applied sciences

MDPI

Article

Deepfake Detection Algorithm Based on Dual-Branch Data Augmentation and Modified Attention Mechanism

Da Wan, Manchun Cai *, Shufan Peng, Wenkai Qin and Lanting Li

College of Information and Cyber Security, People's Public Security University of China, Beijing 100038, China
* Correspondence: caimanchun@ppsuc.edu.cn

Abstract: Mainstream deepfake detection algorithms generally fail to fully extract forgery traces and have low accuracy when detecting forged images with natural corruptions or human damage. On this basis, a new algorithm based on an adversarial dual-branch data augmentation framework and a modified attention mechanism is proposed in this paper to improve the robustness of detection models. First, this paper combines the traditional random sampling augmentation method with the adversarial sample idea to enhance and expand the forged images in data preprocessing. Then, we obtain training samples with diversity and hardness uniformity. Meanwhile, a new attention mechanism is modified and added to the ResNet50 model. The improved model serves as the backbone, effectively increasing the weight of forged traces in the multi-scale feature maps. The Jensen–Shannon divergence loss and cosine annealing algorithms are introduced into the training process to improve the model's accuracy and convergence speed. The proposed algorithm is validated on standard and corrupted datasets. The experiments show that the algorithm proposed in this paper significantly improves effectiveness and robustness, with accuracies of 4.16%, 7.37%, and 3.87% better than the baseline model on DeepFakes, FaceSwap, and FaceShifer, respectively. Most importantly, its detection performance on the corrupted datasets DeepFakes-C, FaceSwap-C, and FaceShifer-C is much higher than that of mainstream methods.

Keywords: deepfake detection; data augmentation; attention mechanism; robustness

check for updates

Citation: Wan, D.; Cai, M.; Peng, S.; Qin, W.; Li, L. Deepfake Detection Algorithm Based on Dual-Branch Data Augmentation and Modified Attention Mechanism. *Appl. Sci.* **2023**, *13*, 8313. https://doi.org/10.3390/app13148313

Academic Editor: Antonio Fernández-Caballero

Received: 8 June 2023
Revised: 6 July 2023
Accepted: 14 July 2023
Published: 18 July 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

With the development of deep learning, especially computer vision technology, digital image tampering and forgery technology have been extensively promoted and widely used in various fields, such as making pornographic videos, spreading fake news, and manipulating political events. Compared with traditional digital image forgery methods, existing technologies are mainly based on automatic encoders and generative adversarial networks (GAN) [1]. This technology using deep learning methods to generate high-quality forged images is called "deepfake" [2]. People can generate realistic fake face images using deepfake that are difficult to detect with traditional techniques. If deepfake is misused, it can potentially bring substantial risks to an individual's reputation, public opinion control, national security, and other aspects, resulting in serious social problems and political threats [3]. Therefore, exploring deepfake detection technology has high practical value and is a hot research direction in academic and industrial circles.

Deepfake detection differs from the general image binary classification task in that deepfake datasets are continuously updated with the development of forgery methods. New forgery methods will solve the previously existing forgery problems, such as noticeable visual artifacts in the generated forged images, making the generated images or videos more realistic. Current forgery generation algorithms are divided into two categories. One replaces the face identity in generating the forged image, as shown in Figure 1b, and the other replaces the person's expression, which does not change the person's identity information, as shown in Figure 1c.



(**a**) (**b**) (**c**)

Figure 1. Face forgery classification: (**a**) Original real face image; (**b**) Face identity replacement; (**c**) Face expression reenactment.

Generating forged images and videos is mainly achieved using generative adversarial networks, explicitly using generators to generate fake data, while the discriminators classify the data and generate enough samples to be identified as fake through adversarial learning. In addition to generative adversarial networks, there are methods based on autoencoders or some face graphics. Regarding identifying forged images, different researchers have exploited forgery traces using the features of forged images, such as pixel features in the spatial domain, and have extracted frequency domain features or noise features using filters. Furthermore, convolutional neural networks have been fully utilized to achieve forgery detection by modifying the network structure. With the development of deepfake detection techniques, the detection accuracy on some current mainstream datasets has peaked. Therefore, more extensive, diverse, and challenging datasets must be built to better simulate the data in real-world application scenarios. In real-world scenarios, forged images may be compressed, have some noise added during transmission, or be cropped and transformed, which does not change the identity or expression information of the person and therefore does not affect the image's authenticity. However, these operations often have a certain degree of negative impact on the classification learning of detection models. The distribution of deepfake datasets in the practical setting is usually complex and unpredictable because some forged images may have been compressed and smoothed in the post-processing process or damaged during transmission. Moreover, some attackers add elaborate interference to the image to induce detection models to make a false judgment [4]. As a result, detection models tend to lose their performance and generally need better robustness when detecting naturally corrupted or artificially perturbed images.

In deepfake detection, researchers need to pay more attention to the impact of hardness and diversity in the training data on the robustness of detection models. Inspired by the fact that data augmentation has been widely used in many fields to increase the diversity of training sets and improve model robustness [5,6], this paper first examines the data augmentation dimension. Researchers can make detection models pay more attention to critical features more effectively by adding attention mechanisms to the backbone networks. However, commonly used channel or spatial attention mechanisms can only capture local information effectively without establishing functional dependency between space and channels, making it challenging to use the global information of forged images entirely. In this paper, the training data and model structure are combined to design an algorithm based on an adversarial dual-branch data augmentation framework and the improved ResNet50 [7], embedded with a modified attention mechanism module to improve the robustness of detection models. The main work and contributions are as follows:

1. The performance of detection models deteriorates when presented with corrupted images. To mitigate this problem of feeble model robustness, we have introduced two data augmentation frameworks—Augmix [8] and Augmax [9]—to enhance and expand deepfake datasets in the preprocessing stage, respectively. The Augmax

framework amalgamates Project Gradient Descent (PGD) [10,11] with the Augmix framework, engendering more diverse and challenging exemplars.
2. In the training phase, we have designed and incorporated a novel attention mechanism module, designated MSE, by modifying and integrating Squeeze-and-Excitation Networks (SENet) [12] into ResNet50, thereby deriving an enhanced model (i_ResNet50) as the backbone. This module augments the model's capability to learn essential information in spatial and channel dimensions across multi-scale feature maps, enabling more efficacious mining of forgery traces in images.
3. We have deployed the cosine annealing algorithm to modulate the learning rate during training and expedite the model's convergence.
4. In the integrated decision-making phase, we have formulated a multi-branch loss function and introduced the Jensen-Shannon [13] divergence loss to improve the correlation between standard and augmented images.

The structure of our treatise is organized as follows. Section 2 reviews the extant literature on deepfake detection algorithms, highlighting limitations that motivate our proposed approach. Section 3 delineates our method's underlying principles and implementation details, encompassing data augmentation during preprocessing and model enhancement during training.. Section 4 provides specifics on the experimental setup, datasets, and results that validate the improved robustness of our deepfake detector. Section 5 concludes with a summary of contributions and suggested directions for future inquiry to advance the field.

2. Related Work

According to the different data-driven methods, deepfake detection algorithms fall into two main categories, including video-level detection and image-level detection.

2.1. Video-Level Detection

Existing forged videos are typically generated in a frame-by-frame manner without inter-frame smoothing, resulting in discrepancies between frames. Prominent video-level detection techniques employ recurrent neural networks (RNNs) to model temporal inconsistencies and diagnose forgeries. David et al. [14] pioneered an adaptive timing method using CNN frame features and RNN timing modeling. Sun et al. [15] introduced LRNet to extract inter-frame timing cues. Sabir et al. [16] aligned faces and learned bidirectional timing networks. Agarwal et al. [17] encoded physiological face/head signals over time and expanded feature correlations via Pearson coefficients before Support Vector Machine (SVM) classification. Zhang et al. [18] designed a 3D CNN (TD-3DCNN) to sample temporal inconsistencies. While video-level methods can capture rich inter-frame dynamics, they require extensive video preprocessing and are sensitive to compression artifacts and lighting changes. Moreover, single-frame detection remains an open challenge. Advances in robust video-level architectures and generalized frame diagnostics would further push the frontier.

2.2. Image-Level Detection

Image-level deepfake detection typically extracts frame features using CNNs for discrimination. Afchar et al. [19] employed compact convolution modules for fine-grained feature learning. Rossler et al. [20] leveraged Xception for full-frame and face features. Face-specific methods often outperform full-frame approaches. Peng et al. [21] detected forgeries from high-freq shallow layer features using Gaussian filtering. Coccomini et al. [22] extracted spatial cues using EfficientNet, Vision Transformer and CrossFit. Nguyen et al. [23] designed a capsule network with VGG-injected face features. Simonyan et al. [24] added attention to focus on tampered regions. Zhou et al. [25] fused RGB and noise cues in a dual-flow Faster R-CNN.

While achieving strong performance on standard datasets, most methods do not consider training diversity and corruption robustness. Few have verified robustness

on distorted inputs, leaving models susceptible in practice. Enhancing diversity and corruption resilience are imperative for generalized deepfake detection. We propose adversarial augmentation and attention mechanisms to improve model robustness. We cordially invite the esteemed research community to provide constructive feedback on this crucial technological challenge. Your insights will help strengthen defenses against this emerging threat.

3. Data Augmentation and Model Design

This section describes the overall structure of our proposed algorithm and explains the principles of the dual-branch data augmentation framework. Then, we present the process of modifying the attention mechanism module and introduce the construction of the multi-branch loss function.

3.1. Integral Structure of the Algorithm

The specific implementation of the algorithm proposed in this paper includes three main stages. First, design a dual-branch data augmentation framework to extend and enhance the standard images in the pre-processing data stage. Second, train the standard and enhanced datasets using the i_ResNet50 network with the modified attention mechanism in the model training stage. Last, construct a multi-branch loss function to decide the authenticity of the input images in the comprehensive decision stage. As can be seen in Figure 2, it shows the overall structure of the proposed algorithm.

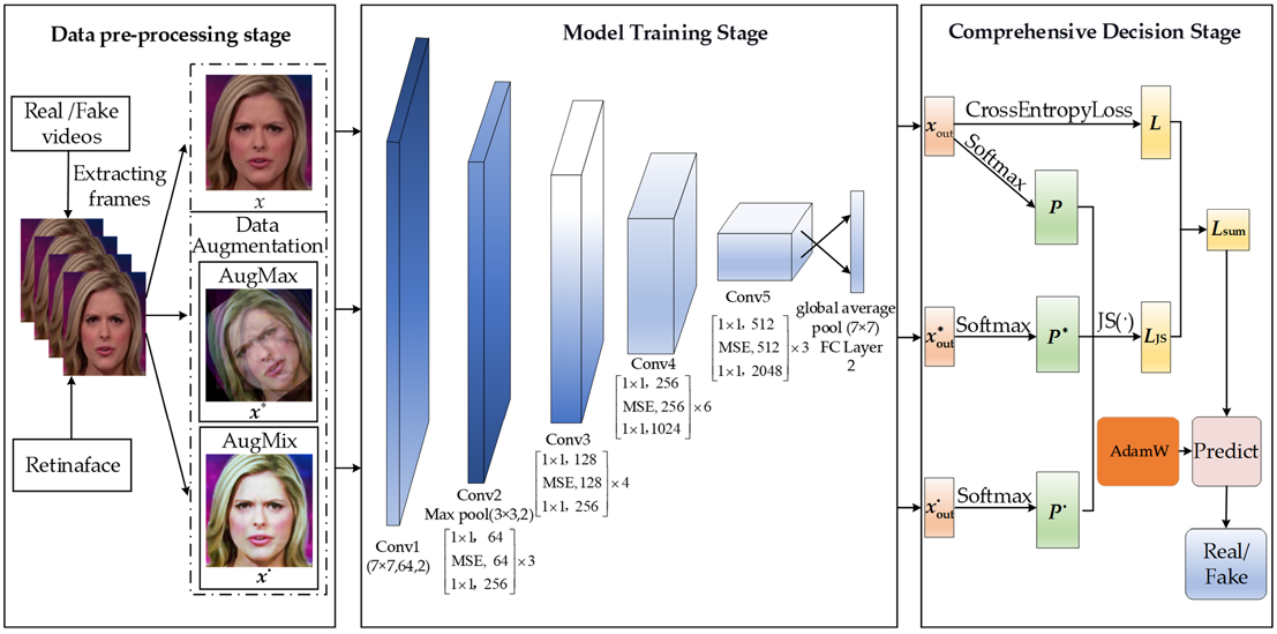


Figure 2. The overall architecture of the proposed deepfake detection algorithm based on dual-branch data augmentation framework and the modified attention mechanism.

We propose a dual-branch augmentation framework for robust deepfake detection. The first branch applies the Augmix algorithm for randomized combination, while the second leverages our novel Augmax algorithm to generate adversarial examples. We modify ResNet50 by integrating the proposed MSE attention modules, replacing 3×3 convolutions to enhance multi-scale feature learning on forgeries. The original and augmented images are fed to the improved i_ResNet50 models for parallel training with shared parameters. Their outputs are combined in the integrated decision stage. We use cross-entropy loss on original data as the raw loss. Critically, we introduce a Jensen–Shannon divergence consistency loss between original and augmented outputs to capture correlations. Their sum provides the total loss. For stable large-scale learning, we use AdamW for optimiza-

tion, which scales down gradients and provides automatic weight decay. To accelerate convergence, we apply cosine annealing to dynamically modulate the learning rate during training. We believe this adversarial data augmentation approach coupled with attention mechanisms will significantly enhance deepfake detection robustness.

3.2. Adversarial Dual-Branch Data Augmentations Framework

In previous deepfake image detection studies, researchers usually used or modified deep neural network models on standard datasets to explore results. However, when deploying detection models to a specific application environment, the samples to be detected are not as evenly distributed as the experimental datasets, and most counterfeit face images detected are dirty. These problems that contain noise and anomalies are mainly reflected in the inaccuracy and inconsistency of the forged images, including natural corruptions (e.g., camera blur or noise), sensory disturbance (e.g., loss in the transmission, electromagnetic interference), domain shifts (e.g., summer $\rightarrow$ winter), and human interference (e.g., confrontation sample). The limited forged image datasets cannot cover enough realistic scenes.

Furthermore, the existing detection models are vulnerable to detecting the images subjected to natural destruction or elaborate design. They are easily influenced to make misclassification. These unforeseeable changes in sample distribution seriously affect the reliability of detection models and make them challenging to play the expected effect in practice. To deal with the enormous risks brought by the continuous development of deepfake, more reliable and robust deepfake image detection algorithms urgently need to be proposed.

Several existing techniques to consolidate model robustness include using robust data augmentations, stability training, pre-training, and building robust network structures. Among these, data augmentations have the characteristics of easy implementation, low computing cost, high reliability, and plug-and-play. Therefore, the data augmentation methods are used to construct a more diverse and hardened training sample, thus improving the model's resistance to interference.

Image augmentations algorithms include two main categories. The first category is to increase the diversity of training samples by combining multiple random transformations, such as the Augmix algorithm. The other is to use the idea of the adversarial sample and induce the error of detection model decision by sampling from the worst-case samples to improve the hardness of the training data. A common way to obtain the adversarial samples is to add an adversarial perturbation. This paper designed a dual-branch data augmentation framework to enhance the forged image by combining Augmix and Augmax. First, use Augmix to enhance the standard input images, improving the robustness and uncertainty estimation of the image classification model by creating a new synthetic image from a clean image sample. Illustrate the specific implementation process in Figure 3.

The multi-branch transformations and select single or multiple augmentation operators are set up in each branch to transform the input image x separately. Then, the multi-branch data mixing results are combined with the original image by randomly sampling linear parameters m and $\boldsymbol{w}$ to obtain the enhanced sample $x^{\bullet}$. The enhanced datasets are more widely distributed and diverse, but their hardness still needs improvement. The datasets generated by simply combining random transformations are not sufficiently stiff to influence the final decision of detection models. Therefore, the latest data enhancement algorithm, Augmax, is introduced in the second augmentation branch. It can unify the diversity and hardness of samples in an augmentation framework by combining the random sampling augmentation method and the adversarial sample idea. The structure of Augmax is depicted in Figure 4.

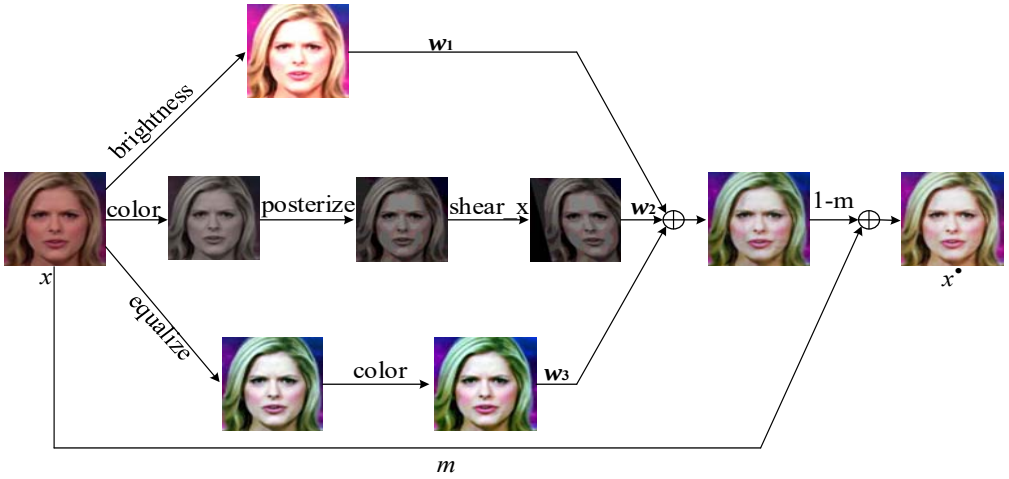


Figure 3. The Augmix framework overview: The brightness, posterize, color, equalize, and shear_x on each branch are augmentation operators; ⊕: Element-wise summation.

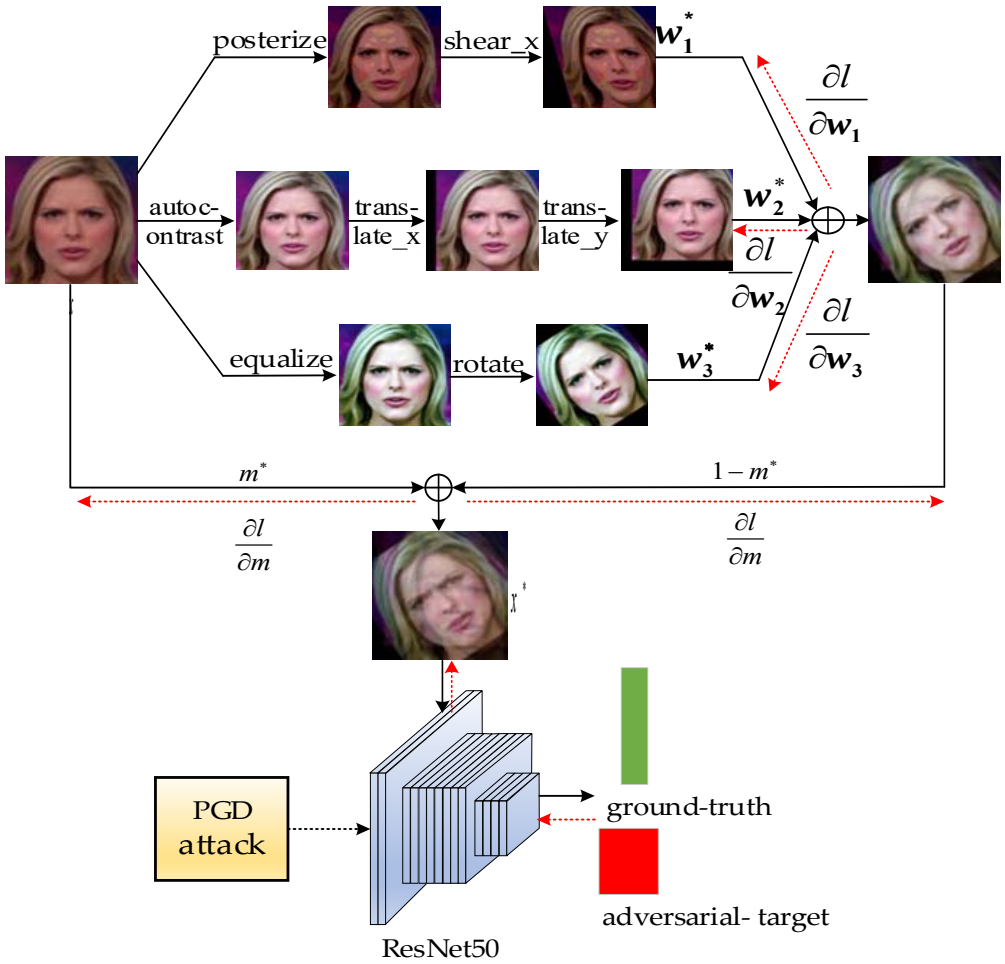


Figure 4. The Augmax framework overview: The auto contrast, posterize, rotate, equalize, translate_x, translate_y, and shear_x are augmentation operators; the red dotted line represents the backpropagation process; PGD: the Project Gradient Descent attack.

The Augmax framework adds an optimization process to the original random transform hybrid framework for selecting image blending weights m and w by introducing the PGD attack. This paper chooses ResNet50 as the attacked model and induces it to

misjudgment during adversarial training. We can obtain new weights m^* and $\boldsymbol{w}^*$ after the backpropagation for adversarial augmentation mixing.

Given a standard deepfake dataset F over images $x \in \mathbb{R}^{\mathrm{d}}$ and labels $y \in \mathbb{R}^{\mathrm{d}}$, we need to train a classifier $f : \mathbb{R}^{\mathrm{d}} \to \mathbb{R}^{\mathrm{c}}$ from x to output Softmax probabilities, parameterized by θ. The classifier f is robust for unforeseeable natural corruption and deliberate manufactured damage. The minimize experience loss is minimized in the training process, as shown in Equation (1).

$$\min_{\theta} E(x, y) \sim {}_{F}\, Loss(f(x; \theta), y) \tag{1}$$

where *Loss* represents loss function (e.g., cross-entropy loss). The PGD algorithm is used to attack ResNet50 during adversarial training, constantly adjusting the parameters m^* and $\boldsymbol{w}^*$ through backpropagation to generate hybrid-enhanced images, inducing the model to make misjudgments. Equation (2) shows the process of generating enhanced images.

$$x^* = g(x; m^*, \boldsymbol{w}^*) \tag{2}$$

where g represents the confrontation training function and x represents the original image. The symbol x^* represents the image enhanced by adversarial augmentation, it not only has the characteristics of wide distribution and diversity of samples $x^{\bullet}$ in Figure 3, but also has higher data hardness. It is challenging to be recognized by detection models and can effectively simulate the "dirty data" encountered in the natural environment. The parameters $\boldsymbol{w}^*$ and m^* maximize the experience loss and drive ResNet50 to make as many misjudgments as possible during adversarial training. The generation process of $\boldsymbol{w}^*$ and m^* is depicted in Equation (3).

$$\boldsymbol{w}^*, m^* = \underset{\boldsymbol{w}, m}{\operatorname{argmax}}(Loss(f(g(x; m, \mathrm{w}); \theta), y)), s.t. m \in [0, 1], \boldsymbol{w} \in [0, 1]^{\mathrm{b}}, \boldsymbol{w}^T \cdot 1 = 1 \tag{3}$$

Finally, the standard dataset and the two sets of enhanced data obtained after the dual-branch image augmentation framework are input into the backbone in parallel for training. The empirical loss of the training process is minimized, as shown in Equation (4).

$$\min_{\theta} \mathrm{E}(x, y) \sim {}_{F}\, Loss(f(x; \theta), y) + \lambda\, \mathrm{JS}(f(x; \theta), f(x^*; \theta), f(x^*; \theta)) \tag{4}$$

The function JS denotes the Jensen–Shannon divergence consistency loss, which serves to regularize the enhanced images to have a similar model output to the original images and improve their feature similarity. The symbol λ is the trade-off parameter. Standard and robustness data enhanced by augmentations have different potential feature distributions and a higher feature heterogeneity, which significantly burdens the model normalization process during training. Therefore, this paper introduces a novel normalization layer named Dual Batch-and-Instance Normalization (DuBIN) [9]. The layer can separate the feature distribution of standard images and two types of enhanced images and carry out normalization operations on data of different distributions, effectively improving the model's efficiency.

3.3. The Modified Attention Mechanism Module

Natural corruptions obscure forgery cues, confounding detection. CNNs inherently under-prioritize high-frequency details—deeper layers learn coarser features missing localized noise traits. Meanwhile, reduced deepfake artifacts further complicate identification. We propose contextual attention mechanisms alongside augmentation to extract corruption-resilient cues. Key discriminate details often persist locally despite interference. Our novel, multi-scale, squeeze-and-excitation module, MSE, recalibrates feature maps through dual-branch convolutions. Unlike mainstream spatial/channel attention, MSE effectively aggregates multi-scale spatial-channel dependencies. It enables mining subtle

forensic traces resilient to corruption. The multi-resolution architecture further enhances representation. MSE strengthens informative features and suppresses irrelevant cues.

We believe attention mechanisms are indispensable alongside augmentation for combating inevitable real-world perturbations. We humbly invite the esteemed community's feedback to advance deepfake detection robustness amidst this pivotal technological challenge collectively.

Figure 5 shows the specific implementation process of the MSE module. In order to extract more spatial information from feature maps, a dual-branch convolutional network is constructed by using two scales of the convolutional kernel to process feature maps. First, we convolve the feature maps $F(i)$ using convolution kernels with scales 3×3 and 5×5, respectively. The input channel dimension of $F(i)$ is C. The dual-branch extracts the spatial information of the feature maps to obtain the feature maps F_0 and F_1. The output channel dimension of each feature map is $C^*/2$. Second, the multi-scale feature map $F^*(i)$ is obtained by connecting F_0 and F_1 in the channel. And the channel of $F^*(i)$ is C^*. In order to decrease the computation of the convolution process, we use group convolution in this paper. The symbol G in the convolution operation represents the group size. Third, we use SENet to extract the attention weights on the dual-branch feature map channels. The implementation process includes F_{sq} (Squeeze) and F_{ex} (Excitation). Equation (5) depicts the implementation of F_{sq}.

$$g_i = F_{sq}(F_i) = \frac{1}{H \times W} \sum_{j=1}^{H} \sum_{k=1}^{W} F_i(j,k), i = 0,1 \tag{5}$$

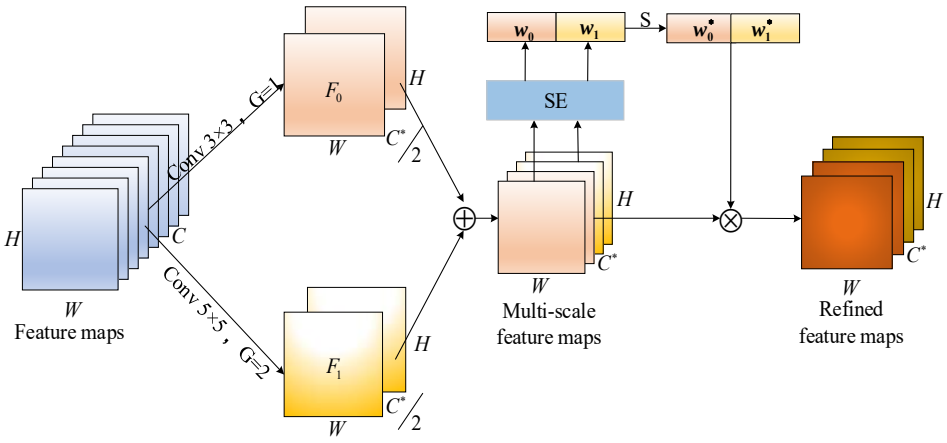


Figure 5. The specific implementation process of Modified Squeeze-and-Excitation(MSE). SE: Squeeze-and-Excitation Network; S: Softmax function; ⊗: Element-wise product.

The squeeze operation compresses features along the spatial dimension of feature maps, turning each two-dimensional feature channel into a $1 \times 1 \times C^*$ feature vector by global averaging pooling. This vector has a global sensitivity field to some extent. The F_0 and F_1 branches' statistics in channels are g_0 and g_1 obtained by F_{sq}. Then, the F_{ex} operation is performed, as shown in Equation (6).

$$\boldsymbol{w}_i = F_{ex}(g_i, W) = \sigma(W_i \delta(W_i(g_i))), i = 0,1 \tag{6}$$

The excitation operation generates weights for each feature channel through the parameter W. It learns to explicitly model the correlation between channels. Where W_0 and W_1 represent the fully connected (FC) layers that play the increasing and decreasing dimension, the symbol δ represents the Rectified Linear Unit (ReLU), and σ represents the Sigmoid function. The channel weights $\boldsymbol{w}_0$ and $\boldsymbol{w}_1$ of the two branches are obtained through F_{ex}, and then they are integrated across dimensions. Equation (7) represents the

multi-scale channel weight vector $\boldsymbol{w}$ of the feature graph obtained by the cross-dimensional fusion of $\boldsymbol{w}_0$ and $\boldsymbol{w}_1$ without destroying the channel attention vector.

$$\boldsymbol{w} = \boldsymbol{w}_0 \oplus \boldsymbol{w}_1 \tag{7}$$

where the symbol $\oplus$ in Equation (7) is the cross-dimensional fusion operation. We recalibrate the weights of the multi-scale feature maps, as shown in Equation (8).

$$\boldsymbol{w}_i^* = \text{Softmax}(\boldsymbol{w}_i) = \frac{\exp(\boldsymbol{w}_i)}{\sum_{i=0}^{1} \exp(\boldsymbol{w}_i)}, i = 0, 1 \tag{8}$$

The Softmax function is used to recalibrate the channel direction attention vector $\boldsymbol{w}_i$ to obtain $\boldsymbol{w}_i^*$, which contains all the spatial position information of the feature maps and the attention weight in channels, realizing the interaction between local and global channel attention. The channel attention vector $\boldsymbol{w}_0^*$ and $\boldsymbol{w}_1^*$ recalibrated by the feature are contacted to obtain the channel attention weight $\boldsymbol{w}^*$ of the whole feature maps, as shown in Equation (9).

$$\boldsymbol{w}^* = \boldsymbol{w}_0^* \oplus \boldsymbol{w}_1^* \tag{9}$$

Last, we multiply the recalibrated multi-scale attention weight $\boldsymbol{w}^*$ by the feature graph $F^*(i)$, obtaining an advanced feature graph F_{out} with rich multi-scale feature information, as shown in Equation (10).

$$F_{\text{out}} = F^*(i) \otimes \boldsymbol{w}^* = \text{Cat}([F_0 \otimes \boldsymbol{w}_0^*, F_1 \otimes \boldsymbol{w}_1^*]) \tag{10}$$

ResNet50 is a residual network consisting of multiple bottleneck modules that contain cascaded convolutional layers. The MSE module is used to replace the Conv 3 × 3 layer in bottlenecks and introduce the DuBIN layer to replace the batch normalization layer (BN), obtaining an improved bottleneck module. The improved ResNet50 (i_ResNet50) serves as our backbone network, as shown in Figure 6.

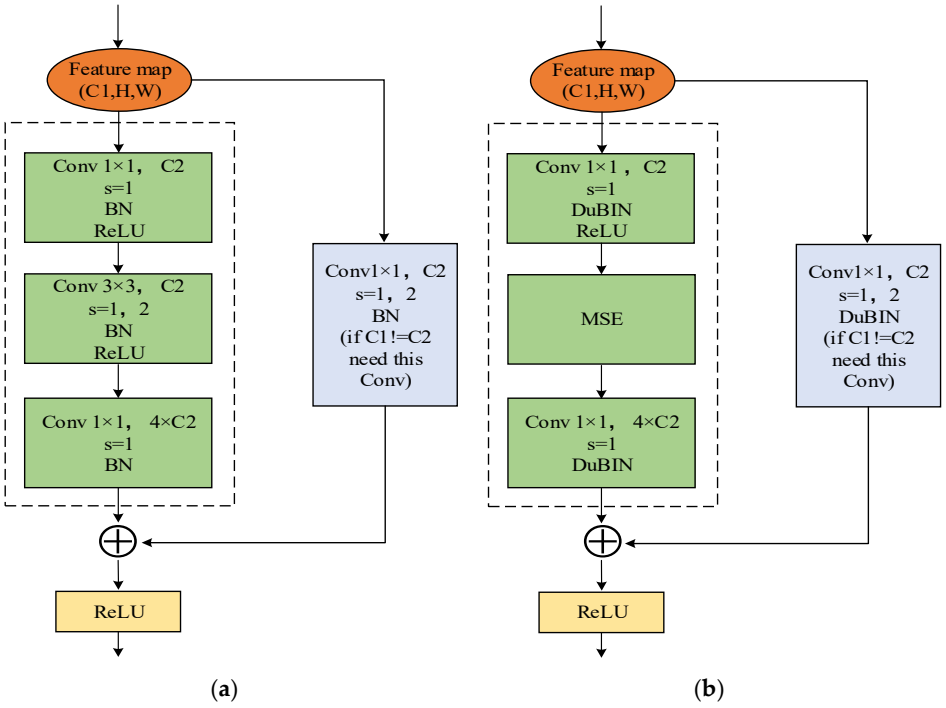


Figure 6. The improved bottleneck module in ResNet50: (**a**) Description of the bottleneck in ResNet50; (**b**) Description of the bottleneck embedded with the MSE module and DuBIN layer.

The i_ResNet50 has a powerful multi-scale feature extraction capability. It can adaptively recalibrate the weights across forged images' channel dimensions. Furthermore, it better obtains the information interaction between local and global channels. Meanwhile, the backbone can disentangle the instance-level heterogeneity and reduce computational effort during the training stage by the DuBIN layer.

3.4. The Multibranch Loss Function

In the comprehensive decision stage, we design the training process in three branches to ensure that the backbone can effectively learn the standard and enhanced images. The images x, $x^{\cdot}$, and x^{*} are inputted to the i_ResNet50. We train models in parallel and share parameters between models. The cross-entropy loss is used in the training process. The Jensen–Shannon divergence is introduced to improve the stability of the model. As shown in Equations (11)–(14), the training outputs of the three input datasets are x_{out}, $x^{\cdot}_{\text{out}}$, and x^{*}_{out}. Their Softmax probability values are calculated to obtain P, $P^{\cdot}$, and P^{*}. The average value of these probability values is P_{avg}.

$$P = \text{Softmax}(x_{\text{out}}) \tag{11}$$

$$P^{\bullet} = \text{Softmax}(x^{\bullet}_{\text{out}}) \tag{12}$$

$$P^{*} = \text{Softmax}(x^{*}_{\text{out}}) \tag{13}$$

$$P_{\text{avg}} = \frac{1}{3}(P + P^{\bullet} + P^{*}) \tag{14}$$

Equation (15) depicts the Jensen–Shannon divergence loss for the three datasets during training, where KL is the Kullback–Leibler divergence.

$$L_{\text{JS}} = \frac{1}{3}(\text{KL}(P \| P_{\text{avg}}) + \text{KL}(P^{\bullet} \| P_{\text{avg}}) + \text{KL}(P^{*} \| P_{\text{avg}})) \tag{15}$$

The backbone optimizes parameters based on the total loss L_{JS} during the training stage by backpropagation, eventually converging the model. The total loss is the sum of the cross-entropy loss L_{JS} and the standard images output x_{out} with the sample label, as shown in Equation (16).

$$L_{\text{sum}} = \text{CrossEntropyLoss}(x_{\text{out}}, label) + \lambda L_{\text{JS}} \tag{16}$$

4. Experiment and Analysis

4.1. Experimental Environment and Datasets Introduction

The experiments in this paper are based on the PyTorch deep learning framework. The specific experimental environment is shown in Table 1.

Table 1. Experimental environment configuration.

Configuration	Parameters
Operating System	Windows10 Professional × 64(10.0, Internal Version 19044)
CPU	Intel(R) Xeon(R) Gold 6346 CPU @ 3.10 GHz (4 CPUs), ~3.1 GHz
RAM	32 GB
GPU	NVIDIA A40
Video Memory	45 GB
Anaconda	4.10.3
PyTorch	1.10.2

We utilize the extensive FaceForensics++ (FF++) [20] with c23 compression for simulations. FF++ contains diverse forgery techniques: DeepFakes auto-encoders, Face2Face graphics, FaceSwap, and NeuralTextures. We focus on DeepFakes, FaceSwap, and the

recent advanced FaceShifter [26] technique from CVPR 2020, which improves facial replacement fidelity through occlusion handling.

As forgeries concentrate in facial regions, we preprocess FF++ by extracting 30 equally spaced frames per video and applying RetinaFace face detection followed by alignment and cropping to 299 × 299. This focuses the detector on the manipulated areas.

The dataset is split 7:3 into train and test sets as elaborated in Table 2. We believe FF++ encompasses the diversity and scale imperative for rigorous deepfake detection research.

Table 2. Introduction of the datasets used in this paper.

Datasets	DeepFakes	FaceSwap	FaceShifter
Training numbers	44784	44648	44784
Testing numbers	18215	18162	18216

4.2. Experimental Setting

In this paper, we abstract the forgery image detection task as a binary classification problem and train a binary classifier for each dataset. The one-dimensional binary vector of the backbone output is converted into the interval (0, 1) using the Softmax. We set the classification threshold to 0.5 and use the accuracy (Acc) and area under the ROC curve (AUC) as model evaluation indexes. The AUC is calculated from the true positive rate (TPR) and true negative rate (TNR), where TPR is also called sensitivity, and TNR is called specificity. The formulas for these indicators are as shown in Equations (17)–(20).

$$\text{Acc} = \frac{(TP + TN)}{(TP + FP + FN + TN)} \tag{17}$$

$$TPR = \frac{TP}{TP + FN} \tag{18}$$

$$TNR = \frac{TN}{TN + FP} \tag{19}$$

$$\text{AUC} = \frac{1}{2}\sum_{i=1}^{n-1}\left((1 - TNR)^{(i+1)} - (1 - TNR)^{(i)}\right) \times \left((TPR)^{(i+1)} + (TPR)^{(i)}\right) \tag{20}$$

The Acc is used to describe the classification accuracy of positive and negative image samples, and the more significant the Acc value, the better the classification effect of the model. The AUC is the value that the probability of actual image samples is greater than that of forged samples and can evaluate the performance index of detection models. The larger the AUC value, the more likely detection models will rank the positive samples ahead of the negative ones, where TP represents the positive samples predicted to be accurate, TN represents the negative samples predicted to be false, and FN represents the positive samples predicted to be false, FP represents the negative samples predicted to be accurate. The symbol n is the total number of positive and negative samples. The cosine annealing algorithm is introduced in model training to adjust the learning rate continuously, as shown in Equation (21).

$$\eta = \eta^i + \frac{1}{2}(\eta^i_{\max} - \eta^i_{\min})(1 + \cos(\frac{T_{cur}}{T_i}\pi)) \tag{21}$$

where i represents the times of the model has run, $\eta^i_{\max}$ and $\eta^i_{\min}$ represent the maximum and minimum values of the learning rate, and define the learning rate range. The symbol T_{cur} represents the number of epochs executed, and T_i represents the total number of epochs during the i-th run. After executing T_i epochs, it will simulate a restart by increasing the learning rate.

4.3. Ablation Experiments

In order to verify the accuracy and anti-interference ability gain of the proposed method on standard and corrupted datasets, two ablation experiments are designed in this paper.

4.3.1. Ablation Experiment on Standard Datasets

The dual-branch data augmentation framework (Aug) and the modified attention mechanism module (MSE) are gradually introduced in experiments. ResNet50 network serves as the baseline. We train standard images and two groups of enhanced images preprocessed by the Aug module in parallel. We experiment on standard datasets to verify the effectiveness of algorithms. We use the Acc and AUC as indicators to test the improvement of the Aug module and MSE module on the performance of the deepfake detection model. Table 3 shows the experimental results.

Table 3. Comparison of the gains generated by different modules under standard datasets based on ResNet50 network. The Metrics are Acc and AUC.

Models	Datasets	Acc/%	AUC
Baseline	DeepFakes	92.33	0.9794
	FaceSwap	88.57	0.9532
	FaceShifer	94.62	0.9869
Baseline + Aug	DeepFakes	93.84	0.9837
	FaceSwap	94.41	0.9866
	FaceShifer	97.64	0.9954
Baseline + MSE	DeepFakes	93.87	0.9855
	FaceSwap	92.91	0.9802
	FaceShifer	96.67	0.9922
Baseline + Aug + MSE (This Work)	DeepFakes	96.49	0.9924
	FaceSwap	95.94	0.9897
	FaceShifer	98.49	0.9974

It can be seen from Table 3 that, after introducing the Aug module, the average Acc of the model tested on three datasets has increased by 3.46%, and the average AUC has increased by 0.015. After introducing the MSE module, the average Acc has increased by 2.64%, and the average AUC has increased by 0.013. By introducing the Aug module and MSE module simultaneously, the average Acc and AUC have increased by 5.13% and 0.02, respectively. The results prove that the datasets enhanced by the Aug module have stronger hardness and broader distribution. While reducing the model's sensitivity to forged images, the detection model can learn in a richer sample space. After adding the MSE module to the backbone, the model can pay more attention to forged traces in images and suppress unimportant features to improve the model's performance. The Aug and MSE modules complement each other by improving the diversity and hardness of training samples and improving the learning ability of critical features in multi-scale feature maps, effectively improving the model's detection ability for forged images.

4.3.2. Ablation Experiment on Corrupted Datasets

The second ablation experiment adds artificial corruptions to standard images, such as Color_contrast, Gaussian_noise_color, and Jpeg_compression, generating three corrupted datasets. These corrupted datasets are Deepfakes-C, FaceSwap-C, and FaceShifer-C. They can simulate natural interference or artificial destruction in the natural environment. The robustness of Aug and MSE modules is verified on corrupted datasets. Table 4 shows the mode and severity of corruption used in experiments.

Table 4. The mode and severity of corruptions used in ablation experiments.

Mode	Severity
None	—
Color_saturation	0.1; 0.2; 0.3; 0.4; 0.5
Color_contrast	0.35; 0.475; 0.6; 0.725; 0.85
Block_wise	16; 32; 48; 64; 80
Gaussian_noise_color	0.001; 0.002; 0.005; 0.01; 0.05
Gaussian_blur	7; 9; 13; 17; 21
Jpeg_compression	2; 3; 4; 5; 6

The DeepFakes, FaceSwap, and FaceShifer are corrupted by different modes and severity of corruption. Some of the original images and their corrupted images in the three datasets are shown in Figure 7.

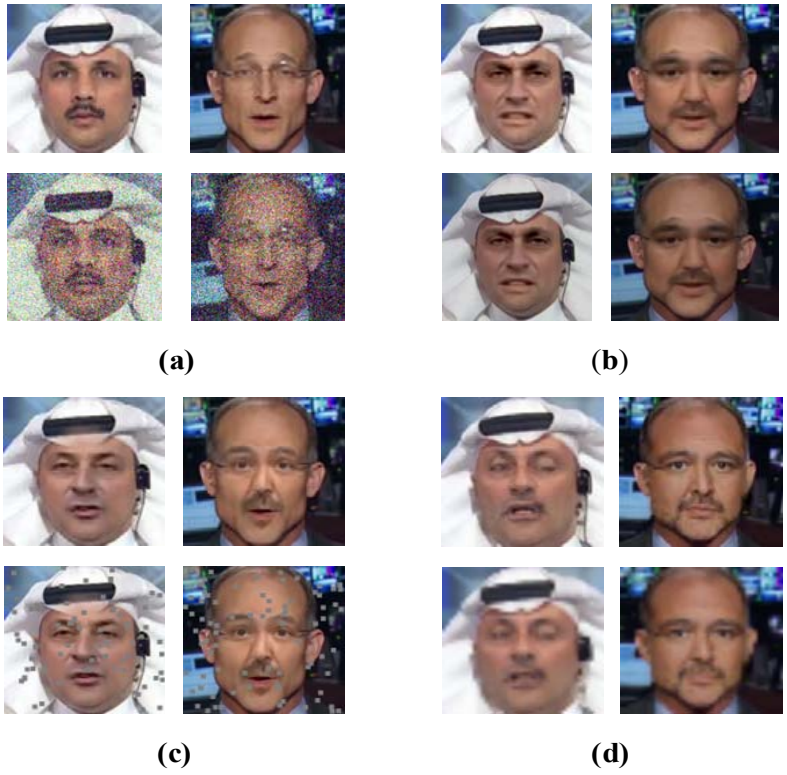


Figure 7. Some of the original and corrupted images: (**a**) The real images corrupted by the Gaussian_noise_color of severity 0.05 in FF++; (**b**) The images corrupted by the Color_contrast of in DeepFakes of severity 0.8 in DeepFakes; (**c**) The images corrupted by the Block_wise of severity 80 in FaceSwap; (**d**) The images corrupted by the Jpeg_compression of severity 6 in FaceShifter.

We introduce the Aug and MSE module to enhance and train standard datasets. Then, we experiment with the pre-trained models on corrupted datasets, verifying the gain of Aug and MSE modules on model robustness. The Acc and AUC of all experiments' results are summed separately as mean values as the final indicators of experiments, as shown in Equations (22) and (23). The results of the second ablation experiment on corrupted datasets are shown in Table 5.

$$\mathrm{Acc}_{\mathrm{avg}} = \frac{\mathrm{Acc}[0.35; 0.475; 0.6; 0.725; 0.85]^{\mathrm{Color_saturast}} + \cdots + \mathrm{Acc}[6; 5; 4; 3; 2]^{\mathrm{Jpeg_compression}}}{31} \tag{22}$$

$$\mathrm{AUC}_{\mathrm{avg}} = \frac{\mathrm{AUC}[0.35; 0.475; 0.6; 0.725; 0.85]^{\mathrm{Color_saturast}} + \cdots + \mathrm{AUC}[6; 5; 4; 3; 2]^{\mathrm{Jpeg_compression}}}{31} \tag{23}$$

Table 5. Comparison the gains generated by different modules under corrupted datasets based on ResNet50 network. The Metrics are Acc and AUC.

Models	Datasets	Acc/%	AUC
Baseline	DeepFakes-C	59.21	0.6121
	FaceSwap-C	60.04	0.6141
	FaceShifer-C	80.97	0.8873
Baseline + Aug	DeepFakes-C	79.29	0.8752
	FaceSwap-C	77.85	0.8544
	FaceShifer-C	89.45	0.9617
Baseline + MSE	DeepFakes-C	77.24	0.8254
	FaceSwap-C	81.36	0.8878
	FaceShifer-C	84.74	0.9329
Baseline + Aug + MSE (This Work)	DeepFakes-C	90.63	0.9654
	FaceSwap-C	91.61	0.9703
	FaceShifer-C	91.03	0.9719

The results show that the performance of ResNet50 on corrupted datasets is significantly lower than that on standard datasets, which indicates that ResNet50 has no practical detection ability for corrupted images. After introducing the Aug module in the preprocessing stage, the average Acc and AUC of detection models on three corrupted datasets have increased by 15.46% and 0.193, which proves that the pre-trained model on enhanced data has excellent robustness. After introducing the MSE module, the average Acc of the detection model has improved by 14.37%. The average AUC has improved by 0.178. These results prove that the MSE module's ability to extract critical features from multi-scale feature maps of forged images, effectively improves the detection ability of the model on corrupted images. While introducing the Aug and MSE modules simultaneously, the average Acc of the baseline on these corrupted datasets has increased by 24.35%, and the average AUC has increased by 0.265.

The improved models that are pre-trained on enhanced data can learn the forgery traces that are not changed due to interference in the forged image. The two modules complement each other to improve the robustness of detection models. The quality of images in FaceShifer is better than the first two. There needs to be more than the mode and degree of corruption set in this experiment to affect the detection performance seriously. The above experimental results prove the effectiveness and robustness of the algorithm proposed in this paper.

4.4. Contrast Experiments

To further verify the proposed algorithm's robustness improvement, a comparison experiment is designed to compare it with mainstream detection models. First, to effectively observe the variation in detection performance of mainstream detection models on standard and interference images, this experiment was set up to test these models on standard images. Four mainstream models, including MesoNet [19], XceptionNet [27], EfficientNet-B4 [28], and F3-Net [29], are selected for pre-training on DeepFakes, FaceSwap, and FaceShifer standard images. The experimental results of these mainstream algorithms on standard datasets are shown in Table 6.

Table 6. The performance of mainstream detection models on standard datasets with the metrics Acc and AUC.

Models	Datasets	Acc/%	AUC
XceptionNet	DeepFakes	95.65	0.9913
	FaceSwap	94.90	0.9897
	FaceShifer	97.36	0.9950
MesoNet	DeepFakes	93.46	0.9891
	FaceSwap	87.33	0.9444
	FaceShifer	96.27	0.9939
EfficientNet-B4	DeepFakes	95.00	0.9937
	FaceSwap	95.23	0.9892
	FaceShifer	98.03	0.9961
F^3-Net	DeepFakes	94.89	0.9912
	FaceSwap	95.15	0.9896
	FaceShifer	98.01	0.9938

The experimental results in Table 6 show that current mainstream deepfake detection models can perform excellently on standard datasets. These models can classify forged images well. Then, these pre-trained models were used to perform robustness tests on the three generated interference datasets. Table 7 shows experimental results.

Table 7. Comparison of the metrics Acc and AUC on corrupted datasets between mainstream models and the proposed model.

Models	Datasets	Acc/%	AUC
XceptionNet	DeepFakes-C	58.25	0.5744
	FaceSwap-C	59.53	0.5797
	FaceShifer-C	82.29	0.8721
MesoNet	DeepFakes-C	75.50	0.8360
	FaceSwap-C	71.35	0.7977
	FaceShifer-C	78.75	0.8934
EfficientNet-B4	DeepFakes-C	60.14	0.5783
	FaceSwap-C	58.98	0.5725
	FaceShifer-C	84.12	0.8721
F^3-Net	DeepFakes-C	57.78	0.5666
	FaceSwap-C	58.98	0.5725
	FaceShifer-C	84.67	0.8885
I_ResNet50 (This Work)	DeepFakes-C	90.63	0.9657
	FaceSwap-C	91.61	0.9703
	FaceShifer -C	91.03	0.9719

As seen in Table 7, The results of these mainstream models on the corrupted test sets are much lower than those on standard datasets. These methods no longer have effective detection performance on post-interference datasets. However, the proposed method in this paper can still perform well. The experimental results illustrate that the robustness of the proposed method is much further ahead compared to mainstream models. To further demonstrate that the work in this paper outperforms mainstream detection models in terms of robustness, the sensitivity (TPR) and specificity (TNR) metrics are used to measure the detection performance of these models on positive and negative image samples after being corrupted. The results are shown in Figure 8.

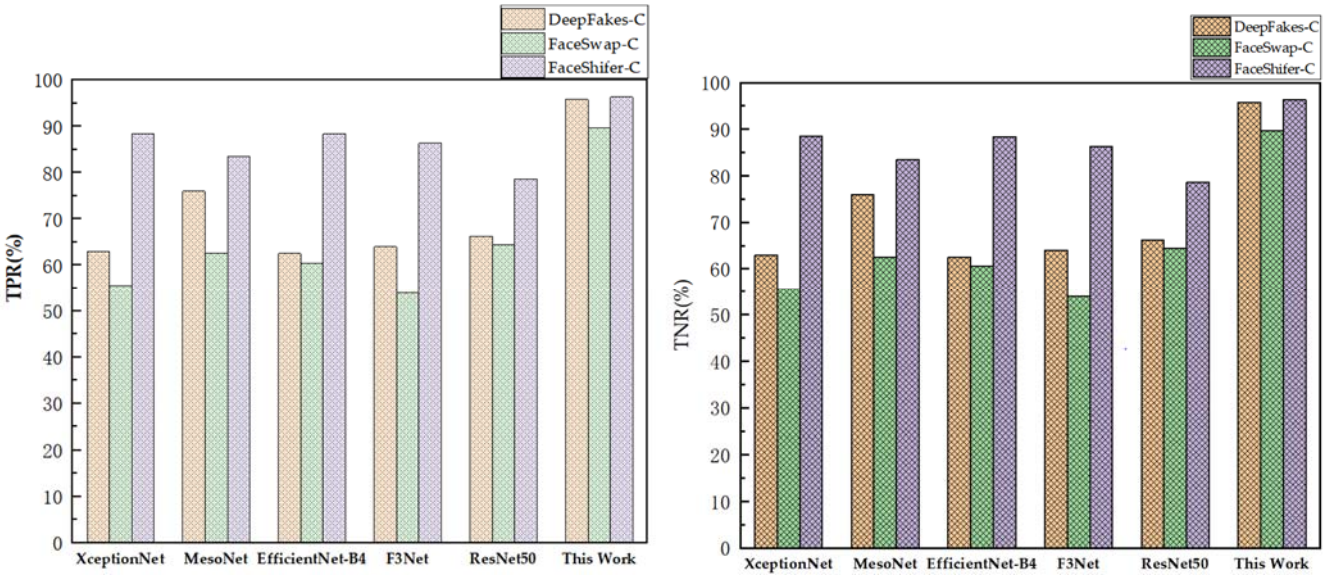


Figure 8. The results of mainstream algorithms and this paper work on robustness test: (**a**) True positive rate (TPR) of each model on different corrupted datasets; (**b**) True negative rate (TNR) of each model on different corrupted datasets.

Figure 8 shows that the algorithm proposed in this paper has higher specificity and sensitivity than mainstream models. It performs well in detecting genuine and fake face images after being disturbed. The Mainstream algorithms are usually constructed and verified on standard datasets, using shallow texture and high-level semantic features of images. When the forged images are subjected to unpredictable corruption, the original performance of these models cannot be guaranteed. Although some researchers have used rotation, cropping, translation, and other data augmentation algorithms to enhance the images, covering all the samples encountered in the natural detection environment is still challenging. The proposed algorithm enhances and expands training datasets from two dimensions of diversity and hardness. The generated training samples are more widely distributed, have higher hardness, and are closer to that in the natural detection environment. In the training stage, the model can focus on learning the critical features of forged data by improving the attention mechanism. Finally, we design the multi-branch loss function to make the model thoroughly learn the corrupted data and enhance the correlation between different datasets, significantly improving the model's robustness.

4.5. Visualization

The above experiments explore the regions of interest in detection models using the Gradient-weighted Class Activation Mapping (Grad-CAM) [30] visual attention maps. The results are shown in Figure 9.

The red part in Figure 9 indicates the area of focus of the neural network. This paper compares the difference in class activation mapping between the proposed method and the baseline model when detecting standard and corrupted images. ResNet50 pays uneven attention and focuses on regions that sometimes exceed the extent of the falsified regions, which is especially evident when detecting images after interference. This result explains the significant degradation of the baseline's detection performance on corrupted datasets. The proposed method focuses on the critical areas of high variances, such as the nose, mouth, and eyes, which can effectively find more subtle forgery clues. Moreover, the proposed method can accurately locate the forged regions of forged images on both standard and corrupted images and has a better detection ability.

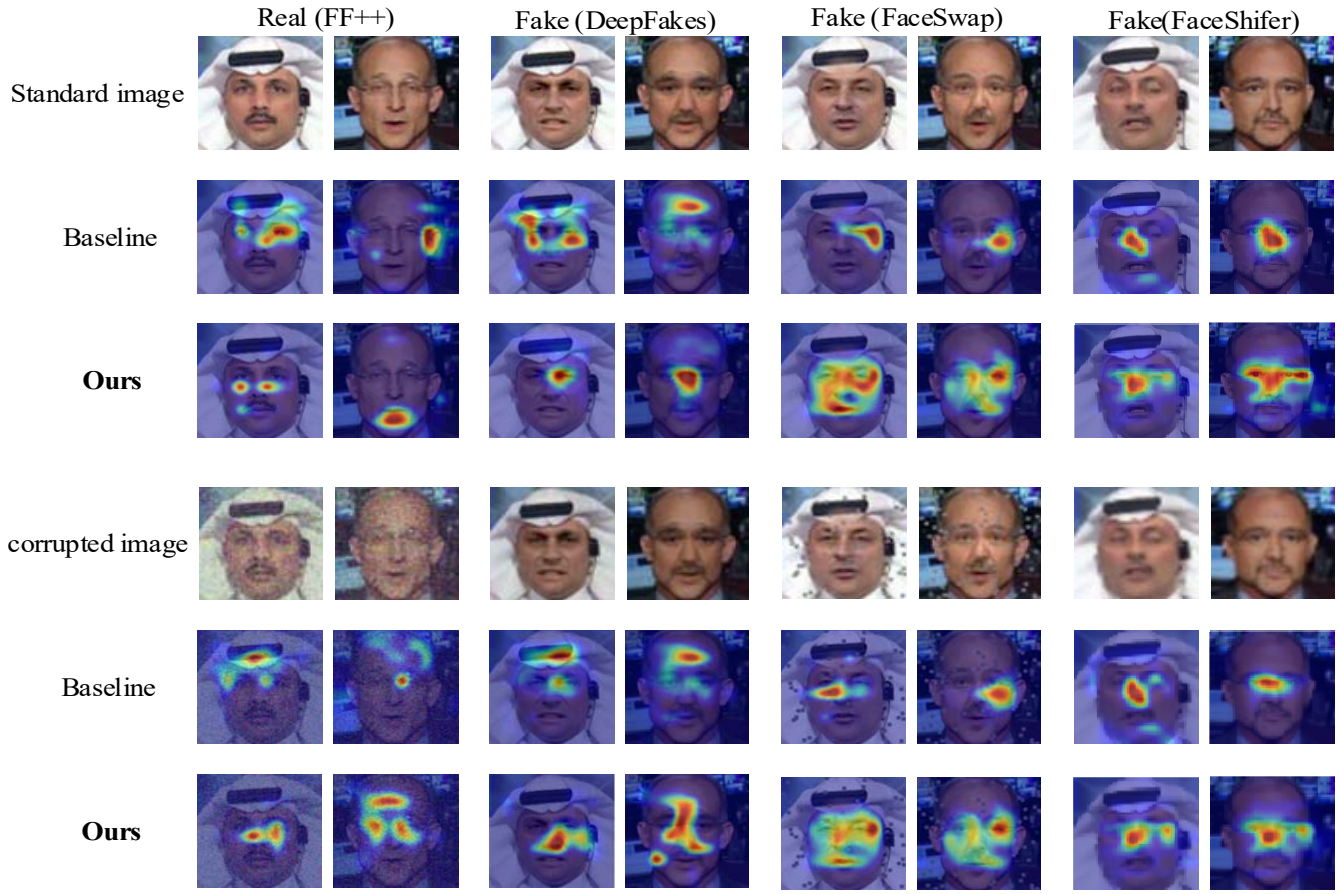


Figure 9. The Grad-CAM visualizes the attention maps of images and explores the detecting regions of interest for the models.

5. Conclusions and Future Work

We propose an algorithm by designing a dual-branch adversarial data augmentation framework to obtain training samples with greater diversity and construct a robust network structure by adding a modified attention mechanism. As demonstrated by the above experiments, we effectively solve the problems of insufficient extraction of critical features and the lower accuracy of detection models. Compared with the current mainstream algorithms, the proposed model has higher robustness and can perform well in unpredictable data distribution. At present, the development of forged image detection algorithms is usually limited to some specific datasets. Although the performance of detection models on these datasets is excellent, it needs more practical effects and is challenging to meet the needs of landing applications. In order to ensure that detection models can achieve ideal results in the practical environment, we should construct more abundant deepfake data samples and improve the cross-database testing ability of the algorithm. The proposed algorithm effectively fills the gap in deepfake detection, where detection models cannot recognize forgery images that suffer from human or natural corruptions. In future research, we can design dual-stream convolutional neural networks to fully use the richer multimodal features in the forged images from the feature fusion perspective, exploring the essential differences between authentic and forged images to improve the generalization of the detection model.

Author Contributions: Conceptualization, D.W., M.C. and S.P.; methodology, D.W. and S.P.; formal analysis, D.W. and M.C.; investigation, D.W., S.P. and W.Q.; resources, D.W. and S.P.; data curation, D.W. and M.C.; writing—original draft preparation, D.W.; writing—review and editing, M.C. and L.L.; visualization, D.W. and S.P.; supervision, M.C. and L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Double First-Class Innovation Research Project for People's Public Security University of China (No. 2023SYL07).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available in a publicly accessible repository.

Acknowledgments: All the face images used in this paper are from publicly available deepfake datasets, including DeepFakes, FaceSwap and FaceShifer.

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

References

1. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
2. Korshunova, I.; Shi, W.; Dambre, J.; Theis, L. Fast face-swap using convolutional neural networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
3. Juefei-Xu, F. Countering Malicious DeepFakes: Survey, Battleground, and Horizon. *Int. J. Comput. Vis.* **2022**, *130*, 1678–1734. [CrossRef] [PubMed]
4. Liu, Y.; Ma, S.; Aafer, Y.; Lee, W.-C.; Zhai, J.; Wang, W.; Zhang, X. Trojaning attack on neural networks. In Proceedings of the 2018 Network and Distributed System Security Symposium, San Diego, CA, USA, 18–21 February 2018; Internet Society: San Diego, CA, USA, 2018.
5. Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F.A.; Brendel, W. ImageNet-Trained CNNs Are Biased towards Texture; Increasing Shape Bias Improves Accuracy and Robustness. *arXiv* **2022**, arXiv:1811.12231.
6. Yun, S.; Han, D.; Chun, S.; Oh, S.J.; Yoo, Y.; Choe, J. CutMix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 10–17 October 2019; pp. 6022–6031.
7. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
8. Hendrycks, D.; Mu, N.; Cubuk, E.D.; Zoph, B.; Gilmer, J.; Lakshminarayanan, B. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. *arXiv* **2020**, arXiv:1912.02781.
9. Wang, H.; Xiao, C.; Kossaifi, J.; Yu, Z.; Anandkumar, A.; Wang, Z. AugMax: Adversarial Composition of Random Augmentations for Robust Training. *Adv. Neural Inf. Process. Syst.* **2022**, *34*, 237–250.
10. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. *arXiv* **2019**, arXiv:1706.06083.
11. Xie, C.; Tan, M.; Gong, B.; Wang, J.; Yuille, A.; Le, Q.V. Adversarial examples improve image recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
12. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2019.
13. Nielsen, F. On a Generalization of the Jensen–Shannon Divergence and the Jensen–Shannon Centroid. *Entropy* **2020**, *22*, 221. [CrossRef] [PubMed]
14. Guera, D.; Delp, E.J. Deepfake video detection using recurrent neural networks. In Proceedings of the 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Auckland, New Zealand, 1 November 2018; pp. 1–6.
15. Sun, Z.; Han, Y.; Hua, Z.; Ruan, N.; Jia, W. Improving the efficiency and robustness of deepfakes detection through precise geometric features. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
16. Sabir, E.; Cheng, J.; Jaiswal, A.; AbdAlmageed, W.; Masi, I.; Natarajan, P. Recurrent Convolutional Strategies for Face Manipulation Detection in Videos. *Interfaces* **2019**, *3*, 80–87.
17. Agarwal, S.; Farid, H.; Gu, Y.; He, M.; Nagano, K.; Li, H. Protecting World Leaders Against Deep Fakes. *CVPR Workshops* **2019**, *1*, 38.
18. Zhang, D.; Li, C.; Lin, F.; Zeng, D.; Ge, S. Detecting deepfake videos with temporal dropout 3DCNN. In Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence; International Joint Conferences on Artificial Intelligence Organization, Montreal, Canada, 19–27 August 2021; pp. 1288–1294.
19. Afchar, D.; Nozick, V.; Yamagishi, J.; Echizen, I. MesoNet: A compact facial video forgery detection network. In Proceedings of the 2018 IEEE International Workshop on Information Forensics and Security (WIFS), Hong Kong, China, 11–13 December 2018; pp. 1–7.
20. Rossler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Niessner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27–28 October 2019; pp. 1–11.

21. Peng, S.; Cai, M.; Ma, R.; Liu, X. Deepfake detection algorithm for high-frequency components of shallow features. *Laser Optoelectron. Prog.* **2022**, *60*, 1–18.
22. Coccomini, D.; Messina, N.; Gennaro, C.; Falchi, F. *Combining EfficientNet and Vision Transformers for Video Deepfake Detection*; Springer: Berlin/Heidelberg, Germany, 2022; Volume 13233, pp. 219–229.
23. Nguyen, H.H.; Yamagishi, J.; Echizen, I. Capsule-forensics: Using capsule networks to detect forged images and videos. In Proceedings of the ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 12–17 May 2018.
24. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
25. Zhou, P.; Han, X.; Morariu, V.I.; Davis, L.S. Two-stream neural networks for tampered face detection. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, 18–23 June 2018.
26. Li, L.; Bao, J.; Yang, H.; Chen, D.; Wen, F. FaceShifter: Towards High Fidelity and Occlusion Aware Face Swapping. *arXiv* **2020**, arXiv:1912.13457.
27. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
28. Tan, M.; Le, Q.V. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *arXiv* **2019**, arXiv:1905.11946.
29. Ye, X.; Xiong, F.; Lu, J.; Zhou, J.; Qian, Y. $\mathcal{F}$3-Net: Feature Fusion and Filtration Network for Object Detection in Optical Remote Sensing Images. *Remote Sens.* **2020**, *12*, 4027. [CrossRef]
30. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.