Detecting GANs and Retouching based Digital Alterations via DAD-HCNN

Anubhav Jain*, Puspita Majumdar*
IIIT-Delhi, India
{anubhav15129, pushpitam}@iiitd.ac.in

Richa Singh, Mayank Vatsa
IIT Jodhpur, India
{richa, mvatsa}@iitj.ac.in

Abstract

While image generation and editing technologies such as Generative Adversarial Networks and Photoshop are being used for creative and positive applications, the misuse of these technologies to create negative applications including Deep-nude and fake news is also increasing at a rampant pace. Therefore, detecting digitally created and digitally altered images is of paramount importance. This paper proposes a hierarchical approach termed as DAD-HCNN which performs two-fold task: (i) it differentiates between digitally generated images and digitally retouched images from the original unaltered images, and (ii) to increase the explainability of the decision, it also identifies the GAN architecture used to create the image. The effectiveness of the model is demonstrated on a database generated by combining face images generated from four different GAN architectures along with the retouched images and original images from existing benchmark databases.

1. Introduction

The availability and affordability of digital cameras have led to the exorbitant generation and usage of images in digital media. Every day millions of images are uploaded or shared through social media channels. According to a survey [1], 68% of adults retouch their images before sharing or posting them on any online platform. With the rapid development of easy-to-use image editing tools, generation of tampered and altered images has become an easy task even for novice users. Along with these "handcrafted" tools, Generative Adversarial Networks (GANs) based tools are also becoming popular [14, 19, 30]. The data-driven deep learning approaches, nowadays, do not require any human level expertise. Once trained, they can automatically retouch the images [7] and add different effects such as changing gender [8] and ethnicity effects. Figure 1 shows samples generated using sophisticated image editing tools to alter/update the image such that it is difficult to visually differentiate a real image from an altered image.

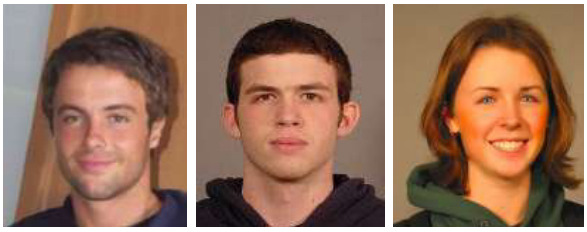


Figure 1. Guess which of these images are original, retouched or generated using GANs? Answer is available at Page 2.

While majority of these images are created for fun, they may be used for deception [25] with malicious intent. For instance, such images can be used for spreading fake news. GANs have been used to create DeepNude, show celebrities with pornographic content by generating an individual's face that closely matches with another face in the video. Fake videos of Mr. Barack Obama were widely circulated on the Internet [34]. To facilitate research in the area of fake image detection, Facebook has recently organized the Deepfake Detection Challenge (DFDC) [2]. Retouching using both handcrafted methods and GANs can affect the biometric identification process as well [5, 10]. These alterations can be undertaken in real-time by swapping faces along with their facial expressions [3]. Some of these ill-intended applications of the society not only has an effect on law enforcement scenarios but can also lead to significant psychological and sociological implications [31].

The adverse consequences of digitally altered images demands an automatic system for the detection and classification of these images. Detection and proper classification of digitally altered images is essential in helping law enforcement agencies in investigation, solving the psychological and sociological issues, and in proper biometric identification process.

1.1. Related Work

Both retouching detection and GAN generated image detection have received increasing interest from the research community. Kee et al. [16] used geometric and photometric features to train non-linear Support Vector Regression

*Equal contribution by the student authors.

Figure 2. With reference to Figure 1, (a) represents the example of a generated image with its corresponding original image shown in bottom, (b) is the original image, (c) shows retouched image with its original counterpart at the bottom.

(SVR) on several celebrity images for detecting the extent of photo retouching. Bharati et al. [5] used a supervised deep Boltzmann machine algorithm for detecting facial retouching. This paper introduced the ND-IIITD dataset with 2600 original and 2275 retouched facial images. Multiple experiments are performed to show the poor recognition performance of commercial off-the-shelf (COTS) and OpenBR face recognition system on retouched face images. Another work by Bharati et al. [6] have shown the variations in the accuracy of retouching detection algorithms with demography of the facial image. They also introduced the Multi-Demographic Retouched Faces (MDRF) dataset with two genders and three ethnicities. A semi-supervised autoencoder is used to improve the classification performance. Jain et al. [15] have proposed a deep learning based architecture to detect retouched images. Portmann et al. [29] detects whether a camera is performing automatic face beautification by comparing camera captured face and rearranged face images.

Recent research has been directed towards the detection of GANs based altered images. Researchers have used color cues [20, 26] for detecting the subtle differences in the images. McCloskey et al. [26] proposed an algorithm, where the Intensity Noise Histograms network is used to classify the histograms formed using R and G chromaticity coordinates as two variables. They trained a Support Vector Machine (SVM) using the features obtained by counting the number of saturated and under-exposed pixels. Li et al. [20] proposed a framework to detect GANs based alterations by training an SVM on the vector of co-occurrence matrix calculated on the high pass filter residuals of the image in RGB, HSV and YCbCr image space. Later, ensemble methods are used by Tariq et al. [36] to detect GANs generated images, and pre-processing techniques are used to detect human tampered images. Dang et al. [9] have proposed a customized convolutional neural network, termed as, CG-Face for GANs generated fake face detection. In order to detect fake videos, Korshunov and Marcel [17] have generated Deepfake videos from the videos of the VidTIMIT database [1] and have shown the effect on existing face recognition and detection algorithms. Li et al. [21] proposed a CNN based algorithm to detect DeepFake videos by learning the artifacts in affine face warping as the discriminating feature. Hsu et al. [13] proposed a deep forgery discriminator which concatenates two classifiers along with a contrastive loss for detecting GANs based fake images. Nataraj et al. [27] have detected GANs generated images using a combination of co-occurrence matrices on the three color channels and deep convolutional neural networks. Li and Lyu [22] proposed a deep learning model that uses affine face wrapping artifacts for deepfake detection. Instead of using deepfake images as negative examples, the proposed approach simulated these images using image processing operations to create artifacts that exist in deepfake content. Amerini et al. [4] detected deepfake videos by exploiting inter-frame dissimilarities, unlike previous work that focuses on single-frame detection. The paper proposed an optical flow-based CNN model to perform this task by using PWC-Net [33] model. Kumar et al. [18] detected face2face facial reenactment in videos using a multi-stream VGGNet based network to detect regional artifacts. Singh et al. [32] discussed the effect of doctored/tampered images generated using GANs on face recognition systems.

1.2. Research Contributions

In the literature, algorithms have been developed for detecting one type of alteration at a time and for closed set attack/alteration detection. However, real-world solutions require a single algorithm to detect multiple alterations. Further, in some cases it could be an unknown attack. The aim of this research is to detect and distinguish between learning (GANs) based digital alterations/generation, handcrafted retouching based alterations, and non-tampered (original) images. For this purpose, a novel framework, Digital Alteration Detection using Hierarchical Convolutional Neural Network (DAD-HCNN) is proposed. The key contributions of this research are:

- Proposing DAD-HCNN framework with three levels of hierarchy to broadly classify an input image as original or digitally altered followed by further classification in the digitally altered class at the subsequent levels of the proposed framework.
- Evaluating the effectiveness of the proposed DAD-HCNN framework by performing experiments on the

[1] http://conradsanderson.id.au/vidtimit/

CMU Multi-PIE [12] and ND-IIITD datasets [5] along with the images generated using different models of GANs [8, 19, 28, 30].

- Analyzing the robustness of the proposed framework in detecting images generated using unknown models of GANs (not seen during training phase) by performing cross-model experiments.

2. Proposed DAD-HCNN Framework

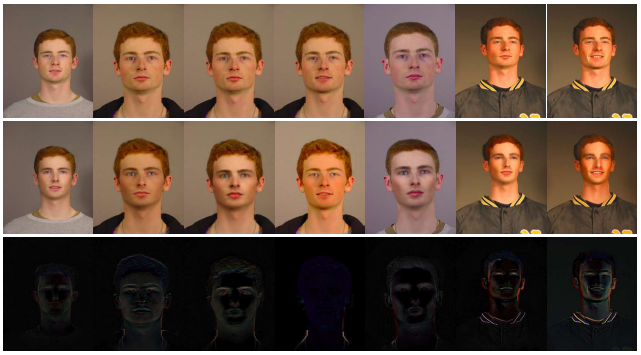
This paper proposes a framework, Digital Alteration Detection using Hierarchical Convolutional Neural Network (DAD-HCNN), for detecting original and digitally altered images, using a three level hierarchical approach. The objective is not only to classify original versus altered images but also to differentiate between retouched versus GANs generated images along with the classification of images generated using different models of GANs. Figure 3 shows the intensity difference map representation for retouching and GANs based digital alterations. It shows that there is a large variability across different kinds of images and models and it is therefore important to learn even the subtle differences caused by different methods of digital alterations. The proposed framework is thus arranged as a 3-level hierarchical network where each level is trained for a specific task.

Block diagram of the proposed DAD-HCNN framework and the expanded view of each level is shown in Figure 4. The CNN network at each level of DAD-HCNN framework consists of five convolutional layers with a wide residual connection. The architecture uses a residual connection [35] which allows deeper neural networks to be effectively trained. Zagoruyko et al. [37] have shown that ResNet [35] performs better when they are wider. Therefore, in this paper, a convolutional block has been introduced into the residual connection of the CNN. Each individual CNN is trained using a patch-based approach with non-overlapping RGB patches of size (64,64,3).

2.1. Level-1 Classification for Original vs Altered

The first level of DAD-HCNN framework performs classification of original and digitally altered images. The CNN and SVM at level-1 are trained with the images of original and digitally altered class. The digitally altered class contains images of the retouched class and the images generated using four different models of GANs, namely, StarGAN [8], SRGAN [19], DCGAN [30], and Context Encoder [28]. The training process of CNN and SVM is discussed below.

Training Convolutional Neural Network: Let L_1 be the level to predict two classes, namely, C_1 and C_2, where, C_1 represents the 'Original' class and C_2 represents the 'Altered' class. Let $\mathbf{X}$ represent the training set with n number of images.



(a)

(b)

Figure 3. Intensity difference maps for the (a) retouched images (ND-IIITD dataset [5]) with first row containing original images, second row containing retouched images, third row containing the intensity difference between the original and its retouched counterpart and (b) GANs generated images with first row containing original images, second row containing images generated using StarGAN [8], third row containing the intensity difference between the original and its generated counterpart.

$$\mathbf{X} = \{\mathbf{X}_1, \mathbf{X}_2, ...\mathbf{X}_n\} \tag{1}$$

where, each image $\mathbf{X}_i$ is divided into non-overlapping patches of size $p \times p$. Let $\mathbf{Z}_i$ be the set of m number of patches corresponding to an image $\mathbf{X}_i$, where m can vary depending upon the size of the input image $\mathbf{X}_i$.

$$\mathbf{Z}_i = \{\mathbf{Z}_{i,1}, \mathbf{Z}_{i,2}, ...\mathbf{Z}_{i,m}\} \tag{2}$$

where, each $\mathbf{Z}_{i,j}$ represents a non-overlapping patch of size $p \times p$ corresponding to an image $\mathbf{X}_i$. The probability of predicting an input patch $\mathbf{Z}_{i,j}$ to class C_k by level L_1 is represented as:

$$P(C_k|\mathbf{Z}_{i,j}) = \phi(\mathbf{Z}_{i,j}, \mathbf{W}, b, L_1) \tag{3}$$

where, $C_k \in C_1, C_2$. $\mathbf{W}$ is the weight matrix and b is the bias. Let $\mathbf{Y}_{i,j}$ represent the true class of the patch $\mathbf{Z}_{i,j}$ in one hot encoding form. Focal loss [23] is used to train the CNN. The loss function can be represented as:

$$Loss = f(C_k, \mathbf{Z}_{i,j}, \mathbf{Y}_{i,j}) \tag{4}$$

$$f(C_k, \mathbf{Z}_{i,j}, \mathbf{Y}_{i,j}) = -\alpha(1 - \mathbf{Y}^t_{i,j} P(C_k|\mathbf{Z}_{i,j}))^\gamma \log(\mathbf{Y}^t_{i,j} P(C_k|\mathbf{Z}_{i,j})) \tag{5}$$

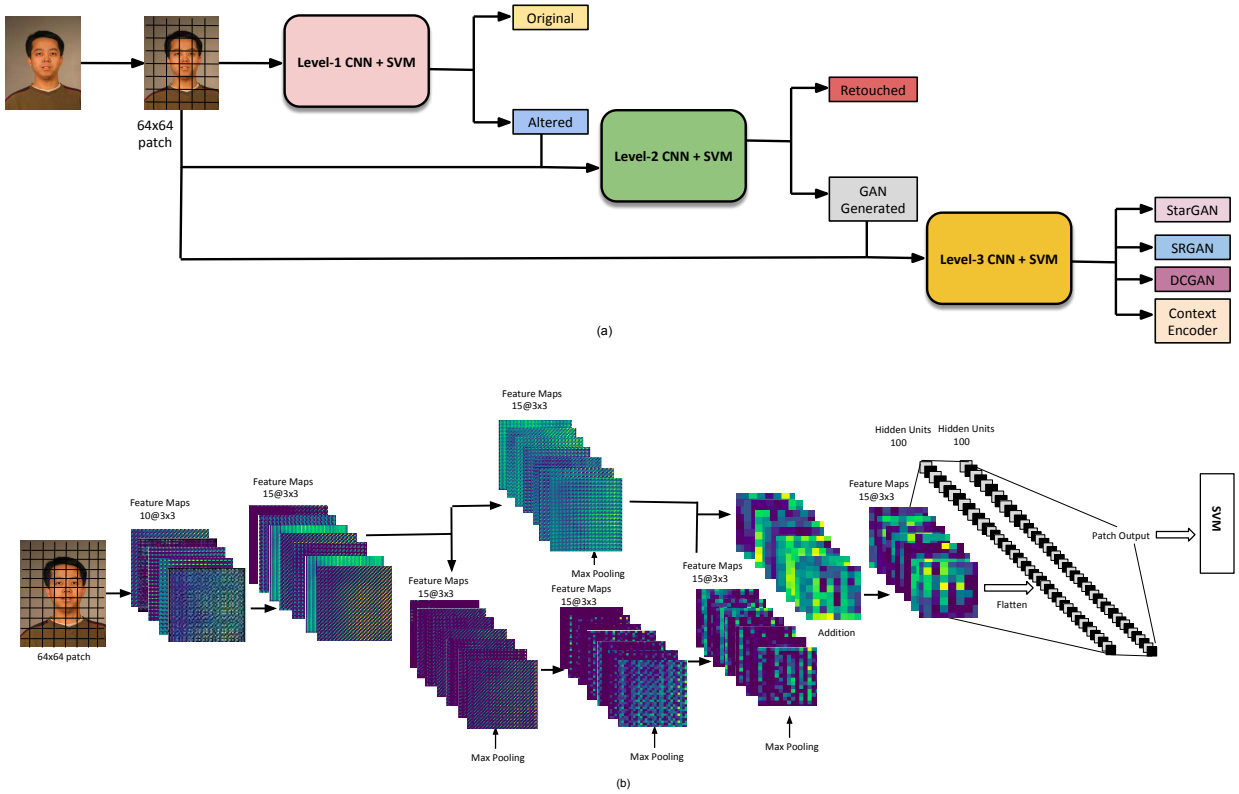


Figure 4. Illustrating the steps involved in the DAD-HCNN. (a) The first level is trained to distinguish between original and digitally altered images. The second level is trained for the classification among digitally altered images to classify retouched and GANs generated images as two separate classes. Third block is trained to further differentiate among the images generated using different models of GANs. (b) Expanded view of the layers of the CNN network at each level of DAD-HCNN framework.

where, $\mathbf{Y}^t_{i,j} P(C_k|\mathbf{Z}_{i,j}) \in [0,1]$ is the probability for the prediction of the positive class. α is taken as 1 and the trainable parameter γ is kept as 5 in all the experiments. Focal loss [23] helps in the localization of objects or regions and is thus suitable for the problem.

Training Support Vector Machine: The patch based predictions of the pre-trained CNN at level-1 of the DAD-HCNN framework are further used for overall image classification. Therefore, an SVM with 'rbf' kernel is trained on top of the predictions of the CNN. To train the SVM, a feature vector is formed using the patch based predictions of the pre-trained CNN. The steps involved in forming the feature vector are discussed below.

Let the predictions of the network be represented by the matrix $\mathbf{S}$, where each row of $\mathbf{S}$ denotes the prediction (logits) of input patch $\mathbf{Z}_{i,j}$. The dimension of matrix $\mathbf{S}$ is (m, k) where m is the total number of patches in $\mathbf{X_i}$ and k is the total number of classes. At level-1, k=2 namely, C_1 and C_2 described in the training process of the CNN. In the next step, the total number of patches predicted as class C_k is normalized using the total number of patches in the input image $\mathbf{X}_i$. Mathematically it is represented as:

$$\mathbf{r_i} = \frac{\sum_{p=1}^{m} \mathbf{S}_{pk}}{m} \qquad \forall k \tag{6}$$

where, $\mathbf{r_i}$ is the normalized feature vector corresponding to each image $\mathbf{X}_i$ that is used by the SVM for image level classification and $\mathbf{S_{pk}}$ denotes the belongingness of a patch corresponding to image $\mathbf{X_i}$ to class k. In the multi-class scenario, images from different classes may have different sizes. In such cases, there is a high probability for the classifier to learn the size difference of the images for distinguishing altered images from the original ones. Therefore, it is important to normalize the feature vector based on the total number of patches m corresponding to the images of each class.

2.2. Level-2 Classification for Retouching vs GANs

The second level of DAD-HCNN framework is used for classifying the images from the digitally altered class into retouched and GANs generated images. After the training of level-1 CNN + SVM, the layers of CNN in level-1 are frozen and level-2 CNN + SVM is trained using the same formulation discussed above. Level-2 is trained with two

Table 1. Protocols used for training CNN + SVM at three different levels of the proposed DAD-HCNN framework for classification.

Experiment	Details				
Experiment 1	Level-1 classification: Distinguishing between original and digitally altered images				
	Level	Class	Training CNN		Training SVM
			# Training Images	# Training Patches	# Training Images
	Level-1 CNN+SVM	Original	2,480	145,738	2,000
		Altered	15,460	153,012	2,000
Experiment 2	Level-2 classification: Distinguishing between retouched with GANs generated images				
	Level-2 CNN+SVM	Retouched	210	109,697	180
		Generated	21,900	132,053	400
Experiment 3	Level-3 classification: Distinguishing between different GANs generated images				
	Level-3 CNN+SVM	StarGAN	17,000	67,597	100
		SRGAN	7,000	82,799	100
		DCGAN	6,200	5,201	100
		Context Encoders	16,800	67,199	100

classes, namely, 'Retouched' class and 'GANs generated image' class. The GANs generated image class contains images generated using StarGAN, SRGAN, DCGAN, and Context Encoder.

2.3. Level-3 Classification for GANs Prediction

The third level of DAD-HCNN framework predicts the GAN used to generate the images. During the training of level-3 CNN + SVM, layers of the CNN in level-1 and level-2 are frozen. The training process follows the same formulation discussed in Section 2.1. Level-3 is trained with four classes, namely, 'StarGAN', 'SRGAN', 'DCGAN', and 'Context Encoder' class.

2.4. Implementation Details

Each CNN of the proposed DAD-HCNN framework is trained using ReLU activation function. The weights of the networks are initialized using Xavier initialization, and batch normalization is performed after every layer. ℓ_1 regularization has been used as it introduces sparsity and is more robust to outliers. Adam optimizer is used with a learning rate of 0.001 and a decay rate of 0.00001. During testing, the input image is divided into patches of size (64,64,3) and given as input to the first level. Depending upon the decision of the first level, the image is further given as input to the subsequent levels. Thus, at each level, a decision is made for the input image. The final classification is therefore dependent on the decision of each level.

3. Experiments and Results

Experiments are performed on multiple datasets and images generated using different models of GANs. Three different experiments are performed to evaluated the performance of the proposed DAD-HCNN framework. First experiment is performed to showcase the classification performance of the proposed DAD-HCNN framework. Second experiment is performed to compare the performance of a GAN discriminator in detecting digital alterations with the proposed DAD-HCNN framework. Third experiment is performed to evaluate the robustness of DAD-HCNN on the images generated using unseen models of GANs. The following subsections discuss the datasets used and experiments performed in detail.

3.1. Details of the Datasets and Generated Images

CMU Multi-PIE dataset [12] contains more than 75,000 images of 337 subjects. The images are taken under different pose, illumination and expression variations.

ND-IIITD dataset [5] contains 2,600 original and 2275 retouched facial images having a total of 4,875 images. PortraitPro Studio Max software is used for applying various retouching operations on original face images from the Notre Dame database, Collection B [11].

StarGAN [8] is trained on the CelebA dataset [24] to learn the transfer of attributes. Nine different attributes namely: black hair; blond hair; brown hair; gender; age; hair and gender combined; hair and age combined; age and gender combined; hair, age, and gender combined are learned corresponding to each image. 2,000 images are used for generating images using the network resulting in a total of 18,000 images. Sample images are shown in Figure 5(a).

SRGAN [19] is trained on the CelebA dataset to generate high-resolution images from its low-resolution counterpart. 14725 low-resolution images are used to generate a total of 14725 high-resolution images. Figure 5(b) shows sample images generated using SRGAN along with their respective original images.

DCGAN [30] takes a random noise as input to generate realistic images. The model is trained on the CelebA dataset and a total of 5700 images are generated. Sample images are shown in Figure 5(c).

Context Encoders [28] is trained to generate an image region conditioned on its surroundings. It is trained on the CelebA database and 113,760 images are generated. Figure 5(d) shows sample images generated using context en-

Figure 5. Images generated using different models of GANs. (a) First column contains original images. The next nine columns contain images generated using StarGAN [8] by changing different attributes, (b) Original (first row) and generated (second row) images using SRGAN [19], (c) Images generated using DCGAN [30], and (d) First row contains original images with mask showing the region to be reconstructed using Context Encoders [28], second row shows the generated images and the third row contains the original images.

coders.

Original images from the ND-IIITD and CMU Multi-PIE datasets form the 'Original' class. Retouched images of the ND-IIITD dataset comprise the 'Retouched' class. Images generated using different GANs models are used for 'GANs generated image' class. Table 1 summarizes the details of the number of images and patches used for training the CNN and the number of images used for training SVM at each level of the DAD-HCNN framework.

3.2. Classification Results

Three experiments are performed to evaluate the performance of the proposed DAD-HCNN framework for classification. The details and results of the experiments are discussed below.

Results for Level-1 Classification - Distinguishing between original and digitally altered images: As shown in Figure 4(a), level-1 of the proposed DAD-HCNN framework is evaluated for classifying whether the image is original or altered. The performance of this level is evaluated on 4000 images, with 2000 images of each class.

It is observed that level-1 of the proposed DAD-HCNN framework achieves an accuracy of 99.95%. Patch-based classification accuracy is also computed using the predictions from the last layer of the network at level-1. On patches, an accuracy of 99.30% is obtained. The result is compared with Bharati et al. [5] that achieves an accuracy of 85.53%. It is observed that the proposed framework performs 13.77% and 14.42% better than [5] using patch-based and image-based approach, respectively. The results are summarized in Table 2.

Results for Level-2 Classification: Distinguishing between retouched and GANs generated images: In order to distinguish between retouched and GANs generated images, the CNN and SVM in the second level of DAD-HCNN framework are trained with the images of retouched and GAN generated classes. The parameters of CNN in the first level of DAD-HCNN framework are frozen and the CNN in the second level is trained. During testing 1000 images are taken from each generative model except DCGAN from which 200 images are used. Along with this, 1000 images of retouched and original classes are also taken. The input

Table 2. Detection accuracy of the proposed DAD-HCNN framework and comparison with other algorithms.

Algorithm	Accuracy (%)
Bharati et al. [5]	85.53
Proposed (Patch Based)	99.30
Proposed (Image Based)	99.95

test image is first divided into patches and given as input to the first level of the proposed framework. Depending upon the decision of the first level on all the patches of an image, it is forwarded to the second level for further processing. The final result is the combined decision of both the levels on all the patches of the input image. The overall accuracy obtained is 99.68% with 99.43% corresponding to the retouched class, and 99.93% corresponding to the GANs generated image class.

Results for Level-3 Classification: Distinguishing between different GANs generated images: This experiment is performed considering the real-world scenario of identifying a digitally altered image along with the source of generation. For this purpose, the CNNs in the first two levels of the proposed DAD-HCNN framework are frozen and the CNN in the third level is trained to distinguish between images generated from different GANs. To test the performance of the DAD-HCNN framework in determining the source of the GANs generated images, the input test image is first divided into patches and given as input to the first level. Depending on the decision of the first level, the patch is forwarded to the next level for further processing and so on. The final result is the combined decision of all the levels on all the patches of the input image. 1000 images are taken from each class namely original, retouched, StarGAN, SRGAN, and Context Encoders with 200 images from DCGAN.

The proposed DAD-HCNN framework achieves an overall accuracy of 99.86% to determine the source of the generated image with 100% accuracy corresponding to the class of images generated using StarGAN, 99.69% corresponding to SRGAN, 99.97% corresponding to DCGAN, and 99.79% corresponding to Context Encoders. The high classification accuracy shows the suitability of the framework in determining the source of the generated images.

Figure 6 shows samples of the correctly classified and misclassified patches by the proposed DAD-HCNN framework. It is observed that most of the patches being misclassified from the CMU Multi-PIE dataset are the ones with poor illumination. Poor illumination leads to lower pixel values which in turn hide the textural properties of the images. It is also observed that the majority of the patches being correctly classified are the ones containing facial regions. Similarly retouched and GANs generated patches being misclassified contains mainly non-facial regions such as clothes, hats, and hairs.

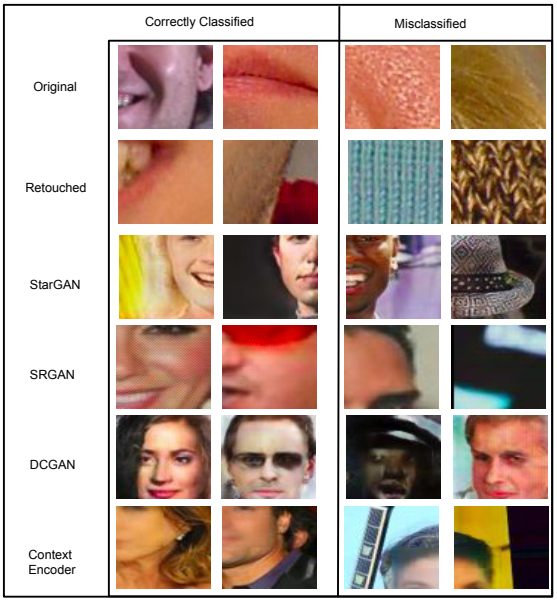


Figure 6. Samples of correctly classified and mis-classified patches of all the classes at different levels of the DAD-HCNN framework.

3.3. Comparing GAN Discriminator with DAD-HCNN

We performed two different experiments to evaluate the performance of using GAN discriminator in this research problem. The discriminator of StarGAN is used to perform the experiments. The first experiment is performed to distinguish between original and StarGAN generated images. The original class contains images from the ND-IIITD and CMU Multi-PIE datasets, whereas the altered class contains images generated using StarGAN. A 50% train test split protocol is followed for performing the experiment. In this experiment, a single block of CNN and SVM of DAD-HCNN architecture is used. The discriminator achieves an accuracy of **97.29%** whereas, the proposed architecture yields an accuracy of **99.65%**.

The second experiment is performed for 3-class classification, where the aim is to distinguish between original, retouched, and GANs generated images by classifying them into three different classes. For this purpose, the last two layers of the discriminator are removed and feature vectors of the input images are extracted. Next, an SVM is trained on these extracted feature vectors to perform 3-class classification. In this experiment, the discriminator achieves a low accuracy of **62.00%**, whereas the proposed DAD-HCNN framework achieves **99.68%** accuracy. Since the discriminator of StarGAN is trained to distinguish the images generated using StarGAN, it is unable to effectively distinguish between retouched images and the images gen-

Table 3. Cross model classification accuracy using the proposed DAD-HCNN framework.

Models used for training	Models used for testing	Accuracy (%)
StarGAN, SRGAN	DCGAN	99.59
StarGAN, DCGAN	SRGAN	99.78
SRGAN, DCGAN	StarGAN	99.98

erated using other models of GANs.

3.4. Robustness Analysis

The experiments performed so far are based on some apriori knowledge about the type of alterations performed or the models used to perform the alterations. However, in a real-world scenario, it is not pragmatic to assume this kind of apriori knowledge. With the advent of new image editing tools and technology, it is possible that the test image contains alterations performed using a tool which is unseen by the model. Therefore, the detection framework must be robust to unseen alterations and newer models of generating digital images.

To evaluate the robustness of the DAD-HCNN framework in detecting digitally generated images by unseen methods, cross-model experiments are performed for binary classification between original and generated images. In these experiments, the CNN network is trained on the images generated using some models of GANs such as StarGAN and SRGAN. During testing, images generated other GANs models such as DCGAN are used to test for evaluation. For training, 2000 images from each of the two GAN models along with 2000 original images are used. For testing, 2000 images generated using an unseen model of GAN and 2000 original images are used. Table 3 shows the classification accuracy of cross model experiments. It is observed that the network achieves above 99% accuracy in all the experiments. This shows the efficiency and robustness of the network of the proposed DAD-HCNN framework in handling alterations made using unseen models of GANs.

3.5. Effect of Residual Connection

The importance of residual connections in the CNN of the proposed DAD-HCNN framework is established by ablating the residual connection and comparing the performance of the network. Figure 7 summarizes the experimental results. It is observed that ablation results in a decrease in the overall training and validation accuracy. This clearly shows the importance of residual connection (RC) in the network and suitability of the network components in the proposed DAD-HCNN framework.

4. Conclusion

This paper presents a framework to detect GANs and retouching based digital alterations. Existing research in the area of detecting digital alterations have been specific to detecting a particular alteration. The proposed hierarchical framework, termed as DAD-HCNN, can detect retouching and GANs generated images with high accuracy which shows its applicability in real life scenarios. Along with detecting alterations, the proposed approach also detects the source GAN model from which the image is generated/altered, i.e. it classifies the images generated using different models of GANs into different classes. Multiple experiments are performed to evaluate the performance of the proposed framework towards alteration detection under open and closed scenarios. The proposed DAD-HCNN showcases superlative performance in different settings and illustrates that digital alterations can be detected with very high confidence.

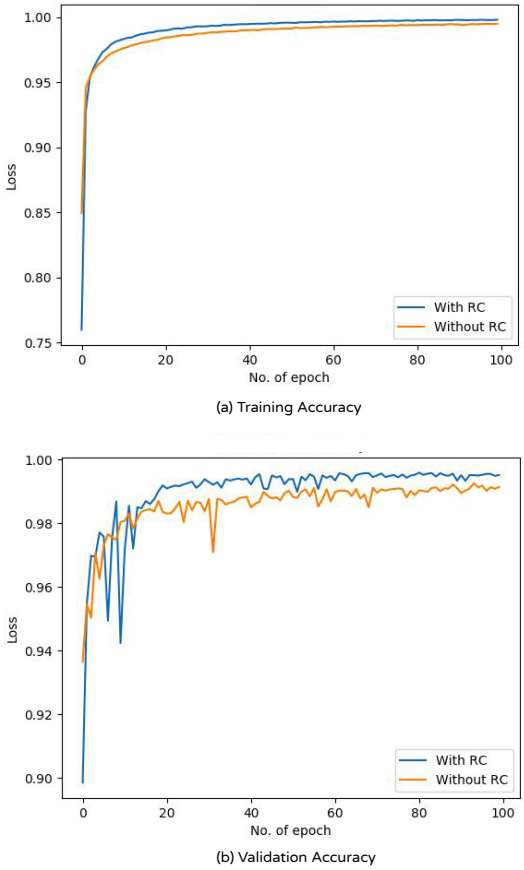


Figure 7. Training and validation accuracy for comparing the performance with and without residual connection (RC).

Acknowledgements

This work is supported through a research grant from MEITY. P. Majumdar is partly supported by DST Inspire PhD Fellowship. M. Vatsa is partially supported through Swarnajayanti Fellowship, Government of India.

References

[1] 68 Percent of Adults Edit Their Selfies Before Sharing Them With Anyone. https://bit.ly/2MYu6Vl. Accessed: 2018-08-31. 1

[2] Deepfake Detection Challenge. https://deepfakedetectionchallenge.ai. 1

[3] Watch a man manipulate George Bush's face in real time. https://bit.ly/2wVgNN4. Accessed: 2018-10-23. 1

[4] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. Deepfake video detection through optical flow based cnn. In *IEEE International Conference on Computer Vision Workshops*, 2019. 2

[5] Aparna Bharati, Richa Singh, Mayank Vatsa, and Kevin W Bowyer. Detecting facial retouching using supervised deep learning. *IEEE Transactions on Information Forensics and Security*, 11(9):1903–1913, 2016. 1, 2, 3, 5, 6, 7

[6] Aparna Bharati, Mayank Vatsa, Richa Singh, Kevin W Bowyer, and Xin Tong. Demography-based facial retouching detection using subclass supervised sparse autoencoder. In *IEEE International Joint Conference on Biometrics*, pages 474–482, 2017. 2

[7] Matthew Brand and Patrick Pletscher. A conditional random field for automatic photo editing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7, 2008. 1

[8] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797, 2018. 1, 3, 5, 6

[9] L Minh Dang, Syed Ibrahim Hassan, Suhyeon Im, Jaecheol Lee, Sujin Lee, and Hyeonjoon Moon. Deep learning based computer generated face identification using convolutional neural network. *Applied Sciences*, 8(12):2610, 2018. 2

[10] Matteo Ferrara, Annalisa Franco, Davide Maltoni, and Yunlian Sun. On the impact of alterations on face photo recognition accuracy. In *International Conference on Image Analysis and Processing*, pages 743–751. Springer, 2013. 1

[11] Patrick J. Flynn, Kevin W. Bowyer, and P. Jonathon Phillips. Assessment of time dependency in face recognition: An initial study. In Josef Kittler and Mark S. Nixon, editors, *Audio- and Video-Based Biometric Person Authentication*, pages 44–51. Springer Berlin Heidelberg, 2003. 5

[12] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-PIE. *Image and Vision Computing*, 28(5):807–813, 2010. 3, 5

[13] Chih-Chung Hsu, Chia-Yen Lee, and Yi-Xiu Zhuang. Learning to detect fake face images in the wild. In *IEEE International Symposium on Computer, Consumer and Control*, pages 388–391, 2018. 2

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1

[15] Anubhav Jain, Richa Singh, and Mayank Vatsa. On detecting gans and retouching based synthetic alterations. In *IEEE International Conference on, Biometrics Theory, Applications and Systems*, pages 1–7, 2018. 2

[16] Eric Kee and Hany Farid. A perceptual metric for photo retouching. *Proceedings of the National Academy of Sciences*, 108(50):19907–19912, 2011. 1

[17] Pavel Korshunov and Sébastien Marcel. Deepfakes: a new threat to face recognition? assessment and detection. *arXiv preprint arXiv:1812.08685*, 2018. 2

[18] Prabhat Kumar, Mayank Vatsa, and Richa Singh. Detecting face2face facial reenactment in videos. In *Winter Conference on Applications of Computer Vision*, pages 2589–2597, 2020. 2

[19] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4681–4690, 2017. 1, 3, 5, 6

[20] Haodong Li, Bin Li, Shunquan Tan, and Jiwu Huang. Detection of deep network generated images using disparities in color components. *arXiv preprint arXiv:1808.07276*, 2018. 2

[21] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 46–52, 2019. 2

[22] Yuezun Li and Siwei Lyu. Exposing DeepFake videos by detecting face warping artifacts. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 2

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE international conference on computer vision*, pages 2980–2988, 2017. 3, 4

[24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision*, December 2015. 5

[25] Puspita Majumdar, Akshay Agarwal, Richa Singh, and Mayank Vatsa. Evading face recognition via partial tampering of faces. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 1

[26] Scott McCloskey and Michael Albright. Detecting gan-generated imagery using color cues. *arXiv preprint arXiv:1812.08247*, 2018. 2

[27] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, BS Manjunath, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H Bappy, and Amit K Roy-Chowdhury. Detecting gan generated fake images using co-occurrence matrices. *Electronic Imaging*, 2019(5):532–1, 2019. 2

[28] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 3, 5, 6

[29] Clemenz Portmann, Yuriy Romanenko, and Vinit Modi. Detection of automated facial beautification by a camera application by comparing a face to a rearranged face. *Technical Disclosure Commons*, 2020. 2

[30] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 1, 3, 5, 6

[31] Salenna 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. 1

[32] Richa Singh, Akshay Agarwal, Maneet Singh, Shruti Nagpal, and Mayank Vatsa. On the robustness of face recognition algorithms against attacks and bias. In *AAAI Conference on Artificial Intelligence*, 2020. 2

[33] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8934–8943, 2018. 2

[34] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics*, 36(4):1–13, 2017. 1

[35] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Association for the Advancement of Artificial Intelligence*, 2017. 3

[36] Shahroz Tariq, Sangyup Lee, Hoyoung Kim, Youjin Shin, and Simon S Woo. Detecting both machine and human created fake face images in the wild. In *International Workshop on Multimedia Privacy and Security*, pages 81–87. ACM, 2018. 2

[37] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 3