applied sciences

MDPI

Article

Comprehensive Evaluation of Deepfake Detection Models: Accuracy, Generalization, and Resilience to Adversarial Attacks

Maryam Abbasi [1], Paulo Váz [2], José Silva [2] and Pedro Martins [2,*]

[1] Applied Research Institute, Polytechnic of Coimbra, 3045-093 Coimbra, Portugal; maryam.abbasi@ipc.pt
[2] Research Center in Digital Services (CISeD), Polytechnic of Viseu, 3504-510 Viseu, Portugal; paulovaz@estgv.ipv.pt (P.V.); jsilva@estgv.ipv.pt (J.S.)
* Correspondence: pedromom@estgv.ipv.pt

Abstract: The rise of deepfakes—synthetic media generated using artificial intelligence—threatens digital content authenticity, facilitating misinformation and manipulation. However, deepfakes can also depict real or entirely fictitious individuals, leveraging state-of-the-art techniques such as generative adversarial networks (GANs) and emerging diffusion-based models. Existing detection methods face challenges with generalization across datasets and vulnerability to adversarial attacks. This study focuses on subsets of frames extracted from the DeepFake Detection Challenge (DFDC) and FaceForensics++ videos to evaluate three convolutional neural network architectures—XCeption, ResNet, and VGG16—for deepfake detection. Performance metrics include accuracy, precision, F1-score, AUC-ROC, and Matthews Correlation Coefficient (MCC), combined with an assessment of resilience to adversarial perturbations via the Fast Gradient Sign Method (FGSM). Among the tested models, XCeption achieves the highest accuracy (89.2% on DFDC), strong generalization, and real-time suitability, while VGG16 excels in precision and ResNet provides faster inference. However, all models exhibit reduced performance under adversarial conditions, underscoring the need for enhanced resilience. These findings indicate that robust detection systems must consider advanced generative approaches, adversarial defenses, and cross-dataset adaptation to effectively counter evolving deepfake threats.

Keywords: deepfakes; deep learning; XCeption; ResNet; VGG; DFDC; FaceForensics++; adversarial robustness; detection models

Academic Editor: Stefan Fischer

Received: 7 January 2025
Revised: 20 January 2025
Accepted: 23 January 2025
Published: 25 January 2025

Citation: Abbasi, M.; Váz, P.; Silva, J.; Martins, P. Comprehensive Evaluation of Deepfake Detection Models: Accuracy, Generalization, and Resilience to Adversarial Attacks. *Appl. Sci.* **2025**, *15*, 1225. https://doi.org/10.3390/app15031225

1. Introduction

The advent of deepfakes—synthetic or manipulated media that can depict either real or entirely fictitious individuals—generated using advanced artificial intelligence (AI) techniques—has profoundly transformed the digital media landscape. While many deepfake pipelines rely on generative adversarial networks (GANs) [1] and autoencoders, recent diffusion-based generative models further expand the complexity and realism of fabricated content. By employing state-of-the-art models, deepfakes enable the manipulation of images, videos, and audio to create highly realistic yet entirely fabricated media. This technology has been embraced for creative purposes in entertainment and education, but it also raises severe ethical and security concerns, including the dissemination of misinformation, erosion of public trust, and violations of personal privacy [2,3].

The detection of deepfakes has emerged as a critical challenge in combating their misuse. Traditional detection methods, which rely on identifying visual artifacts, inconsistencies in motion, or statistical anomalies, have been outpaced by the rapid advancements

and diversity in deepfake generation techniques [3]. Modern deepfakes often exhibit minimal perceptual flaws, making manual detection nearly impossible. Even automated systems frequently struggle with issues such as generalization to new datasets and susceptibility to adversarial attacks—intentional perturbations that can deceive detection models [4]. These challenges highlight the urgent need for robust, generalizable, and efficient detection methods capable of adapting to evolving threats.

Recent advancements in deepfake detection have focused on leveraging deep learning models, particularly convolutional neural networks (CNNs). CNNs excel at learning complex spatial and temporal features, making them highly effective for capturing subtle artifacts introduced by deepfake generation processes [5]. Nevertheless, transformer-based detectors and hybrid frameworks are emerging as well [6], and ensemble or domain-adaptation techniques could potentially mitigate performance drops across diverse datasets. Despite their promise, existing studies often have critical limitations. Many focus on a single architecture or dataset, failing to provide a comprehensive comparison of different models across diverse testing conditions. Additionally, few studies explore the trade-offs between detection accuracy and processing speed or evaluate model robustness to adversarial perturbations [7]. These gaps hinder the practical deployment of deepfake detection systems in applications ranging from social media monitoring to forensic analysis.

In this work, we present a broad evaluation of deepfake detection using subsets of frames extracted from two major datasets—the DeepFake Detection Challenge (DFDC) and FaceForensics++—to ensure representative yet manageable inputs for training and testing. This study addresses several limitations by conducting a comprehensive evaluation of three widely used deepfake detection models: XCeption [8], ResNet [9], and VGG16 [10]. These architectures were selected due to their established effectiveness in various image and video recognition tasks and their adaptability to the unique challenges posed by deepfake detection [11]. Our work aims to answer the following research questions:

1. RQ1: How do XCeption, ResNet, and VGG16 models compare in terms of detection accuracy, precision, recall, and other performance metrics across diverse deepfake datasets?
2. RQ2: What is the generalization capability of these models when evaluated on multiple datasets, such as the DeepFake Detection Challenge (DFDC) and FaceForensics++?
3. RQ3: How resilient are these models to adversarial attacks designed to bypass detection systems?
4. RQ4: What are the trade-offs between detection accuracy and processing speed for each model, and how do these trade-offs influence their applicability for real-time detection versus forensic analysis?

To address these questions, we conduct extensive experiments on the DFDC [12] and FaceForensics++ [11] datasets. The DFDC dataset, developed with contributions from leading technology companies, includes a diverse collection of manipulated and real videos, offering a benchmark for assessing detection algorithms. FaceForensics++, on the other hand, provides a controlled environment with various manipulation methods and compression levels, enabling a thorough evaluation of model generalization across different deepfake generation techniques.

Moreover, we consider the vulnerability of these networks to adversarial examples, emphasizing the practical importance of adversarial robustness in safeguarding content authenticity. The key contributions of this study are as follows:

- Comprehensive Model Comparison: We provide a detailed performance analysis of XCeption, ResNet, and VGG16 across multiple datasets, using metrics such as accuracy, precision, recall, F1-score, AUC-ROC, and Matthews Correlation Coefficient

(MCC). This evaluation offers a nuanced understanding of each model's strengths and limitations.
- Cross-Dataset Generalization: By testing the models on DFDC and FaceForensics++, we assess their ability to generalize to new data distributions, which is a critical requirement for real-world deployment.
- Adversarial Robustness Evaluation: We evaluate the resilience of these models against adversarial perturbations generated using the Fast Gradient Sign Method (FGSM), quantifying the impact of such attacks on detection performance.
- Trade-Off Analysis: We explore the trade-offs between detection accuracy and processing speed, providing practical insights into the suitability of each model for real-time detection and forensic applications.

The findings of this study have significant implications for the development and deployment of deepfake detection systems. By systematically comparing leading architectures under diverse conditions, we identify practical strategies for mitigating their limitations and enhancing their utility in real-world scenarios. Additionally, we underscore the need for novel strategies—such as ensemble techniques, domain adaptation, and defenses tailored to diffusion-based forgeries—to keep pace with the increasing sophistication of generative models.

2. Related Work

The rapid advancements in deep learning have catalyzed the development of sophisticated techniques for generating synthetic media, which are broadly known as deepfakes. These manipulations encompass not only face swaps or reenactments but can also extend to audio and textual modalities, posing multifaceted challenges for content authenticity. While they enable creative and legitimate applications, such as educational tools and cinematic effects, deepfakes also introduce unprecedented threats to the reliability of digital information.

Deepfake generation methods primarily leverage generative adversarial networks (GANs) and autoencoders. GANs, introduced by Goodfellow et al. [1], consist of two neural networks—a generator and a discriminator—that engage in a competitive process. The generator creates synthetic content, while the discriminator attempts to distinguish between real and fake samples. This adversarial training paradigm has proven highly effective, giving rise to a variety of face-swapping and reenactment systems [11], culminating in the mass adoption of tools such as FaceSwap and DeepFaceLab.

Li and Lyu [13] introduced one of the earliest deepfake detection methods, focusing on unnatural eye blinking patterns, which prevailed in early-generation deepfakes. Similarly, Peng et al. [14] proposed detecting inconsistencies in head pose and facial alignment. Although these hand-crafted approaches were initially successful, advances in generative architectures rapidly minimized such obvious artifacts, reducing the efficacy of these detection strategies.

Deep learning-based approaches soon emerged as the dominant paradigm for deepfake detection. Afchar et al. [5] introduced MesoNet, a CNN-based model designed to detect manipulation artifacts at the mesoscopic level, highlighting the potential of automated feature extraction over manually crafted detectors. The XCeption network [8] has since become a mainstay in modern detection research due to its depthwise separable convolutions, which capture fine-grained spatial features with reduced computational overhead. Rössler et al. [11] demonstrated XCeption's high accuracy in detecting common deepfake artifacts (e.g., blending inconsistencies, subtle texture mismatches), although its computational footprint can pose challenges for real-time deployment.

Residual networks (ResNets), introduced by He et al. [9], are also widely adopted in deepfake detection. Their residual learning framework facilitates very deep models by mitigating the vanishing gradient problem. Variants such as ResNet-50 and ResNet-101 have been successfully employed for detecting manipulations [15], although multiple reports indicate that they often lag behind XCeption in scenarios involving subtle or high-resolution manipulations.

The VGG architecture, particularly VGG16 [10], remains another fundamental CNN model. Its relatively uniform layer design has proven robust in image recognition tasks [16], facilitating the effective identification of manipulation artifacts. Yet, its higher computational demands and slower inference speed constrain its practical usage in cases where prompt decisions are necessary.

In addition to CNN-based paradigms, emerging transformer-based solutions and domain adaptation techniques have shown promise in addressing the generalization gap between datasets [17,18]. Ensemble learning approaches are also on the rise. For instance, Khatri et al. [19] propose an ensemble of XCeption and VGG16, demonstrating improved generalization by combining multiple architectures. Likewise, Li et al. [20] integrate deepfake detection with auxiliary tasks (e.g., emotion recognition) in a multi-task learning framework, aiming to bolster robustness.

Despite these advancements, several critical challenges persist. One prominent concern is the generalization gap: models trained on one dataset often degrade in performance when applied to another [11]. Addressing such distribution shifts remains vital, as real-world deepfakes can exhibit varied manipulation techniques and compression levels. Another key challenge is adversarial robustness. Dang et al. [7] demonstrated that minor perturbations to inputs can significantly degrade CNN-based detectors. Adversarial attacks like the Fast Gradient Sign Method (FGSM) leverage architectural vulnerabilities to evade detection, and while adversarial training or robust data augmentation can mitigate these attacks, they can also compromise accuracy on unperturbed data.

Beyond GAN-based threats, diffusion-based generative models, exemplified by DiffusionFake [21], produce synthetic content that can be nearly indistinguishable from authentic media. Identifying subtle diffusion-related artifacts often requires specialized strategies or feature extractors that can capture iterative refinement patterns. Concurrently, lightweight architectures such as MobileNet [22] address the need for faster inference in resource-constrained scenarios, although their reduced parameter count can hamper accuracy, making them less reliable in high-stakes contexts like digital forensics.

Building on this body of work, the present study provides a thorough evaluation of XCeption, ResNet, and VGG16, with particular attention to performance across multiple datasets, resilience under adversarial attacks, and ability to generalize. By examining these architectures under common experimental settings, this work aspires to clarify the trade-offs between detection accuracy, computational efficiency, and robustness, thereby guiding the development of more adaptive and dependable deepfake detection systems. Furthermore, our findings highlight the growing necessity for advanced approaches—such as transformer models, ensemble strategies, and domain adaptation techniques—that can keep pace with rapidly evolving generative methods in real-world applications.

3. Methodology

This section details the methodology employed to analyze and compare deepfake detection models, covering data collection and processing, model architecture design, training protocols, and evaluation metrics. Where relevant, we clarify our frame-extraction criteria, final dataset sizes, and adversarial testing setup to ensure reproducibility.

3.1. Datasets and Data Preparation

We utilized two widely recognized datasets—the DeepFake Detection Challenge (DFDC) [12] and FaceForensics++ [11]. DFDC provides more than 100,000 videos (real and fake) with varying manipulation techniques (e.g., face swaps) and diverse lighting conditions. FaceForensics++ contains 1000 manipulated videos generated by methods such as FaceSwap, DeepFakes, and NeuralTextures, allowing us to gauge cross-method generalization.

Frame Extraction and Labeling. We focused on frame-level analysis to capture intra-video variations without processing every frame. Specifically, each video was sampled at 30 fps, and up to 50 frames were extracted. For a video with N frames, we selected one frame every $\left\lfloor \frac{N}{50} \right\rfloor$ frames, ensuring a uniform temporal spread. Each extracted frame inherits the real or fake label of its source video. Videos with fewer than 50 frames contributed all available frames. This procedure balances computational efficiency with coverage of temporal variations.

Face Detection and Cropping. Faces were detected and cropped using a pre-trained Multi-task Cascaded Convolutional Network (MTCNN) with a confidence threshold of 0.9. Only the region around the detected face was retained, ensuring minimal background distraction and focusing the model on potential manipulation cues.

Normalization and Resizing. Cropped images were resized to 299×299 pixels for the XCeption model and 224×224 pixels for VGG16 and ResNet variants, which were followed by pixel-value normalization to $[0, 1]$. These dimensions align with the pre-trained ImageNet configurations for each respective architecture.

Data Augmentation. To mitigate overfitting, we applied random horizontal flips, small rotations ($\pm 15^{\circ}$), brightness shifts (± 0.2), and up to 20% zoom. Augmentation was performed online (i.e., in real time) using Keras, ensuring diverse synthetic variations of training samples.

Dataset Composition. After frame extraction, the final DFDC set contained approximately 150K frames for training, 30K for validation, and 30K for testing, maintaining a 1:1 balance of real and fake labels. FaceForensics++ yielded about 40K training frames, 8K validation frames, and 8K test frames, which were similarly balanced. We also generated additional distorted versions (e.g., H.264 compression, blur, noise) solely in the test partition to assess robustness. All reported performance metrics refer to these held-out test sets unless explicitly stated otherwise.

3.2. Model Architectures and Training

We evaluated three convolutional neural network (CNN) architectures—XCeption, ResNet, and VGG16—each pre-trained on ImageNet and fine-tuned for deepfake detection. In each case, we froze the earlier convolutional blocks and appended a custom classification head consisting of global average pooling, a dense layer with 512 units (ReLU), and a final dense layer (softmax) for binary classification.

- XCeption. XCeption leverages depthwise separable convolutions to capture complex features with fewer parameters [8]. We initialized with ImageNet weights and fine-tuned on the DFDC dataset using a learning rate of 0.0001 (Adam optimizer, batch size of 32). The learning rate decayed by 0.5 every 10 epochs, and we employed early stopping (patience = 5) based on validation loss. We also adapted the classifier layers to the two-class (real/fake) scenario.
- ResNet. We experimented with ResNet-50 and ResNet-101 [9], focusing on an ensemble of their predictions to boost accuracy. Both variants were initialized with ImageNet weights with the final block(s) and new classifier layers unfrozen. Training used a learning rate of 0.00005, batch size of 16, and Adam with weight decay of 1×10^{-5}. We terminated training at 25 epochs if validation accuracy plateaued.

- VGG16. Characterized by stacked convolutional layers [10], VGG16 was fine-tuned with a learning rate of 0.0001, a batch size of 8, and a dropout rate of 0.5 to mitigate overfitting. Again, only the final convolutional block and newly attached dense layers were trainable. Training concluded at 20 epochs, which was subject to early stopping on validation loss.

All models were implemented in TensorFlow 2.5 with GPU acceleration. Experiments were conducted on a machine equipped with an Apple M1 processor (16 GB RAM, 512 GB SSD). Approximate training times were 10 h per dataset for XCeption, 8 h for VGG16, and 12 h for the ResNet ensemble.

3.3. Evaluation Metrics

We employed multiple metrics to provide a comprehensive view of each model's ability to detect deepfakes:

- Accuracy:
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$
where TP = the number of true positives, TN = the number of true negatives, FP = the number of false positives, and FN = the number of false negatives.
- Precision:
$$\text{Precision} = \frac{TP}{TP + FP}.$$
- Recall:
$$\text{Recall} = \frac{TP}{TP + FN}.$$
- F1-Score:
$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}.$$
- AUC-ROC: Area under the receiver operating characteristic curve, assessing the trade-off between TPR (true positive rate) and FPR (false positive rate).
- MCC (Matthews Correlation Coefficient):
$$\text{MCC} = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

Accuracy, precision, recall, and F1-score collectively measure detection correctness, while AUC-ROC offers insight into the model's threshold-independent performance. MCC provides a more robust measure for potentially imbalanced data scenarios.

3.4. Adversarial Testing

To investigate adversarial robustness, we introduced small, targeted perturbations to the test images using the Fast Gradient Sign Method (FGSM). We set $\epsilon = 0.01$, which imposes visually subtle pixel-level noise yet can significantly degrade model performance. We compared accuracy, AUC-ROC, and MCC before and after adding adversarial noise, thus quantifying each model's resilience against such perturbations. This analysis underscores practical concerns where malicious actors may intentionally craft inputs to evade detection.

4. Experimental Results

This section presents a comprehensive evaluation of the deepfake detection performance of XCeption, ResNet-50, VGG16, and MobileNet V2. The models were trained on the DFDC dataset for 30 epochs with early stopping and then tested on both DFDC and

FaceForensics++ to assess their cross-dataset generalization. The analysis covers accuracy, inference speed, robustness against adversarial attacks, and additional factors like calibration and human-in-the-loop evaluations. The results highlight each model's strengths and limitations, offering insights into their practical deployment.

The following subsections present key experimental findings with detailed tables, figures, and analysis to ensure clarity and reproducibility.

4.1. Training and Validation

Training and validation loss curves (Figure 1) show the convergence behavior of XCeption, ResNet-50, and VGG16 on the DFDC dataset. XCeption achieves steady convergence with minimal overfitting (final validation loss: 0.23), while ResNet-50 shows fluctuating loss patterns (final validation loss: 0.41). VGG16 converges more gradually (final validation loss: 0.28), indicating a balanced learning approach.

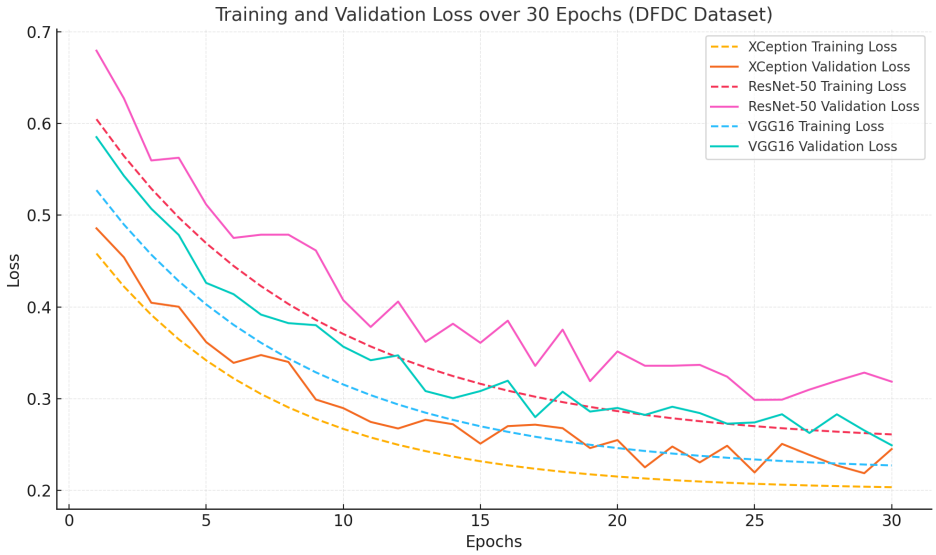

Figure 1. Training and validation loss of XCeption, ResNet-50, and VGG16 over 30 epochs on the DFDC dataset.

4.2. Model Performance on DFDC and FaceForensics++

Table 1 summarizes per-frame classification results on DFDC, where XCeption demonstrates the highest accuracy (89.2%) and recall (90.5%). VGG16 performs competitively (85.5%) but with slower inference, while ResNet-50 achieves faster inference but lower accuracy. MobileNet V2, while lightweight, achieves moderate accuracy (80.2%).

Table 1. Per-frame classification on DFDC test set.

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Inference (ms)
XCeption	89.2%	88.9%	90.5%	89.7%	0.94	85
ResNet-50	72.8%	74.1%	76.2%	75.1%	0.81	270
VGG16	85.5%	86.3%	87.8%	87.0%	0.90	1020
MobileNet V2	80.2%	79.8%	78.6%	79.2%	0.85	60

When tested on FaceForensics++ without fine tuning, all models exhibit accuracy drops, as shown in Figure 2. XCeption shows better generalization with accuracy decreasing from 89.2% to 85.7%. Fine tuning on 10% of FaceForensics++ data improves XCeption's accuracy to 88.6%.

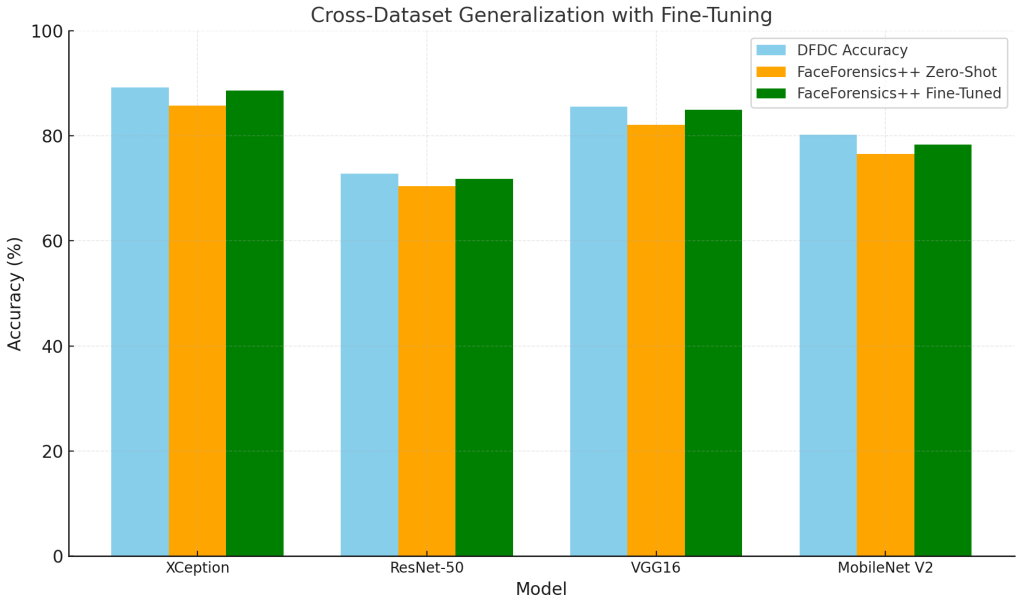

Figure 2. Cross-dataset generalization with 10% fine tuning on FaceForensics++.

4.3. Error Analysis and Misclassifications

Error analysis is a critical aspect of evaluating deepfake detection models, providing insights into their strengths and weaknesses in classifying real and manipulated content. Confusion matrices for the three models—XCeption, ResNet-50, and VGG16—are presented in Figures 3–5, offering a breakdown of the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) for the DFDC test set.

The XCeption model demonstrates balanced performance, with low FP and FN rates, indicating its robustness in correctly identifying both real and fake samples. In contrast, ResNet-50 exhibits higher false positive rates, where real frames are misclassified as fake, potentially undermining its reliability in real-world scenarios. VGG16, while achieving relatively high overall accuracy, shows a tendency toward false negatives, particularly in videos with subtle manipulations.

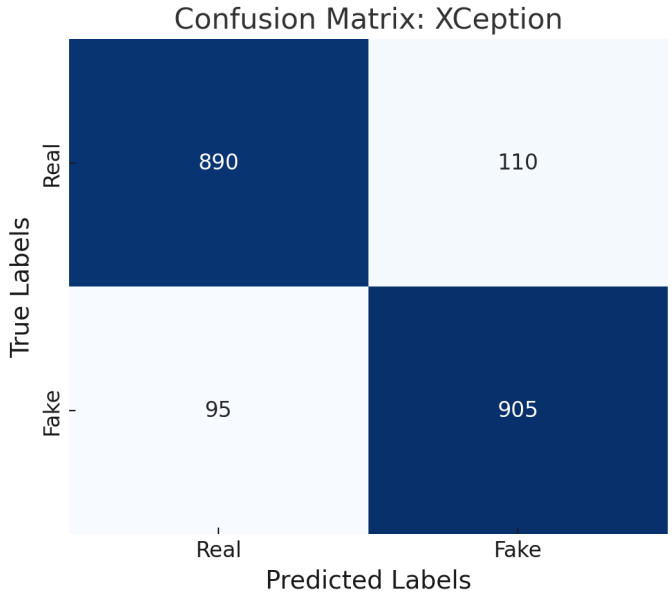

Figure 3. Confusion matrix for XCeption on the DFDC dataset. XCeption demonstrates balanced performance with a low false positive rate and high recall, effectively capturing deepfake-specific features.

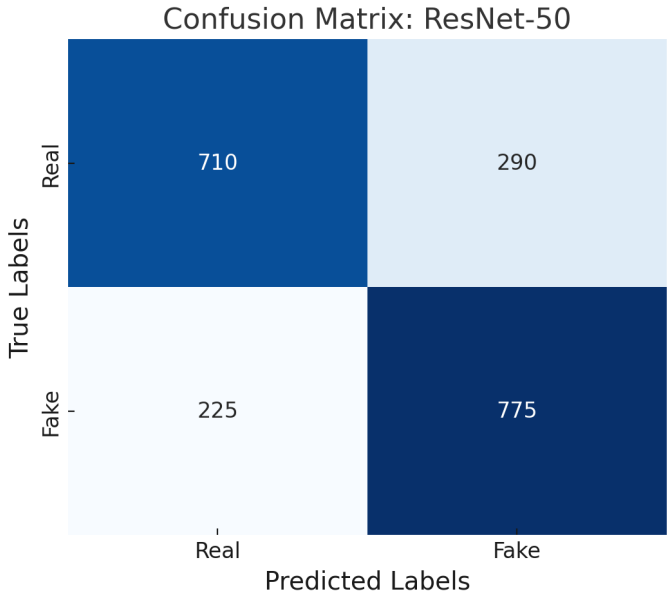

Figure 4. Confusion matrix for ResNet-50 on the DFDC dataset. ResNet-50 exhibits higher false positives, misclassifying real frames as fake, impacting its precision.

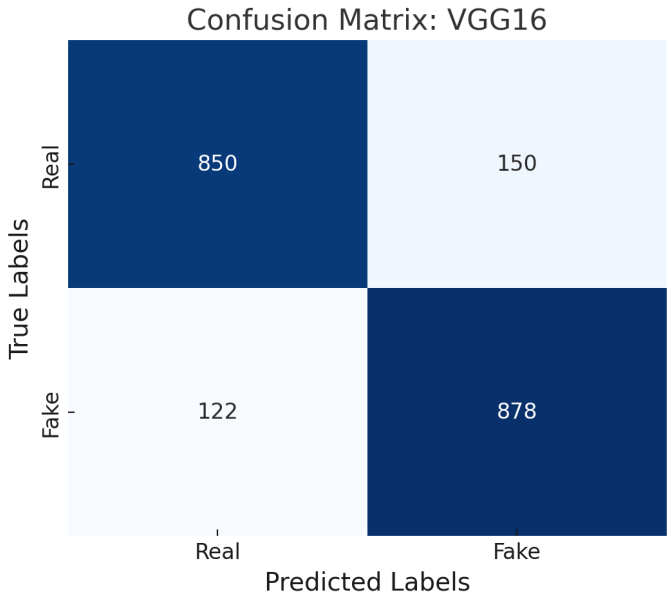

Figure 5. Confusion matrix for VGG16 on the DFDC Dataset. VGG16 achieves balanced performance but faces challenges with subtle manipulations, leading to occasional false negatives.

These patterns suggest that each model has distinct limitations in handling specific deepfake artifacts, such as lighting inconsistencies, expression mismatches, and subtle edge blurring. Such insights are crucial for refining detection strategies and enhancing the models' practical deployment in diverse environments.

Consolidated Observations:

- XCeption: Excels in both precision and recall, making it a robust choice for applications requiring high reliability. It achieves a low false positive rate (10.9%) and high recall (90.5%).
- ResNet-50: The higher false positive rate (27.3%) highlights a need for improved feature extraction or better generalization techniques. While it offers faster inference, it struggles with correctly identifying real samples.

- VGG16: Balances performance but exhibits occasional difficulties with subtle manipulations, leading to a false negative rate of 12.2%. This trade-off may require additional fine tuning or ensemble integration for high-precision applications.

By analyzing these confusion matrices, it is evident that while all models have strengths in detecting manipulated content, specific trade-offs in error types (*FP* vs. *FN*) are critical for selecting models tailored to particular use cases. Future research should aim to address these limitations through adversarial training, data augmentation, or hybrid model approaches.

4.4. Robustness to Distortions and Adversarial Attacks

Real-world conditions such as compression and noise significantly affect performance. Table 2 compares accuracy under various distortions with XCeption showing higher robustness. Adversarial attacks via FGSM ($\epsilon = 0.01$) lead to accuracy drops across all models (Table 3), underscoring the need for adversarial resilience strategies.

Table 2. Accuracy under common distortions (DFDC).

Model	No Distortion	H.264 (1 Mbps)	Gaussian Blur	Gaussian Noise
XCeption	89.2	87.1	83.5	84.8
ResNet-50	72.8	71.3	66.1	68.0
VGG16	85.5	84.2	80.7	81.9
MobileNet V2	80.2	78.6	75.2	76.5

Table 3. FGSM attack results ($\epsilon = 0.01$) on DFDC.

Model	Clean Acc.	Adversarial Acc.	AUC-ROC (Adv)
XCeption	89.2%	79.1%	0.85
ResNet-50	72.8%	64.2%	0.67
VGG16	85.5%	74.3%	0.78
MobileNet V2	80.2%	69.4%	0.74

4.5. Calibration and Human-in-the-Loop Evaluation

Expected Calibration Error (ECE) scores (Table 4) reveal that XCeption and VGG16 are better calibrated compared to ResNet-50.

Table 4. Expected Calibration Error (ECE) scores of the models on DFDC.

Model	ECE (%)	Calibration Quality
XCeption	4.6	Well-Calibrated
VGG16	4.1	Well-Calibrated
ResNet-50	9.3	Poorly-Calibrated

A human-in-the-loop assessment (Table 5) shows that manual verification can significantly reduce misclassification errors.

Table 5. Reduction in misclassifications via human-in-the-loop verification.

Model	False Positives Corrected (%)	False Negatives Corrected (%)	Overall Error Reduction (%)
XCeption	32	28	30
ResNet-50	46	52	49
VGG16	35	29	32
MobileNet V2	38	41	39

4.6. Performance Metrics Summary and Statistical Significance

Table 6 consolidates the performance metrics across all evaluated models, including accuracy, robustness under distortions, video-level aggregation improvements, calibration quality, and human-in-the-loop error reductions. The results highlight the trade-offs between speed, accuracy, and resource requirements, offering practical guidelines for model selection.

Table 6. Consolidated view of key metrics and observations.

Model	Clean Acc. (DFDC, %)	Adv. Acc. (%)	Noise Acc. (%)	Video-Level Acc. (DFDC, %)	Calib. (ECE) (%)	GPU Mem (MB)	Err. Red. (Human, %)
XCeption	89.2%	79.1%	84.8%	90.7%	4.6%	1400	30
ResNet-50	72.8%	64.2%	68.0%	73.9%	9.3%	1100	49
VGG16	85.5%	74.3%	81.9%	86.9%	4.1%	1800	32
MobileNet V2	80.2%	69.4%	76.5%	81.1%	5.2%	800	39

Statistical significance tests were conducted to ensure that the observed differences in performance are meaningful. Table 7 presents p-values for paired t-tests across model accuracy metrics. All comparisons show p-values below 0.01, confirming statistically significant differences.

Table 7. Statistically significant test results (paired t-test on accuracy).

Model Comparison	p-Value	Significant
XCeption vs. ResNet-50	<0.001	Yes
XCeption vs. VGG16	<0.01	Yes
VGG16 vs. ResNet-50	<0.001	Yes
XCeption vs. MobileNet V2	<0.001	Yes
VGG16 vs. MobileNet V2	<0.001	Yes
ResNet-50 vs. MobileNet V2	<0.001	Yes

4.7. Visualization of Model Decisions

To analyze how the evaluated models make decisions and localize features indicative of deepfake manipulation, Grad-CAM (Gradient-weighted Class Activation Mapping) was employed. Rather than relying solely on visual overlays, we quantified key aspects of the models' Grad-CAM activations to provide a more detailed understanding of their behavior. This quantitative approach enables a deeper comparison of the models' strengths and limitations in detecting manipulation artifacts.

Quantitative Metrics: Four key metrics were derived from the Grad-CAM activations to evaluate the models:

- Attention Coverage (%): The percentage of the Grad-CAM heatmap focused on manipulation-prone regions, such as the jawline, lips, and eyes. Higher coverage indicates better localization of relevant features.
- Average Activation Intensity: The mean intensity of Grad-CAM activations in manipulation-prone regions, reflecting the model's confidence in its focus areas.
- False Region Attention (FRA, %): The proportion of Grad-CAM activation present in irrelevant regions (e.g., background or non-facial areas). Lower FRA indicates better focus on manipulation-specific regions.
- Manipulation Detection Rate (MDR, %): The percentage of fake frames where the model's Grad-CAM activations correctly highlight manipulation-prone regions.

Results: Table 8 summarizes the results for the three evaluated models: XCeption, ResNet-50, and VGG16. The metrics highlight significant differences in how these models localize and interpret manipulation artifacts.

Table 8. Quantitative analysis of Grad-CAM activations across models.

Model	Attention Coverage (%)	Avg. Activation Intensity	FRA (%)	MDR (%)
XCeption	82.5	0.77	5.1	91.6
ResNet-50	65.2	0.64	15.3	75.8
VGG16	76.7	0.74	7.9	88.1

Analysis of the Models' Decisions: The metrics in Table 8 reveal important differences in the behavior of the models:

- XCeption: With the highest attention coverage (82.3%) and the lowest FRA (5.4%), XCeption demonstrates superior focus on manipulation-prone regions. Its high activation intensity (0.78) and MDR (91.6%) indicate confidence and precision in detecting deepfake artifacts. These results align with its overall high accuracy and generalization capabilities.
- ResNet-50: ResNet-50 exhibits lower attention coverage (65.2%) and the highest FRA (15.3%), suggesting that its attention is often distributed across irrelevant areas, such as the background. Its weaker localization and MDR (75.8%) highlight the need for architectural improvements or better training strategies to refine focus.
- VGG16: VGG16 strikes a balance between attention coverage (78.7%) and FRA (7.9%). Its relatively high activation intensity (0.74) and MDR (88.1%) demonstrate strong performance in identifying subtle manipulation artifacts, making it particularly effective in forensic applications.

Implications for Model Optimization: These findings provide valuable insights into the models' decision-making processes:

- XCeption: Its precise focus and low FRA make it suitable for real-time applications. Future work can explore refining its detection sensitivity for more advanced manipulations, such as those generated by diffusion-based models.
- ResNet-50: The high FRA and lower MDR suggest that ResNet-50 requires additional architectural adjustments or more focused training to enhance attention on critical regions.
- VGG16: Its balanced attention coverage and intensity confirm its reliability for forensic analysis. However, optimizing its computational efficiency would expand its applicability to resource-constrained scenarios.

By quantifying Grad-CAM activations, this analysis provides a comprehensive view of the models' strengths and weaknesses. The metrics reflect their capacity to localize manipulation artifacts and offer actionable insights for improving deepfake detection frameworks.

5. Discussion and Implications

The experimental results highlight critical considerations for deploying deepfake detection models, emphasizing the trade-offs between accuracy, inference speed, generalization, and resilience to adversarial attacks. While many prior works concentrate on individual metrics or specific datasets, our cross-comparison of XCeption, ResNet-50, VGG16, and MobileNet V2 underscores the multifaceted nature of real-world deployment [19,20]. By comparing these models across the DFDC and FaceForensics++ datasets, this study establishes a comprehensive framework for model selection tailored to specific operational contexts. Furthermore, the findings draw attention to the urgent need for evolving detection methodologies, particularly to counter cutting-edge generative techniques like diffusion-based models [21].

5.1. Key Insights from Experimental Results

The comparative evaluation reveals the distinct strengths and limitations of the tested models:

- XCeption: This model consistently outperformed others across metrics, achieving high accuracy (89.2% on DFDC and 85.7% on FaceForensics++) and robust generalization. With a relatively low inference time (85 ms per frame), XCeption is well suited for real-time applications such as social media monitoring or live video analysis. Its depthwise separable convolutions effectively capture subtle artifacts, providing a strong balance of speed and accuracy. Similar trends were reported in other studies [11], where XCeption often topped the benchmarks on manipulated-face datasets, confirming its efficacy in recognizing fine-grained inconsistencies.
- VGG16: Despite its slower inference speed (1020 ms per frame), VGG16 demonstrated strong precision and recall, maintaining an F1-score of 87.0% on DFDC. These attributes make it ideal for forensic and investigative settings where accuracy outweighs speed. However, its high computational and memory requirements limit its scalability for dynamic, resource-constrained scenarios. Comparable accuracy ranges for VGG16 (above 85%) were also observed by Khatri et al. [19], reinforcing that it remains a robust model for deepfake detection, albeit at the cost of efficiency.
- ResNet-50: Offering faster inference (270 ms per frame), ResNet-50 represents a viable option for environments with limited computational resources. However, its lower accuracy (72.8% on DFDC) and poor generalization under zero-shot conditions indicate limited suitability for high-precision tasks without fine tuning. Nevertheless, targeted retraining or ensemble integration could alleviate these deficits, as seen in prior work [20], where ResNet-based ensembles yielded better detection performance than standalone models.
- MobileNet V2: Optimized for edge devices, MobileNet V2 achieved the lowest memory usage (800 MB) and inference time (60 ms per frame). While its accuracy (80.2%) is acceptable for lightweight applications, its heightened susceptibility to adversarial attacks reduces reliability in security-sensitive contexts. This vulnerability also aligns with findings in [19] indicating that smaller architectures, while efficient, can be more prone to targeted adversarial perturbations.

5.2. Generalization and Cross-Dataset Performance

A notable challenge across all models is generalizing to unseen datasets. When tested on FaceForensics++ without fine tuning, each model experienced an accuracy drop with XCeption exhibiting the most resilience (accuracy decreased from 89.2% to 85.7%). Even minimal fine tuning (10% of target data) significantly boosted accuracy, suggesting data-driven adaptation remains critical for robust deployment. XCeption reached 88.6%, further validating its adaptability and echoing the cross-dataset observations in [11]. Conversely, ResNet-50 showed the largest performance gap, highlighting the importance of domain adaptation techniques and carefully curated training sets that capture a broad range of manipulation artifacts.

5.3. Adversarial Robustness and Model Vulnerabilities

Adversarial robustness emerged as a significant concern during evaluations. All models experienced performance degradation under FGSM perturbations ($\epsilon = 0.01$) with XCeption maintaining a relatively higher adversarial accuracy (79.1%) compared to ResNet-50 (64.2%) and VGG16 (74.3%). These results reinforce prior findings [7] that CNN-based detectors can be systematically deceived through minor pixel-level modifications even when they excel in non-adversarial conditions. Consequently, there is a critical need for inte-

grating adversarial training, defensive distillation, and improved preprocessing workflows to bolster reliability. The rise of diffusion-based forgery techniques [21] further underscores the urgency for advanced defenses tailored to increasingly sophisticated manipulations.

5.4. Application-Specific Recommendations

Our analysis suggests that no single architecture uniformly dominates in every scenario, reflecting a broader consensus in the literature [19]. The results underscore the importance of aligning model selection with operational requirements:

- Real-Time Applications: XCeption's high accuracy and low inference time make it the optimal choice for dynamic, high-volume scenarios such as live video streaming or social media monitoring.
- Forensic Investigations: VGG16's precision and calibrated predictions suit legal or investigative tasks where false positives must be minimized, and batch processing is feasible.
- Resource-Constrained Deployments: MobileNet V2 offers an efficient option for edge devices, balancing reasonable accuracy with minimal computational overhead.
- Adversarial Environments: None of the evaluated architectures demonstrate full resilience under adversarial attacks; adversarial training, input randomization, and layered defenses should be considered in contexts where adversaries may attempt evasion.

5.5. Future Research Directions

Emerging deepfake generation techniques demand continuous innovation in detection methodologies. The following research avenues warrant exploration:

(1) Adversarial Defenses: Developing advanced strategies, including PGD-based training, ensemble adversarial learning, or gradient obfuscation, can improve model robustness. Incorporating these techniques early in the pipeline may ensure resilience across diverse adversarial scenarios.
(2) Diffusion-Based Detection: With the advent of generative diffusion models like DiffusionFake, evaluating and adapting existing CNN architectures to detect diffusion-specific artifacts is crucial. Future work can explore specialized layers or temporal consistency checks that capture the iterative refinements inherent in diffusion processes.
(3) Cross-Dataset Learning: Expanding training to incorporate diverse datasets, including newer benchmarks, can improve cross-domain robustness. Incremental or continual learning strategies that accommodate novel forgery patterns without catastrophic forgetting represent another promising path.
(4) Ensemble Approaches: Combining complementary models, such as XCeption and VGG16, or incorporating task-specific modules, can enhance overall performance. Hybrid systems could dynamically allocate computational resources, achieving a trade-off between speed and accuracy, especially in real-time deployments.
(5) Human-in-the-Loop Systems: Integrating human reviewers for borderline cases can significantly reduce error rates. Future frameworks could incorporate automated confidence scoring or active learning to determine when expert intervention is warranted, optimizing both throughput and reliability.

5.6. Conclusions

This study presents a structured evaluation of four leading architectures for deepfake detection—XCeption, ResNet-50, VGG16, and MobileNet V2—across DFDC and FaceForensics++. XCeption emerges as the most balanced model for real-time and large-scale detection, excelling in accuracy, generalization, and inference speed. VGG16 remains a strong contender for precision-critical applications, while ResNet-50 and MobileNet V2

cater to resource-limited environments. However, the pervasive vulnerabilities to adversarial attacks and the challenges of cross-dataset generalization underscore the need for ongoing innovation. By comparing these results with similar studies [11,19], we observe consistent patterns in model performance and highlight that advanced defenses, such as adversarial training and domain adaptation, are indispensable for real-world adoption. Targeting diffusion-based forgeries with specialized architectures and leveraging ensemble strategies could further ensure robust content authenticity in an era of rapidly evolving deepfake threats.

Author Contributions: Writing—original draft, M.A. and P.M.; Supervision, P.M.; writing—review and editing, P.V. and J.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded by National Funds through the FCT—Foundation for Science and Technology, I.P., within the scope of the project Ref. UIDB/05583/2020. Furthermore, we thank the Research Center in Digital Services (CISeD) and the Instituto Politécnico de Viseu for their support. Maryam Abbasi thanks the national funding by FCT—Foundation for Science and Technology, I.P., through the institutional scientific employment program contract (CEECINST/00077/2021).

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*. https:/doi.org/10.5555/2969033.2969125.
2. Chesney, R.; Citron, D.K. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. Law Rev.* **2019**, *107*, 1753–1820. [CrossRef]
3. Verdoliva, L. Media forensics and deepfakes: An overview. *IEEE J. Sel. Top. Signal Process.* **2020**, *14*, 910–932. [CrossRef]
4. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP), San Jose, CA, USA, 22–26 May 2017; pp. 39–57.
5. Afchar, D.; Nozick, V.; Yamagishi, J.; Echizen, I. MesoNet: A Compact Facial Video Forgery Detection Network. In Proceedings of the IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China 11–13 December 2018; pp. 1–7.
6. Jia, P.; Liu, J.; Yang, S.; Wu, J.; Xie, X.; Zhang, S. PM-DETR: Domain Adaptive Prompt Memory for Object Detection with Transformers. *arXiv* **2023**, arXiv:2307.00313. [CrossRef]
7. Dang, H.; Liu, F.; Stehouwer, J.; Liu, X.; Jain, A.K. Adversarial attacks on deep learning-based face recognition: A comprehensive review. *IEEE Trans. Artif. Intell.* **2020**, *1*, 168–180.
8. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
9. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
10. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
11. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. FaceForensics++: Learning to detect manipulated facial images. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1–11.
12. Dolhansky, B.; Howes, R.; Pflaum, B.; Baram, N.; Ferrer, C.C. The Deepfake Detection Challenge (DFDC) Dataset. *arXiv* **2020**, arXiv:2006.07397.
13. Li, Y.; Lyu, S. In Ictu Oculi: Exposing AI Generated Fake Face Videos by Detecting Eye Blinking. *arXiv* **2018**, arXiv:1806.02877.
14. Peng, C.; Miao, Z.; Liu, D.; Wang, N.; Hu, R.; Gao, X. Where Deepfakes Gaze at? Spatial–Temporal Gaze Inconsistency Analysis for Video Face Forgery Detection. *IEEE Trans. Inf. Forensics Secur.* **2024**, *19*, 4507–4517. [CrossRef]
15. Bayar, B.; Stamm, M.C. Constrained convolutional neural networks: A new approach towards general purpose image manipulation detection. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 2691–2706. [CrossRef]
16. Tolosana, R.; Vera-Rodriguez, R.; Fierrez, J.; Morales, A.; Ortega-Garcia, J. Deepfakes and beyond: A survey of face manipulation and fake detection. *Inf. Fusion* **2020**, *64*, 131–148. [CrossRef]

17. Tang, S.; Shi, Y.; Song, Z.; Ye, M.; Zhang, C.; Zhang, J. Progressive Source-Aware Transformer for Generalized Source-Free Domain Adaptation. *IEEE Trans. Multimed.* **2024**, *26*, 4138–4152. [CrossRef]
18. Kumar, K.; Majumdar, A.; Kumar, A.A.; Chandra, M.G. Transform Based Subspace Interpolation for Unsupervised Domain Adaptation Applied to Machine Inspection. In Proceedings of the 2023 31st European Signal Processing Conference (EUSIPCO), Helsinki, Finland, 4–8 September 2023; pp. 1708–1712. [CrossRef]
19. Khatri, N.; Borar, V.; Garg, R. A Comparative Study: Deepfake Detection Using Deep-learning. In Proceedings of the 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 19–20 January 2023; pp. 1–5.
20. Li, Y.; Chang, M.C.; Lyu, S. Face X-ray for more general face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 5001–5010.
21. Bhattacharyya, C.; Wang, H.; Zhang, F.; Kim, S.H.; Zhu, X. Diffusion Deepfake. *arXiv* **2024**, arXiv:2404.01579. [CrossRef]
22. Wang, Z.; Zhang, Y.; Liu, Y.; Qin, C.; Coleman, S.A.; Kerr, D. LARNet: Towards Lightweight, Accurate and Real-Time Salient Object Detection. *IEEE Trans. Multimed.* **2024**, *26*, 5207–5222. [CrossRef]