Electronic Letters on Computer Vision and Image Analysis 25(2):130-151, 2026

DeepFake Image Detection using Swin Transformer with Attention-guided Feature Fusion and Explainable AI Techniques

Vineela Krishna Suri* and Dr. G.V.S.N.R.V. Prasad+

* *Research Scholar, Department of CSE, JNTUK, Kakinada, Andhra Pradesh, India*
+ *Professor & Director(PGCRD), Department of CSE, SRGEC, Gudlavalleru, Andhra Pradesh, India*

Received 25st of November 2025; accepted 18th of June 2026

Abstract

Deepfakes are synthetic media created using advanced machine learning (ML) techniques, and they pose serious societal challenges by spreading misinformation and undermining trust in the media. With the increasing sophistication of deepfake technologies, distinguishing between genuine and synthetic media has become increasingly difficult. This paper presents a robust deepfake image detection framework using the Swin-B Transformer with an attention-guided hierarchical feature fusion mechanism (AG-Swin). The proposed model extracts multi-stage representations from the hierarchical stages of Swin transformer and dynamically integrates them using an attention module, enabling the model to focus on the most informative feature levels for detecting manipulation cues. To evaluate the efficacy of the framework, a hybrid dataset is constructed by combining real images from the FFHQ dataset with synthetically generated fake images from a publicly available Kaggle dataset, which simulates real-world media scenarios. The proposed AG-Swin model achieved a high accuracy of 98% on the test set, demonstrating a superior generalization to real and synthetic visual data. Grad-CAM based visualization is applied to highlight the spatial segments of the image that the model focuses on during classification, providing insight into the decision-making process. The proposed framework contributes to the improvement of the verification of media authenticity, mitigating the spread of fake news, and strengthening digital trust and safety in modern information ecosystems.

Key Words: Deepfake detection, Vision Transformers (ViT), Generative Adversarial Networks (GAN), Convolutional Neural Networks (CNN), Attention-based fusion

1 Introduction

The pioneering work of Christoph Bregler, Michele Covell, and Malcolm Slaney in developing the video rewrite program in 1997 laid the foundation for Deepfakes. It is marked as the first automated approach to facial reanimation, leading to the more sophisticated deepfake technologies that we encounter today. The term deepfake has since come to encompass both these fake media and the specific method of their creation.The manipulation of digital images and videos has been a promising technique for many years and is done through the use

Correspondence to: vineela.suri@gmail.com

Recommended for acceptance by Angel D. Sappa
https://doi.org/10.5565/rev/elcvia.2480
ELCVIA ISSN:15108-5097
Published by Computer Vision Center / Universitat Autònoma de Barcelona, Barcelona, Spain

of visual effects. However, the inception of deep learning architectures like Generative Adversarial Networks (GANs) has led to a marked increase in the realistic nature of fake content and the ease with which they can be generated [1, 2]. These AI-generated media, commonly referred to as deepfakes, are easily produced using artificial intelligence tools. Through the use of deep learning algorithms, it is now possible to generate fabricated videos and images that are virtually challenging to differentiate from real ones.

Since their emergence in 2017, there has been a proliferation of open source deepfake generation methods and tools, resulting in a growing number of synthetic media clips. FakeApp [3], FaceSwap [4] and FaceShifter [5] are some of the prominent publicly accessible deepfake implementations. Throughout history, there have been numerous instances in which deepfakes have been utilized for malicious purposes, such as creating political tensions, staging fake terrorism events, producing pornographic content, and blackmailing.

As an example, University of Maryland researchers created a bot by the name of Katie Jones on the LinkedIn platform. In order to add more authenticity to this fake profile, a profile picture generated using deepfake technologies was also added, and the bot was then left to make connections on the platform. In just two months, this profile was connected with two US state senators and one Department of Defense (DoD) personnel [6]. It is estimated that, bots make up 65% of the Internet currently. With the advent of deepfake technologies, it might become even more difficult to detect fake identities on the Internet. A study by the AI firm Deeptrace revealed that only in the month of September 2019 they identified 15,000 deepfake videos online, nearly double the increase over the previous nine months. Of these videos, 96% were pornographic in nature, among which 99% of those targeted female celebrities, overlaying their faces on the bodies of some porn stars [7]. This marked the significance of robust and reliable methods for detecting deepfakes thus preventing their abuse.

At the heart of deepfakes lies the intricate process of facial swapping, a technique that allows for seamless overlaying of facial images of a target individual onto videos of a source individual as shown in Fig.1 This mesmerizing artistry results in videos of the target individual appearing to say or do things that were originally performed by the source individual as shown in Fig.2. An example of the influence of deepfakes can be seen in the circulation of such videos related to the Russian-Ukraine conflict, including a deepfake of Russian President Vladimir Putin declaring peace.



Figure 1: Original Image and Corresponding Deepfake Image



Figure 2: An example Deepfake video

The rapid growth of generative models has led to the extensive creation and spread of deepfake images. Although these technologies have creative and commercial applications, they also carry serious risks such as

the spread of false information, identity theft, and a decline in public confidence in digital content [8]. A crucial challenge in the areas of computer vision and social computing is the detection of these misleading media content. The increasing societal impact of deepfake images is another factor driving this study. As deepfakes significantly influence public opinion, identity verification, and integrity of social media, strong detection techniques contribute directly to the broader goals of social computing there by improving the digital trust and authenticity of the content.

Deepfake manipulations are often done at multiple levels, including low-level textual artifacts, mid-level structural deformation, and high-level semantic irregularities. Conventional Transformer-based approaches often consider only the final-layer features may miss subtle forgery inconsistencies captured in earlier layers and ignore multi-scale manipulation artifacts. To address this limitation, this work proposes an Attention-Guided Hierarchical Swin Transformer (AG-Swin) model that fuses transformer features at multiple stages using a learnable attention mechanism.

1.1 Objectives

The key objectives of this research are as follows:

1. To develop an Attention-Guided Hierarchical Swin Transformer (AG-Swin) framework for deepfake image detection that uses multi-stage transformer representations to identify manipulative cues.
2. To design a novel stage-wise attention mechanism that dynamically assigns weights to the hierarchical features extracted from different Swin-B stages
3. To create a hybrid dataset that combines real images from the FFHQ dataset with synthetically generated fake images from a publicly available Kaggle dataset, representing diverse and challenging real-word scenarios.
4. To demonstrate that the proposed approach achieves better generalization to heterogeneous data and outperforms several other existing methods on the benchmark datasets.
5. To analyze and interpret the model's decision making process through attention visualization and stage-wise analysis of features, and evaluate the effectiveness of the hierarchical attention fusion approach.

1.2 Major Contributions

The contributions of this work can be summarized as follows:

1. **Attention-Guided Hierarchical Feature Fusion:** The proposed model introduces an attention-guided fusion mechanism that dynamically combines features extracted from multiple hierarchical stages of the Swin-B transformer.
2. **Utilization of Multi-Stage Transformer Features:** Unlike conventional methods that rely on features extracted from a single layer, AG-Swin extracts and integrates features from four hierarchical stages, capturing subtle manipulated artifacts such as pixel-level inconsistencies, boundary irregularities, semantic anomalies, and structural distortions introduced at different layers.
3. **Dynamic Attention based fusion:** Instead of using static feature fusion strategies such as concatenation or averaging, AG-Swin employs a learnable attention mechanism to dynamically assign weights to each stage representation. These attention weights provide information on which hierarchical feature level contributes the most to the final prediction.
4. **Explainable AI (XAI) based model interpretation:** To enhance the interpretability of the model, GRAD-CAM visualization is applied to highlight the spatial regions of the image at different stages that influence the model's decision.

The subsequent sections are organized as follows: Section 2 reviews existing methodologies for detecting deepfakes. Section 3 explores the data set and Section 4 presents the overall architecture and detailed operational workflow of the proposed approach to detect deepfake images. Section 5 provides a comprehensive review of the results obtained using various evaluation metrics. Section 6 summarizes the paper and provides directions for future research.

2 Literature Survey

In the earliest methods, engineered features obtained from the inconsistencies and artifacts of fake images were used for classification.

2.1 GAN Artifact-Based Methods

In the paper [9] Xuan et al. make use of an image pre-processing step, where they apply Gaussian blur and Gaussian noise to the images in order to extract low-level high frequency cues of GAN generated images. This improves pixel-level statistical similarity between authentic and fraudulent images, allowing forensic classifiers to learn more intrinsic and valuable characteristics compared to existing techniques or networks. Matern and Riess et al. [10], identified several distinguishing features such as missing reflection in the eye, loss of geometry and shape, difference in color of two eyes of the same person, etc., and then created a detection pipeline which segments all the various important features like iris, face crops, etc. These segmented and extracted features were tested for authenticity and used for classification. They achieved their best accuracy of 83.8% with eyes as the primary mode of classification.

2.2 CNN-Based Methods

Hsu et al. introduced a two-phase deep learning method for the detection of deepfakes [11]. It makes use of the Siamese Network architecture based on the Common Fake Feature Network (CFFN) for feature extraction. The Siamese network takes pairs of images along with a pairwise label during training. When both images within a pair are genuine, the label is designated as 1, indicating a ”real-real” pairing. Conversely, if the images belong to different classes, meaning that one is authentic and the other one is manipulated, the label is assigned as 0, reflecting a ”fake-real” or ”real-fake” pairing. During the training phase, a pairwise label is provided, and the model learns the distinguishing features of each class. Further, it can also understand the representative characteristics of real and fake images. Once training is completed, a base image, whose class is known (real or fake), is passed along with the input image given by the user. Depending on the output obtained and the type of base image, we can determine the class of input image. The major setback of this approach is that it uses a non-standard dataset which is generated only using GANs. Hence, cannot work well with other deepfake implementations and can provide results only for GAN generated images.

Another method using pairwise learning was proposed by Zhao et al. for deepfake detection [12], where they made use of the self-consistency of local source features noticed in deepfakes, which are content-independent and spatially local information of images. These local features can be the result of imaging pipelines or encoding methods used in different image synthesis approaches. A CNN-based model is used to learn pairwise self-consistency. The hypothesis behind this paper is that the original image will have consistent features in different localities and a modified image will be inconsistent in generated and original portions of the image. The model is trained using representation learning technique, where, the model, instead of learning about the overall features, looks only at parts of the images. This approach differentiates between genuine and manipulated images by penalizing feature vector pairs with low similarity scores within the same image. Conversely, it also penalizes pairs of feature vectors from different images for exhibiting a higher similarity score. In this paper, cosine similarity is used as an evaluation metric. The feature maps being learned were then provided as input to a classification method for deepfake detection. Although this model performs well on locally modified

images, it struggles with fully generated fake images, where features remain consistent at all positions within the image.

Zhou et al. propose a different approach [13], in which they make use of a two-stream neural network. Instead of pairwise learning to understand latent relations between pairs of images, one stream is used to detect faces and determine their genuineness. The other stream is used to detect traces which might be left due to any steganographic tampering. They make use of FaceSwap and SwapMe apps available on the App Store in iOS to create a custom dataset. However, the customized dataset used in their research is outdated and the results are not updated with recent GAN-based generation methods. Moreover, the evaluation was primarily theoretical and, when tested against a standard Celeb-DF dataset, achieved an AUC score of only 55.7%.

A simpler approach was introduced by Xu Chang et al. [14], where they implemented a modified VGG-16 network called NA-VGG (N: Noise, A: Augmentation). Here, the input is first processed with an SRM filter that highlights the image noise. SRM stands for Style Recalibration Module, which is a fixed weight matrix acquired by training a shallow neural network to modify the image to a specific type, for example, from RGB to HSV etc. In this paper, the SRM layer extracts and highlights latent artifacts in the image which go unnoticed in the case of RGB. After generating this version of the input, augmentation is performed, where more images were generated by flipping, rotating, adjusting the brightness etc. of the image. The augmented images are passed to a standard VGG-16 CNN which contains 16 layers in total: 13 convolution layers and 3 fully connected (FC) layers. In order to downsize the resolution, 5 pooling layers are also used intermediately after convolutions. In their research, they used the Celeb-DF dataset which is created by swapping faces in the CelebA-HQ dataset.

Theerthagiri introduced an approach using InceptionNet [15], a CNN based architecture to capture multiscale features to identify deepfakes. In their study, they used a public dataset downloaded from kaggle consisting of 401 videos. Data Augmentation is applied to generate 3745 images from the dataset. The efficacy of the model was evaluated using various metrics derived from the confusion matrix, achieving an accuracy of 93% in recognizing deepfake images and videos. Another method was introduced by pasupuleti [16], using a Custom Densenet architecture, which is a deep residual neural network. To classify images as *real* or *fake* they employed a CNN based architecture trained using a binary cross-entropy function. The performance of the model is evaluated using various metrics calculated from the confusion matrix.

2.3 Transformer-Based Methods

Ghita et al. proposed an approach for detecting deepfakes using Vision Transformers(ViT) [17]. This ViT based detection model is trained with a dataset consisting of 40000 samples including both real and deepfake images collected from the Kaggle dataset. The model achieved a score of 0.899 on a dataset consisting of 40000 samples. Their initial experiments emphasized the need for a large dataset for training and the fast convergence of the model. Compared to other baseline deepfake detection techniques, the performance of the ViT model was consistent with existing research, indicating its potential for further investigation.

Authors of [18] suggested a shallow vision transformer for recognizing deepfakes, incorporating both an attention mechanism and multi-head attention module. The attention mechanism highlights the critical parts of images, while the multi-head attention module determines how much attention should be allocated to local-level features. Finally, the softmax layer classifies an image as *real* or *not real*. Experiments conducted on two datasets-Real Fake Face (RFF) and Real and Fake Face Detection (RFFD) datasets show that the model achieves higher accuracy of 0.92 on RFF and 0.89 on RFFD datasets. These results outperforms most of the existing models such as GoogleNet, XceptionNet, ResNet50 and other baseline vision transformers. Furthermore, the model demonstrated an accuracy of 0.90 even when trained only on 50% of the RFF dataset.

Shukla et al. introduced a transformer based architecture for the classification of fake and genuine user profiles on social media platforms [19]. In contrast to conventional ML based approaches that depend on manual feature engineering, their model employs the attention mechanism present in transformers to capture complex correlations observed in user profile data. By using the self-attention mechanism present in the transformers,

their model captures long-range dependencies and subtle patterns found in high-dimensional user data.

2.4 Research Gaps

- Most of the existing approaches utilize standard versions of visual transformers or CNN architectures, like EfficientNet, ResNet variants, which may not capture global contextual clues present in high-quality deepfake images.
- Many of the existing deepfake detection models rely on the features extracted from a single layer or a final layer of the neural network and may overlook the subtle manipulations that appear at different levels of representation. Using only the final layer may result in the loss of valuable information present in earlier stages.
- They often utilize a single-source benchmark dataset or domain-specific image collections, which may not fully reflect the diversity and complexity of real-world content shared on social networks.
- Although Swin Transformers have been used in some earlier studies, the Swin-B - which is a larger and more expressive variant of the Swin Transformer, has not been investigated for deepfake image detection.
- In models that use multiple feature levels, features are mostly combined using simple concatenation or averaging. These static fusion methods consider all the features equally and cannot determine which feature level is more informative for the given image.

These days Deepfakes have become very realistic, making local artifacts difficult to detect. In this context, a transformer-based model is needed that will capture global dependencies and can model long-range dependencies of images [18]. The self-attention mechanism in the transformer architecture helps focus attention on minute anomalies that CNNs may overlook. When fine-tuned, they offer better generalization and are robust to the new manipulation methods available. The most informative feature representations are not adaptively focused by traditional models that apply static feature aggregation techniques like simple concatenation or averaging.

To mitigate these issues, this study proposes a novel attention based deepfake detection model that uses the Swin-B transformer, a high-capacity hierarchical vision transformer, applied to a combined dataset consisting of real images from the FFHQ dataset and fake images sourced from a publicly available Kaggle dataset. This hybrid dataset simulates the real-world media environment, improving the model's ability to generalize to unseen manipulations.

3 Dataset Description

To evaluate the effectiveness of the model and for better generalization, a hybrid dataset was constructed combining images collected from two different datasets. Real images were collected from FFHQ Dataset [20] and DeepFake images from a dataset available on Kaggle [22]. Flickr-Faces-HQ (FFHQ) is an image dataset consisting of 70,000 high-quality PNG images of real human faces, each with a resolution of 1024 × 1024, originally created as a standard for GANs [20]. The images in the data set were originally extracted from Flickr.com, an online platform for hosting images and videos, and cropped using the Dlib library [21]. The second dataset [22], is an image dataset consisting of 95,000 JPG images of forged human faces each with a resolution of 256 × 256. Images extracted from both datasets were combined and organized into two main directories: real and deepfake as shown in Figure 3. This combined dataset simulates the variability seen in real-world social media content. Samples of the real and deepfake images from the dataset are shown in Fig. 4a and Fig. 4b.

```
Dataset/
├── real/
│   ├── img1.jpg
│   └── ...
├── deepfake/
│   ├── df1.jpg
│   └── ...
```

Figure 3: Structure of the dataset



(a)



(b)

Figure 4: Example Real (a) and Deepfake (b) images from the dataset.

4 Proposed Model

The proposed model introduces an Attention-Guided Hierarchical Swin Transformer (AG-Swin) for detecting deepfake images. AG-Swin model uses the Swin-B (Base) Transformer, a mid-range model which is a specific type of Swin Transformer [23] designed to provide a good balance of performance, computational cost, and speed. Transformers have shown immense capabilities and work especially well in tasks where pixel-level details might matter. The proposed architecture utilizes the hierarchical representation capability of the Swin-B transformer and integrates a novel attention-guided stage fusion mechanism to effectively combine multi-scale features. Unlike conventional transformer-based models that use only final stage features, the proposed framework extracts features from multiple hierarchical stages of the Swin backbone. These stage-wise representations capture information such as low-level texture artifacts, mid-level structural patterns, and high-level semantic inconsistencies commonly found in manipulated images. AG-Swin is then deployed for demonstration using the Gradio API, where it can also be tested by other people. The overall workflow of AG-Swin, along its various components, is illustrated in Fig. 5

4.1 Data Pre-processing

All images present in the hybrid dataset are transformed so that they have a fixed size of 224×224 pixels, the expected input image size for the models pre-trained on ImageNet, and are normalized. In order to balance both the classes present in the dataset, and to avoid overfitting, data augmentation is applied on real face images.

4.2 Hierarchical feature extraction using Swin-B

The backbone of the proposed AG-Swin model, is the Swin-B transformer, a hierarchical vision transformer that is pre-trained on ImageNet. The pre-trained weights for this model are obtained from the official weights submitted by Microsoft to the Huggingface platform [24]. It contains roughly 28 million trainable parameters and is pre-trained on more than 14 million images present in the ImageNet dataset. The Swin-B Transformer model contains four blocks called "swin layers", each performing similar operations. In the first layer, the image is broken down into non-overlapping patches of size 4 x 4 pixels resulting in 56 x 56 tokens. Swin divides the 56 x 56 tokens into small windows of size 7 x 7. Unlike conventional Vision Transformers, where multi-head attention is calculated globally for the entire image taking a lot of processing time, Swin transformers apply attention only inside local attention window of size 7 x 7. This approach is known as Window-based Multi-head Self Attention (W-MSA) as shown in Eq(1). In the next layer, the windows were shifted by patch size of 3 supporting cross-window communication called as Shifted Window Multi-head Self Attention (SW-MSA). After each group of Swin blocks, a patch merging operation is performed, where 2×2 neighboring patches are concatenated and linearly transformed there by increasing feature depth while decreasing spatial resolution. The overall architecture of the Swin-B transformer consisting of four stages along with the features captured in each layer, is illustrated in Fig. 6. Each stage produces feature maps with increasing semantic abstraction and decreasing spatial resolution, as summarized in Table 1.

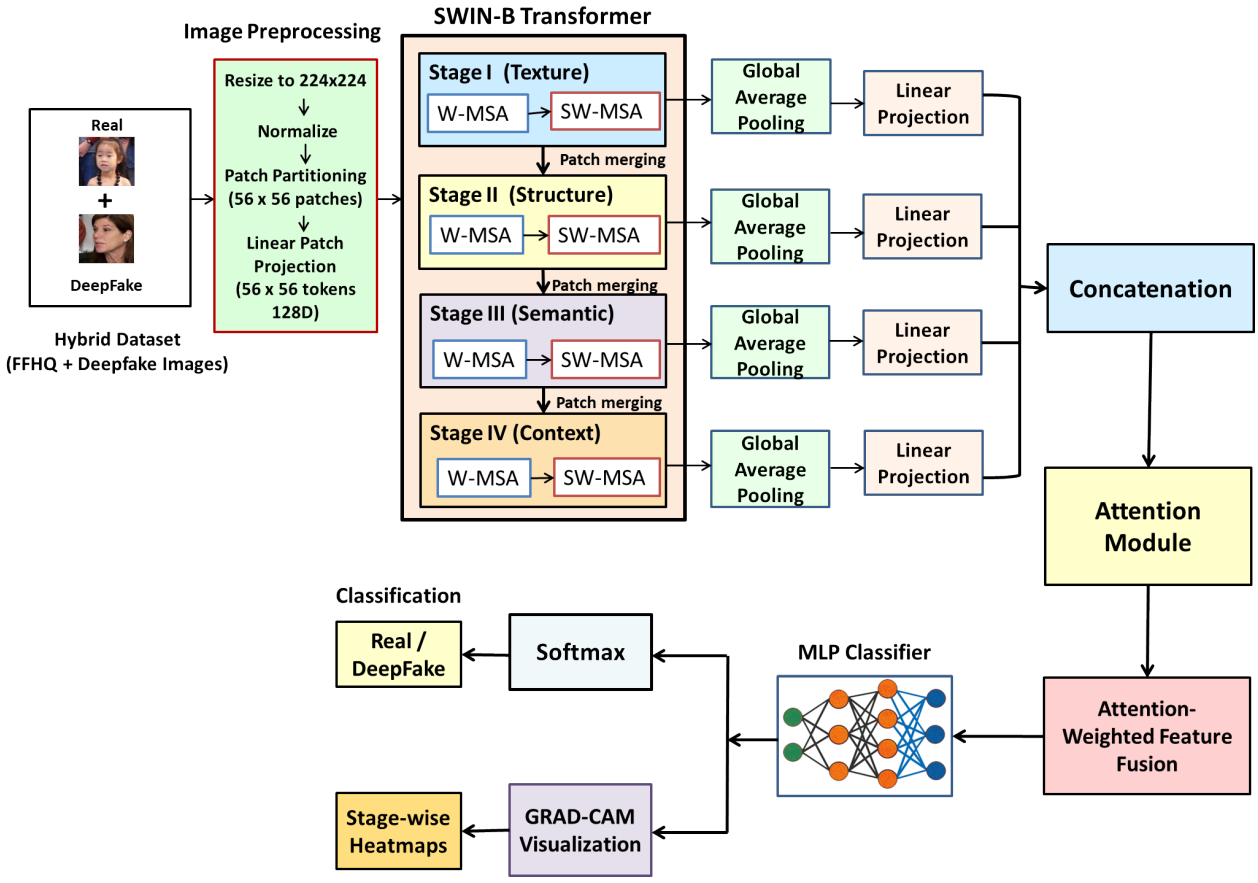


Figure 5: Proposed System architecture

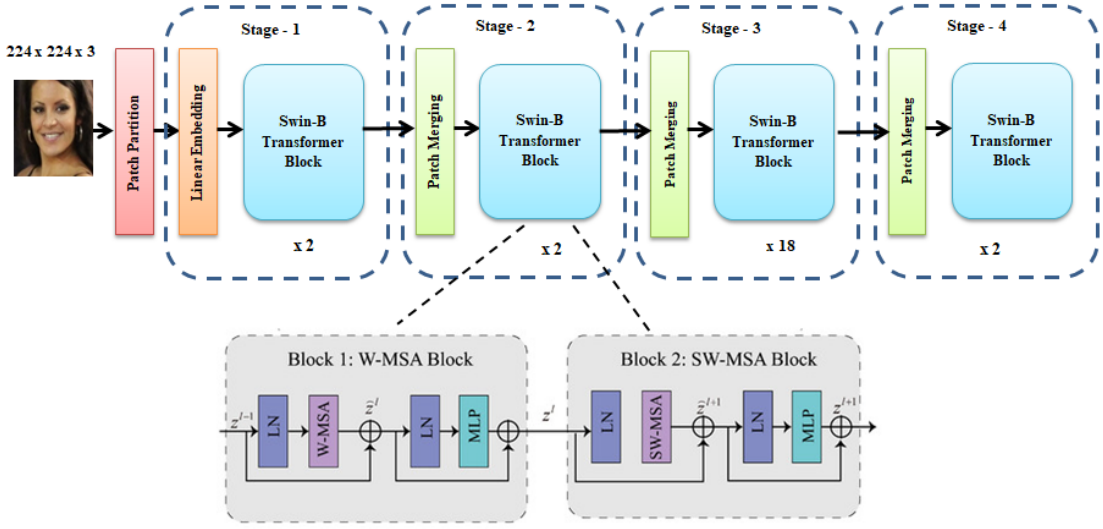


Figure 6: Architecture of Swin-B Transformer

Table 1: Layers inside each Stage of Swin-B Transformer

Stage	Input Size	Layers	Hidden Dimension	Operations	Feature type	Deepfake Feature Representations
1	56×56	2	128	W-MSA + SW-MSA	low-level texture features	Detects pixel level inconsistencies and blending artifacts
2	28×28	2	256	W-MSA + SW-MSA	Structural patterns	Identify boundary irregularities and facial structure distortions
3	14×14	18	512	W-MSA + SW-MSA	Mid-level semantic cues	Detect inconsistencies in facial regions
4	7×7	2	1024	W-MSA + SW-MSA	High-level contextual features	Capture global semantic anomalies

$$MultiHead(Q, K, V) = [head_1, \ldots, head_h]W_0 \tag{1}$$

Where, $head_i$ is self-attention, calculated using Eq(2) and Eq(3)

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2}$$

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

where Queries $Q \in \mathbb{R}^{n \times d_k}$, Keys $K \in \mathbb{R}^{n \times d_k}$, Values $V \in \mathbb{R}^{n \times d_v}$.

Let the feature representations extracted from these stages be denoted as: F_1, F_2, F_3, F_4 where $F_i \in \mathbb{R}^{B \times N_i \times C_i}$, B denotes the batch size, N_i represents the number of tokens at stage i, C_i represents the feature dimension at stage i. To obtain compact feature vectors from each stage, global average pooling (GAP) is applied as shown in Eq.4

$$\hat{F}_i = \text{GAP}(F_i) \tag{4}$$

rs $\hat{F}_i$ have the shape: $\hat{F}_i \in \mathbb{R}^{B \times C_i}$

where B is the batch size, and C_i is the feature dimension at stage i.

Feature vectors $\hat{F}_i$ obtained from the different stages of Swin have different dimensionalities. Directly combining them into a unified representation can lead to inconsistencies in feature space. To overcome this, all features were projected into a unified embedding space using learnable linear transformations as shown in Eq.5. The aligned feature embeddings E_i are obtained by applying learnable linear transformations to the pooled features $\hat{F}_i$:

$$E_i = W_i \hat{F}_i \tag{5}$$

where W_i represents the learnable projection matrix, E_i represents the aligned feature embedding. This ensures that all stage representations share a common dimensional space, enabling effective feature fusion.

4.3 Attention-Guided Feature Fusion

The proposed model (AG-Swin) used attention-guided feature fusion to dynamically combine hierarchical features. Different stages capture different levels of information; it is important to dynamically determine which stage corresponds more to the classification result. The hierarchical features embeddings obtained from the stages of the swin transformer are concatenated using Eq.6.

$$F_{\text{concat}} = [E_1; E_2; E_3; E_4] \tag{6}$$

where E_1, E_2, E_3, E_4 represents the projected embeddings of the four hierarchical stages, each embedding has dimension $d = 256$ and the final vector F_{concat} has dimension $4d = 1024$.

The concatenated feature vector F_{concat} is passed through a light weight attention network that evaluates the relative importance of each stage in determining the final prediction result.

4.3.1 Attention Module

The attention module consists of the following layers:

1. **FC Layer:** The FC layer reduces the dimensionality of the feature vector F_{concat} using Eq.7.

$$H = W_1 \cdot F_{\text{concat}} \tag{7}$$

where $F_{\text{concat}} \in \mathbb{R}^{1024}$, $H \in \mathbb{R}^{64}$, W_1 is the learnable weight matrix. Joint vector H is then passed through layer normalization using Eq.8. Layer normalization is used to stabilize the feature distribution by normalizing the activations across the feature dimension, and to improve training stability.

$$H_{\text{norm}} = LayerNorm(H) \tag{8}$$

2. **LeakyReLU Activation:** The LeakyReLU activation function with leaky slope α = 0.01 is applied to the normalized feature vector H_{norm} (Eq.9), to introduce non-linearity while preserving gradient flow for negative inputs and also to avoid vanishing gradient problem.

$$H_{\text{activation}} = LeakyReLU(H_{norm}) \tag{9}$$

3. **Attention score generation:** The final layer of the attention network projects the feature vector $H_{\text{activation}}$ into a four-dimensional vector using Eq.10.

$$A = W_2 \cdot H_{\text{activation}} \tag{10}$$

where $A \in \mathbb{R}^4$ and each element in this vector represents the attention score of that stage and W_2 is the learnable weight matrix

4.3.2 Weighted Feature Fusion

1. **Softmax Normalization:** Using the softmax function, the raw attention scores were normalized to generate attention weights (Eq.11).

$$\alpha = Softmax(A), \alpha = [\alpha_1, \alpha_2, \alpha_3, \alpha_4] \tag{11}$$

where α represents the attention weights assigned to each stage and the sum of all these weights is equal to 1, which ensures that the contributions of all stages sum to unity.

2. **Attention Fusion:** The normalized attention weights are then used to compute a weighted combination of the hierarchical feature embeddings using Eq.12.

$$F_{fused} = \alpha_1 E_1 + \alpha_2 E_2 + \alpha_3 E_3 + \alpha_4 E_4 \tag{12}$$

4.4 Final Classifier

The resulting fused vector F_{fused} was passed through an MLP classifier with two hidden layers. LeakyReLU activation function with leaky slope α = 0.01 was used after each hidden layer to address the problem of vanishing gradients and a dropout value of 0.3 is applied between layers as a regularization method to avoid overfitting. The final model is trained using an Adam optimizer with learning rate (LR) = 1×10^{-4}. A softmax classifier using Eq. 13 was employed to classify the input into categories: Real, Fake.

$$P(y_i) = \frac{e^{z_i}}{\sum_{j=0}^{1} e^{z_j}} \tag{13}$$

where, $P(y_i)$ = probability of class i

$$L = -\left(y \log(\hat{y}) + (1-y) \log(1-\hat{y})\right) \tag{14}$$

Cross-entropy loss, defined in Eq. 14, with label smoothing = 0.1 was used to calculate the classification errors and improve the accuracy of the model. where, y = true label, $\hat{y}$ = predicted probability for real, 1-$\hat{y}$ = predicted probability for fake.

4.5 Justification of Model Selection

Deepfakes generated using advanced GANs and latent diffusion models have significantly reduced evident local artifacts like color mismatches and boundary alterations. As a result, detection mechanisms must focus on subtle inconsistencies that may span across different parts of the image instead of relying only on local texture inconsistencies. Traditional Convolutional Neural Networks (CNNs) are efficient at capturing local features because of their fixed receptive fields but fail to capture long-range dependencies across distant regions present in a face. Standard Vision Transformers (ViTs) address this limitation and compute attention globally across image patches, resulting in computation scaling quadratically with respect to the number of patches, making them computationally expensive for high-resolution images.

The Swin-B transformer is particularly suitable for deepfake detection because of its hierarchical architecture and its self attention mechanisms (W-MSA, SW-MSA). Unlike conventional CNNs, which are limited by fixed kernel size, and standard ViT that compute only global attention with high processing cost, Swin-B efficiently captures both local and long-range dependencies. The W-MSA focuses on fine-grained local artifacts such as texture inconsistencies, while the SW-MSA enables cross-region interaction, allowing detection of global semantic irregularities. These capabilities make Swin-B highly effective in identifying subtle and distributed deepfake artifacts.

5 Results and Discussions

To carry out this work, images were first extracted from a dataset composed of real faces from the FFHQ real faces dataset [20] and fake images from the Kaggle deepfake images dataset [22]. These images were used to train the model on both real and deepfake images. The dataset is then divided into train, test and validation sets as shown in Fig. 7. The distribution of real and deepfake images across these splits is summarized in Table 2.

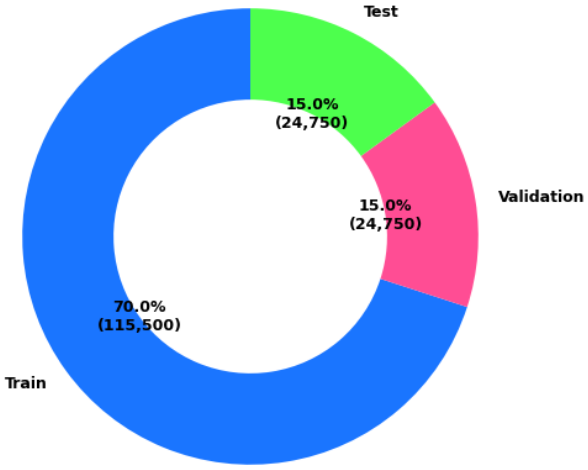


Figure 7: Dataset Proportions

Table 2: Dataset Distribution

Split	Real Images	Deepfake Images
Train	49,000	66,500
Validation	10,500	14,250
Test	10,500	14,250

The model was initialized with pre-trained weights to reduce training time and improve performance. All tests were run on Google Colab Pro, an open-source Python notebook environment featuring a Tesla T4 GPU with 16GB of VRAM and approximately 13GB of system RAM. To ease the T4 GPU's memory constraints, mixed precision training using 'torch.cuda.amp' was employed, which helped reduce both training time and VRAM usage. Checkpoints were saved after every epoch to resume training in case of Colab disconnections. Furthermore, the model was optimized using the Adam optimizer with a learning rate of 0.0001, β_1 = 0.9, and β_2 = 0.999 to ensure a more precise parameter tuning. After training for 10 epochs, the performance of the proposed model on test data was evaluated using various metrics, as shown in Table.3. As shown in Fig. 8, the model demonstrates consistent performance across epochs, indicating robust and reliable deepfake detection capability.

Table 3: Performance Metrics Across Epochs

Epoch	Cross Entropy Loss	Accuracy	Precision	Recall	F1-Score	Matthews Correlation Coefficient
1	0.109	0.962	0.958	0.966	0.962	0.924
2	0.137	0.956	0.937	0.977	0.957	0.912
3	0.163	0.939	0.903	0.985	0.942	0.882
4	0.084	0.969	0.978	0.961	0.969	0.939
5	0.114	0.966	0.949	0.985	0.967	0.933
6	0.110	0.961	0.961	0.962	0.961	0.923
7	0.084	**0.980**	0.968	0.976	0.970	0.939
8	0.130	0.956	0.929	0.988	0.958	0.914
9	0.110	0.965	0.953	0.978	0.965	0.929
10	0.121	0.966	0.954	0.979	0.966	0.932

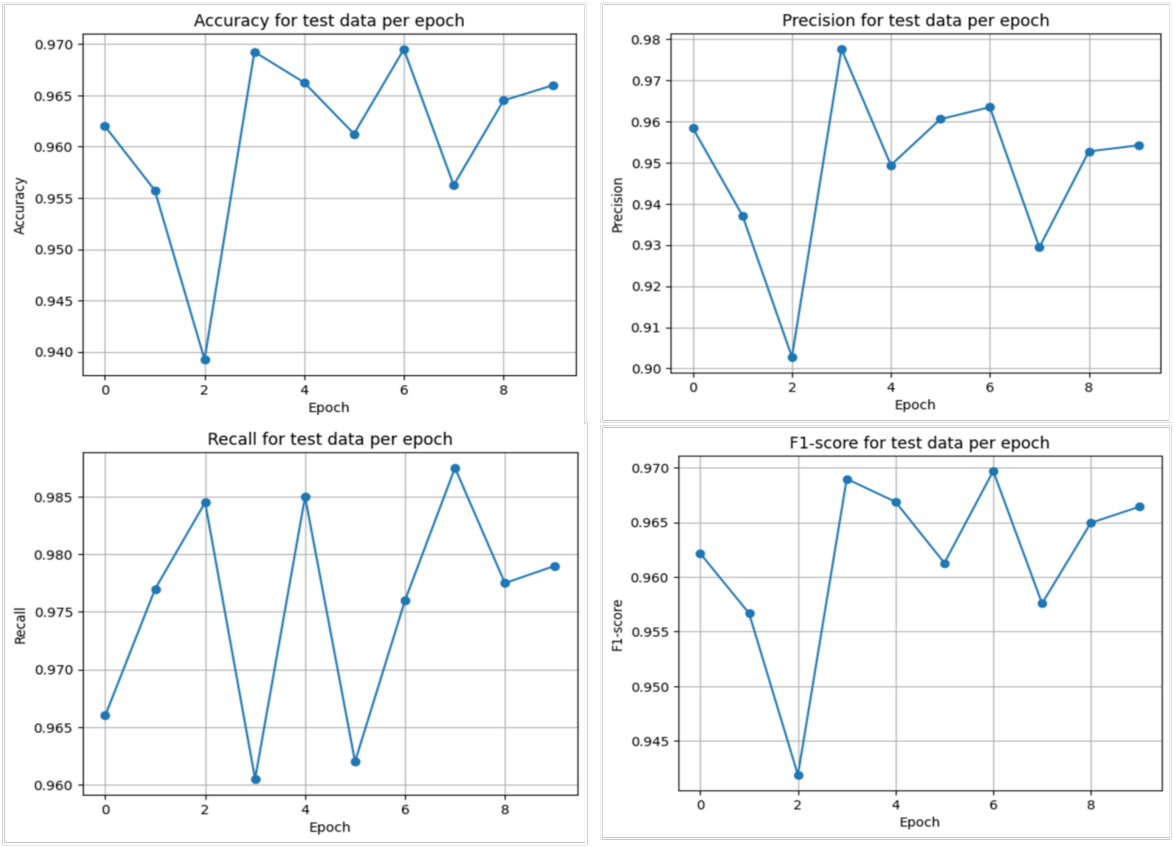


Figure 8: Performance metrics per epoch on test data

5.1 Attention Visualization

Using Gradient-weighted Class Activation Mapping (Grad-CAM) technique, as proposed in [25], heatmaps are generated to visualize the spatial segments of the image on which the proposed model focuses during the classification of real and deepfake images by fusing the feature activations and the corresponding gradients from the layers being selected. These heatmaps were overlaid using a JET color map featuring red, yellow, blue, and green colors, where each color represents the degree of importance of the features:

- **Red/Yellow:** High Activation, most important regions that strongly influence the models decision.
- **Green:** Moderate Activation, regions that contribute partially to the classification.
- **Blue:** Low Activation, regions that have minimal or no influence on the classification.

Figures 9 and 10 illustrate the Grad-CAM heatmaps generated for real and deepfake images at all stages. Heatmap analysis shows that real images exhibit structured and localized activation patterns, while fake images show scattered and irregular attention due to manipulation artifacts.

Fig. 9 exhibits coherent and focused activation patterns that are primarily concentrated on meaningful facial features such as the eyes, nose, and mouth. In Fig. 10 the activation patterns are more diffused and irregular, with high-intensity regions distributed in both the facial and surrounding areas. The scattered activation patterns indicate the presence of inconsistencies introduced during the deepfake generation process, such as manipulated artifacts and unnatural textures. In addition, the attention weights across stages indicate that the proposed model dynamically adapts its focus depending on the input image.

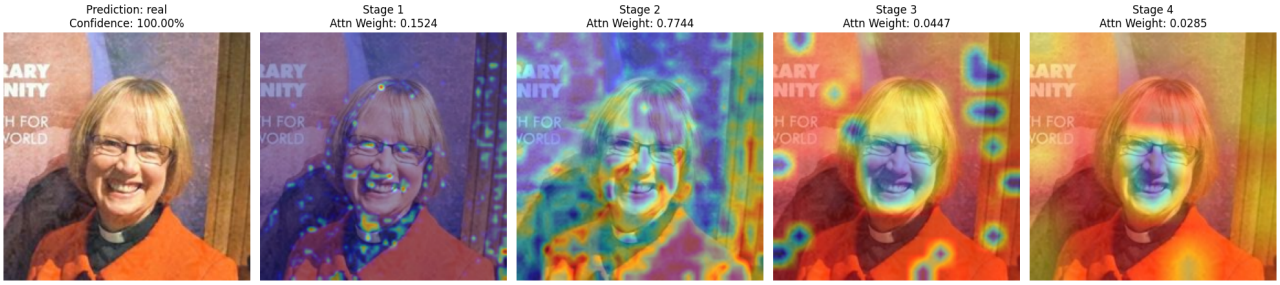


Figure 9: Grad-CAM heatmap generated from Stage 1 to Stage 4 for real image

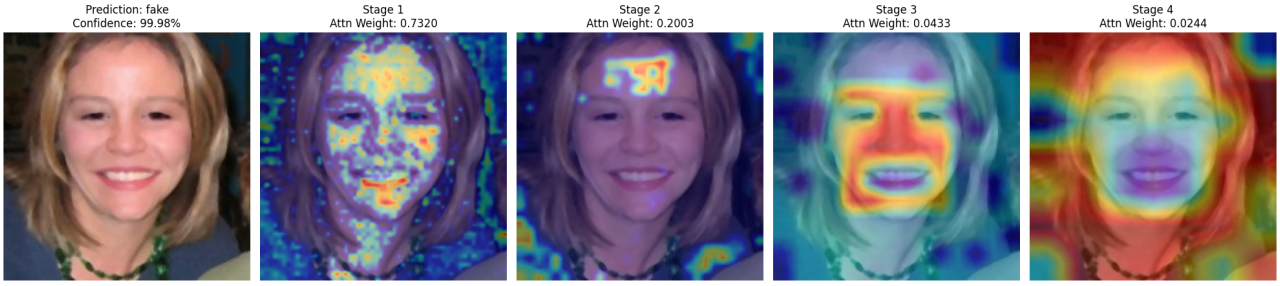


Figure 10: Grad-CAM heatmap generated from Stage 1 to Stage 4 for deepfake image

5.2 Deployment

The proposed model is deployed using the Gradio API [26] to provide a GUI for real-time deepfake detection. A Gradio Interface is built using the Blocks framework, where a prediction function is defined and connected to input and output components. The prediction function contains the logic for generating predictions from an input image.

The application is launched using launch() method. The input to the application is a PIL image, and the outputs include predicted class label, confidence score, stage-wise attention weights. When a user uploads an image, the prediction function is invoked, where the image is first converted to RGB format, resized to 224 x 224 pixels, and transformed to a format compatible with the model. The processed image is then passed to the model to find the predictions, which are then mapped to their corresponding classes. The predictions are formatted appropriately before being returned to the API and displayed to the user. The confidence score is calculated from the predicted probability and displayed as a percentage. Fig.11 shows the User Interface where a random image from the "real" class of the test set was passed as an input. Fig.12 shows the output of a fake image from the dataset, which is correctly classified as Deepfake.

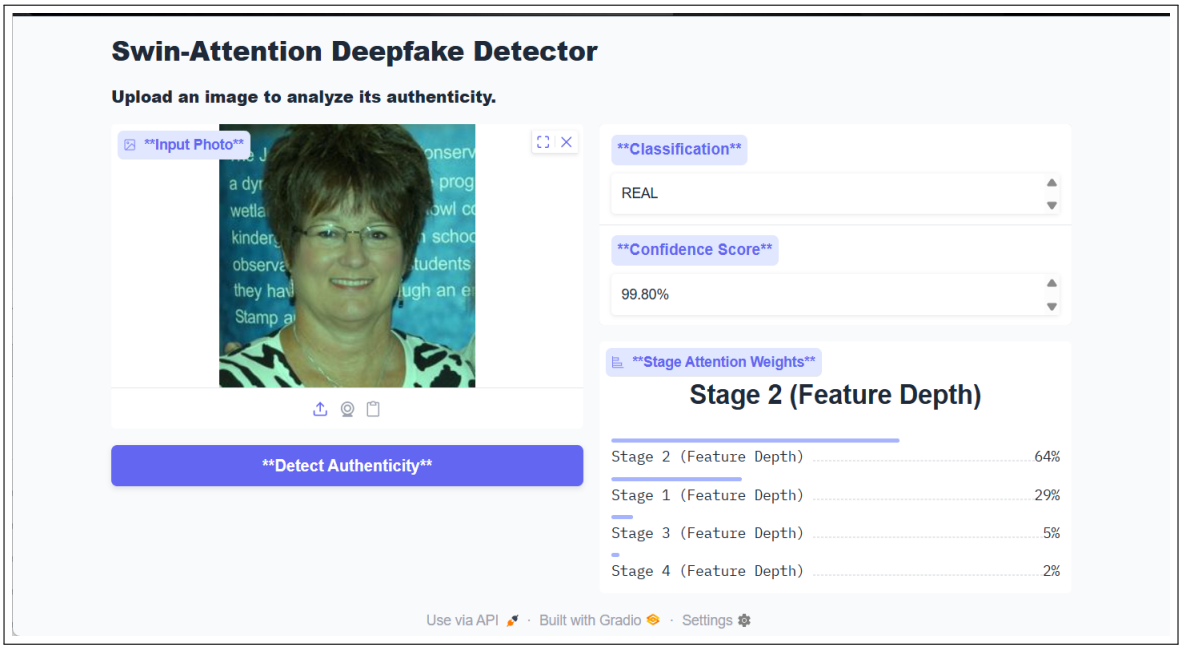


Figure 11: Output of a real image

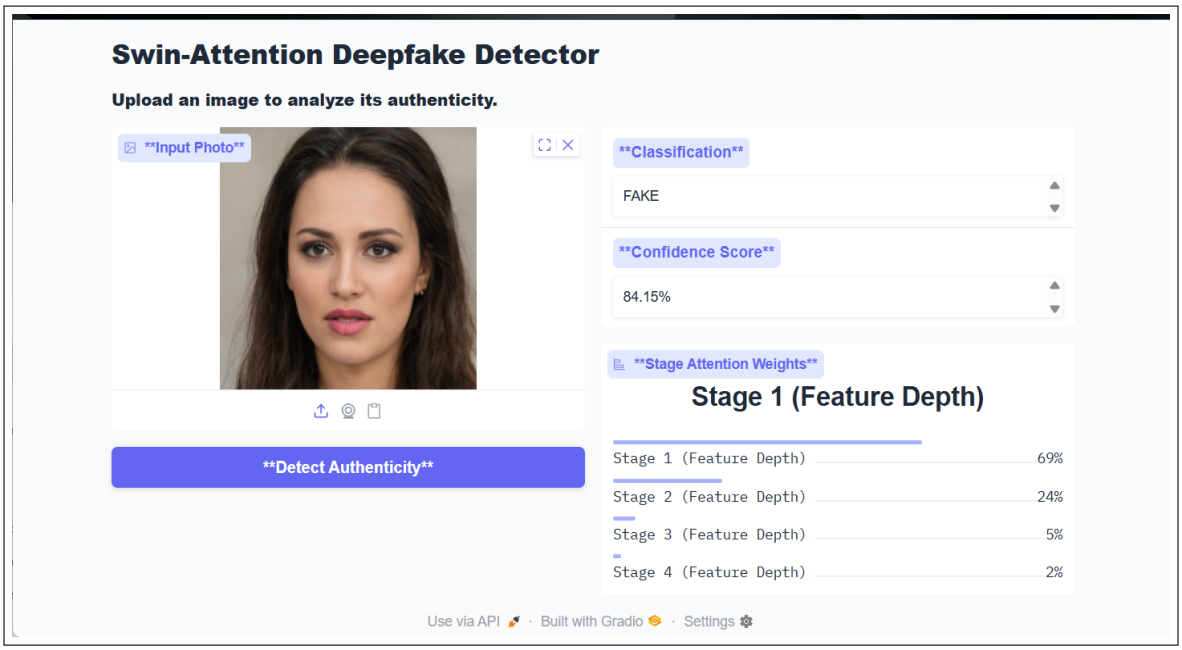


Figure 12: Output of a fake image

5.3 Comparison with other baseline models

Baselines: This section compares the results with existing baseline models for deepfake detection.

- *CNN + ViT (CViT) [27]*: Uses CNN-based architecture combined with a Vision Transformer to identify deepfake faces. The CNN extracts local features such as edges, textures and facial patterns. The features maps generated by CNN are then given as input to ViT. The ViT uses an attention mechanism to understand global relationships and how different parts of the image are related.

- *CapsFake [28]*: Detects instruction-guided deepfakes employing capsule networks. By combining low-level capsules from multiple modalities, CapsFake creates multimodal capsules, which it then uses to predict high-level capsules that represent both real and modified objects.

- *CNN-based [29]*:Employs transfer learning with pre-trained ResNet-50 fine-tuned for deepfake face detection. The original network is modified by adding a fully connected output layer.

- *UAM-Net [30]*: Employs hybrid attention mechanisms to generate Unified Attention Maps. The model concurrently integrates spatial-channel attention with CNN-architecture as a backbone.The attention-based feature maps extracted from both spatial and channel mechanisms were then concatenated to form a unified representation before classification.

- *Swin + ConvNext [31]*: A Hybrid deep learning model that combines the hierarchical attention capabilities of the Swin transformer with the feature extraction mechanism of ConvNext CNNs. ConvNext is used to extract spatial features, and Swin transformer to capture global embeddings. Both of the feature sets are fused before classification.

- *GenConViT [32]*: Generative Convolutional Vision Transformer model consists of two components: generative and feature extraction. The generative component is composed of an auto encoder (AE) and a variational auto encoder (VAE), to understand the latent distribution of training data. The feature extraction component is built using Swin and ViT to extract relevant visual cues.

A comparison with existing systems, shown in Table 4, indicates that the model performs well in most cases. Compared to CNN-based and ViT-based approaches, the proposed AG-Swin model achieves higher accuracy demonstrating its efficacy in capturing both local and global deepfake artifacts as illustrated in Fig. 13. The radar chart as shown in Fig.14 further demonstrate the outstanding and balanced performance of the proposed framework on a standardized set of evaluation metrics. However, AG-Swin incurs high computational cost because of transformer architecture, requires significant GPU memory, and training time. Despite this, the use of window-based attention reduces the computational overhead compared to standard ViT. For real-time deployment, optimization techniques, and the use of quantization techniques may further reduce resource consumption and improve efficiency.

Table 4: Comparison between existing models and AG-Swin

Model	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-score (%)**
Custom Densenet (Pasupuleti et al. 2023 [16])	97	95	-	75
CViT (Soudy et al. 2024 [27])	85	89	86	86
CapsFake (Nguyen et al. 2025 [28])	96.41	97.47	95.3	96.37
CNN-based (Lukas et al. 2025 [29])	96	98	96	97
UAM-Net (Sudarshana et al. 2025 [30])	89.7	85.4	95.8	90.3
Swin + ConvNext (Amarasinghe et al. 2025 [31])	95.6	97.4	93.7	95.2
GenConViT (Deressa et al. 2025 [32])	96.05	-	-	96.12
AG-Swin (Proposed)	**98**	**97.4**	**97.56**	**97.5**

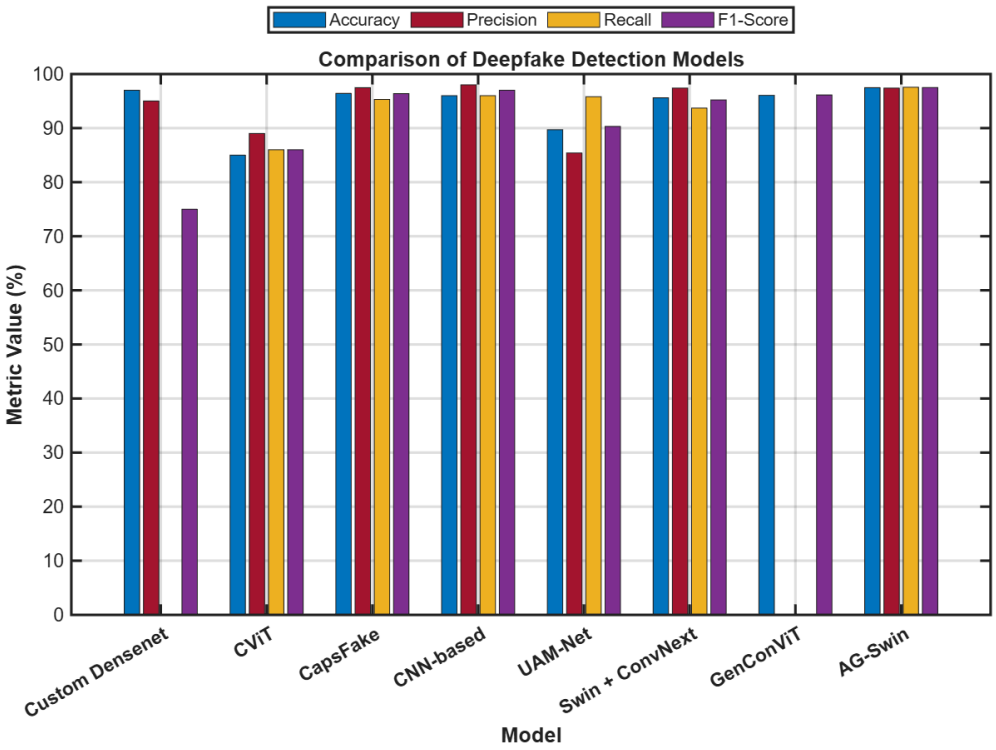


Figure 13: Performance comparison of various deepfake detection models

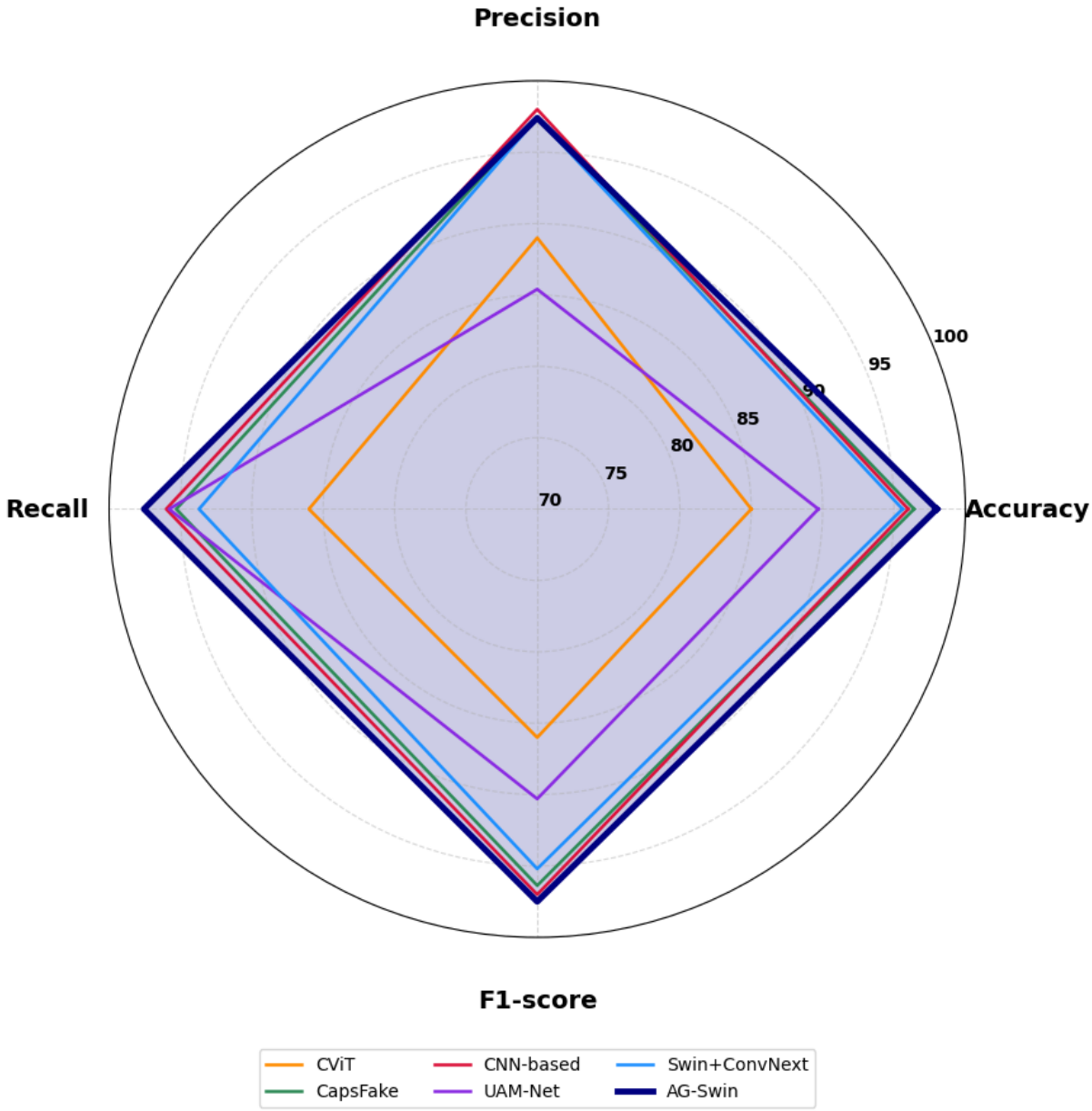


Figure 14: Performance comparison of various deepfake models using a radar chart

6 Ethical and Social Impact

As synthetic media technologies evolve, deepfake images have become a major social and ethical challenge, especially on social media platforms, where images are frequently shared, altered, and used for other purposes. In most of these platforms, information spreads rapidly and user-generated content is often trusted without verification. The ability to generate realistic but deepfake human faces increases the concerns about misinformation, identity fraud, defamation, and influence public opinion leading to the undermine public trust in visual media.

This study technically addresses these challenges by employing the Swin-B Transformer on a hybrid dataset combining real images from FFHQ and fake images from a publicly available Kaggle dataset. The dataset used in our approach reflects the diverse and real-world nature of social media content, improving generalization, and increasing detection accuracy in real-world scenarios. This work has several important social benefits, including the following:

- **Enhanced Content Authenticity:** Assisting social networks in detecting and filtering deepfake images before they mislead users.

- **Support for the control of Fake News:** By detecting deepfake images early, can limit the viral spread of these manipulated visuals particularly in political or sensitive contexts.

- **Ensuring Digital Trust and Safety:** Allows Governments, institutions, platforms, and users to verify the legitimacy of images there by improving public confidence in online content

Despite the high accuracy of the AG-Swin model, there are still a number of challenges with real-world deployment. Adversaries may develop increasingly complex ways to avoid being detected as deepfake detection models evolve, leading to an ongoing arms race. Furthermore, large-scale implementation on social media demands substantial computational resources, raising issues about scalability and cost. The potential for false positives also poses risks, as genuine images can be misclassified as deepfakes, which could have negative social or reputational consequences. Therefore, to ensure ethical and trustworthy usage, effective deployment must prioritize transparency, fairness, and continuous monitoring.

7 Conclusion

In this paper, the proposed model addresses the growing issue of deepfake image generation and its potential negative societal impacts. Deepfakes are an increasing concern due to their ability to spread misinformation, support harassment, and undermine public trust in the visual media. To combat this, we proposed a robust detection framework using Swin-B, a standard variant of the Swin Transformer, which has demonstrated optimal performance in computer vision tasks compared to traditional CNNs. Using the capabilities of this pre-trained model and fine-tuning it for deepfake detection, the model achieved an accuracy of 98% on the test split of the dataset. Furthermore, Grad-CAM analyzes the segments of the image the model is focusing on while classifying the images as *Real* or *DeepFake*. By promoting safer and more trustworthy digital environments, this work not only improves technical accuracy in the detection of deepfakes, but also supports the objectives of social computing.

As the deepfake generation techniques continue to evolve, it will be necessary to re-access the model's performance and potentially retrain the model to ensure its effectiveness. Future work will focus on integrating temporal or multimodal features, extending the detection to deepfake videos, and continuously updating our model with the newly synthetic images, sustaining its accuracy and reliability.

Availability of data and materials

"The data that support the findings of this study can be downloaded from https://github.com/NVlabs/ffhq-dataset and
https://www.kaggle.com/datasets/manjilkarki/deepfake-and-real-images"

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Funding

This research received no external funding.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Grammarly in order to assist with language refinement. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the content of the published article.

References

[1] M. -Y. Liu, X. Huang, J. Yu, T. -C. Wang and A. Mallya, "Generative Adversarial Networks for Image and Video Synthesis: Algorithms and Applications," in Proceedings of the IEEE, vol. 109, no. 5, pp. 839-862, May 2021, doi: 10.1109/JPROC.2021.3049196.

[2] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2022. Transformers in Vision: A Survey. ACM Comput. Surv. 54, 10s, Article 200 (January 2022), 41 pages. doi: https://doi.org/10.1145/3505244.

[3] FakeApp, "FakeApp: Deepfake Creation Tool," [Online]. Available: https://www.fakeapp.org/.

[4] FaceSwap, "Deepfakes FaceSwap GitHub Repository," [Online]. Available: https://github.com/deepfakes/faceswap.

[5] Limin Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. 2019. FaceShifter: Towards high fidelity and occlusion aware face swapping. https://arxiv.org/abs/1912.13457.

[6] J. Vincent, "AI-generated fake faces are being used to trick people online," *The Verge*, Jun. 13, 2019. [Online]. Available: https://www.theverge.com/2019/6/13/18677341/ai-generated-fake-faces-spy-linked-in-contacts-associated-press

[7] A. Hern, "What are deepfakes – and how can you spot them?" *The Guardian*, Jan. 13, 2020. [Online]. Available: https://www.theguardian.com/technology/2020/jan/13/what-are-deepfakes-and-how-can-you-spot-them.

[8] Shukla, P.K., Veerasamy, B.D., Alduaiji, N. et al. Fraudulent account detection in social media using hybrid deep transformer model and hyperparameter optimization. Sci Rep 15, 38447 (2025). doi: 10.1038/s41598-025-24326-8.

[9] X. a. P. B. a. W. W. a. D. J. Xuan, "On the Generalization of GAN Image Forensics," in Biometric Recognition: 14th Chinese Conference, CCBR 2019, Zhuzhou, 2019.

[10] F. Matern, C. Riess and M. Stamminger, “Exploiting visual artifacts to expose deepfakes and face manipulations,” 2019 IEEE Wintere Applications of Computer Vision Workshops (WACVW), pp. 83-92, 2019

[11] C.-C. Hsu, Y.-X. Zhuang and C.-Y. Lee, “Deep Fake Image Detection Based on Pairwise Learning,” Applied Sciences, vol. 10, p. 370, 2020.

[12] T. a. X. X. a. X. M. a. D. H. a. X. Y. a. X. W. Zhao, “Learning Self-consistency for Deepfake Detection,” in IEEE/CVF Conference on Computer Vision (ICCV), 2021.

[13] P. Zhou, X. Han, V. I. Morariu and L. S. Davis, “Two-stream Neural Networks for Tampered Face Detection,” in 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017.

[14] X. Chang, J. Wu, T. Yang and G. Feng, “DeepFake Face Image Detection based on Improved VGG Convolutional Neural Network,” 39th Chinese Control Conference, pp. 7252-7256, 2020.

[15] P. Theerthagiri and G. b. Nagaladinne, ”Deepfake Face Detection Using Deep InceptionNet Learning Algorithm,” 2023 IEEE International Students’ Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 2023.

[16] V. R. Pasupuleti, P. Reddy Tathireddy, G. Dontagani and S. A. Rahim, ”Deepfake Detection Using Custom Densenet,” 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1-5, doi: 10.1109/ICCCNT56998.2023.10307521.

[17] B. Ghita, I. Kuzminykh, A. Usama, T. Bakhshi and J. Marchang, ”Deepfake Image Detection Using Vision Transformer Models,” 2024 IEEE International Black Sea Conference on Communications and Networking (BlackSeaCom), Tbilisi, Georgia, 2024, pp. 332-335, doi: 10.1109/BlackSeaCom61746.2024.10646310.

[18] Usmani, S., Kumar, S. & Sadhya, D., ”Efficient deepfake detection using shallow vision transformer”. Multimed Tools Appl 83, 12339–12362 (2024). doi: 10.1007/s11042-023-15910-z.

[19] Shukla, P.K., Veerasamy, B.D., Alduaiji, N. et al. Encoder only attention-guided transformer framework for accurate and explainable social media fake profile detection. Peer-to-Peer Netw. Appl. 18, 232 (2025). doi: 10.1007/s12083-025-02047-z.

[20] NVLabs, ”Flickr-Faces-HQ (FFHQ) Dataset,” [Online]. Available: https://github.com/NVlabs/ffhq-dataset

[21] Karras, Tero, et al. ”A Style-Based Generator Architecture for Generative Adversarial Networks.” CVPR 2019.

[22] M. Karki, ”Deepfake and Real Images,” Kaggle, [Online]. Available: https://www.kaggle.com/datasets/manjilkarki/deepfake-and-real-images.

[23] Liu, Ze & Lin, Yutong & Cao, Yue & Hu, Han & Wei, Yixuan & Zhang, Zheng & Lin, Stephen & Guo, Baining. (2021). Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. 9992-10002. doi: 10.1109/ICCV48922.2021.00986.

[24] Hugging Face, ”Swin Transformer,” [Online]. Available: https://huggingface.co/docs/transformers/model_doc/swin

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh and D. Batra, ”Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.

[26] Gradio, "Build & share delightful machine learning apps," [Online]. Available: https://www.gradio.app/.

[27] Soudy, A.H., Sayed, O., Tag-Elser, H. et al. "Deepfake detection using convolutional vision transformers and convolutional neural networks", Neural Comput & Applic 36, 19759–19775, 2024, doi: https://doi.org/10.1007/s00521-024-10181-7.

[28] T. Nguyen, N. Khan, and I. Khalil, "CapsFake: A Multimodal Capsule Network for Detecting Instruction-Guided Deepfakes," 2025, doi: https://arxiv.org/abs/2504.19212.

[29] Lukas Kroiß and Johannes Reschke, "Deepfake Detection of Face Images based on a Convolutional Neural Network," 2025, doi: https://arxiv.org/abs/2503.11389.

[30] Kerenalli Sudarshana and Yendapalli Vamsidhar, "UAM-Net: Robust Deepfake Detection Through Hybrid Attention Into Scalable Convolutional Network," *Expert Systems*, vol. 42, no. 2, p. e70009, 2025, doi: https://doi.org/10.1111/exsy.70009.

[31] H. M. S. Amarasinghe and S. P. K. Arachchi, "Deepfake Detection Using a Hybrid Deep Learning Approach with Swin Transformers and ConvNeXt," 2025 International Research Conference on Smart Computing and Systems Engineering (SCSE), Colombo, Sri Lanka, 2025, pp. 1-5, doi: 10.1109/SCSE65633.2025.11030955.

[32] Deressa DW, Mareen H, Lambert P, Atnafu S, Akhtar Z, Van Wallendael G. "GenConViT: Deepfake Video Detection Using Generative Convolutional Vision Transformer. Applied Sciences.", 2025, 15(12):6622. doi: https://doi.org/10.3390/app15126622