.

[5] Rossler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Nießner M. Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF international conference on computer vision; 2019. p. 1-11.

[6] Abbas F, Taeihagh A. Unmasking deepfakes: A systematic review of deepfake detection and generation techniques using artificial intelligence. Expert Systems with Applications. 2024;252:124260.

[7] Gupta G, Raja K, Gupta M, Jan T, Whiteside ST, Prasad M. A comprehensive review of deepfake detection using advanced machine learning and fusion methods. Electronics. 2023;13(1):95.

[8] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations; 2021. Available from: https://openreview.net/forum?id=YicbFdNTTy.

[9] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV); 2021. p. 10012-22.

[10] Chang X, Wu J, Yang T, Feng G. Deepfake face image detection based on improved VGG convolutional neural network. In: 2020 39th chinese control conference (CCC). IEEE; 2020. p. 7252-6.

[11] Bondi L, Cannas ED, Bestagini P, Tubaro S. Training strategies and data augmentations in cnn-based deepfake video detection. In: 2020 IEEE international workshop on information forensics and security (WIFS). IEEE; 2020. p. 1-6.

[12] Deng L, Suo H, Li D. Deepfake Video Detection Based on EfficientNet-V2 Network. Computational Intelligence and Neuroscience. 2022;2022(1):3441549.

[13] Li H, Zhou J, Li Y, Wu B, Li B, Dong J. Freqblender: Enhancing deepfake detection by blending frequency knowledge. Advances in Neural Information Processing Systems. 2024;37:44965-88.

[14] Kumar N, Pranav P, Nirney V, Geetha V. Deepfake image detection using CNNs and transfer learning. In: 2021 International Conference on computing, communication and green engineering (CCGE). IEEE; 2021. p. 1-6.

[15] Coccomini DA, Messina N, Gennaro C, Falchi F. Combining efficientnet and vision transformers for video deepfake detection. In: International conference on image analysis and processing. Springer; 2022. p. 219-29.

[16] Xu Y, Liang J, Jia G, Yang Z, Zhang Y, He R. Tall: Thumbnail layout for deepfake video detection. In: Proceedings of the IEEE/CVF international conference on computer vision; 2023. p. 22658-68.

[17] Ilyas H, Javed A, Malik KM. AVFakeNet: A unified end-to-end Dense Swin Transformer deep learning model for audio–visual deepfakes detection. Applied Soft Computing. 2023;136:110124.

[18] Gong LY, Li XJ, Chong PHJ. Swin-Fake: A consistency learning transformer-based deepfake video detector. Electronics. 2024;13(15):3045.

[19] Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M, et al. The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:200607397. 2020.

[20] King DE. Dlib-ml: A machine learning toolkit. The Journal of Machine Learning Research. 2009;10:1755-8.

[21] Nadimpalli AV, Rattani A. On improving cross-dataset generalization of deepfake detectors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2022. p. 91-9.