Journal of Data Acquisition and Processing

ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7546499

END TO END IMAGE PRIVACY METHOD USING CUSTOMIZED GENERATIVE ADVERSARIAL NETWORKS

Ram Mohan Rao P[1,4,*], AP Siva Kumar[2], Murali Krishna[3,4]

[1]Department of CSE, JNTU Anantapur, Andhra Pradesh India
[2]Professor, Department of CSE, JNTU Anantapur, Andhra Pradesh India
[3] Professor, Department of CSE, SV College of Engineering and Technology, Tirupati, Andhra Pradesh India
[4]Department of CSE, Sphoorthy Engineering College Hyderabad, Telangana India

Abstract:

The importance and need for data privacy is now commonly discussed everywhere and many regulations are also in place to ensure privacy preservation. Personal information like age, gender, marital status, health records etc. are the common data which needs privacy. However, images also contain sensitive information that can compromise individual privacy. With huge number of images being uploaded into public domain, image privacy also became a major privacy concern. The conventional methods of image privacy include image obfuscation which results into images which are not useful enough for further analytics. Modern literature suggests usage of Generative Adversarial Networks (GAN) to generate synthetic data for image privacy protection. The tradeoff between privacy and utility still exists. Our solution is customized GAN based method which is able to generate synthetic face images which can strike a balance between privacy and utility. We have used structural similarity index measure (SSIM) to compare the images generated using various methods and our image privacy method achieved SSIM of 98.3% along with metadata being preserved. The following sections will describe our contributions in detail.

Keywords: Computer vision, image privacy, Generative Adversarial Networks, Structural Similarity Index Measure, Privacy Preservation.

1. INTRODUCTION

It has become a common habit these days, that people go out, spend time with friends and post those pictures to social media platforms and other cloud platforms. Many free cloud based applications are available to upload pictures, manipulate pictures and store them. However, there is no assurance from these cloud providers that privacy of our images is preserved or not. Without consent of the user, these images can be shared with third party vendors also, which is a serious privacy concern [1]. Earlier researchers have developed many image obfuscation techniques like blurring, pixelation, masking etc. These techniques have proved to be inadequate with respect to image utility [2]. In the recent times, deep learning based methods especially using Generative Adversarial Networks (GAN) were widely used for generating synthetic images as discussed by Zhenfei Chen at. al [3]. The GAN based methods are able to generate highly realistic images that replicate the original images by hiding the private and sensitive attributes of the image. The GAN based image privacy has achieved significant progress in the field of image privacy but yet to reach to an optimal solution. Our objective is to extend this work by applying various forms of GANs to image data such that the sensitive information can be preserved while the image is still usable for further analytics. The idea is to take a face image as an input and generate a synthetic image with high resemblance to original image. The fake image thus generated will be such that, one cannot identify who the person is but still the characteristics of the face image are retained [4]. However, there are many challenges involved in achieving a balance between privacy and utility of the image data. Firstly, how to generate a fake image which can replace the original one. Second challenge is how to preserve the sensitive portions of the image. The third challenge is how to ensure utility of the data and also how to preserve the metadata that is being uploaded along with the face image. Latest image privacy concerns include deep fake face creation which leads to cyberbullying and many such cases of cyberbullying were reported in the recent times [5]. 95 million photos are uploaded every day in Instagram which is an alarming number and it will continue to increase every day [6]. On the other hand, GAN is like a double ended sword which can also be misused. Deep fake is an example of a nefarious use of GAN to generate synthetic images and morph them to other face images. Many cyberbullying cases were reported recently where faces of celebrities were morphed with nudity images as discussed by Nagwan Abdel Samee at. al. [7] and Pier Paolo Tricomi at. al. [8]. Differential privacy and federated learning can be combined and used to ensure privacy of the images as discussed by Xu Zheng at. al. [9]. Hence there is a need to generate face images that do not match with any original image but still be good enough for analytics.

We have customized the GAN model by iteratively fine tuning the parameters across various epochs and finally trained the model using custom data. The final output image in our solution is a bitmap format of the image because bitmap format does not include any metadata of the image. All other image formats also upload metadata along with the image. Apart from the metadata preservation, we also ensured that background of the image should not be changed, so that the image resembles the original image.

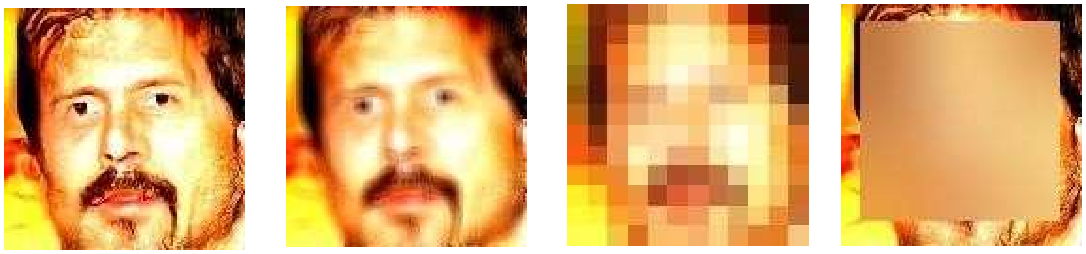
Our solution makes following contributions and advancements to the literature.

1. Able to generate synthetic image which can preserve private and sensitive information.
2. Offer better data utility which can be measured using structural similarity index measure (SSIM)
3. Able to use the generated face images for analytics.

Background and Related Work

We have implemented the conventional methods viz. obfuscation, pixelation, masking and deidentification and tried to measure utility of the modified image along with privacy preservation. Obfuscation, pixelation often depends on the image quality i.e. the size of the block. High resolution images can offer better privacy. If the resolution is less, the protection offered is also less. Masking and deidentification offers maximum privacy but the utility of the image is very less. Hence the conventional methods of privacy are proved to be inadequate.

Figure 1 describes the output of various methods applied on the original image.



a b c
d

Figure 1 a – original image, b- obfuscated (blurred image), c-pixelization and d-deidentification

The conventional methods viz. blurring, pixelization, deidentification etc. are not offering data utility such that the images can be used for any analytical applications. Hence the conventional methods are not used anymore and we proposed a GAN based image privacy method with differential privacy to ensure the privacy of the image is protected while the image is still usable for analytical purposes. Our GAN based end to end privacy method is described in the next section.

Proposed Customized GAN based Method

GAN is widely used to create synthetic data including images. A GAN based image privacy preservation

methods were discussed in Yu wang at. al [9][10]. GAN is a deep neural network model with two components viz. generator and discriminator models. Generator model takes a fixed length random vector as input and generates an image in the domain. The vector is drawn randomly from Gaussian distribution called latent space. The discriminator takes an example from the domain as input and predicts a binary class label of real or fake. After training process, the discriminator model is discarded as we are interested in the generator. The generator learns the joint probabilistic distribution of the input variable and the output variable as mentioned in the equation 1.

$$P\ (X, Y) = P(Y/X)\ P(X)$$

For prediction, the bayes theorem is used internally to find conditional probability of target variable given the input variable as described in the equation 2.

$$P(Y/X) = P\ (X, Y)\ /\ P(X)$$

The two models generator and discriminator are trained together. A single training cycle involves first selected batch of real images from problem domain. A batch of latent points is generated and fed to the generator model to synthesize a batch of images. The discriminator is then updated using the batch of real and generated images minimizing binary cross entropy loss used in any binary classification problem. The generator is then updated via the discriminator model. The generated images are presented to discriminator as if they are real and the error is propagated through the generator model. This will help in updating the generator model towards generating images that are more likely to fool the discriminator. This process is repeated number of times till the discriminator accepts the synthetic images as real one. The entire process of generator and discriminator models is based on minmax strategy. As per minmax strategy applied on GAN, it is an optimization strategy where the generator would want to minimize its loss where as the discriminator would like to maximize its accuracy of comparison. There are several pre trained GAN models available which can be used to generate synthetic data like text, images, audio and video data. PROGAN 128 is a popular image generation GAN model which is freely available and can be trained on custom data also.

The minmax function is described in Figure 2. The loss function describes both generator and discriminator loss functions.

$$\min_{G}\max_{D} V(D,G) = \underbrace{\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})]}_{\text{log prob of D predicting that real-world data is genuine}} + \underbrace{\mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]}_{\text{log prob of D predicting that G's generated data is not genuine}}.$$

Figure 2 Minmax loss function

The GAN model contains generator and discriminator, both will try to optimize themselves by repeated training epochs. They apply the minmax strategy for optimization. Typical GAN architecture is described in the Figure 3.

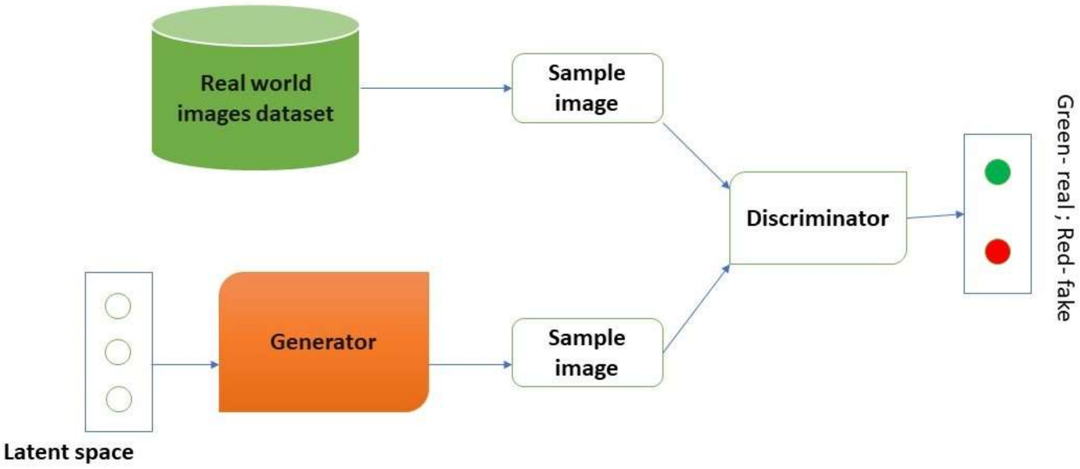


Figure 3 GAN architecture

The regular pre trained models of GAN are not sufficient enough to ensure privacy of an image. GAN can only generate synthetic image of a real image. We have proposed a novel end to end image privacy architecture where the input image is passed to the GAN network to generate a synthetic image of the input and background of the image is preserved to ensure the synthetic image resembles the input image. Using python open AI api the background preservation of the image is achieved. A detailed architecture of the proposed solution is shown in Figure 4.

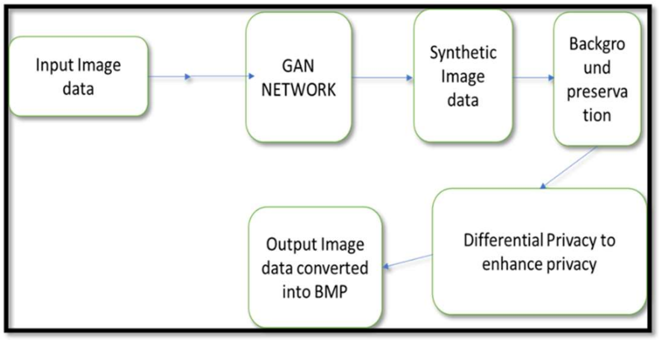

Input Image data
GAN NETWORK
Synthetic Image data
Backgro und preserva tion
Differential Privacy to enhance privacy
Output Image data converted into BMP

Figure 4 Proposed architecture of end-to-end image privacy

Experimental Setup

Based on the architecture described in Figure 4, the following experimental steps were carried out to develop a concrete end-end solution for image privacy.

1. Develop a customized GAN model
2. Background preservation
3. Application of Differential Privacy
4. Image transformation

Customized GAN model:

We have trained a PROGAN 128 (progressive GAN) model using following datasets.
A popular component of computer vision and deep learning revolves around identifying faces for various applications from logging into your phone with your face or searching through surveillance images for a particular suspect. This dataset is great for training and testing models for face detection, particularly for recognizing facial attributes such as finding people with brown hair, are smiling, or wearing glasses. Images cover large pose variations, background clutter, diverse people, supported by a large quantity of images and rich annotations. This data was originally collected by researchers at MMLAB, The Chinese University of Hong Kong.

202,599 number of face images of various celebrities
10,177 unique identities, but names of identities are not given
40 binary attribute annotations per image
5 landmark locations

Data Files

img_align_celeba.zip: All the face images, cropped and aligned
list_eval_partition.csv: Recommended partitioning of images into training, validation, testing sets.
Images 1-162770 are for training, 162771- 182637 are validation, 182638-202599 are used for testing

list_bbox_celeba.csv: Bounding box information for each image. "x_1" and "y_1" represent the upper left point coordinate of bounding box. "width" and "height" represent the width and height of bounding box

list_landmarks_align_celeba.csv: Image landmarks and their respective coordinates.

There are 5 landmarks: left eye, right eye, nose, left mouth, right mouth

list_attr_celeba.csv: Attribute labels for each image.

There are 40 attributes. "1" represents positive while "-1" represents negative

The GAN model was trained for over 1000 epochs in GOOGLE COLLAB GPU environment. The results generated during various epochs is shown in Figure 5 and Figure 6.

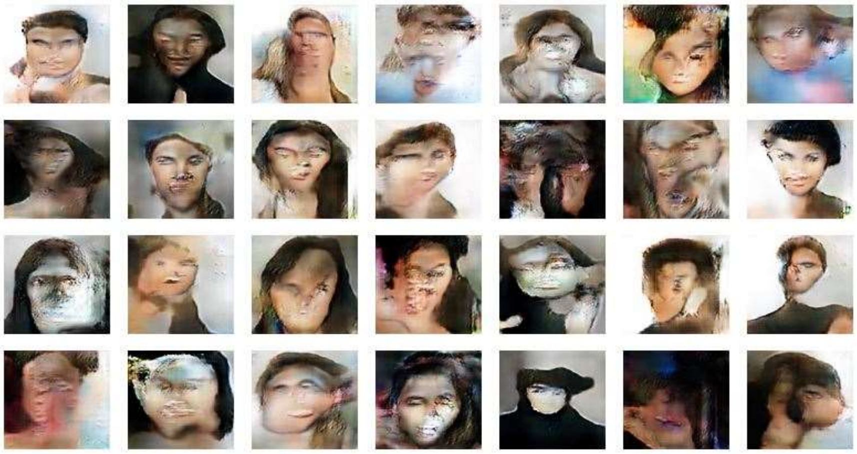
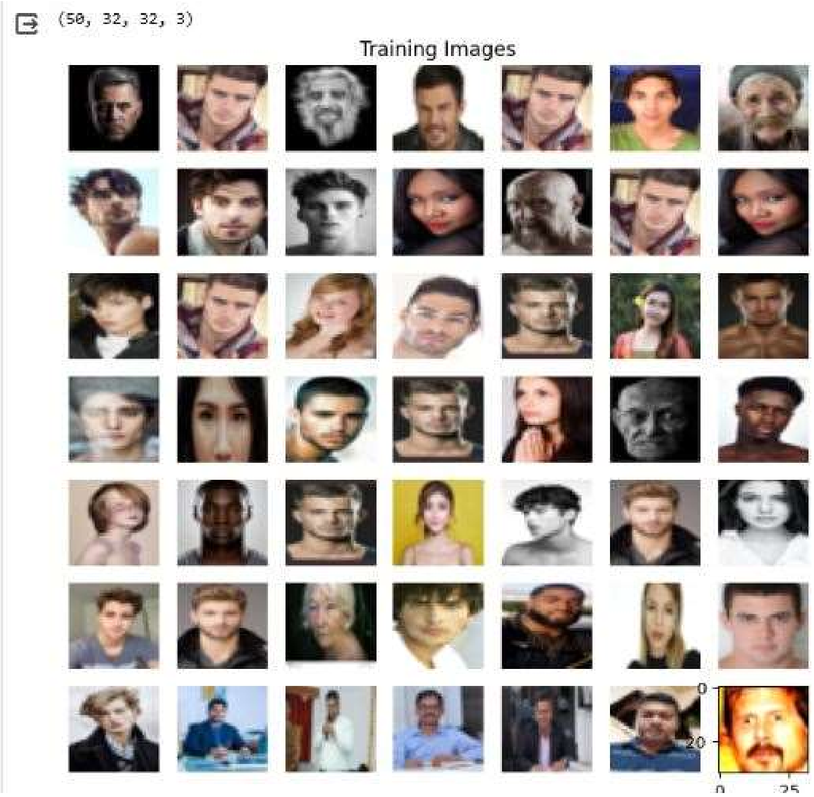


Figure 5 Training images during epoch 10



```
(50, 32, 32, 3)
```

Figure 6 Training images after 1000 epochs

After 1500 epochs, the PROGAN 128 was tested on custom data and it could generate synthetic images for our custom data also. Once the GAN model is trained the output image is processed using Python API for background preservation. We have made following customizations to the GAN model for improving the accuracy.

1. Down sampling using strided convolutions i.e. no pooling layers used.
2. In Up sampling, transpose convolution layer is used
3. Leaky Relu activation function and batch normalization applied
4. Gaussian weights used for initialization i.e. mean=0.0 and stddev=0.02
5. Adam Stochastic Gradient Descent with learning rate=0.0002
6. TANH in the output of generator is used for scaling of images in the range of (-1,1)

For the evaluation of each image of the previous models generated and customized GAN model generated will be done by using SSIM.SSIM stands for Structural Similarity Index Measure. It's a metric used to measure the similarity between two images. SSIM compares the structural information of images rather than just pixel values. It's often used in image processing to evaluate the quality of reconstructed or generated images compared to the originals.

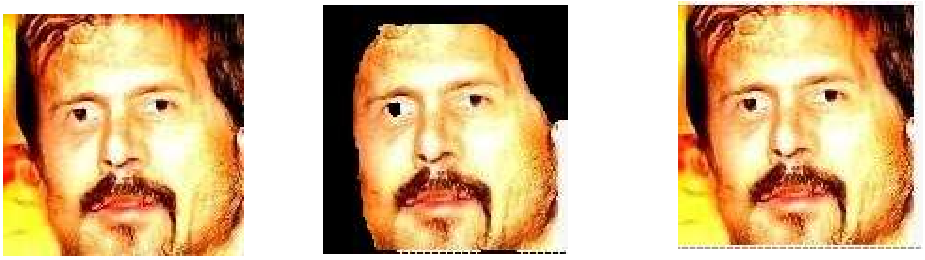
The SSIM index quantifies the similarity between two images based on three components:

1.Brightness (Luminance): Measures the similarity in the overall brightness between the images.

2.Contrast: Measures the similarity in the contrast between the images.

3.Structure: Evaluates the similarity in the structures or patterns present in the images. The SSIM index ranges from -1 to 1, where 1 indicates identical images and -1 indicates completely dissimilar images.

Image Transformation: Finally, the generated image will be converted into bitmap format because the bitmap format does not contain any metadata.



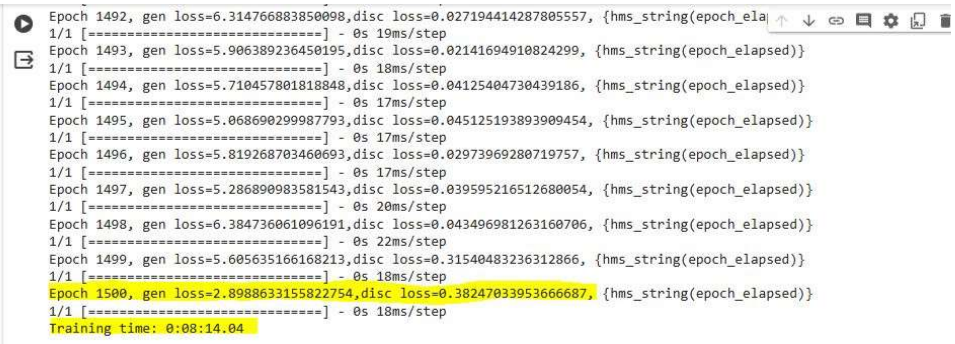
a. original image b. PROGAN 128 image c. customized GAN with background preservation

Figure 7 Images generated by PROGAN 128 and Customized GAN

Figure 7b describes the output generated using PROGAN 128. The background of the original image was not retained and 7c describes the output of the synthetic image generated by our proposed solution with background preservation.

The image generated in 7c is converted into bitmap format because the bitmap format does not contain any metadata. In all other image formats the metadata will also be uploaded along with the image and most of the time, users are not even aware of it. The metadata include information like the device used to capture image, IP address, location, time etc. which is called Exchangeable Information Format (EXIF) data. The EXIF data can also disclose some important information about the user without the user consent and user knowledge. Hence the metadata should be prevented from uploading to public domains along with the image. The common image formats used are .jpeg, .png, .gif etc. all carry the metadata. The bitmap format does not enclose any metadata to the image and hence the metadata can be prevented from disclosure.

The output images are generated in the google drive mounted to the google collab account. The accuracy and loss of both generator and discriminator is described in Figure 8.



```
Epoch 1492, gen loss=6.314766883850098,disc loss=0.027194414287805557, {hms_string(epoch_ela
1/1 [==============================] - 0s 19ms/step
Epoch 1493, gen loss=5.906389236450195,disc loss=0.02141694910824299, {hms_string(epoch_elapsed)}
1/1 [==============================] - 0s 18ms/step
Epoch 1494, gen loss=5.710457801818848,disc loss=0.04125404730439186, {hms_string(epoch_elapsed)}
1/1 [==============================] - 0s 17ms/step
Epoch 1495, gen loss=5.068690299987793,disc loss=0.045125193893909454, {hms_string(epoch_elapsed)}
1/1 [==============================] - 0s 17ms/step
Epoch 1496, gen loss=5.819268703460693,disc loss=0.02973969280719757, {hms_string(epoch_elapsed)}
1/1 [==============================] - 0s 17ms/step
Epoch 1497, gen loss=5.286890983581543,disc loss=0.039595216512680054, {hms_string(epoch_elapsed)}
1/1 [==============================] - 0s 20ms/step
Epoch 1498, gen loss=6.384736061096191,disc loss=0.043496981263160706, {hms_string(epoch_elapsed)}
1/1 [==============================] - 0s 22ms/step
Epoch 1499, gen loss=5.605635166168213,disc loss=0.31540483236312866, {hms_string(epoch_elapsed)}
1/1 [==============================] - 0s 18ms/step
Epoch 1500, gen loss=2.8988633155822754,disc loss=0.38247033953666687, {hms_string(epoch_elapsed)}
1/1 [==============================] - 0s 18ms/step
Training time: 0:08:14.04
```

Figure 8 loss of generator and discriminator after 1500 epochs

Results and Discussions

The existing literature proves that the conventional image obfuscation methods like blurring, pixelation, masking and deidentification do not offer better utility of images and synthetic image creation using GAN has been proved to be better than conventional methods. In our proposed solution we developed a customized GAN based model with improved performance and image quality with respect to privacy preservation and utility. Our proposed solution is evaluated using three evaluation metrics.

Evaluation metrics used are:

1. Inception score (IS): It measures both the quality and diversity of generated images. A higher Inception Score suggests better quality and diversity.
2. Frechet Inception Distance (FID): evaluates the similarity between the distribution of generated and real images. Lower FID values indicate better performance.
3. Structural Similarity Index Measure (SSIM): compares images based on structural similarity.

Table 1 depicts the evaluation of customized GAN, pre trained GAN based on the three evaluation metrics.

Evaluation metric	PROGAN 128	DC GAN	Customized GAN (proposed solution)
IS	0.68	0.6	1.0

SSIM	98 %	97.6 %	98.3 %

Table 1 Comparison of pre trained GAN with customized GAN (proposed solution)

FID: FID was computed on 4 images viz. pixeled image, masked image, blurred image and synthetic image generated through customized GAN. Lower FID found for synthetic image when compared with other images.

Conclusions

The implementation of customized Generative Adversarial Networks (GANs) with bitmap format for image privacy preservation has proven to be a transformative approach, effectively balancing the crucial aspects of privacy protection and utility enhancement. By harnessing the power of GANs, we have successfully demonstrated the capability to generate synthetic images that not only preserve the privacy of individuals but also contribute to the improvement of utility in various applications. However, faceless recognition capability of GAN and deep fake applications of GAN can be misused and will lead to cyberbullying. Recent incidents of such cases were reported where the face of a celebrity is morphed with a nude photograph, using deep fake GAN. Hence there is scope in further research using GAN and its application to enhance privacy in image data with equal emphasis to utility of the image data.

REFERENCES:

1. Dhar, Tribikram, et al. "Challenges of Deep Learning in Medical Image Analysis—Improving Explainability and Trust." IEEE Transactions on Technology and Society 4.1 (2023): 68-75.

2.Tekli, Jimmy, et al. "A framework for evaluating image obfuscation under deep learning-assisted privacy attacks." Multimedia Tools and Applications (2023): 1-33.

3.Chen, Zhenfei, et al. "Privacy preservation for image data: a gan-based method." International Journal of Intelligent Systems 36.4 (2021): 1668-1685.

4.Khojasteh, Mohammad Hossein, Nastaran Moradzadeh Farid, and Ahmad Nickabadi. "GMFIM: A generative mask-guided facial image manipulation model for privacy preservation." Computers & Graphics 112 (2023): 81-91.

5. Samee, Nagwan Abdel, et al. "Safeguarding Online Spaces: A Powerful Fusion of Federated Learning, Word Embeddings, and Emotional Features for Cyberbullying Detection." IEEE Access (2023).

6. https://localiq.com/blog/what-happens-in-an-internet-minute/

7. Samee, Nagwan Abdel, et al. "Safeguarding Online Spaces: A Powerful Fusion of Federated Learning, Word Embeddings, and Emotional Features for Cyberbullying Detection." IEEE Access (2023).

8. Conti, Mauro, Luca Pajola, and Pier Paolo Tricomi. "Turning captchas against humanity: Captcha-based attacks in online social media." Online Social Networks and Media 36 (2023): 100252.

9. Yu, J.; Xue, H.; Liu, B.; Wang, Y.; Zhu, S.; Ding, M. GAN-Based Differential Private Image Privacy Protection Framework for the Internet of Multimedia Things. Sensors 2021, 21, 58. https://doi.org/10.3390/s21010058

10 Yao, Zhexin, et al. "PPUP-GAN: A GAN-based privacy-protecting method for aerial photography." Future Generation Computer Systems 145 (2023): 284-292.