electronics

Article

Generated or Not Generated (GNG): The Importance of Background in the Detection of Fake Images

Marco Tanfoni *,†, Elia Giuseppe Ceroni *,†, Sara Marziali, Niccolò Pancino, Marco Maggini and Monica Bianchini

Department of Information Engineering and Mathematics, University of Siena, 53100 Siena, Italy; sara.marziali@student.unisi.it (S.M.); niccolo.pancino@unisi.it (N.P.); marco.maggini@unisi.it (M.M.); monica.bianchini@unisi.it (M.B.)

* Correspondence: marco.tanfoni@student.unisi.it (M.T.); elia.ceroni@student.unisi.it (E.G.C.)

† These authors contributed equally to this work.

Abstract: Facial biometrics are widely used to reliably and conveniently recognize people in photos, in videos, or from real-time webcam streams. It is therefore of fundamental importance to detect synthetic faces in images in order to reduce the vulnerability of biometrics-based security systems. Furthermore, manipulated images of faces can be intentionally shared on social media to spread fake news related to the targeted individual. This paper shows how fake face recognition models may mainly rely on the information contained in the background when dealing with generated faces, thus reducing their effectiveness. Specifically, a classifier is trained to separate fake images from real ones, using their representation in a latent space. Subsequently, the faces are segmented and the background removed, and the detection procedure is performed again, observing a significant drop in classification accuracy. Finally, an explainability tool (SHAP) is used to highlight the salient areas of the image, showing that the background and face contours crucially influence the classifier decision.

Keywords: fake detection; image forgery; explainability; interpretability; segmentation; SHAP; DeepLabV3+; MobileNetV3 Large

Citation: Tanfoni, M.; Ceroni, E.G.; Marziali, S.; Pancino, N.; Maggini, M.; Bianchini, M. Generated or Not Generated (GNG): The Importance of Background in the Detection of Fake Images. *Electronics* **2024**, *13*, 3161. https://doi.org/10.3390/electronics13163161

Academic Editor: Chiman Kwan

Received: 16 July 2024
Revised: 27 July 2024
Accepted: 2 August 2024
Published: 10 August 2024

1. Introduction

The emergence of the latest and most powerful generative models, such as NVIDIA's StyleGAN [1] and Stable Diffusion models [2], has led to a huge circulation of AI-generated images of human faces on the Internet. Indeed, many of the fakes they generate are almost indistinguishable—to the human eye—from real pictures of existing people, posing an actual threat to privacy and trustworthiness in online environments.

Currently, most of the models proposed in the literature are based on Convolutional Neural Networks (CNNs [3]), which are a class of deep learning (DL) models specifically designed for analyzing data in the form of images and videos [4,5]: CNNs acquire hierarchical spatial features through a series of convolutional, pooling, and dense layers. Starting from foundational research in [6,7], CNNs have transformed numerous areas, such as computer vision [8–10], medical image analysis, and fault detection [11–14], along with autonomous driving systems [15–18]. Beyond image classification, CNNs are effectively utilized in various applications like object detection, semantic segmentation [19–22], and image generation [23]. In this context, the ability to discern generated images from real ones has thus become a critical concern in various fields, such as security, media forensics and content moderation. In particular, the most relevant generative technologies are those based on Generative Adversarial Networks (GANs [24]): these models are remarkably useful in many tasks, but they facilitated the spread of hyper-realistic synthetic images, which are almost indistinguishable from genuine photos to the average, untrained human eye. Among these models, some of the most influential are certainly those in the StyleGAN [1,25,26] and BigGAN [27] families, which have significantly pushed the limits of image quality and resolution. For example, StyleGAN2 improved its predecessor's architecture to reduce artifacts and increase image fidelity, while BigGAN is known for higher resolution image generation with respect to previous GANs. These advancements made the task of distinguishing synthetically generated images from real ones increasingly difficult, thus emphasizing the necessity for effective detection mechanisms.

Synthetic face detection approaches can be classified in three major branches [28]: physical-based methods, physiological-based methods, and DL-based methods. Physical-based methods focus on the identification of real world-related artifacts and inconsistencies such as incongruous image illumination and reflections. Works belonging to this category leverage known problems of synthetic faces: eye pupil illumination, morphology and eyes symmetry are of particular interest, since current GAN generators struggle with this aspect. Physiological-based methods instead rely on the semantic aspect of human faces, searching for irregularities in face symmetry, pupil shape, iris color and texture, and other salient characteristics. DL-based methods use deep neural networks trained to effectively and automatically extract salient features from the images to detect synthetic faces. Usually, large networks pretrained on other image classification tasks are used and fine-tuned, leveraging their already effective feature extraction capabilities. Both physical and physiological-based methods achieve higher-than-human performance (which ranges between 26% and 80% accuracy) and provide built-in interpretability. Their strength is however limited by the strong environmental constraints used to define these methods, such as frontal portrait pose and limitations with regard to face occlusion [28]. DL-based methods can achieve significantly higher performance than physical and physiological methods; however, they completely lack interpretability and operate like a black box. Recently, numerous efforts have been made in the field of DL-based methods for fake face detection based on both specialized datasets and tools, which were created specifically for this purpose. For instance, the OpenForensics dataset, introduced in [29], provides detailed face-wise annotations to aid in training models for multi-face forgery detection and segmentation in uncontrolled settings, enhancing research into deepfake prevention and face detection. Early works on DL-based synthetic faces detection used what was at the time the state-of-the-art pretrained CNN architectures, such as the VGG-Net used in [30]. Other approaches include the use a combination of both fine-grained frequency components and RGB color values of the image [31,32]. Additionally, other datasets and methods that incorporate facial landmark detection and segmentation are introduced in [33], paving the way for a more interpretable and granular solution to the face verification problem. In a similar way, ref. [34] tackles fake detection in a fine-grained classification approach, analyzing facial features to improve detection over multiple datasets. In [35], a novel architecture called Gram-Net is introduced, which bases its detection power on global image texture features, highlighting texture as an important indicator of image authenticity. Other works, such as [36], focus on visual features that make the detection robust to postprocessing procedures by leveraging luminance, chrominance components, and color space characteristics. Finally, in [37], DETER, a method that analyzes neuron activation patterns across layers to identify synthetic images, is proposed, demonstrating its robustness to various GAN-generated images and, in [38], the authors exploit a depth map-guided triplet network for effective deepfake detection, using depth information to enhance the detection accuracy by distinguishing real from fake faces based on discontinuity, inconsistent illumination, and blurring.

While the topic of fake face detection has already been tackled in the literature, there is still room to improve the explainability and interpretability of the classification models. Therefore, this paper addresses the explainability issue in the fake face recognition task, particularly focusing on the importance of the background to detect generated images. Hence, the aim of this study is to highlight an issue that can be considered crucial: current state-of-the-art models tend to identify not fake faces themselves but rather the artificial backgrounds associated with them.

To achieve this goal, a novel model, called Generated or Not Generated (GNG), is introduced, consisting of (1) a widely used semantic segmentation model (DeepLabV3+ [19])

employed for separating the face from the background—this module is also capable of locating various face features such as hair, eyes, nose, lips and so on that could be potentially useful for other applications; (2) a convolutional neural network model (MobileNetV3 [39]), used as a backbone for DeepLabV3+, to obtain a latent representation of the images; (3) a simple Multilayer Perceptron (MLP) for classifying the embeddings obtained in step (2) and, finally, (4) a SHapley Additive exPlanations (SHAP [40]) module, to highlight the most salient areas that guide the classifier's decision.

The rest of this paper is organized as follows: Section 2 presents the materials and methods involved in the study, hence the datasets, the model architectures, and the experimental settings. Moreover, it introduces a widely used explainability method which has been exploited in this study. Section 3 shows the experimental results, which are contextually discussed. Finally, Section 4 collects the conclusions and outlines possible future developments.

2. Materials and Methods

In this section, the data and the experimental methodology used in this work are described. In particular, Section 2.1 presents a precise description of the dataset obtained from online available resources, which is composed of real and fake facial images. Sections 2.2 and 2.3 give an overview of the GAN and CNN family models with a particular focus on the DeepLabV3+ model. Subsequently, Section 2.4 describes a widely used explainability method which has been exploited to study the explainability of the proposed models. Eventually, Section 2.5 shows the experimental setup chosen for the study with specific reference to data setting, model selection, and evaluation.

2.1. Datasets

The complete version of the Mut1ny [41] dataset was employed to train the background removal module. The dataset, comprising 70,621 images, has already been extensively used in the literature [42–44] and features a wide range of subjects of different ethnicities, ages, genders, facial poses and camera angles, with pixel-level labels created by hand by the Mut1ny team and other volunteers, obtaining fourteen different classes, as shown in Figure 1. A free 'community edition' of the dataset is also available [45], containing slightly less than a quarter of the images, fewer classes (no left/right differentiation), and not being updated.

For the fake detection task, the data were extracted from ArtiFact [46] (Artificial and Factual), a large collection of different datasets of real and synthetic images, including multiple categories such as people, animals, vehicles, and more. The dataset comprises 964,989 real images drawn from eight sources, to ensure diversity, and 1,531,749 synthetic images, obtained with twenty-five methods, specifically thirteen GANs [47] (such as StyleGAN, StyleGAN2 [25], StyleGAN3 [26], BigGAN [27], and CycleGAN [48]), seven diffusion models (such as Stable Diffusion and Latent Diffusion [2]), and five miscellaneous generators (such as Taming Transformer [49]), for a grand total of 2,496,738 different samples.

Since most of the real images contained in the dataset depict real-life objects, environments, and animals, a preliminary selection had to be made in order to extract a satisfactory amount of human subjects, specifically close-up images of faces, possibly including shoulders and part of the background. In particular, CelebA-HQ [50] and FFHQ [1] were used for the *Real* class (see Figure 2).The former is a high-quality version of the CelebA [51] dataset, comprising detailed images of celebrity faces, while the latter is a high-quality dataset of human faces taken from Flickr, thus offering a wide range of diverse samples. Synthetic images are generated within the same categories as the real images to maintain consistency in the dataset. Text-to-image and inpainting generators utilize captions and image masks from the COCO [52] dataset, while noise-to-image generators use normally distributed noise with different random seeds.

- **Background** (Class 0) ⬛
- **General face/head** (Class 1) ⬜
- **Left eye** (Class 2) 🟩
- **Right eye** (Class 3) 🟩
- **Nose** (Class 4) 🟦
- **Lips/mouth** (Class 5) 🟥
- **Hair** (Class 6) 🟨
- **Left eyebrow** (Class 7) 🟪
- **Right eyebrow** (Class 8) 🟪
- **Left ear** (Class 9) 🟦
- **Right ear** (Class 10) ⬛
- **Teeth** (Class 11) ⬜
- **Facial Hair** (Class 12) 🟥
- **Specs/sunglasses** (Class 13) 🟦

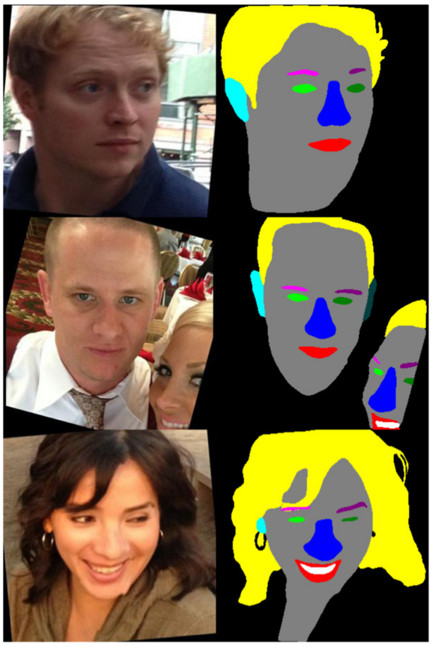

Figure 1. Examples of segmentation masks in Mut1ny dataset. In the experiments, the segmentation module was only used to finely remove the background in the images. To ensure a good variety of training samples, the facial images are drawn from different ethnicities, ages and genders, which are randomly rotated and with a wide facial poses angle range (from –90 to 90 degrees).

(**a**)

(**b**)

Figure 2. The datasets used for the *Real* class samples. (**a**) reports some examples from CelebA-HQ dataset, which provides high-quality images of celebrity faces. Images are retrieved from the official TensorFlow dataset documentation page (accessed on 3 August 2024). (**b**) is a teaser figure for the Flickr-Faces-HQ dataset, which offers a diverse set of human face images sourced from Flickr. Images are retrieved from the NVlabs ffhq-dataset GitHub repository (accessed on 3 August 2024).

As for the *Fake* image class, the images employed in the analyses were generated using StyleGAN2 and StyleGAN3. These two models have been chosen due to their popularity and prevalence in related research works and since they are capable of generating high-quality synthetic facial images.

To accurately reflect real-world conditions, both real and synthetic images in the ArtiFact dataset undergo various impairments. These include random cropping with a ratio of $r = 5/8$ and crop sizes ranging from a minimum of 160 to a maximum of 2048 pixels, resizing to 200 × 200 pixels, and JPEG compression with quality levels between 65 and 100.

2.2. StyleGAN Family

StyleGAN models build on the Progressive Growing GAN concept, previously introduced in [53], where common training instability issues, typical in GANs, are specifically addressed. The proposed training scheme involves the growing of the network during training: the network first learns how to generate low-resolution 4 × 4 images, which grow to high-detail 1024 × 1024 images by adding deep layers to the network, dividing the task in smaller and easier to learn subproblems. The first StyleGAN model has been a major breakthrough in realistic synthetic face generation technology, incorporating style transfer methodologies [54] generating images from a learned vector instead of directly feeding the latent vector to the generator part of the network. StyleGAN2 significantly improved the quality of generated images with respect to the original StyleGAN, especially in regard to the presence of minor but recognizable artefacts such as droplets of color blobs, which was achieved through removing the generation of images at each resolution, an improved regularization scheme for the loss, a better normalization technique and the implementation of skip connections in the network. The latest model of the family, StyleGAN3, improves its predecessor by addressing its propensity to fix finer details (such as hair texture) to specific image coordinates, which was due to the prevalence of image borders and the presence of aliasing patterns in generated faces. StyleGAN3 introduced sufficient zero padding around the image to reduce the border effect, reformulated the operations to work on continuous image representations to avoid the introduction of aliasing artefacts, simplified the network structure with respect to StyleGAN2 and removed some regularization that was previously introduced. Furthermore, transitional and rotational invariance is achieved by substituting the learned constant introduced in StyleGAN2 with information-wise equivalent Fourier features and by substituting 3 × 3 with 1 × 1 convolutions. Finally, low-pass filters were applied to upsample and downsample operations to further reduce the generation of artefacts.

2.3. DeepLabV3+

The background removal module is based mainly on DeepLabV3 +, which is a model belonging to the DeepLab [20] widely used family of semantic segmentation models developed by Google Research. The key innovation with respect to its predecessor is the integration of atrous spatial pyramid pooling (ASPP), allowing the model to precisely capture multi-scale information, crucial for refining segmentation boundaries. More precisely, this technique combines atrous convolution and spatial pyramid pooling (SPP [55]) operations to allow the detection of crucial information independently of the scale, thus making it particularly indicated for semantic segmentation tasks. During the atrous convolution operation, gaps are introduced inside the kernels. This way, the receptive fields of the neurons are increased, allowing to capture features across various scales without increasing the number of parameters. SPP is a particular pooling technique that enables the creation of outputs of uniform length independently of the input images size, overcoming the fixed input size limitations of current CNN models. In SPP, a custom layer on top of the convolutional part of the CNN is inserted, where the input feature maps are segmented into multiple bins at varying scales, pooling each one separately for feature extraction, and then concatenating these features to combine them, resulting in a fixed-size output vector that can be processed by the following fully connected section of the network. Summarizing, the ASPP layer allows for processing images of arbitrary sizes and enhances the network ability to capture both local and global spatial details while remaining robust to changes in object placement and size.

2.4. SHAP

SHapley Additive exPlanations (SHAP) is a powerful explainability method based on the Shapley value, which was introduced by Lloyd Shapley in 1953 to address the problem of fairly distributing the total contribution generated by a group of agents among individuals in a cooperative game [56]. In order to properly introduce the Shapley value, it is necessary to mention some game theory concepts [57].

A *game* is a set of circumstances whereby two or more players contribute to an outcome. Let N be a set of players, $S \subseteq N$ be a coalition and N be the grand coalition. The characteristic function v assigns to each coalition S the best results that the players in S may obtain, i.e., the total payoff, independently of the choices of the other players.

Definition 1. *Given a game with N players and characteristic function v, the* ***Shapley value*** *is the vector $\phi(v)$ whose component $\phi_i(v)$ is the average marginal contribution of player $i \in N$ with respect to all the permutations of the players [58].*

The definition guarantees the existence and uniqueness of the Shapley value, which is one of the most commonly used point solutions and, according to the main equation that can be found in the original paper, represents the average expected marginal contribution of one player after all possible combinations have been considered, thus rewarding each agent for its individual contribution, taking into account cooperation with others [59].

The SHapley Additive exPlanations (SHAP) [40] method that is exploited in this application is based on the calculation of the Shapley value of the conditional expectation function of a model in order to study its explainability. In the case of simple models, the best explanation method consists of the model itself. For more complex architectures, such as estimator ensembles and deep neural networks, however, it is necessary to introduce simpler explanation models that can be defined as interpretable approximations of the original one [40]. Local methods, in particular, explain a prediction $f(x)$, with f being the estimation model and x its input, using a simplified version of the input x' that maps to the original x via the mapping function $x = h_x(x')$. Given a local explanation model g, $g(x') \approx f(h_x(z'))$ must be ensured whenever $z' \approx x'$. Additive feature attribution methods are those for which the explanation model is a linear function of two binary variables

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{1}$$

where $z' \in \{0,1\}^M$, M is the number of simplified input features, and $\phi_i \in \mathbb{R}$ is the effect of feature i on the explanation model g. Methods which have explanation models that match this definition attribute an effect to each feature, and the sum of their attribution effects approximates the output of the original model. Alternatively, the vector z can be interpreted as a coalition vector with M being the maximum size of the coalition. Several popular explanation methods for deep learning, such as LIME [60], DeepLIFT [61] and Layer-Wise Relevance Propagation [62], all satisfy Equation (1), as well as three explanation models based on the Shapley value, Shapley regression values [63], Shapley sampling values [64], and quantitative input influence [65]. Since the Shapley value is the only vector of values that relates to the properties of local accuracy, missingness and consistency defined in the original paper, the only additive feature assignment method that satisfies the statement must be based on the Shapley value, and methods not based on the Shapley value violate local accuracy and/or consistency, thus making SHAP, built on the Shapley value of a conditional expectation function of the original model, the only method adhering to the three desired properties and using conditional expectations to define simplified inputs. The definition of the method aligns perfectly with Shapley regression, Shapley sampling, and quantitative input influence, while allowing connections to LIMe, DeepLIFT, and layer-wise relevance propagation feature attribution techniques. The construction of SHAP values are shown in Figure 3.

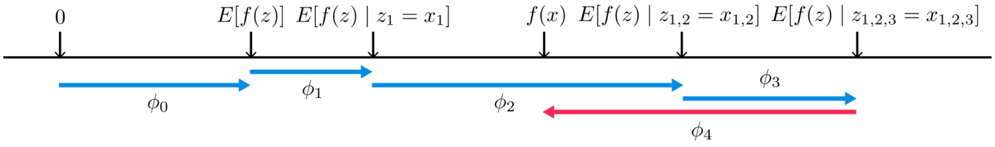

Figure 3. The expected base value $E[f(z)]$ is the predicted value of the model without any known features and $f(x)$ is the current output of the model given the input x. The diagram shows how SHAP values attributed to each feature change the expected model prediction when conditioning on that feature.

2.5. Experimental Setup

For the background removal phase, DeepLabV3+ was trained on the Mut1ny dataset to obtain the semantic segmentation of images—Figure 4 shows the proposed architecture. Augmentation was applied on the training set: specifically, random rotations within a range of ±15 degrees to emulate different head poses, horizontal and vertical translations, shifting images up to 10% of their original dimensions to adjust for alignment discrepancies, shear transformation of up to 0.2 radians to modify the image geometry, imitating a shift in perspective, random zoom by up to 20% to replicate variations in distance from the camera, and horizontal flips to represent different orientations. For the new pixels generated by rotations or shifts in width and height, a "nearest" filling method was applied to maintain local pixel similarity. In the presented version of the model, a MobileNetV3 Large [39] is used as a feature extractor, replacing the original Xception [66] backbone. A ResNet50 backbone was also tested to determine if the significant increase in the number of parameters resulted in a substantial change in performance. The ASPP (Atrous Spatial Pyramid Pooling) module incorporated dilation rates of 1, 4, 8, and 16 to achieve a balance between large contextual information and finer details. Furthermore, the kernels of the ASPP layer have been chosen in {3, 5, 7, 11} to optimize the model ability to capture contextual details.

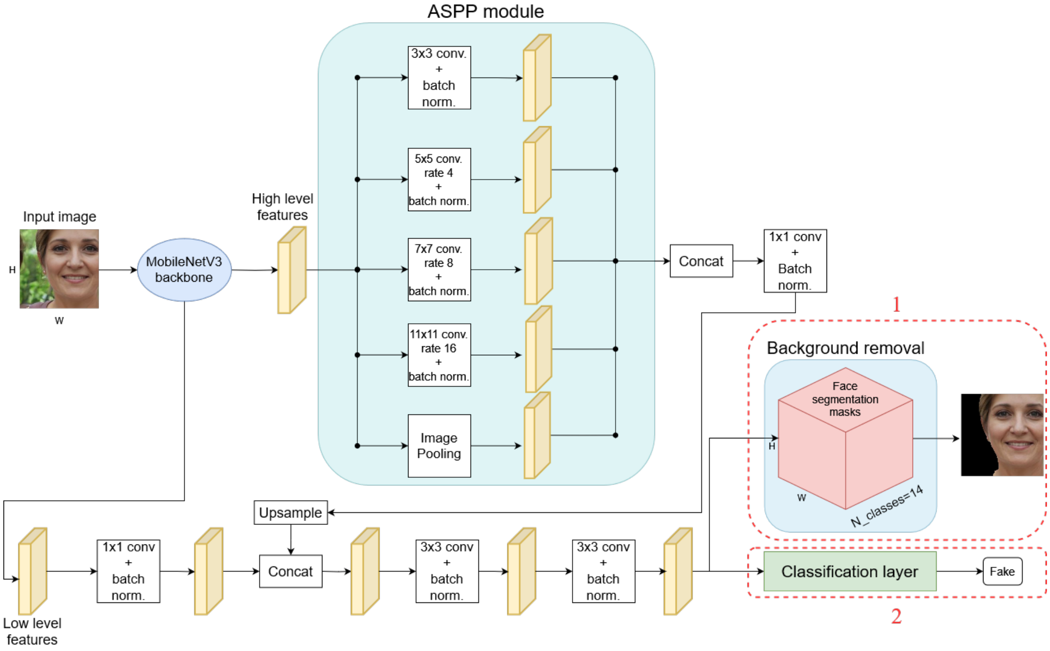

Figure 4. The proposed network architecture. The model serves for both face segmentation and real/fake classification tasks. The model trained this way is used first to infer on the ArtiFact dataset, obtaining its finely cropped version. In a second, separate and successive, phase—represented by the flow ending at the red dotted rectangle labeled '2'—the model performs the classification task on both versions of the dataset.

Then, a subset of the ArtiFact dataset is extracted from the entire collection. Specifically, both FFHQ and CelebA-HQ are used as real samples (*Class 0*, 100,000 elements in total), while images generated from StyleGAN2 and StyleGAN3 are extracted and used as the fake class (Class *1*, 200,000 and 48,867 elements in total, respectively) in the experiments. The previously trained segmentation module is applied to both real and generated images, and the class *Background* is isolated from the others, which are collapsed into one to obtain a fine foreground/background separation and provide the second version of the same datasets needed for the subsequent experiments. Then, both the original and the segmented images are fed separately to the model for the classification phase. The loss functions used for the two different training procedures are the Sparse Categorical Cross-Entropy (on a pixel level) and the Binary Cross-Entropy, respectively. For each training run, class weights were calculated according to the presence of each class across the samples to ensure a balanced contribution to the loss function. Adam optimizer was used with the learning rate initialized as 0.001. Both training sessions were performed on an NVidia RTX 4090 with 24 GB of VRAM. Since this was a binary classification problem, the metrics chosen to evaluate the performance of the model were those usually employed for this kind of tasks, specifically accuracy, precision, recall and F1-score, as defined in Equations (2)–(5).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{F1-score} = \frac{2TP}{2TP + FP + FN} \tag{5}$$

Observe that the F1-score can be interpreted as a weighted average of the precision and recall to provide a more balanced measure in cases of unbalanced datasets.

Additionally, the Matthews Correlation Coefficient (MCC) was considered for evaluation. The MCC takes into account all four quadrants of the confusion matrix (TP, TN, FP, FN) and provides a balanced measure even if the classes are of significantly different sizes. Unlike the F1-score, which primarily focuses on the positive class and balances precision and recall, the MCC provides a comprehensive evaluation of the classifier's performance across both classes.

The MCC is defined as

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

The MCC ranges from -1 to +1, where +1 indicates a perfect prediction, 0 indicates random prediction, and -1 indicates no correct predictions. The confusion matrices for the classification have also been plotted and are available in the Supplementary Materials section at the end of this article.

After the classification was complete, the trained models are saved and used to perform SHAP and obtain some insights into how the model makes its decisions. This approach was particularly useful because it not only highlighted which features are influential but also how much they contributed to the final decision.

3. Results and Discussion

To evaluate the best segmentation result, both the model accuracy and its number of parameters were considered. Since both versions with the ResNet50 and MobileNetV3 Large backbones delivered comparable accuracy (0.9539 and 0.9484, respectively), the latter was preferred due to its much lower number of parameters (15,411,966 and 5,815,646, respectively). Figure 5 shows the results of the background removal process in three StyleGAN2 images, while Table 1 displays the segmentation performance metrics.

Table 1. DeepLabV3+ segmentation models' performance. The feature extraction backbone architecture used is reported together with the total number of trainable parameters and the pixel accuracy of the trained model on the test section of the complete Mut1ny dataset.

Backbone	Number of Parameters	Accuracy
Xception	74,803,174	**0.9601**
ResNet50	15,411,966	0.9539
MobileNetV3 Large	**5,815,646**	0.9484

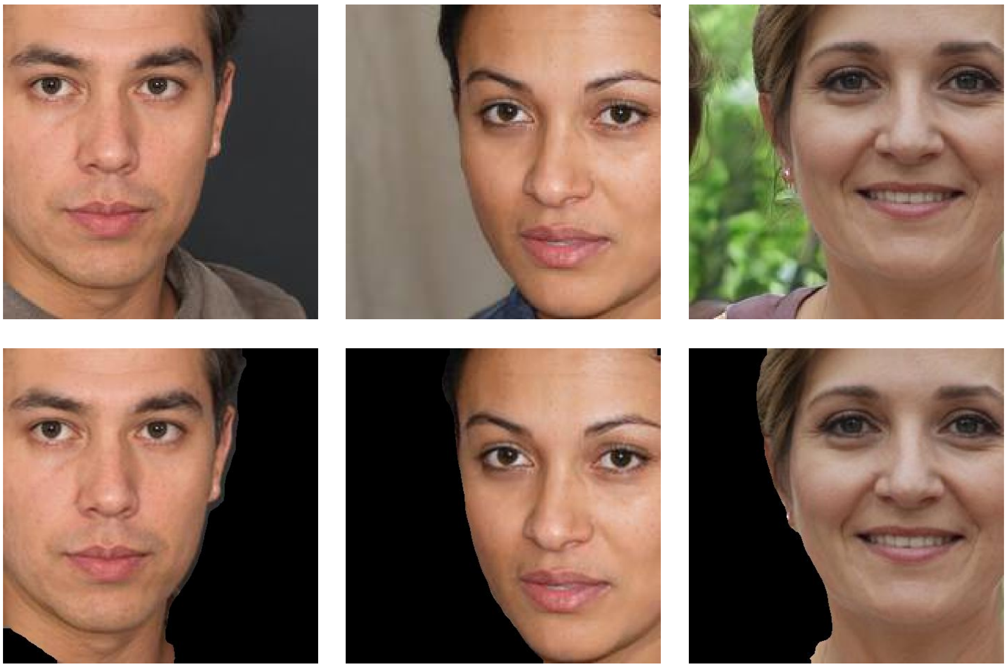

Figure 5. Application of the segmentation module to some StyleGAN2 generated images. The segmentation process provides finely cropped facial images, isolating the main subjects and blackening the background. The same approach is followed for both the generated and the real samples of the dataset, providing alternative and separate versions of the training samples to ensure that the model is only focusing on faces without the help of the background.

Table 2 displays the performance metrics for the classification tasks. These results suggest that for the detection of StyleGAN2-generated images, the background plays a crucial role, as its removal leads to a significant reduction in performance, confirming the initial hypothesis. However, this is not true with StyleGAN3, since, as Table 2 shows, the background seems to be a distracting factor for the model, weakening its ability to detect images obtained with this generator, and this is also highlighted by the drastic reduction in performance across all the metrics considered with respect to the first two experiments (this is likely also due to the much smaller number of generated samples gathered for StyleGAN3).

Finally, SHAP was applied to make the decisions of the model more interpretable. Figure 6 shows how important the background was for the classification task. In particular, the higher the Shapley values (red), the more useful those specific pixel coalitions were in confirming the decision made by the model. Conversely, lower Shapley values (blue) indicated that the corresponding areas were leading the model toward the opposite choice.

Table 2. Fake detection performance metrics for different generators. Notice how the background plays a significant role in StyleGAN2-generated images, aiding, with its presence, the decision process. This is not true for StyleGAN3.

Generator	Background	Performance Metrics				
		Accuracy	Precision	Recall	F1-Score	MCC
StyleGAN2	Yes	**0.9495**	**0.9562**	**0.9687**	**0.9624**	**0.8860**
	No	0.9022	0.9232	0.9307	0.9270	0.7790
StyleGAN3	Yes	0.8373	0.8314	0.6327	0.7186	0.6118
	No	**0.8499**	**0.8448**	**0.6648**	**0.7441**	**0.6492**

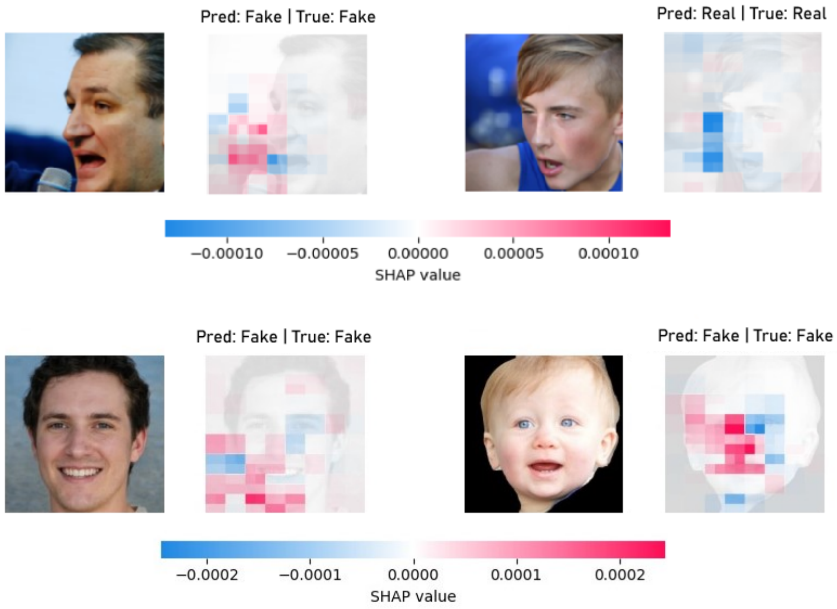

Figure 6. Application of SHAP to four test images. For each sample, the input image and the corresponding SHAP values for the coalitions of pixels are reported. Positive SHAP values (red coalitions) indicate that the group of pixels contributes positively toward the model's prediction, while negative SHAP values (blue coalitions), indicate a negative contribution. In many images, the background plays a crucial role in the decision, both in negative and positive ways (e.g., in the top right and bottom left figures, respectively), thus validating the initial claim and the classification results.

The most important area for the first picture in Figure 6 turned out to be the one at the bottom left, which includes a large portion of the background. On the other hand, in the last picture of Figure 6 the model focused much more in the central portion of the face, specifically on the mouth–nose–eyes region.

4. Conclusions

This paper introduces a novel approach for the detection of AI-generated images and shows the importance of the background for this task, leveraging the feature extraction capabilities of DeepLabV3+ equipped with a MobileNetV3 Large backbone. Multiple experiments were performed, first training the model on the Mut1ny dataset, which contained both real images and StyleGAN2-generated ones, and then repeating the training procedure on the dataset obtained after removing the background from the original images. These experiments showed that for StyleGAN2, the presence of background significantly aids in fake detection, suggesting that this generator does not provide effective and realistic backgrounds. However, the same conclusion does not hold for StyleGAN3-generated images, where the background seems to become a distracting factor rather than a help. The hypothesis of the authors is that the artefact reduction techniques integrated in StyleGAN3 are responsible for these observations. These enhancements could explain both the classifier's lower performance compared to that obtained on the StyleGAN2 dataset—together with the lower number of training samples—and the minimal difference in classification accuracy between images with and without backgrounds. The application of the SHAP explainability technique, although limited to a few images due to the high computational costs, showed that the presence or absence of background does have an impact on what areas of the image the model considers when classifying real and fake faces. It is a matter of future research to improve the presented model performance integrating both new real and generated images, gathering the latter also from a broader selection of generative models, and also expand the collection of SHAP-analyzed images to further confirm the importance of the background on a broader subset of the data. Since the current architecture has proven to be model dependent, employing a more comprehensive collection of generators would enable significantly higher robustness toward them. Another promising direction involves studying the behavior of the detectors against images with a modified background, e.g., faces which are generated and superimposed onto a real background. Both of these possible working branches aim toward a better understanding of how generative models work, thus providing a more interpretable way of detecting AI-generated images and a generally more trustworthy relationship with Artificial Intelligence.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/electronics13163161/s1 which contains the figures reporting the confusion matrices obtained for the four experiments. The code for reproducibility will be available and freely accessible online at https://github.com/mtanf/gng by 1 September 2024.

Author Contributions: Conceptualization, M.T. and E.G.C.; Data curation, M.T. and E.G.C.; Formal analysis, M.T., E.G.C. and S.M.; Funding acquisition, M.M.; Investigation, M.T., E.G.C. and N.P.; Methodology, M.T. and E.G.C.; Project administration, M.B.; Resources, M.T., E.G.C. and N.P.; Software, M.T., E.G.C. and N.P.; Supervision, M.B.; Validation, N.P., M.M. and M.B.; Visualization, M.T. and E.G.C.; Writing—original draft, M.T., E.G.C. and S.M.; Writing—review and editing, M.T., E.G.C., S.M., N.P., M.M. and M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets used in this study are publicly available and freely reusable. The CelebA-HQ dataset can be accessed via its TensorFlow library: https://www.tensorflow.org/datasets/catalog/celeb_a_hq; Mut1ny dataset: https://store.mut1ny.com/product/face-head-segmentation-dataset-community-edition?v=1c2903397d88; FFHQ dataset: https://github.com/NVlabs/ffhq-dataset; Artifact dataset: https://github.com/awsaf49/artifact. All accessed on 1 August 2024.

Acknowledgments: The authors gratefully acknowledge the financial support from the University of Siena for the publication of this article, which covered the publication fees.

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

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
GNG	Generated or Not Generated
GAN	Generative Adversarial Network
CNN	Convolutional Neural Network
MLP	Multilayer Perceptron
SHAP	SHapley Additive exPlanation

ArtiFact	Artificial and Factual
FFHQ	Flickr-Faces-HQ
SPP	Spatial Pyramid Pooling
ASPP	Atrous Spatial Pyramid Pooling
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

References

1. Karras, T.; Laine, S.; Aila, T. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4401–4410.
2. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 10684–10695.
3. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
4. Yamashita, R.; Nishio, M.; Do, R.K.G.; Togashi, K. Convolutional neural networks: An overview and application in radiology. *INsights Into Imaging* **2018**, *9*, 611–629. [CrossRef]
5. Li, Z.; Liu, F.; Yang, W.; Peng, S.; Zhou, J. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 6999–7019. [CrossRef] [PubMed]
6. LeCun, Y.; Boser, B.; Denker, J.S.; Henderson, D.; Howard, R.E.; Hubbard, W.; Jackel, L.D. Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1989**, *1*, 541–551. [CrossRef]
7. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
8. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
9. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
10. Monaci, M.; Pancino, N.; Andreini, P.; Bonechi, S.; Bongini, P.; Rossi, A.; Ciano, G.; Giacomini, G.; Scarselli, F.; Bianchini, M.; et al. Deep Learning Techniques for Dragonfly Action Recognition. In Proceedings of the ICPRAM, Valletta, Malta, 22–24 February 2020; pp. 562–569.
11. Pancino, N.; Graziani, C.; Lachi, V.; Sampoli, M.L.; Ştefănescu, E.; Bianchini, M.; Dimitri, G.M. A mixed statistical and machine learning approach for the analysis of multimodal trail making test data. *Mathematics* **2021**, *9*, 3159. [CrossRef]
12. Landi, E.; Spinelli, F.; Intravaia, M.; Mugnaini, M.; Fort, A.; Bianchini, M.; Corradini, B.T.; Scarselli, F.; Tanfoni, M. A MobileNet Neural Network Model for Fault Diagnosis in Roller Bearings. In Proceedings of the 2023 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Kuala Lumpur, Malaysia, 22–25 May 2023; pp. 1–6. [CrossRef]
13. Stefanescu, E.; Pancino, N.; Graziani, C.; Lachi, V.; Sampoli, M.; Dimitri, G.; Bargagli, A.; Zanca, D.; Bianchini, M.; Mureşanu, D.; et al. Blinking Rate Comparison Between Patients with Chronic Pain and Parkinson's Disease. *Eur. J. Neurol.* **2022**, *29*, 669.
14. Russo, V.; Lallo, E.; Munnia, A.; Spedicato, M.; Messerini, L.; D'Aurizio, R.; Ceroni, E.G.; Brunelli, G.; Galvano, A.; Russo, A.; et al. Artificial intelligence predictive models of response to cytotoxic chemotherapy alone or combined to targeted therapy for metastatic colorectal cancer patients: A systematic review and meta-analysis. *Cancers* **2022**, *14*, 4012. [CrossRef]
15. Lee, S.; Lee, J.; Lee, K.M. V-net: End-to-end convolutional network for object detection. *Expert Syst. Appl.* **2017**, *90*, 295–304.
16. Liang, J.; Li, N.; Sun, X.; Wang, X.; Liu, M.; Shi, J.; Huang, J.; Wang, D.Y. CIRL: Continuous imitation learning from human interaction with reinforcement learning in autonomous driving. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 4038–4052.
17. Chen, T.; Liu, S.; Yang, X.; Shen, J.; Hu, X.; Yang, G. Deepdriving: Learning affordance for direct perception in autonomous driving. In Proceedings of the AAAI Conference on Artificial Intelligence, Austin, TX, USA, 25–30 January 2015; Volume 29, pp. 2722–2728.
18. Bojarski, M.; Del Testa, D.; Dworakowski, D.; Firner, B.; Flepp, B.; Goyal, P.; Jackel, L.D.; Monfort, M.; Muller, U.; Zhang, J.; et al. End to end learning for self-driving cars. *arXiv* **2016**, arXiv:1604.07316.
19. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
20. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 834–848. [CrossRef]
21. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
22. Andreini, P.; Pancino, N.; Costanti, F.; Eusepi, G.; Corradini, B.T. A Deep Learning approach for oocytes segmentation and analysis. In Proceedings of the ESANN, Bruges (Belgium) and Online Event, 5–7 October 2022.

23. Radford, A.; Metz, L.; Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv* **2015**, arXiv:1511.06434.
24. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 2672–2680
25. Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; Aila, T. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8110–8119.
26. Karras, T.; Aittala, M.; Laine, S.; Härkönen, E.; Hellsten, J.; Lehtinen, J.; Aila, T. Alias-free generative adversarial networks. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 852–863.
27. Brock, A.; Donahue, J.; Simonyan, K. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv* **2019**, arXiv:1809.11096.
28. Wang, X.; Guo, H.; Hu, S.; Chang, M.C.; Lyu, S. Gan-generated faces detection: A survey and new perspectives. *ECAI* **2023**, *2023*, 2533–2542.
29. Le, T.N.; Nguyen, H.H.; Yamagishi, J.; Echizen, I. OpenForensics: Large-Scale Challenging Dataset For Multi-Face Forgery Detection And Segmentation In-The-Wild. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 10097–10107. [CrossRef]
30. Simonyan, K. Very deep convolutional networks for large-scale image recognition. *arXiv* **2015**, arXiv:1409.1556.
31. Gu, Q.; Chen, S.; Yao, T.; Chen, Y.; Ding, S.; Yi, R. Exploiting fine-grained face forgery clues via progressive enhancement learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 22 February–1 March 2022; Volume 36, pp. 735–743.
32. Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; Shao, J. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 86–103.
33. Songsri-in, K.; Zafeiriou, S. Complement Face Forensic Detection and Localization with FacialLandmarks. *arXiv* **2019**, arXiv:1910.05455.
34. Nadimpalli, A.V.; Rattani, A. Facial Forgery-Based Deepfake Detection Using Fine-Grained Features. In Proceedings of the 2023 International Conference on Machine Learning and Applications (ICMLA), Jacksonville, FL, USA, 15–17 December 2023; pp. 2174–2181. [CrossRef]
35. Liu, Z.; Qi, X.; Torr, P.H. Global texture enhancement for fake face detection in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8057–8066.
36. Chen, B.; Liu, X.; Zheng, Y.; Zhao, G.; Shi, Y.Q. A Robust GAN-Generated Face Detection Method Based on Dual-Color Spaces and an Improved Xception. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 3527–3538. [CrossRef]
37. Wang, R.; Juefei-Xu, F.; Ma, L.; Xie, X.; Huang, Y.; Wang, J.; Liu, Y. FakeSpotter: A simple yet robust baseline for spotting AI-synthesized fake faces. In Proceedings of the P Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI2020), Yokohama, Japan, 7–15 January 2021.
38. Liang, B.; Wang, Z.; Huang, B.; Zou, Q.; Wang, Q.; Liang, J. Depth map guided triplet network for deepfake face detection. *Neural Netw.* **2023**, *159*, 34–42. [CrossRef]
39. Howard, A.; Sandler, M.; Chen, B.; Wang, W.; Chen, L.C.; Tan, M.; Chu, G.; Vasudevan, V.; Zhu, Y.; Pang, R.; et al. Searching for MobileNetV3. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019; pp. 1314–1324. [CrossRef]
40. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4766–4775.
41. Mut1ny, J.D. *Face/Head Segmentation Dataset Commercial Purpose Edition*; Spriaeastraat: Den Haag, Netherlands, 2024.
42. Hassani, A.; Shair, Z.E.; Ud Duala Refat, R.; Malik, H. Distilling Facial Knowledge with Teacher-Tasks: Semantic-Segmentation-Features For Pose-Invariant Face-Recognition. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 741–745. [CrossRef]
43. Reimann, M.; Klingbeil, M.; Pasewaldt, S.; Semmo, A.; Trapp, M.; Döllner, J. Locally controllable neural style transfer on mobile devices. *Vis. Comput.* **2019**, *35*, 1531–1547. [CrossRef]
44. Khoshnevisan, E.; Hassanpour, H.; AlyanNezhadi, M.M. Face recognition based on general structure and angular face elements. *Multimed. Tools Appl.* **2024,** 1–19. [CrossRef]
45. Mut1ny, J.D. *Face/Head Segmentation Dataset Community Edition*; Spriaeastraat: Den Haag, Netherlands, 2024.
46. Rahman, M.A.; Paul, B.; Sarker, N.H.; Hakim, Z.I.A.; Fattah, S.A. Artifact: A large-scale dataset with artificial and factual images for generalizable and robust synthetic image detection. In Proceedings of the 2023 IEEE International Conference on Image Processing (ICIP), IEEE, Kuala Lumpur, Malaysia, 8–11 October 2023; pp. 2200–2204.
47. Wang, Z.; Zheng, H.; He, P.; Chen, W.; Zhou, M. Diffusion-GAN: Training GANs with Diffusion. *arXiv* **2023**, arXiv:2206.02262.
48. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. *arXiv* **2020**, arXiv:1703.10593.
49. Esser, P.; Rombach, R.; Ommer, B. Taming Transformers for High-Resolution Image Synthesis. *arXiv* **2020**, arXiv:2012.09841.
50. Xia, W.; Yang, Y.; Xue, J.H.; Wu, B. Tedigan: Text-guided diverse face image generation and manipulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2256–2265.

51. Liu, Z.; Luo, P.; Wang, X.; Tang, X. Deep Learning Face Attributes in the Wild. In Proceedings of the International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
52. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014; Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., Eds.; Springer: Cham, Switzerland, 2014; pp. 740–755.
53. Karras, T.; Aila, T.; Laine, S.; Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
54. Huang, X.; Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1501–1510.
55. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
56. Shapley, L.S. 17. A Value for n-Person Games. In *Contributions to the Theory of Games (AM-28), Volume II*; Kuhn, H.W., Tucker, A.W., Eds.; Princeton University Press: Princeton, NJ, USA, 1953; pp. 307–318. [CrossRef]
57. Schmeidler, D. The Nucleolus of a Characteristic Function Game. *SIAM J. Appl. Math.* **1969**, *17*, 1163–1170. [CrossRef]
58. Shapley, L.S.; Shubik, M. A Method for Evaluating the Distribution of Power in a Committee System. *Am. Political Sci. Rev.* **1954**, *48*, 787–792. [CrossRef]
59. Roth, A.E. Introduction to the Shapley value. In *The Shapley Value: Essays in Honor of Lloyd S. Shapley*; Roth, A.E., Ed.; Cambridge University Press: Cambridge, UK, 1988; pp. 1–28.
60. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 1135–1144.
61. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning important features through propagating activation differences. In Proceedings of the International Conference on Machine Learning. PMLR, Sydney, Australia, 6–11 August 2017; pp. 3145–3153.
62. 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] [PubMed]
63. Lipovetsky, S.; Conklin, M. Analysis of regression in game theory approach. *Appl. Stoch. Model. Bus. Ind.* **2001**, *17*, 319–330. [CrossRef]
64. Štrumbelj, E.; Kononenko, I. Explaining prediction models and individual predictions with feature contributions. *Knowl. Inf. Syst.* **2014**, *41*, 647–665. [CrossRef]
65. Datta, A.; Sen, S.; Zick, Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In Proceedings of the 2016 IEEE Symposium on Security and Privacy (SP), IEEE, San Jose, CA, USA, 22–26 May 2016; pp. 598–617.
66. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.

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