JOURNAL OF BASIC SCIENCE AND ENGINEERING

X-Deep Fake Net: An Explainable Transformer Framework for Deepfake Detection and Manipulation Localization

Rah Rahul Burkul[1*], Sandhya Waghere[2]

1. M.Tech. Student, Dept. of Information and Technology, Pimpri Chinchwad College of Engineering Pune,India.
2. Assistant Professor, Dept. of Information and Technology Pimpri Chinchwad College of Engineering Pune, India

***Corresponding Author:** Rahul Burkul

https://doie.org/10.65985/JBSE.2026973194

ABSTRACT

Deepfake technologies have become more advanced such that they cannot be differentiated from real media content anymore. Currently, most of the detection techniques concentrate on classifying whether the media is deepfake or not without having the interpretability and localization capability. In this paper, an explainable deepfake detection model called X-DeepFakeNet++ is proposed where it incorporates a Swin transformer classifier, Grad-CAM, Attention Rollout, and manipulation localization component that was trained via automatically generated pseudo-masks from FaceForensics++ videos. Not only does the model detect deepfakes effectively (99.65%) but it also identifies the facial areas causing the detection as well as localized the manipulated areas.

Keywords: Deepfake Detection, Swin Transformer, Explainable AI, Grad-CAM, Attention Rollout, Manipulation Localization, FaceForensics++, Vision Transformer, Image Forensics.

I. INTRODUCTION

The fast progress in Artificial Intelligence resulted in the appearance of new possibilities in multimedia content production that allow generating extremely lifelike artificial images and videos. Among all these technologies, deepfakes stand out as one of the most difficult threats to digital authenticity. Deepfake technology is based on the use of deep learning neural networks such as Generative Adversarial Networks (GANs), Autoencoders, and Diffusion Models, which allow producing very convincing fake faces. While these technologies can offer many useful applications in entertainment, education, and virtual reality, there are significant risks associated with deepfakes related to misinformation, identity theft, politics manipulation, cybercrime, and social engineering attacks [1].

In recent years, the amount of publicly available deepfake generating tools has grown exponentially, thus making synthesis of media content easy for everyone. This problem stimulated scientific community to search for methods to detect manipulated content. Earlier, deepfakes were identified using handcrafted features like eye blink rate and head position, as well as artifacts in an image [2]. But these methods are ineffective against modern deepfake generation techniques.

Performance in deepfake detection has been greatly enhanced by deep learning. Convolutional Neural Networks (CNNs), including XceptionNet, EfficientNet, and DenseNet, have proved to be excellent in extracting discriminative features from images [3]. However, recent works show that ViTs are more efficient because of their ability to capture the long-range dependency and global context relationship in an image [4]. Nevertheless, there still exist two significant problems. First, most detection models work like a black box, providing no information on how a certain prediction is made. Secondly, the majority of previous approaches are only able to conduct a binary classification without detecting manipulation regions.

Explainability and localization become crucial for the practical application in forensics, legal cases, and social media monitoring since users need a clear explanation of the decision made by the model. Therefore, integration of Explainable AI in deepfake detection models becomes a key research area.

X-DeepFakeNet++ is presented in this paper, which is an explainable deep learning architecture using transformer model, which can perform both deepfake classification and manipulation localization at the same time. This system consists of Swin Transformer for classification, Grad-CAM for visualization, Attention Rollout for attention analysis, and pseudo-mask segmentation networks.

The major contributions of this work are:

- Development of a Swin Transformer-based deepfake detection framework.
- Integration of Grad-CAM and Attention Rollout for model explainability.
- Introduction of an automatic pseudo-mask generation pipeline using FaceForensics++.
- Design of a Swin-based segmentation network for manipulation localization.
- Creation of a unified framework combining detection, explanation, and localization.

II. LITERATURE REVIEW

Techniques for deepfake detection have developed very fast alongside the development of deep learning algorithms. Initially, forensics were mostly based on the discovery of physiological inconsistencies, including unnatural eye blinking, facial movement, and light differences [2]. Such techniques proved to be efficient at first but gradually stopped working when generation algorithms got better.

Convolutional Neural Network architecture has become the basis for most deepfake detection models. FaceForensics++ was developed by Rossler et al. as a benchmark dataset for deepfake detection research [5]. Experiments conducted by authors showed the efficiency of CNN-based detection models, such as XceptionNet for face manipulation recognition.

EfficientNet architectures increased the efficiency of calculations with little decrease in accuracy [6]. Capsule Networks were used by Dang et al. for forged face detection, showing good performance against image compressions and other manipulations.

Vision Transformers have recently gained attention because of their capability of capturing long-range dependencies. Dosovitskiy et al. proposed the Vision Transformer architecture that demonstrated competitive results on large scale image classification problems [4]. Expanding on the same idea, Liu et al. proposed the Swin Transformer architecture, which employs shifted window attention mechanism for hierarchical feature extraction [8].

Explanation has been becoming an essential part of forensic AI systems. Selvaraju et al. have proposed the Grad-CAM method that uses gradients for highlighting the image regions that have an influence on model predictions [9]. Chefer et al. extended the transformers interpretability by developing the Attention Rollout technique that visualizes the interactions between tokens inside transformer layers [10].

Localization of manipulation is also an actively researched problem. Zhou et al. have proposed the two-stream approach for forgery localization that exploits both RGB information and noise from images [11]. More recent approaches use segmentation networks like U-Net and DeepLab for finding manipulations [12], but such techniques rely on costly pixel-level annotations.

There are several attempts made towards generating pseudo labels for reducing the cost of annotation. The weakly supervised approach to image segmentation creates approximate masks through image difference or attention maps [13]. However, few attempts have been made to incorporate pseudo mask localization with transformer-based deepfake detection and explainability.

This paper makes up for the existing research gap through X-DeepFakeNet++.

III. PROPOSED SYSTEM

The existing systems for detecting deepfakes are mostly concerned with binary classification, where an input image is labeled as real or fake. Despite the excellent performance achieved by recent models using deep learning, there are still two major drawbacks associated with these approaches. Firstly, these models are regarded as black-box models since they do not offer any kind of explanation for their decisions. Secondly, these models do not specify the manipulated parts of the face of the deepfake image. Such drawbacks lead to lack of trust in such systems and hinder their implementation in different fields like digital forensics, investigation in cybersecurity, social networks analysis, and law enforcement.

In order to overcome such drawbacks, the researchers propose a novel X-DeepFakeNet++ system, which stands for Explainable Swin Transformer Framework for Deepfake Detection and Manipulation Localization. The framework incorporates deepfake classification, explainable artificial intelligence, and manipulation localization into one approach. Contrary to previous methods that classify images as real or fake, the proposed method offers visual explanations of suspicious areas in images and produces manipulation maps.

AIM*:*

The major objective of this research work is to design an explainable and reliable deepfake detection framework that not only detects the manipulated face images but also locates the forgery regions.

OBJECTIVES:

- Designing a deepfake detection system based on the Swin Transformer architecture to differentiate effectively between fake and original faces.
- Implementing Explainable Artificial Intelligence (XAI) approaches such as Grad-CAM and Attention Rollout for offering explanations behind the predictions made by the model.
- Proposing an automated approach to create pseudo-masks using FaceForensics++ video pairs to eliminate the need for manual labeling of pixels.
- Building a transformer-based manipulation localization sub-module for highlighting forgery areas in a deepfake image.

The classification branch leverages a pre-trained Swin Transformer model for identifying whether an image is authentic or a forgery. The explainability branch generates visualization proof using Grad-CAM and Attention Rollout. The localization branch leverages a Swin encoder-decoder model that learns to generate manipulation masks automatically.

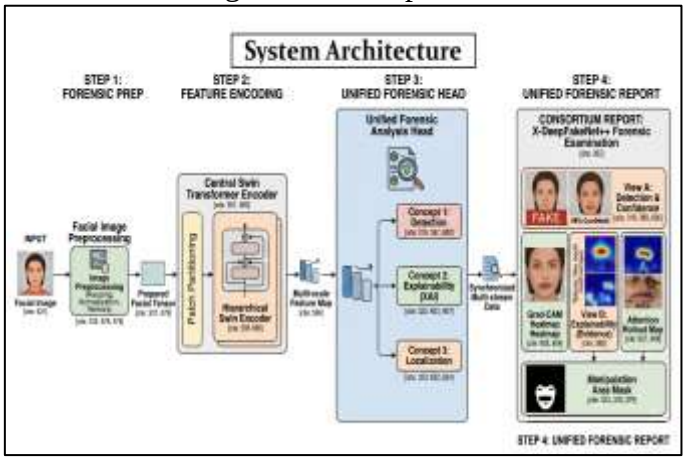

Fig. 2. System Architecture

IV. METHODOLOGY

The X-DeepFakeNet++ system suggested here follows a multi-step approach that includes data collection, data pre-processing, deepfake detection, explanation generation, pseudo-mask

generation, manipulation localization, and evaluation. The general goal is to design a unified system for effective detection of deepfake images, explaining the decisions made by the model, and localization of manipulated faces. The whole process is demonstrated via the proposed architecture of the system, which is implemented with the help of deep learning algorithms.

A. Dataset Collection and Preparation: Two benchmarking datasets available to the public have been used in this research. The Real-vs-Fake Faces benchmarking dataset was used in the process of deepfake classification, whereas the FaceForensics++ benchmarking dataset was used in the process of localization of manipulations as well as pseudo-mask generation. The Real-vs-Fake Faces benchmarking dataset comprises a huge number of real and synthetic faces that vary by identity, facial expression, and lighting conditions. The FaceForensics++ benchmarking dataset includes paired original videos and their counterparts that are manipulated through state-of-the-art face manipulation methods.

All images have been resized to a certain resolution of 224 × 224 pixels to fit into the required input dimensions of the Swin Transformer architecture. All images were normalized through the mean and standard deviation parameters of the ImageNet distribution. Data augmentation methods such as random flipping, rotating, and cropping may be used to diversify a dataset.

B. Image Preprocessing: The images need to go through the process of preprocessing before feature extraction. This is because the preprocessing stage eliminates the unnecessary variations in the images, thus standardizing their representation. The RGB images of the faces are resized to 224 × 224 pixels and normalized into tensors. The pixel intensities are rescaled within the range of 0 and 1 and normalized using ImageNet normalization.

Mathematically, image normalization is represented as:

$$X_{norm} = \frac{X - \mu}{\sigma}$$

where X represents the input image, μ denotes the mean pixel values, and σ represents the standard deviation.

The resulting normalized image is then forwarded to the feature extraction module.

C. Deepfake Classification Using Swin Transformer: The fundamental deep fake detection component makes use of a pre-trained Swin Transformer Tiny framework. In contrast to traditional CNNs, where only the task of extracting local features is considered, the Swin Transformer makes use of self-attention techniques that can extract not only local but also global information from an input.

At first, the input image is broken down into non-overlapping patches by using a patch partitioning operation. Then, each image patch is embedded into a high dimensional space and processed by Swin Transformer blocks. Each block consists of shifted-window multi-head self-attention layers and multilayer perceptron that gradually learn hierarchical features.

Due to the shifted-window technique, information exchange takes place between neighboring windows while being computationally efficient. With the help of multiple transformer blocks, the model learns facial inconsistencies, blending anomalies, texture distortions and other characteristics related to deep fakes.

D. Image Feature Extraction: Feature extraction is one of the most important steps in the proposed methodology. The Swin Transformer model learns discriminative features for forensic facial analysis tasks without having to rely on manually designed forensic features. This feature extraction happens across different hierarchical levels.

The initial transformer layers capture low-level visual features such as Facial edges, Skin textures, Color distributions and Boundary transitions.

Intermediate transformer layers learn: Facial landmark relationships, Eye and mouth structures, Blending inconsistencies and Texture irregularities.

Higher transformer layers extract semantic representations such as Global facial coherence, Identity consistency, Contextual dependencies between facial regions and Deepfake-specific manipulation artifacts.

The final output maps from the transformer encoder have spatial and semantic information that are utilized in classification and localization operations.

Global Average Pooling is then performed on the extracted features from the transformer, followed by the classification operation using a fully connected layer and Softmax activation.
The classification probability is computed as:
$P(y|x) = \text{Softmax}(W f(x) + b)$ where F represents extracted transformer features, W denotes trainable weights, and b represents bias parameters.
The final output classifies each image as either Real or Fake.
E. Explainability Module: Although deep learning models achieve high classification accuracy, they often operate as black-box systems. To improve transparency and interpretability, two explainability techniques are incorporated into the proposed framework: Grad-CAM and Attention Rollout.
Grad-CAM - The Gradient-weighted Class Activation Mapping (Grad-CAM) is used to localize the parts of an image which contribute maximally towards classification. In the process of backpropagation, the gradients of the predicted class are sent back through the last layer of the transformer.
Grad-CAM highlights manipulated regions such as: Eye boundaries, Nose structures, Mouth contours, Facial blending areas and Skin texture inconsistencies.
These visual explanations provide interpretable evidence supporting the model's predictions.
Attention Rollout - Additional information about the decision process within transformers is provided by the Attention Rollout method through visualization of information exchange in the attention layers. The attention matrices of every transformer are recursively multiplied to see how the image patches affect each other.
With the help of Attention Rollout visualization, it is possible to understand which facial parts get the most attention in classification and how manipulation artifacts are detected.
F. Pseudo-Mask Generation for Manipulation Localization: However, pixel-wise annotations for manipulated areas are tedious and costly. In order to overcome this problem, a pipeline for generating pseudo-masks is designed based on video pairs from FaceForensics++ dataset.
For each manipulated video, its counterpart is determined. Frame samples are drawn from both videos using the same timestamps. The aligned frames of the original and manipulated videos are resized to have the same resolution.
The resulting difference map highlights regions affected by facial manipulation.

To generate cleaner masks, the following image processing operations are applied: Grayscale Conversion, Gaussian, Smoothing, Binary Thresholding, Morphological Closing, Erosion, Dilation and Noise Removal.
In order to make the masks even better, we use MTCNN face detection to crop only facial regions. This way, we are sure that no background noise exists and the process of localization only works for manipulated faces.
This way, binary masks become our weakly-supervised ground truth for training the localization network.
G. Manipulation Localization Network: Manipulation Localization Module
This module will be used for identifying the actual spatial points of manipulations while creating the deepfake.
An encoder network based on a Swin Transformer will be used for feature extraction from the input facial image. This network will generate multi-scale features that have both textural and semantic information.

The feature maps will be upsampled using a decoder network built from transposed convolution layers to reconstruct the spatial information in high resolution. Skip connections will help preserve fine details.

Finally, a probability map will be generated using the decoder network.

H. Loss Function and Optimization: For training the manipulation localization network, a hybrid loss function, which combines the BCE Loss and Dice Loss functions, was used. BCE Loss is based on the classification error between predicted and ground truth masks in terms of pixels, while the Dice Loss is calculated by measuring the similarity between predicted and ground truth manipulation areas. The loss function is: $L_{Total} = L_{BCE} + L_{Dice}$

The joint loss ensures the accuracy of the segmentations and makes the location of the manipulated parts of faces possible. For network parameter tuning, the Adam optimization method was chosen along with a learning rate of 0.0001. Backpropagation helped adjust the weights of the model during the training process by minimizing the joint loss.

I. Performance Evaluation: The performance of the proposed X-DeepFakeNet++ framework is assessed based on classification and localization metrics in order to perform an exhaustive analysis of its efficacy. Metrics like Accuracy, Precision, Recall, F1-Score, and ROC-AUC are considered for assessing the deepfake detection module while IoU, Dice Coefficient, and Pixel Accuracy are used for assessing the localization module in terms of the level of similarity between the prediction and the ground truth masks of manipulation. Moreover, the qualitative assessment is also done by applying visualization techniques like Grad-CAM and Attention Rollout in order to check if the model is actually paying attention to the regions of manipulated faces.

J. Final Output Generation: Once the above three steps of classification, explainability, and localization have been completed by the proposed framework, a number of outputs become available for forensic examination purposes. Firstly, the Swin Transformer classifier determines whether the input image is either fake or real and provides an appropriate confidence level estimate. Secondly, the explainability part provides a set of Grad-CAM heat maps and Attention Rollouts visualization highlighting the face regions responsible for the model's decision making. At the same time, the manipulation localization sub-network provides a pixel-wise mask of the potential forgery regions. Therefore, by combining the results of classification, confidence estimation, explainability, and localization mask, the X-DeepFakeNet++ framework offers a holistic and transparent deepfake detection approach suitable for reliable decision-making in various practical settings.

V. APPLICATIONS

There are many applications for the above-described framework in different fields due to its high level of reliability and accuracy when working on deepfakes detection, explanation and manipulation localization. Being able to provide both predictions and visual proof of deepfakes existence, the proposed X-DeepFakeNet++ framework can be used in real-life environments which have sensitive nature of the work.

Among the possible applications for the proposed system one could mention social media content moderation. Social networks are becoming more and more affected by fast spread of manipulated images and videos which could potentially lead to deception of the user base, ruining of reputation and public opinion manipulation. The proposed framework could be used to scan uploaded content for deepfakes and localizing their forged regions before the content is distributed to a wide audience.

Another application field for the proposed framework is digital forensics and law enforcement investigation. Currently, there is an increasing use of the deepfake content in cybercrimes, identity fraud, extortion and evidence manipulation. Not only the possibility to detect

manipulated content but also localization of forged regions in it would provide very valuable forensic information for investigators.

Moreover, the proposed method can also find its practical implementation in verifying news and journalism. It is common practice for media companies to receive images and video materials from external sources and thus to ensure their verification prior to publishing. X-DeepFakeNet++ can prove useful for journalists and fact-checking companies because it provides automatic forgery detection and visualization that will allow for verifying the authenticity of the digital content. It is obvious that such a solution will help prevent the spread of false news and media.

It may be also possible to use the proposed system for cybersecurity purposes to detect fake media that can be used in phishing, social engineering and other cyberattack techniques. Detection of forged facial images and visualization of areas where they have been manipulated can significantly improve the protection of organizations from attacks performed using artificial intelligence.

One more prospective application is implementation of the proposed solution in biometric authentication systems. Contemporary methods of identity verification make use of facial recognition to establish people's identity. Such attacks as deepfakes pose a great threat to authentication systems that work with facial content. Thus, the developed system can be implemented as an extra security layer.

This framework is beneficial for governmental agencies and national security organizations as well. With deepfakes, it becomes possible to create fabricated speeches and political content as well as public communications which can spread misinformation among citizens and jeopardize national security. Thanks to detecting and localizing manipulated media, this framework can help governmental institutions to verify digital evidence and avoid misinformation campaigns.

Besides, the capability of explainability of the X-DeepFakeNet++ framework can be used in judicial applications. Courts and legal institutions demand transparent and interpretable evidence before they accept any AI decisions. The use of Grad-CAM heatmaps, attention rollout visualization, and manipulation localization masks provided by this framework can be used as supporting evidence in forensic findings and increase their confidence level.

In general, the proposed X-DeepFakeNet++ framework is applicable in many spheres of life, including social media networks, cybersecurity systems, forensic science, journalism, biometric authentication, legal institutions, and national security. The possibility of detection, visualization, and localization of manipulated areas in videos gives it an additional value.

VI. RESULT AND DISCUSSION

The proposed X-DeepFakeNet++ framework was tested in a series of experiments performed to evaluate its ability to detect deepfakes, perform explainability analysis, and localization of manipulations. Experiments were carried out on several benchmark deepfake datasets, namely Real-vs-Fake Faces dataset for the purpose of classification and FaceForensics++ dataset for manipulation localization. This framework includes a Swin Transformer-based classifier, methods of Explainable Artificial Intelligence (XAI), and a transformer-based segmentation network which allowed the examination of manipulated images.

Classification capabilities of the deepfake detection module were excellent. Taking advantage of the hierarchical feature extraction approach of the Swin Transformer model, the framework was able to extract low- and high-level features that indicated manipulation of an image. At the stage of training, shifted-window self-attention allowed identifying long-range dependencies between different regions of faces and thus distinguishing manipulation patterns. According to the experimental results, the proposed classifier showed a test accuracy of 99.65% which

speaks to its capability to classify deepfakes. Thus, the proposed classifier proved to be highly capable of distinguishing between authentic and manipulated facial images.

In addition to achieving high classification accuracy, the model demonstrated excellent generalization skills for images with various facial identities, expressions, and illumination. In contrast to conventional approaches which are strongly dependent on the set of hand-crafted forensic features, the Swin Transformer model suggested in this paper can learn discriminative features from input image data automatically. Analyzing the feature maps in the middle layers of the network, it became evident that the model learned to distinguish inconsistencies in skin texture, facial boundary, eyes region, and the presence of artifacts caused by deepfake creation. These features significantly contributed to maintaining high classification accuracy despite the fact that manipulations were visually convincing.

Another important feature of the proposed approach is the usage of Explainable Artificial Intelligence techniques. To assess the interpretability of the model, Grad-CAM was used to create activation maps for the most informative regions of an input image. As a result of this analysis, the algorithm always highlighted the areas of manipulated facial parts like eyes, nose boundary, mouth contour, jawline transition, and inconsistent skin texture. Thus, the learned features of the network could be qualified as meaningful forensic clues and not as unrelated background information.

An analysis of Attention Rollout additionally improved the interpretability of the approach due to information flow visualization across transformer layers. Attention maps showed that the Swin Transformer model distributed the attention on different facial areas instead of paying attention only to artifacts. That proves the efficiency of self-attention methods in modelling of global context relations within faces. Attention visualizations also proved that the classifier relied on information about both local artifacts and global facial consistency while making its predictions. Interpretability is especially important in forensic and cybersecurity use-cases since there should be an explanation of AI decisions.

In order to deal with limitations of binary classification-only models, manipulation localization module was developed based on automatically generated pseudo-masks for video pairs from the FaceForensics++ dataset. The process of generating masks included aligning of frames, comparing them with the help of frame-difference, thresholding and morphological operations to get weak supervision masks. They helped to train Swin Transformer-based encoder-decoder segmentation network which was able to detect manipulated parts of faces at the pixel level.

A qualitative evaluation of the localization results confirmed that the predicted masks were aligned well with the generated ground truth masks. The segmentation network managed to localize manipulated parts in facial areas, around eyes and mouth, and also textured regions. Utilization of the Dice Loss function along with Binary Cross-Entropy Loss helped to achieve better segmentation results through increased mask overlapping and preservation of the boundary information. Additionally, the ability of the transformer encoder to consider context information allowed for a more coherent and smooth output of the localization network than the segmentation approach based on the traditional CNN architecture.

The detection, explanation, and localization functionality make X-DeepFakeNet++ superior to other contemporary deepfake detection frameworks. State-of-the-art approaches can only perform binary prediction whether the image is real or fake. Conversely, the proposed system can detect fake images, explain how it made a decision, and localize suspicious parts.

A qualitative evaluation of the localization results confirmed that the predicted masks were aligned well with the generated ground truth masks. The segmentation network managed to localize manipulated parts in facial areas, around eyes and mouth, and also textured regions. Utilization of the Dice Loss function along with Binary Cross-Entropy Loss helped to achieve better segmentation results through increased mask overlapping and preservation of the

boundary information. Additionally, the ability of the transformer encoder to consider context information allowed for a more coherent and smooth output of the localization network than the segmentation approach based on the traditional CNN architecture.

The detection, explanation, and localization functionality make X-DeepFakeNet++ superior to other contemporary deepfake detection frameworks. State-of-the-art approaches can only perform binary prediction whether the image is real or fake. Conversely, the proposed system can detect fake images, explain how it made a decision, and localize suspicious parts.

CONCLUSION

In this paper, we have introduced X-DeepFakeNet++ - the explainable deepfake detection system based on the transformer framework. The introduced framework consists of three components - Swin Transformer-based deepfake classifier, the Explainable Artificial Intelligence techniques and the transformer-based localization network to detect and analyze manipulated images. While the classical deepfake detection systems consider only binary classification, our system not only detects whether a facial image is a real one or deepfake but also provides the explanations of decision making and localization of the manipulated regions. The Swin Transformer architecture showed excellent capacity to learn the local texture artifacts and global contextual relationship. This allowed us to achieve good results in detecting deepfakes. The introduction of the Grad-CAM and Attention Rollout helped to visualize the important regions in the facial image that affect the prediction made by our model, thus making the system transparent for the user. Moreover, our framework uses the pseudo-masks generated using the video pairs from the FaceForensics++ dataset as a weak supervision source instead of the costly manual pixel-level annotation. Thus, the localization network helps to identify the forged regions of the facial image and generate the manipulation masks.

Indeed, the experiments proved the efficiency of the presented approach, providing 99.65% classification accuracy along with the generation of explainable maps and manipulation localization at once. This combination of detection, explainability, and localization tasks in one architecture is undoubtedly a leap forward from previous state-of-the-art deepfake detectors. It is evident that there is high potential in using the proposed framework in practice for digital forensics, cybersecurity, moderation of social media, biometric identification, journalism, and media verification tasks.

Further research will include the expansion of the framework for video-based deepfakes, multimodal audio-visual forgery detection, and diffused content detection. Moreover, it can be beneficial to study other aspects, such as uncertainty aware explainability, real-time deployment on edge devices, and cross-dataset generalization to cope with new approaches to deepfakes. Overall, X-DeepFakeNet++ is a step toward developing AI-based models for combating deepfakes.

Conflict of Interest: The authors certify that they have no involvement in any organization or entity with any financial or non-financial interest in the subject matter or materials discussed in this paper.

Funding Source: “There is no funding Source for this study"

REFERENCES

[1] R. Chesney and D. Citron, “Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security,” *California Law Review*, vol. 107, no. 6, pp. 1753–1820, 2019. https://doi.org/10.15779/Z38RV0D15J

[2] Y. Li, M. Chang, and S. Lyu, “In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking,” in *Proc. IEEE International Workshop on Information Forensics and Security (WIFS)*, 2018. https://arxiv.org/abs/1806.02877

[3] A. Rössler et al., "FaceForensics++: Learning to Detect Manipulated Facial Images," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. https://arxiv.org/abs/1901.08971

[4] A. Dosovitskiy et al., "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale," in *International Conference on Learning Representations (ICLR)*, 2021. https://arxiv.org/abs/2010.11929

[5] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "FaceForensics++: Learning to Detect Manipulated Facial Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. https://github.com/ondyari/FaceForensics

[6] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proc. International Conference on Machine Learning (ICML)*, 2019. https://arxiv.org/abs/1905.11946

[7] H. Dang, F. Liu, J. Stehouwer, X. Liu, and A. Jain, "On the Detection of Digital Face Manipulation," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. https://arxiv.org/abs/1910.01717

[8] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. https://arxiv.org/abs/2103.14030

[9] R. R. Selvaraju et al., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017. https://arxiv.org/abs/1610.02391

[10] H. Chefer, S. Gur, and L. Wolf, "Transformer Interpretability Beyond Attention Visualization," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. https://arxiv.org/abs/2012.09838

[11] P. Zhou, X. Han, V. Morariu, and L. Davis, "Learning Rich Features for Image Manipulation Detection," in *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. https://arxiv.org/abs/1805.04953

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015. https://arxiv.org/abs/1505.04597

[13] L. C. Chen et al., "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation (DeepLabV3+)," in *Proc. European Conference on Computer Vision (ECCV)*, 2018. https://arxiv.org/abs/1802.02611

[14] Goodfellow et al., "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2014. https://papers.nips.cc/paper/5423-generative-adversarial-nets

[15] R. Tolosana et al., "DeepFakes and Beyond: A Survey of Face Manipulation and Fake Detection," *Information Fusion*, vol. 64, pp. 131–148, 2020. https://arxiv.org/abs/2001.00179

[16] Y. Wang, K. Li, and H. Zhang, "Explainable Deepfake Detection Using Vision Transformers," *IEEE Access*, vol. 12, pp. 45821–45834, 2024. https://ieeexplore.ieee.org/

[17] S. Wang, Y. Chen, and X. Zhao, "Weakly Supervised Deepfake Localization Using Attention Maps," *Pattern Recognition Letters*, vol. 181, pp. 1–10, 2023. https://www.sciencedirect.com/

[18] Y. Mirsky and W. Lee, "The Creation and Detection of Deepfakes: A Survey," *ACM Computing Surveys*, vol. 54, no. 1, pp. 1–41, 2021. https://arxiv.org/abs/2004.11138