CHAPTER 6

A review of techniques to detect the GAN-generated fake images

Tanvi Arora[a] and Rituraj Soni[b]
[a]Department of CSE, CGC College of Engineering, Landran, Mohali, Punjab, India
[b]Department of CSE, Engineering College Bikaner, Bikaner, Rajasthan, India

6.1 Introduction

The generative adversarial network (GAN) is an artificial intelligence-based technique that is based on the deep learning modalities of the machine learning paradigm. It is an unsupervised learning technique. The GANs have been initially created in 2014 to generate new data points from the existing data points. In these, two competing neural networks are made to work against each other to improve their quality. The working principle of the GANs can be best described by taking the example of a generator that is generating some output, and a tester that is testing the generated output, for its authenticity. The tester knows what is correct, thus based on the feedback of the tester, the generator keeps on improving its output. The generator is just like a blind man, which improves its results based on the rejection and selection of its output.

The GANs are used for generative modeling, that is, a model is used to create new instances from the preexisting instances, such as the creation of new images that are quite identical but still different from the already existing images. The GAN-based models work like a gameplay, where both the players try to trick each other and ultimately solve the puzzle. If the GAN-based methods are properly trained then they can be effectively used to create new data items as per the specifications of the items in the training set.

The GAN is composed of two contending neural networks that work against each other in competing mode to investigate, capture, and duplicate the disparities in the dataset. The GANs are composed of three distinct components:

- *Generative*: It is aimed at learning the generative model that can demonstrate the generation of the data concerning the probabilistic model.
- *Adversarial*: The adversarial setting-based training of the model is carried out by this unit.
- *Networks*: The training of this model is carried out using the deep learning-based artificial intelligence methods.

The working of the GANs is shown in Fig. 6.1, which has two components, the generator and the discriminator. The task of the generator component is to create the

Generative Adversarial Networks for Image-to-Image Translation
https://doi.org/10.1016/B978-0-12-823519-5.00004-X

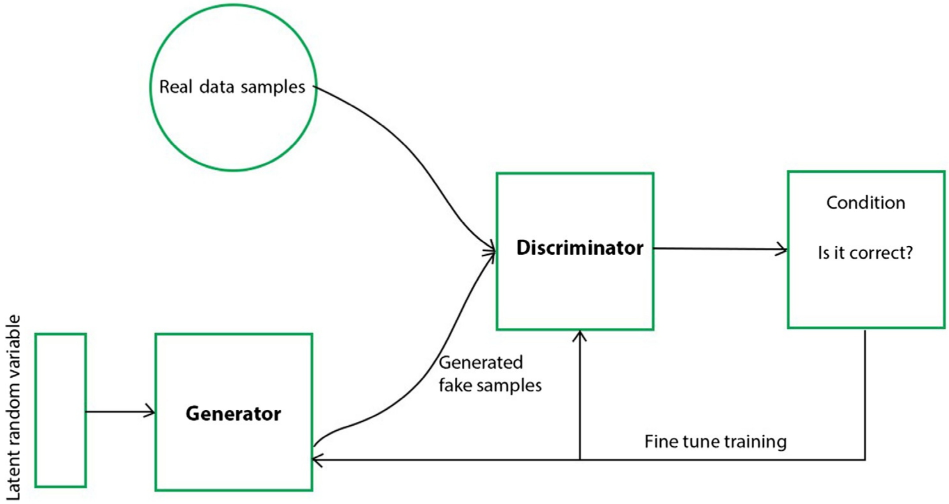

Fig. 6.1 Simple architecture of GANs (https://www.geeksforgeeks.org/generative-adversarial-network-gan/).

self-made illustrations of the data, which may be images, audios, or videos, and the discriminator component tries to discriminate the input given to it as the real data or the self-made illustrations. In the GANs, the generator and the discriminator both are the neural networks, and both of them compete against each other in the training phase. The training of the generator and discriminator is carried over several iterations, and with each subsequent iteration, the capabilities of both the components are enhanced. The generator learns to generate better illustrative samples and the discriminator gets more proficient at judging the illustrations as fake samples. In short, the GANs are based on the minimax game, where the discriminator tries to minimize its gains and the generator tries to minimize the gains of the discriminator or to maximize its losses.

Although GANs have emerged just a few years ago, over a short span of time a large number of variants have emerged. The most commonly used variants of the GANs are

1. *Vanilla GAN*: In this, the basic multilayer perceptron-based generator and discriminator neural networks are used, and it is one of the simplest implementation of the GAN that tries to augment the mathematical functions based on the stochastic gradient descent approach.
2. *Conditional GAN (CGAN)*: This GAN implementation is based on the setting up of the condition-based parameter in the deep learning approach. The extra condition parameter is augmented in the generator component of the GAN to generate the output data. The discriminator is feed with the input data that has labels associated with it to assist it to distinguish between the actual data and the morphed fake data.

3. *Deep convolutional GAN (DCGAN)*: This implementation of the GAN uses the convolutional neural networks in place of the multilayer perceptrons, but these CNNs does not contain the max-pooling layer that has been substituted with the convolutional stride and the layers of the CNN are not completely connected. Over the years, this implantation of the GANs has become the most widely used as well as the most promising implantation of the GANs.
4. *Laplacian pyramid GAN (LAPGAN)*: This implementation of the GANs are mainly used for producing superior quality images. In this initially, the image is down-sampled, and then again it is upsampled until the image comes to its original size. This approach leads to the introduction of noise in the images. This implementation has a large number of generator and discriminator modules of the network along with the distinct levels of the Laplacian pyramid.
5. *Super resolution GAN (SRGAN)*: This implantation aims at producing high-resolution images, and they work at enhancing the low-resolution images, but simultaneously it takes care that the upscaling does not result in introducing noise or error in the images. This implementation is made up of the deep neural and the adversarial networks.

The GAN-based techniques can be used in a large number of applications such as:

- to augment the dataset of images by creating more synthetic images
- to create different facial expressions
- to create real-looking images
- to create the images of the cartoons
- to create images based on text description
- to create emojis from the images
- to create edited photographs
- to improve the resolution of the images
- to transform the clothing of images
- 3D object creation
- completing the incomplete images
- to synthesize the videos

These are just a few of the applications of GANs, but recently they have created a lot of excitement as they have emerged as one of the most fascinating applications of the current AI advancements, and still it is hoped that many more exciting applications will be available in near future.

The GANs have many advantages associated with them, the first and the foremost advantage is that they are unsupervised learning models that have self-learning capability, do not require the labeled data, and can self-learn from the data itself. Moreover, the GAN-based methods can also be used to produce the data that is as good as the real data. The GANs have the capability to not only generate the numeric or alphanumeric data, but also generate multimedia data, that is, images and videos, which are indistinguishable and are at par with the real data. Thus, GAN-generated images that have diverse applications in the field of marketing, gaming, mass media advertisements, and other domains. The GANs not just learn from the data itself, but they also can understand the complex data and they have a wide range of applications in machine learning.

The GANs have gained a lot of hype in the recent times, but they do have their limitations as well, and they are just the product of virtual imagination. They depend on the training data, if the training data is not correct or of good quality, then the GAN-based methods can even fail. They cannot be just used to create novel things, but they can only reformulate the things based on the previous examples. The real strength of the GANs depends on the coordination of the generator and the discriminator component. They both need to be fine-tuned, and the strength of the generator will be of no use if the discriminator is weak or vice versa. They both need to work in synchronization to produce the correct results. The content thus generated by the GANs is termed as DeepFakes.

6.2 DeepFake

A new term DeepFake has emerged in the digital world, which has been derived from the two terms namely deep learning and fake; it is a new product created by artificial intelligence. In layman's terms, DeepFakes can be defined as a false media, images, videos, or sounds, created by using deep learning techniques. Deep learning is a branch of artificial intelligence, which has a large number of layers that can make informed decision making by using a set of algorithms, and work the way the neurons of the human brain work. The intelligence of the deep learning-based algorithms has created the fear of creating something that does not exist, but which mimics the real-world existence of the things that they pose to depict.

Fig. 6.2 shows an example of the DeepFake image created by morphing the original image. Thus, we can say that DeepFakes are the morphed video clips, images, sounds, or other representations in the digital form that are created using the erudite algorithms based on artificial intelligence that creates fabricated media and gives the impression to be realistic. The DeepFakes have emerged just a couple of years ago, but this technology has made many refinements and is posing a great threat to public figures such as celebrities, political leaders, and public figures such as the technology leaders. The very first incidence of the DeepFake caught the attention of the media in 2017, wherein a video footage emerged showing the famous Bollywood figures were made viral, and the reality of that footage was not real, which was the amalgamation of the face of the celebrity and the other details of some other actor and was created using the DeepFake technology. The DeepFakes can be created a large number of images that are available on the internet. Therefore, the popular public figures are the one who can be targeted at large to generate DeepFakes, for whom a large number of images and videos are readily available on the internet, due to media coverage.

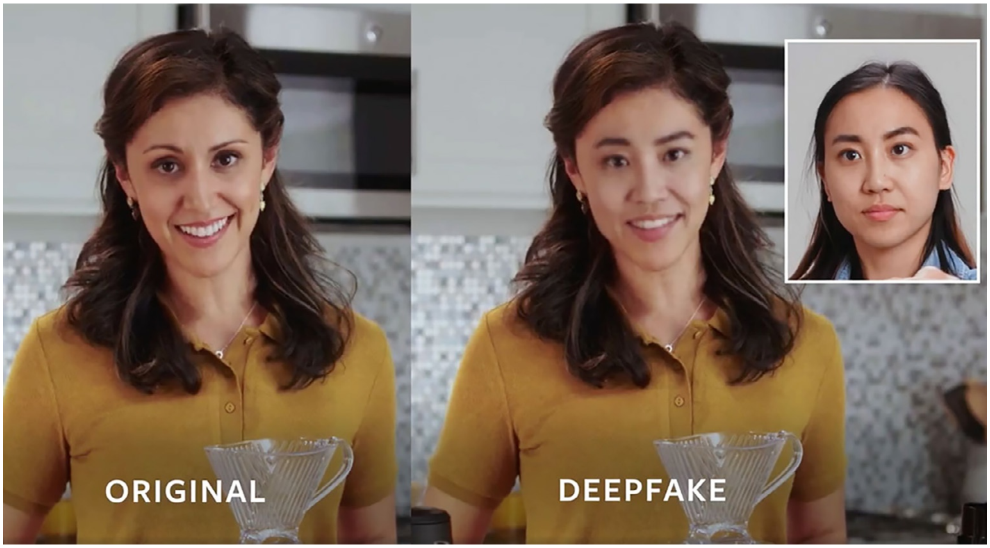

Fig. 6.2 Fake image generated using GAN (https://spectrum.ieee.org).

To reveal how the technology of deep learning can be abused, the scientists at the University of Washington created and posted the DeepFake-generated video of the President Barack Obama over the social media, the scientists proved that they can make the fake video of President Obama speak out whatever they wanted. We all can well imagine what harm this technology can pose to the image of the public figures and it can pose a great threat to the security of the world at large. Thus, fake news and DeepFake can club together and can tar the authentic information and can thus create misunderstandings and miscommunications that are supported by fake facts.

The DeepFakes emerged just a few years ago, but much development has taken place and is improving at a rapid rate. The scientists have developed methods that allow them to even edit the transcripts of the video and do alterations to the words that are being spoken by the person whose video is being altered. In yet another work, the researchers at Stanford University have developed methods that can not only manipulate the facial expressions, but their methods can also do three-dimensional head movement of the characters of the video or make them blink their eyes or let them gaze at particular stuff, all these things can be carried out with the help of the GANs. These features can help the movie industry to easily dub the movies to other languages, as all these things look to be unbelievably photorealistic. But the concern of the research fraternity lies with the counterpoints, what if these techniques are abused and are used for the illegitimate activities.

Initially, it was believed that DeepFakes can only be created for the images if a large number of similar images of the celebrity or the public figures for whom large amounts of images are available in the public domain. But the recent developments by the Samsung's

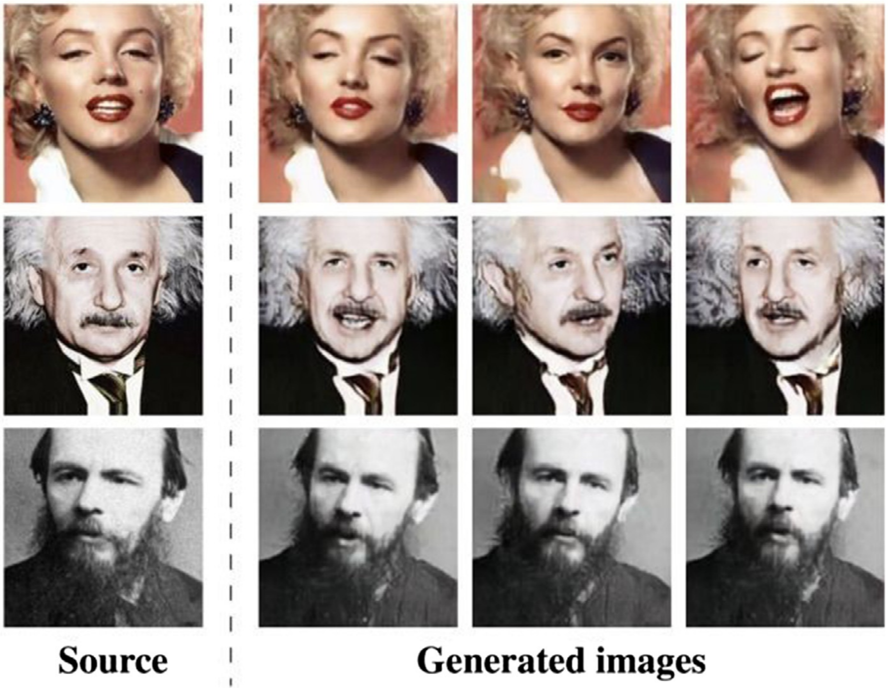

Fig. 6.3 Sample source and generated fake images (https://miro.medium.com).

AI lab has created a living portrait of Salvador Dali, Marilyn Monroe, and many more, moreover, they have created an image of Mona Lisa, who is smiling, and all these things have been achieved by just using a limited number of photographs, as illustrated with examples in Fig. 6.3. The requirement of a limited number of photographs has thus raised concern for the ordinary people, as they initially believed that they are invulnerable to DeepFakes as there are not enough images available to train the computer procedures to create DeepFakes.

After seeing the realistic view of the images created by artificial intelligence, now the big question is how to control and protect the misuse of this technology, which has a large number of advantages associated with it, but still it has few threats associated with its usage.

There are many questions that need to be pondered upon, should legal laws be passed, to bind the social networking websites, to detect the DeepFakes, and subsequently remove them. Moreover, should the intention of creating the DeepFake be given any consideration, while removing the DeepFake images, and apart from that can the DeepFakes can be differentiated based on the intention of their creation either for the entertaining or for perniciousness.

6.3 DeepFake challenges

The rapid development in the domain of artificial intelligence has posed a serious threat to the authenticity of the multimedia content that is being generated. The recent advancements in the field of deep learning that led to the generation of DeepFakes have further intensified the generation of the misinformation and it is being believed that over the years this problem of misinformation supported by fake multimedia content will further intensify. With the development of technology, the approaches that are being used to generate the automated fake multimedia content will further improve and it will become more challenging to discriminate the original and the fake content.

The DeepFake-based multimedia content poses a great deal of challenges in the following domains:

1. create distress during difficult situations
2. can be a threat to the reputation of famous personalities
3. it can spread hatred and disrespect for the innocent
4. can cause loss of faith in the digital content
5. it may turn out to be deceiver's dividend, who can always deny his words by saying, the content is fake, even though it has been said or done by him
6. fake pornography can be created and can cause mental distress to the affected
7. the fake images or even biometrics can be used for financial fraud
8. the DeepFake paradigm can be used to create fake news and hoaxes and may cause social distress
9. harmful to individuals or organizations
10. can cause exploitation
11. lead to sabotage
12. misrepresentation of democratic discourse
13. manipulation of elections
14. corroding belief in establishments
15. aggravating social separations
16. decline public protection
17. discouragement of international relations
18. endangering national safekeeping
19. discouragement of journalism
20. lead to false allegations

These are to name a few domains, but in brief, we can say that DeepFakes are a forthcoming challenge for national security, individual privacy, and egalitarianism. Therefore, there is an urgent need to restraint the spread of fake digital content and to devise methods that can detect the fake content and have the capability to destroy it, without further spread. (DeepFakes: A Looming Challenge for Privacy, Democracy, and National Security.)

6.4 GAN-based techniques for generating DeepFake

There are mainly two broad ways in which the GANs can generate the DeepFake, one uses an image-to-image translation and the other method is text-to-image synthesis, in the following sections, the GAN-based techniques for generating DeepFake are being discussed.

6.4.1 Image-to-image translation

Image-to-image translation aims to convert one image into another image, in this the goal is to learn how the input image can be mapped to an output image. This technique can be used in a variety of ways such as for transfer of style, image super resolution, image inpainting, transfiguration of the objects, transferring the season of the image and for image enhancement.

The image-to-image translation method is also termed as PIX2PIX translation for image generation, a sample is illustrated in Fig. 6.4. In this, the conditional GANs (CGANs) models are used. The image generation was earlier also there but in that case, for each type of translation, a separate model was required for each of the translation types. But the usage of the CycleGAN paved the way for a cycle-based consistency that had loss enabling inverse conversion ability that too without loss of any information. The introduction of the cyclic technique did not require similar image pairs for training but instead, it has the capability for training the GAN networks on two distinct domains, to learn the features of each domain to translate one into another seamlessly. Apart from the CycleGAN, other models have also been developed for image-to-image translation such as BicycleGAN, StarGAN, etc. and they are being discussed in the following section.

6.4.1.1 StarGAN: Unified generative adversarial networks for multidomain image-to-image translation

StarGAN is a scalable technique developed for image-to-image translation and can be used for several domains by just using a single model. It has a unified architecture based on which it can concurrently train many datasets that too of distinct domains by using a single StarGAN network [1]. It can produce images of superior quality and it is very flexible in translating input images into distinct target domains. In this work, the authors have tried to exhibit the results by transforming the facial expressions of the input images.

6.4.1.2 Toward multimodal image-to-image translation

This method proposed a BicycleGAN model for image-to-image translation by joining two distinct GAN models namely Conditional Variational Autoencoder GAN and Conditional Latent Regressor GAN [2]. They have harnessed the good features of both the approaches and the proposed model has the capabilities to implement the interconnect among the hidden encoding and output individually each director jointly and by that it is

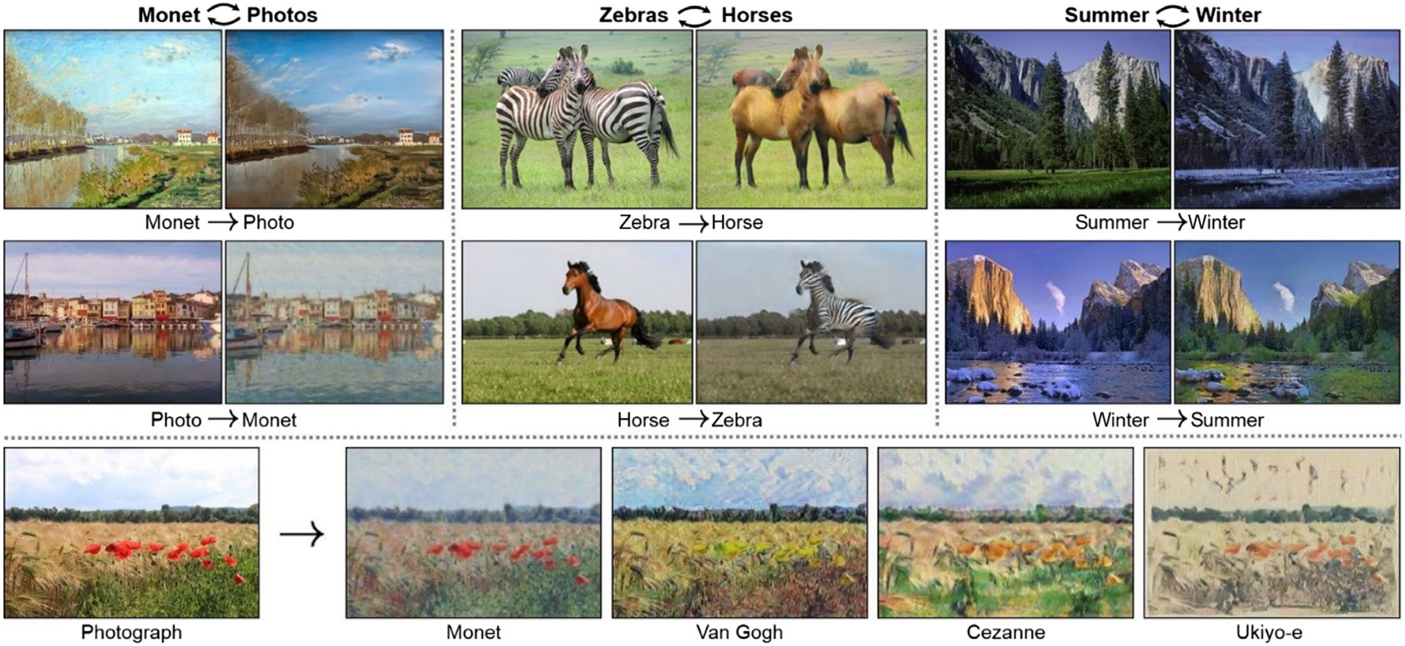

Fig. 6.4 Sample images based on image-to-image translation (https://miro.medium.com).

capable of achieving superior results. This method has been compared against several state-of-the-art encoder methods and is capable of giving superior results in comparison to all of them.

6.4.1.3 U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation

It is an unsupervised image-to-image translation method which has used a novel attention and learnable normalization module, to operate in an end-to-end way. The role of the attention function is to act as a supervisor to the method to pay more attention to the important regions that are distinct in the original and the target image domains that are dependent on the attention map that has been created by the auxiliary classifier [3]. The proposed method is capable of withstanding the geometric or shape changes in the target images. They have also integrated an adaptive layer-instance normalization procedure that supports the attention guided function to easily manage the extent of modification for the shape and texture based on the parameters that it has acquired during the learning phase from the dataset.

6.4.1.4 Image-to-image translation with conditional adversarial networks

The conditional adversarial networks are the baseline methods for carrying out an image-to-image translation, as they can learn from the mapping of the source to target image, and they also learn the loss function to train the further images from this mapping [4]. Due to this method, it is capable of applying the same loss function to a wide variety of images. This method has an associated software with the name PIX2PIX that has been widely used by many artists to experiment with the proposed approach because of the ease of use and varied applicability.

6.4.1.5 Multichannel attention selection GAN with cascaded semantic guidance for cross-view image translation

The multichannel attention selection GAN with cascaded guidance for cross-view image translation aims at the translation of the images with completely distinct views and the images may be suffering from a high degree of deformation, which is a quite challenging task. The proposed work carried out this task with very good precision, in which the system can create natural scene images with random viewpoints that are guided by an input image of the desired scene along with a semantic map that is novel [5]. This method takes input from the semantic maps and is a two-step process, in the first step, the input image and the desired semantic map are fed into the cycled semantic-guided generator to create the initial raw results. In the second step, the initial raw results are further refined by the multichannel attention selection methodology.

6.4.1.6 Cross-view image synthesis using geometry-guided CGANs

In this work, the authors have proposed a cross-view image synthesis method that is based on the geometry-guided CGANs. In this approach, the pixel information is preserved between the two viewpoints, to give a realistic appearance of the input image to the output generated images. To achieve this objective, they have used homography to guide the mapping of the images between the distinct views that are dependent on the overlapping views, so that the details of the image that is input are preserved [6]. To give a realistic image, they have painted the regions that were missing in the image that has been created by transformation by using the GANs. They have used the geometric constraints, due to which the complete minute details can be added to the image thus generated, moreover the proposed approach has given very good results for cross image-based image generation as compared to simple pixel-based image generation methods.

6.4.1.7 Cross-view image synthesis using CGANs

In this work, the authors have proposed a CGANs to generate cross-view images from the natural scene images of aerial to street view and street view to aerial [7], which is a challenging task in the domain of computer vision. It becomes even more challenging when the generation of the new images for a completely different view as the process of conversion of understanding and transforming the image appearance and semantics across different semantic viewpoints is a nontrivial task. In this work, they have used novel architectures namely Crossview Fork and Crossview Sequential that have the capability of generating images that have a resolution of 64 × 64 and 256 × 256. The architecture of Crossview Fork uses one discriminator and one generator. The generator module of the Crossview Fork tries to fantasize about the image along with its semantics for segmentation for the output image. The Crossview Sequential uses two CGANs, out of which the first unit is used for creating the output image that is fed in the second unit to generate the map of the semantic segmentation. To improve the results the feedback is supplied to the first unit from the second unit to improve the quality of the images. The proposed method works well for the generation of the natural scene images by using a cross-view image-to-image translation.

6.4.1.8 WarpGAN: Automatic caricature generation

With the improvement in the GAN-based architectures, automatic caricature generation methods have been developed, which can generate the caricatures from the input image of the face. The WarpGAN architecture cannot only produce the caricatures but can also transform the texture styles [8]. This architecture works by automatically learning to predict a collection of control points that can be further used to transform the image into a caricature and has the capability of preserving the identity of the original photograph as well. The caricatures generated by using the WarpGAN are quite identical to the caricatures that are drawn by using hand, but they have prominent features of the face more exaggerated. This is possible as the WarpGAN uses the identity preserving adversarial loss that helps the discriminator module to differentiate between the distinct images under study and it also gives the option to customize the caricatures thus generated by controlling the styles and the extent of the exaggeration to be produced in the output image.

6.4.1.9 CariGANs: Unpaired photo-to-caricature translation

CariGAN is the first architecture used for the creation of caricature from the input image. It is based upon the two-step process, in the first step, the geometric exaggeration is carried out, whereas in the second step the look and feel style is defined, to carry out these two steps two distinct GAN models are used namely CariGeoGAN and CariStyGAN [9]. The CariGeoGAN carries out the geometrical transformation from the input image to the target caricature and the CariStyGAN translates the look and feel of the caricature to the input photos, but it does not cause any change to the geometrical aspects. This method can easily carry out cross-domain translation by breaking the process into a two-step process, and the output images thus generated closely resemble the hand-generated images, and the caricatures thus generated can be controlled by tuning the parameters to adjust the color and texture of the output images thus generated.

6.4.1.10 Unpaired photo-to-caricature translation on faces in the wild

The unpaired photo-to-caricature translation on faces in the wild is capable of transforming the input photo to the caricature in distinct styles and the same model can be used for other high-end images to image translation applications. Their design uses a two-path approach to detect the overall structure and the local features which are required for carrying out the translation process [10]. They have used two discriminators; one discriminator is coarse and the other is fine. The generator of this model also provides an extra perceptual loss in addition to the loss that is provided by the adversarial and the cycle consistency to attain the learning in two distinct fields. This model can also learn the different styles from the supplementary noise that can be given as input to the model.

6.4.2 Text-to-image synthesis

The GAN-based deep learning architectures have the unique ability to generate the images based on the text descriptors. The system works by giving the text phrase as input, and the GAN model can generate an image based on the description. The sample architecture is shown in Fig. 6.5 that demonstrates the image creation architecture based on the Reed et al. model. The illustrative GAN-based model can successfully convert the text phrases into images. The diagram illustrates the visualization of how the text strings fit so well in the sequential image generation model. The generator network filters the text input using fully connected neural network layers and the random noise is also concatenated in the form of a vector Z, whereas in the discriminator network, the text

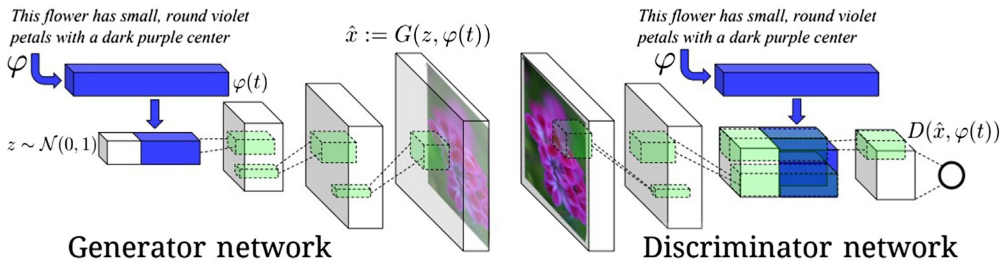
This flower has small, round violet petals with a dark purple center
φ
$\varphi(t)$
$z \sim \mathcal{N}(0,1)$
$\hat{x} := G(z, \varphi(t))$
Generator network
This flower has small, round violet petals with a dark purple center
φ
$D(\hat{x}, \varphi(t))$
Discriminator network

Figure 2. Our text-conditional convolutional GAN architecture. Text encoding $\varphi(t)$ is used by both generator and discriminator. It is projected to a lower-dimensions and depth concatenated with image feature maps for further stages of convolutional processing.

Fig. 6.5 Text-to-image synthesis process (https://www.oreilly.com).

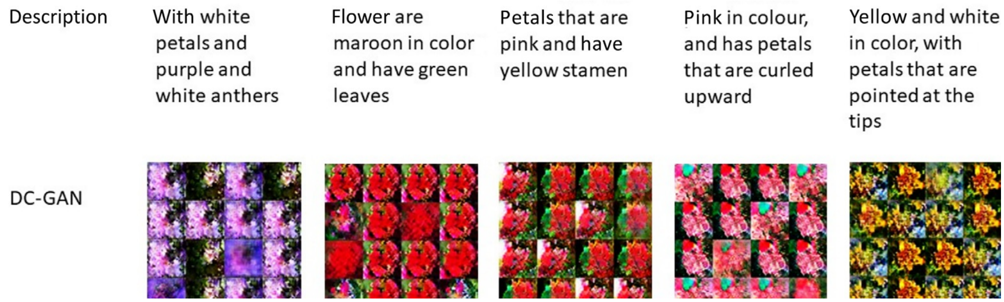
Description
With white petals and purple and white anthers
Flower are maroon in color and have green leaves
Petals that are pink and have yellow stamen
Pink in colour, and has petals that are curled upward
Yellow and white in color, with petals that are pointed at the tips
DC-GAN

Fig. 6.6 Sample text-to-image synthesis (https://cdn-images-1.medium.com).

input is also compressed using a fully connected model as is used in the generator network and then it is recreated and concatenated in the form of an image.

Fig. 6.6 shows a sample of how the text phrases can be converted into the actual images of the flowers. The GAN-based models have been refined and fine-tuned to generate photorealistic images by just taking the text descriptors as input, in the following section, various state-of-the-art methods have been discussed that are being used to generate high-resolution images, based on the textual information.

6.4.2.1 Generative adversarial text-to-image synthesis

The recent advancements in the domain of artificial neural networks have given the power to the computer systems to transform text to pixels. Thereby, facilitating the generation of the images from the text descriptions [11]. All this has been possible due to the recent advancements in the development of a deep convolutional method-based GAN networks. In the proposed work, the authors have tried to generate the images of birds and flowers by giving a detailed description of their structure and features. They have spent considerable effort in creating an efficient GAN model and the training dataset that can create images of birds and flowers from the text descriptors written by humans. They have used five distinct text descriptors along with the dataset of Caltech-UCSD for birds and Oxford-102 for the flowers.

6.4.2.2 StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks

This approach aims at generating the images using the text descriptors. They have used a stacked generative adversarial networks (StackGAN) [12] to produce 256 × 256 images that mimic the realistic images. They have tried to generate photorealistic images by refining the sketches, by decomposing the process of sketch refinement into subtasks. In the initial phase, the GAN-based approach is used to draw the initial shapes of the objects with the colors, based on the text description given, thus yielding a basic low-resolution image. In the second phase, the results of the first stage are combined with the input descriptors in the form of text, to generate more realistic images, and the drawbacks of the first stage are also overcome, thus yielding high-resolution photorealistic images. To give more realistic effects, the proposed model also adopts the conditioning augmentation method that helps to smoothen and condition the image to a great extent. Therefore, this approach is capable of producing images of high resolution, which have very good quality.

6.4.2.3 MC-GAN: Multiconditional generative adversarial network for image synthesis

The proposed method aims at generating an image from the text descriptors when the background base image is already given and the new object can be created at a specified location. This approach is the enhancement of the text-to-image generation phase, as now new objects can be added to the preexisting images and that too at the specified locations, as per the text descriptors. This has been made possible by using multiconditional generative adversarial networks (MC-GAN) [13], which can regulate the background and the desired object simultaneously. This model employs a synthesis block that helps to disassociate the object and the background during the training phase, thereby enabling the MC-GAN to generate as good as real images with a resolution of 128 × 128 by monitoring the extent of the background specifications from the specified base image with the forefront details using the text descriptors. The proposed method is capable of smoothly mixing the possible orientation and the layout of the object with the background image. The method can give excellent results due to the MC-GAN model that can act like a pixel-wise gating function that has the capability of regulating the volume of evidence from the background image with the aid of the text descriptors of the new object that is to be placed in the foreground.

6.4.2.4 MirrorGAN: Learning text-to-image generation by redescription

The authors have developed a three-process method to generate the images from the text descriptors by redescription and have named their method as MirrorGAN [14], the three distinct steps used in this approach are STEM that is an embedding module for the semantic text and it generates word and sentence level embeddings, the second module is GLAM that is based on the cascaded architecture for generating target images from coarse to fine scales, leveraging both local word consideration and global sentence consideration to gradually improve the range and semantic uniformity of the produced images, and the third module is STREAM that aims to regenerate and align the images based on the semantic text.

6.4.2.5 StackGAN++: Realistic image synthesis with stacked generative adversarial networks

This is the improvement of the StackGANs, this approach takes the low-resolution images generated by StackGANs and the text descriptions and then creates high-resolution images that look as real as the realistic images. The proposed method is based on the multistep GAN architecture and is suitable for both conditional and nonconditional methods for GAN-based image generation [15]. In this architecture, there are multiple generators and discriminators organized just like trees, thereby generating multiple-scale images for the similar scenes from the distinct branches of the trees. This approach has a more constant training pattern as compared to StackGAN as it is collaborating with multiple distributions for approximation. They have also integrated the conditioning augmentation to improve the smoothness of the images and also improves the diversity of the images.

6.4.2.6 Conditional image generation and manipulation for user-specified content

In this approach, the authors have created a dataset named CelebTD-HQ that has facial images and the associated text descriptors. The dataset has been generated by using a two-step pipeline, in the first step of the pipeline a textStyleGAN model has been created, which is trained upon the text and in the second part of the pipeline, they have used pre-trained weights of the previously trained textStyleGAN model to carry out the semantic manipulation of the facial images. This approach aims to train the semantic directions based on the latent space [16]. This method is capable of producing conditional images based on the semantic manipulation using the text descriptors.

6.4.2.7 Controllable text-to-image generation

This method can generate high-quality images, from the text descriptors based on natural language using the controllable GANs, in which they have used a generator that is based on the word level spatial and attention, that can generate and manipulate subregions of the image to corresponding relevant textual words [17]. This method also employs a supervisory feedback mechanism based on the text descriptors, which operates by establishing a correlation between the text and the regions of the image, thereby creating an efficient training mechanism that can change particular visual features without disturbing the other content of the image. Thus this method is capable of generating and manipulating the artificially generated images by giving the text-based descriptors. The main focus of this work is to change the category, color, and texture of the images by giving text descriptors.

6.4.2.8 DM-GAN: Dynamic memory generative adversarial networks for text-to-image synthesis

This approach tries to overcome the drawbacks of the initial text-to-image synthesis methods, which greatly rely on the quality of the initial base image, and the contribution of the descriptor works on the different content of the image. In this approach, the authors have used dynamic memory-based GAN approach [18], to synthesize the good quality images from the text descriptors. In this method, the fuzzy contents of the image are improved by using the dynamic memory function. Further, it has two gates named memory writing gate and response gate. The memory writing gate selects the relevant textual information corresponding to the base image content, which further improves the quality of the images generated from the text descriptors and the response gate combines the information gained from the dynamic memory and the attributes of the image. The method has been tested with Caltech-UCSD 200 and Microsoft Common Objects in Context dataset.

6.4.2.9 Object-driven text-to-image synthesis via adversarial training

The ObJGAN method aims to generate realistic images by efficiently capturing the object-level textual information, which is required for the creation of realistic images. This model consists of three components namely the attentive image generator which is driven by the objects, a discriminator based on the objects, and an attention method that is also driven by the objects [19]. In this approach, the text descriptors and a semantic layout that is created well in advance are given as input to the image generator, which is used to create high-resolution synthetic images by a method that refines the coarse images to high-quality images, by an iterative process. In each stage of the iteration, the generator improves the regions of the image by giving due consideration to the words that are associated with the bounding box of that region. The role of the attention layer is to form the labels for the class for each of the words that are used for querying the region and the discriminator checks all the bounding boxes to validate that the objects that are created are at par with the sematic layout of the image that was pregenerated.

6.4.2.10 AttnGAN: Fine-grained text-to-image generation with attentional generative adversarial networks

The proposed method aims to generate the fine-grained synthetic images, by using the attentional GANs [20], which use attention-driven multistep improvement mechanism for generating photorealistic images from the text descriptors. This method subdivides the image into subregions depending on the text descriptors associated with those subregions of the image. It also deploys a deep attentional multimodal that finds out the similarity and finds out the matching of the image and thus trains the generator using the dissimilarity. This method generates a more refined image after each stage and has been tested with the CUB and the COCO dataset.

6.4.2.11 Cycle text-to-image GAN with BERT

In this work, the authors have tried to create the images from the captions of the images, using attention GAN models. Wherein these models learn the attention based on the words to image feature mapping. For the fine-tuning of the model, they have used the cyclic design that can do the mapping of the images back to the caption of the image [21]. The authors have also integrated the pretrained model of the BERT which is based on natural language processing for integrating the initial features of the image in the form of text. The proposed model outperforms the normal attention GAN.

6.4.2.12 Dualattn-GAN: Text-to-image synthesis with dual attentional generative adversarial network

A text-to-image synthesis approach has been described using Dual Attentional Generative Adversarial Network architecture. In this approach, the authors have used double attention methods to improve the local details and the overall structure by considering text descriptors features and the corresponding different regions of the image [22]. There are two different attentions in this work, one is textual attention and the other is the visual attention. The textual attention aims to improve the interface between the text constructs and the visuals, and the visual attention aims to model the internal description of the vision through the spatial axes and the channel, which can help to capture the overall structures of the image in a better way. They have also used the attention embedding method to amalgamate the features from multiple paths. They have stabilized the training of the GAN model by using the spectral normalization and have improved the capability of the CNNs by using the structure based on the inverted residual method.

Throughout just a few years, the GAN-based models have been created that can generate fake images by either transforming the existing images or getting the text descriptors. These models are quite helpful to the researchers to generate the dataset for the training purpose for deep learning-based models, where a large amount of data is required to train the networks, but on the contrary, the same technology can be used for the illegitimate process also, and thus pose a great deal of threat to the mankind. Therefore, methods need to develop that can distinguish between the real and synthetic images. In the following section, the artificial intelligence-based methods are being discussed to detect the DeepFakes.

6.5 Artificial intelligence-based methods to detect DeepFakes

It has been observed that the GAN-based architectures have the capability to produce photorealistic images that can be a concern of security, or it may cheat others by posing false news over social media and falsify the information, thus causing mental agony and revulsion. With the recent advancements in the GANs, the quality of these false images may also improve substantially and may lead to more serious issues. Therefore, it becomes a major issue to devise methods that can distinguish between the real and the GAN-generated false images. Although the GAN-generated images can very well fool the individuals, they cannot escape the computer-based artificial intelligence-powered detectors that are robust and are not vulnerable to the prejudice the humans are. In the following section, we will be discussing the various state-of-the-art methods developed so far to detect the DeepFake images.

6.5.1 Can forensic detectors identify GAN-generated images?

The current work investigates to distinguish between the real images and the GAN-generated fake images. The proposed method verifies the authenticity and originality of the images based on the forensic detectors [23]. In this work, the authors have used two approaches to distinguish between the fake images generated by GAN. The first approach is intrusive, in this case, the detector is created using the GAN architecture; therefore, some of the functions of the GAN are used in the detector to recognize the GAN-generated images. The other approach is nonintrusive, meaning there is no module of the GAN available, and the detector is generated on its own, without any input from the GAN that is used for creating the images. In this work, the authors have verified three nonintrusive methods namely inception scores, face quality assessment method, and the trained VGG 16 network model that is based on the latest features. The intrusive approach can detect the fake images quite efficiently whereas for the nonintrusive approaches, the VGA-based approach is good at detecting the fake images if it has sufficient training data, but the results are not good if there is a mismatch between the training and the test data sets.

6.5.2 Detection of deep network-generated images using disparities in color components

The proposed method aims to detect the fake images using the disparities in the color components of the images [24]. The DeepFake images that are generated by deep networks for the RGB color space and the ones with no specified constraints for the correlation among the color components are very easy to distinguish from the real images. The proposed method detects the statistics of the fake images based on the color components and distinguishes them from the real images. The distinction between the real and fake images has been made based on the feature set that is compact and effective and has been validated on different binary classifiers, and this method works when the generative models are known or unknown.

6.5.3 Detecting and simulating artifacts in GAN fake images

The task of classifying the images, into fake and real is challenging, as the dataset for training is usually unavailable, and the model that is used by the attacker for the generation of the fake images is also not readily available. Therefore, in this approach, the authors have tried to simulate the fake image generation process using an AutoGAN model, and it stimulates the generation of the artifacts of the most common GAN approaches and they have also tried to locate the artifacts that are generated while applying the upsampling operation during the generation of the fake images [25]. Doing this they discovered that the artifacts thus generated are the duplicate copies of the frequency domain spectra, therefore they proposed a spectrum-based classifier rather than a pixel-by-pixel classifier to distinguish between the fake and the real images. This approach has given very good results to detect the CycleGAN-generated fake images.

6.5.4 Detecting GAN-generated fake images using cooccurrence matrices

In this work, the authors have proposed a deep learning and cooccurrence matrices-based combined approach to detect the fake images generated by the GANs. The cooccurrence matrix has been calculated using the color channels of the pixels and then the deep learning-based CNN model has been trained for classification [26]. The pixel-based cooccurrence matrix is directly passed to the deep learning-based model to classify the real and fake images and hence detect the DeepFake images that have been generated using GAN-based models. The proposed method also gave good results, when it was trained and tested on distinct datasets.

6.5.5 Detecting GAN-generated imagery using color cues

The proposed method has been deployed to distinguish between the fake and the real images using the color and saturation-based forensic parameters. For the color-based forensic, they have considered that the GAN-generated images will have a high correlation between the pixels in the chromaticity space as compared to the real-world image. For the saturation-based forensic, the frequency of underexposed and saturated pixels will be reduced as the generator component of the GAN carries out the normalization step [27]. To distinguish between the real images and the GAN-generated images, they have classified the real and fake images using the SVM classifier. This approach was able to achieve the AUC parameter of 70% using the dataset named NIST MFC2018.

6.5.6 Attributing fake images to GANs: Analyzing fingerprints in generated images

This method aims at classifying the images as real or fake based on the GAN attributes that are created with the GAN-generated images. They have also tried to identify the GAN network that has generated the fake image, as each GAN network creates a different fingerprint and a small difference in the GAN training can also change the fingerprint of the GAN-generated fake image [28]. In this, they have used a learning-based method using an attribution network model to map the input image and the fingerprint image comparable to that. For each GAN model, a fingerprint is generated and then the model fingerprint and the actual fingerprint are used for classification to discriminate between the realistic image and the GAN-generated image. The proposed method has been able to achieve 99.5% accuracy using the CelebA dataset and fake images generated using distinct GAN models such as ProGAN, SNGAN, CramerGAN, and MMDGAN.

6.5.7 FakeSpotter: A simple baseline for spotting AI-synthesized fake faces

In this work, the authors have proposed a FakeSpotter that can detect the fake images by analyzing the behavior of the neurons, as it has been observed that for the fake images, the activation function of each layer varies, and that can act as a very strong feature for detecting the fake images, generated by the artificial means. In this approach, the behavior of the neurons has been captured for both the real and the fake images and then an SVM classifier has been trained to classify the real and fake images [29]. The proposed method has been able to achieve an accuracy of 84.7% based on the FaceNet model, they have tested their model on CelebBA-HQ and FFHQ datasets.

6.5.8 Incremental learning for the detection and classification of GAN-generated images

In this work, the authors have proposed a method to detect the unseen images that are fake. They have used a detection model based on multitask incremental learning that can locate and classify the GAN-produced fake images. In this work, they have placed the classifier at different positions based on the iCarl algorithm, to monitor the incremental learning, the two models that are used are named as multitask multiclassifier and multitask single classifier [30]. The proposed model has been tested on five distinct GAN models namely CycleGAN, StyleGAN, StarGAN, ProGAN, and Glow and they used the Xceptionnet model for the process of detection of the GAN-generated fake images.

6.5.9 Unmasking DeepFakes with simple features

Although the DeepFake generation methods have progressed, they have developed to the extent that they can generate an image even from the text descriptions. But the GAN-based methods do leave some artifacts in the fake images that might be missed by the human eye, the artificial intelligence-based methods can catch those artifacts and can easily discriminate between the real and the fake images. In this work, the authors have tried to capture the frequency domain features of the images and a classifier to classify the real and the fake images [31]. The main strength of the method is its capability to give very good results with a limited dataset that has an annotated training set moreover it works very well with unsupervised classifiers. The proposed method is capable of achieving 100% accuracy after being trained with 20 images that are annotated.

6.5.10 DeepFake detection by analyzing convolutional traces

This approach aims at analyzing the human faces present in the DeepFake images, it is based on the fact that the artificially generated images, do leave behind some fingerprints that can be detected by the forensics tools. In this approach, the authors have proposed an expectation-maximization (EM) approach that can extract local features of the images that are specific to the false image generation model [32]. This approach has tested the proposed method with the CELEBA dataset and has been able to detect the fake images along with the different network architectures that have been used to create those fake images.

6.5.11 Face X-ray for more general face forgery detection

This chapter proposes a method to detect the fake images, by converting the input image into its gray-scale image, using which it detects whether the image is the real image or a forged image. If the gray-scale image can be disintegrated into two separate blending images, then it is revealed that it is a fake image as it is having the blending boundary, else it is categorized as the real image [33]. It has been seen that most of the image manipulation methods aim at blending the reformed part into the background of the original image. The method is not able to give good results when it has to detect the low-resolution images, as in that case, the evidence of the blending is less evident, and hence hard to detect.

6.5.12 DeepFake image detection based on pairwise learning

Detecting the GAN-generated fake images is still a challenge, therefore in this approach, the authors have proposed a deep learning approach to detect the fake images based on contrastive loss. In this approach initially, most recent GAN architectures are used to create the fake and real image pairs [34]. In the subsequent step, they have deployed a DenseNet model that is having a two-streamed network model, based on which the pairwise information is fed as input. In this way, a joint network is trained based on the fake features, based on the pairwise learning that can discern the real and artificially created images based on the features. As the last step, a classifier is deployed at the end of the fake feature network to distinguish between the artificially generated and the realistic images.

6.6 Comparative study of artificial intelligence-based techniques to detect the face manipulation in GAN-generated fake images

The GANs have brought in a new era, wherein artificial images can be created by specifying the textual features of the images or by manipulating the already existing images by transforming and manipulation of the pixels of the digital images. The GANs have emerged just a couple of years ago, the fake image generation and fake image detection methods have developed at a rapid rate. In the previous section, we have discussed about the various methods that can detect the GAN-generated fake images. From the literature survey, it has been observed that most of the harm that is done by fake images is attributed to the manipulation of the faces.

This section aims at carrying out a comparative study of the various methods that have been proposed for the detection of the fake images. To carry out the comparative analysis, we have considered four types of methods that can detect the following type of fake images: (1) construction of a new face, (2) swapping of the facial identity, (3) manipulation of facial features, and (4) manipulating the facial expressions. For each type of the fake images, the comparison has been done on the parameters of features and the classifier used. The performance parameters have not been compared as different researchers have used different performance measurement parameters and distinct dataset; therefore, the fair comparison is not feasible.

6.6.1 Techniques for detecting the construction of a new face

The researchers in Ref. [35] have investigated the working of the GAN architectures to trace the different artifacts that can differentiate the original images and the synthetic images. The system has been evaluated using the color-based features and the classification has been carried out using the SVM classifier that is linear. The method can achieve the AUC parameter of nearly 70% using the NIST MFC2018 dataset [36].

Yu et al. [37] discovered that each GAN-based architecture generates a unique fingerprint in the synthetic images, they formulated a learning-based approach using the attribution network model that has the capability of mapping the input image with its equivalent fingerprint image. Therefore, this approach was able to derive a correlation index between the fingerprint of the image and its corresponding model fingerprint that has been used to classify the images. The method has been tested using the dataset named

CelebA [38], which contains the real images and the fake images that have been synthesized using the different GANs as proposed [39–42]. The proposed method has claimed to achieve 99.5% accuracy. Although the system is capable of very good results, but it fails if the images are blur, compressed, noisy, or cropped.

Authors of Ref. [43] inferred that the observation of the neuron behavior can help us to detect the synthetic faces, as the activation of the neuron across different layers generates distinct patterns and can capture the distinct features that can help to detect the manipulated facial attributes, they implemented different face recognition systems based on deep learning [44–46] to learn about the real and fake faces. Based on the features learned, an SVM-based classifier has been trained in order to discriminate between the real and fake images. The proposed work was able to achieve an accuracy of 84.7% using the FaceNet model on CelebA-HQ [39] and FFHQ [47] dataset of real images and InterFaceGAN [48] and StyleGAN [47].

An analysis of the distinct face manipulation approaches has been proposed in Stehouwer et al. [49], where they have proposed that the novel attention mechanisms are capable of giving good results [50] as they help the process as well as enhance the feature maps of CNN architectures. The proposed method can achieve 100% AUC and 0.1% EER for the real face of Refs. [38, 47, 51] datasets and has been tested with the synthetic images created using [39, 47] GAN-based models.

A fake face synthesis approach [52] has been proposed based on steganalysis and the statistics of the real-world natural images. They used a combination of the pixel cooccurrence matrixes and CNN-based deep learning models. They initially validated their approach using the images created using the CycleGAN [53], this method has also been validated using the fake images created using different GAN architectures. They implemented the proposed approach in their work [54] where the validation has been carried using 100K-Face database and can achieve EER value of 7.2%.

Different fake face synthesis systems have been assessed in Neves et al. [54] based on the experimental results using different datasets, they concluded if the experiments are performed in controlled conditions then the results with EER as close to 0.8% are achieved, but if the detection experiments are performed in real-world scenarios then the performance of the proposed systems degrades to a great extent.

To test the methods for the real-world scenarios by Marra et al. [55], experiments have been performed to detect the previously unseen fake images. They used a multitask incremental model based on the learning and have tried to find out the fake images that have been generated using distinct GAN networks.

The comparative analysis of the different techniques for detecting the construction of a new face has been illustrated in Table 6.1 and it can be inferred that most of the work in this domain has been carried out using the CNN classifier and most of the researchers have used image-related features to distinguish between the real and fake images.

Table 6.1 Comparison of different techniques for detecting the construction of a new face.

Work	Features	Classifier
McCloskey and Albright [35]	Color related	SVM
Yu et al. [37]	GAN related	CNN
Kim et al. [43]	CNN neuron behavior	SVM
Stehouwer et al. [49]	Image related	CNN + attention mechanism
Nataraj et al. [56]	Steganalysis	CNN
Neves et al. [54]	Image related	CNN
Marra et al. [55]	Image related	CNN + incremental Learning

6.6.2 Techniques for detecting the swapping of the facial identity

The first study to detect the face swapping has been proposed by Zhou et al. [57], in which the authors have used a two-stream network that can detect the face manipulation. In this, the authors used the fusion of the face classification using CNN based on GoogLeNet [58] and an SVM-based classification approach that used the triplet path which has been trained based on steganalysis features that measure the triplet loss in the patches of the images under consideration for detecting the swapping of the facial identity.

The SwapMe app was evaluated by Li et al. [59] to check the capacity of generalization for the previously trained model that can detect the swapping of the face or the identity. This method turned out be one of the most robust method to detect the swapping of the faces based on the Celeb-DF dataset.

Mesoscopic features of the images were focused using two different neural network models that had different number of layers [60]. In one model, CNN architecture comprising of four convolutional layers and a fully connected (Meso-4) layers were used while in the second model, the Meso-4 layer has been modified that had a different inception module as proposed by Szegedy et al. [58] and it has been named as Mesoinception-4. Initially, the method was tested using the self-created database for detecting the fake images, and it attained the accuracy of 98.4%. Later, it has been tested with the unseen dataset [59] and the proposed method turned out to be robust with other datasets as well as the FaceForensics++ dataset.

The vulnerabilities of the recent face detection approaches namely VGG [44] and FaceNet [46] to DeepFake based on the DeepFakeTIMIT dataset have been described in Korshunov and Marcel [61]. In addition to that they have evaluated the challenges associated with the detection of fake digital content while using the baseline methods. They have used the principal component analysis-based approach for feature reduction and RNN for long short term memory so as to discriminate between the real and fake digital content as proposed in Korshunov and Marcel [62]. They also used image quality measures [63] and the raw faces as the features for the purpose of detection of the fake images. They have used total 129 features based on signal-to-noise ratio, specularity and blur, etc. The PCA with LDA or SVM classifiers have been used for the purpose of classification and they were able to get EER of 3.3% for LQ and EER of 8.9% for HQ using the DeepFakeTIMIT dataset.

The DeepFakes are generally created by merging the synthetic face regions with the real image and doing so leaves certain artifacts that can be traced when the 3D head view of the image is analyzed by Yang et al. [64]. In order to prove their claims, they carried out investigations to find out the differences between the head poses considering the complete set of facial features (68 features were extracted) and the features of the center of the face. The features obtained were normalized and then an SVM classifier has been used for the classification task. This method has been tested with UADFV dataset and an AUC value of 89% has been achieved. They further in Li and Lyu [65] extended this work of the detection of the fake faces using the warping artifacts. In this, they used the CNN to detect the artifacts. The system has been trained using four different variants of CNN as proposed by Refs. [66, 67] and the method has been tested using the UADFV and DeepFakeTIMIT datasets, with very good results.

The authors of Ref. [51] have analyzed the face swapping approaches and evaluated them on distinct detection methods for face swapping and validated the results using the FaceForensics++ dataset. For their evaluation, they considered CNN-based system using steganalysis features [68], CNN-based system with specially tuned layers that can diminish the content in the image [69], a CNN-based system with global pooling layer [70], the CNN Mesolnception-4 [60], and CNN based on XceptionNet [71] using the ImageNet dataset [72]. They concluded that the CNN-based XceptionNet [71] gave the best overall results.

A fake image detection method based on the elementary features of eye color, missing details of eye, teeth, or the reflections, which are generally associated with natural images has been proposed by Matern et al. [73]. They considered the logistic regression and multilayer perceptron [74] for the purpose of classification, and achieved an AUC value of 85.1%.

The fake face detection method has been proposed using the CNN and attention technique by Stehouwer et al. [49], which aims at improving the feature map of the classifiers that are being used. The attention map has the capability to be inserted into any basic neural network, by addition of a convolutional layer, the method was able to achieve the AUC value of 99.43% and EER of 3.1%.

Seeing the popularity and the relevance of the topic, Facebook that contains a huge database of images, has launched a competition named as DeepFake detection challenge, in collaboration with other organizations. They have provided the baseline results using the CNN model, with six convolutional layers, along with a fully connected layer, XceptionNet model trained with face images and with the full images, these base line models have the capability to give precession of 93% with a recall of 8.4%.

Table 6.2 Comparison of different techniques for detecting the swapping of the facial identity.

Work	Features	Classifier
Zhou et al. [57]	Image-related steganalysis	CNN and SVM
Afchar et al. [60]	Mesoscopic level	CNN
Korshunov and Marcel [61]	Lip image-audio speech, image related	PCA + RNN PCA + LDA, SVM
Güera and Delp [75]	Image + temporal	Information CNN + RNN
Yang et al. [64]	Head pose estimation	SVM
Li and Lyu [65]	Face warping artifacts	CNN
Rössler et al. [51]	Image-related steganalysis	CNN
Matern et al. [73]	Visual artifacts	Logistic regression, MLP
Nguyen et al. [76]	Image related	Autoencoder
Stehouwer et al. [49]	Image related	CNN + attention mechanism
Dolhansky et al. [77]	Image related	CNN
Agarwal et al. [78]	Facial expressions and pose	SVM
Sabir et al. [79]	Image + temporal information	CNN + RNN

The comparison of the different techniques for detecting the swapping of the facial identity is presented in Table 6.2. It can be inferred that most of the researchers have used image-related features and a combination of CNN classifier and in some cases, they have used a combination of CNN and some other state-of-the-art classifiers.

6.6.3 Techniques for detecting the manipulated of facial features

In the initial days, the manipulation of the facial attributes was studied to check the robustness of the facial recognition techniques, and the manipulations were tested against the cosmetic surgery, makeup, and the occlusion of the face due to external factors. With the advent of the DeepFakes, the interest for detecting the images with facial attributes manipulated, has again become popular. In Bharati et al. [80], restricted Boltzmann machine-based approach has been used to detect the images that contain manipulated facial features. In this approach, the system for the detection of the manipulated features was given the patches of the face, so as to learn the distinct features of the face and to classify the image as the authentic or the one with manipulated features. The system has been validated using synthetic datasets that have been generated using the ND-IIITD dataset [81] and a set of images of the famous celebrities. The images of the dataset were manipulated using the features such as smile, color of the eyes, shape of the lips, texture of the skin, etc. The system has been able to achieve an accuracy of 96.2% and 87.1% over the celebrity dataset and ND-IIITD dataset, respectively.

The different variants of the CNN architectures have been evaluated by Tariq et al. [82] to detect the manipulation of the facial attributes using the CelebA dataset [38] of real images and they adopted two distinct approaches to generate the fake images, one approach used the ProGAN [39] architecture to generate the fake images and the other set of fake images have been generated using the Adobe photoshop software. The manipulation of images has been done using the cosmetic makeup, adding glasses to the face, changing the hair style, or putting on hats. They considered images of two distinct sizes, namely 32 × 32 and 256 × 256. The GAN-generated images have been detected with 99.99% AUC whereas the Adobe photoshop-generated images have been detected with 74.9% AUC. The CNN model has the capability to detect the machine-generated fake image with very good accuracy, whereas it is capable of giving average results for the images created by the Adobe photoshop software.

An application named as Fake Spotter has been proposed by Kim et al. [43], which is based on the principle that the behavior of the neurons changes across different layers. The activation functions of the neurons across different layers can capture the distinct features, to support the manipulated images. They used face recognition systems as proposed by Parkhi et al. [44], Amos et al. [45], and Schroff et al. [46] for extracting the features and then used the SVM classifier to classify the manipulated and the original images. The proposed method has been tested using the datasets as described in Karras et al. [39] and Karras et al. [47] for the original images and the synthetic datasets generates using InterFaceGAN and StyleGAN approaches, the system was able to achieve an accuracy of 84.7% on FaceNet model.

The facial features manipulation system has been proposed by Jain et al. [83] using the CNN architecture that has six layers for convolution and two layers that are fully connected and it has also used the residual connections as proposed by He et al. [67]. The system has been fed with the nonoverlapping patches of the face, in order to learn the distinct facial features. The classification has been carried out using the SVM classifier, the proposed model has been able to detect the manipulated images with an accuracy of almost 100% using the datasets as proposed by Bharati et al. [80] and the StatGAN [84] generated dataset that has been trained using the CelebA dataset [38].

The attention mechanisms that have the capability to enhance the feature maps of the different CNN architectures have been proposed by Stehouwer et al. [49]. They used the FaceApp to create the fake images with facial features manipulated, using 28 distinct filters that changes the hair style, color of the skin, put or remove the beard, etc., and the fake images that have been synthesized using the StarGAN model using a set of 40 distinct features. They tested the proposed approach using the DFFD dataset and are able to achieve an AUC of 99.9%.

The authors of Ref. [85] have collected the real images and also created the synthetic images using the Adobe photoshop tool named Face Aware Liquify and some manipulated images were created by the professional artists by manipulating the facial features. Then they used the humans to classify the images as real and fake, and the humans were able to classify with almost 50% accuracy, then they subjected the said dataset to deep recurrent networks and thus the automatic system has been able to detect the fake images with an accuracy of 99.8% for the images generated by machines and 99.7% accuracy for the human created fake images.

A steganalysis-based method [56] has been used to detect the fake images with 99.4% accuracy using the StarGAN [84] generated fake images with facial features manipulated and the real images of Liu et al. [38] dataset.

The detection of the fake images has been done by Zhang et al. [86] using the spectrum domain. In this, the RGB channels of the input image are subjected to 2D DFT and a frequency image is generated for each of the RGB channel. The classification has been carried out using the AutoGAN that has the capability to create artifacts similar to GAN without the aid of any trained GAN model. They used the StarGAN [84] and GauGAN [87] for the purpose of evaluation, out of which StarGAN is capable of detecting the images in the frequency domain with 100% accuracy whereas the GauGAN has been able to do so with 50% accuracy.

Table 6.3 illustrates the comparison of different techniques that have been proposed so far for detecting the manipulated facial features. Most of the researchers have used the CNN-based classifiers using the image-related features to distinguish between the real and fake images.

6.6.4 Techniques for detecting the manipulated facial expressions

The advancements in the technology have enabled the computer-based softwares to change the speech of the speaker, along with the facial expressions [88]. In Stehouwer et al. [49], the researchers have proposed a technique to detect the manipulation of the facial features using the DFFD dataset, which achieved the AUC value of 99.4%. It has been observed by Refs. [37, 51, 54] that the results are good in the controlled environments, but most of the methods fail in the real scenarios, therefore new methods need to be explored that have the capability to work in the real scenario wherein there is variation of blur, noise, and compression. The second issue that needs to be addressed is the robustness of the proposed methods against the unseen face manipulation, it has been

Table 6.3 Comparison of different techniques for detecting the manipulated of facial features.

Work	Features	Classifier
Bharati et al. [80]	Face patches	RBM
Tariq et al. [82]	Image related	CNN
Kim et al. [43]	CNN neuron behavior	SVM
Jain et al. [83]	Face patches	CNN + SVM
Stehouwer et al. [49]	Image related	CNN + attention mechanism
Wang et al. [85]	Image related	DRN
Nataraj et al. [56]	Steganalysis	CNN
Marra et al. [55]	Image related	CNN + incremental learning
Zhang et al. [86]	Frequency domain	GAN discriminator

Table 6.4 Comparison of different techniques for detecting the manipulated facial expressions.

Work	Features	Classifier
Zhou et al. [57]	Mesoscopic level	CNN
Rössler et al. [51]	Image-related steganalysis	CNN
Matern et al. [73]	Visual artifacts	Logistic regression, MLP
Nguyen et al. [76]	Image related	Autoencoder
Stehouwer et al. [49]	Image related	CNN + attention mechanism
Sabir et al. [79]	Image + temporal	information CNN + RNN

observed by Refs. [55, 59] that the systems have very poor generalization capability, therefore they fail to give good results in real-world scenarios. The GAN-based methods StyleGAN [47] can detect the face manipulated images with quite good accuracy. The accuracy of the GAN methods is attributed to the fingerprints that are left as artifacts, in the GAN-generated fake images. Further, a research group proposed to eliminate the fingerprints that are generated by the GAN models, so as to make the GAN-generated Fake images, hard to detect [54], they proposed to use autoencoders and degradation of the image quality, but it resulted in the loss of the detection rates.

Table 6.4 illustrates the comparison of different techniques for detecting the manipulated facial expressions. The researchers have mostly used image-related features and CNN classifiers with the benchmark datasets to check the performance of the methods for detecting the fake images generated using different GAN architectures. Most of the methods have been able to achieve high accuracy.

6.7 Legal and ethical considerations

With the rapid development in the domain of artificial intelligence, a situation has emerged that we can now create images by just giving the features of the image in the form of text or we can manipulate any image. The images thus generated are as good as real images and are termed as DeepFakes, as there are generally created using the deep learning-based modalities. The DeepFake images can be quite innovative and can add significant value to the creative and to the education domain, but on the contrary, this innovative methodology have a plethora of threats as well, which can have social, political, and financial harmful implications. The main concern is the DeepFakes are very hard to discern with the human eye as they blur the line between the original and the fake images. Moreover, with the proliferation of the digital media and digital platforms, the DeepFake images can spread like a wild fire on different social platforms.

Therefore, we need to address the legal and the ethical implications attached with the DeepFakes. In order to address this point, the DeepFakes can be categorized into four different categories namely face swapping in order to take revenge and defamation of public figures these two categories are defined under the hard cases and can have hefty legal and ethical complications, on the contrary, the DeepFakes are created for the illustration of creativity or for reducing recapturing, which has social benefit associated with it, they are of the lighter category and have quite few legal and ethical complications associated with them.

The emergence of DeepFakes has posed a serious problem, for which we need to look at the cause of the problem and correct it, instead of just cursing the symptoms associated with it.

The penalties associated with the fake content that is generated are manifold ranging from spreading of misinformation, humiliation of the victims, and propagation of the fake news. The most challenging tasks is how to prevent the propagation of the false information and save the society at large from straightforward implications of DeepFakes. Some countries have passed laws to control the implications of DeepFakes. In which the propagator will be held responsible for posting the DeepFake over the social media that can act like a limiting factor for others from opting the same path. But the harm and humiliation done by once is hard to reverse.

Hence in the era of internet virality and proliferation of social media, the spread of the misinformation is beyond control and the social media platforms have developed as mediums for political and social discourse. In order to curb this menace, either the laws should be framed to tackle the issues, associated with the DeepFakes, or the platforms through which the DeepFakes spread like wild fires, should be equipped with such technologies, which can ascertain the truthfulness of the content that is being posted and censor the fake content, so that it never gets the platform to be launched and hence control the implications associated with it. Table 6.5 illustrates the legalities associated with DeepFake images.

6.8 Conclusion and future scope

Every coin has two sides, the GAN-based systems have a large number of applications, but few of the applications can serve the malicious purpose as well. As it has already been witnessed that how the deep learning-based approaches have been harnessed by the fraudsters, to generate artificial intelligence-based syntactic images and even videos that can be used by the criminals for carrying out scams, fraudulent activities, or to create fake images and even fake news.

On the same line's computational intelligence of the GANs can be harnessed by the fraudsters to use the GAN-generated images and videos for the malicious activities, they can improve their artificial intelligence-based methods by generating the synthetic images of the innocent individuals, whom they have chosen to victimize.

Table 6.5 Legalities associated with DeepFake.

Purpose of DeepFake	Cases	Benefits	Alarms	Affects	Legalities
Face swapping	Swapping the face of the victim with that of others, in order to defame the victim	The person who is swapping is taking revenge, and gets satisfaction	Mental torture and humiliation of the victim	It can have mental torture, abuse, and financial implications to the victim	Criminal proceedings can be initiated
Defamation of public figures	Creating images of the events that never happened	Freedom of expression	Defame a public figure, distort the reputation and even alter the election results	Destroy the international relations, create polarization, and erode the trust in organizations	Public and private law suits can be filled
Reducing recapturing	Dubbing the same video in multiple languages	Reducing the effort of repetitive tasks	May impact the IPR	Redundant data creation	Private law suits can be filled
Creativity	Creation of MEMES	Freedom of expression, creativity	May impact the IPR	People may feel offended	Public and private law suits can be filled

The efforts have been put by the research faternity to find out the techniques to detect the fake images. Most of the works have been carried out using CNN-based classifiers to discern between fake and real images, using the image-related features.

Although much innovations have taken place in the field of artificial intelligence, nobody has given much importance to the security risks. The artificial intelligence-based innovations have posed or may pose in years to come. It is a well-understood fact that in the endeavor to develop intelligent machines, which would mimic the human-like traits and will make the work of the humans easier, but not much importance has been given to the security, privacy, and other risks associated with these advancements.