Grenze International Journal of Engineering and Technology, June Issue

A Comparative Multi-Model Framework for Explainable Detection of AI-Generated Images using Grad-CAM

B. Karthikeya[1], J. Adithya Naik[2], B. Vivek[3] and A. Ravi Kumar[4]
[1-4]Department of IT, Sreenidhi Institute of Science and Technology, Hyderabad
Email: 22311a12k0@it.sreenidhi.edu.in, 22311a12h4@it.sreenidhi.edu.in, 22311a12j1@it.sreenidhi.edu.in, ravi.a@sreenidhi.edu.in

***Abstract*— The rapid advancement of generative artificial intelligence, especially Generative Adversarial Networks (GANs) and diffusion models, has greatly improved the realism of synthetic images, raising concerns about misinformation, identity manipulation, and digital fraud. In this paper, we propose a comparative multi-model framework for explainable detection of AI generated images using three deep learning architectures, namely Baseline CNN, ResNet50 and EfficientNet-B0. The models are trained on a balanced dataset of 50,000 images combining CIFAKE and Tiny GenImage to ensure diversity across multiple generative techniques and to improve generalization capability. To improve interpretability, Gradient-weighted Class Activation Mapping (Grad-CAM) is also integrated to visualize the decision-critical regions. Experimental results show that EfficientNet-B0 achieves the highest performance of 97.09% accuracy which is better than ResNet50 (96.78%) and Baseline CNN (92.89%) and 91.3% accuracy on unseen generators. The system is implemented as a full-stack application with FastAPI and React which enables real-time prediction, model switching and visual analytics. The proposed framework presents an accurate, scalable, and interpretable solution, while also pointing out the challenges of cross-generator generalization in AI-generated image detection.**

***Index Terms*— AI-generated image detection, deep learning, multi-model comparison, Grad-CAM, explainable AI, digital forensics, Deepfake Detection, Computer Vision.**

I. INTRODUCTION

The rapid advancement of generative artificial intelligence (AI), especially Generative Adversarial Networks (GANs) and diffusion-based models, has transformed digital content creation by enabling the generation of photorealistic synthetic images. Modern generative systems such as StyleGAN, BigGAN and Stable Diffusion have enabled applications in media production, virtual content creation, and data augmentation. But these technologies also have serious risks, from deepfakes to identity manipulation to digital misinformation. Such models are widely available, so even non-experts can create convincing synthetic images, and reliable detection is increasingly important.

Despite recent progress, there are still several challenges for existing AI-generated image detection methods. Many approaches work as black-box models, providing predictions without interpretability, thus limiting their reliability in sensitive domains such as digital forensics.

Grenze ID: 01.GIJET.12.2.654_26

Moreover, most models do not generalize well and perform poorly on unseen architecture generated images. The performance is also limited by dataset limitations. Many datasets are not diverse and encourage the models to learn generator-specific artifacts rather than fundamental differences between real and synthetic images. Moreover, there is a lack of comparative evaluation between different deep learning architectures, which leads to ambiguity in choosing the best models for deployment.

To address these challenges, we propose a comparative multi-model framework for explainable AI-generated image detection using three deep learning architectures, Baseline CNN, ResNet50 and EfficientNet-B0. The models are trained on a balanced dataset of 50k images which is a combination of CIFAKE and Tiny GenImage, using multiple generators for diversity and better generalization. We incorporate Gradient-weighted Class Activation Mapping (Grad-CAM) to provide visual explanations, showing the areas that affect the model predictions, thus increasing transparency and trust.

Unlike the current methods focusing only on the performance of a single model, the proposed framework allows for a joint comparative study of heterogeneous architectures within a unified training and evaluation pipeline. The contribution of this work is the systematic evaluation of model behavior across architectures while providing explainability in a deployable system. Additionally, the framework supports dynamic switching between models and real-time inference, bridging the gap between theoretical evaluation and practical application in digital forensics.

Our key contributions are: (i) construction of a balanced multi-generator dataset of 50,000 images, (ii) development of a comparative framework integrating CNN, ResNet50, and EfficientNet-B0 with Grad-CAM, (iii) comprehensive benchmarking across architectures, (iv) evaluation of cross-generator generalization, and (v) deployment of a real-time inference system.

II. Literature Review

Image forgery detection by AI has emerged as a key area of research in computer vision and digital forensics. The first detection mechanisms for this problem were mainly concerned with the detection of artifacts like irregular eye blinking [1], head pose inconsistencies [2], and abnormalities in facial features [3]. With the rise in sophistication of generative models, however, many of these artifacts have disappeared, thus rendering the existing forensic techniques ineffective. This has led researchers to adopt deep learning-based methods that are capable of discerning statistically significant differences between genuine and AI-generated images. Wang et al. [4] found that CNN models are able to detect fingerprints hidden in the images generated by GANs, whereas Marra et al. [5] found that these images possess artificial fingerprints that can be distinguished easily. These approaches, however, tend to rely heavily on generator-specific artifacts and lack generalizability across different generator architectures [6].

Since the introduction of CNNs for image forensics, they have become popular because of their capability to generate feature hierarchies. The authors in [7] proposed a two-stream architecture of CNN which integrates RGB and noise residual data. In addition to this, some other methods involve constrained convolutional neural networks [9] and statistical anomaly detection networks [8]. Moreover, deep residual networks such as ResNet50 [10] have achieved better results due to their capability to train deep neural networks efficiently. EfficientNet architectures have also shown better results in recent times because of the use of compound scaling technique on depth, width, and resolution. However, since previous methods assess individual architectures, it is difficult to select the best performing one.

Recent research has focused on identifying synthetic images produced by diffusion models. According to Corvi et al. [24], diffusion-generated images leave detectable statistical signatures that can be leveraged for forensic analysis. The literature review on this subject emphasizes the urgent need for robust and generalized identification approaches able to deal with unseen generators and evolving image synthesis technology [25]. Transformer-based models, including ViT (Vision Transformers) [21] and Swin Transformers [22], have shown impressive results due to their ability to capture long-term dependencies and global context within images. On another hand, approaches related to the frequency domain identify spectral mismatches caused by image synthesis, offering additional forensic insights apart from spatial-domain features [23]. Still, many of these approaches are computationally complex and, moreover, not interpretable.

Methods such as XAI are gaining ground within the field of deep learning architectures. Grad-CAM [11] generates explanations using heat maps that show the most influential parts of images used in making predictions. Grad-CAM++ [13], on the other hand, enhances the localization process and improves interpretability of explanations. Recent research has stressed the need for explainable detection methods for building trust among users as well as assisting forensic analysis of suspicious images [26]. Notwithstanding

these advancements, the application of explainability approaches to comparative models within AI-generated image detection is relatively underresearched.

Despite the significant advancements that have already been made, there are some limitations that are yet to be addressed. The generalization of current models, their reliance on small sample sizes, as well as interpretability of results, continue to pose problems. The lack of a universal comparison between models prevents one from determining the optimal solution that could be applied in practice. To solve these issues, this study offers a novel approach in which Baseline CNN, ResNet50, and EfficientNet-B0 models with Grad-CAM explainability are compared using an extensive multigenerator dataset.

III. METHODOLOGY

A. Dataset Description

To build a robust and generalizable detection model, a combined dataset strategy was used by merging the CIFAKE and Tiny GenImage datasets. The CIFAKE dataset contains 30,000 training images (15,000 REAL and 15,000 FAKE) generated from CIFAR-10 and StyleGAN2, and 11,908 test images. The Tiny GenImage dataset consists of 20K training images and 5K validation images generated by five different architectures, namely ADM, BigGAN, Midjourney, VQDM and Wukong.

The final training set contains 50,000 balanced images (25,000 REAL and 25,000 FAKE) using six different generative models. Hence, there is both class balance and generator diversity. Such a combined strategy allows the model to learn generalized features rather than generator-specific artifacts, and improves the robustness against unseen generative architectures. The multiple generators reduce the bias to a specific dataset and encourage the model to learn important differences between real and synthetic images. All images were resized to 224 × 224 pixels to fit the input requirements of the deep learning models.

B. Data Preprocessing

Resizing and Normalization

All the images of CIFAKE and Tiny GenImage datasets were resized to 224 × 224 pixels using bilinear interpolation, to fit the input size of ResNet50 and EfficientNet-B0 architectures. The backbone networks were pretrained on ImageNet. After resizing , pixel values were normalized with the mean and standard deviation computed on the ImageNet dataset.

The normalization is obtained:

$$\text{Normalized} = \frac{\text{Image} - \mu}{\sigma} \qquad (1)$$

Whereas $\mu = [0.485, 0.456, 0.406]$ and $\sigma = [0.229, 0.224, 0.225]$ are the mean and standard deviation values for each of the three image channels, respectively. Normalization helps ensure that the input pixels adhere to a particular standard distribution.

Data Augmentation

To make the model more robust and avoid overfitting, the following data augmentation methods were applied only to the training data set:

Data augmentation included random horizontal flipping (p=0.5), random rotation (±10°), and color jittering (brightness and contrast factor 0.2) to improve robustness and reduce overfitting.

This enables the model to extract invariant features from the input data. The validation and test set were not augmented so as to avoid any bias in the model testing.

C. Multi-Model Architecture

In order to study the effects of architecture selection on detecting AI generated images, three different deep learning architectures have been implemented and analyzed. These include a lightweight Baseline CNN, the more commonly used ResNet50, which uses the concept of residual connections and finally the efficient EfficientNet-B0. The training for all models was done using binary classification (REAL vs FAKE). Each model is capable of working with Grad-CAM and all have been trained on the combined 50k images dataset.

Baseline CNN

The Baseline CNN is a lightweight reference architecture that provides a lower bound on performance and allows us to measure the benefits of more complex designs. The model consists of four convolutional layers with batch normalization, ReLU activation, and max pooling after each layer. The architecture progressively increases

the filter sizes from 32 to 256. The flattened feature maps are then passed through two fully connected layers with a dropout regularization (0.5 and 0.3) before the final binary classification layer. The basic architecture is comprised of ~1.2 million trainable parameters, which is computationally efficient for resource-constrained environments.

ResNet50

We choose ResNet50, a 50-layer deep residual network pre-trained on ImageNet, due to its proven effectiveness for image classification tasks. The architecture uses bottleneck residual blocks with identity skip connections to facilitate effective training of deep networks by alleviating the vanishing gradient problem. We replace the original ImageNet classification head with a custom binary classifier consisting of dropout (0.5), a linear layer from 2048 to 256, ReLU, dropout (0.3), and a linear layer from 256 to 2 output classes (REAL and FAKE). The backbone was initialized with pre-trained weights from ImageNet and fine-tuned during training. The model has around 24 million parameters that can be trained.

EfficientNet-B0

EfficientNet-B0 is a compound-scaled architecture that balances network depth, width and input resolution to achieve better performance with fewer parameters. Like ResNet50, the model was initialized with the ImageNet pretrained weights and modified with the same custom binary classifier. EfficientNet-B0 uses mobile inverted bottleneck convolution (MBConv) blocks with squeeze-and-excitation optimization. This allows for efficient feature extraction. The model has ~5.3M trainable parameters, many less than ResNet50, while achieving better accuracy in our experiments, demonstrating the success of efficient scaling for this task.

TABLE I. MODEL ARCHITECTURE COMPARISON

Model	Parameters	Depth	Key Features
Baseline CNN	~1.2M	4 conv layers	Lightweight, simple architecture
ResNet50	~24M	50 layers	Residual connections, pretrained
EfficientNet-B0	~5.3M	Compound scaling	MBConv blocks, efficient scaling

D. Explainability with Grad-CAM

To overcome the black-box nature of deep learning models and to interpret the predictions, Gradient-weighted Class Activation Mapping (Grad-CAM) is incorporated into the detection framework. Grad-CAM gives visual explanations of the model's decisions. It emphasizes the important parts of the image that the network uses to make its decisions. The output of this method is a heatmap that shows the salient features that the model uses for prediction.

For a specific target class c, the importance weights α_k^c for each feature map A_k are computed by globally averaging the gradients of the class score y^c with respect to the feature map activations:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \qquad (2)$$

where Z is the sum of all the pixels in the feature map. This weight reflects the relative importance of each feature map in relation to the desired class.

Grad-CAM heatmap calculation formula:

$$L_{Grad-CA}^c = \mathrm{ReLU}\left(\sum_k \alpha_k^c A_k\right) \qquad (3)$$

Figure 1. Grad-CAM visualization of model attention for real and AI-generated images

In this work, the Grad-CAM is applied to the last convolutional layer of each model. The resulting 7×7 heatmap is upsampled to 224×224 pixels and then overlaid on the input image using a jet colormap with 50% opacity. Warmer colors (red and yellow) correspond to regions with more influence on the model's prediction.

E. Training Strategy

Training Hyperparameters

All three models were trained with the same hyperparameters to make the comparison fair. The training configuration is shown in Table II.

TABLE II. TRAINING HYPERPARAMETERS

Hyperparameter	Model1	Model2	Model3
Batch Size	64	32	32
Learning Rate	0.001	0.0001	0.0001
Optimizer	AdamW	AdamW	AdamW
Weight Decay	0.0001	0.0001	0.0001
Epochs	8	8	8
Early Stopping Patience	5	5	5
Gradient Clipping	1.0	1.0	1.0

Loss Function

The models were trained with the binary cross-entropy loss function, which for a batch of N samples is defined as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)] \quad (4)$$

where y_i is the ground-truth label (0 for REAL, 1 for FAKE) and p_i is the predicted probability of the FAKE class. The class imbalance was handled by using a weighted sampling strategy.

Techniques for Optimization

Various optimization techniques were employed to enhance training efficiency and convergence. Training was accelerated and memory consumption reduced using Automatic Mixed Precision (AMP). A ReduceLROnPlateau scheduler was used to decrease the learning rate by a factor of 0.5 if the validation loss did not improve for 3 consecutive epochs. To prevent overfitting, early stopping was used to restore the best model checkpoint with a patience of five epochs. To stabilize the training (especially for deeper architectures), we used gradient clipping (max norm = 1.0).

All models were trained on NVIDIA GPU hardware (Tesla T4 or equivalent). The average training time was 6 minutes/epoch for ResNet50 and EfficientNet-B0 and 4 minutes/epoch for the Baseline CNN.

IV. SYSTEM ARCHITECTURE

The design for the image recognition system entails adopting the client-server approach, where images will be uploaded and processed by an AI model in real time. It comprises a React front-end, a FastAPI back-end, and a SQLite database.

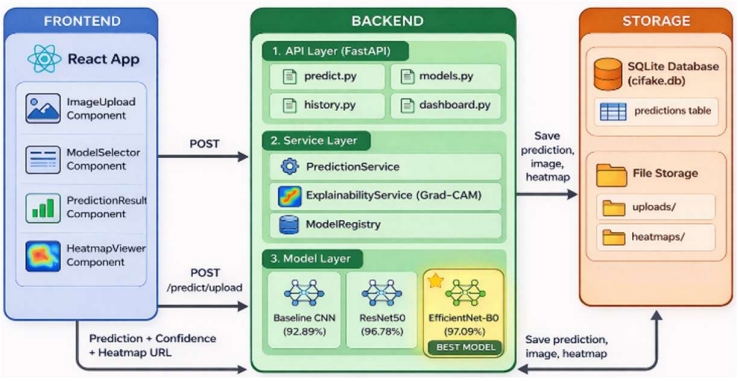

Figure 2. System architecture diagram showing data flow from frontend upload to backend inference, Grad-CAM generation, and result storage with return flow for predictions

As per Fig. 2, the process will include image uploading, pre-processing, model inference, Grad-CAM heatmap generation, and finally visualization of the results.
The FastAPI back-end offers API services including prediction, model management, retrieving prediction history, and analytics. The model registry enables dynamic switching between Baseline CNN, ResNet50, and EfficientNet-B0 models. This is possible without having to restart the service. EfficientNet-B0 will be adopted as the default model owing to its high accuracy.
The React-based front-end entails an application that will allow users to upload an image, select a model to run the prediction, and visualize the results, along with other metrics like confidence score, time taken for the process, and explanations using Grad-CAM. Communication between the two layers is through asynchronous HTTP calls.

A. Proposed Workflow

The suggested architecture uses a sequential pipeline for detecting images generated by AI. The uploaded image is resized and normalized before being passed into the chosen network (Baseline CNN, ResNet50, or EfficientNet-B0), where a prediction, confidence value, and Grad-CAM heatmap are produced. All results are then saved in the SQLite database and shown in the frontend dashboard.

B. Assumptions

This architecture makes an assumption that all the RGB input images are of sufficient quality. All images are resized into 224×224 pixels and the problem formulation is performed in binary classification (REAL vs FAKE). Sufficient computational power is available and the datasets are expected to be correctly labeled.
One of the important features of the system is a dynamic model switching. When a model is selected the backend retrieves its checkpoint and updates the current inference model. This allows users to test different architectures on the same input without breaking the system. This opens the way for comparative studies and practical experiments.
To provide data persistence, prediction records are stored in a lightweight SQLite database with filename, prediction result, confidence score, timestamp, image path, heatmap path and model version. For storage efficiency, uploaded images and generated heatmaps are stored in dedicated directories and served as static files for easy retrieval and traceability of prediction results.

V. Experimental Results

A. Evaluation Metrics

Model performance was evaluated using accuracy, precision, recall, specificity, and F1-score computed from TP, TN, FP, and FN values. These metrics provide a comprehensive assessment of classification effectiveness for AI-generated image detection.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (6)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (7)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (8)$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (9)$$

B. Comparative Model Performance

The three models were tested using the CIFAKE test dataset with 11,908 images (REAL images: 5,906 and FAKE images: 6,002). The comparative results are shown in Table III. EfficientNet-B0 model has the highest accuracy rate of 97.09%. This is due to the success of compound scaling in achieving a good performance in image classification tasks. The recall rate of 0.9805 of the EfficientNet-B0 model suggests that the model has great detection power in identifying FAKE images hence eliminating false negatives. The ResNet50 model has the highest precision rate of 0.9732.

C. Cross-Generator Generalization

To test generalization capability , all models were tested on the Tiny GenImage validation containing images from five unseen generators (ADM , BigGAN , Midjourney , VQDM and Wukong) . The per-generator accuracy is shown in Fig. 3. EfficientNet-B0 achieves the highest average accuracy of 91.3%, showing good generalization ability to different generative architectures.

We observe a performance drop against the CIFAKE test results (97.09% to 91.3%) which is an inherent problem of domain shift across various generative models. Such variations are due to differences in the generation mechanisms, distributions of the training data, and patterns of artifacts across the models. However, the achieved generalization accuracy suggests that the proposed framework is able to capture generator-agnostic features instead of relying on generator-specific artifacts.

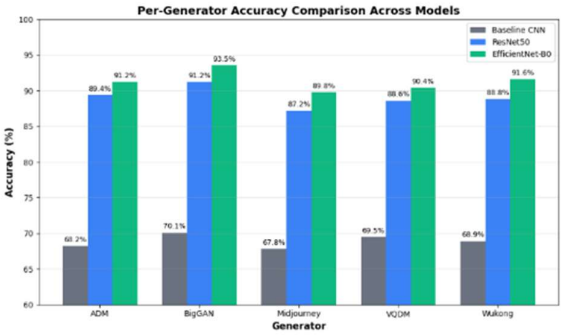

Figure 3. Per-generator accuracy comparison

D. Ablation Studies

Ablation experiments indicate that Grad-CAM introduces negligible performance overhead, while the combined multi-generator dataset significantly improves cross-generator generalization. Data augmentation further enhances robustness and overall detection accuracy.

TABLE III. COMPARATIVE PERFORMANCE OF PROPOSED MODELS ON CIFAKE TEST SET

Model	Accuracy	Precision	Recall	Specificity	F1-Score
Baseline CNN	92.89%	0.9150	0.9469	0.9112	0.9306
ResNet50	96.78%	0.9732	0.9625	0.9738	0.9678
EfficientNet-B0	97.09%	0.9624	0.9805	0.9609	0.9714

VI. LIMITATIONS

However, the promising results obtained by the proposed framework still have several limitations. First, the study does not experimentally compare against transformer-based or frequency-domain approaches, which are recent developments in AI-generated image detection. These approaches have demonstrated encouraging performance, but they often exhibit higher computational complexity and fall outside the scope of this work.

Second, although we use a multi-generator dataset strategy, we only evaluate on CIFAKE and Tiny GenImage datasets. Further assessment of robustness requires additional validation on independent and more diverse benchmarks, especially with recent diffusion based models.

Third, there is a performance gap in the cross-generator generalization, where the model does not learn all generator-invariant features. This exemplifies the intrinsic challenge of domain shift for AI-generated image detection.

Finally, Grad-CAM delivers insightful visual explanations, but provides coarse-grained interpretability and may not capture the fine-grained feature contributions. Further exploration of more advanced explainability techniques may improve model transparency.

VII. CONCLUSION

In this paper, we propose a comparative multi-model framework for explainable detection of AI-generated images using three deep learning architectures namely Baseline CNN, ResNet50 and EfficientNet-B0. We create a balanced dataset of 50k images, class-balanced and diverse across multiple generative models by combining

CIFAKE and Tiny GenImage. We propose an approach that considers classification accuracy and also interpretability by introducing Grad-CAM to visualize the areas affecting model predictions.

The experimental results show that EfficientNet-B0 performs better than the others in overall performance with an accuracy of 97.09% compared to ResNet50 (96.78%) and Baseline CNN (92.89%). However, the high recall value of EfficientNet-B0 demonstrates the effectiveness of the model to detect the AI generated images with less false negatives. Moreover, the cross-generator evaluation on previously unseen generative models reached an average generalization accuracy of 91.3%, demonstrating the model's ability to learn robust and partially generator-agnostic features.

Despite these promising results, there remain challenges in fully generalizing across diverse and changing generative models. Even though the generation techniques and data distributions were different, the domain shifts still had an impact on the performance. In future work, we plan to improve generalization by adding diverse datasets, transformer-based and frequency-domain approaches and by exploring advanced explainability techniques. The improvements are intended to improve the robustness, scalability, and real-world applicability of AI-generated image detection systems.

REFERENCES

[1] Y. Li, M. Chang, and S. Lyu, "In ictu oculi: Exposing AI-created fake videos by detecting eye blinking," in Proc. IEEE Int. Workshop on Information Forensics and Security (WIFS), 2018.

[2] X. Yang, Y. Li, and S. Lyu, "Exposing deep fakes using inconsistent head poses," in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2019.

[3] F. Matern, C. Riess, and M. Stamminger, "Exploiting visual artifacts to expose deepfakes and face manipulations," in Proc. IEEE Winter Applications of Computer Vision Workshops (WACVW), 2019.

[4] S. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot… for now," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2020.

[5] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi, "Do GANs leave artificial fingerprints?" in Proc. IEEE Conf. on Multimedia Information Processing and Retrieval (MIPR), 2019.

[6] D. Cozzolino, J. Thies, A. Rössler, C. Riess, M. Nießner, and L. Verdoliva, "ForensicTransfer: Weakly-supervised domain adaptation for forgery detection," arXiv preprint arXiv:1812.02510, 2018.

[7] P. Zhou, X. Han, V. I. Morariu, and L. S. Davis, "Two-stream neural networks for tampered face detection," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017.

[8] N. Rahmouni, V. Nozick, J. Yamagishi, and I. Echizen, "Distinguishing computer graphics from natural images using convolutional neural networks," in Proc. IEEE Workshop on Information Forensics and Security (WIFS), 2017.

[9] B. Bayar and M. C. Stamm, "A deep learning approach to universal image manipulation detection using a new convolutional layer," in Proc. ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec), 2016.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2016.

[11] 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," in Proc. IEEE Int. Conf. on Computer Vision (ICCV), 2017.

[12] X. Yang, Y. Li, H. Qi, and S. Lyu, "Exposing GAN-synthesized faces using landmark locations," in Proc. ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec), 2019.

[13] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in Proc. IEEE Winter Conf. on Applications of Computer Vision (WACV), 2018.

[14] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," arXiv preprint arXiv:1710.10196, 2017.

[15] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2009.

[16] A. Krizhevsky, "Learning multiple layers of features from tiny images," Technical Report, University of Toronto, 2009.

[17] I. Goodfellow et al., "Generative adversarial networks," in Advances in Neural Information Processing Systems (NeurIPS), 2014.

[18] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Advances in Neural Information Processing Systems (NeurIPS), 2020.

[19] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2022.

[20] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2021.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. on Learning Representations (ICLR), 2021.

[22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, “Swin transformer: Hierarchical vision transformer using shifted windows,” in Proc. IEEE Int. Conf. on Computer Vision (ICCV), 2021.
[23] Y. Durall, M. Keuper, and J. Keuper, “Watch your up-convolution: CNN-based generative deep neural networks fail to reproduce spectral distributions,” in Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2020.
[24] A. Corvi, D. Cozzolino, G. Poggi, K. Nagano, and L. Verdoliva, “On the Detection of Synthetic Images Generated by Diffusion Models,” in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023.
[25] Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, “Detecting AI-Generated Images: A Survey,” in arXiv preprint arXiv:2303.14126, 2023.
[26] H. Guarnera, O. Giudice, and S. Battiato, “Fighting Deepfake by Exposing the Convolutional Traces on Images,” in IEEE Access, vol. 11, pp. 10234-10248, 2023.