Detecting Face Images Generated by AI

Yue Wan[1] Manning Wu[1] Xiuwen Xi [1]
Chen Wang [1]

DDA4210 Advanced Machine Learning Course Project
The Chinese University of Hong Kong, Shenzhen
April 30 2023

yuewan@link.cuhk.edu.cn, xiuwenxi@link.cuhk.edu.cn

chenwang@link.cuhk.edu.cn, manningwu@link.cuhk.edu.cn

Abstract

Considering the increasing prevalence of harmful events caused by the circulation of fake images, detecting fake images has become essential these days. Many research papers that are using fancy ways and give us excellent performance. Our work constructs a novel model with MobileNetV2 as its classifier to enhance efficiency while maintaining high accuracy. The feature we used was based on Liu's (2022) paper. However, instead of directly using the concatenation of the Learned Noise Pattern (LNP), amplitude spectrum, and phase spectrum of images, we first used the Segemant Anything Method (SAM) to extract meaningful objects in the image and then created a Dual Attention Block (DAB) to fuse those three features properly. Our method provides a modification and enhancement of current methods and shares some insights regarding further performance improvements. Code is available at `https://drive.google.com/drive/u/0/folders/1qzfFvCjlVORJd9VEnzbGtaXG90RJ-kVw`.

1 Introduction

With the rapid development of deepfake image synthesis techniques such as GANs and diffusion models recently, there is increasing difficulty in distinguishing between real and deepfake images especially those generated by StyleGAN [3], leading to new crimes like frauds through deceiving AI-generated faces, which has aroused wide public concern. Therefore, it is extremely urgent to develop a promotable method of detecting deepfake faces. While previous work has explored different methods focusing on artifacts detection and data- driven models, limitations of accuracy and generalization issues exist. A prominent artifacts detection method is to use a generator to simulate and classify sampling artifacts on common GANs in both spatial and frequency domains, which demonstrates superior performance in the frequency domain [11]. However, the model is inefficient for unconditional GANs like ProGAN and StyleGAN. A prominent data-driven method is to train ResNet-50 directly on a broad dataset of ProGAN-generated images to learn the common features [9]. However, the model only focuses on the characteristics of generated images, causing poor general applicability. In this work, our goal is to extract the common features of real and fake faces on a broad dataset to develop an improved classification model powerful in solving deepfake face detection problems on new data distributions, contributing to the urgent need for crime prevention and reduction. To achieve these aims, we address the following research question:

> **R.Q.** How to derive an accurate and general applicable face detection model that can be applied to identify real and fake faces regardless of the background and generation architecture?

The main contributions of our work can be summarized as follows:

- We provide sound technical support for fighting against new crimes such as frauds regarding AI-generated faces, contributing to the urgent need of crime prevention and reduction.
- We derive a lighter AI-generated face detection model with superior generalization ability, which

inspires a new way for the future study to achieve high accuracy with shorter training time.

2 Methodology

Our model framework consists of four components: face extraction, feature extraction, feature fusion and classification, illustrated in Figure 1.

Face extraction. We apply Segment Anything Model (SAM), consisting of an image encoder, a prompt encoder and a mask decoder, to extract faces from primary images, preventing disturbance of backgrounds and thus enhancing accuracy (§B). For each image, SAM outputs an object segmentation mask by using an MLP to map given image and prompt embeddings to a dynamic linear classifier computing the mask foreground probability at each location, which is proved to have high accuracy and powerful generalization on various objective types [1]. In our work, we use sparse coordinate points as prompt to locate faces, illustrated in Figure 2.

Feature extraction. For each extracted face, we choose Learned Noise Pattern (LNP), LNP phase spectrum and LNP amplitude spectrum as classification feature, since the spatial and frequency domain features for real and fake images are discrepant [3]. LNP refers to noise representations of an image under high-dimensional spatial mapping through neural networks. In the imaging process of real images, as photons enter the camera, the intensity of incident light at various locations in a negative is irregular, and thus, LNP shows different patterns for a real image depending on light intensities, while exhibits periodic checkerboard-like patterns for a fake image.

We apply CycleISP [10] to extract LNP from images, illustrated in Figure 3. The input noisy image first go through a 3 × 3 convolutional layer to obtain the low-level feature map M_0, which then go through four Recursive Residual Groups (RRG), each composed of two Dual Attention Blocks (DAB) suppressing the uninformative features using spatial and channel attention, to obtain the deep feature map M_1 . Next, M_1 is propagated to another 3 × 3 convolutional layer to obtain the feature map M_2 , 112 and LNP equals $-M_2$. The process can be formulated as follows:

$$M_0 = K_3((INI(x, y))) \tag{1}$$

$$M_1 = RRG(RRG(RRG(RRG(M_0)))) \tag{2}$$

$$M_2 = K_3(M_1) \tag{3}$$

$$LNP = -M_2 \tag{4}$$

where INI denotes the input noisy image and K_3 denotes a 3 × 3 convolutional layer.

We apply Fourier transform to compute phase spectrum and amplitude spectrum of LNP. The Fourier transform of an $M \times N$ image can be formulated as follows:

$$F(u, v) = \frac{1}{MN} \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x, y) e^{-i2\pi(\frac{ux}{M} + \frac{vy}{N})} \tag{5}$$

where $f(x, y)$ denotes the value at spatial domain point (x, y) of the input image, while $F(u, v)$ denotes the value at the corresponding frequency domain point (u, v) of the image after the Fourier transform. Phase spectrum and amplitude spectrum in the frequency domain are then computed as follows:

$$\phi(u, v) = arctan[\frac{I(u, v)}{R(u, v)}] \tag{6}$$

$$A(u, v) = \sqrt{R^2(u, v) + I^2(u, v)} \tag{7}$$

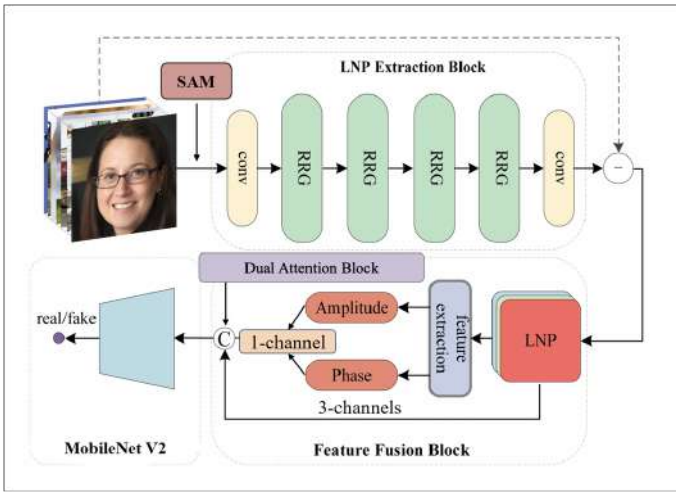

Figure 1. Model Framework Overview

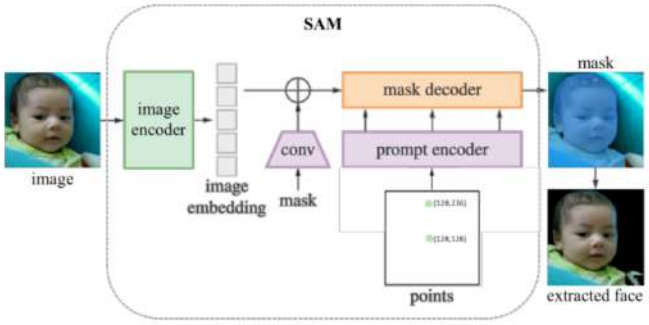

Figure 2. Segment Anything Model (SAM) overview.

where R and I denote the real part and imaginary part of F respectively; $\phi(u, v)$ and $A(u, v)$ denote the value at frequency domain point (u, v) of phase spectrum and amplitude spectrum respectively. Therefore, with LNP demonstrating spatial domain features and LNP spectrums carrying frequency domain features, we extracted the integrated information for each image, several examples of which are shown in Figure 4.

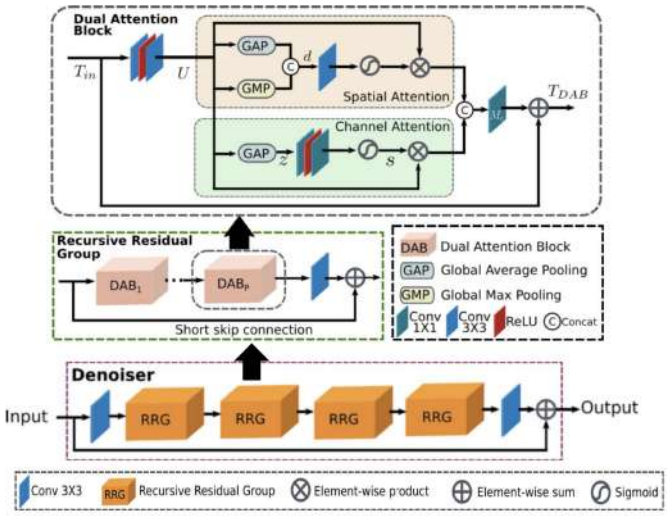

Figure 3. CycleISP denoising network overview

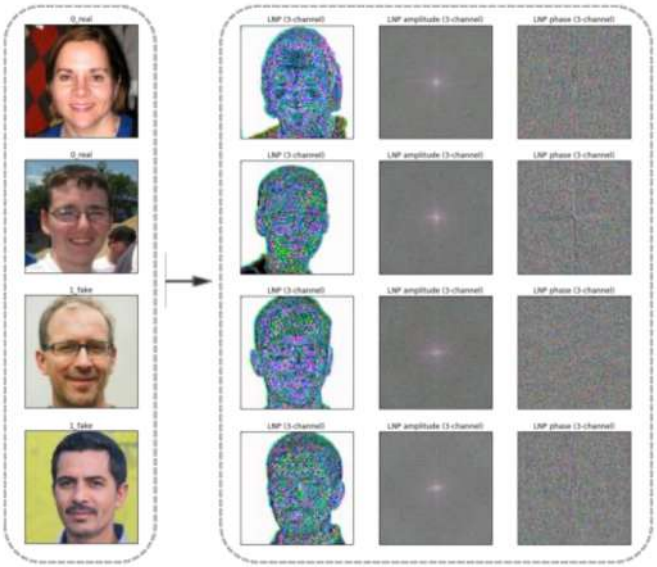

Figure 4. Examples of extracted features.

Feature Fusion. As shown in Figure 4, we obtain the 3-channel RGB images of LNP, LNP amplitude spectrum and LNP phase spectrum for each face image. However, since the spectrums only represent frequency domain features, using all the 3-channel information of spectrums will cause information redundancy and thus reduce accuracy. To overcome this, we transform the spectrums from 3-channel RGB images to 1-channel grey images by allocating the weights as $[r, g, b] = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$.

To further improve feature representation, we apply Dual Attention Block (DAB) to allocate weights to the spectrums with regard to LNP, rather than directly concatenate the obtained 3-channel LNP and 1-channel amplitude and phase spectrums. DAB, a self-attention mechanism originated from DANet [6], contains a position attention module learning the feature spatial interdependencies and a channel attention module learning the model channel interdependencies, illustrated in Figure 5 and Figure 6. Again, since the spectrums do no represent much spatial domain features, we only use the channel attention model as illustrated in Figure 7 to learn the allocated weights.

Classification. After feature fusion, we experiment with three classifiers: ResNet-34, ResNet-50 and MobileNetV2 for the classification task.

Residual Network (ResNet), which adopts deep residual learning and shortcut connections to overcome obstacles brought by deep networks such as degradation and vanishing gradients, is a prominent architecture for image classification [7]. In residual learning, the layers are reformulated as learning residual functions with reference to the layer inputs instead of learning unreferenced functions; short-

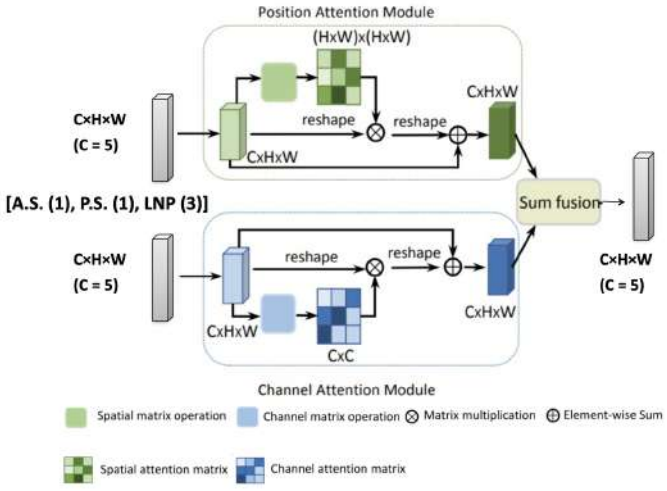

Figure 5. Dual Attention Block (DAB) overview.

cut connections with identity mappings, meanwhile, ensure the low complexity. Thus, ResNet can easily gain accuracy from considerably increased depth. Compared with ResNet-34, as illustrated in Figure 8, ResNet-50 (the suffix number denotes number of layers) has more hidden layers and an additional bottleneck in each residual block [7], allowing for reduced parameters for input layers to improve efficiency.

MobileNetV2, which involves a novel layer module: the inverted residual with linear bottleneck, is a lighter architecture that significantly decreases computational costs while retaining almost the same accuracy [8] and [2]. In the inverted residual structure where shortcut connections are between the thin bottleneck layers, the intermediate expansion layer uses lightweight depthwise convolutions to filter non-linear features, and a linear convolution is subsequently applied to remove non-linearities to maintain feature representations [8], illustrated in Figure 9 and Figure 10.

We train the above-mentioned three classifiers on the training set and validate the general applicability of our model on the validation set. For each classifier, we conduct the experiment for three times and record average performance to reduce bias. Then, we apply the trained classifiers to the test set and obtain the final prediction results. In our work, accuracy and running time are chosen as the performance evaluation metrics.

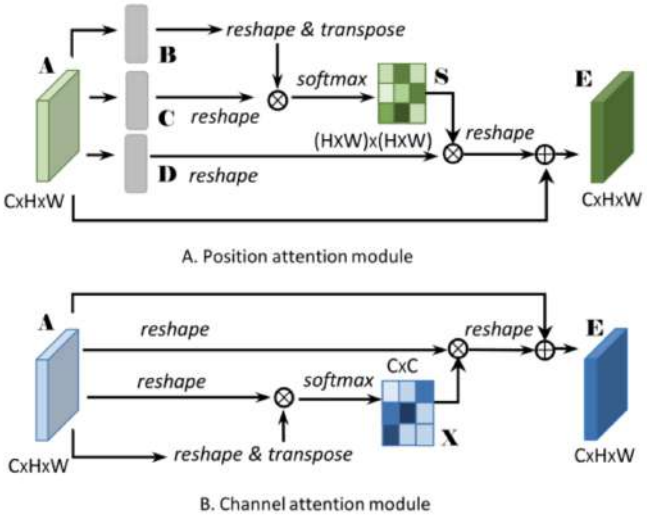

Figure 6. Position and channel attention module.

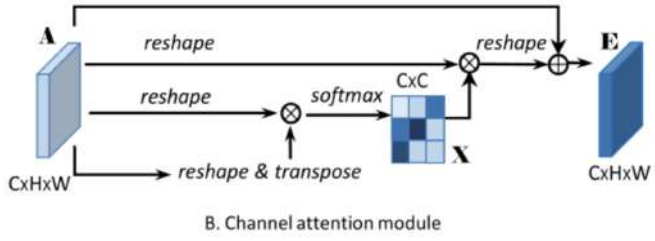

Figure 7. Channel attention module in our work.

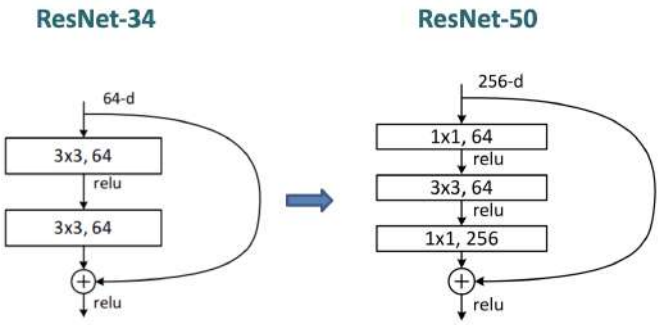

Figure 8. ResNet-34 and ResNet-50.

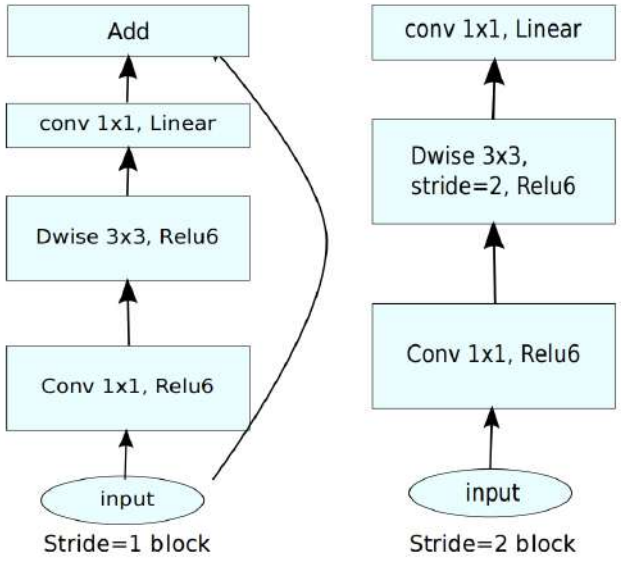

Figure 9. MobileNetV2.

3 Experiments

3.1 Data Collection

Our primary dataset consists of 14k face images with labels of two classes: 70k real faces from Flickr-Faces-HQ by Nvidia and 70k fake faces generated by advanced GANs and diffusion models sampled from the 1-Million-Fake-Faces provided by Bojan (§A). We randomly split each class by a ratio of 5 to 1 to 1 and then combine both classes to form training, validation and test set.

3.2 Experimental Results

Model training. In Table 1 we compare performances on the validation set across different classifiers. Results demonstrate that ResNet-50 has the highest accuracy (99.15%), while the MobileNetV2 has the lowest running time and the second highest accuracy (98.97%). Since the accuracy difference between the two is marginal and MobileNetV2 is significantly faster, we choose MobileNetV2 as our final classifier to achieve a balance between high accuracy and low complexity.

Table 1. Performance on validation set.

Method	ResNet-34	ResNet-50	MobileNetV2
Accuracy(%)	97.20%	99.15%	98.97%
Running Time(h)	4	4	3

Model testing. In Table 2 we compare model performances on the test set across different classifiers, verifying our trained classification model based on MobileNetV2 can gain high accuracy (98.57%) and reduce computational cost on new data, demonstrating superior general applicability. In Table 3 we compare our method with five other deep learning methods from related works: Zhang et al. [11], Frank et al. [5], Wang et al. [9], Gragnaniello et al. [4], Liu et al. [3], and find that our classification model has an outstanding performance.

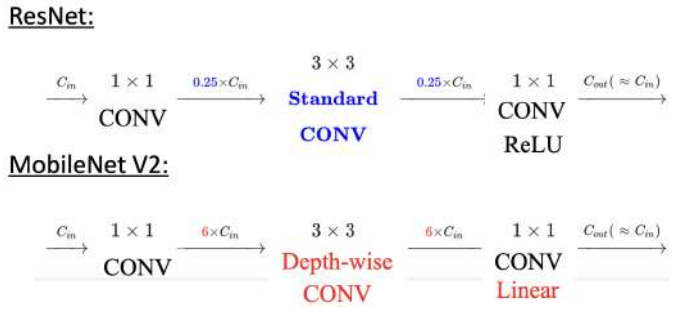

Figure 10. ResNet and MobileNetV2.

Table 2. Performance on test set.

Method	ResNet-34	ResNet-50	MobileNetV2
Accuracy(%)	96.24%	98.92%	98.57%
Running Time(h)	4	4	3

Table 3. Performance Comparison.

Method	Accuracy
AutoGAN-Spec(19')	60.7%
DCT-CNN(20')	43%
Wang(20')	98.5%
Grangnaniello(21')	98.9%
Liu(22')	99.6%
Ours	99.7%

4 Conclusions

To sum up, our work proposes a method to detect AI-generated face images from diverse sources. The novelty and limitations of our work can be summarized as follows:

Novelty

- We use a new objective segmentation model: SAM to precisely extract face regions from primary images rather than use unprocessed images directly, which prevents disturbance of backgrounds and thus the model involving SAM reveals higher accuracy.
- We focus on the common features of real and fake images in both spatial and frequency domains instead of only learning features of generated images, which allows for high generalization ability to new data from various sources and generation architectures.
- We apply dual attention mechanisms for feature fusion, allocating the optimal weights to different extracted features rather than directly concatenate them, which further enhances accuracy.
- We use a lighter network: MobileNetV2 with inverted residual structure for the classification task, which significantly decreases computational costs while retaining almost the same accuracy.

Limitations

- Due to computational limitations, we only randomly sample out part of the whole face dataset, which may lead to sample bias and thus, there is possibility that such high accuracy might not still be maintained if tested on the complete dataset.

- Since we only focus on the face dataset, the discrimination effectiveness is uncertain for other image categories such as scenery and animals.

5 Acknowledgement

Thanks to Professor Jicong Fan, the course instructor of DDA4210 Advanced Machine Learning, his proficient teaching and guidance through the course project is outstanding.

6 Appendix

Table of contents:

6.1 A. Primary Dataset Details

For convenience, all the images are resized to 256 × 256 and compressed in JPEG. We finally obtain 100k images for training, 20k images for validation and 20k images for test, each with equal number of real (labelled 0) and fake (labelled 1) faces. Here are the sources we collect data from:

- Flickr-Faces-HQ (real faces) by Nvidia: `https://github.com/NVlabs/ffhq-dataset`

- 1-Million-Fake-Faces by Bojan: `https://kaggle.com/datasets/tunguz/1-million-fake-faces`

Figure 11shows some sample images of our dataset:

6.2 B. Effectiveness of SAM

In Table 4 we compare performance of our model across applying SAM to primary images or not and using different classification models.

The below experimental results show that applying SVM enhances accuracy of our model, verifying the effectiveness of SVM.

Table 4. Accuracy (%)of applying SVM or not

SAM(without/with)	ResNet-34	ResNet-50	MobileNetV2
Without	96.03%	98.54%	98.21%
With	96.24	98.92	98.57

6.3 C. Feature Fusion Details

We conduct a two-phase experiment to verify the effectiveness of our feature fusion method.

In Phase I, we aim to determine a suitable number of channels to concatenate from amplitude spectrum, phase spectrum, and LNP. We set LNP as 3 channels while spectrums as either 1-channel or 3-channel. The results are shown in Table 5.

Table 5. Accuracy (%) of applying using spectrums with different channels

Spectrums	ResNet-34	ResNet-50	MobileNetV2
3-channel	93.43%	96.21%	95.97%
1-channel	95.92	98.73	98.44

The above experimental results indicate using 1-channel spectrums leads to better performance. In Phase II, we investigate whether applying DAB on channel compression will further improve performance. The results are presented in Table 6, where Att denotes attention module in DAB; AS and PS denote amplitude and phase spectrums. The above experimental results demonstrate applying DAB leads to higher accuracy. Therefore, during the feature fusion, we concatenate 3-channel LNP with attention weighted 1-channel spectrums.

Figure 11. Sample images of our dataset.

Table 6. Accuracy (%) of applying DAB or not

Feature Fusion	ResNet-34	ResNet-50	MobileNetV2
AS+PS+LNP	95.92%	98.73%	98.44%
Att(AS+PS+LNP)	96.16	98.83	98.55
Att(AS+PS)+LNP	96.24	98.92	98.57

References

[1] N. R. H. M. C. R. L. G. T. X.-S. W. A. C. B. W.-Y. L. P. D. Alexander Kirillov, Eric Mintun and R. Girshick. Segment anything. *arXiv: 2304.02643*, 2023.

[2] B. C. W. W. l.-C. C. M. T. G.-C. V. V. Y. Z. R. P. H. A. Andrew Howard, Mark Sandler and Q. Le. Searching for mobilenetv3. *ICCV*, 2019.

[3] X. B. B. X. W. L. Bo Liu, Fan Yang and X. Gao. Detecting generated images by real images. *Lecture Notes in Computer Science*, 2022.

[4] F. M. G. P. Diego Gragnaniello, Davide Cozzolino and L. Verdoliva. Are gan generated images easy to detect? a critical analysis of the state-of-the-art. *ICME*, 2021.

[5] L. S. A. F. D. K. Joel Frank, Thorsten Eisenhofer and T. Holz. Leveraging frequency analysis for deep fake image recognition. *International Conference on Machine Learning*, 2020.

[6] H. T. Y. L. Y. B. Z. F. Jun Fu, Jing Liu and H. Lu. Dual attention network for scene segmentation. *CVPR*, 2019.

[7] S. R. Kaiming He, Xiangyu Zhang and J. Sun. Deep residual learning for image recognition. *CVPR*, 2016.

[8] M. Z. A. Z. Mark Sandler, Andrew Howard and L.-C. Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. *CVPR*, 2018.

[9] R. Z. A. O. Sheng-Yu Wang, Oliver Wang and A. A. Efros. Cnn generated images are surprisingly easy to spot... for now. *CVPR*, 2020.

[10] S. K. M. H. F. S. K. M.-H. Y. Syed Waqas Zamir, Aditya Arora and L. Shao. Cycleisp: Real image restoration via improved data synthesis. *CVPR*, 2020.

[11] S. K. Xu Zhang and S.-F. Chang. Detecting and simulating artifacts in gan fake images. *WIFS*, 2019.