Efficient DeepFake Image Classification Using Lightweight MobileNetV4-Small Architecture

Savita Sidnal, Shradha Kekare, Soujanya Menasagi,
Vaishnavi Tharakar, Uday Kulkarni and Shashank Hegde
School of Computer Science and Engineering, KLE Technological University, Hubli, Karnataka, India

Keywords: Deep Learning, Deepfake Detection, Lightweight Neural Networks, MobileNetV4-Small, Synthetic Media.

Abstract: The rise of Deep Learning (DL) has unlocked a wide range of transformative applications, including the creation of hyper-realistic synthetic images through Generative Adversarial Networks (GANs). While these images demonstrate the immense potential of DL, they also pose significant risks, such as misuse in cybersecurity breaches, political manipulation, and disinformation campaigns. This paper proposes a robust approach for deepfake detection using MobileNetV4-Small, a lightweight and efficient DL model. Leveraging advanced preprocessing techniques, the proposed method enhances the ability to distinguish counterfeit images from authentic ones. The study utilized a dataset containing real and fake images, achieving a notable test accuracy of 89.78%.The model's performance was further analyzed through visual evaluation of classification results. This work underscores the efficacy of lightweight architectures in addressing the challenges posed by deepfake media and provides a comparative analysis with existing approaches. Future enhancements could involve ensemble techniques and expanded datasets to further improve accuracy and generalization. The results affirm the critical role of DL models in mitigating the risks associated with synthetic media.

1 INTRODUCTION

The advancements in Deep Learning (DL)(Mienye and Swart, 2024) have unlocked transformative applications across diverse fields such as education, entertainment, and healthcare. However, these advancements have also given rise to deepfake technology(Croitoru et al., 2024), a highly debated and morally complex innovation. Deepfakes leverage advanced machine learning models, particularly Generative Adversarial Networks (GANs)(Ahmad et al., 2024), to create hyper-realistic images and videos, blurring the boundaries between reality and fabrication. This capability has raised significant concerns in areas such as politics(Appel and Prietzel, 2022), journalism, cybersecurity (Broklyn et al., 2024), and individual privacy.

While deepfakes offer legitimate applications—such as digital education, lifelike animations, and historical preservation—they also pose alarming risks. Malicious actors can exploit them for disinformation, identity theft, or reputational harm. For instance, a deepfake video of a political leader spreading false statements could incite unrest or destabilize societal systems. Consequently, the development of effective detection mechanisms is essential to mitigate these threats and safeguard against potential misuse.

Despite notable progress in deepfake detection, existing methods often rely on resource-intensive models that are unsuitable for deployment on mobile or edge devices. This creates a critical gap, as lightweight and efficient detection systems are urgently needed to ensure accessibility and scalability in real-world scenarios.

To address this challenge, the DeepFake Image Classification explores MobileNetV4-Small (Qin et al., 2025), a lightweight Convolutional Neural Network (CNN) (Patel et al., 2023) designed for efficiency and portability. MobileNetV4-Small balances computational simplicity with high performance, making it an ideal candidate for real-time deployment on resource-constrained devices. By fine-tuning the model to detect subtle artifacts unique to deepfakes and employing optimization techniques, this work aims to deliver an effective yet scalable detection framework.

Section 2 presents the background study, while Section 3 describes the proposed deepfake image detection approach and the MobileNetV4-Small ar-

Sidnal, S., Kekare, S., Menasagi, S., Tharakar, V., Kulkarni, U. and Hegde, S.
Efficient DeepFake Image Classification Using Lightweight MobileNetV4-Small Architecture.
DOI: 10.5220/0013613400004664

In *Proceedings of the 3rd International Conference on Futuristic Technology (INCOFT 2025) - Volume 3*, pages 247-253
ISBN: 978-989-758-763-4

chitecture and preprocessing steps. Section 4 provides details about the dataset and accuracy analysis, demonstrating the effectiveness of MobileNetV4-Small in detecting deepfake images.

2 BACKGROUND STUDY

Deepfake detection has become an increasingly important research area, driven by the rapid evolution of generative technologies such as Variational AutoEncoders (VAEs)(Kingma et al., 2019) and GANs (Goodfellow et al., 2014). These methods allow for highly realistic manipulations, with applications in face swapping, reenactment (Nirkin et al., 2019), and other forms of image tampering (Zheng et al., 2019). To counter these challenges, researchers have developed various detection methods (Heidari et al., 2024), broadly categorized into traditional forensic techniques, CNN-based methods, and lightweight architectures.

Early approaches to detecting manipulated media relied on forensic techniques that analyzed inconsistencies in lighting, pixel values, and compression artifacts. While these methods could identify simple manipulations, they struggle with highly sophisticated deepfake techniques enabled by advanced GANs. For example, techniques like Face2Face (Thies et al., 2016) and Neural Textures (Thies et al., 2019) leverage 3D modeling and photometric reconstruction to produce highly realistic results, making detection through traditional methods increasingly difficult.

The introduction of CNNs revolutionized deepfake detection by automating feature extraction and analysis. Models like ResNet (Targ et al., 2016) and VGG (Tammina, 2019) demonstrated high accuracy in detecting artifacts in manipulated media. ResNet-50 achieved up to 95% accuracy on datasets like Celeb-DF (Li et al., 2020), while EfficientNet (Tan and Le, 2019), with its balance of computational efficiency and accuracy, has been a preferred choice for many applications. Transformers, such as Vision Transformers (Khan et al., 2022), have also emerged as a promising alternative, providing strong generalization across datasets, though at a higher computational cost.

Given the need for real-time and resource-efficient solutions, lightweight models such as MobileNet, ShuffleNet (Zhang et al., 2018), and SqueezeNet (Bhuvaneswari and Enaganti, 2023) have gained popularity. These architectures balance accuracy and computational requirements, making them suitable for deployment on devices with limited resources, such as mobile phones. MobileNetV2 achieved 89% accuracy on datasets like FaceForensics++(Rossler et al., 2019), but its limitations in capturing subtle manipulation details restrict its use in more challenging scenarios. ShuffleNet and SqueezeNet also show promising results, with ShuffleNetV2 achieving 88.7% accuracy and SqueezeNet achieving 87.5% accuracy on similar datasets. However, these models tend to compromise on the ability to detect subtle manipulation artifacts and are limited by the lack of advanced modules for feature extraction.

MobileNetV4 and its compact variant, MobileNetV4-Small, further enhance computational efficiency while maintaining high detection accuracy. MobileNetV4-Small leverages advanced techniques such as depthwise separable convolutions and squeeze-and-excitation modules, optimizing feature extraction and ensuring the model can efficiently process complex data while using fewer resources. In comparison to MobileNetV2, which achieves 89% accuracy, MobileNetV4-Small surpasses this by achieving 89.78% accuracy with a lower test loss of 0.2648. This makes MobileNetV4-Small more efficient and capable of detecting subtle manipulations as well as deployment on resource-constrained devices , a crucial aspect for deepfake detection in real-world applications.

Despite the progress made by lightweight architectures, several challenges remain unresolved. One major issue is the generalization of detection models across different types of manipulations. Many models tend to overfit specific datasets, limiting their ability to detect unseen manipulations. Additionally, resource-intensive models like ResNet and Vision Transformers are impractical for deployment on devices with constrained computational resources. These limitations highlight the need for more efficient, generalizable, and lightweight models that can perform well across diverse manipulation techniques and be deployed on resource-constrained devices.

The proposed preprocessing pipeline improves generalization by applying augmentations like random rotation and color jittering, enhancing the model's robustness to real-world variations. Combined with the compact MobileNetV4-Small architecture, which ensures computational efficiency, this approach makes real-time deepfake detection feasible on lightweight devices. By addressing the limitations of existing methods, our solution offers an efficient and accurate deepfake detection model, optimized for resource-constrained environments, with the next section elaborating on the methodology and key innovations.

3 PROPOSED WORK

The MobileNetV4-Small model is a lightweight architecture optimized for resource-constrained environments like mobile and edge devices. It balances computational efficiency, feature extraction, and classification accuracy, making it ideal for real-time applications such as deepfake image classification. The following subsections highlight the key components and their specific role in detecting deepfake images.

3.1 Model Architecture

The Figure.1 describes the backbone model, MobileNetV4-Small, used for feature extraction, leveraging its specialized convolutional layers for efficient processing. The extracted features are refined through dimensionality reduction before being passed to a dense layer. To enhance generalization and minimize overfitting, techniques like dropout or normalization are applied before the dense layer. Finally, a binary classifier with a sigmoid activation function ensures accurate classification of the input.

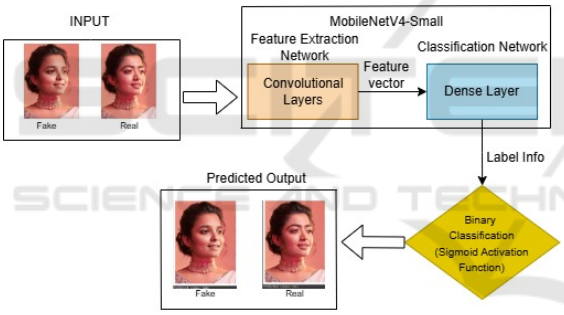

Figure 1: Proposed Work.

3.2 Proposed Architecture Components

Depthwise Separable Convolutions (Kaiser et al., 2017) reduce computational complexity by splitting a standard convolution into two steps: depthwise convolution (spatial filtering) and pointwise convolution (channel-wise interaction). This reduces the computational cost from Equation (1) to Equation (2):

$$O(M \cdot N \cdot K^2) \tag{1}$$

$$O(M \cdot K^2 + M \cdot N) \tag{2}$$

where M is the output channels, N is the spatial dimensions, and K is the kernel size. This optimization allows lightweight architectures to process spatial features more efficiently, making them well-suited for large-scale datasets in deepfake classification.

Universal Inverted Bottleneck (UIB) blocks (Qin et al., 2025), as depicted in Figure 2, expand the input feature dimension F_{in} by a factor t, as shown in Equation (3):

$$F_{\text{exp}} = t \cdot F_{\text{in}} \tag{3}$$

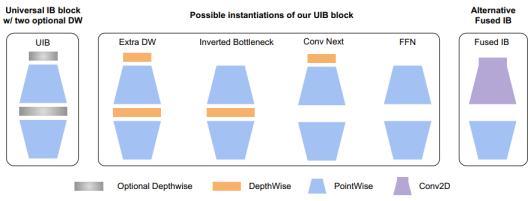

Figure 2: UIB Blocks (Qin et al., 2025).

followed by depthwise convolutions and projection back to F_{out}. These blocks efficiently capture hierarchical structures, including subtle distortions introduced by manipulations, while maintaining computational efficiency.

Fused Inverted Bottleneck Blocks (Tan and Le, 2021), shown in Figure 3, integrate activation and batch normalization directly within the block. These blocks transition between narrow and wide layers, optimizing both computational cost and feature extraction. They effectively identify deepfake artifacts, such as texture irregularities and lighting inconsistencies.

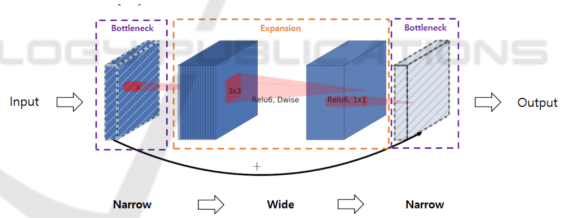

Figure 3: Fused Inverted Bottleneck Blocks architecture (Sandler et al., 2018).

Squeeze-and-Excitation (SE) blocks (Hu et al., 2018) use attention mechanisms to recalibrate channel-wise feature responses. For an input tensor X with dimensions (C,H,W), the global context vector s_c is calculated as shown in Equation (4):

$$s_c = \frac{1}{H \cdot W} \sum_{h=1}^{H} \sum_{w=1}^{W} X_{c,h,w} \tag{4}$$

This vector is used to scale X with learned weights, emphasizing features indicative of manipulations, such as unnatural patterns in facial expressions or mismatched shadows.

The ReLU6 (Howard, 2017) activation function, defined as shown in Equation (5), is employed:

$$f(x) = \min(\max(0,x),6) \tag{5}$$

where x is the input. ReLU6 prevents overflow and ensures stability during low-precision computations, which is crucial for distinguishing subtle differences between real and fake images.

Finally, the head and tail layers (Howard et al., 2019) efficiently manage feature extraction and classification. The head extracts low-level features through convolution, batch normalization, and activation, while the tail reduces spatial dimensions via adaptive average pooling, producing a feature vector F_{tail}. The final predictions are computed using the softmax function, as shown in Equation (6):

$$P(y_i) = \frac{\exp(W_i \cdot F_{\text{tail}})}{\sum_j \exp(W_j \cdot F_{\text{tail}})} \tag{6}$$

This combination ensures robust feature extraction and accurate classification of deepfake artifacts.

3.3 Preprocessing

The dataset preprocessing pipeline was specifically designed to ensure compatibility with the MobileNetV4-Small architecture and improve the model's robustness. The dataset was organized into Train, Test, and Validation directories, each containing subdirectories for real (genuine) and fake (manipulated) images.

For training, the images were resized to 224×224 pixels, which aligns with the input dimensions required by the MobileNetV4-Small architecture. To further improve model generalization, a series of augmentation techniques were applied. These included horizontal flipping with a probability of 0.5, rotation within a range of $\pm 10^\circ$, and color jittering, which adjusted the brightness, contrast, saturation, and hue of the images. These transformations helped the model learn a wider range of features and better handle variations in the data.

Normalization was performed using the ImageNet statistics, specifically the mean $[0.485, 0.456, 0.406]$ and standard deviation $[0.229, 0.224, 0.225]$. This step, represented by the Equation (7) , ensured that the pixel values were standardized to the same scale as the pre-trained weights used by MobileNetV4-Small.

$$x_i' = \frac{x_i - \mu}{\sigma} \tag{7}$$

where x_i is the pixel value, μ is the mean, and σ is the standard deviation.

For the test and validation sets, only resizing and normalization were applied, without any augmentation. This approach was intended to evaluate the model's performance on unaltered images. To streamline the loading and processing of the dataset, PyTorch's ImageFolder class was used to organize the data, and data loaders were configured with a batch size of 32 for efficient training and evaluation.

3.4 Training and Evaluation

The MobileNetV4-Small model was trained over 11 epochs using a deep learning framework, with the training process incorporating the model, training data, loss function, optimizer, and epochs. The model was trained using a cross-entropy loss function as demonstrated in Equation (8).

$$L(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(p_{i,c}) \tag{8}$$

where N is the batch size, C is the number of classes, $y_{i,c}$ is the true label, and $p_{i,c}$ is the predicted probability for class c. The optimizer employed backpropagation and the Adam algorithm to iteratively update model parameters, minimizing the loss as outlined in Equation (9).

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta L(\theta) \tag{9}$$

where η is the learning rate and $\nabla_\theta L(\theta)$ is the gradient. After each epoch, the model's accuracy was calculated using Equation (10).

$$A = \frac{1}{N} \sum_{i=1}^{N} 1(y_i = \hat{y}_i) \times 100 \tag{10}$$

where 1 is an indicator function. Validation was performed at the end of each epoch to assess generalization, and the model was saved after training for testing on a separate test set. During testing, the model's accuracy on the test data was calculated similarly to the training phase. The final test accuracy of 89.78% and loss evaluated the model's effectiveness in deepfake detection, demonstrating its ability to distinguish between real and manipulated images.

The effectiveness of these innovations is evident in the results of the model's evaluation on the test dataset. The following section presents the performance metrics and detailed analysis of the model's ability to detect deepfake images, showcasing its robust capabilities in both qualitative and quantitative assessments.

4 RESULTS

This section presents the experimental results, including both qualitative and quantitative analysis, and provides an overview of the dataset.

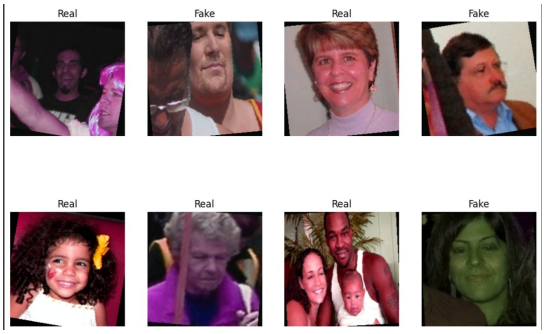

Figure 4: Sample images from the OpenForensics dataset. The figure illustrates examples of genuine (real) images and manipulated (fake) images, highlighting the focus on facial regions where deepfake techniques have been applied.

4.1 Dataset Description

The OpenForensics Dataset (Karki, 2024) for Deepfake Detection is designed for benchmarking deepfake detection methods, organized into Train, Test, and Validation directories with subdirectories for genuine (real) and manipulated (fake) images. Manipulated images focus on facial regions with applied deepfake techniques, while backgrounds remain unchanged, ensuring models focus on facial anomalies. Images in standard formats (JPEG, PNG) ensure compatibility with deep learning frameworks. The dataset includes 140,002 training images (70,001 real, 70,001 fake), 10,905 test images (5,413 real, 5,492 fake), and 39,428 validation images (19,787 real, 19,641 fake), enabling robust model evaluation.The sample images from the dataset,illustrating both real and fake images, are shown in Figure. 4

4.2 Qualitative Analysis

The MobileNetV4-Small model was trained on the training dataset and evaluated on the test dataset to detect real and fake images effectively. The model achieved a test loss of 0.2648 and a test accuracy of 89.78%, indicating its robust performance in distinguishing between real and fake images.

The Figure 5 illustrates the MobileNetV4-Small model's ability to distinguish "Fake" images with facial manipulations from "Real" images, even when backgrounds are altered. The model accurately classifies images as "Fake" when facial features, expressions, or visual artifacts are modified, demonstrating its robustness in detecting manipulations

To further evaluate the model's classification performance, a confusion matrix is presented in Figure 6, which provides insights into the true positive, true negative, false positive, and false negative predictions made by the model.

Figure 5: Top: Real with altered backgrounds. Bottom: Fake (left) vs. Real (right) based on facial edits

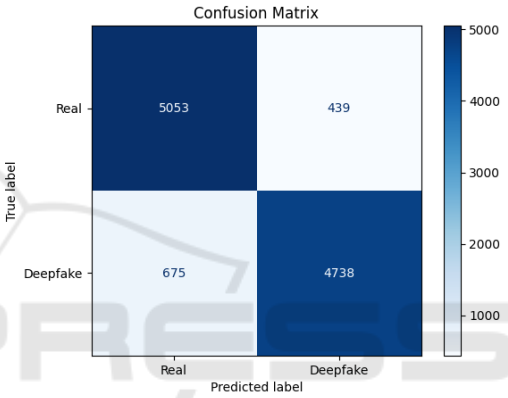

Figure 6: Confusion Matrix for MobileNetV4-Small on the Test Dataset

4.3 Quantitative Analysis

The Table 1 showcases the performance of various models for deepfake image classification, with MobileNetV4-Small achieving an accuracy of **89.78%** on the OpenForensics dataset, which emphasizes facial manipulation detection. In addition to its accuracy, the model exhibits strong quantitative metrics, including a precision of **91.52%**, recall of **87.53%**, F1-score of **89.48%**, and a high ROC-AUC of **96.81%**. These metrics underscore its ability to effectively detect subtle manipulations, such as blending inconsistencies and unnatural textures, even in challenging scenarios.

MobileNetV4-Small's lightweight and efficient architecture makes it ideal for resource-constrained environments, balancing performance and scalability effectively. Its robust design ensures generalization across diverse manipulations, including variations in lighting, facial expressions, and angles. Data augmentation during training further enhances its ability to detect deepfake artifacts. With its compact size

and real-time processing capabilities, MobileNetV4-Small is well-suited for practical deepfake detection applications, addressing both computational efficiency and detection challenges.

Table 1: Comparative Analysis of Lightweight Models for Deepfake Image Classification

Model	Dataset Used	Accuracy (%)
ShuffleNetV2 (Zhang et al., 2018)	DFDC	88.7
SqueezeNet (Bhuvaneswari and Enaganti, 2023)	Celeb-DF	87.5
MobileNetV4-Small	OpenForensics	**89.78**

5 CONCLUSION

The MobileNetV4-Small model proves to be a lightweight and efficient solution for deepfake detection, making it suitable for deployment on resource-constrained devices. Its architecture effectively captures subtle image artifacts, addressing the growing need for real-time detection systems in practical scenarios. Future work could focus on extending this framework to classify specific manipulation techniques, enhancing its robustness across diverse datasets and unseen deepfake types, and exploring integration into real-world applications. Additionally, considerations for dataset biases and challenges with generalization across diverse fake media should be addressed to ensure broader applicability and fairness in detection results.

REFERENCES

Ahmad, Z., Jaffri, Z. u. A., Chen, M., and Bao, S. (2024). Understanding gans: fundamentals, variants, training challenges, applications, and open problems. *Multimedia Tools and Applications*, pages 1–77.

Appel, M. and Prietzel, F. (2022). The detection of political deepfakes. *Journal of Computer-Mediated Communication*, 27(4):zmac008.

Bhuvaneswari, R. and Enaganti, K. K. (2023). Robust image forgery classification using squeezenet network. In *2023 First International Conference on Advances in Electrical, Electronics and Computational Intelligence (ICAEECI)*, pages 1–5. IEEE.

Broklyn, P., Egon, A., and Shad, R. (2024). Deepfakes and cybersecurity: Detection and mitigation authors. *Available at SSRN 4904874*.

Croitoru, F.-A., Hiji, A.-I., Hondru, V., Ristea, N. C., Irofti, P., Popescu, M., Rusu, C., Ionescu, R. T., Khan, F. S., and Shah, M. (2024). Deepfake media generation and detection in the generative ai era: A survey and outlook. *arXiv preprint arXiv:2411.19537*.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Heidari, A., Jafari Navimipour, N., Dag, H., and Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14(2):e1520.

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al. (2019). Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324.

Howard, A. G. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.

Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

Kaiser, L., Gomez, A. N., and Chollet, F. (2017). Depthwise separable convolutions for neural machine translation. *arXiv preprint arXiv:1706.03059*.

Karki, M. (2024). Deepfake and real images dataset. https://www.kaggle.com/datasets/manjilkarki/deepfake-and-real-images/data. Accessed: 2024-12-02.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. (2022). Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s):1–41.

Kingma, D. P., Welling, M., et al. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392.

Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216.

Mienye, I. D. and Swart, T. G. (2024). A comprehensive review of deep learning: Architectures, recent advances, and applications. *Information*, 15(12):755.

Nirkin, Y., Keller, Y., and Hassner, T. (2019). Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7184–7193.

Patel, Y., Tanwar, S., Bhattacharya, P., Gupta, R., Alsuwian, T., Davidson, I. E., and Mazibuko, T. F. (2023). An improved dense cnn architecture for deepfake image detection. *IEEE Access*, 11:22081–22095.

Qin, D., Leichner, C., Delakis, M., Fornoni, M., Luo, S., Yang, F., Wang, W., Banbury, C., Ye, C., Akin, B., et al. (2025). Mobilenetv4: Universal models for the mobile ecosystem. In *European Conference on Computer Vision*, pages 78–96. Springer.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520.

Tammina, S. (2019). Transfer learning using vgg-16 with deep convolutional neural network for classifying images. *International Journal of Scientific and Research Publications (IJSRP)*, 9(10):143–150.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Tan, M. and Le, Q. (2021). Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR.

Targ, S., Almeida, D., and Lyman, K. (2016). Resnet in resnet: Generalizing residual architectures. *arXiv preprint arXiv:1603.08029*.

Thies, J., Zollhöfer, M., and Nießner, M. (2019). Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12.

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395.

Zhang, X., Zhou, X., Lin, M., and Sun, J. (2018). Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6848–6856.

Zheng, L., Zhang, Y., and Thing, V. L. (2019). A survey on image tampering and its detection in real-world photos. *Journal of Visual Communication and Image Representation*, 58:380–399.