Remove glasses diffusion model an innovative conditioned of eye glasses removal with image diffusion model

Yuliza[1], Rachmat Muwardi[1], Galatia Erica Yehezkiel[2], Mirna Yunita[2], Lenni[3]
[1]Department of Electrical Engineering, Universitas Mercu Buana, Jakarta, Indonesia
[2]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
[3]Department of Electrical Engineering, Universitas Muhammadiyah Tangerang, Tangerang, Indonesia

Article Info

Article history:

Received Jun 5, 2024
Revised Nov 18, 2024
Accepted Nov 24, 2024

Keywords:

Conditional modeling
Deep learning
Diffusion model
Facial images
Glasses removal

ABSTRACT

The presence of eyeglasses in facial images poses challenges for image processing, particularly in facial recognition. This paper introduces the remove glasses diffusion model (RGDM), a conditioned denoising diffusion probabilistic model (DDPM) designed for precise glasses removal. RGDM employs conditional modeling to focus on the glasses region while seamlessly restoring facial features. An eyes position accuracy mechanism, leveraging facial landmarks, ensures accurate eye restoration post-removal. Comprehensive evaluations on the CelebA dataset demonstrate RGDM's superior performance, achieving the lowest Fréchet inception distance (FID) of 27.09 and learned perceptual image patch similarity (LPIPS) of 0.299, outperforming state-of-the-art methods such as 3D synthetic, cycle-consistent generative adversarial network (CycleGAN), and eyeglasses removal generative adversarial network (ERGAN). These results highlight the model's effectiveness in producing natural and high-fidelity facial reconstructions. This work advances glasses removal technology and underscores the significance of conditional models in image processing. The proposed approach has practical implications for facial recognition and image enhancement, paving the way for more accurate and robust real-world applications.

Corresponding Author:

Yuliza
Department of Electrical Engineering, Universitas Mercu Buana
Jakarta, Indonesia
Email: yuliza@mercubuana.ac.id

1. INTRODUCTION

Facial images are critical visual data sources in various applications, including facial recognition, image analysis, and digital communication. However, the presence of eyeglasses in facial images introduces significant challenges, particularly in accurately representing facial features, which are crucial for applications such as security systems, where precise identification is paramount. Reflections on the glass surface, lens distortions, and diverse frame styles contribute to visual artifacts that obstruct accurate facial recognition [1], [2]. In applications such as security systems, where precise identification is paramount, the presence of glasses becomes a significant bottleneck [3]. Facial recognition technology [3], [4] also known as FRTs, are one of several biometric tools or modalities developed to detect and identify persons when their photos are taken by a camera lens [5], [6]. It is a cornerstone in contemporary security and identification systems, [7]-[9] which relies heavily on the accurate representation of facial features. The presence of glasses complicates this process, leading to potential misidentifications and reduced system efficacy [10].

Image inpainting has slowly grown in importance as a study area in digital image processing due to the advancement of science and technology, the rapid upgrading of hardware, and the desire for high-quality images [11]. It can restore the texture of an image, extract its high-level abstract properties, as well as semantic imagery like human faces [12], which is why image inpainting is often used in researches where image generation is necessary, such as the removal of glasses. Coccia [13] emphasized in 2017, problem-driven innovations often lead to solutions that provide a competitive edge. The evolution of technology, as discussed by Coccia in 2019 [14], [15], involves the continuous substitution of older methods with more efficient ones. In this context, our research represents an incremental innovation, building upon existing generative adversarial network (GAN)-based approaches and evolving towards a more effective diffusion model framework. While previous studies, such as the one conducted in 2023 by Sadik-E-Tawheed *et al.* [16], have focused on removing specific artifacts like glare and reflections, our approach tackles the root cause by removing the glasses themselves. This comprehensive solution not only enhances the clarity of facial images but also improves facial recognition accuracy, as highlighted by Mao *et al.* [17] in 2021.

This paper introduces the remove glasses diffusion model (RGDM), a novel approach that departs from traditional GAN-based frameworks by leveraging a denoising diffusion probabilistic model (DDPM), which represents a significant advancement in this area. The RGDM integrates a conditioning module that focuses specifically on the glasses area, enabling precise inpainting and preserving the integrity of other facial features. Additionally, an innovative eyes position accuracy mechanism is incorporated, utilizing landmark information to ensure the generated eyes align perfectly with their original positions. This approach not only addresses the limitations of previous methods but also offers a more scalable and adaptable solution for glasses removal in diverse real-world scenarios [18].

In this study, we utilize the aid of segment anything model (SAM) [19] that is solely trained on recognizing glasses in segmenting and masking the glasses. We then combine the use of diffusion model and a conditioning module in order to only conduct the inpainting in the masked glasses area to remove the glasses from the original images. At this stage, after the inpainting area has been confirmed by the conditioning module, the glasses are removed through image inpainting using the diffusion model. The ability to accurately and gracefully remove glasses from facial images addresses a common visual challenge. This technology has widespread applications in photography, personal image enhancement, and even facial recognition, where clear, glasses-free images are often preferred. Therefore, in order to preserve the individual style and facial features, we also corroborated an additional modification towards the proposed method, which is the eyes position accuracy [20]. Through the given dataset attributes, the attributes responsible for the eyes' positions is also inserted into the model, which provides the necessary information for the model to generate the supposed eyes in the correct position.

The proposed model, remove glasses diffusion model, stands at the forefront of eyeglasses removal techniques. Our primary goal is to significantly enhance face recognition accuracy by systematically removing eyeglasses from facial images. The main contributions of our model include:

- Our proposed model leverages the DDPM as its foundational framework, ensuring effective denoising through a diffusion process tailored for intricate data distributions.
- The incorporation of a conditional network employing binary masks to allow for a meticulous and accurate identification of eyeglasses-occluded regions during the diffusion process.
- The model features a robust eye precision accuracy, leveraged from the CelebA dataset's intrinsic landmark attributes, in order to ensure the precise positioning of the generated eyes, maintaining them in the exact same position as in the original image.

The subsequent sections of this paper will elaborate on the RGDM approach, beginning with a discussion of related work and the theoretical foundation of diffusion models in section 2 and 3. The methodology in section 4 will detail the RGDM framework, including the conditioning module, the eyes position accuracy mechanism, dataset details, implementation specifics, and evaluation metrics. Section 5 will present the experimental results of our model. The discussion in section 6 will compare RGDM against state-of-the-art methods, demonstrating its superior performance in glasses removal. Finally, the paper will conclude on the implications of these findings for future research and practical applications in section 7.

2. RELATED WORK

2.1. Diffusion model

Since the publication of diffusion model, the utilization of said method in image processing, particularly image generation or image inpainting, has been steadily rising. Not few publications have also been published with the focus of comparing image generation based on diffusion model to image generation based on previous known models, such as GAN. Focusing on publications released in recent years, Dhariwal and Nichol [21] compared diffusion model to current state-of-the-art GAN models in image synthesis. They conducted the experimentations on both unconditional and conditional image synthesis, where diffusion model proves that the image synthesis it generates achieve superior image sample quality compared to the GANs model. In 2022, Lugmayr *et al.* [22] proposed RePaint as an inpainting approach based on DDPM that can be used with masks. The authors only alter the iterations of the reverse diffusion and does not conduct any modifications or conditionings on the original DDPM model, which proves to generate diverse and high-quality images as a result and outperforms state-of-the-art GAN and autoregressive models. In 2023, Xie *et al.* [23] proposed SmartBrush for object inpainting guided by text and shape based on diffusion model. The authors also proposed a novel object-mask prediction with the goal of better preserving the image's background.

2.2. Image inpainting

Image inpainting is the process of finishing or restoring a missing portion of an image, or eliminating certain elements of an image that were initially inserted into it [24]. The traditional methods of image inpainting are categorized into three groups: diffusion-based techniques, patch-based techniques, and convolution filter-based techniques, while the modern approach relies more upon deep learning-based methods [25]. Thanks to the advancement of image processing tools and the flexibility of digital image editing, automatic image inpainting has found significant applications in computer vision and has grown to be a significant and difficult area of research in image processing [26], [27].

Focusing on the usage of image inpainting for glasses removal, the first method was proposed by Park *et al.* [28] in 2005 using recursive error compensation. This method begins by identifying the areas that the glasses have obscured before creating a face image devoid of glasses with the help of recursive error compensation using principal component analysis (PCA) reconstruction. The resultant image is devoid of any traces of the glasses' frame, reflection, or shading, suggesting that the method successfully addresses the issue of glasses occlusion although there are still some limitations of removing glasses with dark frames and gradated lens colors. Five years later, Wang *et al.* [29] proposed a method to remove glasses from face images based on the active appearance model (AAM) [30]. The model is built by merging shape and appearance variants in a shape-normalized fashion. The results of the experiments reveal that the suggested method is an effective solution for recognizing faces obscured by thick-rimmed eyeglasses, which was a notable problem with [28] method. Another five years after, Liang *et al.* [31] proposed a glasses detection model that was based on Zhu and Ramanan [32] concept of tree-pictorial structure. A double-layered filter made of inpainting and deep learning is used to remove the glasses. After conducting a comprehensive investigation, the authors discovered a drawback of their approach. Given that the edge information on glasses without rims is too "weak" to identify them as glasses, this affects the detection rate for such individuals.

In recent years, a novel image-to-image GAN [33] framework for eyeglasses removal, called ByeGlassesGAN, is proposed by Lee and Lai in 2020 [34]. The components of ByeGlassesGAN are an encoder, a segmentation decoder, and a face decoder. The experiments demonstrate that even for semi-transparent color eyeglasses or spectacles with glare, ByeGlassesGAN may produce aesthetically pleasing results in the facial images with the glasses removed. Then, Hu *et al.* [35] proposed the eyeglasses removal generative adversarial network (ERGAN), a unified eyeglass removal model, in 2021. This method suggested a GAN-based architecture to remove various kinds of glasses in the wild. The proposed model learns to swap the eye region in two faces when given two images of faces wearing and not wearing eyeglasses. That same year, Cheng and Cao [36] proposed another method using the same acronym, called ERGAN (high perform GAN for eyeglasses removal). This method is based on the successful implementation of the face attribute editing task in the GAN. It introduces the more sophisticated GAN inversion model IDInvert [37] and the InterfaceGAN [38] method. The findings of the experiment indicate that this method helps to increase the removal accuracy of glasses. However, it is also clear from the experimental findings that this method requires improvement in terms of how real-world images are processed, as it falls short in several aspects.

Another approach proposed that same year is the multimodal asymmetric dual learning framework by Lin *et al.* [39], which utilizes unsupervised learning for eyeglasses removal. While innovative, this method falls short in terms of precision and quality compared to supervised learning approaches. In 2022, Lyu *et al.* [40] proposed a method using 3D synthetic data to remove eyeglasses and shadows, which provided a foundation for further research. However, this approach is limited by its reliance on synthetic data, which may not fully capture the variability of real-world images. Coccia [41] discussed in 2018 the classification of innovations based on their interaction with existing technologies. Our RGDM model exemplifies an incremental innovation that enhances the interaction between established methods, such as GANs and diffusion models, offering a more precise and adaptable solution for glasses removal in facial images.

From the above passage, we can conclude that current existing glasses removal methods, mainly those that are proposed in recent years, rely on conventional GAN-based frameworks. Our proposed model, RGDM, departs from the conventional use of a GAN [33] by utilizing a DDPM [42], [43] instead. As diffusion models have recently shown state-of-the-art results [43]-[45] while also outperforming GANs in terms of accuracy and variety [21], [46], our method can automatically remove glasses from facial images in a more precise manner.

3. BACKGROUND: DENOISING DIFFUSION PROBABILISTIC MODEL

In this paper, we use denoising diffusion probabilistic model as the base model for our proposed method. DDPM [42], [43], which is shown in Figure 1, is shortly known as diffusion model. It is a model inspired by nonequilibrium thermodynamics [47], which is a class of latent variable models with the form:

$$p_\theta(x_0) := \int p_\theta(x_{0:T}) dx_{1:T}) \qquad (1)$$

Its basis is a parameterized Markov chain trained with variational inference to generate samples that match the data in a finite amount of time.

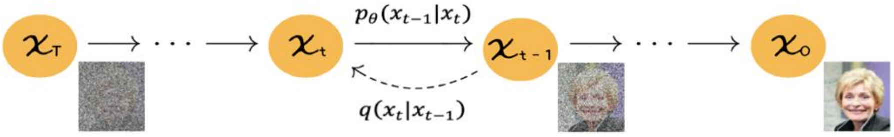

Figure 1. Overview of the denoising diffusion probabilistic model

It starts by sampling from a distribution with noise x_T, gradually producing less-noisy samples x_{T-1}, x_{T-2}, ..., until a final sample x_0 is reached. The reverse process is called the joint distribution $p_\theta(x_{0:T})$, which is a Markov chain with learned gaussian transitions. The forward process, also known as the diffusion process, is the approximate posterior $q(x_{1:T}|x_0)$ that a Markov chain is fixed to, adding gaussian noise to the data gradually based on a variance schedule $\beta_1, \dots, \beta_T$. By optimizing the standard variational bound on negative log likelihood, training can then be carried out.

For a noisy sample x_t, its noise component can be predicted using a function $\epsilon_\theta(x_t, t)$. Given a data sample x_0, we can add noise ϵ to the image by utilizing the forward diffusion Markov process over a scheduled variance of β_t and timestep t. They are randomly drawn together to create a noised sample x_t, which generates every sample in a minibatch to train this function. The forward process may be applied as:

$$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1})) \qquad (2)$$

$$q(x_t|x_{t-1}) := \mathcal{N}\left(\sqrt{1-\beta_t}x_t, \beta_t I\right)) \qquad (3)$$

where T is the total number of steps. The forward process has a notable property of admitting, in closed form, of sampling x_t at timestep t. Hence, x_t may be expressed in a closed form as:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon) \qquad (4)$$

where $\alpha_t := 1 - \beta_t$, $\bar{\alpha}_t := \prod_{j=1}^{t} \alpha_j$, and $\epsilon \sim \mathcal{N}(0, \mathrm{I})$.

In order to create images from random noise, the diffusion process above needs to be inverted. When β_t is small enough, this can be done by learning the gaussian $q(x_{t-1}|x_t)$. Nevertheless, due to the true distribution of x_0 being unable to access the gaussian, $q(x_{t-1}|x_t)$ is not known. Therefore, a neural network p_θ is then trained to estimate the conditional distribution:

$$p_\theta(x_{t-1}|x_t) := \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t))) \qquad (5)$$

where, taken from (4), $x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}\epsilon_t}\right)$ is obtained through training μ_θ to predict it.

The objective of training, considering a simple mean-squared error loss between the actual noise and the anticipated noise as an example, will be $||\epsilon_\theta(x_t, t) - \epsilon||^2$. By using stochastic gradient descent to optimize random terms of L, training is possible to be completed in an efficient manner. In $p_\theta(x_{t-1}|x_t)$, given a gaussian transition $\mathcal{N}(x_{t-1}; \mu_\theta(x_t,t), \sum_\theta(x_t,t))$, the denoising distribution can be modelled with the mean $\mu_\theta(x_t,t)$ computed as a part of $\epsilon_\theta(x_t,t)$. This leads to the following parametrization:

$$\mu_\theta(x_t,t) := \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t,t)\right) \qquad (6)$$

A simplified training objective is discovered by Ho *et al.* [42] to be more beneficial:

$$\mathcal{L} = E_{t,x_0,\epsilon}[\|\epsilon - \epsilon_\theta(x_t,t)\|^2]) \qquad (7)$$

This training objective is obtained by predicting the cumulative noise ϵ_θ that is added to the intermediate image x_t. In short, it removes the need for the weighting in the Langevin dynamics, transforming it into a weighted variational bound which leads to down-weighting the loss terms of the simplified objective corresponding to small t.

4. METHOD

RGDM combines the principles of DDPM with innovative modifications to achieve automatic glasses removal. We first present our approach for conditioning the DDPM on glasses removal. This will be followed by an introduction on an approach to improve the accuracy of the in-painted eyes' position after the removal of glasses.

The model operates through a series of carefully orchestrated steps that integrate conditioning modules, diffusion processes, and landmark-based positioning to achieve high-quality, realistic results. The process begins with an input facial image that contains glasses. This image serves as the starting point for the glasses removal procedure. A modification of the segmentation model, the SAM [19], is applied to the input image. This segmentation model is specifically trained to detect and isolate the glasses from the rest of the facial features, resulting in a binary mask that highlights the glasses region.

Figure 2 illustrates the flow of model explained. The binary mask generated in the previous step is crucial for the conditioning process. It identifies the exact regions of the image where the glasses are located, ensuring that these areas receive focused attention during the subsequent denoising process. In parallel, the model extracts and incorporates landmark data from the input image. This data includes precise coordinates for the eyes and other key facial features, which are essential for maintaining the correct positioning of the eyes during the inpainting process.

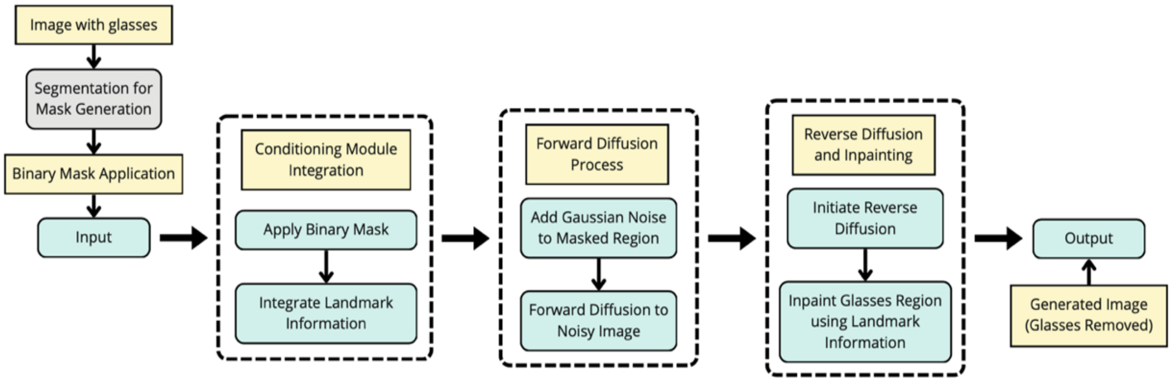
Image with glasses
Segmentation for Mask Generation
Binary Mask Application
Input
Conditioning Module Integration
Apply Binary Mask
Integrate Landmark Information
Forward Diffusion Process
Add Gaussian Noise to Masked Region
Forward Diffusion to Noisy Image
Reverse Diffusion and Inpainting
Initiate Reverse Diffusion
Inpaint Glasses Region using Landmark Information
Output
Generated Image (Glasses Removed)

Figure 2. The flow of the RGDM model

To initiate the diffusion process, gaussian noise is strategically injected into the glasses region of the image, guided by the binary mask. The noise levels are controlled by a variance schedule, ensuring that the noise is gradually added to the identified region while preserving the structure of the surrounding unmasked areas. The model then performs the forward diffusion process, where the noisy image undergoes a series of transformations. Each step of the Markov chain in the diffusion process incrementally adds more noise to the identified region, leading to a progressively noisier version of the original image.

Once the forward diffusion process reaches its peak noise level, the reverse diffusion process begins. The model, conditioned by both the binary mask and the landmark data, systematically removes the noise from the image, effectively reconstructing the pixels in the glasses region. During this reverse process, the model focuses on inpainting the glasses region. The binary mask ensures that the diffusion process is concentrated on the areas that originally contained the glasses, while the landmark data guides the accurate placement of the eyes and other facial features.

The output from each step of the reverse diffusion is gradually refined, with the model filling in the glasses region with plausible facial details. The reverse diffusion process continues until the model generates a final, noise-free image. This output image has the glasses completely removed, with the eyes and other facial features accurately reconstructed and aligned with their original positions.

4.1. Conditioning module

The conditioning module in RGDM is a crucial novel addition, for incorporating information about glasses conditions into the denoising process. It aims to integrate information about the presence or absence of glasses g into the denoising process, allowing the model to adapt its denoising strategy based on this condition. The model’s glasses removal method relies on predicting the missing pixel values of the facial image. This is done by the help of the conditioning module by masking the glasses region. Figure 3 shows the overview of the RGDM approach.

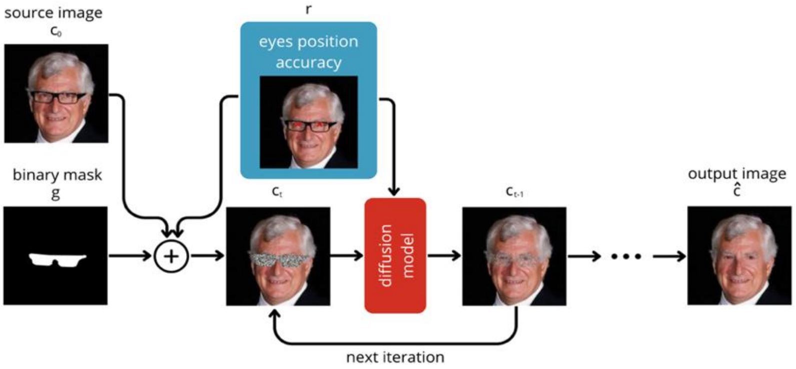

Figure 3. RGDM: glasses removal process overview

In the context of the proposed model, considering an input image denoted as 'c', a glasses mask as 'g' representing the areas of 'c' to be in-painted, and eyes' landmark attribute represented by 'r', the primary objective is to effectively in-paint the glasses mask and generate an output image denoted as '$\hat{c}$'. The generated regions are represented as '$g \odot \hat{c}$', while the known facial regions are expressed as '$(1-g) \odot c$'. Given that noise is exclusively introduced to the masked glasses region, where '$c = c_0$', the forward process at timestep 't' can be mathematically expressed as (8).

$$\hat{c}_t = \sqrt{1-\bar{\alpha}_t}\epsilon + \sqrt{\bar{\alpha}_t}x_0)$$
$$c_t = (1-g) \odot c_0 + g \odot \hat{c}_t \tag{8}$$

The input variables, c_t, g, and r, are fed into the conditioning module, where the module discerns the process to restore the masked glasses region, $c_0 \odot g$, while utilizing the clear and unmasked facial information. This ensures that the in-painted eyes within the masked glasses, g, harmonize with the overall facial image. In (9) encapsulates the training objective derived from this process.

$$\mathcal{L}_{CGR-DDPM} = E_{\epsilon \sim \mathcal{N}(0,I)}[||\epsilon - \epsilon_\theta(c_t, t, g, r)||^2]) \tag{9}$$

Random gaussian noise is generated across the masked area $c_T = (1-g) \odot c_0 + g \odot c_0$. Therefore, to obtain the generated output c_0, the diffusion process is reversed during the inference stage. With the reverse diffusion process successfully conditioned to focus on the glasses area, we successfully in-paint the glasses area with the necessary facial features and eyes, which in return removes the glasses from the facial images. This can be seen in Figure 3, where c_t, which is the source image c_0 combined with binary mask g, undergoes the diffusion model to produce c_{t-1} and so on, which will be iterated until output image $\hat{c}$ is finally produced where the eyeglass is successfully removed.

4.2. Eyes precision accuracy

To ensure the generation of accurate and fair glasses masks, our methodology incorporates a meticulously crafted approach leveraging a modified SAM [19]. This adaptation of SAM is purposefully trained to adeptly detect and segment eyeglasses within facial images, thus enhancing the precision of the generated glasses' masks. The process initiates with inputting the source images into the modified SAM, strategically tailored to focus solely on the segmentation of eyeglasses. The outcome of this segmentation step yields the glasses' masks denoted as 'g' as visually represented in Figure 3. These masks become a pivotal component in the subsequent stages of our methodology, seamlessly integrating into the processes outlined in Section A for further analysis and glasses removal. This nuanced approach ensures the reliability and accuracy of the masks, laying a robust foundation for subsequent phases of the remove glasses diffusion model.

Despite its proficiency in generating an accurate mask for different types of glasses, we found that the segmentation model tends to mask the entire glasses, including their lenses, irrespective of whether they resemble sunglasses with colored lenses or conventional eyeglasses with clear lenses. This is evident in the binary mask 'g' and the masked image 'c_t' portrayed in Figure 3. This comprehensive masking leads to the possibility of the generated eyes being positioned differently than in the original image, in particular concerning the types of glasses that are designed for vision correction (those with clear lenses). To address this potential misalignment and ensure the accurate positioning of the generated eyes, the model is also equipped with information concerning the eyes' landmarks denoted as 'r', which is leveraged from the dataset's intrinsic landmark attributes:

$$r = \left(lefteye_x, lefteye_y \middle| righteye_x, righteye_y\right)) \qquad (10)$$

Through the incorporation of eyes' landmarks denoted as 'r', the model ensures the precise positioning of the generated eyes, maintaining them in the exact same position as in the original image. This strategic use of landmark information acts as a safeguard against potential distortions that could arise during the glasses removal process. Figure 4 provides a visual representation of the impact of including eyes' landmarks, showcasing samples of generated eyes both with and without the utilization of landmark information. This not only exemplifies the effectiveness of the proposed approach in preserving the original eye position but also serves as a testament to the model's ability to produce realistic and accurate results in the absence of glasses. The pseudocode for the glasses removal process can be seen in Algorithm 1.

Algorithm 1. Glasses removal
```
$c_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for $t = T, \dots, 1$ do
    for $u = 1, \dots, U$ do
        $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\epsilon = 0$
        $\hat{c}_t = \sqrt{1 - \bar{\alpha}_t}\epsilon + \sqrt{\bar{\alpha}_t} x_0$
        $z \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $z = 0$
        $r = \left(lefteye_x, lefteye_y \middle| righteye_x, righteye_y\right)$
        $c_t = (1 - g) \odot c_0 + g \odot \hat{c}_t$
    end for
end for
return $c_0$
```

4.3. Experimental setup

4.3.1. Dataset

We conducted our experiments using the CelebA dataset [48], which is a face attribute dataset derived from labelling images selected from the challenging face datasets, CelebFaces. It features 10,177 identities, each with around 20 images, for a total of 202,599 celebrity images. It has each image annotated with forty facial features and five key characteristics, including the presence or absence of glasses and the landmarks of both the left and right eyes. From the annotations, we separated the CelebA dataset into two subsets, one where the images consist of glasses and the other where the images have no glasses. In total, there are 189,406 non-glasses images and 13,193 images wearing glasses. All of the facial images are cropped to 178×178, before being resized to 256×256 each.

4.3.2. Implementation details

The architecture of the remove glasses diffusion model is rooted in the DDPM framework. It is composed of a generator responsible for the denoising process and a conditioning module that employs a binary mask to identify the glasses in the input images. We have implemented the RGDM method using PyTorch 2.0.0, Python 3.8 (Ubuntu20.04), and Cuda 11.8, running on an RTX 3090 (24 GB) GPU. The model undergoes training on the dataset for 300,000 iterations, with T set to 250.

RGDM is based on the DDPM, which is a type of probabilistic generative model that reverses a diffusion process to generate high-quality images. The network is conditioned on a binary mask and landmark data, which guide the diffusion process to identify the glasses region in the input images. We have implemented the RGDM method using PyTorch 2.0.0, Python 3.8 (Ubuntu20.04), and Cuda 11.8, running on an RTX 3090 (24 GB) GPU.

The CelebA dataset was utilized for training and testing the RGDM model. This dataset contains over 200,000 celebrity images with a variety of attributes, including whether the person is wearing glasses. Since the CelebA dataset does not come with detailed segmentation masks, a modified version of the SAM [19] was employed to generate binary masks. These masks isolate the glasses region, helping the model focus on the correct area during the diffusion process. 70% of the dataset (around 140,000 images) is allocated for training the RGDM model. The training set includes images with and without glasses, ensuring the model learns to accurately remove glasses while preserving the underlying facial features. The remaining 30% of the dataset (around 60,000 images) is reserved for testing. This set is used to evaluate the model’s performance, particularly its ability to reconstruct the face without glasses while maintaining naturalism and realism.

The learning rate was set to 0.0001, which is typical for training diffusion models. This rate ensures that the model learns at a pace that balances convergence speed and stability. The model was trained for 300,000 iterations, which is a commonly used measure in diffusion models due to the high number of timesteps involved, with T set to 250. This extensive training allowed the model to learn the complex task of glasses removal and ensured that it generalized well to new images. The Adam optimizer was used for training, with beta1 set to 0.9 and beta2 set to 0.999. Adam was chosen for its robustness and ability to handle the sparse gradients that are often encountered in deep generative models.

The input to the RGDM model is a facial image with glasses, along with a corresponding binary mask that isolates the glasses region. The image is preprocessed to match the input requirements of the model, typically resized to 256×256 pixels and normalized. The final image, with the glasses removed, is generated through the reverse diffusion process. Starting from the noisy image created during the forward process, the model gradually reduces the noise, guided by the conditioning module, until a high-quality, glasses-removed image is produced. The output is a 256×256 pixel image with the glasses removed. The model’s goal is to generate a realistic image where the glasses have been in painted with the underlying facial features, particularly focusing on maintaining the correct placement of the eyes and ensuring that the output is free of artifacts.

4.3.3. Evaluation metrics

We conducted the evaluation of image quality through the utilization of the Fréchet inception distance (FID) [49], [50]. FID serves as a metric for quantifying the dissimilarity between real images and generated images. This calculation is performed using the inception network to extract features and assess the distance between the distributions of real and generated images. FID has established itself as a widely adopted evaluation metric in various image generation tasks [22], [31], [51].

An additional metric employed for assessing the model is the learned perceptual image patch similarity (LPIPS) [52]. LPIPS measures the learned distance within AlexNet [53] deep feature space to evaluate the distance between image patches. A higher LPIPS value indicates greater dissimilarity, suggesting a more substantial difference between image patches. Conversely, a lower LPIPS value signifies increased similarity, indicating a closer match between image patches. This metric operates on the premise that a larger LPIPS distance implies more pronounced distinctions, while a smaller distance signifies heightened visual resemblance between image patches.

5. EXPERIMENTS AND RESULTS

We introduced the extensive experiments conducted on our proposed RGDM model in this chapter. The dataset undergoes a preliminary phase where it is subjected to a modified segmentation model, SAM [19], resulting in the generation of precise glasses binary masks for each image. Subsequently, both the binary masks and the original images are fed into the conditioning module, as expounded in the preceding section. This module orchestrates the combination and preparation of images into masked versions, with the judicious addition of gaussian noise targeted at the masked areas. This preparatory step sets the stage for the subsequent denoising procedure. The combined data, now enriched with gaussian noise, is then processed through the diffusion model, kickstarting the in-painting process to generate versions of the facial images with removed glasses. Crucially, the proposed method showcases its versatility by not only effectively removing clear-lensed glasses but also demonstrating proficiency in handling dark-lensed varieties, including sunglasses. This adaptability underscores the robustness of the RGDM method across a spectrum of eyeglass types, enhancing its applicability in real-world scenarios.

Figure 4 shows the overview process of how the experiment is conducted in order to obtain the necessary results. As can be seen in the figure, the segmentation process provides the necessary glasses binary mask, which is then passed through the conditioning module along with the original image, in order to produce the necessary condition for the DDPM to remove the glasses. The output of the conditioning module is a combination of the original image and the binary mask which has been added with gaussian noise in the conditioning area. Eyes position accuracy, in this case $r = (68, 112|109, 112)$, along with the output from the conditioning module is then passed through the denoising process which finally generates the targeted output image. The result shows that the proposed RGDM method successfully removed the glasses from the original image.

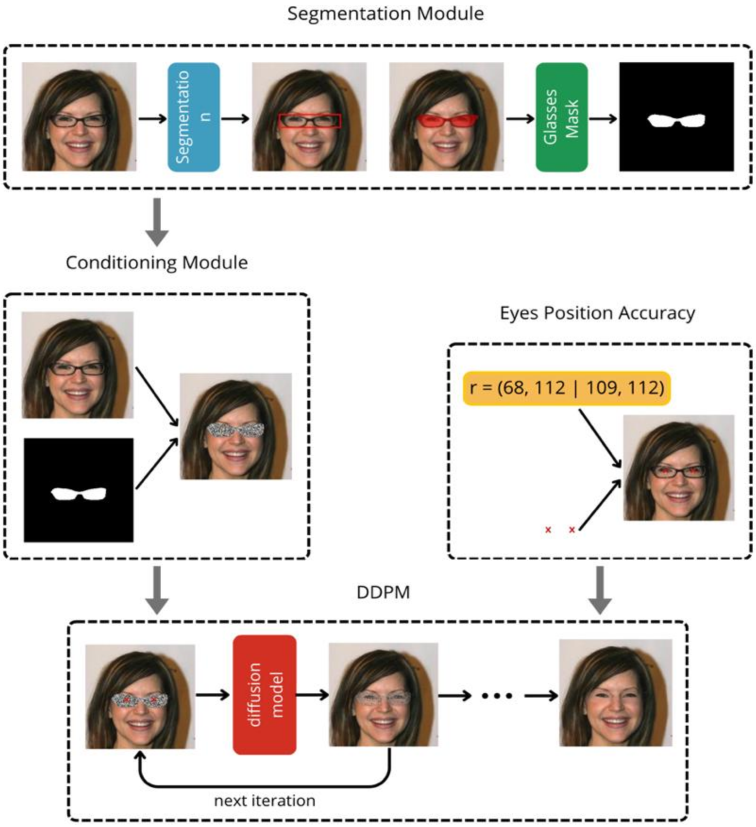

Figure 4. RGDM: glasses removal pipeline

In this detailed analysis, we meticulously assess the performance impact of the introduced eyes position accuracy in the proposed eyeglass removal framework. As elucidated in Figure 5, the implementation of eyes position accuracy plays a pivotal role in ensuring the generated eyes align seamlessly with their intended positions. Particularly, for clear-lensed glasses, the accuracy mechanism guarantees that the generated eyes precisely occupy their original locations. This meticulous alignment becomes equally crucial for dark-lensed glasses or sunglasses, where the generated eyes maintain normal distances and proportions consistent with the average face. In stark contrast, when the model operates without the eyes position accuracy feature, the generated eyes exhibit irregular placements, appearing higher than the original position and uncomfortably close to the person's eyebrows. This divergence results in unnatural and occasionally unsettling images. Conversely, when the model incorporates the eyes position accuracy, the generated eyes retain their authenticity in the original position, culminating in facial images that exude a heightened sense of naturalism and realism. This nuanced evaluation underscores the indispensable role of eyes position accuracy in enhancing the visual fidelity and believability of the generated images, marking it as a critical component in the overall success of the proposed eyeglass removal model.

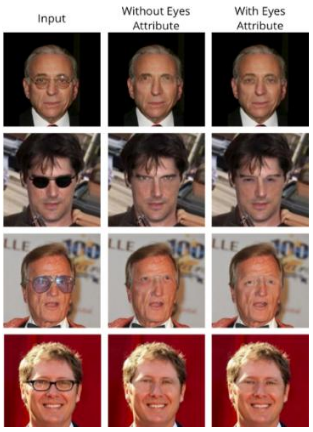

Figure 5. Ablation study. Comparison on glasses removal results with and without eyes position accuracy, on the CelebA dataset

6. DISCUSSION

We also compared our model to other current state-of-the-art models. The proposed RGDM method undergoes a rigorous comparative analysis against contemporary state-of-the-art glasses removal techniques, including CycleGAN [54], ERGAN [35], and 3D Synthetic [40]. To ensure a fair and objective comparison, each method is meticulously trained under identical experimental conditions, mirroring the circumstances employed in training the RGDM model. This meticulous approach to experimentation serves to elucidate the relative strengths and weaknesses of each method, allowing for a comprehensive assessment of the RGDM model's performance in comparison to its counterparts.

6.1. Qualitative evaluations

In a meticulous qualitative assessment, we conducted a comparative assessment of our model against prominent generative methods mentioned above. Figure 6 illustrates the evaluation of the quality of generated glasses-removed images on the CelebA dataset. Through leveraging the open-source codes, the state-of-the-art methods—CycleGAN [54], ERGAN [35], and 3D Synthetic [40]—are reimplemented for benchmarking. As depicted in Figure 6, compared to our approach, CycleGAN, ERGAN, and 3D Synthetic all have their own limitations in removing the glasses from the facial images. Both CycleGAN and 3D Synthetic exhibit limitations, failing to completely eliminate glasses frames and leaving discernible evidence of their presence in the images. Conversely, ERGAN falls short in delivering realistic and clear images while attempting removal, occasionally leaving subtle traces of glasses existence, albeit less prominent than the former two methods. In stark contrast, our approach yields results that are markedly true to life and authentic. The glasses frames are impeccably removed without leaving any traces or anomalous inpainting artifacts. Furthermore, the generated eyes faithfully replicate the original eyes for glasses with clear lenses. For glasses with coloured or dark lenses, the generated eyes harmonize seamlessly with the overall facial context, ensuring a natural and realistic appearance in the generated images. This pronounced superiority in realism and precision underscores the efficacy of the proposed RGDM model in achieving high-fidelity glasses removal.

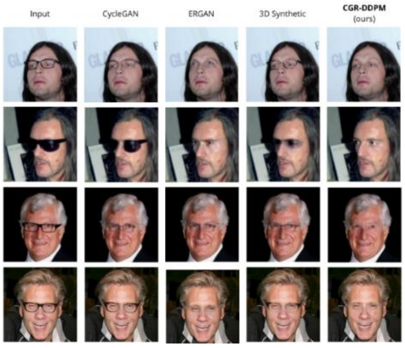

Figure 6. Qualitative results. Comparison against other state-of-the-art methods for glasses removal on the CelebA dataset

6.2. Quantitative evaluations

In the realm of quantitative assessments, we present a comprehensive evaluation of the RGDM method, leveraging the FID and LPIPS metrics to gauge the naturalism and realism of the generated images. The findings, as elucidated in Table 1, underscore the prowess of RGDM by securing the lowest FID and LPIPS values on the CelebA dataset. With an FID value of 27.09 and a LPIPS value of 0.299, RGDM outperforms its counterparts, namely 3D synthetic (with an FID value of 32.96 and a LPIPS value of 0.467), CycleGAN (with an FID value of 40.83), and ERGAN (with an FID value of 41.02 and a LPIPS value of 0.302), positioning itself as the premier method for glasses removal. Notably, the close distribution of generated images to the original images, as indicated by the low FID value, underscores the superior performance of the RGDM method. This not only substantiates its eminence in the domain of glasses removal but also establishes a compelling case for the adoption of diffusion models in image processing, particularly in the nuanced realm of image inpainting. The demonstrated superiority in glasses removal capabilities further bolsters the broader applicability of diffusion models in advancing image manipulation techniques.

Table 1. Quantitative results on different methods

Methods	FID↓	LPIPS↓
CycleGAN [54]	40.83	-
ERGAN [35]	41.02	0.302
3D synthetic [40]	32.96	0.467
RGDM	**27.09**	**0.299**

7. CONCLUSION

This study represents a significant advancement in the field of facial image processing, particularly in the challenging task of glasses removal. The introduction of the RGDM offers a novel solution that departs from traditional GAN-based methods by leveraging the capabilities of a DDPM. This approach, enhanced by a binary mask conditioning mechanism, allows for precise inpainting and ensures that facial features, particularly the eyes, are meticulously preserved and accurately repositioned. In conclusion, RGDM not only advances the state of glasses removal technology but also sets a foundation for future innovations in image processing. By addressing the limitations and exploring new frontiers, this model has the potential to significantly impact both academic research and practical applications in the broader field of artificial intelligence and image enhancement.

ACKNOWLEDGEMENTS

The first special thanks go to Universitas Mercu Buana for supporting foreign collaborative research, the second to Universitas Muhammadiyah Tangerang. Third to Beijing Institute of Technology, for give facility for this research. There will always be a paper collaboration with Universitas Muhammadiyah Tangerang and Beijing Institute of Technology in further research.

ACKNOWLEDGMENTS

The authors would like to thank Multimedia Nusantara University for its full support, facilities, and conducive academic environment during the research process and preparation of this article.

FUNDING INFORMATION

This research was supported by Multimedia Nusantara University through the Independent Research program. Multimedia Nusantara University did not play a role in the design of the research, data collection, data analysis, interpretation of results, or writing of the manuscript.

AUTHOR CONTRIBUTIONS STATEMENT

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Yuliza		✓		✓	✓					✓		✓		✓
Rachmat Muwardi	✓	✓		✓	✓			✓		✓		✓		
Galatia Erica Yehezkiel	✓	✓	✓				✓	✓	✓					
Mirna Yunita		✓	✓	✓						✓		✓		
Lenni		✓		✓						✓				

C	:	**C**onceptualization	I	:	**I**nvestigation	Vi	:	**Vi**sualization
M	:	**M**ethodology	R	:	**R**esources	Su	:	**Su**pervision
So	:	**So**ftware	D	:	**D**ata Curation	P	:	**P**roject administration
Va	:	**Va**lidation	O	:	Writing - **O**riginal Draft	Fu	:	**Fu**nding acquisition
Fo	:	**Fo**rmal analysis	E	:	Writing - Review & **E**diting			

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1] N. Imoh, N. R. Vajjhala and S. Rakshit, “Experimental face recognition using applied deep learning approaches to find missing persons,” in *Proceedings of Second International Conference on Advances in Computer Engineering and Communication Systems*, Singapore, 2022, pp 131–140, doi: 10.1007/978-981-16-7389-4_13.

[2] M. Andrejevic and N. Selwyn, "Facial Recognition," United Kingdom: Polity Press, 2022.

[3] Q. Jiayu, “Occluded Face Recognition with Deep Learning,” in *Computing and Data Science: CONF-CDS 2021*, Singapore, 2021, vol. 1513, pp. 28–35, doi: 10.1007/978-981-16-8885-0_3.

[4] L. Liang, “Face Recognition technology analysis based on deep learning algorithm,” *Journal of Physics: Conference Series*, 2020, vol. 1544, pp. 12158, doi: 10.1088/1742-6596/1544/1/012158.

[5] D. Yeung, R. Balebako, C. I. G. Gaviria and M. Chaykowsky, "Face recognition technologies: designing systems that protect privacy and prevent bias," Homeland Security Operational Analysis Center operated by the RAND Corporation, 2020.

[6] I. Berle, "Face recognition technology: Compulsory visibility and its impact on privacy and the confidentiality of personal identifiable images," Springer Cham, 2020.

[7] G. Guo and N. Zhang, “A survey on deep learning based face recognition,” C*omputer Vision and Image Understanding*, 2019, vol. 189, doi: 10.1016/j.cviu.2019.102805.

[8] H. Du, H. Shi, D. Zeng, X. P. Zhang and T. Mei, “The elements of end-to-end deep face recognition: a survey of recent advances,” *ACM Computing Surveys*, 2022, vol. 54, no 10s, p. 1–42, doi: 10.1145/3507902.

[9] M. Wang and W. Deng, “Deep face recognition: A survey,” *Neurocomputing*, vol. 429, pp. 215-244, 2021, doi: 10.1016/j.neucom.2020.10.081.

[10] K. Zhang, Q. Zhu and W. Li, “Experimental research on occlusion face detection method based on attention mechanism,” *Journal of Physics: Conference Series*, vol. 2258, no 1, p. 12078, 2022, doi: 10.1088/1742-6596/2258/1/012078.

[11] A. Abbad, O. Elharrouss, K. Abbad and H. Tairi, “Application of MEEMD in post-processing of dimensionality reduction methods for face recognition,” *IET Biometrics*, vol. 8, no 1, pp. 59-68, 2019, doi: 10.1049/iet-bmt.2018.5033.

[12] S. Su, M. Yang, L. He, X. Shao, Y. Zuo and Z. Qiang, “A survey of face image inpainting based on deep learning,” in *Cloud Computing: 11th EAI International Conference, CloudComp 2021*, Switzerland, 2022, p. 72–87, doi: 10.1007/978-3-030-99191-3_7.

[13] M. Coccia, “Sources of technological innovation: radical and incremental innovation problem-driven to support competitive advantage of firms,” *Technology Analysis & Strategic Management*, vol. 29, no 9, 2017, doi: 10.1080/09537325.2016.1268682.

[14] M. Coccia, “Theories of the evolution of technology based on processes of competitive substitution and multi-mode interaction between technologies,” *Journal of Economics Bibliography*, vol. 6, no 2, pp. 99-109, 2019, doi: 10.1453/jeb.v6i2.1889.

[15] M. Coccia, “Comparative Theories of the Evolution of Technology,” *Global Encyclopedia of Public Administration, Public Policy, and Governance*, pp. 1-8, 2019, doi: 10.1007/978-3-319-31816-5_3841-1.

[16] M. Sadik-E-Tawheed, T. Ahmed, M. Bhuiyan, A. Hossain and S. Shatabda, “De-Glared: Eyeglasses glare and reflection removal using deep neural networks,” *2023 26th International Conference on Computer and Information Technology (ICCIT)*, Cox's Bazar, Bangladesh, 2023, pp. 1-6, doi: 10.1109/ICCIT60459.2023.10441194.

[17] L. Mao, Y. Xue, Y. Wei and T. Zhu, “An eyeglasses removal method for fine-grained face recognition,” *Journal of Electronics & Information Technology*, 2021, vol. 43, no 5, pp. 1448-1456, doi: 10.11999/JEIT200176.

[18] S. Budiyanto *at al.*, “The automatic and manual railroad door systems based on IoT,” *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 21, no. 3, pp. 1847-1855, 2021. doi: 10.11591/ijeecs.v21.i3.pp1847-1855.

[19] A. Kirillov *et al*., “Segment Anything,” *arXiv:2304.02643*, 2023, doi: 10.48550/arXiv.2304.02643.

[20] Z. Iklima, B. N. Rohman, R. Muwardi, A. Khan and Z. Arifiansyah, “Defect classification of radius shaping in the tire curing process using Fine-Tuned Deep Neural Network,” *SINERGI*, vol. 26, no. 3, pp. 335-342, 2022. doi: 10.22441/sinergi.2022.3.009.

[21] P. Dhariwal and A. Nichol, “Diffusion models beat GANs on image synthesis,” in *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21)*, 2024, vol. 672, doi: 10.5555/3540261.3540933.

[22] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte and L. V. Gool, “RePaint: inpainting using denoising diffusion probabilistic models,” *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022, pp. 11451-11461, doi: 10.1109/CVPR52688.2022.01117.

[23] S. Xie, Z. Zhang, Z. Lin, T. Hinz and K. Zhang, "SmartBrush: text and shape guided object inpainting with diffusion model," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, BC, Canada, 2023, pp. 22428-22437, doi: 10.1109/CVPR52729.2023.02148.

[24] O. Elharrouss, N. Almaadeed, S. Al-Maadeed and Y. Akbari, "Image inpainting: A review," *Neural Processing Letters*, vol. 51, pp. 2007-2028, 2020, doi: 10.1007/s11063-019-10163-0.

[25] D. J. B. Rojas, B. J. T. Fernandes and S. M. M. Fernandes, "A Review on image inpainting techniques and datasets," in *33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, Porto de Galinhas, Brazil, 2020, pp. 240-247, doi: 10.1109/SIBGRAPI51738.2020.00040.

[26] R. Muwardi, H. Qin, H. Gao, H. U. Ghifarsyam, M. H. I. Hajar and M. Yunita, "Research and Design of Fast Special Human Face Recognition System," *2020 2nd International Conference on Broadband Communications, Wireless Sensors and Powering (BCWSP)*, Yogyakarta, Indonesia, 2020. doi: 10.1109/BCWSP50066.2020.9249452.

[27] R. Muwardi, H. Zhang, H. Gao, M. Yunita, Y. Wang and Yuliza, "Design of new traffic system YOLO-LIO: light-traffic intercept and observation," *2024 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET)*, Bandung, Indonesia, 2024. doi: 10.1109/ICRAMET62801.2024.10809042

[28] J. -S. Park, Y. H. Oh, S. C. Ahn and S. -W. Lee, "Glasses removal from facial image using recursive error compensation," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 5, 2005, doi: 10.1109/TPAMI.2005.103.

[29] Y. -K. Wang, J. -H. Jang, L. -W. Tsai and K. -C. Fan, "Improvement of face recognition by eyeglasses removal," in *2010 Sixth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Darmstadt, Germany, 2010, pp. 228-231, doi: 10.1109/IIHMSP.2010.64.

[30] T. F. Cootes, G. J. Edwards and C. J. Taylor, "Active appearance models," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2001, vol. 23, no. 6, pp. 681-685, doi: 10.1109/34.927467.

[31] A. Liang, C. S. N. Pathirage, C. Wang, W. Liu, L. Li and J. Duan, "Face recognition despite wearing glasses," in *2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, Adelaide, SA, Australia, 2015, pp. 1-8, doi: 10.1109/DICTA.2015.7371260.

[32] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," *2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012, pp. 2879-2886, doi: 10.1109/CVPR.2012.6248014.

[33] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, 2014, doi: 10.48550/arXiv.1406.2661.

[34] Y.-H. Lee and S.-H. Lai, "ByeGlassesGAN: Identity preserving eyeglasses removal for face images," in *Computer Vision – ECCV 2020*, 2020, vol. 12374, pp. 243–258,doi: 10.1007/978-3-030-58526-6_15.

[35] B. Hu, Z. Zheng, P. Liu, W. Yang and M. Ren, "Unsupervised eyeglasses removal in the wild," *IEEE Transactions on Cybernetics*, vol. 51, no 9, pp. 4373-4385, 2021, doi: 10.1109/TCYB.2020.2995496.

[36] M. Cheng and X. Cao, "ERGAN: High Perform GAN for Eyeglasses Removal," in *2021 16th International Conference on Intelligent Systems and Knowledge Engineering (ISKE)*, Chengdu, China, 2021, pp. 406-411, doi: 10.1109/ISKE54062.2021.9755402.

[37] J. Zhu, Y. Shen, D. Zhao and B. Zhou, "In-domain GAN inversion for real image editing," in *Computer Vision – arXiv:2004.00049*, 2020, doi: 10.48550/arXiv.2004.0004.

[38] Y. Shen, J. Gu, X. Tang and B. Zhou, "Interpreting the latent space of GANs for semantic face editing," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, 2020, doi: 10.48550/arXiv.1907.10786.

[39] Q. Lin, B. Yan and W. Tan, "Multimodal asymmetric dual learning for unsupervised eyeglasses removal," *MM '21: Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5092 - 5100, doi: 10.1145/3474085.3475559.

[40] J. Lyu, Z. Wang and F. Xu, "Portrait eyeglasses and shadow removal by leveraging 3D synthetic data," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, 2022, pp. 3419-3429, doi: 10.1109/CVPR52688.2022.00342.

[41] M. Coccia, "Classification of innovation considering technological interaction," *Journal of Economics Bibliography*, 2018, vol. 5, no 2, pp. 76-93, doi: 10.1453/jeb.v5i2.1650.

[42] J. Ho, A. Jain and P. Abbeel, "Denoising diffusion probabilistic models," in *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada, 2020, doi: 10.48550/arXiv.2006.11239.

[43] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi and T. Salimans, "Cascaded diffusion models for high fidelity image generation," *The Journal of Machine Learning Research*, vol. 23, no 1, p. 2249–2281, 2020, doi: 10.48550/arXiv.2106.15282.

[44] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet and M. Norouzi, "Image super-resolution via iterative refinement," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, vol. 45, no 4, pp. 4713-4726, doi: 10.1109/TPAMI.2022.3204461.

[45] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu and S. Ermon, "SDEdit: guided image synthesis and editing with stochastic differential equations," *International Conference on Learning Representations (ICLR)*, 2022, doi: 10.48550/arXiv.2108.01073.

[46] G. Müller-Franzes *et al*., "A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis," *Scientific Reports*, vol. 13, 2023, doi: 10.1038/s41598-023-39278-0.

[47] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, vol. 37, pp. 2256-2265, doi: 10.48550/arXiv.1503.03585.

[48] Z. Liu, P. Luo, X. Wang and X. Tang, "Deep learning face attributes in the wild," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 3730-3738, doi: 10.1109/ICCV.2015.425.

[49] M. Seitzer, "pytorch-fid: FID Score for PyTorch," August 2020. [Online]. Available: https://github.com/mseitzer/pytorch-fid.

[50] 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," *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 6629 - 6640.

[51] Q. Mao, H. -Y. Lee, H. -Y. Tseng, S. Ma and M. -H. Yang, "Mode seeking generative adversarial networks for diverse image synthesis," arXiv:1903.05628, 2019, doi: 10.48550/arXiv.1903.05628.

[52] R. Zhang , P. Isola, A. A. Efros, E. Shechtman and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in 2*018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 2018 pp. 586-59, doi: 10.1109/CVPR.2018.00068.

[53] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84 - 90, 2017, doi: 10.1145/306538.

[54] J. -Y. Zhu, T. Park, P. Isola and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 2242-2251, doi: 10.1109/ICCV.2017.244.

BIOGRAPHIES OF AUTHORS

Yuliza, S.T., M.T. is currently an assistant professor in Department of Electrical Engineering, Universitas Mercu Buana, Jakarta, Indonesia. She completed master from Universitas Mercu Buana, Jakarta, Indonesia. She can be contacted at email: yuliza@mercubuana.ac.id.

Rachmat Muwardi, B.Sc., S.T., M.Sc. is currently a lecturer in the Department of Electrical Engineering, Universitas Mercu Buana, Jakarta, Indonesia. He graduated from the Beijing Institute of Technology in 2020 with a Master's in electronic science and technology. Currently, he declared as a recipient of a China Scholarship Council (CSC) to continue his doctoral program at the Beijing Institute of Technology in September 2022, majoring in optical engineering. During his undergraduate, he received a double degree scholarship from Universitas Mercu Buana and Beijing Institute of Technology in electrical engineering and computer science. His research interest is object detection, target detection, and embedded system. He can be contacted at email: rachmat.muwardi@mercubuana.ac.id.

Galatia Erica Yehezkiel, B.Sc. is currently a Master's student in Beijing Institute of Technology, Beijing, China, courtesy of the China Scholarship Council (CSC). She received her undergraduate degree in the same university, majoring in computer science. Her areas of interest include image processing, machine learning, and internet of things (IoT). She can be contacted at email: galatia.erica@bit.edu.cn.

Mirna Yunita, S.Kom., M.Sc. received a master’s in computer science and technology from Beijing Institute of Technology, Beijing, China. Currently, as a Ph.D. student at the School of Computer Science and Technology, Beijing Institute of Technology, China. She is interested in related topics in machine learning, web development, data mining, and bioinformatics. She was a frontend and mobile application developer in a Logistics and Supply Chain Company in Jakarta, Indonesia. She can be contacted at email: mirnayunita@bit.edu.cn.

Dr. Lenni currently an assistant professor in the Department of Electrical Engineering, Muhammadiyah Tangerang University, Tangerang, Indonesia. She completed master from Trisakti University. She completed doctoral education at Institute Pertanian Bogor, Indonesia. She can be contacted at email: lenni@umt.ac.id.