Pixel-Wise Information in Fake Image Detection

Nhat-Khang Ngo[1,2,3] and Xuan-Nam Cao[2,3](✉)

[1] Advanced Program in Computer Science, University of Science,
Ho Chi Minh City, Vietnam
nnkhang19@apcs.vn

[2] Faculty of Information Technology, University of Science,
Ho Chi Minh City, Vietnam
cxnam@fit.hcmus.edu.vn

[3] Vietnam National University,
Ho Chi Minh City, Vietnam

Abstract. In recent years, generative adversarial networks have generated high-quality images that are difficult to differentiate by human eyes. Aside from the positives, improper use of this technology might have severe consequences for society. As a result, digital picture forensic techniques are critical to preventing these damages in our lives. In this paper, we describe a technique for detecting fraudulent face images generated by StyleGAN. Using a U-net-based classifier, we can integrate both global and local information of an image to determine if it is real or fake. A huge testing set of 20,000 pictures is used to confirm the model's efficacy. For comparison, we do tests on different CNN-based models. The results demonstrate that among the models on the same dataset, the U-net-based classifier has the greatest accuracy (98.43%). Finally, prediction maps are shown to demonstrate the relevance of pixel-level information in detecting real/fake images.

Keywords: Digital image forensics · Generative adversarial networks · StyleGAN · U-net

1 Introduction

Generative adversarial networks (GANs), which were first proposed by Goodfellow et al. [5], have recently gained a lot of attention in both academia and industry. As the name suggests, a GAN-based model tries to generate new data by simulating an adversarial game between a generator and a discriminator in the model. In computer vision, many GANs variants are developed to synthesize more realistic images. Several significant variants of GANs [7,8,16] can generate high-quality images which are usually indistinguishable by human eyes. In particular, StyleGAN can synthesize human faces with high reality and perceptual degrees. Figure 1 displays fake images generated by StyleGAN compared to the

T. K. Dang et al. (Eds.): FDSE 2021, CCIS 1500, pp. 436–443, 2021.
https://doi.org/10.1007/978-981-16-8062-5_30

real ones. Apart from the benefits that this revolutionized technology provides, malicious actors may use it to spread fake news to fool users or even to propaganda inciting violence. GANs generated image detection, a sub-topic of digital image forensics is a must-do task to verify the authenticity and integrity of a digital image.

There are two observations that we consider to detect fake faces in the images generated by GAN-based models. First, we argue that face images generated by GANs often contain suspicious backgrounds. In other words, the backgrounds of fake images are not realistic and tend to be recognized by detailed investigation. Then, it is essential to notice that GAN-based models focus on generating faces as identical as natural faces. Hence, high detailed synthesized faces are more difficult to recognize than the backgrounds in fake images. This conclusion is a motivation for us to experiment on fake face images detection.

In this paper, we present a deep learning method to detect fake faces generated by StyleGAN. Besides detection, we also aim at pointing out that pixel-wise information plays a role in face image authentication. To do that, we address the problem as a binary classification problem. However, there are two types of classifications in our definition, i.e., image-wise classification and pixel-wise classification. Image-wise classification can be done by extracting global information in an image. In addition, it is necessary to figure out which locations in an image can make the model decide this image is real or fake. We consider this information as pixel-level or local information. To combine local and global information, Schonfeld et al. [12] present an U-net-based discriminator that performs pixel-wise classification. The team synthesizes high-quality images by training a GAN-based model consisting of a U-net-based discriminator. Inspired by this work, we apply a U-net-based model to our real/fake faces classification problem. We use the architecture of U-net as a classifier to perform binary classification tasks at two levels, i.e., image-wise and pixel-wise.

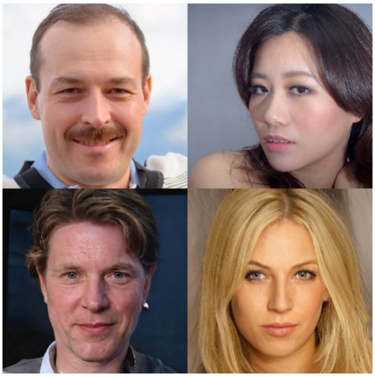

Fig. 1. Real images are displayed in the first row, and fake faces are in the second row

2 Related Work

Images generated by GAN-based models can be either synthesized from a random code or manipulated from a real image. In particular, fake faces generated by deep GAN-based models are known as deep fakes. Several models, including StyleGAN [8], StarGAN [1], StarGAN2 [2], can synthesize realistic faces by injecting natural face styles into the generators. Furthermore, StyleGAN successfully synthesizes fake faces that consist of various styles of humans. These faces vary continuously among hairstyles, ages, and gender. Consequently, high-quality synthesized faces pose a challenge for digital image forensic. Images synthesized by computers can easily fool a simple fake face detection system because these images do not contain any modifications from real faces, which are believed to be uncomplicated to recognize.

GAN-generated image detection, specifically deep fakes detection, has appeared in many studies in digital image forensics. Several methods combine traditional computer vision techniques and deep neural networks to detect differences between fake and real images. Frank et al. [3] and Zhang et al. [15] utilize spectrum as an input for the classifier. The teams claim that there are significant differences between authentic images and GAN generated images when they are interpreted in a frequency space. Moreover, Goebel et al. [4] and Nataraj et al. [10] compute a co-occurrence matrix of an image before using it as an input for deep learning networks. On the other hand, Xuan et al. [14] uses several techniques, e.g., Gaussian Blur or Gaussian noise, to preprocess the input before using it to train a CNN-based classifier. In our experiments, there are neither no preprocessing steps nor transforming inputs (except normalization). We, instead, train an end-to-end classifier to detect fake face images.

3 Method

This section is divided into two subsections. First, we describe in detail the U-net-based model for real/fake classification. We, then, demonstrate how to calculate the loss when training the model.

3.1 U-Net-Based Classifier

U-net [11] includes two modules, i.e., encoder, and decoder. Besides, there is a bottleneck layer between the two modules. The encoder of U-net extracts global information from the input image, while the decoder generates a prediction map that contains specific information of each pixel. The skip-connection mechanism enables the decoder to use the output of each layer in the encoder for pixel-wise information aggregation. U-net has high performance in image segmentation as it is good at representing local information. In addition, we observe that each pixel can act as a role to determine whether the image is real or fake. Prediction maps consisting of the confidence score of each pixel to be real or faker allow the classifier to learn the differences between real and fake images at the local level [12].

There are two types of outputs in a U-net-based classifier, i.e., scalar outputs and prediction maps. Figure 2 illustrates the model and its prediction mechanism. More precisely, a scalar value is a global prediction of an image. This value determines an image's label when it is downsampled. On the other hand, a prediction map, which is the same size as the input image, contains prediction values for each pixel. Each value demonstrates whether the corresponding pixel is real or fake. In a U-net-based classifier, the output of the bottleneck goes in two parallel ways. We, first, pass this output to a fully connected network (FC) to get a scalar value, i.e., scalar prediction. The bottleneck output also simultaneously moves to the decoder to generate a prediction map, and this map acts as a pixel-wise prediction map. The two final outputs are used to calculate the total loss for the model. In this paper, we use Sigmoid function to compute the confidence scores of the outputs.

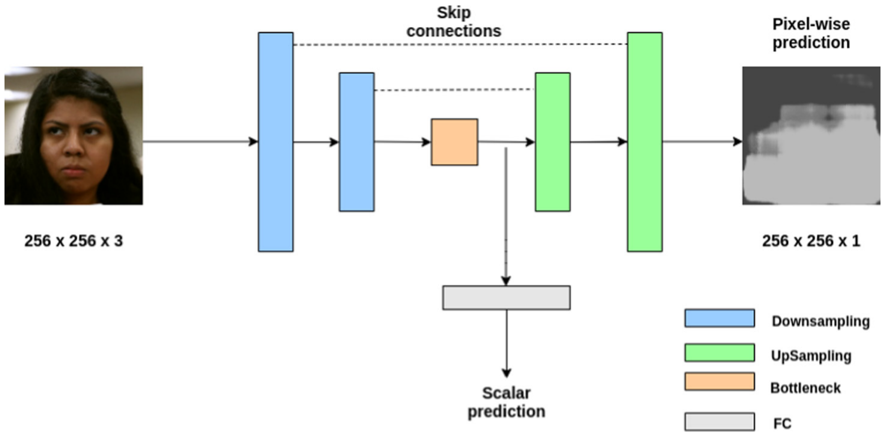

Fig. 2. U-net-based classifier

3.2 Loss Function

We apply binary cross-entropy loss at two levels. First, the original binary loss is calculated between the image's label and the models' scalar prediction. We also calculate a pixel-wise binary loss. More precisely, we calculate the loss for each pixel in the output map. The pixels in a prediction map from a real input image have ground-truth values of 1. Otherwise, these values from fake images are 0. After calculating the loss of each pixel, we take the average value of these losses. In total, the loss function is a sum of the image-wise loss and pixel-wise loss.

The equations below describe the loss functions that we use in training. N and M denote the number of images and number of pixels in each image, respectively.

Moreover, y_i, $\hat{y_i}$ are the ground-truth value and prediction score of an image, whereas y_{ij}, $\hat{y_{ij}}$ are the ground-truth value and prediction score of a pixel in an image. L_g refers to the image-wise loss (global loss), and L_p refers to the pixel-wise loss.

$$L_g = -\frac{1}{N}\sum_{i=1}^{N} y_i \log \hat{y_i} + (1 - y_i) \log (1 - \hat{y_i})$$

$$L_p = -\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} \log \hat{y_{ij}} + (1 - y_{ij}) \log (1 - \hat{y_{ij}})$$

$$L = L_g + L_p$$

4 Experiments and Results

4.1 Dataset

Our study uses a dataset consisting of 70,000 real images and 70,000 fake images [1]. Real images are from Flickr Dataset, and the fake ones are generated by StyleGAN [8]. We use 100,000 images for training, 20,000 for validation, and 20,000 for testing. The images illustrated in Fig. 1 are samples in the dataset. Additionally, all of the images are resized to 256 × 256.

4.2 Network Structure and Implementation Details

We use the architecture of Pix2Pix's generator [7] which is a U-net-based architecture, to build the classifier. We append a fully connected network (FC) right after the bottleneck. The output of the bottleneck has the size of $(B, 1, 1, 512)$, where B is the batch size. Hence, we need to flatten this output before passing it to the fully connected network. The network has two layers, whose sizes are 512,256, respectively, and ReLU as activation functions. We empirically train our model with different layer sizes and obtain the highest accuracy on 512 and 256 for two layers. In addition, the output size of the decoder is also resized to $256 \times 256 \times 1$ in which each cell represents a confidence score of being real of each corresponding pixel in the input image. To train the model, we create two types of labels for each image, i.e., scalar value (0 or 1) for global classification and a label mask with a size of 256 × 256 × 1 for pixel-wise classification. Only fake and real images are included in the dataset, i.e., there are no mixed images that are made up of both real and fake pixels. Thus, the label mask either contains full of 1s or 0s based on the image's label. By default, we use Adam optimizer [9] with a learning rate of 2×10^{-4}, and the batch size is 128 for model training.

[1] Link to the dataset: https://www.kaggle.com/xhlulu/140k-real-and-fake-faces.

4.3 Results

Besides training a U-net-based model, we also train and test other models for comparison and evaluation. We conduct experiments on DenseNet [6], VGG [13], and vanilla CNNs. After testing the models on the testing set of 20,000 images, we attain accuracy for each model as displayed in Table 1. The table reveals that our model has the highest accuracy (98.43%). We speculate that adding pixel-wise loss helps increase the accuracy. In the prediction stage, we use three kinds of confidence scores, i.e., scalar prediction, map prediction, and an average of the two predictions, to determine the label of tested images. First, we only use scalar prediction as a confidence score and achieve 98.43% of accuracy. Then, we take the average of values in the prediction map to get a scalar score for the corresponding image, and this also yields an accuracy of 98.43%. Finally, we take the average of the two scores and acquire accuracy of 98.4275%. The three results reveal a relevance between global information and local information in real/fake image detection.

Table 1. Test accuracy on U-net-based model and CNN-based models

	Accuracy
DenseNet	97.00%
VGG	95.00%
CNNs	92.00%
U-net classifier	**98.43%**

As mentioned above, a U-net-based classifier can differentiate real images from fake images at the pixel level. Figure 3 displays images that we use in the testing phase with their corresponding prediction maps. The left image in the first row is real, whereas the below one is fake. As we consider brighter colors as higher confidence scores of pixels to be real, the figure reveals that our model can distinguish real and fake pixels in images. The map of the real image is entirely white, which means that every pixel in the image has an expensive confidence score of being real. The score of the fake image with the dark map, on the other hand, is low.

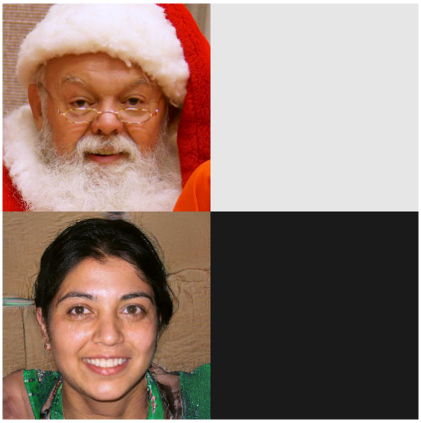

Fig. 3. Images and their prediction maps

5 Conclusion

This paper presents a U-net-based classifier to differentiate GAN-based generated faces. The model is inspired by the U-net-based discriminator in [12]. We train the model to classify real and fake faces generated by StyleGAN. After training, we test our method on the testing set of 20,000 images and attain the highest accuracy of 98.43%. Besides accuracy, we show that a U-net-based classifier can recognize and distinguish fake and real pixels. High results demonstrate the importance of local information in fake face detection or fake image detection in general. In the future, we plan to test our model on other types of images, e.g., scenery images.

Acknowledgement. This research is supported by research funding from Faculty of Information Technology, University of Science, Vietnam National University - Ho Chi Minh City.

References

1. Choi, Y., Choi, M., Kim, M., Ha, J.W., Kim, S., Choo, J.: StarGAN: unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8789–8797 (2018)
2. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: StarGAN v2: diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8188–8197 (2020)
3. Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. In: International Conference on Machine Learning, pp. 3247–3258. PMLR (2020)

4. Goebel, M., Nataraj, L., Nanjundaswamy, T., Mohammed, T.M., Chandrasekaran, S., Manjunath, B.: Detection, attribution and localization of GAN generated images. arXiv preprint arXiv:2007.10466 (2020)
5. Goodfellow, I., et al.: Generative adversarial nets. In: Advances in Neural Information Processing Systems 27 (2014)
6. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4708 (2017)
7. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1125–1134 (2017)
8. 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, pp. 4401–4410 (2019)
9. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
10. Nataraj, L., et al.: Detecting GAN generated fake images using co-occurrence matrices. Electron. Imaging **2019**(5), 532–1 (2019)
11. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
12. Schonfeld, E., Schiele, B., Khoreva, A.: A U-Net based discriminator for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8207–8216 (2020)
13. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
14. Xuan, X., Peng, B., Wang, W., Dong, J.: On the generalization of GAN image forensics. In: Sun, Z., He, R., Feng, J., Shan, S., Guo, Z. (eds.) CCBR 2019. LNCS, vol. 11818, pp. 134–141. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-31456-9_15
15. Zhang, X., Karaman, S., Chang, S.F.: Detecting and simulating artifacts in GAN fake images. In: 2019 IEEE International Workshop on Information Forensics and Security (WIFS), pp. 1–6. IEEE (2019)
16. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)