

Performance evaluation of generative adversarial networks for generating mugshot images from text description

Nur Nabilah Bahrum, Samsul Setumin, Nor Azlan Othman, Mohd Ikmal Fitri Maruzuki, Mohd Firdaus Abdullah, Adi Izhar Che Ani

Electrical Engineering Studies, College of Engineering, Universiti Teknologi MARA, Cawangan Pulau Pinang, Permatang Pauh, Pulau Pinang, Malaysia

Article Info

Article history:

Received Jan 30, 2023

Revised May 24, 2023

Accepted Jul 13, 2023

Keywords:

Face recognition

Face sketch recognition

Face sketch synthesis

Forensic sketch

Generative adversarial network

Text to photo

ABSTRACT

The process of identifying photos from a sketch has been explored by many researchers, and the performance of the identification process is almost perfect, particularly for viewed sketches. Suspect identification based on sketches is one of the applications in forensic science. To identify the suspect using these kinds of methods, a face sketch is required. Hence, the methods require skilled artists to sketch the suspect based on descriptions provided by eyewitnesses. However, the skills of these artists are different from one another, which results in different rendered sketches. Therefore, this work attempts to propose a new identification method based only on forensic face-written descriptions. To investigate the feasibility of the proposed method, this study has evaluated the performance of some text-to-photo generators on both viewed and forensic datasets using three different models of GAN which are SAGAN, DFGAN, and DCGAN. Then, the generated images are compared to the real photo contained within those datasets to evaluate how well the proposed method recognizes the faces. The results demonstrated that the recognition rate for the generated photos by the DCGAN models is better than the other two models which achieve a 38.3% recognition rate at rank-10 for mugshot identification.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Samsul Setumin

Electrical Engineering Studies, College of Engineering, Universiti Teknologi MARA

Cawangan Pulau Pinang, 13500 Permatang Pauh, Pulau Pinang, Malaysia

Email: samsuls@uitm.edu.my

1. INTRODUCTION

Face sketch recognition systems have grown in popularity worldwide due to technological advancements. When it comes to the utilization of face recognition algorithms, one of the most common approaches that are utilized in the field of forensic science is the creation of a forensic sketch in order to identify a suspect. The suspect identification becomes harder when there is no image capture of the suspect available. It becomes challenging for the police to identify the suspect without the image sources. In these cases, a forensic artist is frequently used to collaborate with the witness in order to generate a forensic sketch that portrays the suspect's facial appearance based on the verbal description. After the sketch image of the suspect is done, it will be distributed to law enforcement officers and media outlets, hoping that someone knows who the suspect is [1]. Forensic sketches need to be done because there is no medium that is used to translate the eyewitness description into an image. Therefore, an early attempts of face sketch recognition relied on a human observer, a laborious procedure whose precision was subject to individual differences in skill. Even though the sketches contain all of the necessary information about the suspect face's spatial

topology, it is still a difficult task to perform and very time-consuming to identify and match the forensic sketch to the corresponding photo manually.

Other than that, the most difficult aspect of face sketch recognition is comparing images of different modalities [2]. A face photograph is captured by a digital camera, whereas a face sketch is created by an artist using a different level of detail. Even for the same human subject, the photograph and sketch of the face could be different. The face shape may be exaggerated by the artist, or the texture may be eliminated or replaced [3]. This issue has become a serious problem for forensic investigations when the eyewitness cannot recall the suspect's face in exact detail. From past studies, face sketch recognition could be categorized into two main categories which are intra-modality and inter-modality approaches. Intra-modality approaches attempt to synthesize sketch to the pseudo photo in order to recognize face sketches in the same modality (sketch or photo) and these approaches have been proven to have a great performance as compared to the inter-modality approaches [4]. Even though this existing method is effective in recognizing forensic sketches but it has one major drawback, which is the forensic sketch has to be created before the transformation task (i.e., synthesis) can be performed. If the photo can be directly generated from the description that was provided by the eyewitness, then the artist's sketching style and transformation loss might be reduced, which could contribute to better recognition accuracy.

In recent years, generative adversarial networks (GANs) have been developed with the advent of deep learning. GANs are algorithmic structures that pit two neural networks against one another to generate new, synthetic data instances that can pass as real data. Typically, they are used to generate images, videos, and voices. Reed *et al.* [5] introduced text-to-image synthesis to the public for the first time in 2016 as a result of breakthroughs in GAN it is a significant and pioneering study topic in computer vision [6]. It is similar to reverse image captioning in that it attempts to construct visual characteristics based on the input words. Text-to-image synthesis, which involves the creation of picture captions, explores the visual semantic process in the human brain by exploiting the link between text and image. In addition, it possesses immense potential for creating works of art, forensic sciences, computer-aided design, image searching, and other areas. However, since text-to-image synthesis is still a new approach in computer vision, so the majority of existing studies in this domain are focused on generating images such as flowers or birds from textual descriptions, particularly for testing generative models based on GAN versions and for recreational purposes [7].

Other than flowers and birds, there is also a study that used text-to-image synthesis for generating face images. However, the existing study only focuses on the generation of high-quality images and the generation of images that are congruent with the input descriptions, which are not focusing on the forensic recognition domain. Therefore, this study proposed text-to-image synthesis using three different models of GAN that focus on generating photo-realistic images that are aligned with the input descriptions from the eyewitness. In this study, the GANs model will be used to generate a face photo-realistic image, which could eliminate the sketching process of the criminal suspect. Therefore, this study has the potential to provide an accessible instrument in the forensic image recognition field, which might be helpful and easy to use by the police in identifying the crime suspect from the eyewitness description. In this context, this study will develop a method and approach for identifying the crime suspect without the sketches. This proposed method was implemented using Python language, and the performances of the GAN models were evaluated using Kernel inception distance (KID), fr chet inception distance (FID), and clean-FID, while the recognition rate of the generated images was evaluated using cumulative match curve (CMC).

2. RELATED WORK

In 2018, with the advancement of deep learning, GAN became popular in generating high-quality realistic photos from sketches and sketches from photos. GAN-based methods have shown great results on image-to-image translation problems and in particular, photo-to-sketch synthesis. However, it is limited in generating high-resolution, realistic images. Yu *et al.* [8] proposed a conditional CycleGAN to generate a pseudo image. Particularly, this study designed a feature-level loss to lead the network to produce high-quality pseudo photos. In addition, by combining the benefits of CycleGAN and conditional GANs, the synthetic images were suitable for face recognition against a gallery of images. Next, Wang *et al.* [9] proposed a study using photo-sketch synthesis utilizing multi-adversarial networks (PS2-MAN). This innovative synthesis framework generates low-resolution to high-resolution images in an adversarial style using an iterative process. The hidden layers of the generator are supervised to produce lower-resolution images, then the network is refined to produce higher-resolution images. In addition, as photo sketch synthesis is a paired translation problem, the CycleGAN framework makes utilizes the pair information. In this study, image quality evaluation, and photo-sketch matching were conducted to demonstrate that the proposed framework outperforms existing state-of-the-art systems.

On the other hand, Sannidhan *et al.* [10] conducted a study on face sketch generation using GANs. This study emphasized the use of a conditional generative adversarial network (cGAN) to generate colour

photo images from facial sketches, allowing for efficient classification while avoiding cross-domain problems in sketch identification. The precision of the GAN-generated image is dependent on the quality of the sketches used to train the adversarial network. A trained convolution neural network is used for sketch generation, while a cGANs pix2pix model is used for colour photogeneration. This study demonstrates that the quality of photos generated from trained CNN sketches is better as compared to the actual dataset sketches.

Furthermore, GAN has made considerable strides in recent years. As a result, both the quality and diversity of images available for production have improved. GANs have lately demonstrated tremendous potential in the multimodal production of text and images. Text-to-image synthesis is the reverse task of image captioning in that it generates photorealistic and semantically reliable images from natural text descriptions. The goal of this research area is to generate images that correspond to the meaning of the text. Because of the enormous potential for practical applications, text-to-image conversion has emerged as one of the most active research areas in recent years. There are several researchers have conducted a study on text-to-face images from text descriptions. Sabae *et al.* [11] conducted a study on generating the human face from the textual description using StyleGAN2. This study utilized the latent space of the stylegan in order to manipulate the different facial features. This study had been creating its face image dataset and text dataset and is divided into three major stages. The facial attribute values are extracted from the input text by first processing the text. Then, the values of these features are utilized to manipulate the latent space of the StyleGAN2 model in order to sample the latent code that represents the features. Then the latent code that was extracted is sent to the StyleGAN2 synthesis network so that the final face image can be produced. However, as stated in the publication, the system of this study is not yet stable due to the presence of many failure instances resulting from contradictory input facial characteristics, consecutive latent manipulation, or excessive navigation in particular directions.

Next, Deorukhkar *et al.* [12] conducted a comparative study of three models of face generating from the textual description which are deep convolutional generative adversarial network (DCGAN), deep fusion generative adversarial network (DFGAN), and self attention generative adversarial network (SAGAN) by employing phrase embedding and latent noise as inputs for each model. This study utilized the CelebA dataset, which consists of 200 k celebrity photographs, and constructed their captions by splitting the CelebA dataset's properties into six categories which are face structure, facial hair, hairstyles, facial features, appearance, and accessories. In this study, when compared to DFGAN, the SAGANs and the DCGANs produced a high-quality image however, due to the complex nature of the model architectures it requires more time to train. Wang *et al.* [13] also proposed a study for text-to-face image generation by developing the text-to-face model system that produces images in high resolution, which is $1,024 \times 1,024$, with text-to-image consistency. Additionally, this work generates several varied looks to cover a broad range of undefined facial characteristics organically. This work obtains the vectors and picture embeddings required to alter the normal distribution-sampled input noise vector. A pre-trained high-resolution image generator is then fed the modified noise vector to construct a series of faces with the necessary facial attributes. This is achieved by adjusting the multi-label classifier and image encoder.

3. METHOD

In this section, a detailed explanation of the method used in this study will be presented to see how feasible the proposed method could be used to give a reasonable performance of criminal suspect identification using three different GAN models which is SAGAN, DFGAN, and DCGAN. In this study, the GAN models were used to generate the suspect images based on eyewitness descriptions. Previously, if no image capture sources of the criminal suspect were available, the police would typically find a witness of the crime. Then, the eyewitness will provide the forensic artist with a verbal description of the suspect based on their memory, and the forensic artist will produce a forensic sketch of the suspected person. Then, the forensic sketch will be synthesized by converting the sketch into a pseudo-photo and matching it with the mugshot images from the database. However, in this study, the eyewitness description will be converted to text, which will then be used to directly generate a pseudo-photo of the suspect. Figure 1 shows the system overview of the proposed method in this study. From Figure 1, the traditional matching method requires a forensic sketch before the transformation task (i.e., synthesis) can be performed and the generated photo (i.e., pseudo-photo) will eventually be used to find the potential suspect. While for the proposed method, the stage of producing the forensic sketch could be eliminated, and the pseudo-photo of the suspect could be directly generated from the verbal description that was provided by the eyewitness. In terms of the system performance, the proposed method had been evaluated using KID, FID, and clean-FID, while the performance of the recognition rate was evaluated using CMC.

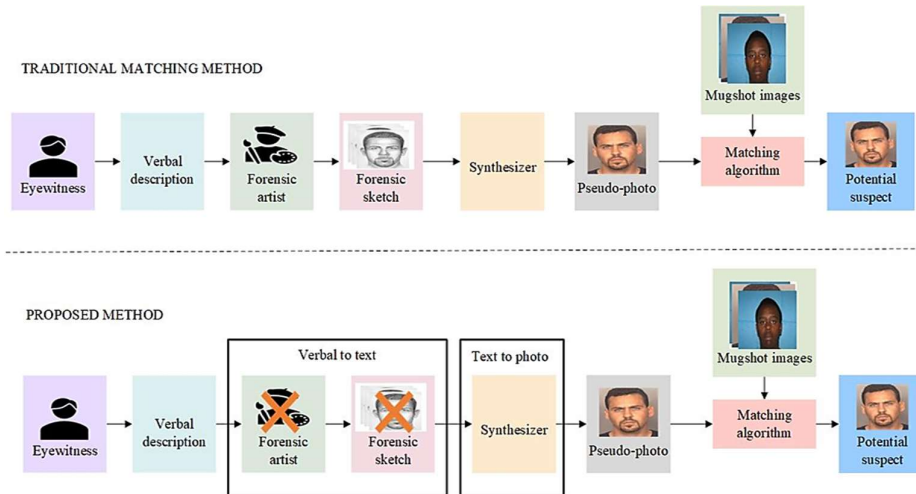


Figure 1. System overview of the proposed method in this study

3.1. Data collection

The facial sketch and photo were collected from the Chinese University Hong Kong (CUHK) database [14], pattern recognition and image processing (PRIP) hand-drawn composite (PRIP-HDC) [15], and memory gap database (MGDB) [16]. The CUHK dataset consists of 188 pairs of viewed sketches and their corresponding photo. MGDB dataset consists of 100 pairs of viewed sketches and their corresponding photo. While the PRIP-HDC dataset consists of 47 pairs of forensic sketches and their mugshot photo. It should be noted that all of the face images from these datasets have different geometrical placements, orientations, and sizes [17]. Directly using these images in this study will result in poor photo generation. To make this better, all of the photo images were geometrically aligned so that the fiducial points of each face image fit into fixed reference points. These photos were lined up by moving, rotating, and resizing the images in two fiducial point alignments. When there are more than two fixed points, the affine transformation is used. Face alignment places similar face components from distinct photos in the same location. In this study, the fiducial points that had been used during training and testing is similar to the study in [4]. Three fiducial points had been selected in this study, which are at the left and right face edge and chin tip. By using affine transformation on these three points, the images are eventually cropped to size 175×140 and the fiducial points are changed into fixed reference points.

3.2. Training and testing datasets

In this study, the collected datasets were divided into two: viewed photos and mugshot photos. For viewed photos, the CUHK dataset is used while for the latter, the MGDB dataset is combined with the PRIP-HDC dataset because the PRIP-HDC dataset has a limited number of samples (i.e., 47 photos). Next, on one hand, for generating the viewed photo based on the text description (i.e., generated from the annotation of the attribute list based on the viewed sketch), 100 photos from the viewed photo dataset had been used as training samples, and the remaining 88 photos as testing samples. On the other hand, generating the mugshot photo based on the text description, the MGDB dataset had been used to train the GANs models. This is because the sample images in the PRIP-HDC and MGDB datasets are matched (i.e., consisting of western human faces) and appropriate for this generating task. It should be noted that using different races of human faces for training will result in poorly generated images of the other races. Once the GANs models are ready, the PRIP-HDC dataset was used to test the system's performance as the sample images in this dataset are the mugshot images of the real crime suspect.

3.3. Text description generation

In this study, for generating the text description of a photo, the attribute list was annotated by three persons. For the training dataset, the attribute list was annotated based on the photo. While for the testing dataset, the attribute list was annotated based on the sketches. Six groups of features that consist of 36 attribute lists had been created similar to the study in [12], [18] in order to convert the attribute list that consists of images from the CUHK dataset, MGDB dataset, and PRIP-HDC datasets into meaningful text descriptions. Starting with the outline of the face and going up to the features that make the face look better, these characteristics describe the face in ascending order. In addition to these characteristics, the person is

also referred to using gender-specific terms, such as “she” and “he”. The first facial characteristics consist of a round face, double chin, oval face, and prominent cheekbones. The facial hairstyle includes a five o’clock shadow, goatee, mustache, and sideburns. Aside from that, this study also focuses on a hairstyle that includes baldness, straight hair, black hair, blonde hair, brown hair, grey hair, bangs, curly hair, and a receding hairline. The following group describes the other facial feature, which includes big lips, a big nose, a pointed nose, narrow eyes, arched eyebrows, bushy eyebrows, and a slightly open mouth. Other than that, the attributes that enhance appearance are youth, attractiveness, a pleasant smile, pale skin, rosy cheeks, and heavy makeup. Accessories, such as earrings, hats, necklaces, neckties, eyeglasses, and lipstick, are the final group of features. The groups of features are set up so that the generator in GANs can build a face by first learning how to make the outline of the face, then adding hair in the specified hairstyle, then making eyes, nose, and other features, then improving the look with words like “young” and “attractive,” and finally adding the accessories listed in the text description. Table 1 shows the group of features with the attribute list that had been used in this study.

Table 1. Group of facial features with the attribute list

Group of facial features	Attribute list
Face outline	Round face, double chin, oval face, and prominent cheekbones
Facial hairstyle	Five o’clock shadow, goatee, mustache, and sideburns
Hairstyle	Baldness, straight hair, black hair, blonde hair, brown hair, grey hair, bangs, curly hair, and a receding hairline
Facial feature	Big lips, big nose, pointed nose, narrow eyes, arched eyebrows, bushy eyebrows, and slightly open mouth
Facial appearance	Youth, attractiveness, a pleasant smile, pale skin, rosy cheeks, and heavy makeup
Accessories	Earrings, hats, necklaces, neckties, eyeglasses, and lipstick

3.4. Sentence encoding

Bidirectional encoder representations from transformers (BERT) established new benchmarks in a variety of sentence classification and sentence-pair regression tasks [19]. BERT utilizes a cross-encoder, whereas two phrases are relayed to the transformer network and the aim value is determined based on these sentences. In this study, as an input to the generator, the semantic vector of the sentence needs to be provided. To accomplish this, this study used sentence-BERT, which is similar to the study in [12]. Sentence-BERT, also known as SBERT, is a variant of the BERT network that makes use of siamese and triplet networks. It is able to generate embeddings for sentences that are semantically significant. In 2019, Reimers and Gurevych [19] proposed a study that demonstrated the various approaches to producing sentence embeddings through the use of BERT that produced inadequate results when used for tasks such as textual similarity. In addition to this, they evaluated the computational efficiency of SBERT in comparison to GloVe embeddings, InferSent, and the universal sentence encoder. Reimers and Gurevych [19], claim that the SBERT algorithm is very efficient computationally. This is because SBERT was approximately 55% faster than universal sentence encoder and approximately 9% faster than InferSent when running on a GPU. This is because of the keen batching method that is utilized, in which sentences of like length are placed together in order to expedite the process. Therefore, this study chose SBERT to transform the text description into an embedding. Each sentence from the description is sent to the model to be analyzed, and the embeddings of those sentences are added to a list. Following the completion of the parsing process for each sentence, the embedding list is then averaged and reshaped (32, 768). For batches of images, the output from (32, 768) is stacked to produce a batch of embeddings with the shape ($|B|$, 768), where B is the batch set.

3.5. Network architectures

In this study, to generate a pseudo-photo from the text description of the viewed sketch and forensic sketch, three different models of GAN had been used which are DCGAN, DFGAN, and SAGAN. The network architecture of DCGAN, DFGAN, and SAGAN that had been used in this study is similar to the study in [12]. GAN is a framework for learning a function or programmed that may produce samples that are quite close to samples obtained from a specified training distribution. GAN is an algorithmic architecture that employs two neural networks, which are generator (G) and discriminator (D), that pit one against the other to generate new, synthetic data instances that can pass for actual data. In this study, the generator network is responsible for generating new data which are generated viewed photo and generated mugshot photo that had been produced by learning from training data distribution in a way that makes it impossible for a discriminator to tell that the generated photo or data is not from the training dataset. For generating the viewed photo, the generator had been trained using 100 photos from the CUHK dataset, while for generating the mugshot photo, the generator had been trained using 100 photos from the MGDB dataset. This is to make

sure the generator could figure out how the training data is spread out and is also capable of generating new photos or data that look like they came from the training dataset. In this study, the generator had been trained by feeding the generator network with text embeddings and random noise. Next, the discriminator network is responsible for figuring out whether the generator's output comes from the training data or not. The discriminator also will give the probability of the sample that was made by the generator, which shows whether the sample is from the same distribution as the training data or not. The discriminator has also been trained with original training samples to better understand how training samples are different.

The training process of GAN is structured like the min-max two-player game [20]. In this game, the discriminator tries to tell the difference between the photo that had been generated by the generator and the training sample. Similarly, in this study, the generator tries to trick the discriminator into thinking that the generated viewed photo and generated mugshot photo came from the training sample. Training process is done when the discriminator predicts that each output from the generator has a 50% chance of generating viewed photo and mugshot photo that is similar to the training sample. It can be said that the model has reached a point where the generator has almost figured out how the training samples are spread out and can turn random data into something that looks like it came from the training data. Both the discriminator and the generator are called multilayer perceptron, and backpropagation is used to train the whole system. In this study, both the generator and the discriminator are being trained at the same time. During each epoch of total data, the discriminator will be trained to get the most accurate predictions of the generated viewed photo and generated mugshot photo. At the same time, the generator is trained to minimize its loss function by making as accurate viewed photo and mugshot photo as possible, which can trick the discriminator into mislabeling it. So essentially, the generator and discriminator participate in a min-max two-player game. In (1) shows the formula for the loss function of the GAN model that describes this min-max two-player game, where G is the generator, D is the discriminator, $Pdata(x)$ is the original data distribution and $Pz(z)$ is input noise min-max two-player game distribution.

$$\min G \max D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} [\log D(x)] + \mathbb{E}_{z \sim p_z(z)} [\log(1 - D(G(z)))] \quad (1)$$

3.6. Performance evaluation

In order to evaluate the performance of the proposed method, FID, clean-FID, KID, and CMC had been used. In this study FID, clean-FID, and KID are utilized to evaluate the quality of images [21] (from text description) that are generated by the selected GAN models. On the other hand, CMC is used to measure the matching identification performance [22] of the generated images with the mugshot images to identify the criminal suspect. The FID metric is based on the assumption that the features computed by a pre-trained inception network have a Gaussian distribution for both real and generated images. FID, in particular, employs the Fréchet distance between two multivariate Gaussian functions, which has a closed-form formula. To compute the FID score for both real and generated images, the Gaussian distributions had been fit to the features extracted by the inception network at the pool3 layer [23]. FID score is calculated by using (2), where the score is referred to as d^2 , μ_1, μ_2 , is presented as the feature-wise mean of the real image and generated images, while C_1 and C_2 is the covariance matrix of the real and generated feature vectors that usually known as sigma and Tr is presented as the trace linear algebra operation [24]. Next, in order to perform the FID evaluation, the input image of the inception network needs to be set to a fixed size where it is resized the output images that had been generated by the generator to match the network's input dimension. The process of resizing the generated images before performing the FID score could leads to different result of the FID score [21]. Therefore, to overcome this problem, this study use clean-FID which has been introduced by Parmar *et al.* [21] as an evaluation of the GAN models.

$$d^2 = \|\mu_1 - \mu_2\|^2 + \text{Tr}(C_1 + C_2 - 2 * \text{sqrt}(C_1 * C_2)) \quad (2)$$

Other than that, KID is also used in this work to evaluate the trained GAN models. KID has been proposed as a replacement for FID [23]. It is said that the FID lacks an unbiased estimator, resulting in a greater expected value on smaller datasets. Therefore, KID is appropriate for smaller datasets because its predicted value is not depending on sample size. It is also less computationally heavy, more numerically stable, and easier to implement. KID is used to measure the squared maximum mean discrepancy (MMD) between the inception representations of the real and generated samples using a polynomial kernel to evaluate generative models [23]. This is a non-parametric test, so it doesn't make the strict Gaussian assumption. Instead, it just assumes that the kernel is a good way to measure how similar two things are. Since the dataset of this study is fewer so the KID evaluation is the most suitable method to evaluate GANs models because it doesn't have to fit the quadratic covariance matrix. In FID, clean-FID, and KID, a lower score indicates that the images generated are closer to the real image.

Finally, to evaluate the matching performance of the proposed method, CMC is used. The CMC performs a cumulative measurement across the ranks to determine the percentage of correct identities. In rank-1, accuracy merely refers to the percentage of right matches made purely based on the shortest distance, which is comparable to the retrieval rate. The rank-k accuracy gives the retrieval rate of the correct match over the first k shortest distances. For example, if the rank-k percentage is 100%, which indicated that the proposed method is capable of shortlisting k face candidates without an error.

4. RESULTS AND DISCUSSION

This section focuses on the results obtained in this study. It will include an in-depth discussion of the findings in this study. The discussion of the discriminator and generator losses for generating viewed and mugshot photo had been discussed in subsection 4.1. While for the performance evaluation of GANs models had been discuss in subsection 4.2, and the performance evaluation of the recognition rates had been discussed in subsection 4.3.

4.1. Discriminator and generator losses

The discriminator and generator loss patterns for the three different models of GANs, which are DCGAN, DFGAN, and SAGAN, had been generated during the training progress as shown in Figure 2. In this study, in order to generate the viewed photo based on the text description, the dataset used during the training progress is from the CUHK dataset, while the dataset used to generate the mugshot photo is from the MGDB dataset. A GAN has converged when both the generator and discriminator losses have converged to a stable value [25]. In this study, the generator and discriminator pair have been trained with learning rates of 0.0001 and 0.0004. Then, the Adam optimizer is utilized, with the parameters $\beta_1 = 0.5$ and $\beta_2 = 0.5$ for DCGAN models, while $\beta_1 = 0$ and $\beta_2 = 0.9$ for DFGAN and SAGAN models. After watching the convergence of loss values during training for all the GAN model, the number of training iterations was determined to be 1,000.

Figure 2 shows the loss pattern for the training progress of DCGAN, DFGAN, and SAGAN for generating viewed photos and mugshot photos. As shown in Figures 2(a) and 2(b), the generator and discriminator of DCGAN model losses began to converge after only 200 iterations and produce a consistent result until 1,000 iterations. However, for the DFGAN and SAGAN model only discriminator gets better after 200 iterations, but generator performs worse and could not achieved a stable losses pattern even had been trained until 1,000 iterations. In this study both of generator and the discriminator were trained in a 1:1 optimization step ratio. One discriminator optimization step is followed by one generator optimization step in a single iteration. From Figure 2 also, the pattern losses of the discriminator and generator for the DCGAN model indicate better and more stable for both training datasets, the CUHK and MGDB datasets, compared to the loss pattern of the DFGAN and SAGAN models.

4.2. Performance evaluation of generative adversarial networks models

The evaluation of GAN models had been performed using three evaluation matrices which are FID, clean-FID, and KID. Table 2 shows the performance of GAN models in generating viewed photos and mugshot photos. As shown in Table 2, the DCGAN models outperformed the DFGAN and SAGAN models for generating the viewed photo by having the lowest scores for FID, clean-FID, and KID values. It should be noted that in FID, clean-FID, and KID, a lower score indicates that the images generated are closer to the real image. This could be visualized in Figure 3, which shows a comparison of a real viewed photo and a generated viewed photo that had been generated based on the text description using the DCGAN, DFGAN, and SAGAN models. Based on Figure 3, the generated viewed photo from the DCGAN model is better and closer to the real viewed photo when it is compared to the generated viewed photo that had been generated from DFGAN and SAGAN models.

On the other hand, the performance of the DCGAN model in generating the mugshot photo also had been outperformed the DFGAN and SAGAN models with the FID value is 137.988, clean-FID value is 139.175, and KID value is 0.093, which are the lowest score among the three models in generating the mugshot photo. Figure 4 shows the generated mugshot photo that had been generated using DCGAN, DFGAN, and SAGAN models. From Figure 4, the generated mugshot photo that had been generated using DCGAN is closer to the real mugshot photo compared to the DFGAN and SAGAN models. Other than that, to compare the performance of the DFGAN and SAGAN models, the DFGAN model performs better than the SAGAN model since the FID, clean-FID, and KID score of the DFGAN model is better than the SAGAN model and based on the generated mugshot photo and viewed photo, DFGAN model produced a better generated images as compared to the SAGAN models.

4.3. Performance evaluation of recognition rate

The recognition performance of the generated images had been evaluated using CMC. In this study, the CMC accumulates the rate of correct identification recognition between the generated images and their corresponding photos across ranks-1 to rank-10. Table 3 shows the accuracy of the recognition rated at rank-10 for both generated viewed photo and generated mugshot photo using DCGAN, DFGAN, and SAGAN models. From Table 3, for both generated viewed photo and generated mugshot photo, it can be seen that the DCGAN model performs better than DFGAN and SAGAN models. However, using the DCGAN model, the generated mugshot photo has a better recognition accuracy at rank-10 which is 38.30% when it is compared to the generated viewed photo which only reaches 23.86%. This accuracy gap is affected by the number of the testing images (i.e., generated mugshot photo only has 47 generated images with their corresponding photo while the generated viewed photo has 88 generated images with their corresponding photo). To elaborate further, even though the generated viewed photo has a lower accuracy as compared to the generated mugshot photo, 21 correct identities can be recognized using the generated viewed photo at rank-10 while the generated mugshot photo is only able to recognize 18 correct identities at rank-10.

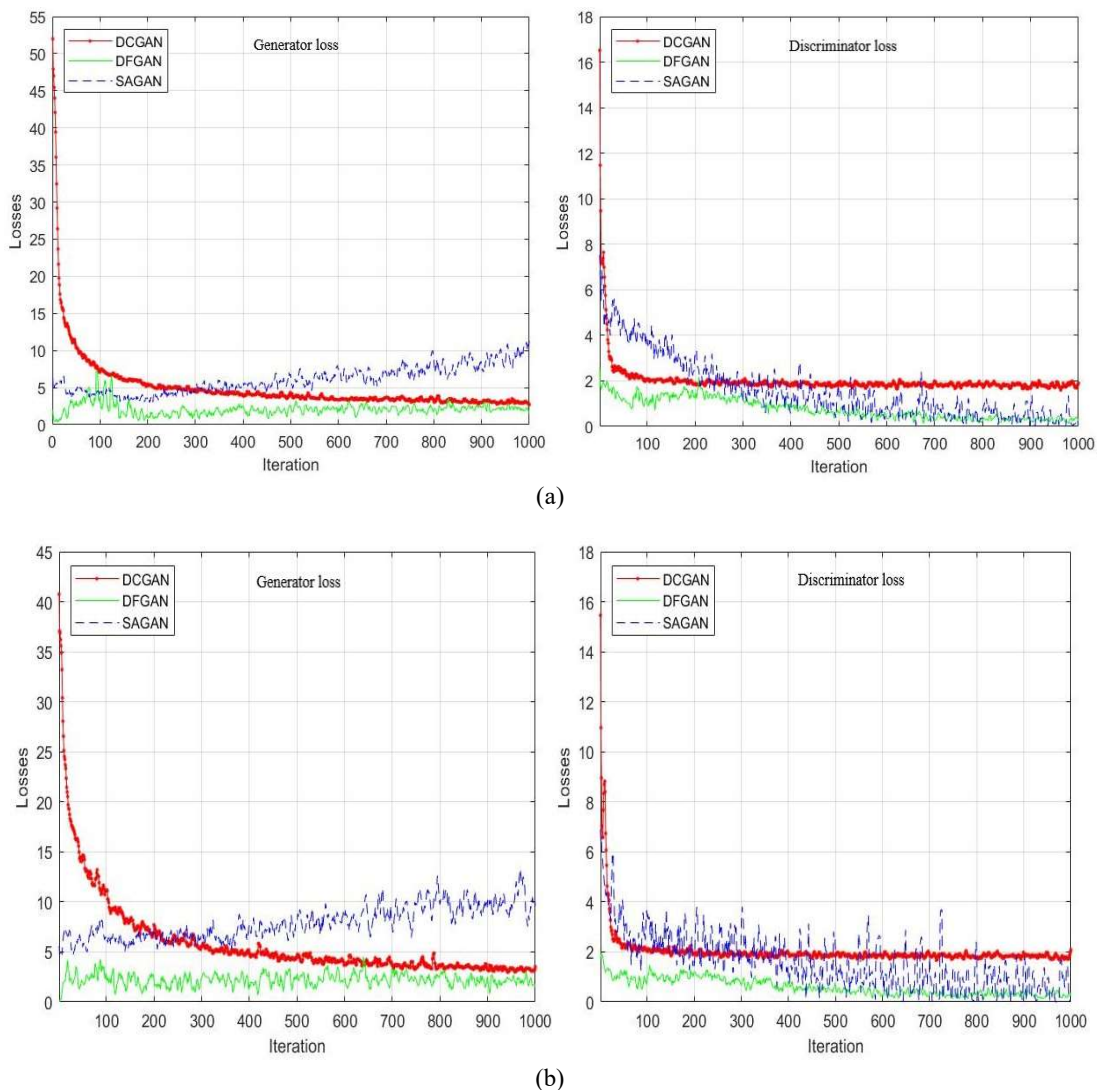


Figure 2. Generator and discriminator training losses for the DCGAN, DFGAN, and SAGAN model; (a) losses for generated viewed photo and (b) losses for generated mugshot photo

Table 2. The performance of GAN models in terms of the generated image quality

GAN models	Generated viewed photo			Generated mugshot photo		
	FID	Clean-FID	KID	FID	Clean-FID	KID
DCGAN	92.799	91.791	0.086	137.988	139.175	0.093
DFGAN	177.466	192.245	0.243	144.384	157.213	0.113
SAGAN	228.144	225.902	0.291	236.146	251.025	0.237

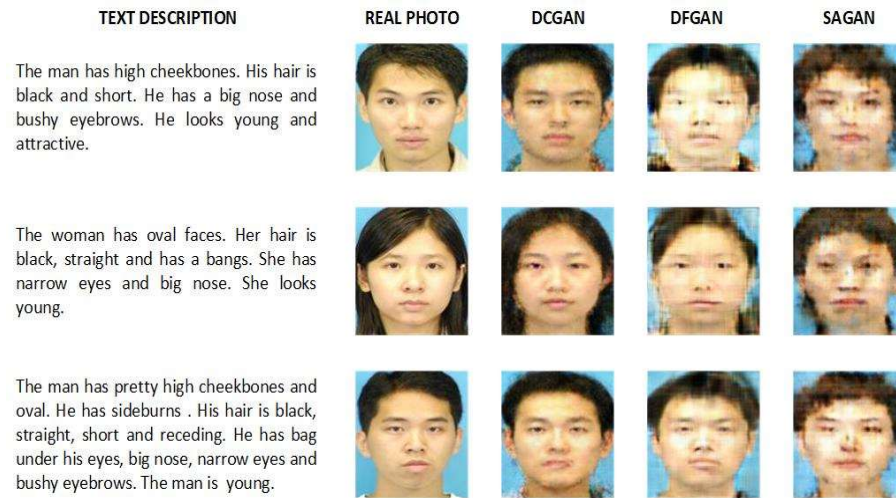


Figure 3. Comparison of real viewed photos and generated viewed photos with the text description

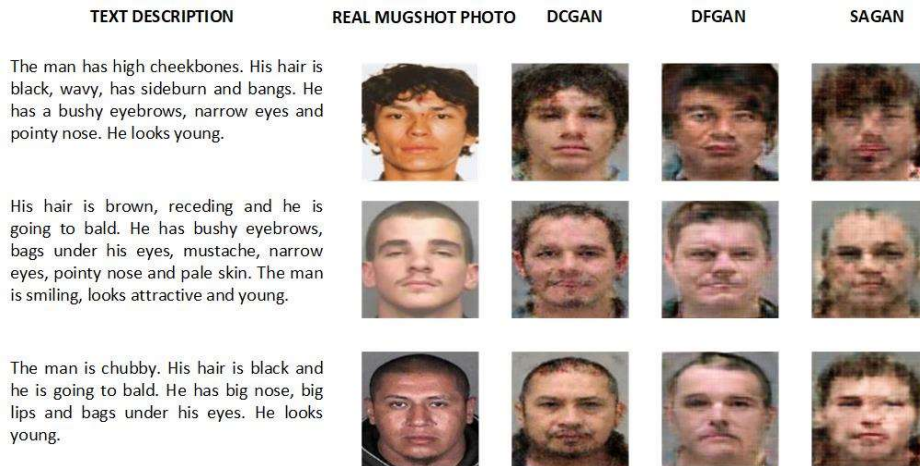


Figure 4. Comparison of real mugshot photos and generated mugshot photos with the text description

Table 3. The recognition accuracy of GAN models at rank-10

GAN models	Generated viewed photo	Generated mugshot photo
	Recognition rate (%)	Recognition rate (%)
DCGAN	23.86	38.30
DFGAN	20.45	27.66
SAGAN	13.64	19.15

Other than having a less testing sample, the occlusions that exist in the forensic sketches themselves could lead to the poor generation of the mugshot photo. The majority of the suspects will typically cover their identities during committing a crime by wearing a face mask, eyeglasses, a hoodie, and a cap. Since the text description in this study is based on the forensic sketch, the face part of the suspect that has an occlusion could not be annotated. In addition to the Table 3, Figure 5 depicts the accuracy across the first ten ranks of

both methods. As can be seen in Figure 5(a), the DCGAN models for recognizing the generated viewed photo perform better at ranks-3 until rank-10 as compared to the DFGAN and SAGAN models. While for recognizing the generated mugshot photos which had been display in Figure 5(b), DCGAN model performs better than DFGAN and SAGAN at rank-2 until rank-10. Therefore, based on the results obtained, the generated images from DCGAN models having better recognition rate for both generated viewed photos and generated mugshot photos as compared to the generated photos from the DFGAN and SAGAN models.

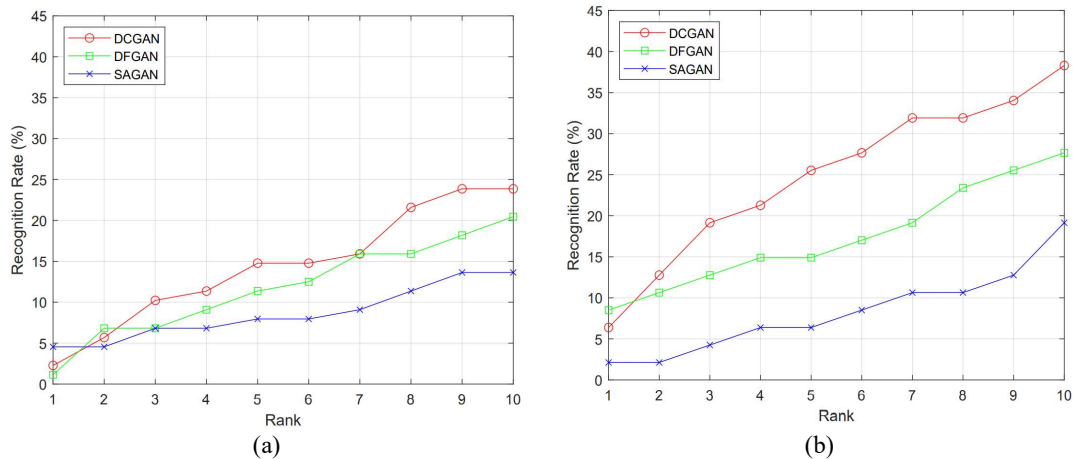


Figure 5. The accuracy across the first ten ranks of DCGAN, DFGAN, and SAGAN models; (a) recognizing the generated viewed photos and (b) recognizing the generated mugshot photos

5. CONCLUSION

In conclusion, the performance evaluation of GANs for generating mugshot images from text descriptions has been done in this study. The study attempts to propose a new method to identify the suspect based only on the text description of the suspect. By doing this, the stage of producing the forensic sketch could be eliminated, and the pseudo-photo of the suspect could be directly generated from the verbal description provided by the eyewitness. For that, three different GANs models which are DCGAN, DFGAN, and SAGAN have been evaluated to see the feasibility of these kinds of models to be used in the proposed method. Based on the results obtained, the DCGAN models were able to produce better-generated mugshot photos and generated viewed photos that are demonstrated by the best FID, clean-FID, and KID score values as compared to the DFGAN and SAGAN models. In addition to that, by using the generated images from DCGAN models, the recognition rate for both generated viewed photos and generated mugshot photos are better as compared to the generated photos from the DFGAN and SAGAN models. Finally, it can be deduced that the proposed method has the potential to be used by the police and law enforcement agencies to identify a criminal suspect based only on eyewitness descriptions in cases where there is no forensic artist available.

ACKNOWLEDGEMENTS

The authors would like to thank Universiti Teknologi MARA, Cawangan Pulau Pinang, especially the members of Machine Learning Research Group (MLRG), Electrical Engineering Studies, College of Engineering for their support and assistance in completing this research work.




REFERENCES

- [1] B. Klare, Zhifeng Li, and A. K. Jain, "Matching Forensic Sketches to Mug Shot Photos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 3, pp. 639–646, Mar. 2011, doi: 10.1109/TPAMI.2010.180.
- [2] X. Tang and X. Wang, "Face Sketch Recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 1, pp. 50–57, Jan. 2004, doi: 10.1109/TCSVT.2003.818353.
- [3] S. Setumin and S. A. Suandi, "Cascaded Static and Dynamic Local Feature Extractions for Face Sketch to Photo Matching," *IEEE Access*, vol. 7, pp. 27135–27145, 2019, doi: 10.1109/ACCESS.2019.2897599.
- [4] S. Setumin, M. F. C. Aminudin, and S. A. Suandi, "Canonical Correlation Analysis Feature Fusion with Patch of Interest: A Dynamic Local Feature Matching for Face Sketch Image Retrieval," *IEEE Access*, vol. 8, pp. 137342–137355, 2020, doi: 10.1109/ACCESS.2020.3009744.




- [5] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative Adversarial Text to Image Synthesis," 2016, doi: 10.48550/ARXIV.1605.05396.
- [6] R. Zhou, C. Jiang, and Q. Xu, "A survey on generative adversarial network-based text-to-image synthesis," *Neurocomputing*, vol. 451, pp. 316–336, Sep. 2021, doi: 10.1016/j.neucom.2021.04.069.
- [7] S. Pande, S. Chouhan, R. Sonavane, R. Walambe, G. Ghinea, and K. Kotecha, "Development and deployment of a generative model-based framework for text to photorealistic image generation," *Neurocomputing*, vol. 463, pp. 1–16, Nov. 2021, doi: 10.1016/j.neucom.2021.08.055.
- [8] S. Yu, H. Han, S. Shan, A. Dantcheva, and X. Chen, "Improving Face Sketch Recognition via Adversarial Sketch-Photo Transformation," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, Lille, France: IEEE, May 2019, pp. 1–8, doi: 10.1109/FG.2019.8756563.
- [9] L. Wang, V. Sindagi, and V. Patel, "High-Quality Facial Photo-Sketch Synthesis Using Multi-Adversarial Networks," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, Xi'an: IEEE, May 2018, pp. 83–90, doi: 10.1109/FG.2018.00022.
- [10] M. S. Sannidhan, G. A. Prabhu, D. E. Robbins, and C. Shasky, "Evaluating the performance of face sketch generation using generative adversarial networks," *Pattern Recognition Letters*, vol. 128, pp. 452–458, Dec. 2019, doi: 10.1016/j.patrec.2019.10.010.
- [11] M. S. Sabae, M. A. Dardir, R. T. Eskarous, and M. R. Ebbad, "StyleT2F: Generating Human Faces from Textual Description Using StyleGAN2," 2022, doi: 10.48550/ARXIV.2204.07924.
- [12] K. Deorukhkar, K. Kadamala, and E. Menezes, "FGTD: Face Generation from Textual Description," in *Inventive Communication and Computational Technologies*, G. Ranganathan, X. Fernando, and F. Shi, Eds., in Lecture Notes in Networks and Systems. Singapore: Springer Nature, 2022, pp. 547–562, doi: 10.1007/978-981-16-5529-6_43.
- [13] T. Wang, T. Zhang, and B. Lovell, "Faces à la Carte: Text-to-Face Generation via Attribute Disentanglement," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA: IEEE, Jan. 2021, pp. 3379–3387, doi: 10.1109/WACV48630.2021.00342.
- [14] X. Wang and X. Tang, "Face Photo-Sketch Synthesis and Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 1955–1967, Nov. 2009, doi: 10.1109/TPAMI.2008.222.
- [15] S. J. Klum, H. Han, B. F. Klare, and A. K. Jain, "The FaceSketchID System: Matching Facial Composites to Mugshots," *IEEE Trans. Inform. Forensic Secur.*, vol. 9, no. 12, pp. 2248–2263, Dec. 2014, doi: 10.1109/TIFS.2014.2360825.
- [16] S. Ouyang, T. M. Hospedales, Y.-Z. Song, and X. Li, "ForgetMeNot: Memory-Aware Forensic Facial Sketch Matching," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 5571–5579, doi: 10.1109/CVPR.2016.601.
- [17] S. Setumin and S. A. Suandi, "Difference of Gaussian Oriented Gradient Histogram for Face Sketch to Photo Matching," *IEEE Access*, vol. 6, pp. 39344–39352, 2018, doi: 10.1109/ACCESS.2018.2855208.
- [18] O. R. Nasir, S. K. Jha, M. S. Grover, Y. Yu, A. Kumar, and R. R. Shah, "Text2FaceGAN: Face Generation from Fine Grained Textual Descriptions," in *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, Sep. 2019, pp. 58–67, doi: 10.1109/BigMM.2019.00-42.
- [19] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China: Association for Computational Linguistics, 2019, pp. 3980–3990, doi: 10.18653/v1/D19-1410.
- [20] I. Goodfellow *et al.*, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020, doi: 10.1145/3422622.
- [21] G. Parmar, R. Zhang, and J.-Y. Zhu, "On Aliased Resizing and Surprising Subtleties in GAN Evaluation," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, Jun. 2022, pp. 11400–11410, doi: 10.1109/CVPR52688.2022.01112.
- [22] R. M. Bolle, J. H. Connell, S. Pankanti, N. K. Ratha, and A. W. Senior, "The Relation between the ROC Curve and the CMC," in *Fourth IEEE Workshop on Automatic Identification Advanced Technologies (AutoID'05)*, Buffalo, NY, USA: IEEE, 2005, pp. 15–20, doi: 10.1109/AUTOID.2005.48.
- [23] E. Betzalel, C. Penso, A. Navon, and E. Fetaya, "A Study on the Evaluation of Generative Models," 2022, doi: 10.48550/ARXIV.2206.10935.
- [24] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium," *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 6627–6638, 2017, doi: 10.48550/ARXIV.1706.08500.
- [25] N. Kodali, J. Abernethy, J. Hays, and Z. Kira, "On Convergence and Stability of GANs," pp. 1–18, 2017, doi: 10.48550/ARXIV.1705.07215.

BIOGRAPHIES OF AUTHORS






Nur Nabilah Bahrum    received her B.Eng. (Hons.) in Electrical and Electronic Engineering from Universiti Teknologi MARA, Malaysia in 2022. She is currently pursuing a Master of Science in Electrical Engineering at the School of Electrical Engineering, College of Engineering, University Teknologi MARA (UiTM) in Pulau Pinang, Malaysia, focusing on deep learning approach in forensic sketch recognition. Her research interests include deep learning, generative adversarial networks (GAN), image processing, and computer vision. She can be contacted at email: numabilahbahrum@gmail.com.






Samsul Setumin    received a B.Eng. degree (Hons.) in Electronic Engineering from the University of Surrey, in 2006, and an M.Eng. degree in Electrical-Electronic and Telecommunication from the Universiti Teknologi Malaysia, in 2009. He obtained his Ph.D. degree from Universiti Sains Malaysia in 2019 in the imaging field. Since 2010, he has been a lecturer with the Universiti Teknologi MARA, Malaysia. He was a test engineer with Agilent Technologies (M) Sdn. Bhd., and the industrial attachment staff at Intel Microelectronics (M) Sdn. Bhd., for one year. His research interests include computer vision, image processing, pattern recognition, and embedded system design. He can be contacted at email: samsuls@uitm.edu.my.






Nor Azlan Othman    received his B.Sc. (Hons.) in Electrical and Electronics Engineering from Universiti Tenaga Nasional (UNITEN), Malaysia. He was awarded the MARA Excellence Scheme Program to pursue M.Sc. in Control Systems Engineering at the University of Sheffield, United Kingdom. He worked as an R&D Engineer for Sony and Motorola Malaysia for several years. In 2015, he received his Ph.D. in Bioengineering from the University of Canterbury, New Zealand. He is currently a senior lecturer at UiTM Pulau Pinang's Faculty of Electrical Engineering. His research interests include physiological modelling, parameter identification for type 2 diabetes, renewable energy, and control systems. He can be contacted at email: azlan253@uitm.edu.my.






Mohd Ikmal Fitri Maruzuki    received the B.Eng. degree in Computer Engineering from Ehime University, Japan, in 2004, and the master's degree in Communication and Computer Engineering from Universiti Kebangsaan Malaysia (UKM), in 2010. He is currently employed as a lecturer at the School of Electrical Engineering, College of Engineering, Universiti Teknologi MARA (UiTM), Penang Branch, Malaysia. His research interests include deep learning, image processing, and embedded systems. He can be contacted at email: ikmalf@uitm.edu.my.



Mohd Firdaus Abdullah    received his M.Sc. degree in Science (Electrical Engineering) from the Faculty of Electrical Engineering, Universiti Teknologi MARA, Malaysia in 2012. Currently, he is doing his Ph.D. at Universiti Teknologi MARA, Malaysia. His areas of research focus include the image processing of medical imaging, specifically in analyzing CT scan images as well as deep learning. He can be contacted at email: f.abdullah@uitm.edu.my.



Adi Izhar Che Ani    is a senior lecturer at the Centre for Electrical Engineering, Universiti Teknologi MARA, Cawangan Pulau Pinang (UiTM CPP), with a master's degree in Engineering from Universiti Malaya Malaysia (2012). He obtained his bachelor's degree in Electrical and Electronics Engineering from the University of Miyazaki (Japan) in 2007. His research interests are the fields of artificial intelligence. He can be contacted at email: adiizhar@uitm.edu.my.