International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI: 10.55041/ISJEM06011
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Face Swap Detection using Deep Learning Techniques

Dr.M.V.Krishna Mohan,Associate Professor
Ragolu Vineetha, Vindula Devi Priyanka, Pappu Siva Sai Prasad, Padala sai venkata Prasad

Department of Computer Science and Engineering
Visakha Institute of Engineering and Technology, Visakhapatnam, Andhra Pradesh, India

Abstract

Face swap and deepfake manipulations have become a *serious threat to the authenticity of digital images and videos. Such manipulations can be misused for spreading misinformation, identity fraud, and damaging public trust. Manual checking is slow and not suitable for large-scale use. This paper presents* a *deep learning-based system for automatic face swap detection. The system uses OpenCV for face detection and alignment combined with* a *fine-tuned ResNet50 model for classification. It distinguishes real faces from swapped faces and achieves* a *validation accuracy of 98.75%. Gradient-weighted Class Activation Mapping (Grad-CAM) is integrated to highlight manipulated regions such as blending artifacts and texture inconsistencies, making the results more understandable. The model is deployed as* a *simple web application that provides real-time predictions, confidence scores, visual explanations, and useful recommendations. The system also supports multiple languages to improve accessibility. Experimental results show that the proposed approach is accurate, interpretable, and practical for real-world digital forensics and content moderation applications.*

Keywords:Face *Swap Detection, Deepfake Detection, Deep Learning, ResNet50, OpenCV, Grad-CAM, Digital Forensics*

1. Introduction

In recent years, digital images and videos have become the pnmary medium of communication across social media, news platforms, and online services. However, the rapid development of face swap and deepfake technologies has made it increasingly difficult to trust visual content. Face swap tools can seamlessly replace a person's face in an image or video with another person's face, often producing highly realistic results that are difficult to detect with the naked eye. These manipulations are frequently misused for creating fake news, political propaganda, revenge pornography, financial fraud, and identity theft, leading to serious social, legal, and security concerns.Traditional manual verification methods are highly time-consuming, subjective, and impractical when dealing with the massive volume of media uploaded every day. Early detection techniques based on handcrafted features such as color histograms or edge detection often fail under varying lighting conditions, compression, or different angles.

Deep learning approaches, particularly Convolutional Neural Networks (CNNs), have shown promising results by automatically learning complex patterns and subtle manipulation artifacts like unnatural skin textures, inconsistent lighting between face and background, blending boundaries around the face, and landmark mismatches. Among various CNN architectures, ResNet50 is widely preferred due to its residual learning framework, which allows training of deeper

networks without degradation in performance and enables effective extraction of hierarchical features from face images.To make the system robust for real-world scenarios, OpenCV is integrated for essential preprocessing tasks such as accurate face detection using its DNN module or Haar cascades, facial landmark extraction for proper alignment, and normalization. This preprocessing step significantly improves the quality of input data fed to the deep learning model.Furthermore, accuracy alone is not sufficient for practical applications. Users, especially non-technical ones such as journalists, law enforcement officers, and social media moderators, need to understand why the system classified an image as swapped. To address this, Gradient-weighted Class Activation Mapping (Grad-CAM) is incorporated to generate visual heatmaps that highlight the specific regions of the face contributing to the model's decision.

The proposed system is developed as a complete end-to-end web-based application that allows users to upload images or short video frames and receive instant results. This work aims to provide an accurate, interpretable, and user-friendly solution that can be practically deployed for digital forensics and online content moderation.

2. Design

The proposed system is designed as an end-to-end pipeline for face swap detection. It includes image input, OpenCV-based preprocessing, feature extraction using ResNet50, binary classification (Real or Swapped), Grad-CAM explainability, and result display through a web interface.

Figure

Architecture of the proposed face swap detection system.

[Start]
Input Image/Video **Face Detection**
Preprocessing

Feature Extraction (CNN) **Classification Layer**
Output (Real/ Fake)
t
[End]

The workflow begins when a user uploads a face image or a short video clip through the web interface. OpenCV first detects the face region using its robust DNN face detector or Haar cascade classifier. It then extracts 68 facial landmarks to align the face properly, crops the aligned face, resizes it to 224x224 pixels, and normalizes the pixel values using ImageNet mean and standard deviation. These steps ensure consistency and remove variations caused by pose, scale, and background noise.

The preprocessed face image is then passed to a fine-tuned ResNet50 model based on transfer learning. The pre-trained convolutional layers extract rich spatial and texture features, while the custom top layers (a fully connected layer with 512 neurons followed by dropout and a 2-neuron softmax output) perform binary classification into "Real" or "Swapped". After classification, Grad-CAM is applied on the last convolutional layer to produce a heatmap that visually highlights manipulated areas such as blending seams around the jawline, eyes, or forehead.

The Flask-based web application integrates all modules and displays the original image, predicted class, confidence score, Grad-CAM heatmap, manipulation analysis (describing detected artifacts), and practical recommendations (such as further

forensic verification). Multilingual support is added using translation libraries so that results can be viewed in English, Hindi, Telugu, and other languages.

Table 1: Comparison of detection accuracy with existing methods

Method	Model	Accuracy (%)
MesoNet	MesoNet	195.59
Rossler et al.	Xception	98.60
Proposed Method	ResNet50 + OpenCV	198.75

3. Analysis

The methodology of the proposed system involves careful dataset preparation, robust preprocessing, model training, and integration of explainable Al. The dataset is constructed by combining real and manipulated face images extracted from two well-known public benchmarks: FaceForensics++ and Celeb-DF. Celeb-DF contains 590 real celebrity videos and 5,639 high-quality deepfake videos, providing diversity in age, gender, ethnicity, lighting conditions, and compression levels. This diversity helps the model learn generalized features rather than overfitting to specific manipulation styles.

All images are preprocessed using OpenCV. Face detection is performed first, followed by landmark-based alignment to standardize the face orientation. The aligned faces are cropped and resized to 224x224 pixels to match the input size of ResNet50. Data augmentation techniques such as random rotation (±15°), horizontal flipping, brightness and contrast adjustment, and addition of mild Gaussian noise are applied only during training to improve the model's robustness against real-world variations.

The ResNet50 model is initialized with pre-trained lmageNet weights. Transfer learning is applied in two phases. In the first phase, only the newly added classification layers are trained with a learning rate of 0.001 using the Adam optimizer. In the second phase, the entire network is fine-tuned with a lower learning rate of 0.0001 to refine the feature representations. Binary cross-entropy loss is used, and the model is trained with a batch size of 64 on GPU-enabled environment for faster convergence.

After training, Grad-CAM is implemented to generate class activation heatmaps from the final convolutional layer. This helps in understanding whether the model is focusing on meaningful manipulation artifacts or irrelevant background regions. The complete system is deployed as a Flask web application with an interactive frontend developed using **HTML,** CSS, and JavaScript. A statistical dashboard and crop-specific (face forensic) information are also included to assist users in better decision-making.

4.Result

The proposed deep learning model was evaluated on the validation set and achieved a training accuracy of 99.20% and a validation accuracy of 98.75%. The close gap between training and validation performance indicates minimal overfitting and strong generalization capability across different face poses, lighting conditions, and manipulation techniques.

The confusion matrix analysis revealed high diagonal dominance, meaning most real and swapped faces were correctly classified with very low misclassification rates. Precision, recall, and F1-score values were balanced for both classes, confinning the reliability of the system in practical scenarios where both false positives and false negatives carry significant cost.

Figure

****Table 2:** Classification performance metrics**

I Class I Precision I Recall I Fl-Score I

|----------|----------|--------| ---------- I

I Real I 0.99 I 0.98 I 0.985 I

I Swapped I 0.98 I 0.99 I 0.985 I

Grad-CAM heatmaps clearly highlight blending artifacts and texture inconsistencies m swapped faces while showing uniform activation in real faces. The system runs efficiently in real time on the web platform.

4. Conclusion

.This paper presented a deep learning-based face swap detection system using OpenCV preprocessing and a fine-tuned ResNet50 model. The system achieves 98.75% validation accuracy and provides clear visual explanations through Grad-CAM. It is deployed as a
user-friendly web application with multilingual support, making it suitable for practical use in digital forensics and media authentication.Future work will focus on full video analysis, mobile deployment, and improving robustness on low-quality real-world images.

References

[l] D. Afchar et al., "MesoNet: A compact facial video forgery detection network," in Proc. IEEE WIFS, 2018.

[2] A. Rossler et al., "FaceForensics++: Leaming to detect manipulated facial images," in Proc. IEEE ICCV, 2019.

[3] Y. Li et al., "Celeb-DF: A large-scale challenging dataset for deepfake forensics," in Proc. IEEE CVPR, 2020.

[4] K. He et al., "Deep residual learning for image recognition," in Proc. IEEE CVPR, 2016.

[5] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE ICCV, 2017.