Fake Faces Identification via Convolutional Neural Network

Huaxiao Mo
Sun Yat-sen University
Guangzhou, China
mohx5@mail2.sysu.edu.cn

Bolin Chen
Sun Yat-sen University
Guangzhou, China
chenbl8@mail2.sysu.edu.cn

Weiqi Luo*
Sun Yat-sen University
Guangzhou, China
luoweiqi@mail.sysu.edu.cn

ABSTRACT

Generative Adversarial Network (GAN) is a prominent generative model that are widely used in various applications. Recent studies have indicated that it is possible to obtain fake face images with a high visual quality based on this novel model. If those fake faces are abused in image tampering, it would cause some potential moral, ethical and legal problems. In this paper, therefore, we first propose a Convolutional Neural Network (CNN) based method to identify fake face images generated by the current best method [20], and provide experimental evidences to show that the proposed method can achieve satisfactory results with an average accuracy over 99.4%. In addition, we provide comparative results evaluated on some variants of the proposed CNN architecture, including the high pass filter, the number of the layer groups and the activation function, to further verify the rationality of our method.

KEYWORDS

Image Forensics, Deep Learning, Generative Adversarial Networks (GAN), Convolutional Neural Network (CNN)

ACM Reference Format:
Huaxiao Mo, Bolin Chen, and Weiqi Luo. 2018. Fake Faces Identification via Convolutional Neural Network. In *IH&MMSec '18: 6th ACM Workshop on Information Hiding and Multimedia Security, June 20–22, 2018, Innsbruck, Austria.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3206004.3206009

1 INTRODUCTION

With the rapid development of image processing technology, modifying an image without obvious visual artifacts becomes much easier. Nowadays, seeing is no longer believing. Image forensics have attracted widely attention in the last decade, and many forensic methods based on hand-crafted features such as [4, 15, 16, 19] have been proposed until now.

Different from conventional methods based on hand-crafted features, deep learning can exploit cascaded layers to adaptively learn hierarchical representations from input data. Some novel models in deep learning such as CNN and GAN have been extensively studied and have achieved great success in many image related applications, such as image style transfer [8, 11], image super-resolution [13, 18], image inpainting [10, 22] and image steganalysis [6, 21]. Up to now, several deep learning based works have been proposed for image forensics. For instance, Chen et al. [5] proposed a median filtering forensic method based on CNN; Bayar et al. [2] proposed a new CNN architecture to detect several typical image manipulations; Rao et al. [17] proposed a CNN based method to detect image splicing and copy-move; Choi et al. [7] proposed a CNN based method to detect composite forgery detection. Recently, some studies have shown that we can obtain fake face images with high visual quality (refer to Section 2 for details) based on GAN model. Since these fake face images can successfully cheat our eyes, identifying fake images becomes an very important issue in image forensics. In this paper, we propose a CNN based method to identify the fake images generated by the work [20]. In our method, we carefully design the CNN architecture, with particular attention to the high pass filter for the input image, the number of layer groups and the activation function, and then provide extensive experimental results to show the effectiveness and rationality of the proposed method. To the best of our knowledge, this is the first work to investigate this forensic problem.

The rest of the paper is organized as follows. Section 2 describes two recent GAN based works on face generation. Section 3 gives the proposed detection method based on CNN. Section 4 shows experimental results and discussions. Finally, the concluding remarks of this paper and the future works are given in Section 5.

2 FACE GENERATION WITH GAN

Generative adversarial networks (GAN) [9] is a prominent generative model, which learns the distribution form high-dimension data and produces novel sample. Typically, a GAN contains two parts: a generator and a discriminator. The generator learns to create fake data indistinguishable from the real data, while the discriminator learns to determine whether the input data is real or fake. They contest against each other during training until the generator can produce high quality fake data.

Recently, several GAN based methods are proposed to generate high quality fake face images. For instance, in [3], Berthelot et al. proposed a novel equilibrium method for balancing the two parts of GAN to generate visually pleasing face images. However, this method can only produce fake face images in low resolutions, such as 256×256. In [20], Karras et al. proposed a progressive strategy to construct and train GAN for generating high quality images. The progressive strategy is illustrated in Fig. 1. Instead of training the whole GAN on high-resolution image, they first construct a simple GAN training on low-resolution images in the beginning, and then gradually add more layers to adapt the model to high-resolution images during the training stage. Based on the experiments, most fake face images (1024×1024) generated by this method are difficult to identify with the naked eye, as illustrated

*Corresponding author.

IH&MMSec'18, June 20–22, 2018, Innsbruck, Austria

ACM ISBN 978-1-4503-5625-1/18/06…$15.00
https://doi.org/10.1145/3206004.3206009

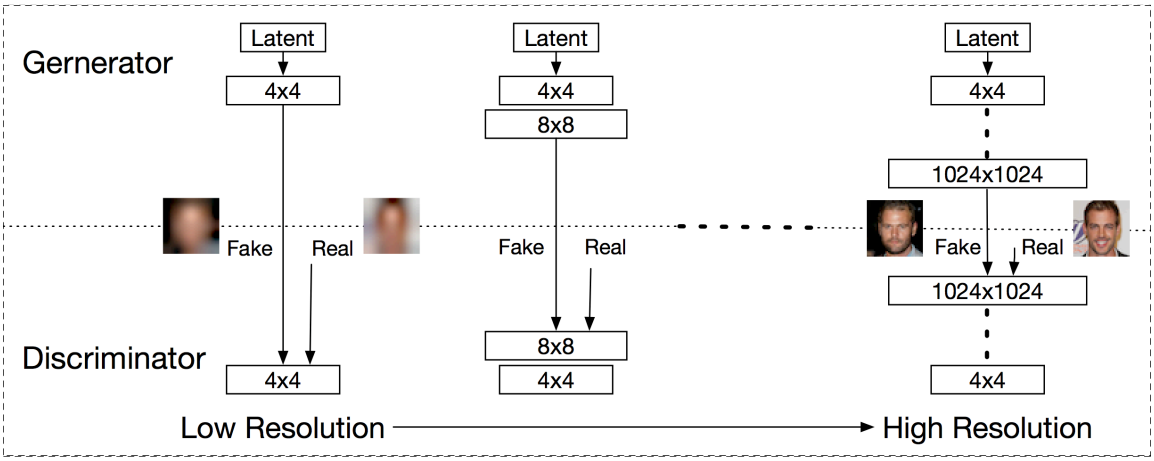

Figure 1: The progressive training strategy employed in [20]. Here $N \times N$ refers to layers operating on images of $N \times N$ resolution.

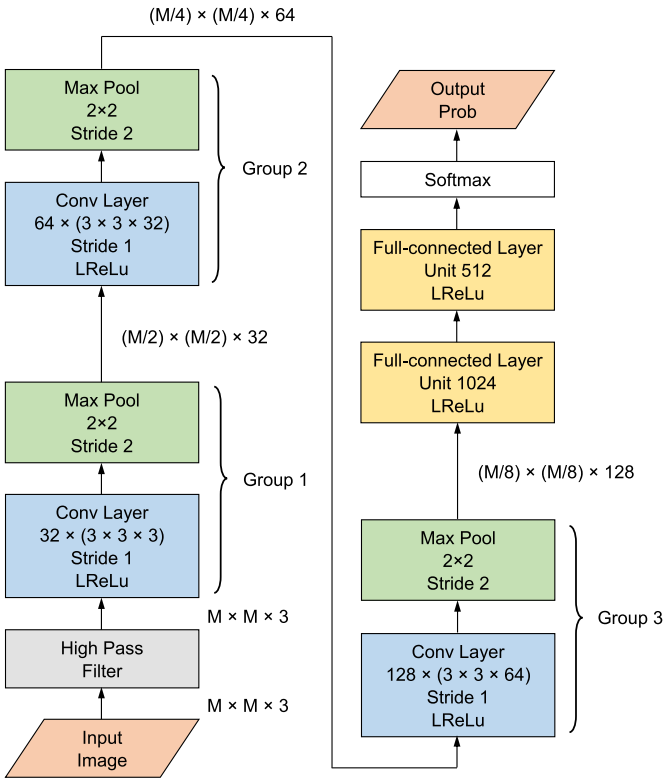

Figure 2: The proposed architecture

in the first row of Fig.8. However, some poorer results are also obtained using this method, as illustrated in the second row of Fig. 8. In this paper, we first propose a method to identify those good fake face images generated using the method [20].

3 THE PROPOSED DETECTION METHOD

Since the generator and discriminator employed in [20] are mainly based on CNN, it is natural to use a CNN based method to detect the resulting fake face images. To this end, we carefully design the architecture of the proposed CNN model, as illustrated in Fig. 2. The model input is an RGB color image with size $M \times M \times 3$. Since the contents of fake and true face images are quite similar, it is expected that the main difference between the two kinds of images would be reflected on the residual domain according to the previous research [14]. Therefore, we first transform the input images into residuals using a high pass filter. The resulting residuals are then fed into three layer groups. Each group includes a convolutional layer (3×3 size, 1×1 stride) equipped with LReLu and a max pooling layer (2×2 size, 2×2 stride). The output feature map number of the convolutional layer in the first group is 32, while that of the other convolutional layers is twice the corresponding input feature map number. The output feature maps of the last group are then aggregated and fed into two fully-connected layers. They both equipped with LReLu and consist of 1024, 512 units, respectively. Finally, the softmax layer is used to produce the output probability.

In our experiments, we implement the proposed CNN model using Tensorflow [1] and train it using Adam [12] with a learning rate of 0.0001. All the weights are initialized using a truncated Gaussian distribution with zero mean and standard deviation of 0.01. The biases is initialized as zero. L2 regularization is enabled in the fully-connected layer with the λ of 0.0005. In the training stage, we use a batch size of 64 and train the proposed CNN for 20 epochs. In addition, we shuffle the training data between epochs.

4 EXPERIMENTAL RESULTS

In this section, we first describe the image data set used in our experiments. Then we present some experiments to show the effectiveness of the proposed method in identifying fake face images. In addition, we conduct extensive experiments to show the rationality of the proposed model.

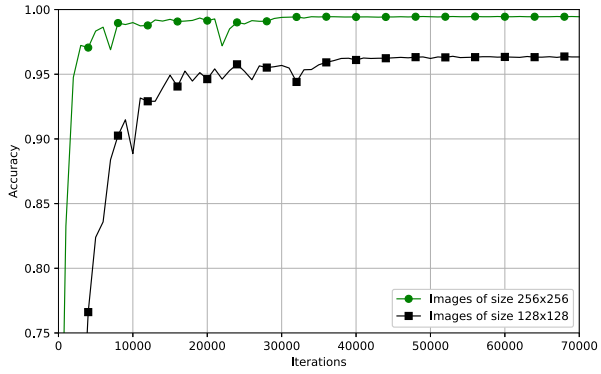

Figure 3: Comparison of different image sizes

4.1 Image Data Set

In our experiments, we use 30,000 true face images from CELEBA-HQ dataset and select 30,000 fake ones with good visual quality from the fake face image database [1] generated by [20]. All images are with 1024×1024 and stored in PNG format. In our experiments, we resize all images into 256×256 using bilinear interpolation, and compress them using lossy JPEG compression with a quality factor of 95. Finally, we divide the resulting images into training set (12,000 pairs of true-fake face images), validation set (3,000 pairs) and test set (15,000 pairs). To achieve convincing results, we randomly split the training, validation, and test set three times and report the average results in the following experiments.

4.2 Fake Face Identification

In this section, we aim to identify whether a given face image is real or generated one. As shown in the blue box in Fig. 9, we found that some background regions in some fake face images looks unnatural, which may help increase the detection performance. To reduce the influence of image backgrounds, we crop a small segment (128×128) from every image in the original image set (256×256), and ensure that each cropped segment mainly includes some facial key-points (such as eyes, nose, and mouse), as illustrated in red box in Fig. 9. Finally, we obtain two different image data set for experiments, that is original ones including face and background, and the cropped ones just including main facial region.

The experimental results evaluated on the two validation set are shown in Fig. 3. From Fig. 3, we observe that during training stage, the proposed model on both original images (green line) and cropped images (black line) can converge within 70,000 iterations, and both detection accuracies would be over 95% after 40,000 iterations. For a more convincing result, we evaluate the trained model on the test set, and obtain accuracies of over 99.4% and 96.3% on the original images and cropped images respectively, which means that we can still obtain satisfactory results after removing unnatural parts in the background.

[1] Available at: https://drive.google.com/drive/folders/0B4qLcYyJmiz0TXY1NG02bzZVRGs

4.3 Comparison of Variants of the Proposed Model

In this section, we present some results to validate the rationality of the proposed model in Fig.2. Three parts of our model are considered, including the high pass filter, the number of the layer groups and the activation function. The corresponding results evaluated on the validation set are shown in the following subsections.

4.3.1 High Pass Filter: In this experiment, three following high pass filters are evaluated in the model.

$$\underset{\text{(a) filter A}}{\begin{bmatrix} -1 & 1 \end{bmatrix}} \quad \underset{\text{(b) filter B}}{\begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}} \quad \underset{\text{(c) filter C}}{\begin{bmatrix} -1 & 2 & -2 & 2 & -1 \\ 2 & -6 & 8 & -6 & 2 \\ -2 & 8 & -12 & 8 & -2 \\ 2 & -6 & 8 & -6 & 2 \\ -1 & 2 & -2 & 2 & -1 \end{bmatrix}}$$

Figure 4: Three high pass filters

The corresponding results are shown in Fig. 5. From Fig 5, we observe the proposed model (i.e. using the filter B) can achieve highest accuracy among the three test filters. We also observe that the model using the filter A can achieve similar results with our proposed model, while the model with filter C has the lowest detection accuracy. In addition, the detection accuracy of removing the high pass filter is nearly 98%, which means that using suitable high pass filter can help improve detection performance.

4.3.2 Number of layer groups: In this experiment, we evaluate the influence of adding/removing one layer group in the proposed model. The results are shown in Fig. 6. From Fig. 6, we observe that adding one group to the proposed model does not increase the detection performance, while removing one group decreases the detection performance slightly, which means that using three layer groups in the proposed model is sufficient for the investigated problem.

4.3.3 Activation functions: Activation function is another important factor in CNN. In this experiment, six commonly used activation functions are considered in the proposed model. They are TanH, ReLu, and four variants of ReLu, including PReLu, LReLu, ELU and ReLu6. The experimental results are shown in Fig. 7. From Fig. 7, we observe that LReLu ReLu, PReLu, and ELU can achieve similar accuracy. Among six activation functions, LReLu obtains the best performance while TanH shows the worst performance.

5 CONCLUSION

In this paper, we first propose a CNN based method to identify fake face images generated with the state-of-the-art method [20], and provide extensive experimental results to show that the proposed method can effectively identify fake face images with a high visual quality from real ones. Our experimental results also indicate that even though the current GAN based methods can generate realistic looking faces (or other image objects and scenes), some obvious statistical artifacts would be inevitably introduced and can serve as evidences for fake ones.

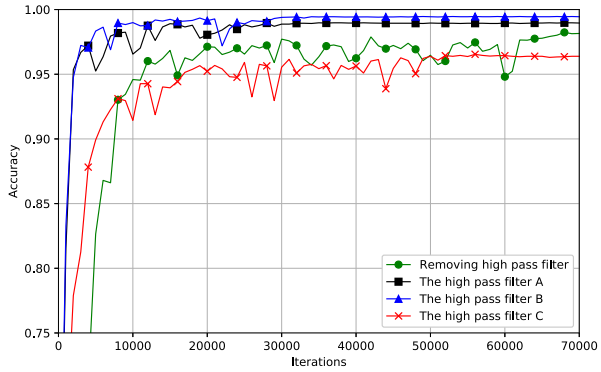

Figure 5: Comparison of using different high pass filters/without high pass filter

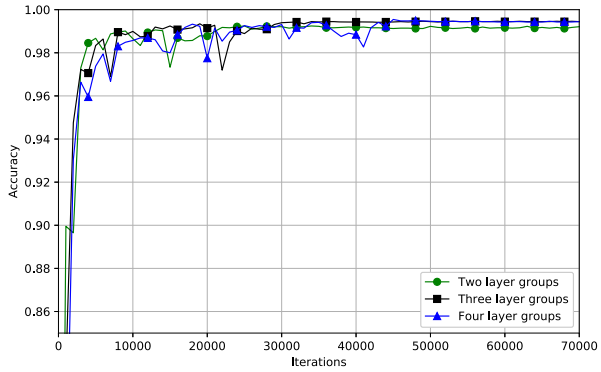

Figure 6: Comparison of different numbers of groups

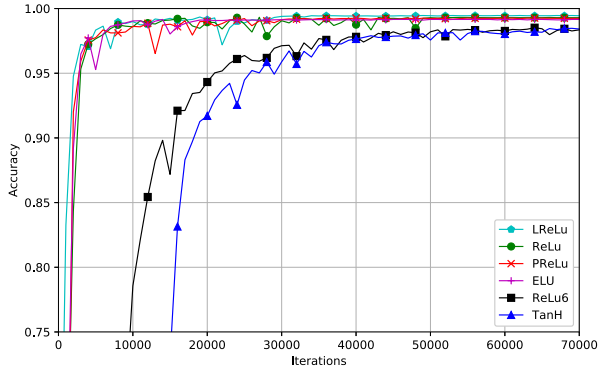

Figure 7: Comparison of different activation functions

In future, we will further investigate some inherent artifacts left by the GAN in [20] for image forensics. On the other side, we will try to propose a wise face generation method that can avoid detection.

ACKNOWLEDGMENTS

This work is supported in part by the NSFC (61672551), the Special Research Plan of Guangdong Province under Grant 2015TQ01X365, and the Guangzhou Science and Technology Plan Project under Grant 201707010167.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).

[2] Belhassen Bayar and Matthew C Stamm. 2016. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*. 5–10.

[3] David Berthelot, Tom Schumm, and Luke Metz. 2017. Began: Boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717* (2017).

[4] Gang Cao, Yao Zhao, Rongrong Ni, and Xuelong Li. 2014. Contrast enhancement-based forensics in digital images. *IEEE transactions on information forensics and security* 9, 3 (2014), 515–525.

[5] Jiansheng Chen, Xiangui Kang, Ye Liu, and Z Jane Wang. 2015. Median filtering forensics based on convolutional neural networks. *IEEE Signal Processing Letters* 22, 11 (2015), 1849–1853.

[6] Mo Chen, Vahid Sedighi, Mehdi Boroumand, and Jessica Fridrich. 2017. JPEG-Phase-Aware Convolutional Neural Network for Steganalysis of JPEG Images. In *ACM Workshop on Information Hiding and Multimedia Security*. 75–84.

[7] Hak-Yeol Choi, Han-Ul Jang, Dongkyu Kim, Jeongho Son, Seung-Min Mun, Sunghee Choi, and Heung-Kyu Lee. [n. d.]. Detecting composite image manipulation based on deep neural networks. In *IEEE International Conference on Systems, Signals and Image Processing*. 1–5.

[8] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. 2017. A learned representation for artistic style. In *Proceedings of International Conference on Learning Representations*.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[10] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2017. Globally and locally consistent image completion. *ACM Transactions on Graphics* 36, 4 (2017), 107:1–107:14.

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*. Springer, 694–711.

[12] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *CoRR* abs/1412.6980 (2014). arXiv:1412.6980 http://arxiv.org/abs/1412.6980

[13] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4681–4690.

[14] Haodong Li, Weiqi Luo, Xiaoqing Qiu, and Jiwu Huang. 2018. Identification of various image operations using residual-based features. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 1 (2018), 31–45.

[15] Lu Li, Jianru Xue, Zhiqiang Tian, and Nanning Zheng. 2013. Moment feature based forensic detection of resampled digital images. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 569–572.

[16] Xiaoqing Qiu, Haodong Li, Weiqi Luo, and Jiwu Huang. 2014. A universal image forensic strategy based on steganalytic model. In *Proceedings of the 2nd ACM workshop on Information hiding and multimedia security*. ACM, 165–170.

[17] Yuan Rao and Jiangqun Ni. 2016. A deep learning approach to detection of splicing and copy-move forgeries in images. In *IEEE International Workshop on Information Forensics and Security*. 1–6.

[18] Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. 2017. Amortised map inference for image super-resolution. In *Proceedings of International Conference on Learning Representations*.

[19] Matthew Stamm and KJ Ray Liu. 2008. Blind forensics of contrast enhancement in digital images. In *IEEE International Conference on Image Processing*. IEEE, 3112–3115.

[20] Samuli Laine Jaakko Lehtinen Tero Karras, Timo Aila. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *International Conference on Learning Representations* (2018). https://openreview.net/forum?id=Hk99zCeAb accepted as oral presentation.

Figure 8: Fake face examples from the work [20]. The first row shows examples with a good visual quality, while the second shows ones with a poor visual quality that would be removed in our experiments.

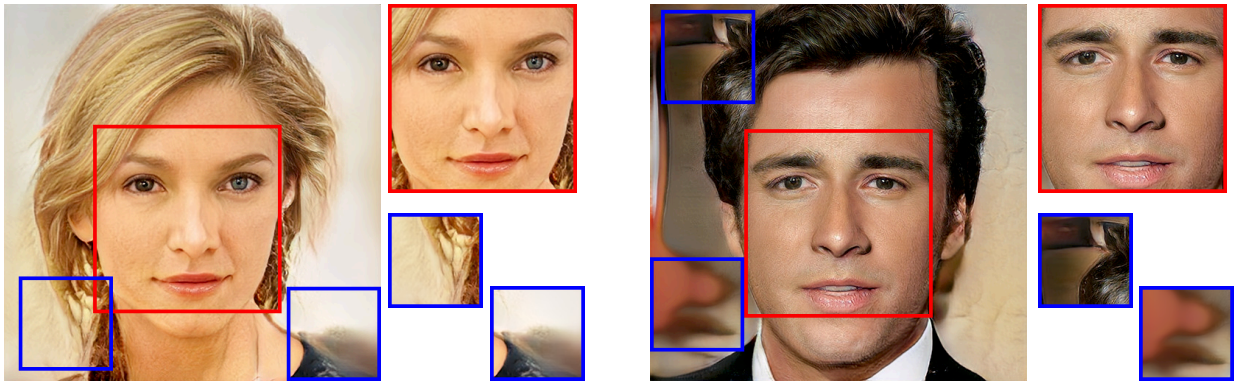

Figure 9: Fake face vs. Background. The region in the red box includes some facial key-points; while the blue ones are located at the background with poor visual artifacts.

[21] Guanshuo Xu, Han Zhou Wu, and Yun Qing Shi. 2016. Structural Design of Convolutional Neural Networks for Steganalysis. *IEEE Signal Processing Letters* 23, 5 (2016), 708–712.

[22] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. 2017. High-resolution image inpainting using multi-scale neural patch synthesis. In *The IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 1. 3.