INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT (IJSREM)
VOLUME: 10 ISSUE: 04 | APRIL - 2026 SJIF RATING: 8.659 ISSN: 2582-3930

ENHANCING TRUST: EXPLAINABLE AI FOR IDENTIFYING GENERATED IMAGES

Dr. B. Jalender[1] , Chalumuri Yuvakishor[2], Mulagundla Akshith[2],

Pallapu Krishna Kanth[2], Sripada Abhinav[2]

[1]*Associate Professor, Department of Information Technology, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India*

[2]*Department of Information Technology, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India*

***Abstract*— Generative models have achieved rather rapid progress, which has simplified distinguishing between real and fake photos. The existing detection techniques are typically not very practical or provide unclear explanations, making them less applicable in practice. This paper offers a DL-based model of image classification with explainable identification to address this problem. The system works using a tagged set of actual and AI-generated photos that have been changed in size, normalized, and augmented by flipping, rotating, and changing the color. Some of the models we use and test include a custom CNN, a longer CNN with dropout, and a pretrained Xception network. The Flask-based web app uses the best-performing model to make predictions in real time. Also, a Grad-CAM mechanism is included in to make visual explanations that show crucial areas that affect the model's choice. The accuracy of the Xception model is 99.92% that indicates that it is superior to other models in experiments. The proposed method is an effective combination of high classification and interpretability, making it suitable to real-world synthetic image detection tasks.**

"Keywords— AI-generated images, Image classification, Deep learning, Transfer learning, Explainable artificial intelligence (XAI), Grad-CAM."

I. INTRODUCTION

Generative AI is transforming the creation of digital content very rapidly. It is now capable of producing very realistic images which appear to be much closer to real-life photos. Large-scale generative models, including latent diffusion and large-scale diffusion models, have improved diffusion-based and text-to-image models recently, demonstrating their ability to generate high-resolution and photorealistic content [2], [5], and [6]. The existence of huge datasets like LAION-5B has sped up the progress of these models even further, leading to their widespread use in many fields [7]. Such new technologies are quite handy in the creative sphere, yet they raise significant questions of authenticity, disinformation, and abuse. Indicatively, AI-generated content has been confused with real media content in the past [1]. Also, the psychological impact of misinformation and influence on the perceptions of people make the necessity of a good detection system even more significant [3], [4].

Although generative modelling has made a significant step forward, it is still difficult to distinguish AI-generated images as they appear more real and in more styles. Traditional methods of assessing visual plausibility often rely on manual creation of features or multimodal analysis, which may not be relevant to modern generative techniques [4]. This limitation indicates the importance of effective and scalable solutions that can effectively identify authentic and fake images. CNNs, in particular, and DL in general have emerged as a powerful tool to solve image identification problems since it is capable of automatically constructing hierarchical feature representations on the basis of data [8]–[10]. The models have proven to be rather effective in various computer vision tasks, and have been applied to synthetic picture detection.

To this end, the primary purpose of this research is to deliver a model of categorizing real and fake photos created by AI that is functional and user-friendly. The proposed solution involves DL techniques that perform binary classification and simplified to be easier to comprehend to enable people to have more trust in it. The task is to design, train and test a great number of CNN-based architectures and deploy the most effective model to a web application that operates in real time. The key objectives are to achieve high classification accuracy, ensure the model is powerful through appropriate preprocessing and augmentation techniques and provide visual explanations of how the model reached its decisions.

There are three contributions made to this work. To start with, a fully developed DL pipeline is developed in the classification of images. This involves data preparation, model training and testing its performance. Second, a systematic comparison of several models, such as bespoke CNN architectures and a pretrained network, is done to find the best one. Third, the deployed system contains an explanation mechanism that is built in which makes the predictions of the model easier to comprehend by presenting them in a visual format. Overall, our research provides a valuable and practical means of coping with the increasing issue of the detection of AI-generated fake photos.

II. RELATED WORK

Other recent studies in the area of AI-generated and altered media have received a lot of attention due to rapid advances in generative models. Numerous research have concentrated on the detection of synthetic pictures generated by contemporary text-to-image frameworks.This demonstrates that discriminative characteristics that are specific to generative pipelines can be obtained [11]. Corvi et al. also examined specific to diffusion-based picture synthesis detection methods. They discussed the difficulty of being aware of such images as they are increasingly realistic and the necessity of having a good detection system [12].

The methods that were used in the past relied heavily on statistical and low-level feature analysis. Li et al. proposed a method that works based on color component differences to identify images generated by deep networks, exploiting variances that are often not visible to humans [13]. Categorizing GAN-generated pictures according to the Benford rule, Bonettini et al. made use of statistical anomalies in pixel distributions [14]. In spite of the practical results of these approaches, they are often unable to be generalized to a large range of generative models, especially with the emergence of advanced diffusion algorithms.

A number of studies have focused on the detection of deepfakes in video, which is basically similar to the detection of synthetic images. Amerini et al. used optical flow-based convolutional neural networks to identify motion inconsistencies in deepfake movies [15]. Gueara and Delp proposed an approach which is based on a recurrent neural network which makes use of time dependencies to discover edited video content [16]. M2TR by Wang et al. is a more modern method that incorporates both spatial and temporal information to achieve the better identification [17]. Another area of research has been hybrid architectures that utilize CNN and LSTM models to utilize both spatial and sequential information to enhance deepfake detection [18]. These are effective with video based tasks, but cannot be applied directly to the task of classifying pictures which are static because they rely on temporal information.

With increasingly complex DL models, the ability to interpret them has been a valued aspect of AI systems that can be relied upon. Gunning et al. stressed how important XAI is for making decision-making systems more open and trustworthy for users [19]. In this context, Grad-CAM, which Selvaraju et al. came up with, is a popular way to show how a model makes judgments by showing the parts of an image that have the biggest impact on a prediction [20]. Such types of strategies prove much assistance in making judgment that are sensitive or highly implication such as search of false images.

Although much progress has been made in this field, gaps in studying exist. Many of the existing techniques merely examine handmade or statistical properties, which are not very robust to new generative techniques, or have complex structures, difficult to apply to real-time systems and expensive to execute. Moreover, much of existing research is on the task of identifying deepfakes in videos, implying that there are no simple and efficient methods to categorize non-video images. Also, despite the existence of numerous studies on detection performance, they often disregard explainability, thus limiting the interpretability and potential practical use of the models.

To address these issues, the proposed system is aimed at resolving those problems, as it is a deep learning-based binary classification of real and fake images, which is based on convolutional architectures to deliver powerful feature extraction. Its methodology contrasts with other existing ones as it relies on Grad-CAM to offer an explainability mechanism to provide visual information regarding model predictions. The system is also configured to operate on the web that enables users to interact with it and make decisions real time. The proposed work is unique in comparison to the other approaches: it uses high-performance classification, interpretability, and real-world use. It also fills important gaps that have been found in the literature.

III. MATERIALS AND METHODS

A) System Overview and Architecture

The proposed solution provides the entire system to categorize and describe the way of creating AI-generated fake photos. The entire process begins with an input image, which undergoes preprocessing steps such as scaling and normalization to ensure that it can be used with the trained DL models. Once the picture has been processed, it is subjected to classification model. Various architectures are tried during the training process and the most suitable one is selected to be deployed. The algorithm provides a guess of the possibility of the image being real or false and a confidence score. It has a Grad-CAM module to help make the decisions made by the model easier to interpret, as those parts of the picture influencing them are displayed. The framework should operate in two processes; first, offline training stage to construct and test the model and second, an online deployment stage in which a Flask-based web application will be used to make real-time inferences.

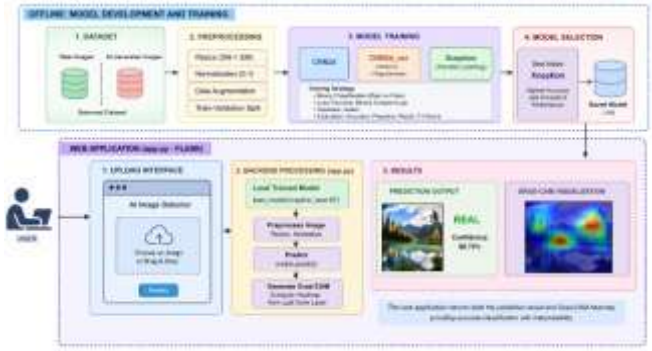

Fig.1 System Architecture

The general layout of the system of the proposed framework is represented in figure 1. The architecture begins with the user entering the information via the web interface, whereby he/she uploads a picture which is relayed to the preprocessing module. Once it has been processed, the image is forwarded to the trained classification model to obtain the likelihoods of all the predictions. The Grad-CAM module, then, creates a heatmap highlighting the important parts and overlays it on the original image. The interface of the web application then presents the user with the prediction results

as well as a visual explanation of the same. This complicates the decisions making process and enables it to be understood easily.

B) Dataset Description

The data set used in this study is based on the AI versus Human Generated Images data set available on Kaggle, which was firstly introduced as a part of the Women in AI 2025 competition. This study is done with only the train subset of the dataset. It contains photos that are organized in CSV file and file names and labels accompanying these photos. There are two types of photos in the dataset: Real (real images) and Fake (AI-generated images). Each image is annotated with a label, and a structured file of annotations connects each image with the label, thus supervised learning is available. The images belong to real-life platforms and are contrasted with the fake images created with the help of the strong generative algorithms. To test the effectiveness of the model, the dataset is divided into 70% and 30% training and validation/testing respectively.

C) Pre-processing

Preprocessing stage is quite critical in ensuring consistency of the input data and also in ensuring the model performs better. It ensures that picture sizes and formats and distributions are identical, and it employs augmentation techniques to enable the model to learn to generalize among various visual patterns.

a) Data Processing: To keep things consistent and work with DL architectures, all input photos in the proposed system are first downsized to a constant size of 224×224 pixels. Then, the images are converted into a form of a tensor representation such that they can be processed by the PyTorch framework. This is followed by normalization with the values of mean and standard deviation in ImageNet. This makes sure that the pixel intensity distributions are the same across the whole dataset. This phase of normalizing is very crucial for using pretrained models since it makes the input data match the distribution that the models were initially trained on. This speeds up convergence and performance.

b) Data augmentation: The data is further increased in the course of training in order to produce even a more resilient model. Some of these include random rotation within a given range, random flipping in horizontal and vertical direction as well as colour jittering (contrast and brightness changing). These kinds of fluctuations lower the likelihood of overfitting, provide the training data diversity, and enable the model to learn qualities that are constant. Augmentation enhances generalization capabilities of the model to new input by feeding the model with numerous photos of the same image. In the study, resizing and normalizing are the only methods that are used to keep consistency.

D) Model Architectures

The suggested approach uses several DL networks to efficiently classify both actual and artificial intelligence-generated pictures. In order to evaluate the performance and select an optimal architecture to be used in the task, a pretrained model and specially developed convolutional neural networks are implemented.

a) CNN2d (Baseline): CNN2d model is a customized convolutional neural network with four convolutional layers. These layers are based on batch normalization, ReLU activation, and max-pooling. These layers slowly extract spatial features of input photographs and finally completely connected layers are employed to classify them. It is a simple model which can be utilized to test ability to extract simple features.

$$S(i,j) = \sum_{m}\sum_{n} I(i+m, j+n) \cdot K(m,n) \quad (1)$$

b) CNN2d_ext (Improved CNN): To enhance the CNN2d model, CNN2d-ext model introduces dropout layers that follow the convolutional and fully connected layers. This extra step helps stop neurons from co-adapting, which helps decrease overfitting. The idea of the model is to improve generalization and retain the same fundamental structure as the base CNN.

c) Xception: The Xception model is created using the timm library and is a pretrained deep convolutional architecture. It employs transfer learning whereby learnt feature representation involves large datasets. We adapt the final classification layer to suit the task of binary classification. The choice of this paradigm to be deployed is based on the fact that other designs are less effective than this paradigm.

$$O_{i,j,k} = \sum_{m,n} X_{i+m,j+n,c} W_{m,n,k} \quad (2)$$

E) Training Strategy

The training procedure will aim at ensuring that convergence is stable and that discriminative features are learnt effectively. To make predictions more generic and less certain, a cross-entropy loss function with label smoothing is used. The AdamW optimizer, which has a learning rate of 1e-4 and weight decay for regularization, is used to improve the models. To make the learning even more effective, there is a cosine annealing learning rate scheduler and 10 epochs. A batch size of 16 is used to train the models and this is a good compromise between speed and precision. This is a way of ensuring that every design which was considered is optimized.

F) Performance Evaluation Metrics

The performance of the proposed models is measured with the help of standard classification measures to provide a complete picture. All metrics are macro averaged i.e. they treat each class equally, regardless of the number of examples of the class.

Accuracy: Accuracy is the proportion of samples correctly categorized of the overall number of samples. It

provides an approximate of the success of the model in all classes.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

Precision: Precision examines the number of the anticipated positives that are positive. It demonstrates the ability of the model to avoid false positives, which is essential to discriminate between genuine and fake pictures.

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

Recall: Recall informs you of the number of the actual positive samples which were correctly identified as positive. It demonstrates the ability of the model to find all the required examples.

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

F1-Score: F1-score is the weighted average of the accuracy and recall, i.e. it provides a reasonable evaluation of the accuracy and the recall. It can be of great assistance where a balance between false positives and false negatives is required.

$$F1\ Score = \frac{2 \times (Precision \times Recall)}{Precision + Recall} \quad (6)$$

G) Explainability using Grad-CAM

The proposed solution has Gradient-weighted Class Activation Mapping (Grad-CAM) that eases model predictions. The method identifies the final convolutional layer of the network and establishes forward and backward hooks to the network, where feature activations and their gradients are captured. They are combined to form weight of importance in each area of the image as a heatmap that is displayed by the image in each important area. The heatmap is then scaled in order to give you a clear visual understanding and overlaid on the original image. This approach adds transparency and reliability to synthetic picture categorization by providing individuals with an understanding of the way the machine makes such decisions.

H) Deployment

The proposed system is configured as a web-based application based on Flask framework to allow individuals to communicate with it and access it online. Users may upload input photographs using the interface. These images are then processed using the trained classification model to make predictions and assigning confidence scores. Grad-CAM visualization is also used to display heatmaps highlighting important areas that influence the decision in the system. Registration and authentication in user administration is safely managed by using a SQLite database. Such an implementation makes the model a handy one that allows individuals to learn and comprehend the authenticity of photographs on-the-fly.

IV. EXPERIMENTAL RESULTS

A) Experimental Setup

The proposed system is experimented to determine the ability of various DL models to distinguish between real and fake photos. We compare three architectures: CNN2d, CNN2d_ext and Xception using the standard criteria of performance such as accuracy, precision, recall and F1-score. These measures provide a complete view of the ability of the models to categorize things in a reasonable testing condition. The primary purpose of the study is to compare the performance of the baseline and advanced architectures and discover the most promising model of fake image finder, which will be accurate and reliable.

B) Performance Evaluation

To bring the how well worked of the models to a common classification measure, we were able to clearly compare the how well worked of the model to find real and AI generated photos.

Table.1 Performance Evaluation of Models

ML Model	Accuracy	Precision	Recall	F1-Score
CNN2d	0.9933	0.9933	0.9933	0.9933
CNN2d Ext	0.9699	0.9712	0.9700	0.9699
Xception	0.9992	0.9992	0.9992	0.9992

All assessment measures show that the Xception model outperforms CNN-based architectures with an accuracy of 0.9992, which is depicted in Table 1. The CNN2d baseline, too, performs well, whereas CNN2d_ext does not perform as well, which indicates that the pretrained model is superior.

C) Comparative Analysis

The bar graphs are used to indicate the performance of the models in various measures so that you can have a better idea of their effectiveness relative to the other models.

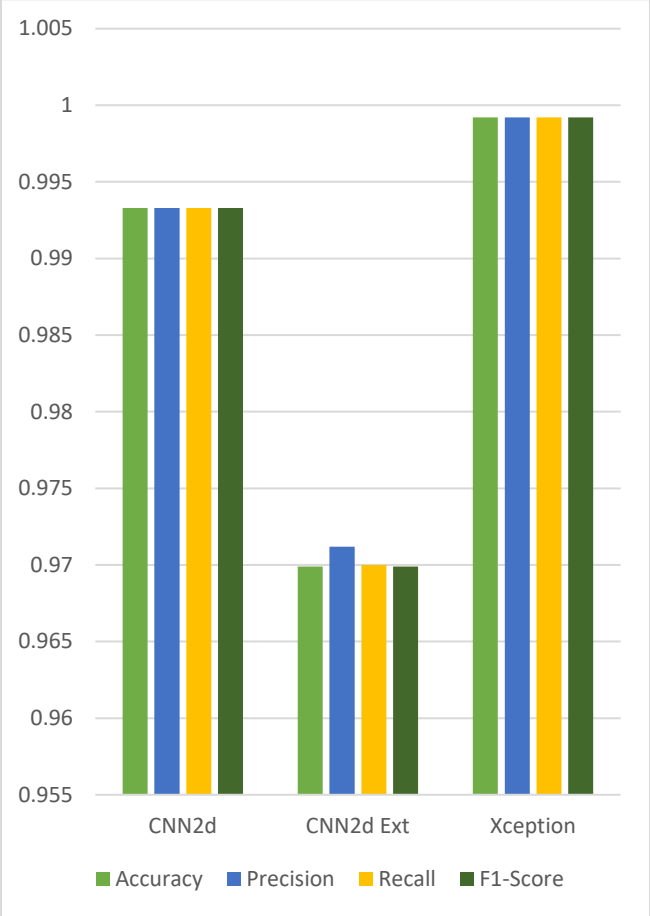

Fig.2 Comparison Graph

Figure 2 shows how accurate, precise, recall, and F1-score are for all three models. Xception model always outperforms all other models in all aspects, demonstrating that it is more stable and generalizes more effectively. The CNN2d model is also effective, providing the same results. Conversely, CNN2d_ext model does not perform as well, possibly due to being more regularized. Overall, the graph indicates that transfer learning is more effective compared to specifically designed systems to do so.

D) UI Demonstration

Fig.3 Upload Interface

The user interface is visible in Fig. 3, and here people can post a picture. The system takes picture files and gets them ready for preprocessing and categorization.

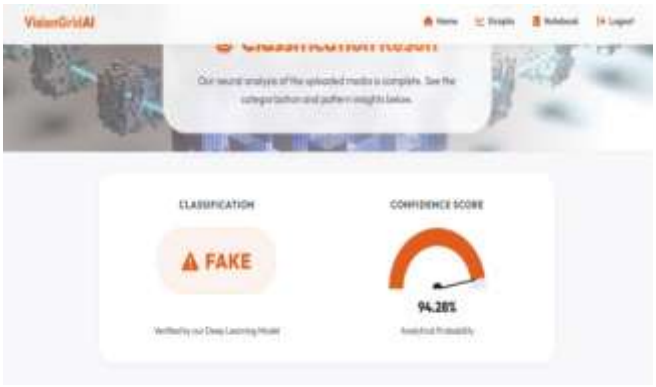

Fig.4 Prediction

The figure 4 demonstrates a prediction result of an uploaded image. It displays the categorical (Real/Fake) and the score that the model used to classify it with confidence.

Fig.5 Grad-CAM

The Grad-CAM visualization presented in Fig. 5 demonstrates key components of the image that assisted the model in coming up with its prediction. This makes it easier to understand the classification result.

Fig.6 Upload Interface

Figure 6 displays another example of the upload interface, which shows that it works the same way for all types of input photographs that need to be classified in the web app.

Fig.7 Prediction

In figure 7, the prediction output of another input picture is presented. It demonstrates the model categorization and confidence that indicates that the model is consistent with a wide range of inputs.

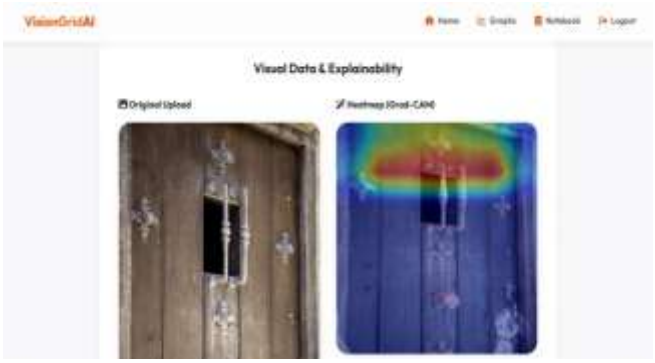

Fig.8 Grad-CAM

The output of the Grad-CAM with the prediction is presented in Figure 8. It demonstrates that there are certain areas which influenced the choice, which demonstrates the extent to which the system can explain itself.

E) Discussion

The experimental results indicate that Xception model is effective compared to the customized CNN architectures. This is mostly because it has a deep architecture and pretrained feature representations learnt from large-scale datasets. This assists the model in identifying more complex patterns and minute differences between real and AI-generated images. CNN2d_ext model, however, is a bit worse. The reason is that, despite dropout layers preventing overfitting, the presentation of the data can cause a more challenging time for the model to learn fine-grained features. The results show that transfer learning is important in enhancing generalization and classification accuracy. Also, the Grad-CAM usage enhances interpretability, with the visual representation of the areas that influence model judgments, which increases the system openness and trustworthiness.

V. CONCLUSION

This study developed a strong framework on how to classify and identifiable recognise AI-created synthetic images through DL techniques. The suggested solution used different convolutional architectures, such as bespoke CNN models and a pretrained Xception network, to tell the difference between real and fraudulent photos. To ensure that our model was performing well, we used a complete preprocessing pipeline and an effective training method. In an experimental test, the Xception model was the best in all criteria, indicating that transfer learning is an effective method to learn complex visual patterns. The model predictions explained visually through the use of Grad-CAM along with good classification accuracy highlighted significant areas in the input image giving a visual explanation of the model prediction. This simplified the model and gave more confidence to the user. Moreover, by embedding the model into a Flask-based web application, it became feasible to have people interact with the model and obtain predictions in real-time, thus rendering the system applicable in practice. The proposed method manages to bring together precision, understandability, and practicality, giving a holistic solution to the increasing issue of locating AI-created bogus images in the real-life.

The suggested approach may be improved and made more useful in real-life situations by adding more features. The future directions could include the inclusion of additional data sets of numerous sources and the addition of more complex synthetic images to promote generalization. The model can also be used to locate other types of generative models by using it to do multi-class classification. Grad-CAM applied in conjunction with advanced explainability techniques could enable us to comprehend model decisions better. Moreover, it can be made easier to use by just making the system more amenable to faster inference and deploying it on cloud or mobile platforms.

VI. REFERENCES

[1] K. Roose, "An AI-generated picture won an art prize. Artists aren't happy," *The New York Times*, Sep. 2022.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10684–10695.

[3] G. Pennycook and D. G. Rand, "The psychology of fake news," *Trends Cognit. Sci.*, vol. 25, no. 5, pp. 388–402, May 2021.

[4] B. Singh and D. K. Sharma, "Predicting image credibility in fake news over social media using a multi-modal approach," *Neural Comput. Appl.*, vol. 34, no. 24, pp. 21503–21517, Dec. 2022.

[5] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021, pp. 8821–8831.

[6] C. Saharia et al., "Photorealistic text-to-image diffusion models with deep language understanding," *arXiv preprint arXiv:2205.11487*, 2022.

[7] C. Schuhmann et al., "LAION-5B: An open large-scale dataset for training next generation image-text models," *arXiv preprint arXiv:2210.08402*, 2022.

[8] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[9] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022.

[10] J. Gu et al., "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.

[11] Z. Sha, Z. Li, N. Yu, and Y. Zhang, "DE-FAKE: Detection and attribution of fake images generated by text-to-image models," *arXiv preprint arXiv:2210.06998*, 2022.

[12] R. Corvi et al., "On the detection of synthetic images generated by diffusion models," *arXiv preprint arXiv:2211.00680*, 2022.

[13] H. Li, B. Li, S. Tan, and J. Huang, "Identification of deep network generated images using disparities in color components," *Signal Process.*, vol. 174, Art. no. 107616, Sep. 2020.

[14] N. Bonettini, P. Bestagini, S. Milani, and S. Tubaro, "On the use of Benford's law to detect GAN-generated images," in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 5495–5502.

[15] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, "Deepfake video detection through optical flow based CNN," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2019, pp. 1205–1207.

[16] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in *Proc. IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2018, pp. 1–6.

[17] J. Wang et al., "M2TR: Multi-modal multi-scale transformers for deepfake detection," in *Proc. Int. Conf. Multimedia Retr. (ICMR)*, Jun. 2022, pp. 615–623.

[18] P. Saikia, D. Dholaria, P. Yadav, V. Patel, and M. Roy, "A hybrid CNN-LSTM model for video deepfake detection using optical flow features," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 1–7.

[19] D. Gunning et al., "XAI—Explainable artificial intelligence," *Sci. Robot.*, vol. 4, no. 37, Art. no. eaay7120, Dec. 2019.

[20] 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. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.