

Lightweight Model for Occlusion Removal from Face Images

Sincy John* and Ajit Danti

Department of Computer Science and Engineering, Christ University, India

sincyjohn2@gmail.com; ajitdanti@yahoo.com

*Correspondence: sincyjohn2@gmail.com

Received: 28th December 2022; Accepted: 14th February 2024; Published: 1st April 2024

Abstract: In the realm of deep learning, the prevalence of models with large number of parameters poses a significant challenge for low computation device. Critical influence of model size, primarily governed by weight parameters in shaping the computational demands of the occlusion removal process. Recognizing the computational burdens associated with existing occlusion removal algorithms, characterized by their propensity for substantial computational resources and large model sizes, we advocate for a paradigm shift towards solutions conducive to low-computation environments. Existing occlusion riddance techniques typically demand substantial computational resources and storage capacity. To support real-time applications, it's imperative to deploy trained models on resource-constrained devices like handheld devices and internet of things (IoT) devices possess limited memory and computational capabilities. There arises a critical need to compress and accelerate these models for deployment on resource-constrained devices, without compromising significantly on model accuracy. Our study introduces a significant contribution in the form of a compressed model designed specifically for addressing occlusion in face images for low computation devices. We perform dynamic quantization technique by reducing the weights of the Pix2pix generator model. The trained model is then compressed, which significantly reduces its size and execution time. The proposed model, is lightweight, due to storage space requirement reduced drastically with significant improvement in the execution time. The performance of the proposed method has been compared with other state of the art methods in terms of PSNR and SSIM. Hence the proposed lightweight model is more suitable for the real time applications with less computational cost.

Keywords: *Dynamic Quantization; Generative adversarial network (GAN); Occlusion; Pix2pix*

1. Introduction

Use of facial recognition technology among the various biometric methods has increased in the age of digital information. Face being obscured by a mask has made facial recognition difficult, leading to stringent identification requirements. Occlusion is the term used to describe obstruction of the face image by external objects, such as glasses, a mask, a scarf, or hair. Occlusion causes the loss of details or features that are necessary for identifying an image. For acquiring the facial images from occluded images, a variety of occlusion removal methods, including machine learning [1] and GAN-based method [2], are available. The goal of this work was to develop an intelligent system that would reduce the model size while maintaining accuracy by using pix2pix to remove occlusion. Due to their promising performance in the unsupervised mode of operation, GAN algorithms were examined.

Deep Learning approaches, when compared to other state-of-the-art methods, have a significant impact on a variety of fields due to their higher accuracy and efficiency. Generative adversarial networks (GAN) are deep neural networks that automatically detect and learn the patterns in input data. Deep Learning is a subset of machine learning in which model building is the crux. In deep learning, a model is

created from experience, where the model is put through a series of tasks and learns from them. Deep learning is more of predictive modelling, and the complexity of learning increases with how deep the model is with the number of layers. As the number of layers' increases, delicate and detailed prediction are made based on learning [3]. This, in turn, is directly proportional to precision. When the number of layers' increases, the model size also increases.

Most state-of-the-art deep neural network approaches focus on accurate models with the increase in the number of layers, parameter nodes, and availability of graphics processing units (GPUs) with solid computing capabilities, making the model complex but flawless. As a result, the large model frequently becomes resource-intensive, time-consuming, and memory-intensive. Hence, model compression is required to integrate artificial intelligence into embedded systems without affecting the model's accuracy. Its application in low computation devices as a compressed model decreases execution time, making it suitable for real-time applications. Compressed models are more effective in scenarios where models are updated in real-time or in larger size than the initial model, and model inference adds deployment overhead. The model size can affect the deployment cost if the model is hosted on a server [4]. As the scale of the model grows, so does the cost of deployment.

To address this problem, this study aims to introduce a novel contribution of compressed model which can be utilized for removal of occlusion in face images. In this approach, dynamic quantization and a GAN is used to create a lightweight model. It works by quantizing the weights of the model to a lower precision. The model in our work was able to generate images without occlusion with minimal loss and lightly reducing execution time. This is important for applications where storage space and/or computational resources are limited. After compression, the model size was reduced considerably without compromising efficiency. The experimental analysis demonstrates that the suggested model outperforms current state-of-the-art approaches with a structural similarity (SSIM) of 0.895 and a peak to signal ratio (PSNR) of 27.18dB.

2. Literature Survey

Deep neural networks have received a lot of interest and have been the topic of extensive research and study. With many parameters, nodes, and layers, neural networks are becoming increasingly complicated. The output of the complexity has got both pros and cons. The advantage is that as the number of layers' increases, the features become more apparent, the model becomes more promising with the problem or task given, and the outcome becomes more precise. The drawback is that as the number of layers, nodes, and parameters rises, the model grows in size, requiring more computation, memory use, calculation time, and high computational hardware. Several approaches have been developed to stabilize model compression.

We provide an overview of recent literature examining occlusion removal and compression strategies utilizing diverse methodologies. This review encompasses several approaches for addressing occlusions in facial images, along with GAN-based techniques specifically designed for occlusion removal. Additionally, it delves into the multifaceted realm of compressing and enhancing Generative Adversarial Networks (GANs) to cater to a range of facial image processing tasks. Each article proposes a different technique for compressing GANs, such as iterative pruning, co-evolutionary techniques, self-growing and pruning methods, knowledge distillation, and quantization. These techniques vary in complexity, effectiveness, and suitability for different GAN architectures and applications.

Face completion is done with the aid of a deep generative network. A generator with two adversarial loss functions, two discriminators, and a parser network make up the GAN-based network [5]. After being initially masked with noise pixels on a randomly selected region, the input is then sent into an auto-encoder, which is the generator part of a GAN. Two discriminators are utilized to distinguish between genuine and false synthesized contents in the mask and the complete produced image. The model is able to generate realistic face completion results, which makes it a valuable tool for a variety of applications. The drawbacks of the work are that some misaligned faces are too complex for the model to handle. The addition of 3D data can help to relieve this. The spatial connections between nearby pixels are not adequately taken advantage of by the model.

Face image unmosaicing is done by an approach called UMGAN [6], which is an image-to-image translation method. A Generative Adversarial Network (GAN) based methodology using perceptual loss for generating a unmosaiced facial image with fine details from a mosaiced image. The authors suggest a potential new technique for extracting latent semantic structure from mosaiced images called the GAN approach to image unmosaicing. As performance measures, SSIM, L1 loss, and perception loss are employed. The CelebA and MIT-CBCL picture datasets were used by the authors to test their methodology and achieved a low level l1 loss and high level SSIM loss.

Both articles [5] and [6] focus on image restoration tasks using GAN-based methods. These methods leverage GANs and perceptual loss functions to generate realistic and detailed images from corrupted or incomplete inputs. Face completion and image unmosaicing have demonstrated efficacy but are hindered by limitations in handling complex image structures and spatial connections between nearby pixels. There is a need for novel approaches that can effectively address these challenges, potentially leveraging 3D data to enhance performance.

An image-to-image translation technique based on a Generative Adversarial Network, the microphone from the facial picture is eliminated by an approach called MRGAN [7]. The main aim is to remove the microphone from the image and fill the hole using facial semantic details, and this is carried out in two phases by an inpainter and refiner. To train the model, the authors combined adversarial loss, reconstruction loss, and perceptual loss. This makes sure that the generated images are accurate and are similar to the ground truth. The model was trained using a dataset of artificially created facial photos and microphones. The drawbacks of the work are that users must basically provide the microphone region according to the approach. This might be a challenging task, particularly for pictures with intricate backgrounds. Only a synthetic dataset of microphone-capable facial photos is used to train the algorithm. As a result, the approach might not work as effectively when used with real-world facial photos and microphones. The method is computationally expensive. This means that it may not be practical for real-time applications.

Unmasking of the face image is done using a GAN based network using two discriminators [8]. The authors train their model on a GAN-based network. Two discriminators are present in the network: one is used to learn the overall structure of the face, and the other is used to learn the deep missing region. Here, for the mask region, the model automatically creates a binary segmentation. Removes the mask while maintaining the overall coherence of the facial structure and synthesizes the afflicted area with precise details. The disadvantage of the work is that a paired dataset of photos with and without masks is needed for the procedure. This can be difficult to obtain, especially for real-world images. The method is computationally expensive. This means that it may not be practical for real-time applications.

Articles [7] and [8] address similar challenges in image manipulation, specifically microphone removal from facial images and unmasking of faces, respectively. While article [7] focuses on removing microphones from facial images, article [8] deals with unmasking faces obscured by masks. Both methods employ GAN-based networks and utilize discriminators to ensure the accuracy and coherence of the generated images.

Model compression is an optimization technique for shrinking the network's size without compromising performance. The model can be compressed in two ways: the first is by using a method for compressing neural networks during the training process, and the second is by using a technique for compressing trained neural networks. Different approaches for model compression include pruning, quantization, low-rank approximation and sparsity, neural architecture search, efficient architecture design, and knowledge distillation [3,9,10]. Several survey studies on deep learning compression exist, but those studies, on the other hand, tackle general deep learning algorithms rather than GAN-specific concerns.

Various research has been carried out to see how compressed GAN may be used in an embedded setting. To maintain network quality, the authors [11] used multiple iterative pruning methods in StarGAN during and after training. In this method, a compressed generator's training is supervised by a trained discriminator. The authors demonstrate how their method can produce a convincing performance even in highly sparse settings. This means that even if the compressed generator is much smaller than the original generator, it can still generate images of good quality. The shortcomings of the work are that not all GAN applications respond well to the approach. This shows that the strategy might not be suitable in

all situations. The technique is less efficient than other compression methods. The authors discovered that while other strategies can get a compression rate of 50% or more, their method only yields a compression rate of 25%. The approach hasn't been fully explored yet. The authors admit that they haven't yet tried to obtain very aggressive compression rates or to apply the technique to other GAN applications.

In paper [12], the Cycle-consistent generative adversarial network (CycleGAN) is compressed using a co-evolutionary technique based on a genetic algorithm, with discriminator loss being utilized to find unnecessary filters for pruning. The paper's authors suggest a co-evolutionary method for compressing GANs. In order to repeatedly investigate the most crucial convolution filters, this method works by encoding the generators for two picture domains as two populations and synergistically optimizing them. The number of parameters, a discriminator-aware regularization, and cycle consistency are used to determine each population member's fitness. This guarantees the compressed generators' effectiveness and compactness. Extensive tests performed on reference datasets show how effective the suggested approach is. The authors demonstrate how their approach may significantly increase compression rates while preserving the quality of the resulting photos.

In self-growing and pruning GAN (SP-GAN) [13], pruning technique minimizes excessive network growth by removing feature maps with higher correlation in each layer. For the creation of realistic images, the authors suggest a new SP-GAN. This approach can enhance the stability and effectiveness of network training by dynamically adjusting the size and design of a network throughout the training phase. Two seed networks that will serve as the generator and discriminator in the SP-GAN approach are initially trained. These seed networks are manageable and compact. The convolution kernels of each seed network are then duplicated in a self-growing process to increase the scale of the network. In order to achieve the ideal network scale, they apply a pruning approach to lessen the redundancy of the enhanced network. The downside of the paper is that it can be expensive computationally to implement the self-growing and pruning methods. This is due to the network's ongoing updating and the loss function's dynamic adaptation to various training phases.

Extra learnable layers are avoided using Knowledge Distillation with Kernel Alignment [14], which makes feature representations from the two models identical. In their approach, a teacher network serves as a search area to locate effective network architectures. Additionally, the authors provide a one-step pruning technique that looks for student architecture in the teacher model. This approach eliminates the need for l1 sparsity regularization and significantly lowers the cost of searching. The maximization of feature similarity between the instructor and student models is another method the authors suggest for condensing knowledge. They gauge feature similarity using an index called Global Kernel Alignment (GKA). A limitation of their work is that the ability of generative models to synthesize high-quality images under highly constrained computational budgets remains uncertain.

The authors [15] have used channel pruning and knowledge distillation to condense unconditional GANs. In this study, the authors suggest compressing unconditional GANs. Their strategy combines content awareness, knowledge distillation, and channel trimming. The authors' method significantly outperforms the most recent compression technique. They cut StyleGAN2's flops by 11 with almost perceptible image quality degradation. According to the authors, the content-aware compression method may not be effective for all types of content and the compressed models may not be as effective for all tasks as the full-size models. The proposed compression pipeline is more complex than previous methods, according to the authors. This suggests that deployment and implementation may be more challenging.

The proposed quantization [16] scheme aims to improve efficient and accurate on-device inference for deep learning models, especially on mobile devices. Utilizes integer arithmetic to approximate floating-point computations in neural networks, leading to a 4× reduction in model size and improved inference efficiency, particularly on ARM CPUs. This advancement enhances the tradeoff between latency and accuracy in computer vision models, making real-time visual recognition feasible on low-end phones. The synergy between quantization and efficient architecture design suggests integer-arithmetic-only inference as a key enabler for widespread deployment of AI technologies on resource-constrained devices. One potential disadvantage of the proposed quantization scheme is the risk of accuracy degradation during the quantization process. While the co-designed training procedure aims to preserve model accuracy post-quantization, there may still be some loss of precision, particularly in complex models or tasks requiring fine-grained distinctions.

The quantization technique [17] is applied to different GANs like the StyleGAN, Self-Attention GAN, and Cyclic GAN, and models are successfully quantized, maintaining the quality of the original model. Modern quantization methods were thoroughly experimentally studied by the paper's authors using three different GAN structures. They discovered that while maintaining the accuracy of the original full-precision models, they could successfully quantize these models for inference using 4/8-bit weights and 8-bit activations. The shortcoming of the work was that the study only looked into post-training quantization and quantization-aware training methods were taken into account in the study. Other quantization methods might work better. The study only took 4/8-bit quantization into account. Other bit depths might be more useful. The effect of quantization on the training process was not taken into account in the study. Quantization could make the training process more challenging or unsteady.

From the above literature survey, articles [3, 9-17] discuss model compression techniques aimed at reducing the computational complexity and memory footprint of GAN models. These techniques include pruning, quantization, knowledge distillation, and co-evolutionary methods. While some articles propose novel approaches for model compression [11-17], others provide comprehensive surveys or experimental evaluations of existing methods [3, 9-10]. Model compression techniques offer a pathway to reduce computational complexity and memory requirements, enabling the deployment of GAN models in resource-constrained environments.

While recent advancements in compression techniques, such as iterative pruning and co-evolutionary methods, show promise in reducing network size while preserving image quality, there is a need for further exploration and refinement of compression methods to overcome these challenges and enhance their applicability across different GAN architectures and tasks. It can be observed that there is a need for more efficient compression techniques, the lack of exploration of aggressive compression rates, limited applicability of certain techniques to specific GAN architectures or applications, and the uncertain impact of compression on image quality. The methods for compressing GANs are computationally expensive and may not be effective for all types of content. The methods may not be as effective for all tasks as the full-size models, and may not be able to produce results as high quality as the original GAN.

3. Methodology and Material

This paper aims to provide an approach for effectively removing occlusion from a facial image for facial recognition. Using the GAN approach, it is possible to decipher the facial picture from the obscured photographs [18]. The primary motivation is to obtain an un-occluded image with a minimum error rate and to reduce computational time.

Generative adversarial networks are used for different image transformations, conversion, generation, and formations for various problems in computer vision. In deep learning, having just one network is computationally demanding; GANs have a complex architecture with more than one network and compete against one another. The architecture includes a generator and a discriminator. As a result, the model is vast, necessitating more training time and memory usage. High-performance processors are required to shorten training time, and significant memory is required for deploying the model. This high computation and memory requirement is unsuitable for devices with low computation since they have limited resources. Lightweight or compressing the model can reduce memory usage and computation time. This lightweight model is suitable for processing in low computation devices.

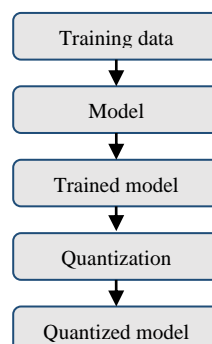


Figure 1. Post quantization process

The discriminator is a neural network whose function is to create an output that is merely a decision or a prediction. In contrast, the generator, generates the data. As a result, the generator network's output is comparatively more extensive than the input. Even though a GAN comprises two networks, the discriminator is critical during model training. At the same time, the generator often takes care of most of the work once the model is deployed. When dealing with model compression in GAN, the emphasis is on efficient generator deployment by utilizing state-of-the-art compression approaches.

The basic steps of model compression using post quantization are depicted in figure 1. The model is given with the input data and training is carried out for fine-tuning of the model. After the model has been trained, it is quantized to minimize its size, and the resultant quantized model is compressed.

3.1. Model compression in Pix2pix

Pix2pix is a GAN in which the generator and discriminator compete to produce a model that is both efficient and accurate. Figure 2 portrays the steps involved in the model compression. In occlusion removal in face images using pix2pix and thereby, the algorithm for occlusion removal. A U-NET is used in the pix2pix generator, and the discriminator is a patch GAN. The discriminator has been trained to distinguish between bogus and real images. To imitate the discriminator, the generator creates a fake image.

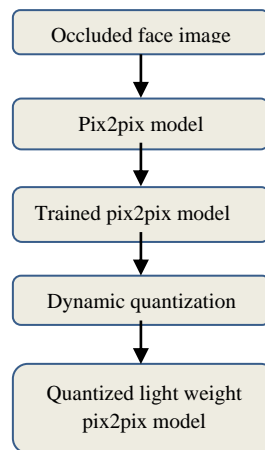


Figure 2. Pix2pix quantized model

Algorithm 1. Algorithm for Occlusion Removal

Input: Webface OCC database

Output: Occlusion Removed

1: Start

2: Loading the occluded image.

3: Resizing and normalizing the image according to the pix2pix model input condition

4: Loading the trained Pix2pix Model for occlusion removal

5: Passing the occluded image to trained pix2pix model.

6: Perform down sampling by convolutional layers and then features are up sampled by the up-sampling layers.

7: Apply concatenation on down sample layer output and up sampled layer to generate the occlusion removed image.

8: Stop

During the training phase, the generator module takes occluded image as input and generates non occluded images. The discriminator takes in occluded image and real image without occlusion, trying to figure the classification matrix of the given image. Both the generator and discriminator compete and learn against each other. The discriminator determines the probability of a class with the provided features. The goal of GAN is to make the output created by the generator look like the real distribution. The occlusion removal training method is determined by the following algorithm.

Algorithm 2. Algorithm for Occlusion Removal Training

Input: Batches of Occluded images

Output: Image from generator

1: Start

2: Load the occluded images from Webface OCC database

3: Create suitable batches of images for training

- 4: Pass the occluded image to the generator
- 5: Output image from generator is passed to the discriminator.
- 6: Determine error rate between real and generated output image using the eqn (1).
- 7: Update the weights of Discriminator using discriminator loss using the eqn (1).
- 8: Update the weights of Generator using generator loss using the eqn (2).
- 9: Repeat step 4 to step 8 for each batch of images.
- 10: Stop

The discriminator loss is calculated based on output from the generator and the input image. The Binary Cross Entropy (BCE) loss [19] equation yields the discriminator loss as given in Eq. (1).

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^m [y^{(i)} \log h(x^{(i)}, \theta) + (1 - y^{(i)}) \log (1 - h(x^{(i)}, \theta))] \quad (1)$$

Where the parameters are, x and y stand for features and labels. The prediction is given by h , and is the discriminator parameter.

Generator loss is the sum of BCE loss [20] and L1 loss multiplied by λ . L1 loss is the mean absolute difference between the generated output and the true output. The generator loss is given in Eq. (2).

$$\text{Generator Loss} = \text{BCE Loss} + \lambda \sum_{i=1}^n |\text{generated output} - \text{real output}| \quad (2)$$

After the training step, the model learns the information from the training data and based on this information, the weights of the model are updated. The weights of the model are optimized during the quantization step, which minimizes the model's size. This optimized model can be suitable for running in low computation devices like mobile phones, Raspberry pi etc. During the optimization step, the weights of the trained model are quantized. The algorithm for dynamic quantization is depicted below.

The U-Net section of the generator has 12 layers, 6 of which are contracting and 6 of which are expanding. Weights are kept as tensors in learned models. There are 117438275 total learnable parameters. The model is trained up to 250 epochs. The quantization method takes this trained model as input. Prior to quantization, we must configure the backends for running quantized operators, with the architecture set to x86 CPUs. We'll use Qconfig to describe how weights and activations will be quantized. Qengine provides the backend after the quantized model has been done. In dynamic quantization, the weights and activation are dynamically quantized.

Algorithm 3. Algorithm for Dynamic Quantization

Input: Trained Model

Output: Quantized model file

1: Start

2: Load the trained model

3: Determine Scale factor dynamically for the activation function based on the data range.

4: Convert the model weights into precision integer by multiplying floating point number by a scale factor and round off to whole number to obtain the Quantized model using the eqn (3).

5: Stop

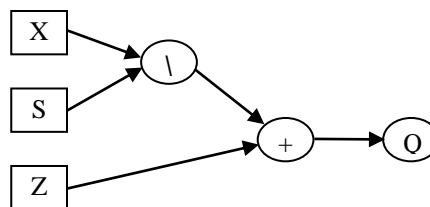


Figure 3. Flow chart of the quantization

Figure 3 shows the quantization flow chart, with X representing the trained model's weights, S is the scaling factor, and Z representing the zero point. As a result, the floating-point values are transformed to integers, yielding quantized weights and an activation function.

3.2. Quantization

Weights in the deep neural network are stored as 32-bit floating point numbers. The size of the neural network can be significantly reduced by reducing the number of bits utilized to represent the weights [21]. The idea behind the quantization technique is to reduce the number of bits used to express these weights.

The number of bits can be reduced to 16-bit, 8-bit, 4-bit, or 1-bit. Quantization limits the bits of precision of the model parameters, resulting in a reduction in the amount of data that needs to be saved. Technically, it is the process of clustering or rounding off weights in order to represent the same number of connections with less memory. The following parameters are included in the basic quantization equation: floating points represented as x , scale factor, and zero point. The zero point is used to convert negative integers to positive values. Quantization is given below in Eq. (3);

$$(x, \text{scale}, \text{zero point}) = \text{round}(x \text{scale} + \text{zero point}) \quad (3)$$

In the current, situation most of the deep learning models are client-server based. This end-to-end communication depends on network bandwidth, speed and latency. After deploying the model, the user must provide the data to the server side for inference. The data is given to the client after the inference is formed. The client-server communication will be eliminated using the lightweight models. This will shorten the execution time and deliver an instant result. A lightweight and low computational model helps in removing this client server communication.

Quantization can be done during training known as quantization, aware training, and after training known as post quantization. Quantization is applied to reduce the model size and can be downloaded to handheld devices for testing. In our approach, we use post-training quantization, which involves first training the model using the input data and then applying quantization to the trained model. There are three types of quantization based on the reduction of bits used to indicate weights in the network.

The following are the benefits and drawbacks of quantization in convolutional neural networks: It can be utilized for both fully connected and convolution layers, according to the pros. The quantization procedure can be used during or after training. Execution time can be reduced and processing of data can be done faster. The size of the model can be lowered without sacrificing quality. The cons include quantized weights make neural networks more difficult to converge [15], and back propagation becomes harder with quantized weights.

3.2.1. Dynamic Quantization

When a network is compressed through quantization, the weights and activation are converted to a lower precision integer form. To cut down on the number of bits required to store the weights and activation. The weights are converted from floating points to integer points. This method involves dividing a floating-point value by a scaling factor and rounding the result to a whole number. The scale factor for activation in dynamic quantization is determined dynamically by the data range recorded at runtime. During model conversion, the model parameters are known and are converted and saved in INT8 format ahead of time.

The weights of neural network architecture are normally recorded in 32 bits, but when utilizing dynamic quantization, they are converted to 8 bits, resulting in a fourfold reduction in the size of the model. This reduction in model parameters also aids in upto 50% faster execution for convolution models and upto 3 times faster execution for fully connected and RNN based models.

4. Experimental Study and Results

In this experiment we used the pix2pix method for the occlusion removal in face images and the database used is the webface-OCC. Webface-OCC [22], an occluded face recognition data set of 804, 704 face images of 10, 575 subjects. PyTorch version 1.9.0+cu102 was used to create the model. From the Webface-OC dataset, 46262 training samples with a size of 256*256 were used for training. On a Tesla P100-PCIE with 25.346GB of RAM, the model was trained and evaluated. Model were trained upto 250 epochs. Pytorch dynamic quantization was used on the original model for model compression.

The simulation results for occlusion removal in face images using the quantized pip2pix are shown in Table 1. The various assessment metrics, such as Peak Signal to Noise Ratio (PSNR) and Structural SIMilarity (SSIM), were compared to different state of the art image editing methods, such as GLCI [23], GICA [24], EdgeConnect [25], MRGAN [7], and GUMF Method [8].

The experimental results in Table 1 show that the quantized pix2pix method outperforms the other state-of-the-art image editing methods for occlusion removal in face images. This is evident from the higher values of SSIM and PSNR obtained for the quantized pix2pix method.

Table 1. Performance comparison of proposed work with other work

Methods	SSIM	PSNR
GLCI [23]	0.752	21.20db
GICA [24]	0.791	17.55db
EdgeConnect [25]	0.819	19.78db
MRGAN [7]	0.847	24.02db
GUMF [8]	0.854	25.17db
Quantized Pix2pix (Proposed)	0.896	27.32db

SSIM is a measure of how similar two images are, while PSNR is a measure of the amount of noise in an image. The higher the values of SSIM and PSNR, the more similar the two images are and the less noise there is in the image.

In this case, the quantized pix2pix method achieved a SSIM of 0.896 and a PSNR of 27.32dB, which is significantly higher than the other methods. This means that the quantized pix2pix method was able to restore the original image more accurately than the other methods.

The reason for the improved performance of the quantized pix2pix method is due to the use of a quantized generative adversarial network (GAN). GANs are a type of machine learning model that can be used to generate realistic images. The results of this experiment suggest that the quantized pix2pix method is a promising new approach for occlusion removal in face images. The method is able to achieve high SSIM and PSNR values, which indicates that it is able to restore the original image accurately. Additionally, the method is relatively efficient, which makes it suitable for real-world applications.

Figure 4 and figure 5 show the SSIM and PSNR of various methods. The quantized pix2pix method has higher value than the existing methods GLCI, GICA, EdgeConnect, MRGAN, and GUMF Method.

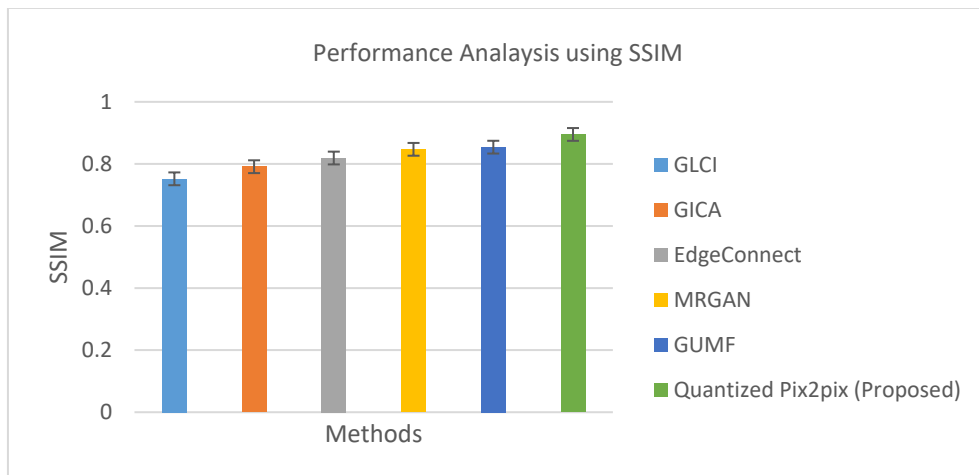
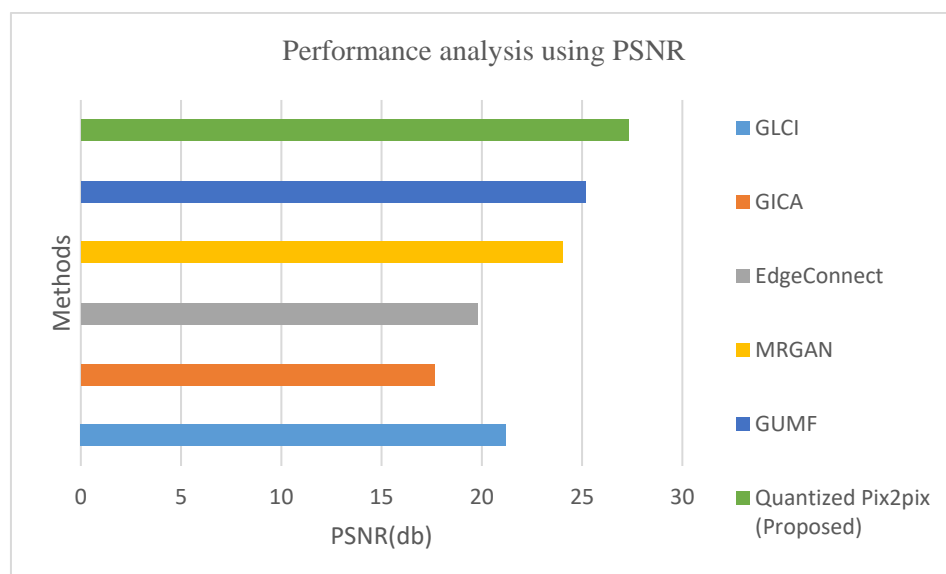
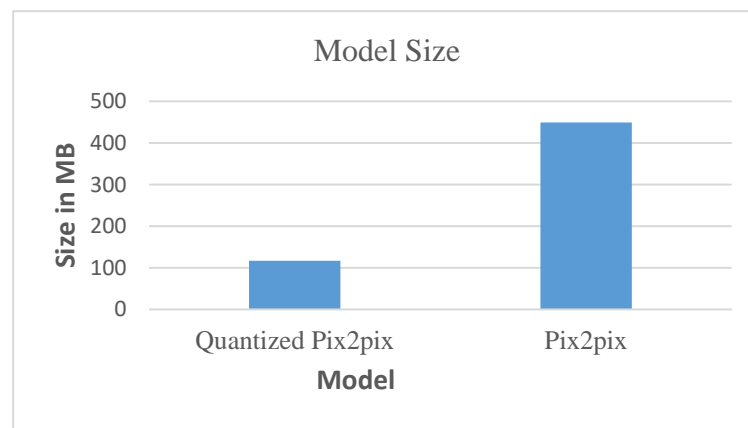
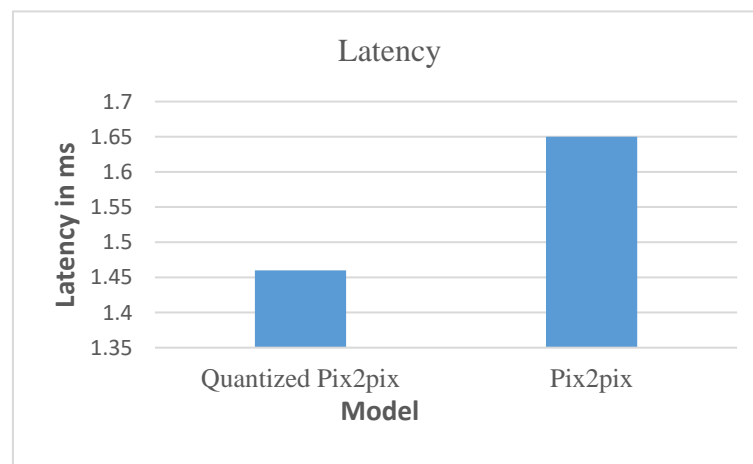
**Figure 4.** Performance metrics of SSIM**Figure 5.** Performance metrics of PSNR

Table 2. Model Compression results based on different parameters

Model	Model Size (MB)	Latency (ms)	L1 Loss
Pix2pix model [18]	449	1.65	0.0463
Quantized Pix2pix (Proposed)	117	1.46	0.0452

Quantized pix2pix is also compared with our previous work on occlusion removal by pix2pix [18]. Table 2 shows the results obtained based on model size, latency and L1 loss. The model size was compressed from 449 mb to 117 mb. The L1 loss is the pixel wise distance between the original image and the occlusion removed image. The calculated L1 loss for a Quantized model resulted in quantitative loss of 0.0452 and the pix2pix model is of 0.0463. The L1 loss is a measure of the pixel-wise difference between two images. A lower L1 loss indicates that the two images are more similar. In this case, the quantized pix2pix model has a lower L1 loss than the pix2pix model, which means that the quantized model is more similar to the original image. The reason for this is that the quantized model has been compressed, which means that it has fewer parameters. This makes the quantized model less complex, which can lead to a lower L1 loss.

**Figure 6.** Comparison based on Model Size**Figure 7.** Comparison based on Latency

Latency is the amount of time to complete one inference referred to as T , T_2 is the end time of the execution and T_1 is the start of the execution. The execution time taken by the quantized pix2pix model is 1.46ms and the pix2pix model is 1.65ms. The execution time of a model is the time it takes for the model to make a prediction on a single image. The execution time of the quantized pix2pix model is 1.46 ms, while the execution time of the pix2pix model is 1.65 ms. This means that the quantized pix2pix model is able to make predictions 11% faster than the pix2pix model. The reason for the difference in execution time is due to the fact that the quantized pix2pix model has fewer parameters. This means that the quantized model can be processed more quickly by the Central Processing unit (CPU) or GPU. L1 loss and the latency is given in Eq. (4) and Eq. (5):

$$L1\ Loss = \sum_{i=1}^n |generated\ output - real\ output| \tag{4}$$

$$T = T2 - T1 \tag{5}$$

The graph was plotted with various metrics of the Quantized pix2pix model and pix2pix model. Figure. 6 represents the graph with y-axis as the model size and x-axis as the Quantized model and pix2pix model. Similarly, Figure. 7 and Figure 8 shows the latency and L1 loss observed for the quantized pix2pix model and the pix2pix model, respectively.



Figure 8. Comparison based on L1 Loss

The figure. 9 represents the original image, the occlusion region in the mouth, and the results obtained with quantized pix2pix model and pix2pix model.




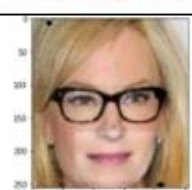
Original Image	
Occluded Image	
Occlusion removed Image using Quantized Pix2pix model	
Occlusion removed Image using Pix2pixmodel	

Figure 9. Output obtained using quantized pix2pix and pix2pix model

5. Discussion

The quantized pix2pix model is able to achieve better image quality than the pix2pix model, while also being significantly smaller and faster. This makes the quantized pix2pix model a promising new approach for occlusion removal in face images. The quantized pix2pix model has a lower model size, lower latency, and lower L1 loss than the pix2pix model. This is because the quantized model has been

compressed, which means that it has fewer parameters. This makes the quantized model less complex, which can lead to a lower latency. Additionally, the quantized pix2pix model also achieves higher SSIM and PSNR values than the other state of the art methods. This indicates that the quantized pix2pix model is able to restore the original image more accurately, while also being significantly smaller and faster. Overall, the results show that the quantized pix2pix model is able to achieve better image quality than the pix2pix model, while also being significantly smaller and faster. This makes the quantized pix2pix model a promising new approach for occlusion removal in face images.

6. Conclusion

In this study key ideas necessary for model compression of pix2pix model for removal of occlusion are emphasized. It is demonstrated that applying the hybrid quantization technique to trained GANs can result in compressed generators without sacrificing quality. This technique resulted in a considerable reduction in the generator's size, which was reduced from 449mb to 117mb i.e 74%. After quantization the execution time was slightly reduced without compromising accuracy. According to our literature review, this is the first time a light weight model has been used to remove occlusion in the mouth region utilizing pix2pix. The size of the model has been reduced by 74%. There is compressed model for different applications like image classification, image enhancement, segmentation and image to image translation. The proposed quantized pix2pix provides 0.896 as the SSIM and 27.32db as the PSNR value which is higher when compared to various state of the art methods. This makes the quantized pix2pix model a good choice for applications where image quality is important, but where speed and/or storage space are also important considerations. The quantized pix2pix model could be used in a variety of applications, such as facial recognition, augmented reality, and virtual reality. The model could be used to improve the quality of images that have been partially obscured by objects or other people. The model could be used to create new images that are based on existing images.

References

- [1] G. Rajeswari and P. Ithaya Ran, "Face occlusion removal for face recognition using the related face by structural similarity index measure and principal component analysis", *Journal of Intelligent & Fuzzy Systems: Application in Engineering and Technology*, pp. 5335-5350, Vol. 42, No. 6, 1st January 2022, Published by IOS Press, DOI: 10.3233/JIFS-211890, Available : <https://dl.acm.org/doi/abs/10.3233/JIFS-211890>.
- [2] Diksha Khas, Sumit Kumar and Satish Kumar Singh, "Facial Occlusion Detection and Reconstruction Using GAN", in *Communications in Computer and Information Science: Computer Vision and Image Processing*, Singapore: Springer Nature, 2021, Print ISBN: 978-981-16-1091-2, Vol. 1377, Ch. 2, pp. 255-267, DOI: 10.1007/978-981-16-1092-9_22, Available: https://link.springer.com/chapter/10.1007/978-981-16-1092-9_22.
- [3] Yu Cheng, Duo Wang, Pan Zhou and Tao Zhang, "Model Compression and Acceleration for Deep Neural Networks: The Principles, Progress, and Challenges", *IEEE Signal Processing Magazine*, Print ISSN: 1053-5888, pp. 126-136, Vol. 35, No. 1, 10th January 2018, Published by IEEE, DOI: 10.1109/MSP.2017.2765695, Available: <https://ieeexplore.ieee.org/abstract/document/8253600>.
- [4] Tejalal Choudhary, Vipul Mishra, Anurag Goswami and Jagannathan Sarangapani, "A comprehensive survey on model compression and acceleration", *Artificial Intelligence Review*, Vol. 53, pp. 5113-5155, 8th February 2020, Published by Springer Nature, DOI: 10.1007/s10462-020-09816-7, Available: <https://link.springer.com/article/10.1007/s10462-020-09816-7>.
- [5] Yijun Li, Sifei Liu, Jimei Yang and Ming-Hsuan Yang, "Generative face completion", in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 21-26 July 2017, Honolulu, USA, Electronic ISBN: 978-1-5386-0457-1, Print ISSN: 1063-6919, pp. 3911-3919, Published by IEEE, DOI: 10.1109/CVPR.2017.624, Available: <https://ieeexplore.ieee.org/document/8100107>.
- [6] Kamran Javed, Nizam Ud Din, Seho Bae, Rahul S. Maharjan, Donghwan Seo *et al.*, "UMGAN: Generative adversarial network for image unmosaicing using perceptual loss", in *Proceedings of the 2019 16th International Conference on Machine Vision Applications (MVA)*, 27-31 May 2019, Tokyo, Japan, Electronic ISBN: 978-4-901122-18-4, pp. 1-5, Published by IEEE, DOI: 10.23919/MVA.2019.8757902, Available: <https://ieeexplore.ieee.org/document/8757902>.
- [7] Muhammad Kamran Javed Khan, Nizam Ud Din, Seho Bae and Juneho Yi, "Interactive removal of microphone object in facial images", *Electronics*, Print ISSN: 2079-9292, Vol. 8, No. 10, p. 1115, 2nd October 2019, Published by MDPI, DOI: 10.3390/electronics8101115, Available: <https://www.mdpi.com/2079-9292/8/10/1115>.

- [8] Nizam Ud Din, Kamran Javed, Seho Bae and Juneho Yi, "A novel GAN-based network for unmasking of masked face", *IEEE Access, Electronic*, ISSN: 2169-3536, Vol. 8, pp. 44276 – 44287, 2nd March 2020, Published by IEEE, DOI: 10.1109/ACCESS.2020.2977386 , Available: <https://ieeexplore.ieee.org/document/9019697>.
- [9] Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney *et al.*, "A Survey of Quantization Methods for Efficient Neural Network Inference", *Low-Power Computer Vision*, 1st ed. New York, USA: Taylor & Francis, 2022, E-ISBN: 9781003162810, Ch. 13, pp 291-326, DOI: 10.1201/9781003162810-13.
- [10] Dina Tantawy, Mohamed Zahran and Amr Wassal, "A survey on GAN acceleration using memory compression techniques", *Journal of Engineering and Applied Science*, Vol. 68, No.1, 19th December 2021, pp. 1-23, Published by Springer Nature, DOI: 10.1186/s44147-021-00045-5, Available: <https://jeas.springeropen.com/articles/10.1186/s44147-021-00045-5>.
- [11] Chong Yu and Jeff Pool, "Self-supervised Generative Adversarial compression", *Advances in Neural Information Processing Systems*, ISBN: 9781713829546, Vol. 33, pp. 8235-8246, 3rd July 2020, Available: <https://proceedings.neurips.cc/paper/2020/hash/5d79099fcd499f12b79770834c0164a-Abstract.html>.
- [12] Han Shu, Yunhe Wang, Xu Jia, Kai Han, Hanting Chen *et al.*, "Co-Evolutionary Compression for Unpaired Image Translation", In *Proceedings of 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 27 October - 02 November 2019, Seoul, South Korea, Electronic ISBN: 978-1-7281-4803-8, Print ISSN: 1550-5499, pp. 3235–3244, Published by IEEE, DOI: 10.1109/ICCV.2019.00333, Available: <https://ieeexplore.ieee.org/document/9010692>.
- [13] Xiaoning Song, Yao Chen, Zhen-Hua Feng, Guosheng Hu, Dong-Jun Yu *et al.*, "SP-GAN: Self-growing and pruning generative adversarial networks", *IEEE Transactions on Neural Networks and Learning Systems*, Print ISSN: 2162-237X, pp. 2458 – 2469, Vol. 32, No. 6, 10th July 2020, Published by IEEE, DOI: 10.1109/TNNLS.2020.3005574, Available: <https://ieeexplore.ieee.org/document/9138445>.
- [14] Qing Jin, Jian Ren, Oliver J. Woodford, Jiazhou Wang, Geng Yuan *et al.*, "Teachers Do More Than Teach: Compressing Image-to-Image Models", in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20-25 June 2021, Nashville, USA, Electronic ISBN: 978-1-6654-4509-2, Print ISSN: 1063-6919, pp. 13600-13611, Published by IEEE, DOI: 10.1109/CVPR46437.2021.01339, Available: <https://ieeexplore.ieee.org/document/9578627>.
- [15] Yuchen Liu, Zhixin Shu, Yijun Li, Zhe Lin, Federico Perazzi *et al.*, "Content-Aware GAN Compression", in *Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20-25 June 2021, Nashville, USA, Electronic ISBN: 978-1-6654-4509-2, Print ISSN: 1063-6919, pp. 12151-12161, Published by IEEE, DOI: 10.1109/CVPR46437.2021.01198, Available: <https://ieeexplore.ieee.org/document/9578495>.
- [16] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang *et al.*, "Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference", in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18-23 June 2018, Salt Lake City, USA, Electronic ISBN: 978-1-5386-6420-9, Electronic ISSN: 2575-7075, pp. 2704-2713, Published by IEEE, DOI: 10.1109/CVPR.2018.00286, Available: <https://ieeexplore.ieee.org/document/8578384>.
- [17] Pavel Andreev and Alexander Fritzler, "Quantization of Generative Adversarial Networks for Efficient Inference: A Methodological Study", in *2022 26th International Conference on Pattern Recognition (ICPR)*, 21-25 August 2022, Montreal, Canada, Electronic ISBN: 978-1-6654-9062-7, Print ISSN: 1051-4651, pp. 2179-2185, Published by IEEE, DOI: 10.1109/ICPR56361.2022.9956041, Available: <https://ieeexplore.ieee.org/document/9956041>.
- [18] Sincy John and Ajit Danti, "Removal of Occlusion in Face Images Using PIX2PIX Technique for Face Recognition", in *Lecture Notes on Data Engineering and Communications Technologies: Congress on Intelligent Systems*, Singapore: Springer Nature, Print ISBN: 978-981-16-9112-6, Vol. 111, Ch. 5, pp. 47-57, DOI: 10.1007/978-981-16-9113-3_5, Available : https://link.springer.com/chapter/10.1007/978-981-16-9113-3_5.
- [19] Mesay Belete Bejiga and Farid Melgani, "Gan-Based Domain Adaptation for Object Classification", in *Proceedings of the 2018 IEEE International Geoscience and Remote Sensing Symposium (IGARSS 2018)*, 22-27 July 2018, Valencia, Spain, Electronic ISBN: 978-1-5386-7150-4, Print ISSN: 2153-6996, pp. 1264-1267, Published by IEEE, DOI: 10.1109/IGARSS.2018.8518649, Available: <https://ieeexplore.ieee.org/document/8518649>.
- [20] Kusam Lata, Mayank Dave and K N Nishanth, "Image-to-Image Translation Using Generative Adversarial Network", in *Proceedings of the 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*, 12-14 June 2019, Coimbatore, India, Print ISBN: 978-1-7281-0168-2, pp. 186-189, Published by IEEE, DOI: 10.1109/ICECA.2019.8822195, Available: <https://ieeexplore.ieee.org/document/8822195>.
- [21] Tailin Liang, John Glossner, Lei Wang, Shaobo Shi and Xiaotong Zhang, "Pruning and Quantization for Deep Neural Network Acceleration: A Survey", *Neurocomputing*, Print ISSN: 0925-2312, Online ISSN: 1872-8286, Vol. 461, 21st October 2021, pp. 370-403, Published by Elsevier B.V., DOI: 10.1016/j.neucom.2021.07.045, Available: <https://www.sciencedirect.com/science/article/abs/pii/S0925231221010894>.
- [22] Baojin Huang, Zhongyuan Wang, Guangcheng Wang, Kui Jiang, Kangli Zeng *et al.*, "When Face Recognition Meets Occlusion: A New Benchmark", in *Proceedings of the ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 06-11 June 2021, Toronto, Canada, Electronic ISBN: 978-1-7281-

- 7605-5, Print ISSN: 1520-6149, pp. 4240-4244, Published by IEEE, DOI: 10.1109/ICASSP39728.2021.9413893, Available: <https://ieeexplore.ieee.org/document/9413893>.
- [23] Satoshi Lizuka, Edgar Simo-Serra and Hiroshi Ishikawa, "Globally and locally consistent image completion", *ACM Transactions on Graphics*, Vol. 36, No. 4, 20th July 2017, pp. 1-14, Published by ACM, DOI: 10.1145/3072959.3073659, Available: <https://dl.acm.org/doi/abs/10.1145/3072959.3073659>.
- [24] Jiahui Yu, Zhe in, Jimei Yang, Xiaohui Shen, Xin Lu *et al.*, "Generative image Inpainting with contextual attention", in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18-23 June 2018, Salt Lake City, USA, Electronic ISBN: 978-1-5386-6420-9, Print ISSN: 1063-6919, pp. 5505-5514, Published by IEEE, DOI: 10.1109/CVPR.2018.00577, Available: <https://ieeexplore.ieee.org/document/8578675>.
- [25] Kamyar Nazari, Eric Ng, Tony Joseph, Faisal Z. Qureshi and Mehran Ebrahimi, "EdgeConnect: Structure Guided Image Inpainting using Edge Prediction", in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, 27-28 October 2019, Seoul, South Korea, Electronic ISBN: 978-1-7281-5023-9, pp. 3265-3274, DOI: 10.1109/ICCVW.2019.00408, Available: <https://arxiv.org/abs/1901.00212>.



© 2024 by the author(s). Published by Annals of Emerging Technologies in Computing (AETiC), under the terms and conditions of the Creative Commons Attribution (CC BY) license which can be accessed at <http://creativecommons.org/licenses/by/4.0>.