electronics

MDPI

Article

Efficient Vision-Based Face Image Manipulation Identification Framework Based on Deep Learning

Minh Dang

Department of Information Technology, FPT University, Ho Chi Minh City 70000, Vietnam; minhdl3@fe.edu.vn

Abstract: Image manipulation of the human face is a trending topic of image forgery, which is done by transforming or altering face regions using a set of techniques to accomplish desired outputs. Manipulated face images are spreading on the internet due to the rise of social media, causing various societal threats. It is challenging to detect the manipulated face images effectively because (i) there has been a limited number of manipulated face datasets because most datasets contained images generated by GAN models; (ii) previous studies have mainly extracted handcrafted features and fed them into machine learning algorithms to perform manipulated face detection, which was complicated, error-prone, and laborious; and (iii) previous models failed to prove why their model achieved good performances. In order to address these issues, this study introduces a large face manipulation dataset containing vast variations of manipulated images created and manually validated using various manipulation techniques. The dataset is then used to train a fine-tuned RegNet model to detect manipulated face images robustly and efficiently. Finally, a manipulated region analysis technique is implemented to provide some in-depth insights into the manipulated regions. The experimental results revealed that the RegNet model showed the highest classification accuracy of 89% on the proposed dataset compared to standard deep learning models.

Keywords: image manipulation; face manipulation; deepfakes; deep learning; face manipulation

Citation: Dang, M. Efficient Vision-Based Face Image Manipulation Identification Framework Based on Deep Learning. *Electronics* **2022**, *11*, 3773. https://doi.org/10.3390/electronics11223773

Academic Editors: Hyunsoo Yoon and Namgi Kim

Received: 25 October 2022
Accepted: 16 November 2022
Published: 17 November 2022

1. Introduction

The human face is one of the most crucial factors in social interaction because it can be used to identify an individual's identity and convey many different emotions. Therefore, the human face has been used extensively in our daily life for identity identification in applications, such as security, payment, and law enforcement [1]. However, these advancements also draw various attacks from malicious people aiming to bypass the systems as genuine users. In addition, technological advancements have simplified the process of manipulating face images, enabling them to spread like wildfire on the Internet [2]. As a result, it is critical to introduce effective and robust strategies to detect manipulated images created by facial forgery attacks to mitigate the negative impact and improve both privacy and security.

Common face image manipulation techniques are image splicing and copy-move forgery [3]. The copy-move forgery is performed by copying regions of an image and pasting them into other regions of that image. With a different approach, image splicing refers to the process of taking some regions of a source image and putting them into a target image to create a composite image. Finally, various post-processing methods can be implemented to rectify the manipulated regions to make distinguishing between genuine and manipulated images more challenging [4]. Moreover, it has become simpler for regular users to perform face manipulation via open-source software and mobile apps. The manipulated face images have been proved to fool face recognition models easily because those models recognize identity from the manipulated images as in authentic images [5].

As manipulated images have become more straightforward to create and become highly realistic, research devoted to detecting face image manipulation has increased

sharply in recent years. However, conventional machine learning (ML)-based models had various drawbacks because they relied mainly on manually extracting essential features from a specific dataset. Therefore, deep learning technology has been increasingly employed over traditional methods during the last few decades because it obtained state-of-the-art performance in standard computer vision (CV) topics, including vision-based manipulation detection [1,6].

This study introduces an automated manipulated face image identification system based on deep learning. We propose a large face manipulation database which was manually generated and validated. After, a fine-tuned RegNet-based framework for face manipulation classification was proposed. Finally, several test scenarios were introduced to verify the RegNet-based model's performance on the collected dataset and against other models.

The remainder of the study is arranged as follows. Section 2 briefly discusses previous face manipulation generation and detection approaches. After that, the fine-tuned face manipulation identification proposed in this study is explained in Section 3. Various experiments are offered in Section 4 to thoroughly evaluate the performances of the suggested model on the collected dataset and against previous models. Section 5 shows a brief discussion about the proposed framework. Finally, we summarize and discuss the proposed framework in Section 6.

2. Related Work

2.1. Conventional Face Manipulation Detection

Traditional ML-based methods for image manipulation have previously been extensively studied. However, these studies required a manual investigation and extraction of representative hand-crafted image features based on specific characteristics of the dataset under consideration, which was complicated and time-consuming. Some common techniques that are usually used to perform image forgery analysis are local noise estimation, illumination analysis, color filter array (CFA), steganalysis features, and double JPEG localization [7]. For instance, Yao et al. revealed inconsistent noise levels from manipulated regions of RGB images using the noise level function (NLF) approach. The proposed method achieved high accuracy using the Bayesian inference on the dataset [8]. Zhang et al. checked whether the quantization matrixes of the input images match well after implementing the double JPEG compression in order to detect manipulated images [9].

Conventional ML models show poor performance when post-processing procedures, such as noise reduction and image blending, are performed to reduce noise consistency. Moreover, conventional models are well-fitted to small datasets and usually performed poorly with unseen data.

2.2. Deep Learning-Based Face Manipulation Detection

Deep learning is part of ML that uses artificial neural networks (ANN) with representation learning. Over recent decades, it has showed state-of-the-art performances in various applications, including classification, segmentation, and other domains [6]. Current image manipulation identification studies have started to implement deep learning. For example, Mayer et al. introduces an image forgery detection framework using deep learning that automatically extracted representative features from input images [10]. The obtained results revealed that the system identified three types of image manipulation with an average accuracy of 99%. With a different approach, Zhou et al. [11] represented a novel deep learning-based image manipulation detection framework that learned two different features, including RGB color and noise features. The authors proved that the extracted two features significantly improved the model's performance and robustness on images generated by various tampering methods.

Recently, neural architecture search (NAS), a procedure for automating deep learning design, has gained popularity [12]. It relies on a neural structure search technique to propose an optimal structure based on the search space settings, such as computing

expense and optimal parameter set. Representative models based on NAS technology that have obtained state-of-the-art performances are EfficientNet [13] and MobileNetV3 [14]. Compared to these NAS-based models, the novel NAS-based RegNet model is different because the authors designed the network design spaces that parametrize populations of networks instead of concentrating on developing individual network instances [15]. RegNet has shown higher performance and fivefold faster speeds than the state-of-the-art EfficientNet model. As a result, this research proposes to fine-tune the RegNet model for performing face manipulation identification.

3. Methodology

The overall description of the proposed manipulated face classification framework is illustrated in Figure 1, which has three main processes. The first process is data collection, which gathers images containing human faces from the Internet. Then, various regions from the faces of the genuine images are altered to make manipulated face images. After that, both genuine and manipulated images are fed into the preprocessing process, including two sub-tasks. First, faces from those images are extracted and rescaled to a predefined size. Next, various image augmentation methods are implemented on both the authentic and manipulated datasets to enhance the model's performance and robustness. Finally, the preprocessed images are used to train different models to identify manipulated face images.

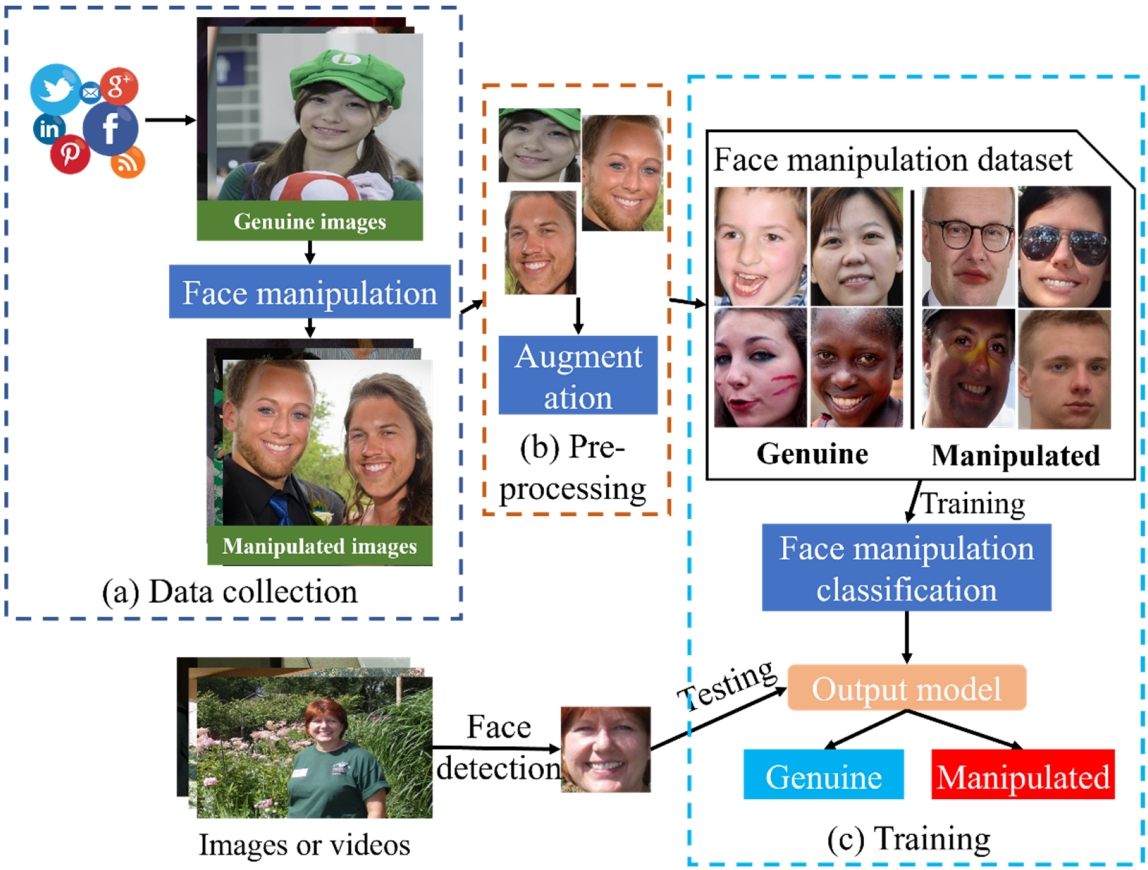

Figure 1. Overall architecture of the proposed face manipulation identification framework. Three main processes of the framework include (**a**) data collection, (**b**) preprocessing, and (**c**) training and evaluation.

3.1. Data Collection

This section explains the process of collecting images for the dataset with their statistics. Even though various face manipulation datasets have been proposed, they contain only a limited number of images and people. For example, a high-resolution DSI-1 face ma-

nipulation dataset was proposed by De Carvalho et al. [16], which only contains 25 images. On the other hand, large and challenging image manipulation databases, such as the CASIA V2 dataset [17] and the Columbia image splicing detection dataset [18], mainly comprise various objects without human faces. As a result, a huge and high-quality face manipulation dataset was proposed in this study, which supports the development of robust manipulated face image classification models.

A crawler was initially developed to automatically collect human face images from social media sites and search engines, such as Google Images and Baidu. A total of 5650 authentic images were downloaded, which involved varying poses under different illumination and surrounding environments. Moreover, the downloaded face image collection also covered a broad range of nationalities, genders, accessories, and facial hair.

Based on the collected genuine images, a team of 5 members took part in a manipulated face image generation task using various image manipulation approaches that lasted for two months. Each person made about 20 images per day on average. At the end of this process, 4188 manipulated face images were generated and validated manually. The image size ranges from 82 × 82 to 1098 × 1098. Figure 2 displays nine sample images for authentic and manipulated face images from the proposed dataset. Advantages of the dataset over existing datasets include (i) having the primary purpose of detecting face manipulation, and (ii) high quality manipulated regions of the faces and realistic manually made images.

Figure 2. Examples of genuine face images (**left**) and manipulated face images (**right**) generated from the collected genuine face images.

3.2. Preprocessing

3.2.1. Face Detection

The faces are the main priority of this study as they contain all manipulated regions. Therefore, other unnecessary parts of the image can be removed to improve the training efficiency. A face detection approach using facial landmarks was implemented by an open-source toolkit named dlib to detect all of the faces [19]. This library uses a pretrained support vector machine (SVM) to extract all faces in an image trained on Histogram of Oriented Gradients (HOG) features.

3.2.2. Data Augmentation

Three standard data augmentation methods, including color jitter, random rotation, and random perspective, were implemented to generate more training images and improve the framework's performance and robustness. Sample images created by the data augmentation process are displayed in Figure 3. Finally, all detected face images are rescaled to 224 × 224.

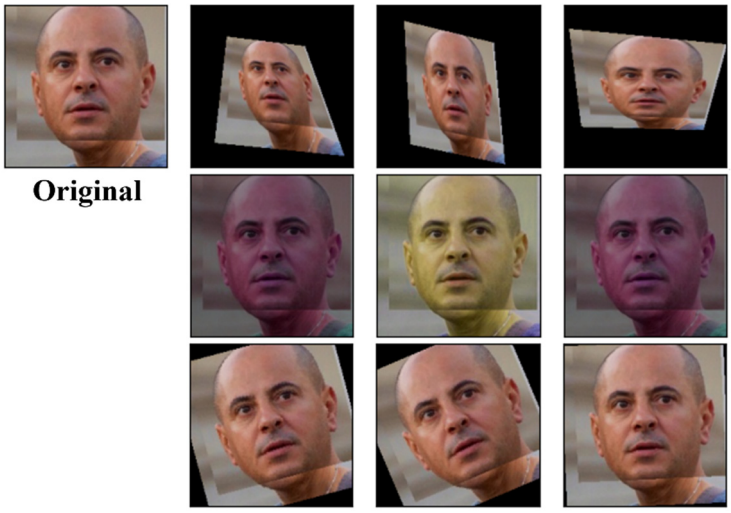

Figure 3. Image generated by various augmentation methods, such as random perspective, color jitter, and rotation.

3.3. Manipulated Face Identification Model

As illustrated in Figure 4, the proposed manipulated face identification framework was motivated by the RegNet network architecture [15], which includes three main modules: stem, layer, and head. The input of the model was 224 × 224 RGB images, whereas the output was a binary classification result of two classes for a genuine or manipulated image.

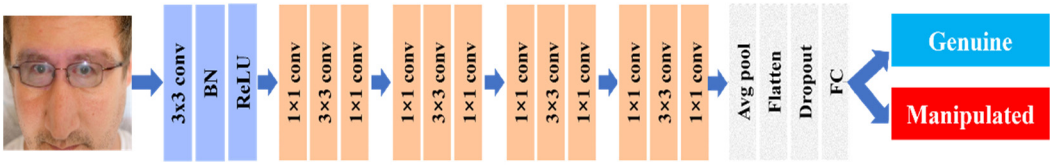

Figure 4. Architecture of the face manipulation identification model (motivated by the RegNet structure). Note: BN stands for batch normalization; conv is a convolutional layer; avg pool is the average pooling; FC indicates a fully connected layer.

The stem module is a crucial part of the generic ResNet/ResNeXt architecture, which includes a 3 × 3 convolutional layer (Conv) and a batch normalization (BN) to shrink the spatial size of the input quickly. Next, the layer module consists of a set of residual blocks. Each residual block includes a set of 1 × 1, 3 × 3, and 1 × 1 convolutional layers with BN and a ReLU activation function following each convolutional layer. Finally, the extracted features go through the head module containing the pooling and fully connected layers. The pooling layer is normally applied to shrink the spatial size of the extracted feature maps. This study used an adaptive average pooling with an output size of 1. Finally, a fully connected layer (FC) was implemented to predict one of the two classes, genuine or manipulated, based on the output features.

A cost function has the main objective of minimizing the training loss during the training stage. Cross-entropy loss, which is a standard cost function for classification problem, is selected as the primary cost function in this study [20]. The problem under consideration in this study is a binary classification problem with two classes—manipulated and genuine. The binary cross-entropy can be defined as follows.

$$L = -\sum_{i=1}^{2} g_i \log(p_i) \quad (1)$$

where g_i is the ground truth value taking the value of 0 for genuine class and 1 for manipulated class. p_i is the linear probability for the ith class. To optimize this loss value, Adaptive Moment Estimation (Adam) optimizer [21] was implemented. Adam is an adaptive learning rate optimization approach that considers both scaling and momentum. The method is effective and requires less memory when coping with complex topics involving a lot of data or parameters. The Adam weight update rule w_t with respect to the parameter w at time step t is performed as follows.

$$w_t = w_{t-1} - \eta \frac{\widehat{m_t}}{\sqrt{\hat{v}_t} + \epsilon} \tag{2}$$

where $\widehat{m_t}$ and $\hat{v}_t$ are bias-corrected first moment (the mean) and the second moment (the uncentered variance) of the gradients respectively. We follow the default hyperparameter values proposed in the original paper [21] and not using parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$. Finally, the step size η is set to 0.001 to optimize the binary cross-entropy loss value.

4. Experimental Results

The experimental results section represents various test scenarios carried out to verify the proposed model's performance. All experiments were performed on an Ubuntu 18.04 NVIDIA deep learning workstation with an Intel®Core i7-5930K processor, 128GB of DDR4 RAM, and two 12GB Titan X GPUs. The face manipulation identification framework was programmed and trained using PyTorch, an open-source and well-known deep learning library. In order to ensure fairness in the following investigations, the transfer learning technique was implemented for all the models using the corresponding pre-trained models on the ImageNet dataset [22]. The batch size during the training process of all models was fixed to 64. The models were trained for 40 epochs.

The first experiment is implemented in Section 4.1 to test the sensitivity of some crucial hyperparameters. Section 4.2 show the proposed model performance on the proposed dataset using a cross-validation setting. After that, the model performance compared to previous deep learning models using the proposed dataset was also de-scribed. Finally, we show a visualization of the detected manipulated regions that helped the model in predicting the class in Section 4.3.

4.1. Hyperparameter Analysis

The hyperparameters are essential factors enabling high classification performance and fast convergence speed. Table 1 demonstrates how different sets of loss functions (cross-entropy and focal loss), optimizers (SGD and Adam), and learning rates (0.01 and 0.001) affect the RegNet model performance. The variations using the focal loss function and the Adam optimizer showed slightly higher classification accuracy than the cross-entropy loss and the SGD optimizer. Therefore, the loss function and the optimizer influenced the model performance. It can be concluded that the RegNet model showed the highest performance of 89.4% when focal loss, Adam optimizer, and a learning rate of 0.001 were applied. As a result, we used them as the default hyperparameter settings for the following experiments.

4.2. Face Manipulation Classification Performance Analysis

In this experiment, 5-fold cross-validation was implemented on the proposed database in order to eliminate the overfitting issue and boost the classifier robustness. Table 2 shows the number of genuine and manipulated images for training/validation for each fold and the corresponding validation accuracy of each fold. Overall, the proposed model achieved stable performance for five folds with acceptable oscillations of 3.8% and an average accuracy of 87.95%. Fold 5 showed the highest face manipulation identification accuracy of 89.71%.

Table 1. Model performance comparison using different combinations of hyperparameters, including loss function, optimizer, and learning rate.

Loss	Optimizer	Learning Rate	Accuracy
Cross-entropy	SGD	0.01	83.3
		0.001	84.1
	Adam	0.01	84.2
		0.001	86.5
Focal	SGD	0.01	85.4
		0.001	86.1
	Adam	0.01	87.3
		0.001	89.4

Table 2. Numbers of genuine and manipulated face images for training and validation tasks for each fold and their corresponding validation accuracy using a 5-fold cross-validation.

Fold	Genuine Images		Manipulated Images		Validation Accuracy (%)
	Training	Validation	Training	Validation	
#1	4522	1128	3348	840	85.92
#2	4512	1138	3358	830	87.85
#3	4511	1139	3359	829	87.49
#4	4532	1118	3339	849	88.82
#5	4523	1127	3348	840	89.71
Average					87.95

Figure 5 represents the training and validation accuracy and loss for Fold 5, which achieved the highest performance among the five folds. The proposed model's training and validation accuracy increase sharply to above 80% and 75%, respectively, after epoch 5 and stop at 97% for training and 89% for validation at the end of the training process. The computed error values decline gradually and stop at 0.4 for the training and 0.2 for validation at epoch 40.

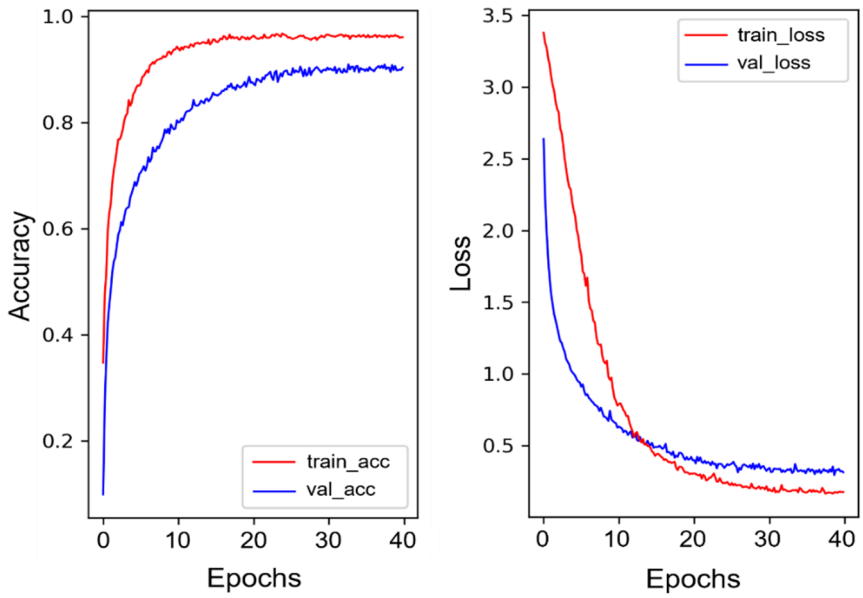

Figure 5. Training and validation accuracy/loss performance of the fine-tuned RegNet model on Fold 5.

The main goal of the following investigation is to demonstrate the superiority of the suggested model over existing models, which includes EficientNet [13], ResNet-50 [23], and

MobileNetV3 [14], on the collected manipulated face dataset. Other required hyperparameters for each model were set strictly following the descriptions from the original papers.

Table 3 describes the experimental results of all models in terms of classification accuracy. Among the models, the proposed fine-tuned RegNet-based model obtained the highest classification accuracy of 89.71%, which was 6.47% better than the MobileNetV3 model. On the other hand, the NAS-based EfficientNet and ResNet50 achieved slightly slower performances than the proposed model at 88.05% and 86.85%, respectively. Both RegNet and EfficientNet contain a similar number of trainable hyperparameters. RegNet required 21 ms to process a single image, whereas the inference time was slightly faster for EfficientNet at 19 ms.

Table 3. Performance of the proposed manipulated face image identification model compared to the other models. The best performance for each metric is highlighted in bold.

Model	Accuracy (%)	Parameters	Inference Time (ms)
EfficientNet [13]	88.05	12M	19
ResNet-50 [23]	86.85	23.5M	28
MobileNetV3 [14]	83.24	4.2M	13
RegNet (Our)	89.71	11.2M	21

4.3. Qualitative Evaluation and Manipulated Region Analysis

Figure 6 quantitatively describes the outputs of the proposed model for four genuine face samples and four manipulated face samples. The fine-tuned RegNet model correctly distinguished between genuine and manipulated face images with high classification confidence.

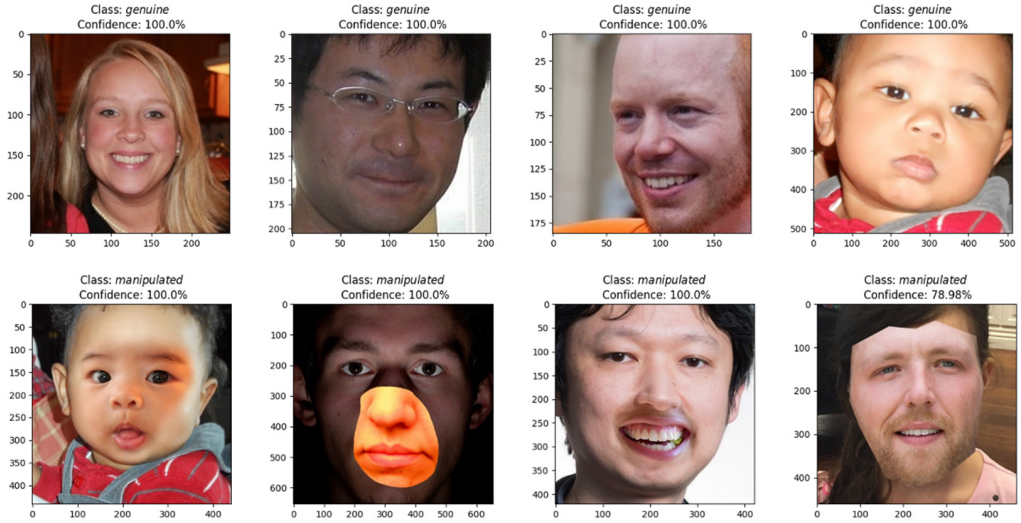

Figure 6. The model predictions for four genuine face images and four manipulated face images with the corresponding prediction confidence.

In order to further interpret how the model made a specific prediction, a class activation map (CAM) technique was implemented. CAM reveals how the model correctly utilizes the trained weights to identify manipulated input images. It projects weights for each class from the previous layers back to feature maps of the final FC layer to highlight crucial manipulated regions. Figure 7 reveals that the manipulated areas from the original images were accurately projected in the corresponding CAM results. As a result, it can be concluded that the model identified a manipulated image based on analyzing related regions.

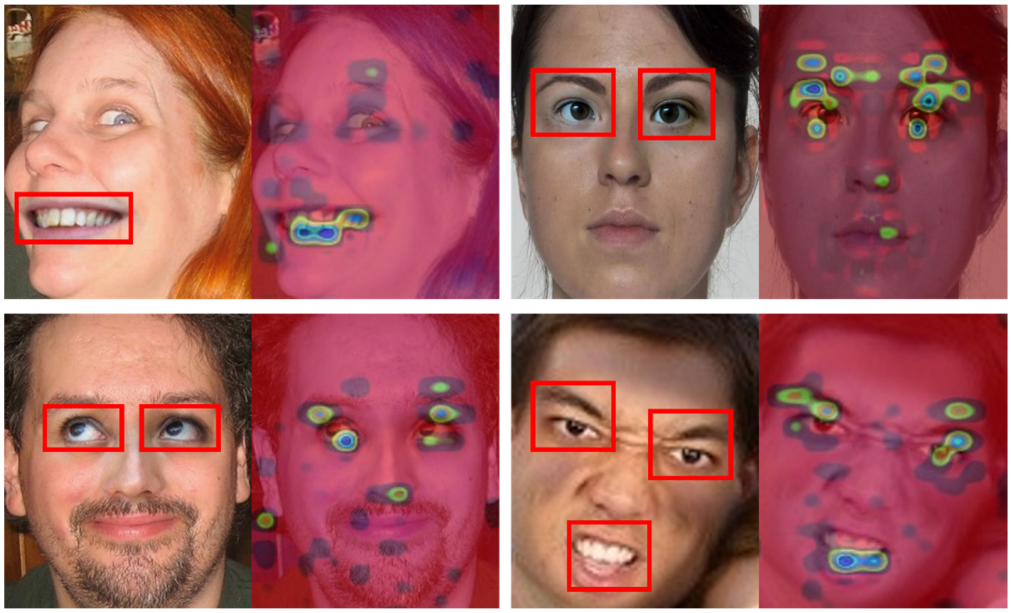

Figure 7. Class activation map (CAM) interpretation of the model predictions.

5. Discussion

The experimental results revealed that the RegNet-based manipulated face identification framework and the proposed dataset are essential to promote developments in face forgery detection. First, this study introduced a manually collected and validated face manipulation dataset that mainly contains high-quality images that can be used for future studies as a benchmark dataset. Second, the framework was based on the state-of-the-art RegNet model, which is straightforward to deploy in popular deep learning libraries, such as Tensorflow and PyTorch. Third, the proposed model achieved the highest manipulation classification performance of 89.7% compared to other state-of-the-art deep learning models (EfficientNet [13], ResNet-50 [23], and MobileNetV3 [14]), which was appropriate for applications that focus more on performance than computational complexity. Finally, we proved that the model correctly recognized crucial manipulated regions by projecting weights for each class from the previous layers back to the feature maps of the final FC layer.

6. Conclusions

This study proposed a RegNet-based face manipulation identification framework. First, a manipulated face dataset containing a considerable number of high-quality images that were manually validated was proposed. The fine-tuned RegNet model overcame the challenges of conventional approaches by automatically extracting representative abstract features. In addition, several techniques were implemented to improve the performance of the proposed model, such as preprocessing and transfer learning. The proposed model set a high face manipulation classification performance with the highest accuracy of 89.7%.

The following weaknesses can be solved to boost the model's performance further. First, even though the framework showed high performance on RGB images, it is essential to study manipulated images using other color channels, such as infrared, in order to explore potential practical inputs for face forgery identification. Second, more preprocessing methods, such as rescaling and whitening transformation, can be applied during the collection of the manipulated face dataset to make it more challenging. Finally, the current framework was unable to reveal the exact locations of the manipulated regions, which can be addressed by training an object detector.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available on request due to privacy restrictions.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Dang, L.M.; Hassan, S.I.; Im, S.; Lee, J.; Lee, S.; Moon, H. Deep learning based computer generated face identification using convolutional neural network. *Appl. Sci.* **2018**, *8*, 2610. [CrossRef]
2. Minh, D.; Wang, H.X.; Li, Y.F.; Nguyen, T.N. Explainable artificial intelligence: A comprehensive review. *Artif. Intell. Rev.* **2021**, *55*, 1–66. [CrossRef]
3. Jaiswal, A.K.; Srivastava, R. A technique for image splicing detection using hybrid feature set. *Multimed. Tools Appl.* **2020**, *79*, 11837–11860. [CrossRef]
4. Chen, J.; Liao, X.; Wang, W.; Qian, Z.; Qin, Z.; Wang, Y. SNIS: A Signal Noise Separation-based Network for Post-processed Image Forgery Detection. *IEEE Trans. Circuits Syst. Video Technol.* **2022**. [CrossRef]
5. Parveen, A.; Khan, Z.H.; Ahmad, S.N. Block-based copy–move image forgery detection using DCT. *Iran J. Comput. Sci.* **2019**, *2*, 89–99. [CrossRef]
6. Nguyen, T.N.; Lee, J.; Dinh-Tien, L.; Minh Dang, L. Deep learned one-iteration nonlinear solver for solid mechanics. *Int. J. Numer. Methods Eng.* **2022**, *123*, 1841–1860. [CrossRef]
7. Saber, A.H.; Khan, M.A.; Mejbel, B.G. A survey on image forgery detection using different forensic approaches. *Adv. Sci. Technol. Eng. Syst. J.* **2020**, *5*, 361–370. [CrossRef]
8. Yao, H.; Cao, F.; Tang, Z.; Wang, J.; Qiao, T. Expose noise level inconsistency incorporating the inhomogeneity scoring strategy. *Multimed. Tools Appl.* **2018**, *77*, 18139–18161. [CrossRef]
9. Zhang, Y.; Song, W.; Wu, F.; Han, H.; Zhang, L. Revealing the traces of nonaligned double JPEG compression in digital images. *Optik* **2020**, *204*, 164196. [CrossRef]
10. Mayer, O.; Bayar, B.; Stamm, M.C. Learning unified deep-features for multiple forensic tasks. In Proceedings of the 6th ACM Workshop on Information Hiding and Multimedia Security, Innsbruck, Austria, 20–22 June 2018; pp. 79–84.
11. Zhou, P.; Han, X.; Morariu, V.I.; Davis, L.S. Two-stream neural networks for tampered face detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu, HI, USA, 21–26 July 2017.
12. Elsken, T.; Metzen, J.H.; Hutter, F. Neural architecture search: A survey. *J. Mach. Learn. Res.* **2019**, *20*, 1997–2017.
13. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019.
14. Qian, S.; Ning, C.; Hu, Y. MobileNetV3 for image classification. In Proceedings of the 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Nanchang, China, 26–28 March 2021.
15. Radosavovic, I.; Kosaraju, R.P.; Girshick, R.; He, K.; Dollár, P. Designing network design spaces. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020.
16. De Carvalho, T.J.; Riess, C.; Angelopoulou, E.; Pedrini, H.; de Rezende Rocha, A. Exposing digital image forgeries by illumination color classification. *IEEE Trans. Inf. Secur.* **2013**, *8*, 1182–1194. [CrossRef]
17. Dong, J.; Wang, W.; Tan, T. Casia image tampering detection evaluation database. In Proceedings of the 2013 IEEE China Summit and International Conference on Signal and Information Processing, Beijing, China, 6–10 July 2013.
18. Ng, T.-T.; Chang, S.-F. Columbia Image Splicing Detection Evaluation Dataset. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Singapore, 24–27 October 2004; Volume 2.
19. Nguyen, L.Q.; Li, Y.; Wang, H.; Dang, L.M.; Song, H.-K.; Moon, H. Facial Landmark Detection with Learnable Connectivity Graph Convolutional Network. *IEEE Access* **2022**, *10*, 94354–94362. [CrossRef]
20. Gordon-Rodriguez, E.; Loaiza-Ganem, G.; Pleiss, G.; Cunningham, J.P. Uses and abuses of the cross-entropy loss: Case studies in modern deep learning. *PMLR* **2020**, *137*, 1–10.
21. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
22. Recht, B.; Roelofs, R.; Schmidt, L.; Shankar, V. Do imagenet classifiers generalize to imagenet? In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019.
23. 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, Las Vegas, NV, USA, 27–30 June 2016.