ORIGINAL RESEARCH
published: 08 July 2022
doi: 10.3389/frsip.2022.918725

An Eyes-Based Siamese Neural Network for the Detection of GAN-Generated Face Images

*Jun Wang *, Benedetta Tondi and Mauro Barni*

Department of Information Engineering and Mathematics, University of Siena, Siena, Italy

Generative Adversarial Network (GAN) models are nowadays able to generate synthetic images which are visually indistinguishable from the real ones, thus raising serious concerns about the spread of fake news and the need to develop tools to distinguish fake and real images in order to preserve the trustworthiness of digital images. The most powerful current detection methods are based on Deep Learning (DL) technology. While these methods get excellent performance when tested under conditions similar to those considered for training, they often suffer from a lack of robustness and generalization ability, as they fail to detect fake images that are generated by "unseen" GAN models. A possibility to overcome this problem is to develop tools that rely on the semantic attributes of the image. In this paper, we propose a semantic-based method for distinguishing GAN-generated from real faces, that relies on the analysis of inter-eye symmetries and inconsistencies. The method resorts to the superior capabilities of similarity learning of extracting representative and robust features. More specifically, a Siamese Neural Network (SNN) is utilized to extract high-level features characterizing the inter-eye similarity, that can be used to discriminate between real and synthetic pairs of eyes. We carried out extensive experiments to assess the performance of the proposed method in both matched and mismatched conditions pertaining to the GAN type used to generate the synthetic images and the robustness of the method in presence of post-processing. The results we got are comparable, and in some cases superior, to those achieved by the best performing state-of-the-art method leveraging on the analysis of the entire face image.

Keywords: deep learning for forensics, semantic analysis, siamese networks, synthetic media detection, generative adversarial networks, image forensics

OPEN ACCESS

Edited by:
Roberto Caldelli,
Consorzio Nazionale Interuniversitario
Per Le Telecomunicazioni, Italy

Reviewed by:
Cecilia Pasquini,
University of Trento, Italy
Alireza Sepas-Moghaddam,
Queen's University, Canada

****Correspondence:***
Jun Wang
j.wang@student.unisi.it

Specialty section:
This article was submitted to
Image Processing,
a section of the journal
Frontiers in Signal Processing

Received: *12 April 2022*
Accepted: *03 June 2022*
Published: *08 July 2022*

Citation:
Wang J, Tondi B and Barni M (2022)
An Eyes-Based Siamese Neural
Network for the Detection of GAN-
Generated Face Images.
Front. Sig. Proc. 2:918725.
doi: 10.3389/frsip.2022.918725

1 INTRODUCTION

Since the emergence of Generative Adversarial Networks (GANs) (Goodfellow et al., 2020), research on image manipulation using GANs had seen tremendous advances (Creswell et al., 2018). The generated images have become extremely realistic, very high quality, and can easily deceive a human observer. Furthermore, building a fake image is nowadays an easy task, hence very realistic fakes can be generated by non-expert users. In addition to image editing, modification of attributes and style transfer, GANs are used to create completely synthetic images from scratch. A website called *This person does not exist* (link: http://thispersondoesnotexist/) allows anyone to generate a synthetic face image in very few seconds. Alongside benign uses of this technology, the possible misuse of the synthetic contents generated by GANs represents a serious threat calling for the development of image forensic techniques capable to distinguish between real and fake (GAN-generated) images.

Several methods have been proposed in the forensic literature to reveal whether an image has been generated by a GAN or it is a natural one. Most recent methods are based on Deep Learning (DL), which can significantly outperform feature-based methods (Verdoliva, 2020). Despite the excellent performance of these methods when tested under conditions that are similar to those considered for training, they fail to generalize to images that are generated by different unseen models (Gragnaniello et al., 2021). In many cases, the robustness against processing is also poor. This problem is exacerbated by the fact that DL-based solutions often lack interpretability, so a clear understanding of what the network is seeing and a solid explanation of why decisions are made is not possible.

Relying on semantic attributes of the image can help to devise more general, robust and explainable tools. As pertains to the specific task of GAN face image detection, the eyes can provide relevant clues. Recent literature has shown that some artifacts, like the color difference between the two eyes (Matern et al., 2019) and the inconsistencies of the corneal specular highlights (Hu et al., 2021) can be a spot to reveal the GAN nature of the face images. Besides these simple artifacts, there can be more sophisticated forensic traces that can be revealed through eyes inspection, e.g., in the iris pattern, which can lead to the development of a more general and robust GAN face detector.

In this paper we propose a semantic-based method for GAN-generated face image detection that relies on the analysis of inter-eye symmetries and inconsistencies and resorts to the superior capabilities of similarity learning of extracting robust features from the images. The proposed method relies on the underlying assumption that GANs are not perfect in reconstructing symmetries between the eyes and then it is possible to look at the presence of inconsistencies between the patterns in the left and right eyes to detect if the image is real or fake. In the proposed architecture, two identical branches of Siamese Neural Networks (SNN) are fed with the right and left eye. The purpose of these two branches is to extract high-level features, characterizing the inter-eye similarity, that permit to discriminate between real and synthetic pairs of eyes. A modified version of the XceptionNet architecture is used as the backbone network for the two branches.

We carried out an extensive experimental campaign to assess the performance of the proposed method, both in matched and mismatched conditions pertaining to the GAN type used to generate the synthetic images. We also carried out robustness tests to assess the performance of the method in presence of global post-processing and local manipulations. The robustness performance of the tool is also assessed in the challenging scenario of rebroadcast attacks. Given that rebroadcast attacks are very effective in hiding *any* type of forensic trace from the image, addressing the capability of our tool of discriminating between re-acquired GAN and real images allows getting insights into the degree of robustness of the method.

Our experiments reveal that, by relying on semantic-related features, the proposed method achieves not only very good performance on StyleGAN2, that is, the same generation method used for training, but also very good generalization capability, when the fake is obtained using a different GAN architecture, namely, the ProGAN (Karras et al., 2018) and the new StyleGAN3 (Karras et al., 2021) model. In particular, the performance we got is comparable to those achieved by the best performing state-of-the-art method in (Gragnaniello et al., 2021), which relies on the analysis of the entire image, with our method having overall better generalization capability to unseen GAN models and better robustness against the rebroadcast attack. Moreover, the characteristic of relying on the analysis of a semantic attribute of the image, and not the entire image, makes our method naturally more robust against local manipulation compared to (Gragnaniello et al., 2021).

The paper is organized as follows: after an overview of the state-of-the-art in **Section 2**, in **Section 3** we ddsludsescribe the proposed method. Then, **Section 4** details the methodology that we followed for the experiments, whose results are reported in **Section 5**. We conclude the paper with some final remarks in **Section 6**.

2 RELATED WORK

In this section, we review the most relevant approaches for GAN face synthesis. Then, we discuss the methods that have been developed in the literature for the detection of GAN-generated faces.

2.1 Synthetic Face Generation

In the past recent years, many GAN models have been developed for face image generation and manipulation. Face editing models modify the attributes of a face such as the colour of the hair or the skin, gender, and age, see for instance IcGAN (Perarnau et al., 2016) and attGAN (He et al., 2019). Besides face editing, a line of research focuses on the development of GAN models for face to face translation. This is the case of StarGAN (Choi et al., 2018), where a scalable approach is implemented to get high quality translated face images. Another line of research focuses on generating face images from random noise. In this case, the face images are synthesized from scratch.

While early attempts could only generate high quality low resolution images (Radford et al., 2016; Zhao et al., 2017; Berthelot et al., 2017), more recent models have been proposed that can generate realistic high quality large resolution images, up to 1,024 × 1,024. In particular, ProGAN (Karras et al., 2018) was the first generative model able to synthesize high resolution (1,024 × 1,024) images by exploiting a progressive training approach. StyleGAN (Karras et al., 2019) further improved the quality of the generated high resolution images by utilizing an alternative generator architecture, borrowed from style transfer literature. In addition to progressively increasing the resolution of the generated images, StyleGAN incorporates 'style' features in the generative process. The quality of StyleGAN images has been further improved by the StyleGAN2 model (Karras et al., 2020), that redesigns the normalization used in the generator.

Very recently - just a few months ago - a new GAN architecture was released by NVIDIA, named StyleGAN3 (Karras et al., 2021), which solves the problem of "texture sticking" (a.k.a, aliasing) in the images generated by StyleGAN2, performing architectural changes that guarantee that unwanted information does not leak into the hierarchical synthesis process. StyleGAN3 impressively improves the quality of synthetic images, paving the way for generative models better suited for video and animation. Despite the similar name, the StyleGAN3 architecture is very different from StyleGAN2 (which is similar to the original StyleGAN), and, in fact, the two architectures have been shown to learn different internal representations (Karras et al., 2021).

2.2 GAN Face Detection

Several methods have been proposed in the literature to discriminate between GAN generated faces and real ones. Some early approaches exploited very specific facial traces, e.g., irises color or borders of face, that are left behind by the earlier GAN architectures (Matern et al., 2019). In Yang et al. (2019), the authors showed that it is possible to reveal whether a face image is GAN generated or not by considering the locations of facial landmark points and using them to train an SVM. A line of research exploits color information to reveal GAN contents (McCloskey and Albright, 2019; Li et al., 2020). In McCloskey and Albright (2019), two metrics based on the correlation between color channels and saturation are defined and considered for the detection. The approach in Li et al. (2020), instead, combines color channel analysis and Subtractive Pixel Adjacency Matrix (SPAM)-like features to perform GAN image detection. In particular, co-occurrences are extracted from color channels and combined into a feature vector used to train an SVM. Leveraging on the superior performance of Convolutional Neural Networks (CNNs), in Nataraj et al. (2019) the authors have shown that improved performance can be achieved by feeding a CNN with co-occurrence matrices computed directly on the input image. The work has been extended in Barni et al. (2020), where cross-band co-occurrence matrices have also been considered as input to the CNN.

Besides the methods in Nataraj et al. (2019) and Barni et al. (2020), a wide variety of techniques based on CNNs have been proposed (Marra et al., 2018; Marra et al., 2019; Hsu et al., 2020; Hulzebosch et al., 2020), that achieved improved performance compared to previous methods based on standard machine learning and hand-crafted features. Most of these methods are fully supervised approaches. Notably, in Marra et al. (2018), off-the-shelf very deep modern networks, like Xception, Inception and DenseNet, pre-trained on ImageNet, are shown to achieve excellent performance for GAN detection, by directly training on the pixel image.

The fully supervised approaches are all very effective under matched conditions between training and testing, and in particular when the test GAN images are from the same model considered during training. However, they often fail to generalize to images generated by different, unseen, GAN types. In Liu et al. (2020), the authors presented a new CNN-based detector, called Gram-Net, that leverages global image texture representations to improve the generalization and the robustness of GAN image detection. An approach that relies on semantic segmentation and perform detection based on multiple semantic fragments getting remarkable generalization capability has been recently presented in Chen and Yang (2021). A different solution to improve generalization is proposed in Xuan et al. (2019): the idea is to carry out augmentation by Gaussian blurring so as to force the discriminator to learn more general features. A similar approach is followed in Wang et al. (2020), wherein a standard pre-trained model, namely ResNet50, is further trained with a strong augmentation based on compression and blurring. Experiments show that, by training on a single GAN architecture, the learned features can generalize to unseen architectures, datasets, and training methods. The generalization performance is further improved by the method in Gragnaniello et al. (2021), where successful augmentation and training strategies, as well as key architectural design choices (among them, in particular, the insertion of an initial layer for residual extraction and the removal of the down-sampling operation in the first layers), are identified.

3 THE PROPOSED EYES-BASED GAN-GENERATED FACE IMAGE DETECTOR

In this section we describe the proposed Eyes-based GAN-generated face detector. As we mentioned, the goal of our method is to distinguish GAN images from real images by exploiting dissimilarities and inconsistencies between the eyes in GAN synthesized faces, that are not present in real faces. To address this binary decision problem, we decided to resort to an SNN architecture and exploit the excellent capabilities of these structures in finding similarities between paired inputs. SNNs have been widely used in the related fields, e.g. in the field of face verification (Hu et al., 2014; Guo et al., 2021), person re-identification (Varior et al., 2016; Fang et al., 2019), and even object tracking (Bertinetto et al., 2016; Son et al., 2017) with very good results.

In addition, it is known that, by relying on the similarity learning paradigm, in many cases, SNN can improve the generalization capability of the models on unseen data, since they tend to learn more robust features (Krishnagopal et al., 2018; Roth et al., 2020; Agarwal et al., 2021; Zhou et al., 2021; Fonseca and Guidetti, 2022).

The scheme of the proposed Siamese Eyes-based detector of GAN-generated faces is provided in **Figure 1**. The detector consists of three modules: a pre-processing module, the feature extraction module and the final classifier (these two last modules building the SNN). The purpose of the pre-processing module is to localize the eyes within the face and extract the two bounding boxes of the eyes that constitute the paired input of the SNN. More specifically, the Dlib (King, 2009) face detector is used to locate the face, followed by a landmark predictor that outputs 12 landmark points (6 feature points per eye) whose coordinates indicate the locations of the left and right eyes. The bounding boxes (BB) of the left eye E_l and right eye E_r are cropped exploiting the coordinates of feature points for each eye. Then, the BBs are paired and fed to the two branches of the feature extraction module, after resizing them to the same fixed size

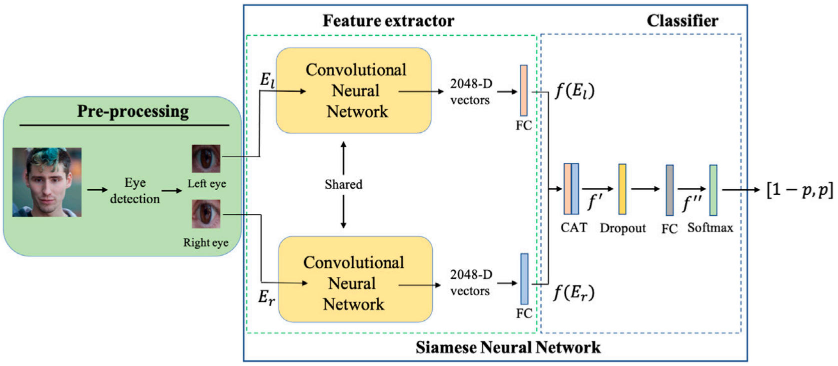

FIGURE 1 | The proposed Siamese Eyes-based detector of GAN-generated face images.

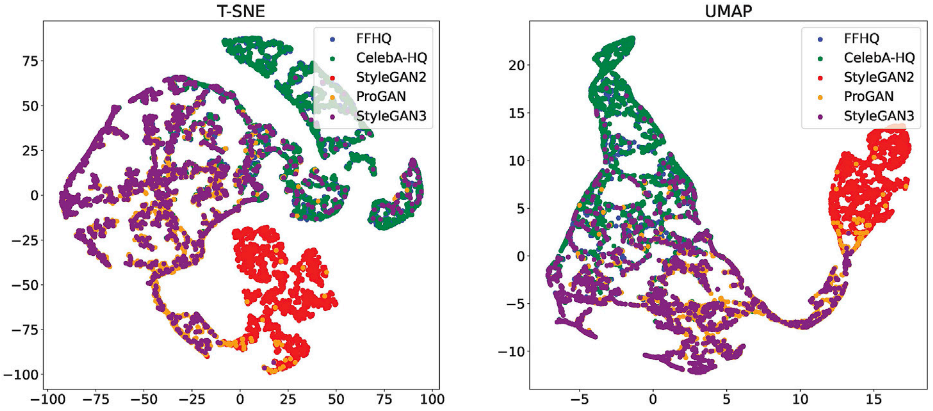

FIGURE 2 | Feature space distribution for each dataset using T-SNE and UMAP reduction methods. 2,000 images per each dataset have been considered.

(the input network size). We denote with $X = (E_l, E_r)$ the input of the SNN.

The feature extractor is based on a modified version of the XceptionNet architecture (Chollet, 2017). XceptionNet is a particular version of Inception network (Szegedy et al., 2016), that relies on a modified depthwise separable convolution. Such a network has been proved to achieve great performance for the deepfake detection (Rossler et al., 2019). Inspired by the work in (Barni et al., 2018), in order to retain as much spatial information as possible (which is especially relevant in the presence of strong processing and JPEG compression), we remove the sampling operation in the first convolutional layer of the network, setting the stride parameter to 1. In addition, following (Basha et al., 2020), we replaced the 1000-dim FC layer in the original network with an FC layer of size 512. Then, the FC layer takes as input the 2048-dim feature vector obtained by the final Global Average Pooling (GAP) layer of the convolutional part and outputs a 512-dim feature vector. The parameters of the two modified XceptionNet branches extracting the features from the left and right eye respectively are shared.

The two 512-dim feature vectors from the two branches, namely $f\ (E_l)$ and $f\ (E_r)$, are then concatenated to get a 1024-D feature vector f'. Then, $f' = CAT(f\ (E_l)f\ (E_r))$, where CAT denote the concatenation operation. To reduce the overfitting, the concatenation is followed by a dropout layer, where the nodes are dropped out with a probability of 0.5. The output of the dropout layer goes as input to another FC layer, with 2 output nodes, namely, the number of classes. Finally, a softmax layer is applied to the output of the FC f'', in order to get the output probability vector, characterizing the probability of the output being real and GAN respectively. We denote with p the probability that the input is GAN: the input is deemed GAN if $p >$ 0.5, pristine otherwise. Then, the two output probability scores are $p_0 = 1 - p$ and $p_1 = p$, where we use the convention label 0 for pristine and label 1 for GAN. Without loss of generality, in the following, we refer to the GAN class as the positive class and to the real class as the negative class.

It is proper to mention that the idea behind the proposed method of exploiting eyes clues to perform GAN detection is not totally new, and has already been considered in the literature, see (Hu et al., 2021). The technique in (Hu et al., 2021) is based on statistical hand-crafted features manually extracted from the eyes, and relies on an estimation of the corneal region. However, due to the limitations inherent in the process of estimation, the method lacks robustness and generality, and eye localization can be successfully achieved only in not-so-difficult cases (high contrast and good illumination conditions in the eyes region). By resorting to a simple bounding-box extraction procedure and exploiting the superior capabilities of SNNs to learn discriminative features, the proposed method is instead very general and can get robustness and generalization results that are comparable and, in some cases, superior, to the state-of-the-art for GAN-generated face image detection.

4 METHODOLOGY

In this section, we discuss the datasets used for the experiments and report the setting of the training of our Eyes-based GAN detection model. The state-of-the-art GAN detection method that we considered as the baseline for the comparison is also presented. Finally, we discuss the metrics used for evaluating the performance.

4.1 Datasets

The datasets used for our experiments are described in the following.

To create the synthetic images, we considered several generative models. Training is performed considering only the StyleGAN2 model (Karras et al., 2020), while the ProGAN (Karras et al., 2018) and StyleGAN3 (Karras et al., 2021) are considered, in addition to StyleGAN2, to generate the images used for the tests.

As we pointed out in **Section 2.2**, despite the similar name, the architecture of StyleGAN3 is very different from the architecture of StyleGAN2 (which is similar to the original version of the StyleGAN model), thus making non-trivial the generalization of StyleGAN2 detectors to the case of synthetic images generated by StyleGAN3. More in detail, the following datasets of faces are considered in our experiments.

- A collection of 100,000 real face images: 30,000 images are taken from the CelebA-HQ dataset (Karras et al., 2017) while the remaining 70,000 images come from the FFHQ dataset (Karras et al., 2019)[1]. Then, 90,000 images are used for training (5,000 of those are left for validation), and 10,000 images for the tests. The same 3:7 proportion of CelebA-HQ and FFHQ images are considered in training and test datasets. Therefore, among the images in the training set, 27,000 images come from CelebA-HQ and 63,000 from FFHQ, while the test set consists of the remaining 3,000 CelebA-HQ and 7,000 FFHQ images.
- A dataset of StyleGAN2 fake images consists of 100,000 images in total, where 90,000 images are used to train the model (85,000 training and 5,000 validation), and 10,000 images are used for the tests. We use the official released code[2] to generate synthetic faces with different generation parameters in order to increase the diversity of the dataset. More specifically, we considered several values of the truncation parameter of the network and generated 10,000 StyleGAN2 (Karras et al., 2020) faces for each value in the set \{0.3, 0.4, 0.5, 0.6, 0.65, 0.7, 0.75, 0.8, 0.9, 1\}. Both the training and test sets contain images generated with all the truncation parameters in equal proportions.
- A collection of 10,000 images generated by ProGAN and 10,000 images generated by StyleGAN3, was used for the tests.
- A large-scale dataset of printed and scanned pristine and GAN images, called VIPPrint (Ferreira et al., 2021), consisting of 10.000 recaptured real (from the FFHQ dataset) and 10.000 recaptured GAN (StyleGAN2) images, obtained by printing the digital images and then scanning them. More details can be found in Ferreira et al. (2021). The printing and scanning operation can be used to hide the traces of image manipulation, then, arguably, also the synthetic nature of images. In the following, we refer to this dataset as Print&Scan image dataset.

A summary of the datasets used in our experiments is provided in **Table 1**.

4.2 Training Setting

The input size of the two branches of the SSN is set to 66 × 100 × 3. Rescaling is performed as a pre-processing step to the bounding boxes of the eyes, in order to resize them to the input size. The same rescaling is also applied to the images during testing.

The network was trained using the cross entropy loss function, with the Adam optimizer (Kingma and Ba, 2014), with default parameters ($\beta_1 = 0.9$ and $\beta_2 = 0.999$). The batch size was set to 64. Training is performed with a constant learning rate of 0.0001 for 10 epochs.

A strong level of augmentation is considered during training in the attempt to enhance the robustness of the model, including JPEG compression - with quality factors ranging from 40 to 100, flipping, scaling - with the scale factors in [0.8,1.3], contrast and brightness adjustment, and Gaussian blurring. The above processing operations are applied with probabilities 0.8, 0.1, 0.8, 0.8, and 0.2, respectively for JPEG compression, flipping, scaling, contrast adjustment and Gaussian blur. The parameters of the convolutional neural networks in the two branches are initialized using the solution pre-trained on ImageNet dataset.

The framework is implemented in the *Python*3 environment using Tensorflow library. An NVIDIA GTX2080Ti GPU was utilized for model training and testing.

4.3 Comparison With the State-of-the-Art

In order to demonstrate the effectiveness of our method, comparison is carried out with the current best performing method for GAN detection from the state-of-the-art, that is the method in Gragnaniello et al. (2021). The method implements a series of strategies that are experimentally proven to be effective to enforce learning of features that can generalize better to unseen GAN types. Among the most relevant strategies implemented by the method: 1) an initial layer is added

[1] https://github.com/NVlabs/ffhq-dataset

[2] https://github.com/NVlabs/stylegan2

TABLE 1 | Datasets used for training and testing.

Datasets	CelebA_HQ	FFHQ	StyleGAN2	ProGAN	StyleGAN3	Print&Scan
Class	Pristine	Pristine	GAN	GAN	GAN	Pristine,GAN
Training	63,000	27,000	90,000	—	—	—
Test	3,000	7,000	10,000	10,000	10,000	10,000, 10,000

for residual extraction; 2) down-sampling and pooling in the first layers are removed; and 3) a specific chain of augmentations is implemented that includes Gaussian noise addition, geometric transformations, brightness and contrast changes. Notably, in the proposed structure, the first layer is an adaptive layer, that automatically adapts to the size of the input. In this way, the image is entirely fed as input to the network, without performing any resizing. The best results are achieved using ResNet50 as the backbone network. For the details of the training, we refer to (Gragnaniello et al., 2021). The method improves previous methods, e.g., the method in Marra et al. (2018) and Wang et al. (2020). We used the model released by the author[3] to carry out the tests. In the following, we refer to this method as ResNet50-NoDown.

As mentioned in **Section 2.2**, a GAN detection technique that has close ties with the proposed approach is the method in (Hu et al., 2021), which, similarly to the proposed method, exploits eyes clues to perform GAN detection. The method relies on statistical hand-crafted features. Given that the performance of this detector are much lower than those achieved by ResNet50-NoDown, the results of these methods are not reported for comparison in **Section 5**. In particular, the method in Hu et al. (2021) suffers from the poor performance of the estimation of the corneal region, which does not provide accurate localization of the eye region in many cases. With reference to the datasets used in our experiments, accurate localization can be achieved only in 29% of the CelebA-HQ test images, 48% of the FFHQ, and 55% of the StyleGAN2. The development of a robust method for corneal extraction is in fact still an open research problem.

4.4 Metrics

With regard to the metrics used for evaluating the performance, we consider the True Positive Rate (TPR) and True Negative Rate (TNR) of the decision made by the SNN network (0 when $p < 0.5$, 1 otherwise), where the positive event refers to the output being GAN (label 1) while the negative event refers to the output being real (label 0). Since we always consider balanced sets, the accuracy can be measured as ACC = (TPR + TNR)/2. We also report the Area Under Curve (AUC) of the Receiver Operating Curve (ROC) of the classification, measuring the discrimination capabilities of the method and providing an indication of the best performance that can be achieved on the test set by adjusting the decision threshold. A more practical measure is evaluated by measuring the probability of correct detection at a fixed False Positive Rate (FPR) of 5%, indicated as Pd@5%. Specifically, we use the pristine images in the validation set (5,000 images) to set the threshold of the SNN-based detector, by fixing the false positive error of the decision at 5%. The detection performance are then evaluated on the test set using this threshold. Both raw (uncompressed) and JPEG compressed images are considered to determine the threshold, to get a general operating point for the detector. More specifically, the 5,000 images in the pristine validation set are compressed with quality factors \{70, 80, 90,100\}, for a total of 25,000 images used to set the threshold.

5 EXPERIMENTAL RESULTS

We conducted several experiments to assess the performance of the proposed Eyes-based GAN-generated facial image detector, particularly in terms of robustness and generalization. The experiments confirm that, thanks to the use of semantic-related features and the similarity learning paradigm, robust and general features are learned by the network.

5.1 Performance Analysis, Generalization and Robustness

The results of the proposed method in matched and mismatched conditions, that is, when the fake images generated with the same and different GAN models, are reported in **Table 2**, where the TNR/TPR, the AUC, and the Pd@5% are reported for both the proposed and baseline method. For the Pd@5%, the FPR measured on the test set is also reported among brackets. Since the TNR refers to the pristine class, the TNR values are the same in all the columns.

Both methods achieve perfect detection results on StyleGAN2 (TPR = 100%). Our method is the one that can achieve the best overall generalization results. In particular, without threshold adjustment, it achieves TPR = 84.4% on StyleGAN3 and TPR = 87.2% on ProGAN. For the ResNet50-NoDown, the method can get perfect results on ProGAN (with a gain of 12.8% in the TPR with respect to our method), however, it can not generalize to StyleGAN3, in which case TPR = 1.1%, even if the AUC is good. The Pd@5% is also poor, being equal to 28.0%. These results indicate that the ResNet50-NoDown method can not work on StyleGAN3 without re-calibrating the tool on the same StyleGAN3 images used for testing.

In **Figure 2**, we visualize the distribution in the feature space for each image dataset for our method. Dimensionality reduction is performed to a 2-dim space by means of T-SNE (Van der Maaten and Hinton, 2008) and UMAP (Mclnnes et al., 2018) technique, shown respectively in the left and right plot. The separability of the pristine

[3] https://github.com/grip-unina/GANimageDetection

TABLE 2 | TPR/TNR (%), AUC (%) and Pd@5% (%) of the proposed method and the baseline on unprocessed images. Tests are carried out in matched (StyleGAN2) and mismatched (ProGAN and StyleGAN3) conditions.

Processing type	ResNet50-NoDown			Ours		
	TPR/TNR	AUC	Pd@5% (FPR)	TPR/TNR	AUC	Pd@5% (FPR)
StyleGAN2	100/100	100	100 (0)	100/100	100	100 (2.7)
ProGAN	100/100	100	100 (0)	87.2/100	99.7	97.2 (2.7)
StyleGAN3	1.1/100	100	28.0 (0)	84.4/100	99.6	96.7 (2.7)

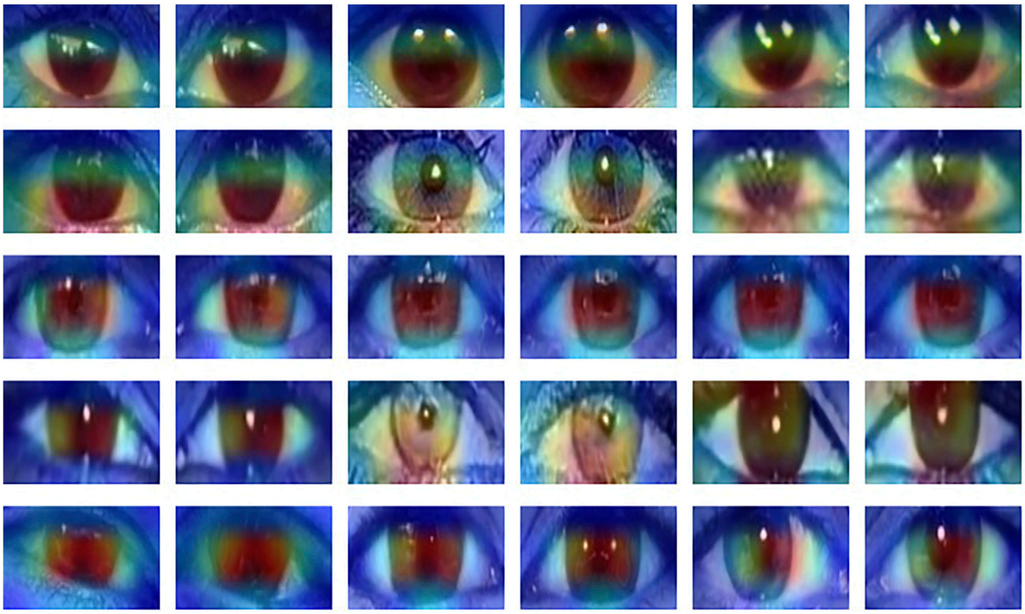

FIGURE 3 | GradCAM visualization for the proposed detector. From top to bottom row: FFHQ, CelebA-HQ, StyleGAN2, ProGAN and StyleGAN3. In each row, 3 sample pairs are visualized (left and right eyes).

TABLE 3 | TPR/TNR (%) and AUC (%) of the proposed method and the baseline under various image processing operations. Tests are carried out in matched (StyleGAN2) and mismatched (ProGAN and StyleGAN3) conditions.

Processing type	StyleGAN2				ProGAN				StyleGAN3			
	ResNet50-NoDown		Ours		ResNet50-NoDown		Ours		ResNet50-NoDown		Ours	
	TPR/TNR	AUC	TPR/TNR	AUC	TPR/TNR	AUC	TPR/TNR	AUC	TPR/TNR	AUC	TPR/TNR	AUC
JPEG100	100/100	100	100/100	100	100/100	100	82.0/100	99.6	1.9/100	100	79.3/100	99.5
JPEG90	100/100	100	100/100	100	99.4/100	100	69.1/100	98.7	2.1/100	100	69.0/100	98.8
JPEG80	100/100	100	100/100	100	94.3/100	100	55.8/100	97.0	10.0/100	99.6	61.1/100	97.7
JPEG70	100/100	100	100/100	100	92.1/100	100	47.4/100	94.6	10.1/100	97.6	55.2/100	96.3
Gaussian Noise	70.2/74.0	70.8	41.1/90.0	84.6	71.3/74.0	80.4	14.0/90.0	69.9	55.1/74.0	70.8	8/90.0	46.6
Resize- 2	100/100	99.8	99.3/100	100	100/100	100	85.4/100	99.6	1/100	99.8	84.2/100	99.9
Resize- 1.3	100/100	99.8	100/100	100	100/100	100	85.8/100	99.7	1/100	99.8	83.3/100	99.6
Resize - 0.5	92.1/100	97.0	100/99.2	100	100/100	100	85.4/99.2	99.4	3.1/100	97.0	71.2/99.2	98.3
Gaussian blur- 3 × 3	100/100	99.9	100/99.3	100	100/100	100	72.2/99.3	98.6	5.4/100	99.9	73.3/99.3	98.6
Gaussian blur -5 × 5	100/100	99.9	100/99.3	100	100/100	100	63.1/99.3	96.9	8.9/100	99.9	65.0/99.3	97.2
Median filter -3 × 3	88.1/100	65.0	99.0/99.1	100	99.1/100	100	74.4/99.1	97.9	9.3/100	84.0	35.2/99.1	91.4
Contrast enhancement	99.9/100	100	96.1/98.4	99.7	100/80.4	100	60.4/98.4	96.5	0.01/100	94.3	53.7/98.4	94.7

and GAN classes is good, with only some overlap between pristine and the StyleGAN3 images, not considered for training. Interestingly, StyleGAN2 and StyleGAN3 images get clustered separately both with the t-SNE and UMAP reduction techniques, while the distribution of ProGAN overlaps with them. As expected, the pristine images from FFHQ and CelebA-HQ gets clustered together.

TABLE 4 | Pd@5% (FPR) for the proposed method and the baseline under various image processing operations. The FPR (%) measured on the test set is reported among brackets. Tests are carried out in matched (StyleGAN2) and mismatched (ProGAN and StyleGAN3) conditions.

Processing type	StyleGAN2		ProGAN		StyleGAN3	
	ResNet50-NoDown	Ours	ResNet50-NoDown	Ours	ResNet50-NoDown	Ours
JPEG100	100 (0)	100 (2.7)	100 (0)	96.3 (2.7)	14.0 (0)	96.1 (2.7)
JPEG90	100 (0)	100 (3.6)	100 (0)	92.0 (3.6)	96.2 (0)	93.3 (3.6)
JPEG80	100 (4.1)	100 (5.9)	100 (4.1)	85.4 (5.9)	98.1 (4.1)	89.4 (5.9)
JPEG70	100 (18.7)	100 (8.0)	100 (18.7)	79.1 (8.0)	100 (18.7)	87.4 (8.0)
Gaussian Noise 0.01	94.2 (54.7)	61.1 (12.8)	93.4 (54/7)	25.2 (12.8)	88.2 (54.7)	2 (12.8)
Resize - 2	100 (0)	100 (3.0)	100 (0)	97.3 (3.0)	26.0 (0)	97.2 (3.0)
Resize - 1.3	100 (0)	100 (2.6)	100 (0)	97.2 (2.6)	23.0 (0)	97.3 (2.6)
Resize - 0.5	99.2 (0)	100 (6.1)	99.2 (0)	98.3 (6.1)	25.0 (0)	94.4 (6.1)
Gaussian blur- 3 × 3	100 (0)	100 (6.0)	100 (0)	93.3 (6.0)	55.4 (0)	94.3 (6.0)
Gaussian blur - 5 × 5	100 (0)	100 (9.0)	100 (0)	90.1 (9.0)	67.3 (0)	92.2 (9.0)
Median filter - 3 × 3	100 (21.1)	100 (9.3)	100 (21.1)	93.1 (9.3)	79.3 (21.1)	72.3 (9.3)
Contrast enhancement - 1.5	100 (0)	100 (13.8)	100 (0)	93.2 (13.8)	10.0 (0)	89.2 (13.8)

Figure 3 shows some examples of attention maps obtained for our method with the GradCAM algorithm (Selvaraju et al., 2017). The activation maps reveal that our Siamese network indeed looks the eyes region of the bounding-boxes to take the decision, confirming the explainability and the semantic nature of the tool. In particular, we see that the attention is focused on the iris region of the eyes.

The performance in the presence of processing, that is, when the real and fake images are subject to post-processing operations, are reported in **Table 3**, in terms of TNR/TPR and AUC, and in **Table 4**, in terms of Pd@5%, for several types of processing and processing strength. For the case of Gaussian noise addition, the parameter we report in the table is the variance of the noise. For the resize, the scale factor is reported. For the case of Gaussian blurring, the parameters refer to the size of the Gaussian kernel, while for the median filtering, it refers to the window size. Finally, for the contrast enhancement, the image contrast is increased by a factor of 1.5. All these processing operations correspond to global manipulations of the image since they affect all the pixels of the image. A local manipulation is considered in **Section 5.2.1**.

Looking at the AUC results and at the Pd@5% we see that both the proposed method and the ResNet50-NoDown show good robustness against processing, and in particular JPEG compression, resizing, filtering, blurring and contrast adjustment. In particular, these experiments confirm the same trend, with the baseline that outperforms our method on ProGAN, with an improvement of a few percent in many cases (and less than 10% in all the cases), but can not generalize to StyleGAN3, where our methods largely outperform ResNet50-NoDown. Both methods suffer from Gaussian noise addition. Given that noise addition has been considered among the augmentations during the training of ResNet50-NoDown, but not for the training of our method, it is not surprising that the performance with respect to this type of processing for our method is lower.

5.2 Other Results

5.2.1 Performance in the Presence of GAN Splicing

A noticeable strength of the proposed Eyes-based GAN detection method is that relies on semantic information for the discrimination. This is not the case with the ResNet50-NoDown method, which bases the decision on features automatically extracted by the network from the entire image. **Figure 4** (top row) shows some examples of attention maps obtained with the GradCAM algorithm (Selvaraju et al., 2017) for ResNet50-NoDown. As it is often the case with self-learned CNN architectures, the regions highlighted by the maps - that mostly affect the decision - are many and spread over the whole image, lying also in the background, confirming that the method in Gragnaniello et al. (2021) also suffers from well-known drawback of poor explainability of CNN-based solutions.

Due to this behavior of the baseline method, we then expect an advantage of our tool based on semantic features under local manipulations, e.g., in the GAN splicing scenario, when the GAN object (the face, in this case) is pasted on a real background.

To run some tests in the image spicing scenario, we generated a number of forged images for the case of real and synthetic faces, by cutting the foreground person from the image and pasting it on a real background. A total of 30 GAN spliced images for each GAN type (StyleGAN2, ProGAN and StyleGAN3) and 30 real spliced images are obtained in this way. An example of a local GAN spliced image is illustrated in **Figure 5**. Some examples of attention maps for ResNet50-NoDown obtained by running the GradCAM algorithm on the spliced images are provided in **Figure 4** (bottom row).

The results of the tests are provided in **Table 5**, where we report the Pd@5% obtained using the same threshold as before, set on the validation set. Obviously, the performance of our method is not affected by the splicing operations, given that the eyes region remains the same. Regarding the performance of the baseline, we observe that, although the evidence that can be found in the foreground is enough for the method to perform correct discrimination in the StyleGAN2 case (that is when the pasted foreground corresponds to a StyleGAN2 face), the presence of the real background affects the performance of generalization. In the case of StyleGAN3, where the performance was already poor in the non splicing case, the method gets Pd@5% = 3% on the GAN spliced images (i.e., the GAN spliced images are detected as real with a probability of 97%). In the ProGAN case, the Pd@5% decreases by 10%.

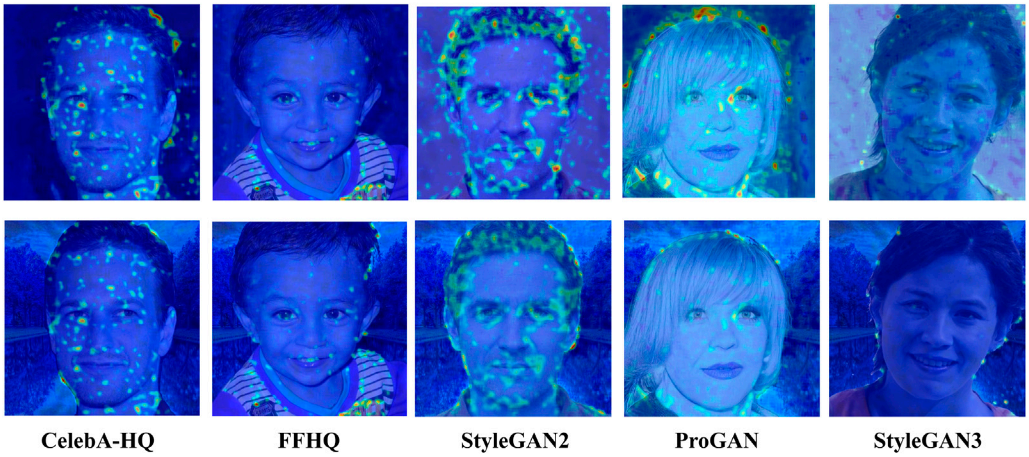

FIGURE 4 | GradCAM visualization for ResNet50-NoDown (Gragnaniello et al., 2021) before (top) and after (bottom) image splicing.

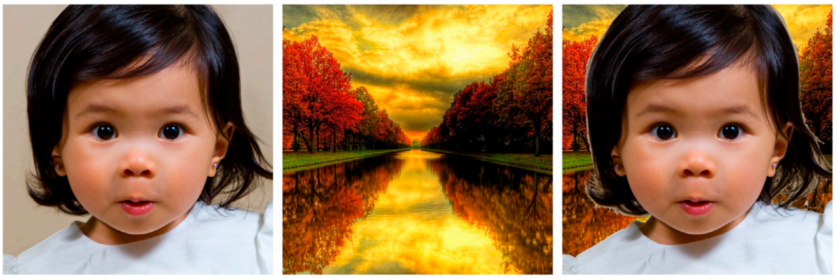

FIGURE 5 | Example of splicing operation in FFHQ dataset. From left to right: original image, real background, spliced image.

TABLE 5 | Results in terms of Pd@5% (FAR) achieved on spliced images.

Datasets	StyleGAN2	ProGAN	StyleGAN3
ResNet50-NoDown	100 (0)	90 (0)	3 (0)
Ours	100 (0)	97 (0)	100 (0)

5.2.2 Performance on Print&Scan Images

We also run some tests on the Print&Scan dataset in order to investigate the robustness of the proposed method to the rebroadcast operation and assess whether the features the detectors look at to reveal the GAN nature of the image survive recapturing. Being our method based on semantic attributes (eyes clues), we expect a better robustness against recapturing compared to the state-of-the-art method based on features automatically learnt from the full image. Some examples of recaptured GAN and real images, alongside the original digital versions, from the Print&Scan dataset are reported in **Figure 6**. We can see a noticeable quality degradation in the recaptured images. In particular, noisy textures are visible and the colors are changed.

The ROC curve of the proposed Eyes-based GAN-generated facial image detector and the baseline method on the Print&Scan image dataset are reported in **Figure 7**. We see that our method can achieve some noticeable discrimination capability. In particular, using the same threshold fixed on the validation set for digital images we get Pd@5% = 76% with a FPR = 22%, that is relevant results given the significant difference in the test image domain in this case. Adjusting the threshold on recaptured data helps improve the performance of the detector by 4% in the Pd@5% for the same FAR. Given that the Print&Scan dataset (Ferreira et al., 2021) is very challenging and recapturing attacks are very strong attacks, the results achieved by our method are good ones. From **Figure 6** (right), we see that the recapturing operation completely destroys the (weak) features that the ResNet50-NoDown method looks at, and no discrimination between real and fake can be obtained using this network, the AUC being around 61.6%.

5.3 Ablation Study

In **Table 6**, we report the performance of the Eyes-based GAN-generated facial image detector when different network architectures are considered to implement the two branches of the SNN. The performance achieved by training a standard XceptionNet on the entire image (Marra et al., 2018) is also reported for completeness. This table shows that the SNN with

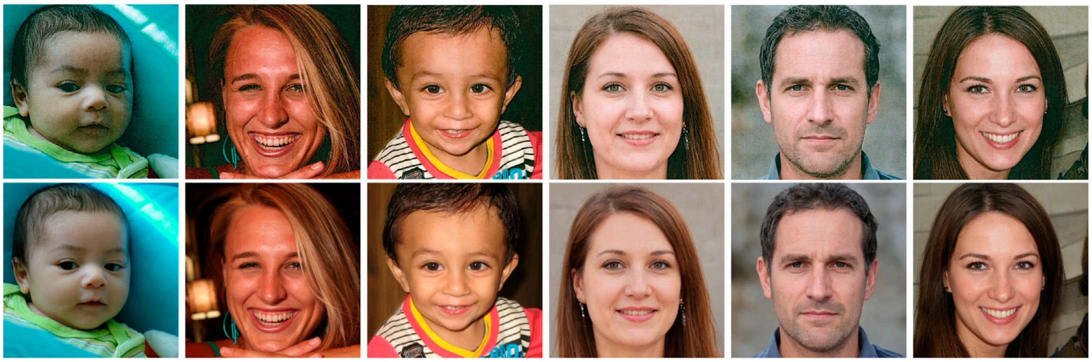

FIGURE 6 | Examples of images from the Print&Scan image dataset (Ferreira et al., 2021) (top), and corresponding digital image (bottom). From left to right, the first 3 images are real, the last 3 are GAN.

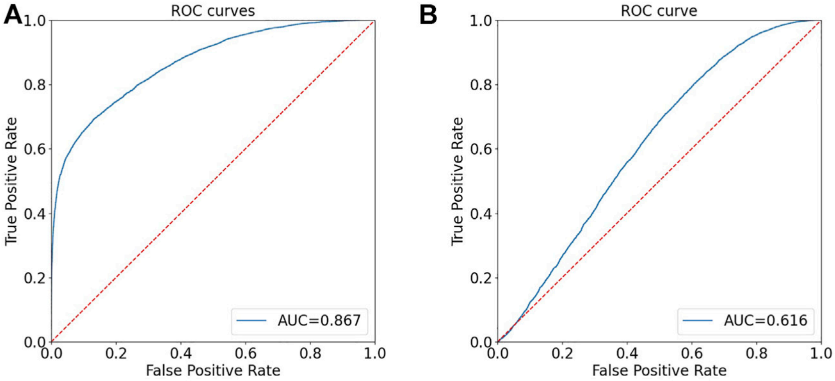

FIGURE 7 | Performance (AUC) on the Print&Scan image dataset: the proposed SNN-modifiedXceptionNet **(A)** and ResNet50-NoDown **(B)**.

TABLE 6 | Performance comparison among different network architectures.

Architectures	StyleGAN2		ProGAN		StyleGAN3	
	TNR/TPR	AUC	TNR/TPR	AUC	TNR/TPR	AUC
Xception Marra et al. (2018)	100/100	100	100/21.0	93.0	100/16.0	85.8
SNN-Xception	100/100	100	100/43.0	96.0	100/73.0	99.0
SNN-modifiedXception	100/100	100	100/62.0	98.7	100/80.0	99.3
SNN-modifiedResNet50	90.2/100	100	90.2/30.1	69.0	90.2/90.0	96.5

the modified XceptionNet corresponds to the best choice, getting better results than the standard XceptionNet. Moreover, better results are achieved using XceptionNet as the backbone, with respect to ResNet50, that is the network considered in Gragnaniello et al. (2021). Interestingly, all the models get a TPR of 100% on StyleGAN2 images (the TNR is also 100%), and the difference among the trained models relies on the results with ProGAN and StyleGAN3 images, that is, in the generalization performance. In particular, the SNN with the modified version of XceptionNet (best choice) can improve the generalization on ProGAN and StyleGAN3 images by 39 and 64% in the TPR, respectively, with respect to the standard XceptionNet model trained on the entire image. These results justify our choice of the modified XceptionNet as the backbone network for the two convolutional branches of the SNN.

6 CONCLUSION

We proposed a semantic-based method for GAN-generated face images detection that reveals the synthetic nature of a face image based on the analysis of eye clues, exploiting the similarity learning paradigm and Siamese Neural Networks

(SNNs). The method relies on the underlying assumption that GANs are not perfect in reconstructing inter-eye symmetries between the two eyes. The SNN is implemented by considering a modified XceptionNet as backbone network. Our experiments showed the good performance of the method both in terms of robustness against processing and generalization to unseen GAN architectures. In particular, the method can achieve comparable results to the best performing state-of-the-art method for GAN detection that works on the entire image, with the generalization and robustness performance being even superior in some cases.

Similarly to what has already been done for the detectors working full image (Bondi et al., 2020; Wang et al., 2020; Gragnaniello et al., 2021), it would be interesting to study the impact of the augmentation strategy on the performance of the method and see experimentally if there is an optimized chain of augmentations that allow to get better robustness and generality.

Future work will try to further improve the generalization and robustness capability of the detector by performing fusion with other detectors resorting to semantic analysis, that is, looking at other semantic facial attributes [e.g., the mouth (Suwajanakorn et al., 2017; Haliassos et al., 2021), or the nose (Chen and Yang, 2021)]. The existence of inconsistencies in symmetries that might come from other facial attributes, e.g., eyebrows or mouths shapes, is also worth investigation. Such inconsistencies could be exploited using an architecture based on Siamese Networks similar to the one used in this paper to get facial attributes-based detectors with improved robustness, thanks to the similarity-learning paradigm.

Finally, the behavior of the proposed network, obtained by inspecting activation maps, suggests that an algorithm could also be designed that looks specifically at the inconsistencies between the irises region by means of a Siamese Network architecture similar to the one proposed in this paper. To that purpose, the development of a method for the estimation of the (internal and external) circular corneal region, possibly resorting to the superior capabilities of CNNs for iris segmentation, is a necessary first step and also an interesting future research.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. The data used for training and testing our method have been made publicly available for reproducibility and can be found in the author's repository at the following links: https://github.com/tkarras/progressive_growing_of_gans; https://github.com/NVlabs/stylegan2; https://github.com/NVlabs/stylegan3; https://github.com/NVlabs/ffhq-dataset.

AUTHOR CONTRIBUTIONS

JW implemented the method, carried out the experiments, and wrote a first draft. BT contributed to planning the experiments, paper writing and positioned the work in the current state of research. MB supervised the study and helped interpreting the results.

FUNDING

This work has been partially supported by the China Scholarship Council (CSC), file No. 202008370186, by the PREMIER project under contract PRIN 2017 2017Z595XS-001, funded by the Italian Ministry of University and Research, and by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) under agreement number FA8750-20-2-1004. The US. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA and AFRL or the US. Government.

ACKNOWLEDGMENTS

The authors want to acknowledge their reviewers for their contributions in making this manuscript better and publishable.

REFERENCES

Agarwal, R., Machado, M. C., Castro, P. S., and Bellemare, M. G. (2021). Contrastive Behavioral Similarity Embeddings for Generalization in Reinforcement Learning. *arXiv preprint arXiv:2101.05265*

Barni, M., Costanzo, A., Nowroozi, E., and Tondi, B. (2018). "Cnn-Based Detection of Generic Contrast Adjustment with Jpeg Post-processing," in 2018 25th IEEE International Conference on Image Processing (ICIP) (IEEE), 3803. doi:10.1109/icip.2018.8451698

Barni, M., Kallas, K., Nowroozi, E., and Tondi, B. (2020). "Cnn Detection of gan-Generated Face Images Based on Cross-Band Co-Occurrences Analysis," in 2020 IEEE International Workshop on Information Forensics and Security (WIFS) (IEEE), 1–6. doi:10.1109/wifs49906.2020.9360905

Basha, S. H. S., Dubey, S. R., Pulabaigari, V., and Mukherjee, S. (2020). Impact of Fully Connected Layers on Performance of Convolutional Neural Networks for Image Classification. *Neurocomputing* 378, 112–119. doi:10.1016/j.neucom.2019.10.008

Berthelot, D., Schumm, T., and Metz, L. (2017). Began: Boundary Equilibrium Generative Adversarial Networks. *arXiv preprint arXiv:1703.10717*

Bertinetto, L., Valmadre, J., Henriques, J. F., Vedaldi, A., and Torr, P. H. S. (2016). "Fully-Convolutional Siamese Networks for Object Tracking," in European conference on computer vision (Springer), 850–865. doi:10.1007/978-3-319-48881-3_56

Bondi, L., Cannas, E. D., Bestagini, P., and Tubaro, S. (2020). "Training Strategies and Data Augmentations in Cnn-Based Deepfake Video Detection," in 2020 IEEE International Workshop on Information Forensics and Security (WIFS) (IEEE), 1. doi:10.1109/wifs49906.2020.9360901

Chen, Z., and Yang, H. (2021). "Attentive Semantic Exploring for Manipulated Face Detection," in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE), 1985–1989. doi:10.1109/icassp39728.2021.9414225

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). "Stargan: Unified Generative Adversarial Networks for Multi-Domain Image-To-Image Translation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 8789–8797. doi:10.1109/cvpr.2018.00916

Chollet, F. (2017). "Xception: Deep Learning with Depthwise Separable Convolutions," in Proceedings of the IEEE conference on computer vision and pattern recognition, 1251–1258. doi:10.1109/cvpr.2017.195

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2018). Generative Adversarial Networks: An Overview. *IEEE Signal Process. Mag.* 35, 53–65. doi:10.1109/msp.2017.2765202

Fang, P., Zhou, J., Roy, S. K., Petersson, L., and Harandi, M. (2019). "Bilinear Attention Networks for Person Retrieval," in Proceedings of the IEEE/CVF international conference on computer vision, 8030–8039. doi:10.1109/iccv.2019.00812

Ferreira, A., Nowroozi, E., and Barni, M. (2021). Vipprint: Validating Synthetic Image Detection and Source Linking Methods on a Large Scale Dataset of Printed Documents. *J. Imaging* 7, 50. doi:10.3390/jimaging7030050

Fonseca, N., and Guidetti, V. (2022). Similarity and Generalization: From Noise to Corruption. Interational Conference of Machine Learning. *arXiv preprint arXiv:2201.12803*. doi:10.3204/PUBDB-2022-00850

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2020). Generative Adversarial Networks. *Commun. ACM* 63, 139–144. doi:10.1145/3422622

Gragnaniello, D., Cozzolino, D., Marra, F., Poggi, G., and Verdoliva, L. (2021). "Are gan Generated Images Easy to Detect? a Critical Analysis of the State-Of-The-Art," in 2021 IEEE International Conference on Multimedia and Expo (ICME) (IEEE), 1–6.

Guo, W., Tondi, B., and Barni, M. (2021). A Master Key Backdoor for Universal Impersonation Attack against Dnn-Based Face Verification. *Pattern Recognit. Lett.* 144, 61–67. doi:10.1016/j.patrec.2021.01.009

Haliassos, A., Vougioukas, K., Petridis, S., and Pantic, M. (2021). "Lips Don't Lie: A Generalisable and Robust Approach to Face Forgery Detection," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5039–5049. doi:10.1109/cvpr46437.2021.00500

He, Z., Zuo, W., Kan, M., Shan, S., and Chen, X. (2019). Attgan: Facial Attribute Editing by Only Changing what You Want. *IEEE Trans. Image Process.* 28, 5464–5478. doi:10.1109/tip.2019.2916751

Hsu, C.-C., Zhuang, Y.-X., and Lee, C.-Y. (2020). Deep Fake Image Detection Based on Pairwise Learning. *Appl. Sci.* 10, 370. doi:10.3390/app10010370

Hu, J., Lu, J., and Tan, Y.-P. (2014). "Discriminative Deep Metric Learning for Face Verification in the Wild," in Proceedings of the IEEE conference on computer vision and pattern recognition, 1875–1882. doi:10.1109/cvpr.2014.242

Hu, S., Li, Y., and Lyu, S. (2021). "Exposing gan-Generated Faces Using Inconsistent Corneal Specular Highlights," in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE), 2500–2504. doi:10.1109/icassp39728.2021.9414582

Hulzebosch, N., Ibrahimi, S., and Worring, M. (2020). "Detecting Cnn-Generated Facial Images in Real-World Scenarios," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 642–643. doi:10.1109/cvprw50498.2020.00329

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2018). Progressive Growing of Gans for Improved Quality, Stability, and Variation. *arXiv preprint arXiv:1710.10196*

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., et al. (2021). Alias-free Generative Adversarial Networks. *Adv. Neural Inf. Process. Syst.* 34, 852–863.

Karras, T., Laine, S., and Aila, T. (2019). "A Style-Based Generator Architecture for Generative Adversarial Networks," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4401–4410. doi:10.1109/cvpr.2019.00453

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). "Analyzing and Improving the Image Quality of Stylegan," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8110–8119. doi:10.1109/cvpr42600.2020.00813

King, D. E. (2009). Dlib-ml: A Machine Learning Toolkit. *J. Mach. Learn. Res.* 10, 1755–1758. doi:10.5555/1577069.1755843

Kingma, D. P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*

Krishnagopal, S., Aloimonos, Y., and Girvan, M. (2018). Imilarity Learning and Generalization with Limited Data: A Reservoir Computing Approach. *Complexity* 2018, 15. doi:10.1155/2018/6953836

Li, H., Li, B., Tan, S., and Huang, J. (2020). Identification of Deep Network Generated Images Using Disparities in Color Components. *Signal Process.* 174, 107616. doi:10.1016/j.sigpro.2020.107616

Liu, Z., Qi, X., and Torr, P. H. (2020). "Global Texture Enhancement for Fake Face Detection in the Wild," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8060–8069. doi:10.1109/cvpr42600.2020.00808

Marra, F., Gragnaniello, D., Cozzolino, D., and Verdoliva, L. (2018). "Detection of gan-generated Fake Images over Social Networks," in 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (IEEE), 384–389. doi:10.1109/mipr.2018.00084

Marra, F., Saltori, C., Boato, G., and Verdoliva, L. (2019). "Incremental Learning for the Detection and Classification of gan-generated Images," in 2019 IEEE International Workshop on Information Forensics and Security (WIFS) (IEEE), 1–6. doi:10.1109/wifs47025.2019.9035099

Matern, F., Riess, C., and Stamminger, M. (2019). "Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations," in 2019 IEEE Winter Applications of Computer Vision Workshops (WACVW) (IEEE), 83–92. doi:10.1109/wacvw.2019.00020

McCloskey, S., and Albright, M. (2019). "Detecting gan-Generated Imagery Using Saturation Cues," in 2019 IEEE International Conference on Image Processing (ICIP) (IEEE), 4584–4588. doi:10.1109/icip.2019.8803661

Nataraj, L., Mohammed, T. M., Manjunath, B. S., Chandrasekaran, S., Flenner, A., Bappy, J. H., et al. (2019). Detecting gan Generated Fake Images Using Co-occurrence Matrices. *Electron. Imaging* 31, 532–541. doi:10.2352/issn.2470-1173.2019.5.mwsf-532

Perarnau, G., Van De Weijer, J., Raducanu, B., and Álvarez, J. M. (2016). Invertible Conditional Gans for Image Editing. *arXiv preprint arXiv:1611.06355*

Radford, A., Metz, L., and Chintala, S. (2016). "Unsupervised Representation Learning With Deep Convolutional Generative Adversarial Networks," in 4th International Conference on Learning Representations, (ICLR) 2016, San Juan, May 2–4, 2016. *arXiv preprint arXiv:1511.06434*

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. (2019). "Faceforensics++: Learning to Detect Manipulated Facial Images," in Proceedings of the IEEE/CVF International Conference on Computer Vision, 1–11.

Roth, K., Milbich, T., Sinha, S., Gupta, P., Ommer, B., and Cohen, J. P. (2020). "Revisiting Training Strategies and Generalization Performance in Deep Metric Learning," in International Conference on Machine Learning (PMLR), 8242–8252.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). "Grad-cam: Visual Explanations from Deep Networks via Gradient-Based Localization," in Proceedings of the IEEE international conference on computer vision, 618–626. doi:10.1109/iccv.2017.74

Son, J., Baek, M., Cho, M., and Han, B. (2017). "Multi-object Tracking with Quadruplet Convolutional Neural Networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 5620–5629. doi:10.1109/cvpr.2017.403

Suwajanakorn, S., Seitz, S. M., and Kemelmacher-Shlizerman, I. (2017). Synthesizing Obama. *ACM Trans. Graph.* 36, 1–13. doi:10.1145/3072959.3073640

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). "Rethinking the Inception Architecture for Computer Vision," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2818–2826. doi:10.1109/cvpr.2016.308

Varior, R. R., Haloi, M., and Wang, G. (2016). "Gated Siamese Convolutional Neural Network Architecture for Human Re-identification," in European conference on computer vision (Springer), 791–808. doi:10.1007/978-3-319-46484-8_48

Verdoliva, L. (2020). Media Forensics and Deepfakes: an Overview. *IEEE J. Sel. Top. Signal Process.* 14, 910–932. doi:10.1109/jstsp.2020.3002101

Wang, S.-Y., Wang, O., Zhang, R., Owens, A., and Efros, A. A. (2020). "Cnn-generated Images Are Surprisingly Easy to Spotfor Now," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 8695–8704.

Xuan, X., Peng, B., Wang, W., and Dong, J. (2019). "On the Generalization of gan Image Forensics," in Chinese conference on biometric recognition (Springer), 134–141. doi:10.1007/978-3-030-31456-9_15

Yang, X., Li, Y., Qi, H., and Lyu, S. (2019). Exposing gan-synthesized Faces Using Landmark Locations. *Proc. ACM Workshop Inf. Hiding Multimedia Secur.*, 113–118. doi:10.1145/3335203.3335724

Zhao, J., Mathieu, M., and LeCun, Y. (2017). Energy-Based Generative Adversarial Network. 5th International Conference on Learning Representations, ICLR 2017. *arXiv preprint arXiv:1609.03126*

Zhou, F., Jiang, Z., Shui, C., Wang, B., and Chaib-draa, B. (2021). Domain Generalization via Optimal Transport with Metric Similarity Learning. *Neurocomputing* 456, 469–480. doi:10.1016/j.neucom.2020.09.091

Van der Maaten, L., and Hinton, G. (2008). Visualizing Data Using t-SNE. *J. Machine Learning Res.* 9 (11), 2579–2605.

Mclnnes, L., Healy, J., and Melville, J. (2018). Umap: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv preprint arXiv:1802.03426

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Wang , Tondi and Barni. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.