On Detecting GANs and Retouching based Synthetic Alterations

Anubhav Jain, Richa Singh, Mayank Vatsa
IIIT Delhi, India
{anubhav15129, rsingh, mayank}@iiitd.ac.in

Abstract

Digitally retouching images has become a popular trend, with people posting altered images on social media and even magazines posting flawless facial images of celebrities. Further, with advancements in Generative Adversarial Networks (GANs), now changing attributes and retouching have become very easy. Such synthetic alterations have adverse effect on face recognition algorithms. While researchers have proposed to detect image tampering, detecting GANs generated images has still not been explored. This paper proposes a supervised deep learning algorithm using Convolutional Neural Networks (CNNs) to detect synthetically altered images. The algorithm yields an accuracy of 99.65% on detecting retouching on the ND-IIITD dataset. It outperforms the previous state of the art which reported an accuracy of 87% on the database. For distinguishing between real images and images generated using GANs, the proposed algorithm yields an accuracy of 99.83%.

1. Introduction

Digital images have become an essential part of our daily lives. With the availability of sophisticated image processing tools and techniques, the Internet is filled with fake images. While some of these images are harmless, others have been used for creating forged legal documents, presenting doctored evidence in court, and manipulating historic incidences. Today, social media is also flooded with re-touched images which makes someone's skin look flawless. Social media websites have also started promoting retouching, by introducing image filters to enhance your appearance. These filters let the user remove wrinkles, pimples, change basic facial structures, and add texture, along with altering facial color i.e. forging skin tones to have fairer skin or adding unnatural tanning effect. Similarly, as shown in Figure 1, beauty or celebrity magazines which are giving people unrealistic expectations with altered appearances.

One of the most challenging aspects of image forgery is that, when done carefully, it can be visually imperceptible. Russello [15] showed that such altered appearances lowers self-esteem of people trying to adhere to the societal norms on attractiveness. It also leads to body dissatisfaction due to unrealistic body images being portrayed. In 2013, Israel announced its plans to enact the Photoshop Law. This law makes it mandatory for advertisers and magazines to label photo-shopped images [3]. It shows the necessity of algorithms to detect tampering and also the extent to which this issue is prevalent.

Figure 1. The image on the left is the altered body image of model Countess Filippa Hamilton and the image on right is an unaltered image. Image taken from [4]

Apart from the health and moral effects of image retouching, synthetic alterations also affect biometric system used for identification of individuals. There is a plausibility that the doctored image might be unrecognizable or incomparable with its original version. This could hinder the identification process or automatic matching with original faces. Recent studies have shown that face recognition models suffer in the presence of retouching or make-up [8]. While facial images are being used in identification cards, there is a need for an automatic system which can detect retouched images.

More recently, with the emergence of Generative Adversarial Networks [7, 11, 21], researchers have been exploring image generation as well. With these algorithms becoming more sophisticated, generated images are now looking exceptionally realistic. Various GANs such as CycleGANs [21] are used for learning image to image transla-

Figure 2. The first row consists of original images and the second row consists of their altered approaches [1, 2]. The first two samples show retouching and the remaining two samples are generated using StarGANs.

tions. Pix2pix GANs [11] network is able to generate images using label maps. It also uses an image to image translation approach and is able to color images and even generate images from edge maps. While the use of generated images has not been reported for manipulation, it possesses similar powers as tampered images, if not more. This makes it important to also have a mechanism to detect such fake images.

1.1. Related Work

Retouching, makeup detection, face spoofing and morphing are widely studied areas, that can be considered similar to retouching detection. Recent work by Bharati et al. [5] makes use of supervised deep Boltzmann machine algorithm for detecting retouching on the ND-IIITD database. It also introduces the ND-IIITD dataset which consists of 2600 original and 2275 retouched images. It uses different facial parts to learn features for classification. In 2017, Bharati et al. [6] proposed an algorithm which uses semi-supervised autoencoders. The paper has reported results on the Multi-Demographic Retouched Faces (MDRF) dataset. Earlier research by Kee and Farid [12] learned a support vector regression (SVR) between the retouched and original images. They used both geometric and photometric features for training the SVR on various celebrity images.

Research in the broad area of facial image forensics, include the paper by Kose et al. [13] which uses SVM and alligator classifier on a feature vector consisting of shape and texture characteristics. The have reported accuracies for makeup detection on the YMC and VMU datasets. Singh et al. [16] presents an algorithm which detects tampered face images. The algorithm makes use of a gradient based approach for classification.

1.2. Contributions

While existing research has primarily focuses on one of the challenges, this paper proposes a convolutional neural network based algorithm to detect retouching and image generation using GANs. The results are demonstrated using images generated from StarGAN [7] and facial retouching on the ND-IIITD dataset [5].

2. Proposed Detection Algorithm

Different kinds of tampering/retouching algorithms introduce different kinds of irregularities in the face image. As the alterations visually blend inside the image there is a need to focus on local regions, boundary regions and texture. Convolutional neural networks such as ResNet [17] have demonstrated effective results for different image classification challenges by encoding local and global features. Therefore, we have proposed CNN based architecture for detecting alterations.

2.1. Convolutional Neural Network

As shown in Figure 3, the proposed approach is built on the CNN architecture, where the first step consists of extracting non-overlapping patches of size (64,64,3) or (128,128,3) (only in the case of detecting retouching) from the image. The extracted patches are used as an input for the convolutional neural network which detects various features such as edges, texture and objects. The architecture consists of 6 hidden convolutional layers and 2 fully connected layers. The convolutional layers use kernel size of (3,3,D) where D is the depth of the filter.

Inspired from the wider architecture of ResNet [20] and the residual connections, the proposed algorithm uses a residual connection. Mathematically, the residual connection for wider networks can be summarized as:

$$y = F_1(x, \{W_{1_i}\}) + F_2(x, \{W_{2_i}\}) \tag{1}$$

The functions in equation (1) represent the mapping to be learned. The "addition" operation refers to an element wise addition with the shortcut connection. The algorithm uses a similar residual connection to connect the second and the fifth layers with the help of a pooling layer i.e., the output of the second layer is added to the output of the 5th layer. The algorithm takes inspiration from the idea in [20] and it introduces a 15 layered convolutional block on the residual connection. With this, the output getting added is the output/feature map of this convolutional block.

To localize retouching in patches, the model is trained using focal loss, which is mathematically represented in Equation (2).

$$FocalLoss(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t) \tag{2}$$

$$p_t = \begin{cases} p & if \;\; y = 0 \\ 1-p & otherwise \end{cases} \tag{3}$$

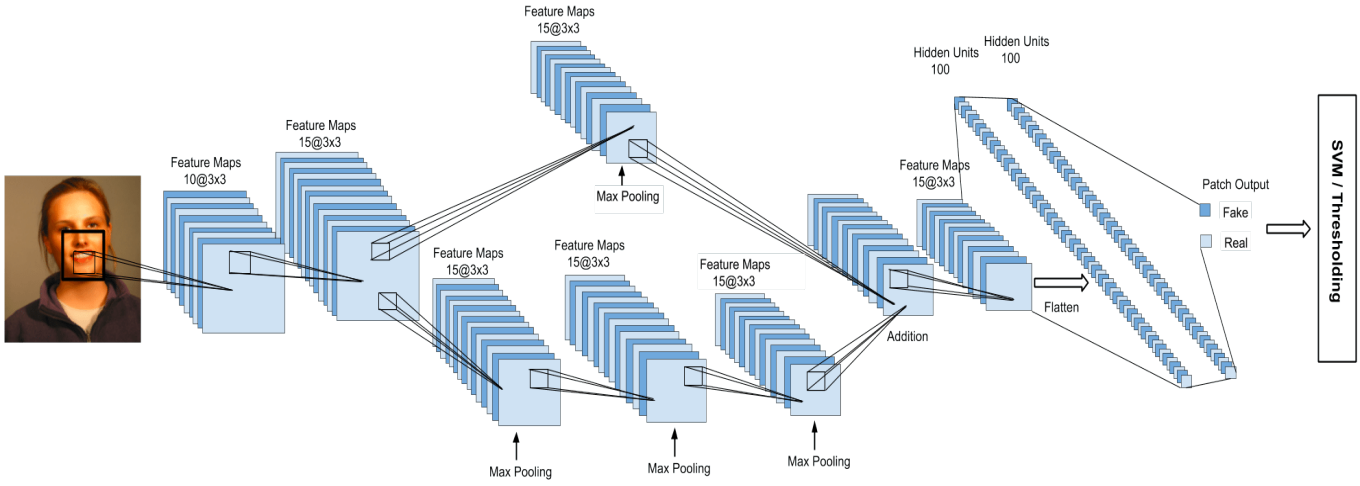

Figure 3. Illustrating the steps involved in the Convolutional Neural Network Architecture for alteration detection.

This is derived by modifying the expression for the conventional cross entropy loss. It introduces an additional term of $(1-p_t)^{\gamma}$ which contains the tunable parameter γ. The algorithm sets the value of the trainable parameter γ to be equal to 5. The benefit of using focal loss is two fold, it helps in localizing objects or regions and addresses large class imbalance. The latter is quite relevant for this application, as retouching or tampering is only present in specific regions. This can lead to large class imbalances in tampered and authentic patches. One of the major advantages of focal loss over cross entropy loss is that focal loss helps in localizing retouched or distinguishing objects/ textures in the patches.

2.2. Image Classification

The classification results on patches are combined to classify the images. The proposed algorithm looks at two methods of classification, namely thresholding and support vector machine (SVM). The predictions are preprocessed to take into account the differences in the number of patches predicted as fake. This is primarily because if retouching is only introduced in specific regions, the difference in the number of authentic patches would not be large. This is even more prevalent in image tampering techniques such as splicing and cloning where only specific regions are tampered. As the CNN network performs better at classifying original patches, we focus on detecting differences in tampered patches. The prediction of the CNN pipeline is therefore post-processed using:

$$\text{Output} = \frac{\text{Total no. of patches predicted as tampered}}{\text{Total no. of patches}} \times 100 \tag{4}$$

The ratio in the equation ensures that for images of different size, a common threshold can be obtained. For large differences in image size, a common threshold of tampered patches would not be able to distinguish whether the difference is due to alterations or size difference. This is the motivation behind using the above mentioned method which normalizes all values for efficient classification.

For thresholding based classification, the best threshold is searched in the range of 1 to 10 using grid search, and the most optimal threshold is observed to be 4. For decision making, if an image contains more number of tampered patches than the threshold, the image is classified as retouched or fake. Otherwise, it is classified as authentic.

For SVM based classification, radial basis function kernel is used for training the SVM. The input to the SVM is the output obtained from equation (4). As retouching has been introduced in non-facial regions of the image in the ND-IIITD database [5], all the patches of the retouched images are considered as tampered. For GANs based evaluation, specific regions that are classified using GANs are labeled as tampered. SVM based classification learns the decision boundary based on the training samples and has shown superior results compared to thresholding based approach.

2.3. Implementation Details

The model weights are initialized using Xavier's method as shown by Glorot and Bengio [10]. L1 regularization is used because of its robustness and its ability to select features. Normalization is performed as a preprocessing step to make the data comparable across all the features. To prevent the issue of internal covariate shift while training deep neural networks we normalize the data in each mini-batch.

Figure 4. Original and retouched image from each probe. (i) original image; retouched image from (ii) probe 1, (iii) probe 2, (iv) probe 3, (v) probe 4, (vi) probe 5, (vii) probe 6, (viii) probe 7.

ReLU activation function [19] is used due to its ability to accelerate the convergence of stochastic gradient descent. The CNN network is pre-trained with training data, while testing the test patches are passed through the pre-trained network to get the predictions of the patches. A portion of the training data is used to train the support vector machine and to grid search the threshold value.

3. Dataset

The paper presents results on two sets of fake images: retouched images and generated images using StarGAN [7]. For the former, the algorithm is trained on the ND-IIITD [5] dataset.

3.1. ND-IIITD Dataset

The dataset consists of total 4875 face images, out of which 2600 are original collected from the Notre Dame database, Collection B [9] and 2275 are retouched. The alterations have been introduced using a sophisticated software, PortraitPro Studio Max. There are seven sets of probe images each varying in terms of the characteristics and extent of retouching. In the first two probe sets the level of alterations are lesser compared to other probes and they are only present in local regions. This increases with different probes, with the 7th probe having maximum deviation from the original images. Each probe set contains 325 facial images out of which 211 are males and 114 are females. The following protocols are used for classification:

1. To be consistent with the protocols followed in literature, 50% train-test split protocol is followed. The model is tested on 106 males and 57 females from each retouched and original probe.

2. To learn intra-probe (preset) variations the algorithm is trained on 50% of the images within a probe (preset) consisting of about 185,000 blocks of size (64,64,3). This is tested on the remaining 163 images of the same preset.

3. To determine whether the trained models are generalizable across different kinds of probe, the model is learned on the 7th probe and tested on the rest. In other words, 325 retouched and 325 original images from probe 7 are used for training, and the remaining dataset is used for testing.

3.2. Generated Images

StarGANs [7] was trained using the CelebA dataset to learn attribute transfer, such as black hair; blond hair; brown hair; gender; aging; hair and gender together; hair and age together; age and gender together; hair, age and gender together. Using StarGANs, 18,000 images are created corresponding to a set of 2000 face images and 9 attributes. Additionally, 15500 original non-overlapping images from the CelebA dataset are used. All the images are of size (128,128,3) and are cropped from the center of the original CelebA dataset images. The GANs model is trained for 20 epochs. Barring some generated images, the rest can deceive the human eye in terms of whether they are generated or original. They are exceedingly similar to original images in terms of facial features, textures and colors. The total database thus comprises 35,500 images. For automated classification of generated images, 2500 images are used for testing, in which 1500 images are from the authentic class and 1000 are from the generated class. Further, 500 images from each class are used for validation. The model is trained on the remaining 32000 images.

4. Results

The result section is divided into three parts according to the experiments. The first subsection discusses the results of detecting facial retouching on the ND-IIITD database and comparison to state-of-the-art reported in literature. The second subsection reports the results of the proposed algorithm in detecting synthetic alterations made using GANs. The third and last subsection examines the impact of image compression on the performance of the proposed algorithm.

4.1. Detecting Facial Retouching

The proposed architecture with SVM classification yields an overall accuracy of 99.65% for protocol 1. The class-based accuracies range from 99.38% to 100% for probe 1 to 7, respectively. Setting a threshold manually gives slightly lower classification result of 99.48%. On the same database, Bharati et al. [5] achieved 87.1% and Kee and Farid [12] achieved 48.8% accuracy. The same experiment is also performed using image patches of size (64,64,3). It yields an accuracy of 99.42% with SVM and

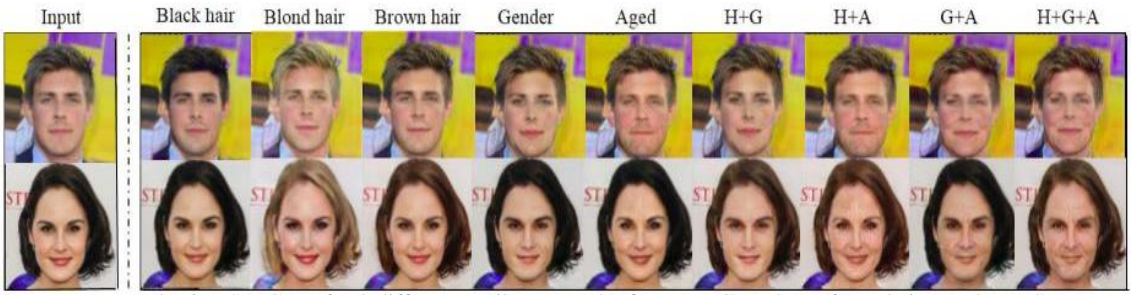

Figure 5. Images generated using StarGAN for 9 different attributes. In the figure, H,G, and A refer to hair, gender and age respectively

Table 1. Retouching detection accuracy when training and testing sets pertain to the same database

Input	**Method**	**Probe 1**	**Probe 2**	**Probe 3**	**Probe 4**	**Probe 5**	**Probe 6**	**Probe 7**
Images	Thresholding	98.75	98.44	98.44	96.89	95.11	99.68	96.89
Images	SVM	97.90	97.95	98.96	97.95	98.75	98.96	93.88
Patches	CNN	99.51	99.22	99.02	99.54	99.53	99.79	99.41


Figure 6. Some examples of misclassified patches

Table 2. Normalized confusion matrix for patch predictions using CNN for detecting retouching and generated images.

			Predicted Labels	
			Real	Fake
True Labels	**Retouching**	Real	0.99	0.01
		Fake	0.01	0.99
	Generated	Real	1.00	0.00
		Fake	0.04	0.96

99.70% while using thresholding. These results are summarized in Table 3. In this experiment, all the authentic patches are correctly classified by the CNN architecture. Some of the misclassified patches of size (128,128,3) are shown in Figure 6. For retouched images, the architecture does not perform well on patches which contained either clothing articles or hair. This is because the retouching texture differences gets hidden in clothing or hair texture.

The network is tested for its performance in detecting inter-probe and intra-probe variations. For intra-probe variations, protocol 2 is followed and it achieved an accuracy of around 98% for probes 1,2 and 3; 96% for probes 4 and 5; and slightly higher accuracy for the probe 6, i.e., 99%. However, the predicted accuracy for probe 7 is slightly lower than others. It yields an accuracy of about 96.89% using thresholding and 93.88% using SVM for image classification from patches. This shows that the data required to train the model is not very high because of the patch based approach and moreover this performs well with individual probes. The results are summarized in Table 1. The accuracies are lesser as compared to the overall accuracy on the dataset, due to scarcity of data for training for individual probes.

To test the network for inter-probe variations, one probe set is used for training and the others are used for testing. As probe 7 contains maximum retouching, it is used for training. The overall testing accuracy on probe 1-6 is 99.73% using thresholding and 99.91% using SVM.

To assert the importance of the residual connection, it is ablated (i.e. removed). A significant dip in the performance of the model is observed. The patch based accuracy is only 97.83% and the classification accuracy of the image using thresholding is 96.80% and 99% using SVM. The SVM model is finding a high threshold value for distinguishing the two classes as the number of misclassified patches increases. Thus, to ensure a lower false positive rate (FPR), the residual connection is required in the model.

4.2. Detecting Generated Images

The overall accuracy of the proposed algorithm on generated images from STARGAN [7] is 99.83% using thresholding and 99.73% using support vector machine for image classification from patches. The images are converted to JPG format to test the models performance in presence of lossy compression. The images are compressed using a quality factor of 50. The patch classification accuracy on compressed images is 95.6% and image classification using SVM yields an accuracy of 96.33%. The accuracy for JPG compressed images using thresholding is significantly

Table 3. Overall image retouching detection accuracy and comparison with existing reported results in literature

Algorithm	Accuracy
Kee and Farid [12]	48.8%
Bharati (Unsupervised DBM) [5]	81.9%
Bharati (Supervised DBM) [5]	87.1%
Proposed (Thresholding) - (64,64,3)	99.70%
Proposed (SVM) - (64,64,3)	99.42%
Proposed (Thresholding) - (128,128,3)	99.48%
Proposed (SVM) - (128,128,3)	99.65%

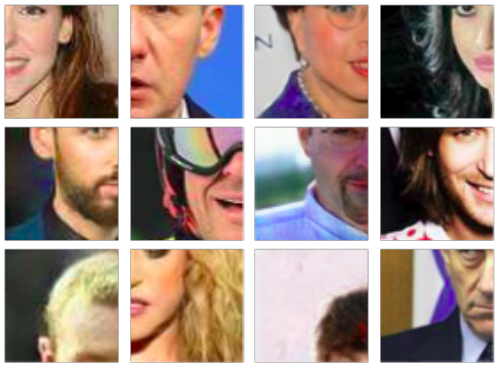

Figure 7. Column 1: Correctly classified authentic patches; Column 2: Correctly classified generated patches; Column 3: Misclassified authentic patches; Column 4: Misclassified generated patches

Table 4. Classification accuracies for generated images for different image formats

Compression	SVM	Thresholding
PNG images	99.73%	99.83%
JPG images	96,33%	88.89%

Table 5. Overall generated image detection accuracy and comparison with other algorithms for PNG format images

Algorithm	Overall Accuracy
Bharati [5]	91.83%
Proposed (Thresholding)	99.83%
Proposed (SVM)	99.73%

lesser, 88.89%. One of the major reason is the higher false positive rate. This also shows the added advantage of using support vector machines for predicting the labels of images as generated or real. These results are summarized in Table 4 below. Figure 7 shows some examples for patches which are correctly and falsely classified. Table 5 compares the performance of proposed algorithm for detecting synthetic alterations with Bharati et al. [5] which yields an accuracy of 91.83%.

4.3. Compression Analysis for Retouching

Deep learning models perform well in detecting double compressed JPG images [14, 18]. Introducing tampering/retouching in an already compressed JPG image followed by re-compression afterwards leads to double compression. To ensure that the model is not learning such variations, JPG format images present in the retouching dataset are converted to PNG format and the model is fine tuned for PNG format images. Similar results confirmed that the network is not learning the properties of image compression.

5. Conclusion and Future Work

The paper presents a convolutional neural network architecture for detecting digital manipulations in terms of retouching and GANs based alterations. The results are demonstrated on two databases: ND-IIITD database and images generated using StarGANs. The proposed algorithm shows significant improvements compared to the results reported in the literature. This paper also opens another possible avenue for research in classification of generated images and testing photorealism as well. While this paper analyzes the images generated using StarGANs, the idea of unified automatic detection could be an interesting extension.

6. Acknowledgement

Vatsa and Singh are partly supported through Infosys Center for Artificial Intelligence at IIIT-Delhi. The research is also partly supported through a grant from MEITY, Government of India.

References

[1] Age Defying Techniques. `https://bit.ly/2LzDpLa`. Accessed: 2018-04-26.

[2] Jennifer Lopez Retouched Image. `https://bit.ly/2NWdn1A`. Accessed: 2018-04-26.

[3] New Israeli law bans use of too-skinny models in ads. `https://cnn.it/1mNTiY1`. Accessed: 2018-04-26.

[4] Ralph Lauren apologises for digitally retouching slender model to make her head look bigger than her waist. `https://dailym.ai/1zOvgAh`. Accessed: 2018-04-18.

[5] A. Bharati, R. Singh, M. Vatsa, and K. W. Bowyer. Detecting facial retouching using supervised deep learning. *IEEE Transactions on Information Forensics and Security*, 11(9):1903–1913, Sept 2016.

[6] A. Bharati, M. Vatsa, R. Singh, K. W. Bowyer, and X. Tong. Demography-based facial retouching detection using subclass supervised sparse autoencoder. *CoRR*, abs/1709.07598, 2017.

[7] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *arXiv preprint arXiv:1711.09020*, 2017.

[8] M. Ferrara, A. Franco, D. Maltoni, and Y. Sun. On the impact of alterations on face photo recognition accuracy. In A. Petrosino, editor, *Image Analysis and Processing*, pages 743–751. Springer Berlin Heidelberg, 2013.

[9] P. J. Flynn, K. W. Bowyer, and P. J. Phillips. Assessment of time dependency in face recognition: An initial study. In J. Kittler and M. S. Nixon, editors, *Audio- and Video-Based Biometric Person Authentication*, pages 44–51. Springer Berlin Heidelberg, 2003.

[10] X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[11] P. Isola, J. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CoRR*, abs/1611.07004, 2016.

[12] E. Kee and H. Farid. A perceptual metric for photo retouching. *Proceedings of the National Academy of Sciences*, 108(50):19907–19912, 2011.

[13] N. Kose, L. Apvrille, and J. L. Dugelay. Facial makeup detection technique based on texture and shape analysis. In *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, May 2015.

[14] B. Li, H. Luo, H. Zhang, S. Tan, and Z. Ji. A multi-branch convolutional neural network for detecting double JPEG compression. *CoRR*, abs/1710.05477, 2017.

[15] S. Russello. The impact of media exposure on self-esteem and body satisfaction in men and women. *Journal of Interdisciplinary Undergraduate Research*, 1(1):4, 2009.

[16] A. Singh, S. Tiwari, and S. K. Singh. Face tampering detection from single face image using gradient method. *International Journal of Security and its Applications*, 7(1), 2013.

[17] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, volume 4, page 12, 2017.

[18] Q. Wang and R. Zhang. Double jpeg compression forensics based on a convolutional neural network. *EURASIP Journal on Information Security*, 2016(1):23, Oct 2016.

[19] B. Xu, N. Wang, T. Chen, and M. Li. Empirical evaluation of rectified activations in convolutional network. *arXiv preprint arXiv:1505.00853*, 2015.

[20] S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[21] J. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *CoRR*, abs/1703.10593, 2017.