Journal of Innovative Image Processing (ISSN: 2582-4252)
www.irojournals.com/iroiip/

Survey of Fake Image Synthesis and its Detection

Thiruvaazhi Uloli[1], R. M. Koushal Akash[2], A. G. Keerthika[3], K. S. Dhanwanth[4]

Information Science and Engineering, Kumaraguru College of Technology,Anna University, Coimbatore, India

E-mail: [1]thiruvaazhi.u.ise@kct.ac.in, [2]koushalakash.19is@kct.ac.in, [3]keerthika.19is@kct.ac.in, [4]dhanwanth.19is@kct.ac.in

Abstract

In recent times, image synthesis has attracted significant attention of people for both positive and negative reasons. Images can be easily synthesized using various techniques. This paper surveys various techniques for image synthesis as well as its detection in a unique structured manner, to enable a perspective on this iterative phenomenon. The paper describes both advantages and limitations starting from simple fake image detection to AI synthesized image detection approaches that are available in the literature. Generative Adversarial Network (GAN) is the trending algorithm for artificial image synthesis, because the faces generated by GAN are highly realistic. As discriminators are already present in the GAN's structure, any attempt to create a distinguisher that detects fake images synthesized by GAN, needs to structure itself to detect all existing patterns of fake image synthesis including that of GAN.

Keywords: GAN, Deepfake, CNN, deep learning, fake face detection

1. Introduction

In the current information age, digital images act as a major resource across many fields including movies, medicine, architecture, journalism, scientific publication, digital forensics, etc. and are used for numerous purposes. These images are being manipulated for hiding some important piece of information within that image, or for creating fake forged images which act as a threat. A digital image can be manipulated using tools for image editing like GNU Image Manipulation Program (GIMP) or Adobe Creative Cloud Suite, etc. Art forgery is one of the oldest manipulation techniques which was started well over two thousand years ago. Due to the Internet and its wide access, the availability of enormous number of photos and its editing tools over the cloud, anybody can possibly manipulate and

DOI: https://doi.org/10.36548/jiip.2022.4.006
Received: 15.11.2022, received in revised form: 24.12.2022, accepted: 10.01.2023, published: 21.01.2023

create the forgery of these digital images. This has led to various kinds of problems for different kinds of people including popular celebrities and for those of normal users, who post pictures on social media.

Digital and technological advancements, rather than providing solutions to these problems, in fact, is only aiding this process of creating fake images which for an unsuspecting user will for all practical purposes look like a true image. There is an increasing difficulty in distinguishing between a real image and a forged one. Image synthesis is the process used for artificially generating images that contains features desired by its author. It is described by [1] as the inverse of the classification problem. Here synthesis is used to create many images with visual contents that are associated with or can be classified with a specific label.

This paper is organized as follows: Section II presents the traditional (non-GAN or non-AI) methods of image synthesis. Section III describes the methods to detect such fake images. Section IV goes further in presenting the recent methods for synthesizing images using AI and GAN. Section V presents the available methods of detection such AI synthesized images. The results and the discussion are summarized in Section V, and Section VI concludes the paper.

2. Non-GAN / AI image synthesis

Morphing is defined as the process used for transforming states of appearance of a feature of image which involves the usage of various operations such as basic modification and rotation, changes in colour and texture, and shapeshifting [2]. Morphing can cause a high level of impact due to the features available in the variety of image editing tools. The idea of image processing itself is being kept under question due to the adverse effects of morphing. The rest of this section presents the traditional methods of image synthesis or fake image generation.

2.1 Use of Image Manipulation Tools

The true images are manipulated and transformed into desired output by using image processing / editing tools like Adobe Photoshop / GIMP. Any normal user does use a few of such tools for resizing the images or in this age of exploding usage of mobile phones and their apps, use various image filters to make them look in a variety of entertaining ways. There are several use cases for these positive manipulations. Sometimes they are used to

create art pieces involving new techniques which do not cause any threat. Some digital artists use Photoshop to create their workpieces which are a contribution to art. Several techniques and software are used for image manipulation. These software applications can be used by anybody, right from a normal user to an expert user. Though the utility of such image edits using these tools does not require detailed evidence, some applications of photograph manipulation is considered unethical because some manipulations are used to cheat the people who use social media and other platforms. Worser is the case where people use digital images of others and manipulate them in the ways to meet their ulterior objectives at the cost of innocent people or celebrities whose photos has been negatively manipulated.

2.2 Copy move forgery

This a simple technique proposed by [3] where the images are manipulated by copying one part from a particular location of the image and pasting it in some other location. This is mainly done to hide some important information in the image that is undesirable to the person who is editing the image. This may also contain some duplicate portion of the image. Since the image contains a part of the same real image, the background and texture naturally blend in the image and hence it becomes difficult to find if it is real or forged [4].

2.3 Image Slicing

This is similar to the copy and move technique. Paper [3] proposed to again cut and paste, in this case, not from the same image, but from many different images. If splicing is carried out with high level of accuracy, then it becomes difficult to identify the forgery with our visual perception.

3. Non-AI/GAN synthesized fake image detection techniques

3.1 Authentication of original images to detect forgery

Another approach to detect fake images is to require each image to be distributed with authentication information, in such a manner that edited images will not satisfy the authenticity verification requirements. Paper [4] described one such authentication mechanism for JPEG images, by applying Genetic Algorithms. Following is the sequence of steps. At the first step, the original image is divided into 8x8blocks. Each block is marked Block(i,j). Authentication information for each block is generated by Cyclic Redundancy Check. Each block value is modified using Genetic Algorithm and is compressed in such a

way that authentication information is not lost. Now all modified blocks with authenticated information are combined and formed as image B. Finally, this image B undergoes JPEG compression with the Quantization Factor, and the image C is obtained with hidden water mark on it, containing authentication information. So, if any further modification is done on the image, the watermark will get tampered and therefore modification can be detected on a forged or manipulated image.

3.2 Copy and move forgery detection

Paper [5] proposed radix sort method to detect the locations where the image has been forged. First the image is split into b blocks, each of dimension 16x16. Each block is arranged as a 9-dimensional vector starting from v1, v2…. v9. Each block is again divided into 4 blocks s1, s2,s3, and s4. The average intensity of each block is found. This will indicate any modification in the image, if any. The radix sort is performed on each matrix formed as given below:

RADIX-SORT(A,d):

for j ← 1 to d sort array A based on j.

The values range from 0 to 255. Then the shift vector is found which indicates the position of the vector which has been relocated in the image. The shift vector identifies the position of modification with reduced complexity. Hence this method is much more optimal compared to watermark technique. The original and copy pasted image are compared by sorting and the degree of relocation is also detected with higher accuracy.

3.3 Detection based on Image Quality Metrics

This technique [6] is mainly used in place of the existing techniques for blind fake detection. The algorithm initially extracts Image Quality Metrics. The image is first divided into four regions. All the features are extracted from each region. The following step involves applying wavelet transform to every feature and the corresponding sub-bands are being calculated. The characteristic function for each sub-band is calculated by applying histogram. The moments are calculated by applying Discrete Fourier Transform. Similarly, 2-D histograms are applied for 2-D sub-bands, and 2-D characteristic function is obtained. All the functions are calculated and finally prediction error 2-D array is found. Then marginal moments are calculated again. The 2-D array is represented as 2x2, 4x4 and 8x8 and blocks. Discrete cosine transforms are applied. These values are rounded to the nearest integer

values. The same steps are repeated for all the features. The best parameters for training SVM model are created and the model is trained and used for test. While 85% is reported for the data part of their dataset, for new unseen data the accuracy is 65%.

3.4 Singular Value Decomposition

Singular Value Decomposition (SVD) is highly efficient compared to the previous techniques. Paper [7] proposed an especially important method that uses SVD to detect the forgery of fake images. This technique is mainly used for 2D image detection. Here a protected image is created from the original image and released to the public. If any changes are done in the protected image, then the SVD is implemented to detect the level of forgery. This is detected by finding the distance of the eigen vector from the orthogonal spaces. SVD is based on a linear algebra theorem where a rectangular matrix is a product of three matrices, an orthogonal matrix U, a diagonal matrix S, and the transpose of an orthogonal matrix V. It is represented as $A= USV^T$. Two secret column vectors v1 and v2 are calculated such that v1.v2 = 0 and protected image is created such that $A^1= US^1V^T$. Where, A1 and S1 are pre-processed values from A and S respectively. Then the pre-processed image is released to the public. The threshold is not auto set, and it keeps changing for every image. For forgery detection, a fake factor is computed based on secret column vectors v1, and v2. For any forged image A^, if the detected fake factor P is greater than threshold, then that image is concluded to be fake.

The authors have tried performing on four different scenarios of fake image.

- Face
- Rotation
- Date
- Question paper

In all these scenarios, the algorithm was able to predict the fake factor with 80% accuracy.

3.5 Error Level Analysis

This technique is used in image forensics. It is mainly used to check the level of image modification that has been done. This method works by analysing the pattern of the error in the images and compares between the original and modified image. Every image is split into a size of 8x8 pixel sized block. Now each block is checked in both original and modified image. If the block is not present, then the position is said to be modified. If the error pattern on all blocks is same and there is local minima, then the image is not modified.

Higher the chance of error pattern, then more the image is fake. The analysis can also be performed by saving the image to a different compression quality level, and then observing the changes between the levels. This analysis is mainly carried out on JPEG images and one such analysis is provided in [8].

3.6 Statistical Fourier Analysis

Image splicing can be detected by implementing Fourier statistics, as the technique of image slicing disturbs the series spectrum of the original image. Hence statistical analysis may be implemented to detect such image forgery. Fast Fourier Transform based methods as described in [9] can also be used to detect such image forgery.

4. Image synthesis techniques using AI and GAN

4.1 Using Machine Learning for Image Synthesis

In [10], a development of 3D face model with elevated level of photo realism was illustrated. The method built a machine learning model to create faces with different features like expressions, hair, clothing, and other accessories. The skin was also done using meso-displacement to achieve realism in skin texture. The domain of face analysis requires large dataset with rich features and classifications. Real images with rich features in large numbers and labelled, are difficult to be obtained. The paper provided a method for synthesizing such images. The advantages include the ability to control the variety and diversity of the dataset and provide noiseless annotation and accurate labelling. The synthesis was done by a procedure to integrate parametric face model with artist created assets including hair and textures. The position of eyes, its movement tracking, and dense landmarks were also found. These were implemented in neural networks with Pytorch packages and Image to Image translation was done for face parsing. With a C classed colour image x as input, the output was a prediction of C-channeled label image. The training was provided only with synthetic data which minimizes binary cross entropy loss.

4.2 Generative Adversarial Network Variants

Generative Adversarial Network (GAN) is a new framework of generative model which was proposed in 2014. Images can be synthesized in better ways using GAN. Therefore, it has become one of the popular research areas. There are two main fields of

research on GAN. One is the application of GAN in Computer Vision and another is for Natural Language Processing.

A General Adversarial Network is composed of two Neural Networks a Generator and a Discriminator, and the two networks compete. Generator tries to synthesize realistic images such that they do not get detected by the Discriminator, whereas the Discriminator's job is to distinguish the real samples from the synthesized ones.

Paper [11] proposed Control GAN specially to control the random distribution of samples which are produced. Data augmentation methods were used in Control GAN. Though the techniques used hinder learning the structure of GAN, they were useful for classification purposes. To enhance the quality of produced pictures, [6] proposed a novel model known as decoder-encoder GAN which works based on a) adversarial training and b) variational Bayesian inference. When compared with multistage GAN implementation and other hierarchical approaches, the suggested method can be executed easily and produced good results. A common problem with traditional GAN architecture is the mode collapse problem. The generator generates more realistic image with the target of fooling the discriminator. But the problem is that during the training, the generated samples become limited to only few modes instead of all modes of the dataset. To overcome this, [12] proposed that, instead of using traditional single generator, multiple ones could be used. This method not only increases the diversity of generated samples solving the problem, but it also does it in an operationally efficient manner.

A modified form of Deep Convolutional GAN (DCGAN) [13] was used in [14]. Here the generated samples and noise vectors were fed into certain convolution layers and the results were adjoined to form a single tensor that further passed through the additional transport layers. This fully connected network is called Structure-GAN, which transforms an image into typical map. The structure GAN's biggest setback is its necessary requirement of additional features in order to get the ground truth for surface mapping. To enhance the further generation skills, Complete Representation-GAN [15] was proposed. This CR-GAN incorporated generator units in addition to the single reconstruction path for continual learning. These two-learning structure worked in a sequential way to share the parameters for improvising the generation skills. [16] proposed a Star-GAN which is a Scalable model and specifically designed to perform image to image translations for many domains. Star- GAN consists of well-defined integrated model design which help to achieve training multiple datasets in a parallel under various domains of a single network. This model Is known for

temporal alignment, transcription, and parallel operations. But this is comparatively less efficient.

To generate a realistic picture along with foreground and other background details, [17] proposed a model with fusion approach known as SF-GAN. The SF-GAN has a geometry synthesizer and an appearance synthesizer. The geometry synthesizer starts to sketch the local geometry of the background images by using it, where it could transform the foreground objects into background image. The appearance synthesizer takes the responsibility to tune the colours, brightness of the foreground object that matches to the background image without conflicts. GAN also has guidance filter to retract the detail loss. The geometry synthesizer and the appearance synthesizer work in an interconnected way and with a little supervision, it achieves concurrent geometry and appearance synthesis realism. Synthesis of face images with detailed textures and high perceptual quality is not trivial. For this purpose, [18] proposed Conditional Adversarial Autoencoder (CAAE), which is now mostly used to transform the facial image from one age to another. The CAAE consists of an Encoder E, a Generator G, and a pair of Discriminators (Dz and Dimg). The CAAE is usually trained in the following way: The Encoder E maps the sample input face image X to vector Z which indicates the personal features of the image X. The output vector Z is associated with the conditional vector C indicating age which is fed as an input to the Generator G to generate new sample face image. On the other hand, the vector Z is fed as an input to the discriminator Dz which makes Z to be uniformly distributed. Discriminator Dimg forces the Generator G to synthesize the face to be realistic and verifies if it is suitable to the input age. Apart from that, there are two adversarial losses namely L2 content loss and Total Variation (TV) loss of the two discriminators mentioned above. The content loss forces the generator to synthesize the similar face, and on other hand TV loss helps to remove ghosting effects.

Similarly, article [19] proposed a method for Face aging which is known as Age-conditional Generative Adversarial Network (Age-cGAN). It is composed of encoder E and cGAN. Similar to the CAAE, it uses the encoder E to map a sample face image x to the latent vector Z and the conditional generator G tries to map with latent vector Z along with the conditional vector C to synthesize new face image. On the other hand, cGAN and the encoder were first trained with the help of pair of latent vectors and generated face images. After the training phase, the input face X0 with input age C0 was fed as an input to the model. At first x0 was given as input to the encoder which gave the latent vector Z0 as output. In the next

step, by using the identity preserving optimization, Z0 was iteratively updated to the latent vector Z. Using this Z, ensures synthesis of face image which is close to the sample input face image. In the next step, the latent vector Z and the target age were given as input to the Generator where the respective new synthesised face image was obtained as an output.

To create a different perspective of views of facial images, [20] proposed Disentangled Representation learning- GAN (DRGAN). The generator of this model consists of an encoder-decoder architecture, which is used to learn the disentangled representation of the facial images. The encoder starts working by mapping a face image X to the identity future vector F. The output of the encoder is the input of the decoder. The decoder consists of two parts. One is Identity Classification which contains the addition identity class for the fake images and the other is for pose classification. The aim of the Discriminator is to realize and identify the identity and pose correctly from the sample input facial image and verify the identity whether it is fake or real. The aim of the generator is to outsmart the discriminator. DRGAN is one of the superior poses invariant face recognitions models.

Paper [21] proposed a model known as Two-Pathway Generative Adversarial Network. It is especially designed for frontal face image synthesis. The network consists of two-pathway architecture. It has two Generators. One is a global generator for generating global structures and the other is local generator which is used to extract the details around the facial landmarks. Its objective function for training also contains different losses: L1 pixel-wise content loss which helps to measure the difference between the synthesized face image and sample input face image, and symmetry loss which helps by forcing the synthesized facial image to be horizontally symmetric. In addition, few of the earlier described losses like adversarial loss, identity preserving loss and TV loss are also in scope. Weight sharing convolution layers were also used to avoid expensive computation during synthesis.

Article [22] proposed an image-to-image translation framework known as Pix2Pix using cGAN. Its network follows the DCGANS architecture guidelines and implements some additional features. The modules were applied in the form of convolution-BatchNormRelus. It skips the connection between deep layer and shallow layer for the generator. Its discriminator uses the Patch GAN architecture which makes the network to run faster and penalize the fake structure at patch scale. To synthesize the image which not only is realistic but is also close to the ground truth, the above-mentioned structure contains not only adversarial but also content loss (that is measure of distance between the synthesized image

and ground truth image) in its objective function. The Pix2Pix has a disadvantage that it requires a paired image (that is image before translation and the respective image after translation). Hence training this model is a tedious task because most of such image pairs do not exist.

To overcome this problem, [23] proposed cycle-consistent Generative Adversarial Networks (cycle GAN). It is an unpaired image to image translation framework. It was designed with two GAN network. One is to map the image from one domain to another $X_{trans}=G(x)$, whereas the other Generator performs inverse translation $x=G_{inv}(x_{trans})$. Its networks architecture follows feed forward network [24], which is based on perceptual losses which makes it effective on style transfer. Similar to the Pix2Pix, its discriminator also uses the patch GAN. This GAN was also combinedly trained following the regulation of LSGAN. Here the adversarial losses instead of using traditional log function were implemented using square function to achieve more stable training, apart from the two adversarial losses of two GANs. The objective function also consists of L1 cycle consistency loss, which forces to perform dual translation (that is image translates itself after translation). By using this method, several translations can be achieved: for example, for collection transfer, season transfer, etc.

The Adversarial Inverse Graphics Network (AIGN) [25] is another image-to-image translation unpaired framework. This model consists of three main architecture components namely a Generator G, a discriminator D, and a task specific render P. Here the Generator translates input image X to output image G(x), whereas the Renderer P maps the output of the Generator to its respective input. Training this model objective function consists of two types of loss: one is adversarial loss and other is re-constructional loss. AIGN not only is used for image translation but also for 3D human pose estimation, image inpainting, face super resolution, etc. The above-mentioned methods of image translation are based on supervised learning.

Paper [26] proposed a Distance Gan model which is especially designed to perform unsupervised image translation. The working of DistGAN is mostly similar to the cycle GAN. DistGAN uses the self-distance constraint to support stochastic gradient instead of self-consistency loss. Distance GAN can generate image by calculating the distance in RGB space that is more realistic than traditional GAN methods. The main setback in Distance GAN is expecting to have higher correlation between two images of two separate domains which would have equivariance after the process of translation. One could perform the image

translation from one domain to another targeted in order to achieve image translation on multiple domains. But the problem is that, for each domain there is unique G and as the number of domains increases then concurrently number of times calculating G also increases and the same requires more computational power.

To overcome this problem, [16] proposed Star GAN model. By using this, the unsupervised multi-domain image to image translation can be achieved. The Star GAN has separate auxiliary domain classifier to overcome multidomain translation problem. The auxiliary classifier is used to classify the input images based on their domains. The Star GAN is capable to take two inputs. One as conditional input as real image, and the other is the label for the target Domain. The Star GAN is designed in a way to use one hot vector which is a combination of all label vectors. This makes Star GAN possible to train on multiple datasets with different class labels on it. This model uses the cycle-consistency loss which enforces to synthesize the image that is similar to the sample input image.

5. GAN/AI Synthesized Image Detection

5.1 Auto GAN

Paper [27] presented a Deepfakes classification method based on spectra of the frequency domain as input. A GAN simulation framework, called Auto GAN was proposed. The working of Auto GAN involves passing the given real image through a generator which produces an output. The working of Auto GAN is like image generating GAN. The decoder includes up-sampling techniques such as transposed convolution and interpolation. In this GAN architecture, only the knowledge about model weights and parameters were assumed. The same also applies for other image detection GAN models. This is similar to a black box solution, where there was zero knowledge about the details that were being used to detect the attack model. The white box is said to have full details. The Auto GAN uses a discriminator and l1 norm loss. Instead of matching the distribution of image output to the image of another semantic category, the output result produced from the generator was matched with the original image itself. The loss function is given by:

$$L=n\sum i=1\log(D(Ii)) +\log(1-D(G(Ii))) +\lambda\|Ii-G(Ii)\|1 \quad (1)$$

Here, D(·) is the discriminator, G(·) is the generator, G(Ii) is the output of G(·) when taking I as input and λ is the trade-off parameter between two different losses. The first two terms are similar to the GAN loss. The discriminator wants to distinguish between the

generator output and the real image, while the generator wants to fool the discriminator. The third term is the 1-norm loss function to make the input and output similar in terms of 1-distance.

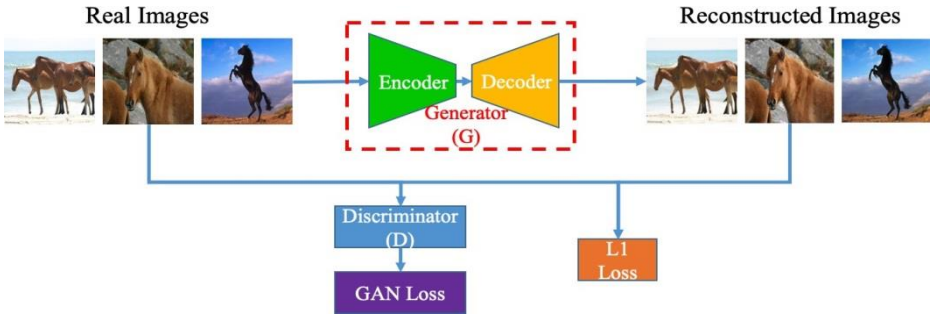

Figure 1. Pipeline of Auto GAN

The real image and its augmented constructed image are being compared as shown in Fig. 1. The frequency spectrum of both are analyzed. Although both images look similar, the frequency domain of both images are different, and the reconstructed image has all GAN included features. The classifier focuses on the artifact and thus can generalize to other GAN fake images that have similar artefacts. Auto GAN has some major benefits like, it does not need any fake image as input during training. The required artifacts can be given along with input real image to create the fake image. Next advantage is that Auto GAN can take input image form any semantic category. This contrasts with image2image, where only the same semantic category can be given as inputs (e.g., zebra and horse).

5.2 CNN with Traditional GAN

Paper [28] studied the understanding of human behavior and the working of CNN models and how well they are trained in discriminating fake/real faces. This was done by training with GAN generated images. A series of studies and experiments were conducted to diagnose the discriminator working in CNN and perform low-level statistics analysis for verification. Although CNN based training models are much better performing compared to other techniques in fake face detection, it is still not robust enough, as it cannot handle real world scenarios. In these scenarios, images may be modified from different unknown sources which are unidentified. Several studies were conducted between humans and CNN in discriminating fake/real faces, and it was concluded that fake faces have texture different from the real ones. Low-level texture statistical analysis also showed that large texture information is more robust to image editing and are invariant among different GANs. Then grey level co-occurrence matrix was employed to check for the homogeneity, contrast, and

other features among images. This helped in determining the difference between texture of real and fake images.

5.3 CNN with VGG-16

GAN synthesized faces can be efficiently distinguished by Deep Learning based detection methods [29]. One of the most common methods is to train the Deep Neural Network by extracting the signal level features. One of the earliest techniques by [30], is using VGG-Net [31] for GAN-face detection. It used CelebA face dataset [32] to train the network. Fake faces generated using DC-GANs [13] and PG-GANs were also fed as input for training. Work of [32], showed that VGG-16 architecture used with pre-trained weights of VGG-Face [33], can efficiently distinguish real and GAN synthesized faces. The input face images were first processed with the high pass filters and then the resulting residuals were fed into DNN. Next, a feature set was extracted to capture colors image statistics followed by using the concatenated features to train the classifier.

5.4 CFFN (Common Fake Feature Network) with pairwise learning

In [34], a Deep Learning Framework combined with pairwise learning was proposed. Deepfake image detection with pairwise learning is a method for detecting manipulated images using deep learning techniques. It is based on the concept of pairwise learning, which involves training a model to predict the similarity between pairs of images. The model is trained on a dataset of pairs of original and manipulated images, and learns to differentiate between the two based on their similarities and differences. The pairwise learning approach allows the model to learn from a large number of positive and negative image pairs, which can be used to improve the model's performance. The goal is to learn a feature representation of each image that captures the underlying content, which can then be used to detect manipulated images. The deepfake image detection method can be used to detect manipulated images in various applications such as social media, news, and entertainment. The model can be trained on a large dataset of images, and can be fine-tuned to improve its performance on specific types of manipulated images.

5.5 Deepfake detection by RGB analysis

Paper [35] compares the colour spaces in the fake generated images. Postprocessing of images after attack creates an abnormal trace in the space of RGB and hence appears to be very unreal. The statistical analysis of chrominance information in other colour space will be

much more robust and can be easily used to distinguished between real and fake images. 2 steps were carried out in the experiment: at first, the features were extracted, and then Random Forest Algorithm [33] was used for classifying real and fake images.

5.6 Detection by Frequency Spectrum Analysis

In [27], two different models were considered for demonstrating the detection. First one is the ridge regression and next is a CNN model. Both were trained on the DCT frequency spectrum from natural and generated images. The CNN model differentiated five different classes, where four of them were GAN models and the fifth class contained natural images. The regression is totally different. It acts as a binary classifier. The classifier was trained for each GAN model. Out of these two, it was concluded that the regression model is better. It was shown that the patterns in the frequency spectrum were captured by the weights of the regression model. This suppresses the frequency pattern of all the attacks that occur and helps to give fingerprint based counter attacks.

5.7 Deepfake Detection model with Mouth Features

There are some other techniques which extract certain facial features or texture and the same could be used in some deep learning mode. Deep Fake detection model with mouth features was designed based on this technique. Its approach focuses on the open mouths with teeth and is fed into the standard CNN model to detect Deepfake by isolating, analyzing, and verifying lip and mouth movements. Later, [36] employed a pretrained well-designed Xception [37] network (which is a primitive convolution neural network that is trained on ImageNet by separable convolutions with residual connections). It was found to have better results over the existing other approaches.

5.8 Deep Fake Detection using DCL

Paper [38] studied the positive and negative paired data and used it for Deep Fake Detection. This method is known as Dual Contrastive Learning (DCL). By using this framework, the simultaneous contrast features can be extracted between different instances (Inter-Instances) and within the instances (Intra-Instances). At first DCL was used to train the network with the help of contrastive learning framework in a supervised way. The DCL consists of few important modules. One is Data Views Generation, which is responsible to generate different views of inputs by using the special designed data augmentation. This is followed by feature extraction from well-designed supervised contrastive learning

architecture. Next Inter-Instance Contrastive Learning module was used to arrange the feature distribution and Intra-Instance Contrastive Leaning module was used to enhance the inconsistency of forged faces.

5.9 Deepfake Detection with SRM model

The SRM [39] model uses a technique called Self-Regularization, which involves adding a term to the loss function that encourages the model to learn a compact and robust feature representation. The model was trained on a dataset of manipulated and non-manipulated images, and learns to differentiate between the two based on their feature representations. The SRM model can be used to detect manipulated images in various applications such as social media, news, and entertainment. The model can be trained on a large dataset of images, and can be fine-tuned to improve its performance on specific types of manipulated images.

5.10 Deepfake Detection with MAT model

The MAT model [40] was trained with a combination of original and manipulated images, as well as with different types of manipulation methods, like splicing, copy-moving and deep learning-based manipulation methods. This allows the model to learn to detect a wide range of manipulation methods. The MAT model can be used to detect manipulated images in various applications such as social media, news, and entertainment. The model can be trained on a large dataset of images and can be fine-tuned to improve its performance on specific types of manipulated images.

6. Results and Discussion

The creation of fake artifacts has been from the early ages which always creates a delusion about the authenticity of the work; a similar problem is fake image synthesis which has been creating a major impact in recent days. At the early stage of fake image creation, most of them used the image manipulation tools like adobe, GIMP tools, and copy move forgery techniques. The works created by the above-mentioned techniques can be easily distinguished by analyzing few image quality metrics and error patterns in the image's pixel and series spectrum of color distribution and also by using watermark techniques. But in recent years, vast interest in the GAN has led to a great evolution in the field of fake image synthesis. Gradual improvements in architectural design to attain different objectives and goals such as, to increase the diversity of generated samples replacement of single generator

units with multiple ones, and introducing deep convolutional layers along with multiple generator units known as CRGAN architecture. To attain the realistic geometry of the real image, SF -GAN has been used. To replicate similar facial features DR GAN architecture has been used. To perform image-to-image translations on multiple domains, the Star GAN architecture has been used. These different architectures create a more realistic fake images than the traditional fake images (Non-GAN/AI synthesized). Though a few detection techniques which perform well are listed above, the fake ones are synthesized by different architectures and each architecture is based on different strategies. The same underlying research problems of distinguishing between a human user and a machine (bot) which is a key problem in fake news, fake chat (chatGPT), and deepfake video domains are faced, which is a great setback on the authenticated issue on various social media networks. Therefore, solving this generic problem would be a significant initial step towards solving the above-mentioned varieties of problems, which would be of greater benefits to the people in this digital era.

7. Conclusion

The risk posed by deep fake image is becoming high. One reason is because of the social and psychological impact that it can create to those who are personally affected by it. The other factor is that the probability of occurrence of such attacks is on a continual increase. With both these factors increasing, the net risk that it poses is significantly high that this problem is paid adequate attention by the community. Anybody can get access to anyone's social pictures and videos. This is aggravated by the fact that open-source tools are available for usage and they can use it without much restrictive controls. Social media also enables distribution of the created fake content without adequate checks and controls, though some work has just started to happen in that direction. Deep learning methods such as GAN is gaining a lot of popularity. Though some deep learning-based methods have been proposed to address this issue, this will be largely inadequate to convincingly solve the problem. This is because deep learning just like GAN, raises the platform for both generators and discriminators in a similar way. Hence, as technology increases the dimension of fight between the police and the thief, but never solves the basic problem, deep learning also appears to be inadequate to solve the problem of indistinguishability between digital true and fake images. Hence, a systematic study of existing methods of fake image detection and fake image synthesis with an interesting perspective is needed to get the clarity that, this paper brings out as a conclusion that it needs more than a deep learning mechanism to

fundamentally distinguish a fake image from a real image. While this paper makes the problem definition and potentially eliminates deep learning to be an absolute solution to the problem, it calls for more innovative ways of effective solutions to this problem, which will be the scope of future work.

References

[1] A. Kumar, L. Bi, J. Kim, and D. D. Feng, "Machine learning in medical imaging," Biomedical Information Technology, pp. 167–196, 2020, doi: 10.1016/B978-0-12-816034-3.00005-5.

[2] R. Nasina and S. Kommareddy, "FACE IMAGE METAMORPHOSIS OR MORPHING."

[3] R. Shankar, A. Srivastava, G. Gupta, R. Jadhav, and U. Thorate, "Fake Image Detection Using Machine Learning," 2020. [Online]. Available: www.ijcrt.org

[4] V. Gopal Edupuganti and F. Y. Shih, "Authentication of JPEG Images Based on Genetic Algorithms," 2010.

[5] H.-J. Lin, C.-W. Wang, and Y.-T. Kao, "Fast Copy-Move Forgery Detection." [Online]. Available: http://int.chihlee.edu.tw

[6] S. Math and R. C. Tripathi, "IMAGE QUALITY FEATURE BASED DETECTION ALGORITHM FOR FORGERY IN IMAGES," 2011.

[7] A. Al-Rammahi, N. el Abbadi, A. M. Hassan, and M. Al-Nwany, "Blind Fake Image detection," 2013. [Online]. Available: www.IJCSI.org

[8] N. A. N. Azhan, R. A. Ikuesan, S. A. Razak, and V. R. Kebande, "Error Level Analysis Technique for Identifying JPEG Block Unique Signature for Digital Forensic Analysis," Electronics (Switzerland), vol. 11, no. 9, 2022, doi: 10.3390/electronics11091468.

[9] N. Kanwal, J. Singh Bhullar, A. Girdhar, and L. Kaur, "Detection of Digital Image Forgery using Fast Fourier Transform and Local Features."

[10] Asif Razzaq, "Microsoft AI Research Introduces A Huge Synthetic-Face Dataset Along With A Face Analysis Method Using Synthetic Data Alone," Marktechpost LLC, Oct. 04, 2021.

[11] M. Lee and J. Seok, "Controllable generative adversarial network," IEEE Access, vol. 7, pp. 28158–28169, 2019, doi: 10.1109/ACCESS.2019.2899108.

[12] G. Zhong, W. Gao, Y. Liu, Y. Yang, D. H. Wang, and K. Huang, "Generative adversarial networks with decoder–encoder output noises," Neural Networks, vol. 127, pp. 19–28, Jul. 2020, doi: 10.1016/j.neunet.2020.04.005.

[13] A. Radford, L. Metz, and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," Nov. 2015, [Online]. Available: http://arxiv.org/abs/1511.06434

[14] P. Shamsolmoali et al., "Image Synthesis with Adversarial Networks: a Comprehensive Survey and Case Studies," Dec. 2020, [Online]. Available: http://arxiv.org/abs/2012.13736

[15] Y. Tian, X. Peng, L. Zhao, S. Zhang, and D. N. Metaxas, "CR-GAN: Learning Complete Representations for Multi-view Generation," Jun. 2018, [Online]. Available: http://arxiv.org/abs/1806.11191

[16] 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," Nov. 2017, [Online]. Available: http://arxiv.org/abs/1711.09020

[17] F. Zhan, H. Zhu, and S. Lu, "Spatial Fusion GAN for Image Synthesis."

[18] Z. Zhang, Y. Song, and H. Qi, "Age Progression/Regression by Conditional Adversarial Autoencoder." [Online]. Available: https://zzutk.github.io/Face-Aging-CAAE

[19] G. Antipov, M. Baccouche, and J.-L. Dugelay, "Face Aging With Conditional Generative Adversarial Networks," Feb. 2017, [Online]. Available: http://arxiv.org/abs/1702.01983

[20] L. Tran, X. Yin, and X. Liu, "Disentangled Representation Learning GAN for Pose-Invariant Face Recognition."

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in 32nd International Conference on Machine Learning, ICML 2015, 2015, vol. 1, pp. 448–456.

[22] P. Isola, J.-Y. Zhu, T. Zhou, A. A. Efros, and B. A. Research, "Image-to-Image Translation with Conditional Adversarial Networks." [Online]. Available: https://github.com/phillipi/pix2pix.

[23] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, and B. A. Research, "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks Monet Photos." [Online]. Available: https://github.com/junyanz/CycleGAN.

[24] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual Losses for Real-Time Style Transfer and Super-Resolution."

[25] H.-Y. Fish Tung, A. W. Harley, W. Seto, and K. Fragkiadaki, "Adversarial Inverse Graphics Networks: Learning 2D-to-3D Lifting and Image-to-Image Translation from Unpaired Supervision."

[26] S. Benaim and L. Wolf, “One-Sided Unsupervised Domain Mapping.” [Online]. Available: https://github.com/sagiebenaim/DistanceGAN.

[27] X. Zhang, S. Karaman, and S.-F. Chang, “Detecting and Simulating Artifacts in GAN Fake Images,” Jul. 2019, [Online]. Available: http://arxiv.org/abs/1907.06515

[28] Z. Liu, X. Qi, and P. H. S. Torr, “Global Texture Enhancement for Fake Face Detection In the Wild.”

[29] X. Wang, H. Guo, S. Hu, M.-C. Chang, and S. Lyu, “GAN-generated Faces Detection: A Survey and New Perspectives (2022),” Feb. 2022, [Online]. Available: http://arxiv.org/abs/2202.07145

[30] H. Chi, I. Seop Na, S. Kim, N.-T. Do, I.-S. Na, and S.-H. Kim, “Forensics Face Detection From GANs Using Convolutional Neural Network,” 2018. [Online]. Available: https://www.researchgate.net/publication/327905310

[31] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” Sep. 2014, [Online]. Available: http://arxiv.org/abs/1409.1556

[32] X. Liu and X. Chen, “A Survey of GAN-Generated Fake Faces Detection Method Based on Deep Learning,” Journal of Information Hiding and Privacy Protection, vol. 2, no. 2, pp. 87–94, 2020, doi: 10.32604/jihpp.2020.09839.

[33] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, “VGGFace2: A dataset for recognising faces across pose and age,” in Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018, Jun. 2018, pp. 67–74. doi: 10.1109/FG.2018.00020.

[34] C. C. Hsu, Y. X. Zhuang, and C. Y. Lee, “Deep fake image detection based on pairwise learning,” Applied Sciences (Switzerland), vol. 10, no. 1, Jan. 2020, doi: 10.3390/app10010370.

[35] V. Wesselkamp, K. Rieck, D. Arp, and E. Quiring, “Misleading Deep-Fake Detection with GAN Fingerprints,” May 2022, [Online]. Available: http://arxiv.org/abs/2205.12543

[36] M. T. Jafar, M. Ababneh, M. Al-Zoube, and A. Elhassan, “Digital Forensics and Analysis of Deepfake Videos,” in 2020 11th International Conference on Information and Communication Systems, ICICS 2020, Apr. 2020, pp. 53–58. doi: 10.1109/ICICS49469.2020.239493.

[37] T. Wang, H. Cheng, K. P. Chow, and L. Nie, “Deep Convolutional Pooling Transformer for Deepfake Detection,” Sep. 2022, [Online]. Available: http://arxiv.org/abs/2209.05299

[38] K. Sun, T. Yao, S. Chen, S. Ding, J. Li, and R. Ji, "Dual Contrastive Learning for General Face Forgery Detection." [Online]. Available: www.aaai.org

[39] Y. Luo, Y. Zhang, J. Yan, and W. Liu, "Generalizing Face Forgery Detection with High-frequency Features," Mar. 2021, [Online]. Available: http://arxiv.org/abs/2103.12376

[40] H. Zhao, T. Wei, W. Zhou, W. Zhang, D. Chen, and N. Yu, "Multi-attentional Deepfake Detection," in 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2185–2194. doi: 10.1109/CVPR46437.2021.00222.