computers

MDPI

Article

Feature Focus: Towards Explainable and Transparent Deep Face Morphing Attack Detectors †

Clemens Seibold [1,*], Anna Hilsmann [1] and Peter Eisert [1,2]

[1] Department of Vision and Imaging Technologies, Fraunhofer Heinrich-Hertz-Institute, 10587 Berlin, Germany; anna.hilsmann@hhi.fraunhofer.de (A.H.); peter.eisert@hhi.fraunhofer.de (P.E.)
[2] Department of Visual Computing, Humboldt Universität zu Berlin, 10117 Berlin, Germany
* Correspondence: clemens.seibold@hhi.fraunhofer.de
† This paper is an extended version of our paper published in WACV xAI4Biometrics Workshop 2021.

Abstract: Detecting morphed face images has become an important task to maintain the trust in automated verification systems based on facial images, e.g., at automated border control gates. Deep Neural Network (DNN)-based detectors have shown remarkable results, but without further investigations their decision-making process is not transparent. In contrast to approaches based on hand-crafted features, DNNs have to be analyzed in complex experiments to know which characteristics or structures are generally used to distinguish between morphed and genuine face images or considered for an individual morphed face image. In this paper, we present Feature Focus, a new transparent face morphing detector based on a modified VGG-A architecture and an additional feature shaping loss function, as well as Focused Layer-wise Relevance Propagation (FLRP), an extension of LRP. FLRP in combination with the Feature Focus detector forms a reliable and accurate explainability component. We study the advantages of the new detector compared to other DNN-based approaches and evaluate LRP and FLRP regarding their suitability for highlighting traces of image manipulation from face morphing. To this end, we use partial morphs which contain morphing artifacts in predefined areas only and analyze how much of the overall relevance each method assigns to these areas.

Keywords: face morphing attacks; DNN explainability; face image forgery detection

Citation: Seibold, C.; Hilsmann, A.; Eisert, P. Feature Focus: Towards Explainable and Transparent Deep Face Morphing Attack Detectors. *Computers* **2021**, *10*, 117. https://doi.org/10.3390/computers10090117

Academic Editor: Paolo Bellavista

Received: 13 August 2021
Accepted: 14 September 2021
Published: 18 September 2021

1. Introduction

Face morphing is a simple way to obtain a digitally generated face image that resembles two different subjects. Such images can fool biometric verification systems [1] as well as trained experts [2] into confirming that any of the two different subjects is the person shown on the image. Such an image, if used in an identification document, would break the unique link between the intended owner and the document. The ownership of this document could additionally be claimed by a person different from the one it was issued for. This person would be able to travel as or use benefits bound to the legitimate identification document owner. Using such a method to adopt a different identity constitutes a face morphing attack. Such an attack can be performed without expert knowledge. In order to generate a morphed face image, one image of each of the subjects is needed. Using a standard image manipulation tool or one of the freely available morphing tools, the images only need to be aligned and blended to generate an image that looks similar to both subjects. While the theoretical possibility of this attack has already been shown in 2004 [3], Ferrara et al. demonstrated the feasibility of this attack in 2014 [1], also releasing a manual on how to create morphed face images using a common image manipulation program. Since several countries allow their citizens to provide an image for passports and national ID-cards, e.g., USA, France or Germany, a face morphing attack requires no further attacks on IT-systems. The consequences of the feasibility of such attacks to the integrity of automated and manual identity verification checks, e.g., at country borders,

have motivated several research groups to develop detectors and analyze the problem of morphed face images in detail [4,5]. The importance of solving this problem has also been stressed by governmental agencies such as Frontex (European Border and Coast Guard Agency) [6], the European Commission [7] or by the National Institute of Standards and Technology of the USA [8]. In addition to facial recognition systems at border crossing points, face morphing attacks are also a threat to other applications in the big and growing market of facial recognition systems [9,10].

Within recent years several face morphing attack detectors based on different concepts with different requirements have been proposed to tackle this problem. Approaches based on learned features like deep neural networks can achieve a very high accuracy in this task [11], but are not as transparent as approaches based on handcrafted features that describe physical properties such as sensor noise [12,13]. Explainability approaches like LRP [14] help to get a better understanding of the decision-making process of DNNs and to determine which structures and regions are important for the detection. However, applying LRP to DNN-based face morphing detectors involves new challenges, making LRP's results difficult to interpret and need further investigations [15]. For example, when asking LRP to highlight forged regions, it highlights in some cases/for some detectors the genuine regions instead of the morphed ones.

In this paper, we propose Feature Focus, a new transparent DNN-based face morphing attack detector, as well as Focused LRP (FLRP) [16], an extension of LRP [17], to tackle the previously mentioned problems of LRP. Feature Focus is based on a modified VGG-A architecture with a reduced number of neurons in the last convolutional layer and an additional loss function for the output neurons of the DNN's feature extraction component. This loss function shapes the neurons to focus on a class of interest, e.g., having a strong activation if a morphed face image is presented and otherwise no activation. This detector in combination with FLRP provides a reliable explainability component that highlights traces of forgery in morphed face images with high accuracy. We compare the new detector with a naïvely-trained network [11] and a network that was pre-trained on images with morphing artifacts present only in up to four pre-defined areas [15]. The latter has been shown to be more robust against attacks on the decision making process of DNNs such as adversarial attacks, but also against image quality improvement methods applied on morphed face images, which can be used as counter forensics against some face mophing attack detectors [18]. We analyze the learned features and characteristics using FLRP in combination with partially induced artifacts and based on the discrimination power of selected neurons in the feature output of the DNN. Furthermore, we perform a quantitative comparison between FLRP and LRP for three different DNNs for face morphing attack detection. FLRP was developed to highlight traces of forgery more accurately than LRP. For its evaluation, we thus use partial morphs which contain morphing artifacts only in predefined regions, and analyze whether the relevance is only assigned to the forged regions. Figure 1 shows an example of relevance distributions calculated by FLRP and LRP and differently trained DNNs for a partially morphed face image which contains only artifacts in the areas of the left eye and mouth.

The contributions of this paper are:

- Feature Focus: A more accurate and transparent face morphing attack detector based on a new loss function and modified architecture
- Quantitative analysis of FLRP and comparison with LRP using morphs that contain artifacts only in known predefined areas (partial morphs)
- Analysis of the features' discrimination power learned by DNNs for face morphing attack detection and its relation to interpretability via FLRP and LRP
- Reliable and accurate explainability component for DNN-based face morphing attack detectors based on FLRP

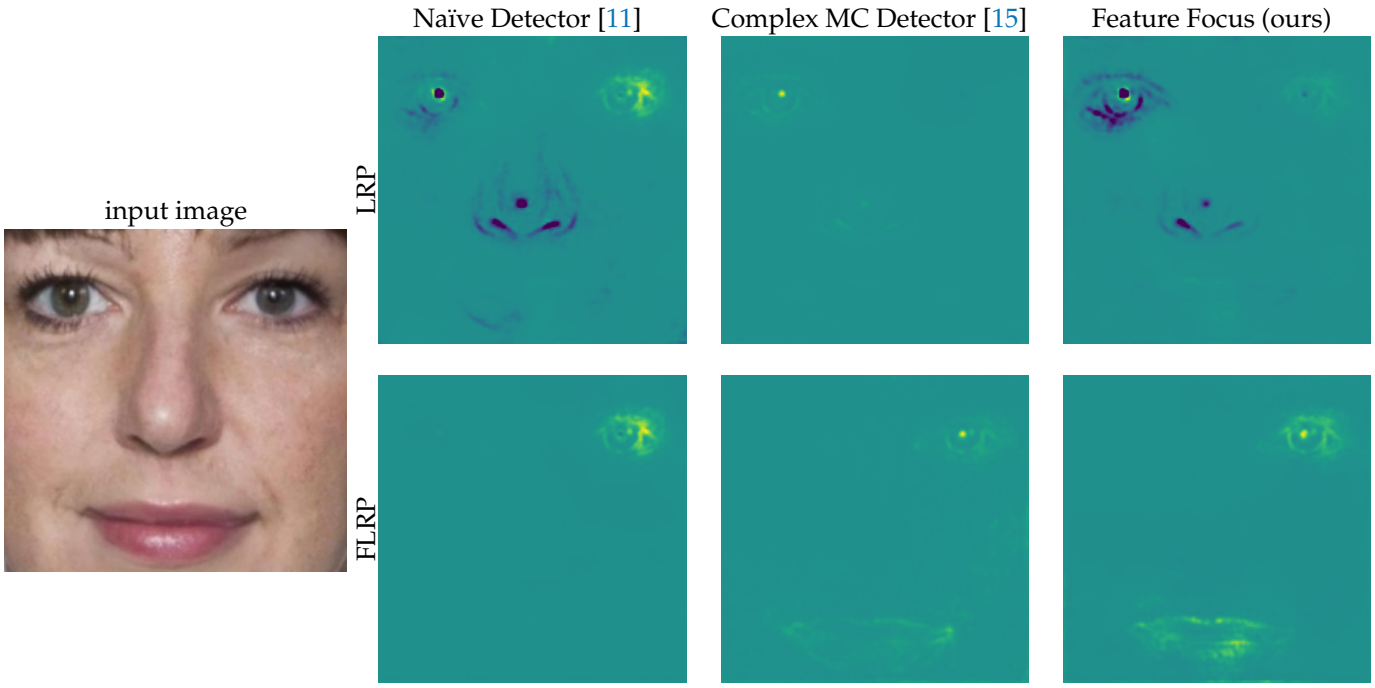


Figure 1. FLRP and LRP relevance distributions for a partial morph (the left eye and the mouth have been morphed) on differently trained DNNs for face morphing attack detection. A yellow colored region contributes to the decision "morphed face image", a blue color denies it and a olive green colored region contains no information about the decision. All except LRP for the Complex MC Detector highlight the right eye as a region that contributes to the class morph. The Complex MC Detector highlights the non-forged left eye. This relevance distribution of the Complex MC Detector is caused by a complex structure that compares different regions of the face, for further details see [15]. The expected relevance distribution, which assigns relevance to all morphed parts is only achieved by FLRP for the Complex MC Detector and our proposed detector.

This paper is structured as follows. In the next section, we provide an overview on existing face morphing attack detectors and interpretability methods for DNNs. Subsequently, the LRP extension FLRP [16] is described in detail in Section 3. In Section 4, we introduce our new training method for DNN-based face morphing attack detectors without a reference image. The details on experimental data and training of the three different detectors are summarized in Section 5. The metrics that are used to evaluate the differently trained networks and LRP and FLRP are described in Section 6, followed by the results in Section 7. Finally, we discuss our results and finish with a conclusion on the advantages of FLRP and our new training method.

2. Related Work

A common way to classify face morphing detectors is to divide them into blind and differential face morphing attack detectors. To decide if the presented image is a genuine or a morphed face image, differential face morphing attack detectors need a reference image, which is used for comparison [19,20] or to demorph the image [21,22], or a 3-D model [23] of the subject that claims to be shown on the image. On the other hand, blind detectors do not depend on additional data to make this decision. Comprehensive overviews on face morphing detectors and their characteristics can be found in [4,5]. In this paper, we focus on blind detectors. Blind detectors usually consist of a feature extraction step followed by a classifier. The features can be handcrafted and describe effects such as the statistical properties of JPG coefficients after double compression [24] or the image impairment/change of spatial frequency distribution that arises from the warping and blending steps during the generation of a morphed face image [25,26]. Furthermore, the noise pattern characteristics of camera sensors [12] can be analyzed for detecting manipulated images. The features can also be derived from image statistics [27] or learned

specifically for the problem of detecting morphed face images by using a Deep Neural Network (DNN) [11]. Such learned features are very powerful, but it is difficult to analyze what they describe or represent. Other blind detection methods are based on compositions of different detectors and fuse their predictions to obtain a more robust and accurate face morphing attack detector [28,29].

With the increasing use of DNNs for computer vision tasks, researchers developed different approaches that identify which regions are important for image classification tasks. One simple method to evaluate whether a part in an image is relevant, is occlusion sensitivity. A part is occluded, e.g., by a random color for each pixel or a pre-defined color, and the change of the DNN's decision with respect to the occluded region is analyzed. It assumes that a DNN's decision will change if an important region is occluded. The authors of [30] use this method to reveal which parts of an image are important and study its correlation with the positions of strong activations in the feature map with the strongest activation. A more sophisticated approach is presented in [31]. The authors propose a new DNN architecture that identifies prototypical parts that are important for the decision-making in the analyzed image and finds similar prototypical parts in a set of reference images. This approach only requires training on image-level labeled data and learns without further supervision to identify the prototypical parts. Another approach to produce visual explanations is Gradient-weighted Class Activation Mapping (Grad-CAM) [32]. It assigns a relevance score to each feature map in the last convolutional layer based on the gradients of a class with respect to these feature maps and calculates a weighted sum of the feature maps. The resulting map with width and height of the last convolutional layer is upsampled to be mapped into the input image. Thus, Grad-CAM can only mark coarse regions as relevant for the DNN's decision.

In contrast to Grad-CAM, Layer-wise Relevance Propagation (LRP) [17] is not based on gradients of the class of interest, but on neurons that result in its activation. Furthermore, it considers the whole structure of the DNN, the classification part as well as the activations and weights of the convolutional layers. By that, it can create finer heatmaps and assigns a relevance score to every pixel, describing its influence on the class of interest which can either contribute to or inhibit activation. These approaches can in generally help to understand which pixels of an image are important for the decision making process of a machine-learning model and reveal undesired properties of the model. For example, Lapuschkin et al. [14] showed that a famous model for a well-known image recognition challenge looked for the signature of a photographer to detect if an image depicts a horse. This strategy achieves quite a high degree of accuracy, since all images with this signature within this dataset are images of horses. In the case of such simple examples, in which the presence of a structure leads directly to an activation of a class, LRP is an excellent method to highlight the area responsible for the activation and uncover such problems. In the case of DNNs for face morphing attack detection, the DNNs are always confronted with face images and traces of the morphing steps can be very subtle and appear in different regions of the face. In addition, differences between image regions can be an indication for morphed face images, but cannot be explained by LRP without further investigations [15]. Thus, an artifact induced by the face morphing process might not be deemed relevant by this method. Furthermore, LRP focuses on the whole model and information about traces of forgery detected by some features might get lost. Contrarily, the recently proposed Focused Layer-wise Relevance Propagation (FLRP) focuses only on the learned features, ignoring the information from the fully-connected layers, and has experimentally been found to highlight traces of forgery better than LRP [16]. In this paper, we show that FLRP highlights traces of forgery with high accuracy and without undesired relevance assignments to non-forged parts of the image. This makes FLRP a perfect tool to support non-technical experts (e.g., border guards) in understanding and arguing why an image is a forgery.

3. Focused Layer-Wise Relevance Propagation

The interpretability method LRP [17] assigns relevance to each pixel of the input image. This leads to a heatmap that indicates which image regions are important for the network's decision. The relevance is assigned such that regions that lead to a high activation of the class of interest receive a positive value and regions that inhibit its activation a negative value. The mathematical background of LRP is based on a "deep Taylor decomposition" of the neural network for a class of interest [14]. In a first step, it assigns a starting relevance value to this class. Next, this relevance is propagated layer-by-layer into the input images. To this end, different rules exist that define how to map relevance from a neuron to all neurons in the previous layer that are connected to this neuron. These rules are intended to assign relevance to neurons in the previous layer that are responsible for an activation or inhibition of this neuron. If a neuron inhibits an activation, the relevance is negated. LRP is usually used with different rules depending on the type and position of the layer. In our experiments, we use the epsilon-decomposition rule for the fully-connected layers and the $\alpha\beta$-decomposition with $\alpha = 2$ and $\beta = -1$ for all convolutional layers except the first one, which is subject to a flat decomposition. This has been shown to be a good practice for similar structured DNNs [33]. While the ϵ-decomposition treats activating and inhibiting relevance similarly, the $\alpha\beta$-decomposition considers them separately. With $\|\alpha\| > \|\beta\|$, which is a recommended setting, this rule focuses more on activating relevance, leading to more balanced results. The flat-decomposition propagates the relevance of a neuron equally distributed to all neurons in the previous layer that have an influence on this neuron. For a more detailed explanation of these methods, we refer to [33].

In contrast to LRP, FLRP does not investigate the complete decision-making process of DNNs, but focuses on discriminative neurons in the output of its last convolutional layer (feature output), see Figure 2. For the VGG-A architecture with an input size of 224×224, which we use in our experiments, this feature output is a tensor of size $N \times N \times M$, with $N = 7$ and $M = 512$. We use this standard input size and feature output size as proposed by the authors of VGG-A since it has been shown to be a suitable setting for classification tasks and pre-trained models are available [34]. FLRP assigns relevance to neurons of interest in this layer and propagates this relevance into the image using LRP rules. The selected neurons are different neurons for each class of interest. They are selected to have a strong activation if an image of the class of interest is presented and a small or no activation at all otherwise and thus can identify images of the class of interest. Since FLRP starts the relevance assignment at the output of the DNN's feature extractor, a spatial relation between these neurons and coarse regions in the input image is already given by the network's architecture. By applying relevance propagation from these neurons, the regions can be refined to highlight exactly the structures that led to their activations. FLRP restricts these neurons to N^2. Interpreting the $N \times N \times M$ feature output as a $N \times N$ pixel image with M channels, FLRP assigns a starting relevance to one channel for each pixel (neuron). For the starting relevance we use the activation of the neuron for the image that should be analyzed. Regarding DNNs for face morphing attack detection, we are interested in detecting and highlighting traces of forgery. Thus, the class of interest is the class of morphed face images, and the neurons of interest have a strong activation if a morphed face image is presented to the DNN. In the follow we explain the single steps to determine the neurons of interest in our experiments:

In a first step, we calculate the output of the feature extraction component of the DNN for each image in the training data. This output consists of a $N \times N \times M$ tensor for each image. It can be interpreted as an image with M channels and a size of $N \times N$ pixels. For each pixel, we select the channel that has a larger value when the input is a morphed face image and is best suited to distinguish between genuine and morphed face images. To this end, for each neuron we calculate a threshold such that the number of morphed face images that lead to activation values above this threshold is equal to the number of genuine face images that lead to activation values below that threshold. Based on these thresholds, we select the channel that is most suitable to separate between genuine and morphed face

images for each pixel in the M-channel "image". This yields N^2 neurons which we will use to initialize our relevance propagation. In contrast to common LRP or sensitivity maps, which start from a single neuron and changing the starting value only scales the resulting relevance values, FLRP needs to assign suitable initial values for these neurons. To do so, we pass the image that should be inspected through the DNN and use the resulting activation values as starting relevance. The idea behind this initialization method is to assign starting relevance mainly to neurons that did detect face morphing related artifacts and thus have large activation values. Starting with this assignment of relevance in the last layer of the feature extractor, we use the $\alpha\beta$-rule from LRP with $\alpha = 2$ and $\beta = -1$ for all but the first convolutional layer to propagate the relevance into the input image. For the first convolutional layer, we use flat decomposition.

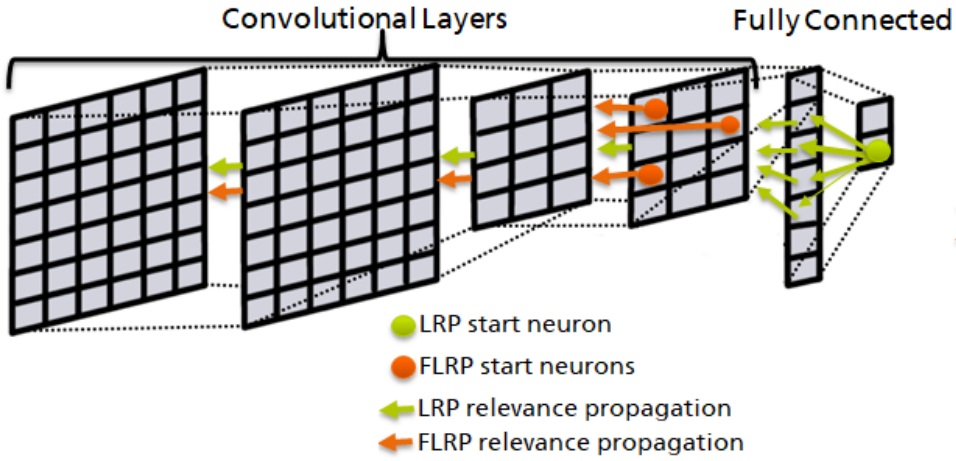


Figure 2. Concept of LRP and FLRP. While LRP starts the relevance propagation from a neuron that describes the likelihood of a class, FLRP starts at selected discriminative neurons in the feature output of a DNN.

4. Feature Shaping Training

For FLRP, we have to select one out of M channels/neurons for each position in the feature output. There might be different neurons that indicate different morphing artifacts and there is no guarantee that there will be discriminative neurons in all cases. To overcome the problem of selecting suitable neurons and to already shape the neurons during training, we modified the VGG-A architecture and added another loss function after the last pooling layer. The goal of this modification and loss function is to have only two feature maps with opposite behaviors as output of the feature extraction component of the DNN. One feature map is morph-aware, meaning that it has a strong activation if a morphed face image is fed to the DNN and no activation otherwise. The other feature map behaves exactly the other way around and thus has a strong activation if a genuine face image is presented to the DNN.

The details are described in the following. We reduce the number of channels of the last convolutional layer to two and remove the fully-connected layers. Thus, we can apply to these two feature maps a loss function that considers neurons independently of one another, similar to loss functions for segmentation tasks [35]. After going through a ReLU and maximum pooling-layer, the output of the convolutional part is directly fully connected to two neurons, one for the class morph and the other one for the class genuine image. During training a drop-out layer was added between the maximum pooling-layer and the fully connected layer.

In addition to the negative-log likelihood loss for the neurons that represent the two classes, we add a soft margin loss to train one of the two features maps to have a strong

activation if the presented image is a morphed face image and otherwise no activation, and vice versa the other feature map. This loss function for the feature map can be written as:

$$L(\mathbf{f}, y) = -y \sum_i \ln\left(\frac{1}{1+\exp(-\mathbf{f}[i])}\right) + (1-y)\sum_i \left(-\ln 0.5 + \ln\left(\frac{1}{1+\exp(-\mathbf{f}[i])}\right)\right), \quad (1)$$

with $\mathbf{f}$ representing the neurons of a feature map (after applying a ReLU and a max-pooling layer on the output of the last convolutional layer) and y a variable to select if an activation should be favored or penalized. The loss function (1) shapes the features output such that it has morph-aware and genuine-aware neurons, but it does not shape the final output of the DNN, so we need to combine it with another loss function for this purpose. To this end, we use the negative-log likelihood loss, which is a common loss function for classification tasks. Thus, the final loss function for our optimization can be written as:

$$L(y, \mathbf{x}, \mathbf{f_g}, \mathbf{f_m}) = -\ln(\mathbf{x}[y]) + \tfrac{1}{N^2}(L(\mathbf{f_m}, y) + L(\mathbf{f_g}, (1-y))) \quad (2)$$

with $\mathbf{f_g}$ or $\mathbf{f_m}$ being the feature map that should have a strong activation if the presented image is a genuine or morphed face image, respectively, $\mathbf{x}$ being output of the DNN after applying a softmax layer and y being 1 if the input image is a morphed face image and 0 otherwise. We scaled the loss function on the feature maps by $\frac{1}{N^2}$, since they contain N^2 neurons each. This way, the feature shaping loss and the class loss have similar values and a similar influence on the training.

5. Training of the Detectors and Experimental Data

We acquired face images from different public and in-house datasets for our experiments. In total, we collected about 2000 face images from different subjects including images from the public datasets BU4DFE [36], Chicago Face Database [37], FERET [38], London Face Database [39], PUT [40], scFace [41] and Utrecht [42]. We pre-selected the images such that the inter-eye distance is at least 90 pixels and the subject is neutrally looking into the camera in full front view. We split our data into a training (70% of all genuine face images), a validation (10%) and a test set (20%) and generated the same amount of morphed face images using only images from the same set. The pairs for the morphed face images were selected such that both have the same gender, are from the same dataset and that all subjects are morphed with the same frequency. For the generation of morphed face images, we used two different fully automated face morphing pipelines. One is based on an alignment via field morphing as described in Seibold et al. [11] and the other uses a triangle based alignment [15]. Both approaches differ only in the method used to align the two input images. For the testing and validation sets, we also use the horizontally flipped version of the face images to augment the data.

Before feeding an image into the network, we pre-process the face images to avoid unnecessary variance in the data and to have a standardized input. To this end, we use the method proposed in [11]: First, we estimate facial landmarks and rotate the images so that the eyes are on a horizontal line, then we crop the inner part of the face including eyebrows and mouth, as shown in Figure 3. During training, horizontal image flipping and random shifting of the image of up to two pixels were used for data augmentation.

We trained three different detectors based on the VGG-A architecture [34]. The first detector is directly trained to distinguish between morphed and genuine face images using two output neurons and a negative log likelihood loss function. We refer to this detector as Naïve Detector. The second detector is first pre-trained on partial morphs to predict which regions of the morphed face image are forged using a multi-label training with a multi-label soft margin loss with four neurons, one for each region. After that, these four neurons were replaced by two neurons, each for one class, and the last two layers have been retrained using the same loss as used by the naïve training approach. In [15,18], the authors showed that this training method shapes the DNN to consider more regions of the face for the decision making and leads also to a detection that is more robust against

adversarial attacks and image improvement methods applied to the morphs. We refer to this detector as Complex MC Detector in the following. A dropout with a probability of 0.5 is applied to the Naïve Detector and the Complex MC Detector during training. The third detector is based on our proposed new training method, described in Section 4, which we refer to as Feature Focus. In addition, we trained an Xception [43] network to detect morphed face images. The Xception architecture has been shown to be suitable for the detection of deep fakes [44]. We evaluated the Xception network regarding its suitability for detecting morphed face images and showed its drawbacks regarding interpretability with different examples.

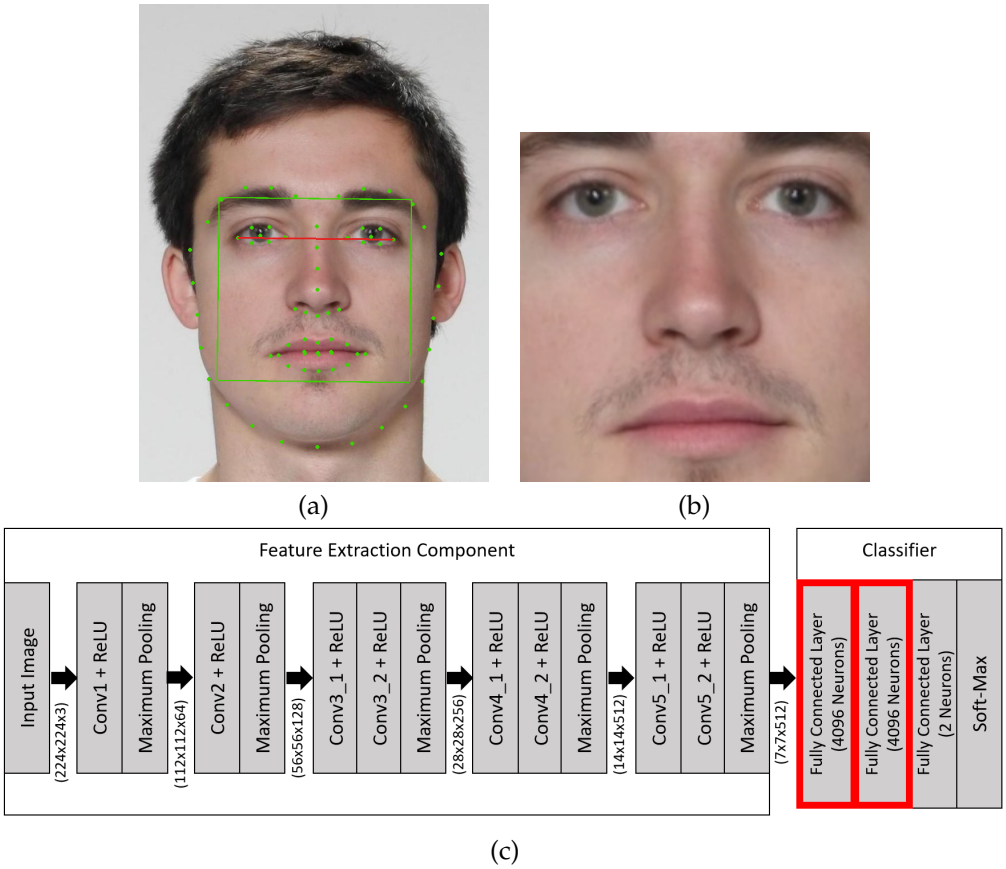


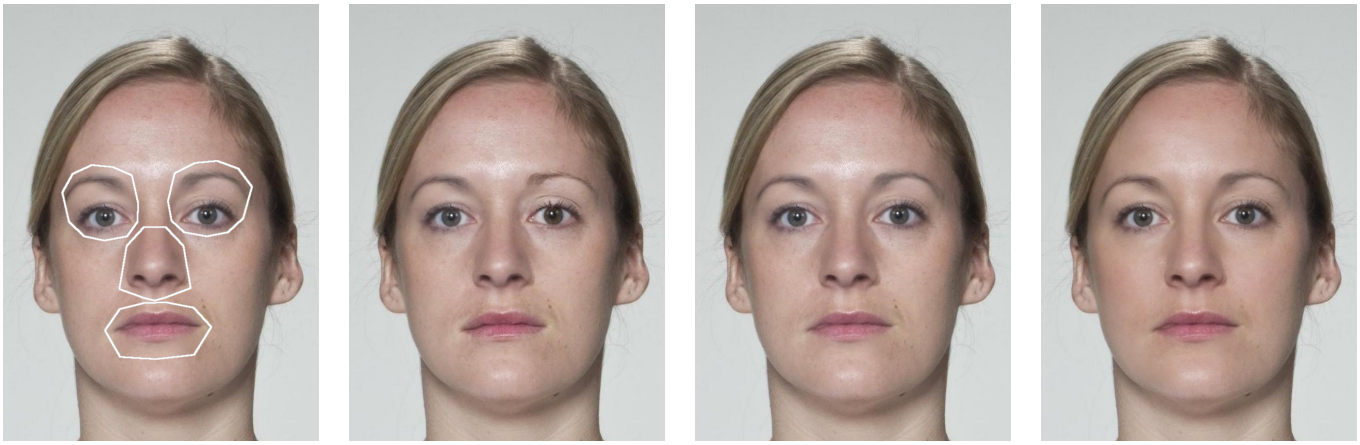
Figure 3. Single steps of the analyzed DNN-based face morphing attack detectors. (**a**) shows the input images with estimated facial landmarks (green points). First, the image is rotated such that the eyes (red line) are parallel to the horizontal image border. Afterwards, the inner part of the face (green box) is defined by a bounding box, which is calculated based on the facial landmarks. (**b**) shows the cropped and aligned region of the face image that is fed into the VGG-A based DNN shown in (**c**). The layers framed in red are removed for our Feature Focus Detector and the output of the feature extraction component is reduced from 512 channels to two channels.

6. Evaluation Methods

6.1. (F)LRP Evaluation with Partial Morphs

For a quantitative evaluation of FLRP and comparison to LRP, we use partial morphs [15] and analyze the distribution of the relevance produced by each method. A partial morph is a morphed face image for which only certain parts, e.g., one eye or the nose, have been morphed and the rest originates from a genuine face image. We use the same regions that were defined in [15] for our partial morphs. These are the right eye, the left eye, the nose and the mouth. In order to generate a partial morph, two input face images are aligned, but only the selected regions are blended and inserted into the input image using Poisson Image Editing [45]. Figure 4 illustrates the regions that might be blended for a partial morph and shows examples of such morphs. Since FLRP and LRP each claim to

be "a method that identifies important pixels by running a backward pass in the neural network" [46], we expect them to assign most of the relevance to the morphed areas of partial morphs when applied to the class morphed face image.



(a) (b) (c) (d)

Figure 4. Partial Morphs. A partial morph is a morphed face image for which only certain parts have been morphed and the rest originates from a genuine face image. We use such morphs in our experiments to evaluate if LRP/FLRP assigns relevance to the forged part only. This figure shows which parts of the face can be forged in case of our partial morphs and examples of partial morphs. (**a**) shows the four regions that can be blended for a partial morph. (**b**) shows a partial morph with a morphed right eye and nose. A partial morph with all four regions blended is shown in (**c**,**d**) shows a complete morphed face image.

For each morphed face image in the test set, we generate all possible combinations of partial morphs with one to four morphed regions. We run LRP and FLRP on these partial face morphs for each of the trained networks. We use the ϵ-decomposition rule for the fully connected layers and the $\alpha\beta$-decomposition with $\alpha = 2$ and $\beta = -1$ for all convolutional layers except the first one, for which we take the flat decomposition rule. After calculating the relevance, we use a cleaning method, as described in the following, to suppress small relevance values over large regions and relevance that occurred due to border padding effects. To this end, we remove all relevance values smaller than 5% of the maximal relevance value and all relevance in a 2 pixel wide border of the image. By removing all relevance values below 5% of the maximal relevance value, we ensure that only the meaningful relevance is considered and small, meaningless relevance values are ignored. This is especially important if only one region of the image is morphed and thus the traces of forgery are only in a small area. The removal of all relevance at the image borders eliminates meaningless relevance that can arise from border padding [47]. Finally, we calculate the relative amount of relevance that falls into the morphed region for each image.

In addition, we compare for some examples the results of FLRP with those of the interpretability methods Grad-CAM [32] and LayerCAM [48] and discuss the limitations of the later two.

6.2. Evaluation of Features' Discrimination Power

Another metric that we use to analyze the characteristics and significance of the learned features is their discrimination power. In [49], the authors show that even a few neurons of the feature output of a VGG-based face morphing detector are sufficient to distinguish between morphed face images and genuine images with high accuracy. However, no further investigations into the spatial distribution of the features or if they have a strong activation for morphed face images or genuine images have been performed. We analyze these features in more detail to evaluate the differently trained detectors and reveal correlations between the discrimination power of neurons and the LRP-/FLRP-based

relevance distributions of the partial morphs. For this analysis, we study the ability to distinguish between genuine images and morphed face images of single neurons in the feature output of the DNNs. For the Naïve and Multiclass MC Detector, we analyze the neurons that are selected by FLRP for the class morph and for the class genuine face image. For the Feature Focus Detector, we can directly use the two output feature maps. The ability to distinguish based on the single neurons is assessed based on the equal error rate on the testing set. Further analyses show that the discrimination power of a neuron is also correlated to the spatial distribution of the relevance calculated by LRP and FLRP.

7. Results

7.1. Accuracy

Table 1 shows the accuracy of the detectors using the ISO metrics for the evaluation of presentation attack detection: Attack Presentation Classification Error Rate (APCER) and Bona-fide Presentation Classification Error Rate (BPCER) [50]. These are also commonly used for face morphing attack detectors. The APCER is the relative number of undetected attacks and the BPCER the relative number of genuine face images that have been predicted as morphed face images. In addition, we provide the Equal-Error-Rate (EER). Our proposed method performs best for all three metrics, while the other two VGG-A-based detectors show a similar performance, with the Complex MC Detector being slightly worse. The Xception network performs similarly to the Naïve Detector and the Complex MC Detector, but is outperformed by our Feature Focus Detector.

Table 1. Accuracy of the different detectors in terms of BPCER, APCER and EER. The Naïve Detector, the Complex MC Detector and the Xception network perform similarly in terms of EER and differ mostly in terms of APCER and BPCER, which can be adjusted by shifting the decision threshold. Our proposed Feature Focus Detector outperforms all other detectors.

Training Type	BPCER	APCER	EER
Naïve [11]	0.8%	2.2%	1.4%
Complex MC [15]	1.8%	1.8%	1.8%
Xception [43]	1.7%	1.2%	1.5%
Feature Focus (ours)	0.5%	1.2%	0.6%

7.2. Discrimination Power of Single Features

Figure 5 shows the discrimination power of selected neurons in terms of equal error rate as described in Section 6.2. In these maps, the entry at position x, y represents one neuron at the position x, y in a feature map. A yellow color indicates a low equal error rate and thus a high discrimination power, a blue color a high equal error rate, and green a rate in between. Note that a different scale was used for the Feature Focus Detector, since its equal error rates are significantly smaller than for the other two detectors. We can observe that the Naïve Detector has neurons that can distinguish between morphed and genuine eyes very well and, albeit with less accuracy, detect forged noses too. The Complex MC Detector has a more balanced distribution, but a stronger discrimination power for genuine-aware neurons than for morph-aware neurons. For the Feature Focus Detector, the features are very balanced and higher equal error rates appear only for the neurons that represent regions in the image with less content, e.g., at the regions of the cheeks at the border of the image. It also shows significantly lower equal error rates for classifications based on a single neuron for both the genuine-aware neurons and the morph-aware ones compared to the other models.

7.3. Relevance Distribution

Table 2 shows how much relevance on average is assigned to the morphed region(s) of the partial morphs for the differently trained networks, both for FLRP and LRP. In most cases, FLRP assigns more relevance to the region that contains morphing artifacts than LRP.

For the Naïve Detector and the partial morphs of eyes, the similarity of FLRP and LRP is quite strong and the discrimination power of neurons that are responsible for these regions is also quite high. For the partial morphs of mouth and nose, the difference between FLRP and LRP is very strong and the discrimination power much lower than for the eyes. The Complex MC Detector shows a very interesting behavior for the partial morphs. Here LRP assigns only little relevance to morphed eye regions compared to the other detectors. This, and the fact that the discrimination power is stronger for the genuine-aware neurons, supports the evidence in [15] that the Complex MC Detector does not focus mainly on artifacts induced by the face morphing process, but compares different regions and checks if they fit to the same model for properties of genuine faces. These comparison structures in the DNN result in relevance assignments to the non-morphed parts and LRP does not highlight the important artifacts. A detailed explanation of this phenomenon is described in [15]. The DNN based on the Xception architecture has a similarly poor performance as the Complex MC Detector. The Feature Focus Detector with FLRP performs best or nearly similar in all cases. It assigns 83.3% of relevance correctly in the worst case and 94% correctly on average. Moreover, the discrimination power of this detector is much higher compared to the other two.

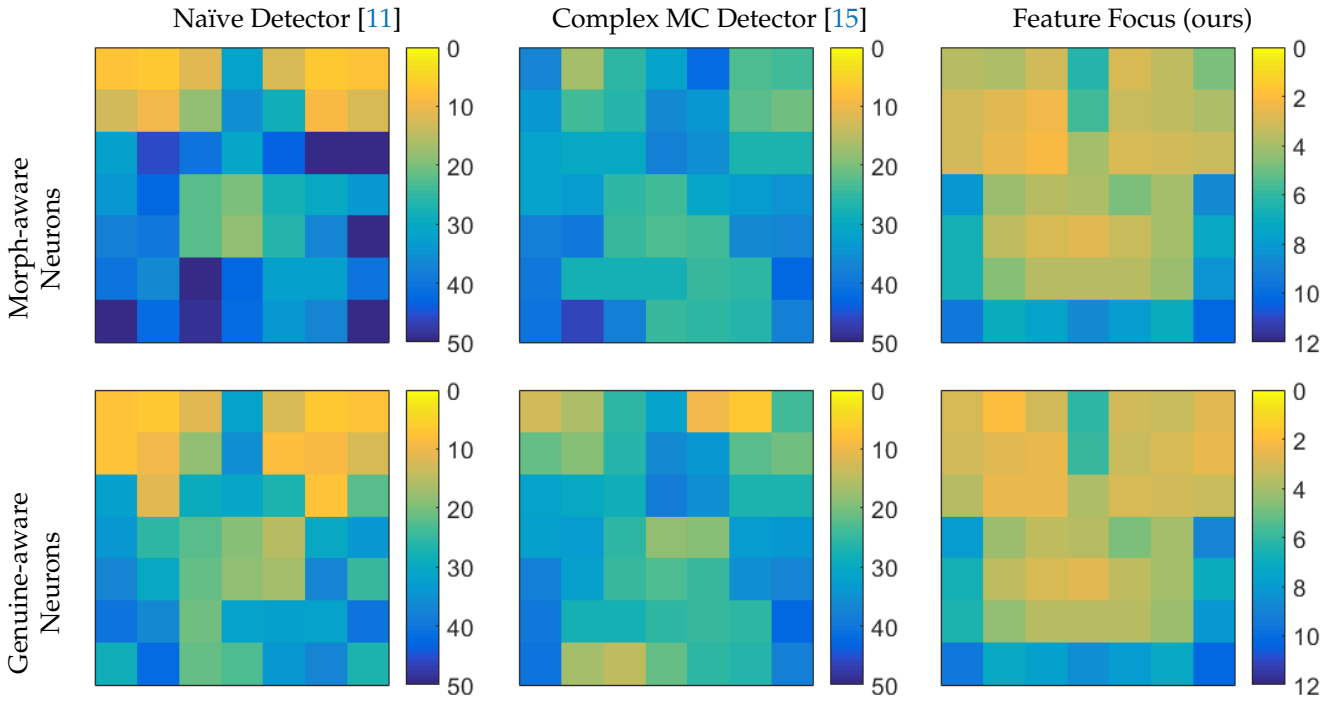


Figure 5. Visualization of the discrimination power of selected features of the different DNNs. The features in the top left and top right areas describe the regions of the left and right eye, the features in the center describe the nose and those in the bottom center describe the mouth. While the Naïve Detector can distinguish between genuine and morphed face images mostly by the eyes, the Multiclass MC Detector and the Feature Focus Detector are more balanced and can also make predictions based on other parts of the face. The features of the Feature Focus Detector outperform the other two detectors in terms of discrimination power. Visualization of the discrimination power of selected features of the different DNNs. The exact numbers are shown in Appendix A. The column Feature Focus has a different scale, since the EER is far smaller compared to the other two.

7.4. Sample Results

Figures 6 and 7 show relevance distributions for LRP and FLRP applied to a morphed face image and a partial morph for the differently trained detectors. The morphed face image in Figure 6 contains visibly strong morphing artifacts around the nostrils and some minor artifacts around the oral fissure, but no striking artifacts in the eyes. However, LRP

assigns only a small amount of relevance or even negative relevance to the mouth and in case of the Naïve and Complex MC Detector it assigns negative relevance to the nostrils. Contrarily, FLRP assigns more positive relevance to the clearly visible artifacts as one would expect from a component that should highlight traces of forgery in morphed face images. For the Naïve and the Complex MC Detector, FLRP also assigns a significant amount of relevance to the region of the eyes. As shown in our experiments on the relevance assignment using partial morphs, these two detectors are known to also assign relevance to non-forged regions. Especially for images with a morphed mouth and nose, these two detectors in combination with both relevance propagation methods tend to assign a significant amount of relevance to non-forged regions. This behavior is also visible for this morphed face image. Figure 7 shows the relevance distributions for a partial morph of the left eye and the mouth. In the FLRP case, relevance is mainly assigned to the left eye, which is obviously forged. It also assigns some relevance to the nose and to the mouth for the Naïve and Complex MC Detector. FLRP's relevance distribution for the Feature Focus Detector, however, is as expected only on the mouth and left eye. The relevance assignment of LRP for the eyes are plausible for the Naïve and Feature Focus Detector. Positive relevance is assigned to the morphed eye and negative to the genuine eye. For the Complex MC Detector, however, it is contrary to what we would expect. According to LRP, the genuine eye contributes to the class morphed face image and the morphed eye inhibits its activation. This phenomenon and its cause are described in detail in [15]. Furthermore, LRP's relevance assignment for the nose region is plausible only for the Naïve Detector and Feature Focus Detector. Figure 8 shows LRP relevance assignments for the detector based on the Xception architecture and Grad-Cam and LayerCAM results for the Naïve Detector. LRP assigns predominantly relevance to genuine parts of the face or parts without obvious artifacts. Image structures such as the border of the iris or nostrils are not immediately recognizable by the VGG-A-based detectors. Malolan et al. [44] showed also that LRP does not highlight recognizable structures in the image for DNNs based on the Xception architecture. The methods Grad-CAM [32] and LayerCAM [48] can only mark coarse regions as relevant for the decision-making process. Thus, these methods are not suitable for localizing the exact structures corresponding to traces of forgery.

Table 2. Relevance distribution on partial morphs. While LRP and FLRP assign relevance to genuine areas to a notable extent for the Naïve Detector, the Complex MC Detector and Xception-based Detector, FLRP assigns nearly all relevance to the forged part only for our proposed Feature Focus Detector.

Morphed Regions	**Naïve [11]**		**Complex MC [15]**		**Feature Focus (Ours)**		**Xception [43]**
	LRP	**FLRP**	**LRP**	**FLRP**	**LRP**	**FLRP**	**LRP**
left eye	74.1%	76.7%	8.8%	53.6%	80.0%	89.0%	21.7%
right eye	78.9%	86.1%	25.7%	45.4%	72.1%	87.7%	29.2%
nose	36.7%	71.4%	5.3%	40.5%	57.4%	85.2%	13.6%
mouth	18.2%	43.2%	22.3%	48.2%	16.6%	83.3%	3.8%
both eyes	91.8%	92.9%	44.9%	68.6%	91.2%	94.7%	72.2%
left eye, nose	84.8%	92.0%	18.7%	68.9%	89.6%	95.2%	36.6%
right eye, nose	87.1%	95.2%	32.5%	63.3%	87.6%	95.9%	54.2%
left eye, mouth	77.5%	84.1%	11.4%	71.4%	83.7%	96.7%	30.9%
right eye, mouth	82.5%	90.6%	61.3%	66.8%	78.9%	95.8%	39.9%
mouth and nose	48.8%	85.1%	38.2%	63.2%	57.0%	94.2%	19.6%
all but left eye	89.5%	97.4%	73.0%	75.8%	87.9%	98.4%	62.4%
all but right eye	86.7%	95.1%	34.1%	79.1%	90.2%	98.0%	45.0%
all but nose	92.7%	95.0%	33.5%	79.6%	94.2%	97.8%	83.2%
all but mouth	95.0%	97.6%	76.4%	79.0%	96.1%	97.0%	86.7%
all	95.9%	98.9%	89.4%	85.5%	97.2%	98.9%	91.3%

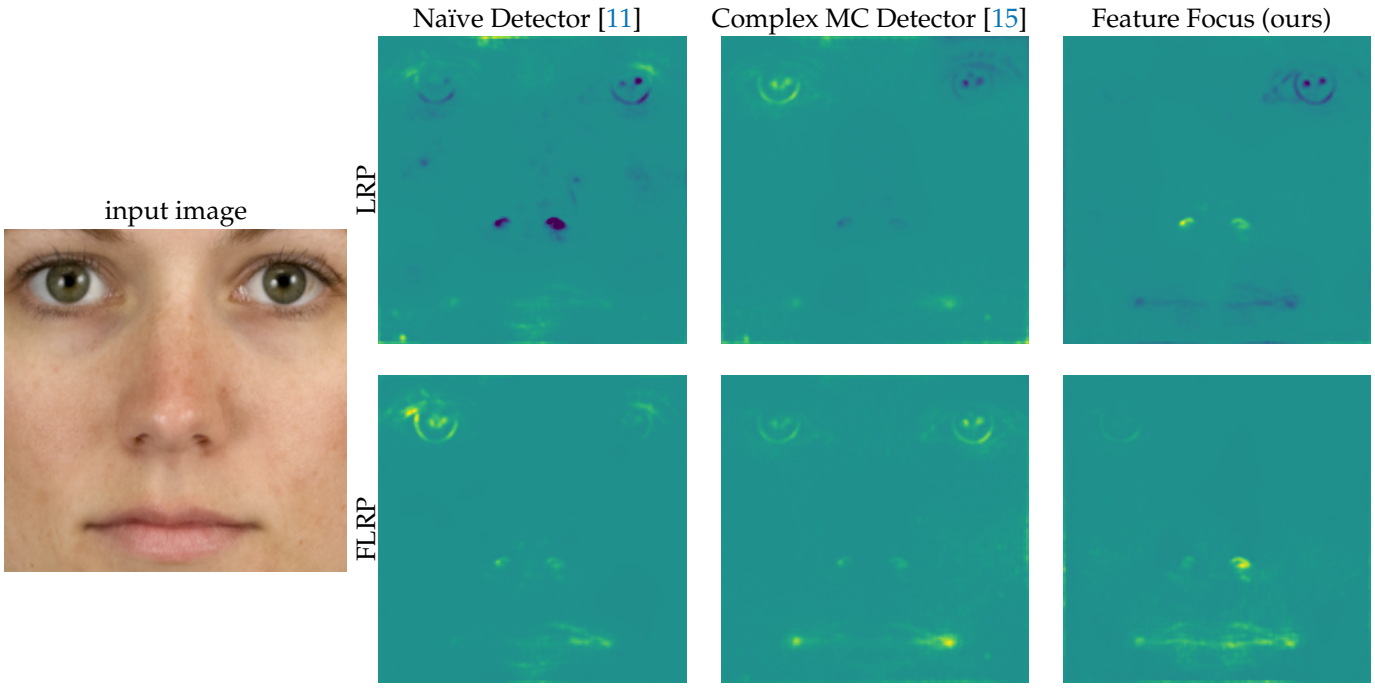


Figure 6. FLRP and LRP relevance distributions for a morphed face image. A yellow colored region contributes to the decision "morphed face image", a blue color denies it and a olive green colored region contains no information about the decision. In this example we would expect LRP and FLRP to highlight mostly the nose and mouth as relevant for the decision, since they contain obvious morphing artifacts. However, only FLRP in combination with the Feature Focus Detector does so. The other methods highlight also other parts as relevant to a strong extent or mark the mouth or nose as denning the decision.

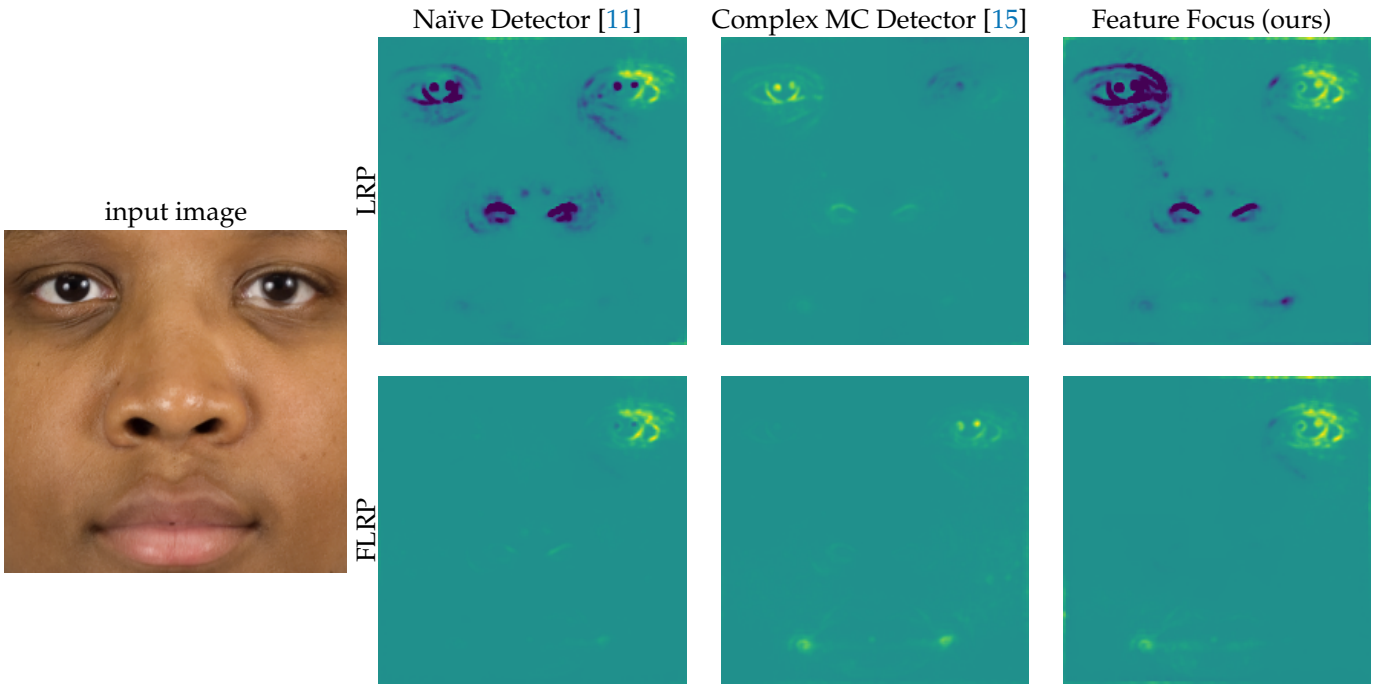


Figure 7. FLRP and LRP relevance distributions for a partial morph of the left eye and the mouth. A yellow colored region contributes to the decision "morphed face image", a blue color denies it and a olive green colored region contains no information about the decision. In this example LRP and FLRP should highlight the mouth and the left eye as relevant for the decision, since only these regions contain traces of forgery. Only FLRP in combination with the Feature Focus Detector or the Complex MC Detector does so. The other methods highlight also other regions as contributing or the mouth as denning the decision.

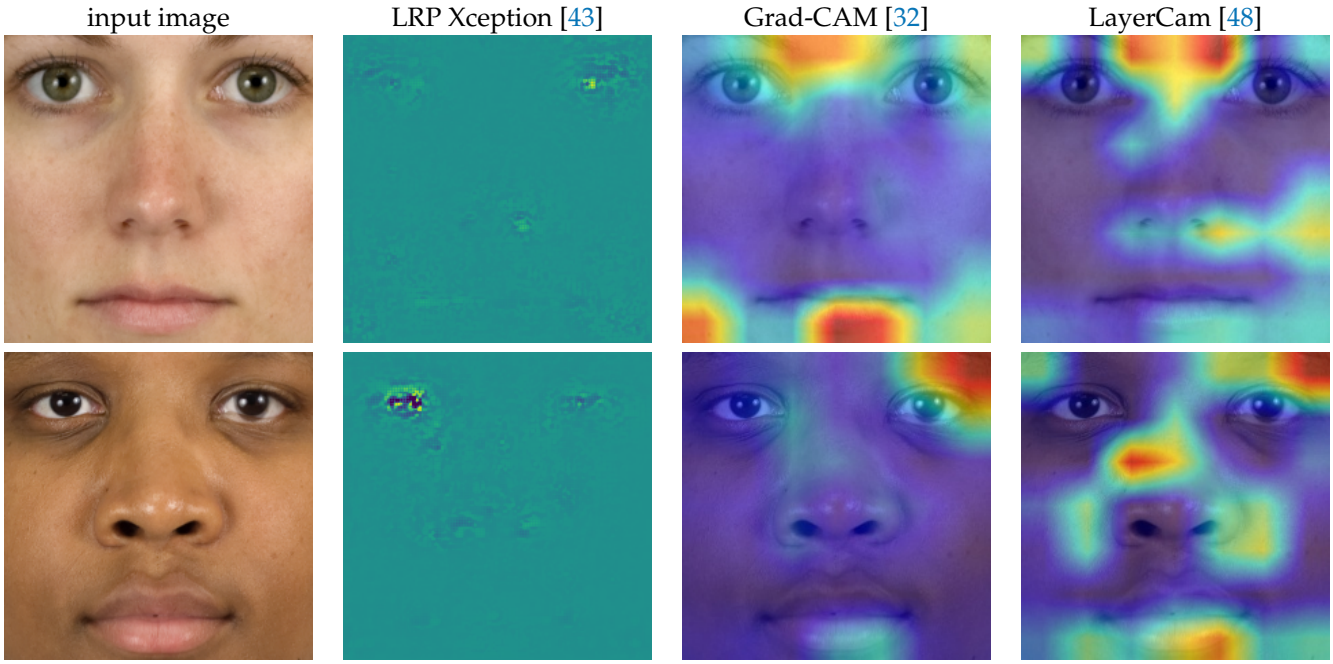


Figure 8. LRP relevance assignment results for the Xception architecture DNN and Grad-CAM and LayerCAM for the Naïve Detector applied on the same images shown in Figures 6 and 7. LRP assigns relevance predominantly to regions without traces of forgery for both images. Furthermore, the structures of the facial features are not as visible as for the VGG-A-based architectures. Grad-CAM and LayerCAM show only the coarse regions that are relevant according to these methods and highlight also genuine regions as relevant.

8. Discussion

Applying LRP might fail for some DNN-based face morphing attack detectors if the intention is to highlight regions that contain structures that are typical for a class of interest. Complex structures in the fully-connected layers of a DNN can make LRP highlight non-forged regions as relevant structures for causing activations of the class of morphed face images [15]. To tackle this problem, we propose two mutually beneficial concepts, Feature Focus and FLRP. Using partial morphs, which contain traces of forgery only in pre-defined regions, we show that FLRP is more accurate in highlighting these traces of forgery than LRP, but it is still not optimal. Especially for forgeries of the mouth region, it performs much worse than for other parts of the face. One problem of FLRP is that it has to identify a small set of relevant neurons, which represent traces of forgery, in the feature output of the DNN. There is no guarantee that such a set of morph-aware neurons, which have a strong activation if the input image is a morphed image, exists in a DNN's feature output. A DNN might assign different kinds of artifacts to different neurons or learn a different approach to detecting morphed face images. The Multiclass MC Detector, for example, has better genuine-aware neurons, which show a strong activation if and only if the input image is a genuine face image, than morph-aware neurons. Feature Focus solves this problem by design. Its feature output has only two channels and a loss function shapes them during training, such that one of them is morph-aware and the other one genuine-aware. Thus, the selection of relevant neurons for FLRP becomes trivial. This proposed change of architecture and the proposed loss function for the Focus Feature Detector increase morph detection accuracy. Furthermore, LRP and FLRP more accurately assign relevance to morphed image regions for this detector. An analysis of the DNN feature output's discrimination power shows that it has much more accurate morph-aware and genuine-aware neurons than other detectors. Furthermore, the simplification of the

network reduces the memory required for its weights from about 515 Megabytes to only 27 Megabytes.

9. Conclusions

In this paper, we propose Feature Focus, a transparent and accurate DNN-based face morphing detector and FLRP, an extension of the explainability method LRP. We quantitatively prove the advantages of FLRP over LRP for the analysis of DNN-based face morphing attack detectors and show the advantages of the new detector, especially in combination with FLRP. The new detector overcomes FLRP's problem of selecting relevant neurons. Furthermore, it results in a better relevance assignment for FLRP such that most of the relevance is assigned to the morphed region only. In addition, the accuracy in detecting morphed face images increases compared to the Naïve Detector and the Multiclass MC Detector [15]. FLRP, our modification of the network architecture, and our proposed loss function increase the interpretability of DNN-based face morphing attack detectors significantly. These improvements allow highlighting structures that are typical of morphed face images with high accuracy and by this, they constitute a well-suited tool to explain why an image is a forgery and to get a better understanding of the detector's decision.

Author Contributions: Conceptualization, C.S., A.H. and P.E.; methodology, C.S., A.H. and P.E.; software, C.S.; validation, C.S.; formal analysis, C.S.; investigation, C.S.; resources, P.E. and A.H.; data curation, C.S.; writing—original draft preparation, C.S.; writing—review and editing, C.S., A.H. and P.E.; visualization, C.S.; supervision, P.E and A.H.; project administration, P.E. and A.H.; funding acquisition, P.E and A.H. and C.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partly funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 833704 (D4FLY) and by the German Federal Ministry of Education and Research (BMBF) under grant number FKZ: 13N15735.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

Abbreviations

The following abbreviations are used in this manuscript:

APCER	Attack Presentation Classification Error Rate
BPCER	Bona-fide Presentation Classification Error Rate
DNN	Deep Neural Networks
EER	Equal-Error-Rate
FLRP	Focused Layer-wise Relevance Propagation
LRP	Layer-wise Relevance Propagation

Appendix A. Discrimination Power of Single Features

Table A1. EERs for Morph-/Genuine-aware Neurons of the Naïve Detector.

Morph-Aware Neurons							Genuine-Aware Neurons						
7.3%	6.7%	11.2%	31.4%	12.2%	6.8%	7.2%	7.4%	6.7%	11.2%	31.5%	12.1%	6.8%	7.2%
13.3%	9.8%	18.3%	35.7%	28.9%	9.3%	11.9%	7.0%	9.7%	18.1%	35.6%	8.3%	9.2%	11.8%
32.6%	46.0%	40.3%	30.7%	43.0%	53.1%	51.0%	32.6%	11.0%	29.2%	30.5%	27.2%	7.7%	21.9%
34.3%	42.5%	22.6%	19.7%	27.6%	30.3%	33.8%	34.2%	25.6%	22.4%	19.4%	15.1%	30.4%	33.7%
37.5%	39.4%	22.0%	18.7%	26.3%	36.9%	51.8%	37.4%	30.1%	21.8%	18.7%	16.5%	36.8%	24.2%
40.1%	36.5%	62.8%	42.8%	32.2%	32.1%	40.3%	40.1%	36.4%	21.0%	31.9%	32.2%	31.8%	40.4%
58.9%	41.6%	49.1%	41.5%	34.0%	37.0%	67.4%	28.7%	41.7%	21.1%	23.3%	34.1%	36.9%	27.3%

Table A2. EERs for Morph-/Genuine-aware Neurons of the Multiclass MC Detector.

Morph-Aware Neurons								Genuine-Aware Neurons					
37.4%	16.5%	25.1%	32.0%	42.1%	22.8%	24.1%	13.2%	16.0%	25.0%	31.9%	9.7%	6.5%	24.0%
33.8%	24.1%	26.6%	36.5%	33.8%	22.2%	20.6%	21.7%	19.5%	26.5%	36.1%	34.0%	22.2%	20.6%
31.7%	30.0%	30.0%	38.3%	35.6%	26.8%	27.1%	31.9%	30.1%	28.9%	38.3%	35.5%	26.8%	27.1%
32.0%	33.3%	25.5%	23.6%	30.5%	33.0%	34.5%	32.2%	33.2%	25.5%	18.5%	19.5%	33.1%	34.2%
38.1%	39.3%	24.2%	22.9%	24.1%	36.1%	37.1%	38.2%	32.0%	24.4%	22.8%	24.5%	35.8%	37.3%
39.4%	28.0%	28.0%	27.6%	24.7%	25.0%	42.6%	39.5%	27.8%	27.9%	25.8%	24.6%	25.0%	42.7%
40.5%	46.8%	38.1%	24.4%	25.3%	26.2%	38.3%	40.4%	16.5%	14.6%	21.5%	25.0%	26.1%	38.0%

Table A3. EERs for Morph-/Genuine-aware Neurons of the Feature Focus Detector.

Morph-Aware Neurons								Genuine-Aware Neurons					
3.6%	3.9%	3.1%	6.2%	3.0%	3.5%	4.8%	2.8%	2.0%	3.1%	6.2%	3.1%	3.2%	2.7%
3.1%	2.7%	2.3%	5.8%	3.2%	3.5%	3.9%	3.1%	2.7%	2.4%	5.9%	3.2%	2.8%	2.4%
3.1%	2.6%	2.2%	4.1%	3.0%	3.1%	3.4%	3.7%	2.6%	2.4%	3.9%	2.8%	3.1%	3.3%
8.1%	4.3%	3.7%	3.9%	4.8%	4.0%	8.8%	8.0%	4.2%	3.5%	3.7%	4.7%	4.0%	8.8%
6.6%	3.5%	3.0%	2.7%	3.4%	4.0%	7.2%	6.7%	3.5%	3.0%	2.7%	3.5%	4.0%	7.0%
6.6%	4.5%	3.6%	3.6%	3.6%	4.3%	8.3%	6.5%	4.4%	3.6%	3.6%	3.7%	4.2%	8.2%
9.5%	7.1%	7.6%	8.8%	8.0%	7.1%	10.2%	9.5%	7.2%	7.8%	8.6%	8.0%	7.2%	10.1%

References

1. Ferrara, M.; Franco, A.; Maltoni, D. The magic passport. In Proceedings of the IEEE International Joint Conference on Biometrics, Clearwater, FL, USA, 29 September–2 October 2014. [CrossRef]
2. Ferrara, M.; Franco, A.; Maltoni, D. On the Effects of Image Alterations on Face Recognition Accuracy. In *Face Recognition across the Imaging Spectrum*; Springer: Cham, Switzerland, 2016. [CrossRef]
3. Lewis, M.; Statham, P. CESG Biometric Security Capabilities Programme: Method, Results and Research challenges. In Proceedings of the Biometrics Consortium Conference (BCC), Arlington, VA, USA, 20–22 September 2004.
4. Scherhag, U.; Rathgeb, C.; Merkle, J.; Breithaupt, R.; Busch, C. Face Recognition Systems Under Morphing Attacks: A Survey. *IEEE Access* **2019**, *7*, 23012–23026. [CrossRef]
5. Makrushin, A.; Wolf, A. An Overview of Recent Advances in Assessing and Mitigating the Face Morphing Attack. In Proceedings of the EUSIPCO 2018: 26th European Signal Processing Conference, Roma, Italy, 3–7 September 2018.
6. van der Hor, T. Developing harmonized automated border control (ABS) training capabilities. *ICAO TRIP Mag.* **2017**, *12*, 2.
7. European Commission. *Action Plan to Strengthen the European Response to Travel Document Fraud*; Council of the European Union: Brussels, Belgium, 2016.
8. Ngan, M.; Grother, P.; Hanaoka, K.; Kuo, J. *Face Recognition Vendor Test (FRVT) Part 4: MORPH—Performance of Automated Face Morph Detection*; NIST Interagency/Internal Report (NISTIR); National Institute of Standards and Technology: Gaithersburg, MD, USA, 2020. [CrossRef]
9. Adjabi, I.; Ouahabi, A.; Benzaoui, A.; Taleb-Ahmed, A. Past, Present, and Future of Face Recognition: A Review. *Electronics* **2020**, *9*, 1188. [CrossRef]
10. Minaee, S.; Luo, P.; Lin, Z.L.; Bowyer, K. Going Deeper Into Face Detection: A Survey. *arXiv* **2021**, arXiv:2103.14983.
11. Seibold, C.; Samek, W.; Hilsmann, A.; Eisert, P. Detection of Face Morphing Attacks by Deep Learning. In Proceedings of the 16th International Workshop, IWDW 2017, Magdeburg, Germany, 23–25 August 2017. [CrossRef]
12. Debiasi, L.; Rathgeb, C.; Scherhag, U.; Uhl, A.; Busch, C. PRNU Variance Analysis for Morphed Face Image Detection. In Proceedings of the 9th IEEE International Conference on Biometrics Theory, Applications and Systems, Redondo Beach, CA, USA, 22–25 October 2018.
13. Scherhag, U.; Debiasi, L.; Rathgeb, C.; Busch, C.; Uhl, A. Detection of Face Morphing Attacks Based on PRNU Analysis. *IEEE Trans. Biom. Behav. Identity Sci.* **2019**, *1*, 302–317. [CrossRef]
14. Lapuschkin, S.; Wäldchen, S.; Binder, A.; Montavon, G.; Samek, W.; Müller, K.R. Unmasking Clever Hans Predictors and Assessing What Machines Really Learn. *Nat. Commun.* **2019**, *10*, 1096. [CrossRef] [PubMed]
15. Seibold, C.; Samek, W.; Hilsmann, A.; Eisert, P. Accurate and robust neural networks for face morphing attack detection. *J. Inf. Secur. Appl.* **2020**, *53*, 102526. [CrossRef]
16. Seibold, C.; Hilsmann, A.; Eisert, P. Focused LRP: Explainable AI for Face Morphing Attack Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops, Waikola, HI, USA, 5–9 January 2021; pp. 88–96.

17. Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.R.; Samek, W. On Pixel-wise Explanations for Non-Linear Classifier Decisions by Layer-wise Relevance Propagation. *PLoS ONE* **2015**, *10*, e0130140. [CrossRef]
18. Seibold, C.; Hilsmann, A.; Eisert, P. Style Your Face Morph and Improve Your Face Morphing Attack Detector. In Proceedings of the International Conference of the Biometrics Special Interest Group, Darmstadt, Germany, 18–20 September 2019.
19. Scherhag, U.; Rathgeb, C.; Merkle, J.; Busch, C. Deep Face Representations for Differential Morphing Attack Detection. *IEEE Trans. Inf. Forensics Secur.* **2020**, *15*, 3625–3639. [CrossRef]
20. Banerjee, S.; Ross, A. Conditional Identity Disentanglement for Differential Face Morph Detection. In Proceedings of the 2021 IEEE International Joint Conference on Biometrics (IJCB), Shenzhen, China, 4–7 August 2021; pp. 1–8. [CrossRef]
21. Ferrara, M.; Franco, A.; Maltoni, D. Face Demorphing. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 1008–1017. [CrossRef]
22. Peng, F.; Zhang, L.; Long, M. FD-GAN: Face De-Morphing Generative Adversarial Network for Restoring Accomplice's Facial Image. *IEEE Access* **2019**, *7*, 75122–75131. [CrossRef]
23. Seibold, C.; Hilsmann, A.; Eisert, P. Reflection Analysis for Face Morphing Attack Detection. In Proceedings of the 26th European Signal Processing Conference (EUSIPCO), Roma, Italy, 3–7 September 2018.
24. Makrushin, A.; Neubert, T.; Dittmann, J. Automatic Generation and Detection of Visually Faultless Facial Morphs. In Proceedings of the VISIGRAPP—Volume 6: VISAPP, Porto, Portugal, 27 February–1 March 2017; pp. 39–50.
25. Neubert, T. Face Morphing Detection: An Approach Based on Image Degradation Analysis. In Proceedings of the 16th International Workshop on Digital-forensics and Watermarking (IWDW 2017), Magdeburg, Germany, 23–25 August 2017; pp. 93–106.
26. Neubert, T.; Kraetzer, C.; Dittmann, J. A Face Morphing Detection Concept with a Frequency and a Spatial Domain Feature Space for Images on eMRTD. In Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, Paris, France, 3–5 July 2019.
27. Ramachandra, R.; Raja, K.B.; Busch, C. Detecting morphed face images. In Proceedings of the 8th IEEE International Conference on Biometrics Theory, Applications and Systems, Niagara Falls, NY, USA, 6–9 September 2016; pp. 1–7.
28. Damer, N.; Zienert, S.; Wainakh, Y.; Saladié, A.M.; Kirchbuchner, F.; Kuijper, A. A Multi-detector Solution Towards an Accurate and Generalized Detection of Face Morphing Attacks. In Proceedings of the 2019 22th International Conference on Information Fusion (FUSION), Ottawa, ON, Canada, 2–5 July 2019; pp. 1–8.
29. Makrushin, A.; Kraetzer, C.; Dittmann, J.; Seibold, C.; Hilsmann, A.; Eisert, P. Dempster-Shafer Theory for Fusing Face Morphing Detectors. In Proceedings of the 2019 27th European Signal Processing Conference (EUSIPCO), A Coruna, Spain, 2–6 September 2019; pp. 1–5. [CrossRef]
30. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In Proceedings of the ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014.
31. Chen, C.; Li, O.; Barnett, A.; Su, J.; Rudin, C. This looks like that: deep learning for interpretable image recognition. *arXiv* **2018**, arXiv:1806.10574.
32. 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 ICCV IEEE Computer Society, Venice, Italy, 22–29 October 2017; pp. 618–626.
33. Kohlbrenner, M.; Bauer, A.; Nakajima, S.; Binder, A.; Samek, W.; Lapuschkin, S. Towards best practice in explaining neural network decisions with LRP. In Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020.
34. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
35. Ouahabi, A.; Taleb-Ahmed, A. Deep learning for real-time semantic segmentation: Application in ultrasound imaging. *Pattern Recognit. Lett.* **2021**, *144*, 27–34. [CrossRef]
36. Yin, L.; Chen, X.; Sun, Y.; Worm, T.; Reale, M. A high-resolution 3D dynamic facial expression database. In Proceedings of the 8th IEEE International Conference on Automatic Face Gesture Recognition, Amsterdam, The Netherland, 17–19 September 2008; pp. 1–6. [CrossRef]
37. Ma, D.; Correll, J.; Wittenbrink, B. The Chicago face database: A free stimulus set of faces and norming data. *Behav. Res. Methods* **2015**, *47*, 1122–1135. [CrossRef] [PubMed]
38. Phillips, P.J. *Color FERET Database*; National Institute of Standards and Technology: Gaithersburg, MD, USA, 2003.
39. DeBruine, L.; Jones, B. Face Research Lab London Set. *figshare* **2017**. [CrossRef]
40. Kasiński, A.; Florek, A.; Schmidt, A. The PUT face database. *Image Process. Commun.* **2008**, *13*, 59–64.
41. Grgic, M.; Delac, K.; Grgic, S. SCface—Surveillance Cameras Face Database. *Multimed. Tools Appl.* **2011**, *51*, 863–879. [CrossRef]
42. Hancock, P. *Utrecht ECVP Face Dataset*; School of Natural Sciences, University of Stirling: Stirling, UK, 2008. Available online: http://pics.stir.ac.uk/ (accessed on 13 April 2017).
43. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 1800–1807. [CrossRef]
44. Malolan, B.; Parekh, A.; Kazi, F. Explainable Deep-Fake Detection Using Visual Interpretability Methods. In Proceedings of the 2020 3rd International Conference on Information and Computer Technologies (ICICT), San Jose, CA, USA, 9–12 March 2020; pp. 289–293.
45. Pérez, P.; Gangnet, M.; Blake, A. Poisson Image Editing. *ACM Trans. Graph.* **2003**, *22*, 313–318. [CrossRef]

46. Fraunhofer HHI and TU Berlin. Available online: http://heatmapping.org (accessed on 20 July 2021).
47. Lapuschkin, S. Opening the Machine Learning Black Box with Layer-Wise Relevance Propagation. Doctoral Thesis, Technische Universität Berlin, Berlin, Germany, 2019. [CrossRef]
48. Jiang, P.T.; Zhang, C.B.; Hou, Q.; Cheng, M.M.; Wei, Y. LayerCAM: Exploring Hierarchical Class Activation Maps. *IEEE Trans. Image Process.* **2021**. [CrossRef] [PubMed]
49. Seibold, C.; Hilsmann, A.; Makrushin, A.; Kraetzer, C.; Neubert, T.; Dittmann, J.; Eisert, P. Visual Feature Space Analyses of Face Morphing Detectors. In Proceedings of the 2019 IEEE International Workshop on Information Forensics and Security (WIFS), Delft, The Netherlands, 9–12 December 2019; pp. 1–6. [CrossRef]
50. International Organization for Standardization. *Information Technology—Biometric Presentation Attack Detection—Part 3: Testing and Reporting*; ISO Standard No. 30107-3:2017; International Organization for Standardization: Geneva, Switzerland, 2017. Available online: https://www.iso.org/standard/67381.html (accessed on 20 July 2021).