IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 May 2022, accepted 16 June 2022, date of publication 22 June 2022, date of current version 27 June 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3185206

Privacy-Preserving Deep Learning With Learnable Image Encryption on Medical Images

QI-XIAN HUANG[1], WAI LEONG YAP[1], MIN-YI CHIU[1], AND HUNG-MIN SUN[2]

[1]Institute of Information Systems and Applications, National Tsing Hua University, Hsinchu 300, Taiwan
[2]Department of Computer Science, National Tsing Hua University, Hsinchu 300, Taiwan

Corresponding author: Hung-Min Sun (hmsun@cs.nthu.edu.tw)

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Project MOST 110-2221-E-007-040-MY3; and in part by Senao Network Corporation.

ABSTRACT The need for cloud servers for training deep neural network (DNN) models is increasing as more complex architecture designs of DNN models are developed. Nevertheless, cloud servers are considered semi-honest. With great attention to the privacy issues of medical diagnoses using a DNN, previous studies have proposed the idea of learnable image encryption. Though some methods have been presented to partially attack previous encryption schemes, there is still some space for improvement. We proposed a learnable image encryption scheme that is an enhanced version of previous methods and can be used to train a great DNN model and simultaneously keep the privacy of training images. We conducted an experiment on medical datasets from open sources and the result demonstrates the effectiveness of our proposed method in performance and privacy-preserving.

INDEX TERMS Deep neural network, learnable image encryption, medical analysis, privacy-preserving.

I. INTRODUCTION

Due to the great success achieved by Deep Neural Networks (DNNs) in various aspects, such as computer vision, voice recognition, natural language processing, DNNs have recently been used in medical fields, which help the healthcare industries in image diagnosis. For example, IBM Watson has entered the imaging domain after their successful acquisition of Merge Healthcare..[1] Google DeepMind Health and National Health Service in the UK have signed an agreement to process the medical data of 1 million patients [1].

As more complex architecture designs of DNN models have been developed to enhance the performance, more computing resources are needed to train such models. The cloud server can be one of the key solutions in providing large computing capacity to train those complicated DNN models. Unfortunately, the cloud server itself is considered semi-honest since people who are accessible to the server can do whatever to the images that have been sent to the server.

Meanwhile, despite the data is increasing day by day and posing threats to privacy violation, a few regulations have been issued, such as the Health Insurance Portability and Accountability Act of 1996 (HIPAA), which provides legal rights to patients for protecting their medical records. Therefore, sharing of medical data is severely complex and challenging compared to other datasets. In this case, we are inspired to find a solution to address this problem; that is, to invent a method that can send data to the cloud server for the training of models while still keeping the privacy of data.

Such problems bring out the idea of learnable image encryption, which aims to provide an encryption scheme so that the DNN models can directly learn from encrypted images without decrypting them in advance. Recently, Tanaka [2] and Sirichotedumrong *et al.* [3], [4] have proposed state-of-the-art encryption schemes (named as Tanaka scheme and the SKK scheme respectively) for privacy-preserving deep learning, and these schemes can be used in the medical analysis in which DNNs are used.

Since such schemes exist, there are some studies focused on the security evaluation of learnable encryption. For example, [5], [6] evaluate the robustness of the SKK scheme against ciphertext-only attacks (COAs). In particular, methods from [7] have succeeded in partially attacking the images encrypted by the Tanaka scheme and SKK scheme. The two proposed methods, leading bit attack and minimum difference attack, can recover enough features of the images

The associate editor coordinating the review of this manuscript and approving it for publication was Rongbo Zhu.

[1] https://www.techrepublic.com/article/ibm-watson-bets-1-billion-on-healthcare-with-merge-acquisition/

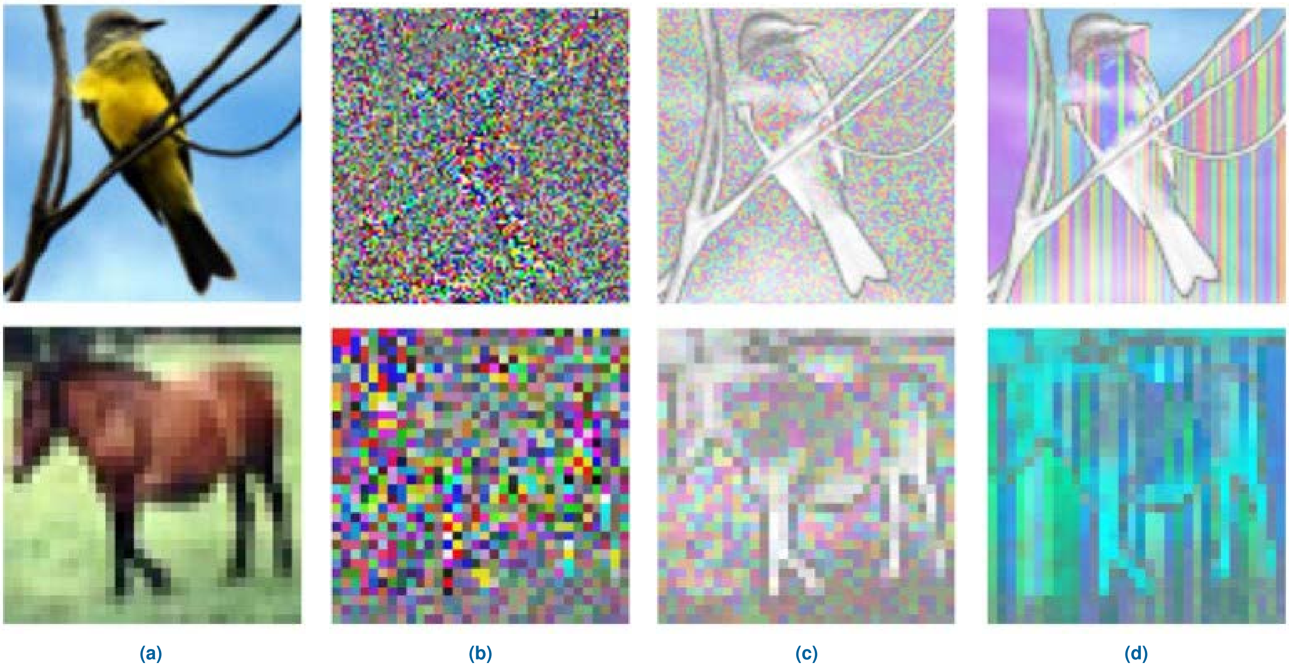

FIGURE 1. (a) Example of 8-bit images with size 32 × 32 from the CIFAR10 dataset; (b) The results after undergoing the SKK scheme encryption; (c) and (d) showing the results after applying leading bit attack and minimum difference attack on (b), respectively.

encrypted with the SKK scheme. Figure 1 demonstrates some of the examples using the methods to attack the images that are encrypted with the SKK scheme.

We proposed an improved version of the SKK scheme, which adds up some of the image statistical smoothing techniques to the previous method and applies such encryption methods to the medical datasets. The results showed that without dropping much accuracy, our proposed scheme is better in protecting the information of images from the stated attacks than using the original SKK scheme. To sum up, our contributions are as follows:

1) We designed the learnable image encryption which can applied to medical images for the purpose of training privacy-preserving deep neural network models in medical fields.
2) We proposed an improved version of the SKK image encryption scheme, which provides a parameter to control the trade-off between the safety of images and the performance.
3) We have also examined the learned features of the privacy-preserving deep neural network model with learnable image encryption via GradCAM [8].

The rest of this paper is organized as follows. Section II recaps some related works in machine learning used in medical fields, medical image encryption, and learnable image encryption. Section III depicts our methodology and approach. Section IV evaluates the performance of our experiment. All the results and conclusions are summarized in Section V. Finally, the recommendations for future work are given in Section VI

II. RELATED WORK

The literature related to our work is recapped in this section. Our work draws on recent researches based on machine learning applied in medical fields, medical image encryption, and learnable image encryption.

A. MACHINE LEARNING IN MEDICAL FIELDS

Wang *et al.* [9] have conducted the detection of breast cancer in digital mammography using traditional machine learning (ML) on the dataset from Tumor Hospital of LiaoNing Province. Two ML approaches are involved: the single-layered neural network (ELM) and the traditional support vector machine (SVM). Although this study did not apply a DNN based approach, it paved the way for the idea to use deep learning models to execute automatic breast cancer detection.

Shen *et al.* [10] have employed the DCNN on mammographic images to make breast cancer detection better. They used the dataset from CBIS-DDSM [11], consisting of 2478 mammography images, and trained it with Resnet-50 and VGG-16. With the use of spatial attention module, CBAM. [12], in the ResNeSt [13], Zhang *et al.* [14] introduced an attention guided deep convolution neural network, ResNetSAt to detect the brain tumor from the new brain MR dataset provided by Ruijin Hospital, Shanghai Jiao Tong University School of Medicine.

Akilan [15] proposed a Deep Convolutional Neural Network with slow feature learning strategy for the diagnosis of COVID19 using the chest X-rays datasets. However, if the models are too complex, a cloud server will be needed to solve the bottleneck of computation of physical machines. On the other side, this can lead to another issue because once the medical images have been uploaded to the server, people who have the privilege may have the full control of the images.

B. MEDICAL IMAGE ENCRYPTION

An effective encryption method for DICOM medical images using Advanced Encryption Standard (AES) proposed by Natsheh *et al.* [16] can narrow the time used to encrypt and decrypt these images.

With the characteristic of pseudorandomness, ergodicity, and initial value sensitivity in chaotic maps, the generated sequences by the chaotic maps consist of great characteristics of security keys. Kanso *et al.* [17] proposed a selective chaos-based image encryption scheme that suitable for medical images. This approach includes several rounds and each round consists of two block-based phases: shuffling phase and masking phase. Chaotic cat maps are used to shuffle and mask an input image. Chong *et al.* [18] presented a chaos-based encryption scheme for medical images. To protect the images, this scheme uses a substitution mechanism in the permutation process through a bit-level shuffling algorithm.

Ding *et al.* [19] proposed a deep neural network, called as DeepEDN to do the encryption and decryption of the medical images. With the Cycle-Generative Adversarial Network (Cycle-GAN) is adopted as the main learning network, the medical images are transferred from the plain domain into the target domain, where the target domain is used to guide the learning model for encryption. The decryption process is executed through a reconstruction network. Instead of decrypting the whole images, a region of interest (ROI)-mining network is also employed to extract the objects of interest from the encrypted image.

However, the approaches mentioned above are not suitable when it is required to send the encrypted image to the cloud server for a DNN model training. If the decryption is needed to be done before inputting to the DNN model, those who are accessible to the server can look through all the original images. To our best knowledge, there is no study yet on investigating whether or not the DNN model can learn from the images encrypted by above-mentioned approaches.

C. LEARNABLE IMAGE ENCRYPTION

As dictated in the introduction, Tanaka [2] proposed the state-of-the-art learnable image encryption scheme (named as Tanaka scheme) for the privacy-preserving deep neural networks. This scheme aims to encrypt the images only for humans instead of machines. Therefore, the encrypted images can be straightly learned by the network.

Sirichotedumrong *et al.* [3], [4] presented another state-of-the-art learnable image encryption scheme (named as SKK scheme). Instead of using the same encryption keys as the Tanaka scheme, the SKK scheme uses independent encryption keys to decide whether the corresponding pixel is applied with negative-positive transformation and color shuffling operation. The use of separate keys improves the robustness against DNN-based attacks and provides a large keyspace that is robust against brute-force attacks. Moreover, they also found that models trained with the images encrypted by the SKK scheme can maintain the accuracies under the use of data augmentation.

Sirichotedumrong *et al.* [20] recently proposed an image transformation scheme for privacy-preserving deep neural networks that using the Generative Adversarial Networks (GANs), proving that the need to manage encryption keys no longer existed.

Several studies have focused on the security evaluation of learnable image encryption [5], [6], which evaluates the robustness of the learnable image encryption scheme against ciphertext-only attacks (COAs). In particular, the study in [7] proposed one attack algorithm that can completely decrypt the images encrypted with the Tanaka scheme, and two attack algorithms that have the ability to restore a certain amount of features of the image encrypted with the SKK scheme to image in the original and grayscale formats respectively. Figure 1 shows some examples of using methods to attack the images that are encrypted by the SKK scheme.

Enlightened by the SKK scheme, we therefore have articulated a learnable image encryption under the medical images for privacy-preserving DNNs and provided some improvements for the SKK scheme by adding the image smoothing techniques.

III. METHODOLOGY

Our method is depicted thoroughly in this section. First, the formulation and validation of our proposed image encryption are explained. Then, the implementation details and procedure are provided.

A. PROPOSED IMAGE ENCRYPTION

1) SKK SCHEME WITH STATISTICAL SMOOTHING

Our proposed encryption scheme is enlightened by the original SKK scheme [3], [4]; thus, the negative-positive transformation and color shuffling operation are included in our proposed scheme. After completing all the steps in the original SKK scheme, we added up the process of statistical smoothing into the images. We used the four most commonly used statistical smoothing filters in our method: median filter, mean filter, maximum filter, and minimum filter, which aim to fill all elements of a block with its median, mean, maximum, and minimum values respectively. Figure. 2 demonstrates the idea of applying a maximum filter on a 3 × 3 sized block. The procedure can be summarized as follows:

1) Divide image I_{RGB} with $U \times V$ pixels into individual pixels.
2) The negative-positive transformation is applied to each channel of each pixel, p_R, p_G, p_B, with a random binary integer generated by secret keys $K_{np} = \{K_{np,R}, K_{np,G}, K_{np,B}\}$. Hence, the value of the i-th pixel,

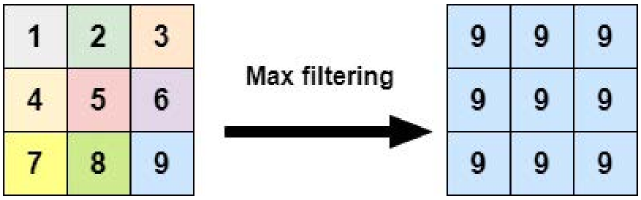

FIGURE 2. The idea of applying a max filter on 3 × 3 block.

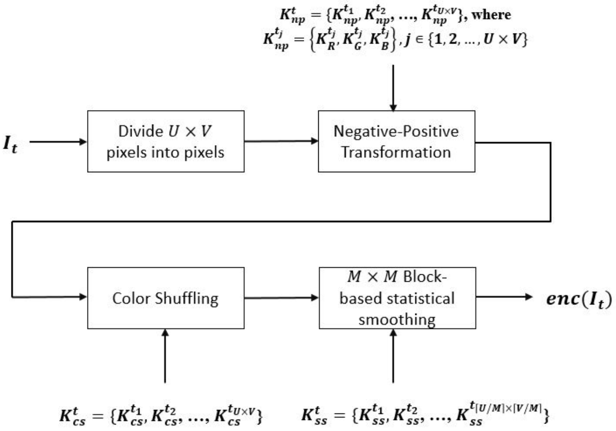
$K_{np}^{t} = \{K_{np}^{t_1}, K_{np}^{t_2}, \ldots, K_{np}^{t_{U\times V}}\}$, where
$K_{np}^{t_j} = \{K_R^{t_j}, K_G^{t_j}, K_B^{t_j}\}, j \in \{1, 2, \ldots, U \times V\}$
I_t
Divide $U \times V$ pixels into pixels
Negative-Positive Transformation
Color Shuffling
$M \times M$ Block-based statistical smoothing
$enc(I_t)$
$K_{cs}^{t} = \{K_{cs}^{t_1}, K_{cs}^{t_2}, \ldots, K_{cs}^{t_{U\times V}}\}$
$K_{ss}^{t} = \{K_{ss}^{t_1}, K_{ss}^{t_2}, \ldots, K_{ss}^{t_{\lceil U/M\rceil \times \lceil V/M\rceil}}\}$

FIGURE 3. The overall process of our proposed scheme.

TABLE 1. Six different integers correspond to a unique possible channel shuffling permutation.

The secret key of color shuffle, K_{cs}	location of channels		
	p_R	p_G	p_B
0	p_R	p_G	p_B
1	p_R	p_B	p_G
2	p_G	p_R	p_B
3	p_G	p_B	p_R
4	p_B	p_R	p_G
5	p_B	p_G	p_R

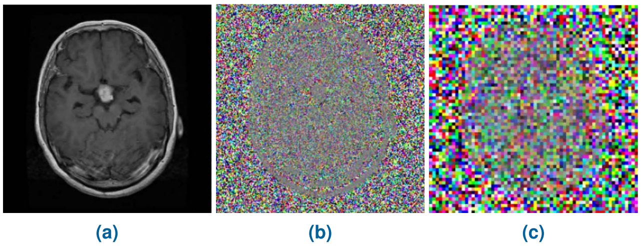

FIGURE 4. (a) One of the images from Magnetic Resonance Imaging(MRI) datasets; (b) The encrypted images with original SKK scheme; (c) The encrypted images with our proposed scheme.

p_c, is calculated using:

$$p_c = \begin{cases} p_c \oplus (2^L - 1), & \text{if } K_{np,c} = 1 \\ p_c, & \text{otherwise} \end{cases}$$

where $c \in \{R, G, B\}$ and L represents the number of bits of the image. The occurrence probability of $K_{np,c}$ is uniformly distributed, denoting as $P(K_{np,c}) = 0.5, \forall c \in \{R, G, B\}$

3) ***Color Shuffling Process***: The color components of each pixel are shuffled by using a pseudorandom key chosen from six possible integers (0 to 5), in which each of the integers refers to a unique color channel shuffling permutation. Table 1 displays all the possible permutations of color shuffling.

4) After all the steps dictated above have been processed, the image I_{RGB} is split into blocks with sized $M \times M$. For each block, we randomly apply one of the statistical smoothing filters with that size. Those smoothing filters are median, mean, maximum, and minimum filters respectively. The smoothing filters used for each block are not necessarily the same as each other.

The flowchart of our proposed image encryption scheme and its algorithm are shown in both Figure 3 and Algorithm 1. Figure 4 demonstrates an encrypted image by the SKK and our proposed scheme.

2) KEY SPACE

Since the encrypted images can be directly learned from DNN models, there is no need to manage the secret keys used for encryption. Each step in our proposed scheme uses independent security keys, which make the adversaries more difficult to carry out collision attacks as well as known-plaintext attacks (KPAs).

Consider an image I with $U \times V$ pixels is divided into individual pixels, during the negative-positive transformation, three independent keys are generated for each channel belonging to a certain pixel; therefore, the keyspace of negative-positive transformation is given by

$$N_{np}(U, V) = 2^{3UV} \tag{1}$$

While being in the color shuffling process, independent keys are generated from six possible permutations for each pixel; hence, the keyspace of color shuffling is represented by

$$N_{cs}(U, V) = 6^{UV} \tag{2}$$

In the final step, at which the block-based statistical smoothing is employed, image I is split into blocks of size $M \times M$. For each block, a key is generated randomly from 4 possible integers that stand for the choices of smoothing filter to be applied to the block; hence, the keyspace of block-based statistical smoothing is given by

$$N_{ss}(U, V) = 4^{\lceil * \rceil U/M \times \lceil * \rceil V/M} \tag{3}$$

Therefore, the keyspace of images encrypted using our proposed encryption scheme is

$$\begin{aligned} N(U, V) &= N_{np}(U, V) \cdot N_{cs}(U, V) \cdot N_{ss}(U, V) \\ &= 2^{3UV} \cdot 6^{UV} \cdot 4^{\lceil * \rceil U/M \times \lceil * \rceil V/M} \end{aligned} \tag{4}$$

Algorithm 1 Proposed Encryption Scheme

```
Input: $U \times V$ sized image; number of bits $L$
Output: Encrypted Image $I_{enc}$
foreach $p = (u, v) \in I$ do
    procedure Negative-Positive Transformation
        foreach $c \in \{R, G, B\}$ do
            $K_{np,c} \in \{0, 1\}$
            if $K_{np,c} = 1$ then
                $p_c \leftarrow p_c \oplus (2^L - 1)$;

    procedure Color Shuffling
        $K_{cs} \in \{0, 1, 2, 3, 4, 5\}$
        if $K_{cs} = 0$ then
            $p = \{p_R, p_G, p_B\}$
        else if $K_{cs} = 1$ then
            $p = \{p_R, p_B, p_G^t\}$
        else if $K_{cs} = 2$ then
            $p = \{p_G, p_R, p_B\}$
        else if $K_{cs}^t = 3$ then
            $p = \{p_G, p_B, p_R\}$
        else if $K_{cs}^t = 4$ then
            $p = \{p_B, p_R, p_G\}$
        else
            $p = \{p_B, p_G, p_R\}$

procedure Block-based statistical smoothing
    Divide $I$ into blocks with sized $M \times M$
    foreach block $b$ do
        $K_{ss}^b \in \{0, 1, 2, 3\}$
        if $K_{ss}^b = 0$ then
            $value \leftarrow average(b)$
        else if $K_{ss}^b = 1$ then
            $value \leftarrow median(b)$
        else if $K_{ss}^b = 2$ then
            $value \leftarrow max(b)$
        else
            $value \leftarrow min(b)$
        foreach $elem \in b$ do
            $elem \leftarrow value$
Return $I$
```

For example, an image with a size of 256 × 256 and $M = 4$ used in our experiment has the keyspace of an encrypted image calculated by

$$N(256, 256) = 2^{3*256*256} \cdot 6^{256*256} \cdot 4^{64*64}$$

which is 4^{64*64} times more than that of the keyspace of the image encrypted by the original SKK scheme.

In addition, the choices of smoothing filters do not need to be the same as filters stated above. One could also change the number of choices by adding up more filter types, such as interquartile range filters and 3$^{\text{rd}}$ quartile filters that search for the interquartile range and 3$^{\text{rd}}$ quartile of elements of the corresponding block respectively.

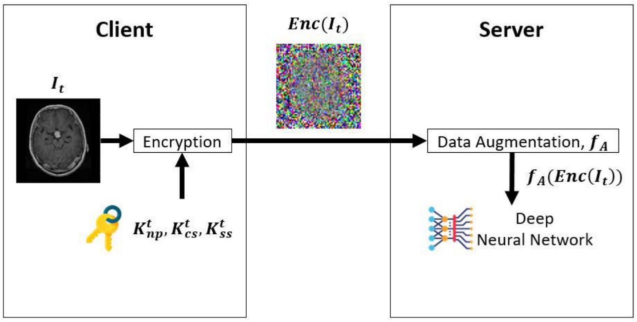

FIGURE 5. Implementation of our experiment.

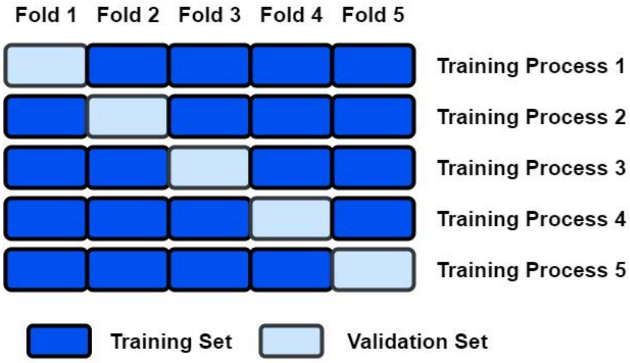

FIGURE 6. The training process of our experiment with cross-validation (kfold = 5).

Due to the large keyspace, our proposed image encryption scheme can be robust to stand the brute force attacks.

B. MODEL IMPLEMENTATION

Figure 5 shows the procedure of implementation of our experiment. The images are first encrypted on the client side with our proposed scheme and then those encrypted images will be sent to the server to execute the data augmentation and finally be used as the input to the DNN model.

Two models, DenseNet-121 [21] and XceptionNet [22], have been used in our experiment. The images are resized into 256 × 256 before inputting to the models with batch size set to 35 and the optimizers set to the stochastic gradient descent(SGD) with the momentum of 0.9 respectively. During the training process, we schedule the learning rate to start with 0.1, then decrease it by a factor of 10 for every 10 epochs run and maintain it to 1e-5 after 25 epochs. In addition, the training process is running 5 times with applying cross-validation (5 folds in total, as shown in Figure 6) on the training dataset. The data augmentation techniques have also been applied to the training dataset with width shifting, height shifting, and horizontal flipping.

IV. EXPERIMENTAL RESULT

In this section, we evaluate our proposed method and compare it with the existing SKK scheme on two open-source datasets, including the MRI brain tumor dataset and COVID19 dataset.

A. DATASETS

There are two types of the dataset from open sources have been selected for our experiment, of which the details are shown as follows:

1) MAGNETIC RESONANCE IMAGING(MRI) BRAIN TUMOR DATASETS

Three MRI brain tumor datasets are combined, which are:

1) The dataset [2] contains 3064 T1-weighted contrast-enhanced images from 233 patients with three kinds of brain tumors: meningioma, glioma, and pituitary tumor. Since all of these images are 16-bits images and they are in mat files, we first convert them into 8-bits images, then equalize the histogram of the images and save them into jpg files.
2) The dataset [3] from GitHub contains 3264 images with four classes—no tumor, meningioma, glioma, and pituitary tumor.
3) The dataset [4] from Kaggle contains 98 normal images and 155 images with a brain tumor.

We mixed 98 normal images from dataset [4] (labeled as no tumor) with all data in dataset [2] and dataset [3], accumulating 6426 images in total. The mixed dataset is then split into training and testing datasets with a ratio of 80:20.

2) COVID-19 X RAY DATASET

This dataset [5] consists of the samples of COVID-19 retrieved from different sources, which contains 6939 X-ray images with three classes: COVID, normal, and pneumonia. The dataset is then divided into 5898 training samples and 1041 testing samples.

During the training process, the training samples above is split in a ratio of 80:20 (4112 training and 1028 validation images for brain tumor dataset, 4718 training and 1180 validation samples for COVID-19 dataset) in each cross-validation with 5-folds, and the whole process is shown in Figure 6. Figure 7 and Figure 8 present some image samples in the MRI datasets and COVID-19 dataset we used, respectively.

B. PERFORMANCE

The performance metric used in this experiment is the F1 score, which can be represented by

$$precision = \frac{true\ positive}{true\ positive + false\ positive}, \tag{5}$$

$$recall = \frac{true\ positive}{true\ positive + false\ negative}, \tag{6}$$

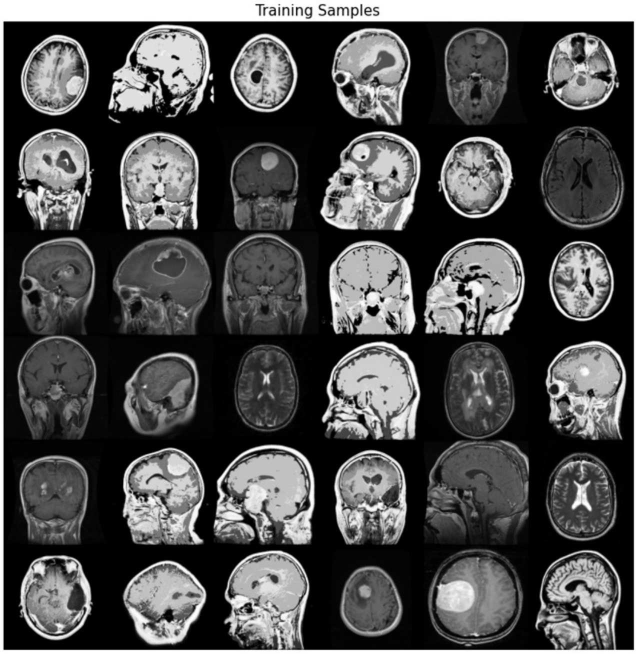

FIGURE 7. The sample images in the brain tumor MRI dataset.

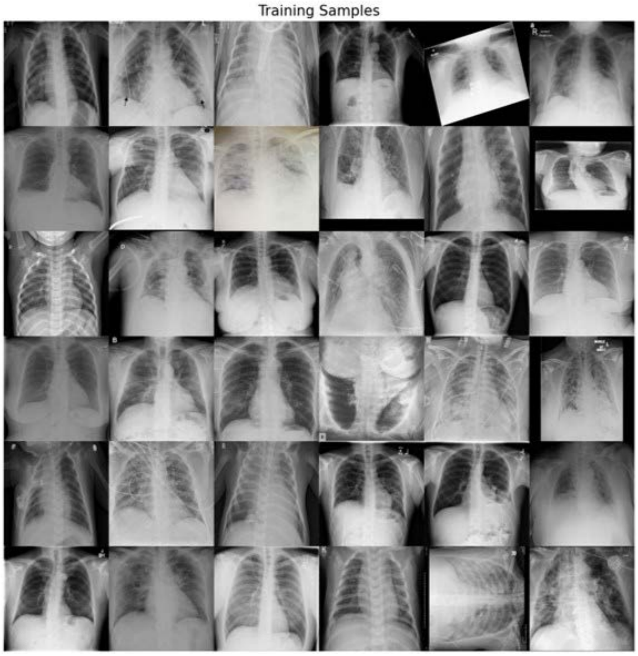

FIGURE 8. The sample images in the COVID-19 X-ray dataset.

$$f1\ score = \frac{2 \times precision \times recall}{precision + recall} \tag{7}$$

As shown in Table 2, for the MRI brain tumor images, the F1 scores achieved by DenseNet and XceptionNet using plain images are similar to each other and are exceeding 98%. While DenseNet and XceptionNet trained with the SKK scheme encrypted images achieved 93.36% and 91.42% respectively, which are dropped around 5% ∼ 7% from

[2] https://figshare.com/articles/dataset/brain_tumor_dataset/1512427?file=7953679

[3] https://github.com/sartajbhuvaji/brain-tumor-classification-dataset

[4] https://www.kaggle.com/navoneel/brain-mri-images-for-brain-tumor-detection

[5] https://www.kaggle.com/amanullahasraf/covid19-pneumonia-normal-chest-xray-pa-dataset

models trained with plain images. By using the images encrypted by our proposed scheme, DenseNet and XceptionNet are able to achieve 89.30% and 88.96% respectively, which are 3% less than that of using the SKK scheme.

On the other hand, similar results are achieved when training with X-ray COVID 19 dataset. Both DenseNet and XceptionNet trained with plain images are exceeding 96%. The F1 scores are slightly decreased when applying SKK scheme, that is, 95.01% for DenseNet and 94.63% for XceptionNet. While training with the images encrypted by our proposed scheme, DenseNet and Xception achieved 94.71% and 93.66% separately, which are only 1% less than the F1 scores achieved using the SKK scheme.

Note that different filter sizes (size of blocks) are applied on different datasets, block size of 4 × 4 is applied in the MRI images and block with sized 6 × 6 are used during the encryption of X-ray images using our proposed scheme.

C. SECURITY EVALUATION

1) ATTACK ALGORITHMS

As described in the first section, the leading bit attack and minimum difference attack proposed in [7] can be used to reconstruct the encrypted images into original images. Therefore, the robustness against the methods is discussed in this section.

Before comparing the results of applying the attacks to the SKK and our proposed scheme encrypted images, each attack method proposed in [7] is dictated below:

1) ***Leading bit attack***: Authors of [7] observed the fact that the areas between edges have similar color component values, which implies that the gradient magnitude at each pixel is close to the minimum. In addition, the combined magnitude of the color components of each pixel does not change during the color shuffling process, so they are only concerned with the negative-positive transformation.

   To reconstruct a grayscale version of the original image that is recognizable from an encrypted image using the SKK scheme, the leading bit attack simply guesses between 0 or 1 as the leading bit of each color component and changes all color components of every pixel to have the same leading bit. The whole algorithm of the leading bit attack is represented in Algorithm 2.

2) ***Minimum difference attack***: Authors of [7] present that minimizing the change in color component values of pixels can be an effective way to recover the original image from an encrypted one using the SKK scheme while minimizing the quantity:

   $$\sum_{c\in\{R,G,B\}} \mid q_c - p_c \mid \qquad (8)$$

   where p is a pixel that has not yet been decrypted, and q is a nearby pixel that has been decrypted. Therefore, for every pixel p in the encrypted image, there are $6 \times 2^3 = 48$ options of possible permutations of color shuffling and negative-positive transformation. The attack method calculates the numbers of each option and assigns the option that minimizes the quantity to the decrypted version of p. The whole algorithm steps are given in Algorithm 3.

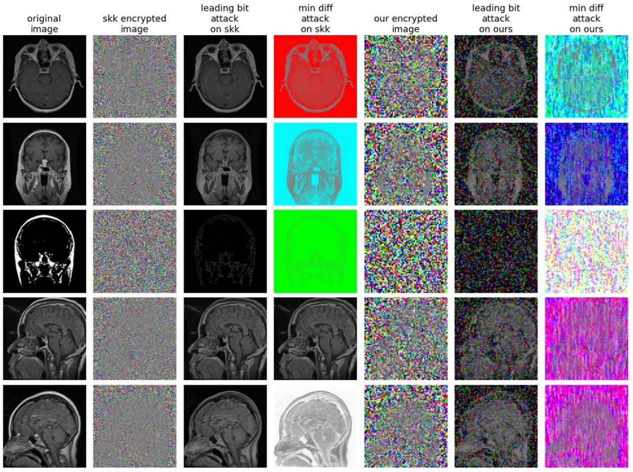

FIGURE 9. The results of applying leading bit attack and minimum difference attack on the encrypted MRI image with SKK scheme and our proposed scheme.

Algorithm 2 Leading Bit Attack

```
Input: U × V sized encrypted image $I_{enc}$; number of bits $L$; leading bit $b \in \{0, 1\}$
Output: Attacked grayscale image $I_{attack}$
foreach $p = (u, v) \in I_{enc}$ do
    foreach $c \in \{R, G, B\}$ do
        if $\lfloor p_c/2^{L-1} \neq b \rfloor$ then
            $p_c \leftarrow p_c \oplus (2^L - 1)$;
```

Algorithm 3 Minimum Difference Attack

```
Input: Encrypted image $I_{enc}$ of size $U \times V$; number of bits $L$
Output: Attacked image $I_{attack}$
foreach $p = (u, v) \in I_{enc}$ do
    Choose nearby pixel $q$;
    p_min ← $p$;
    diff_min ← $\sum_{c\in\{R,G,B\}} |q_c - p_c|$;
    foreach option $p*$ of $p$ do        ▷ Each neg-pos and shuffle permutation
        if $\sum_{c\in\{R,G,B\}} |q_c - p*_c| <$ diff_min then
            p_min ← $p*$;
            diff_min ← $\sum_{c\in\{R,G,B\}} |q_c - p*_c|$;
    $p$ ← p_min;
```

2) COMPARISON OF THE EXPERIMENT RESULT

Judging from our experiment result in Figure 9 and Figure 10, the leading bit attack can recover enough details

TABLE 2. Comparison on F1 scores achieved by models using plain images, encrypted images with SKK scheme, and our proposed encrypted images.

Model	Dataset	Epochs	Filter Size	Plain	SKK	Proposed
DenseNet	Brain Tumor	100	4×4	98.36	93.36	89.30
DenseNet	COVID 19	50	6×6	96.25	95.01	94.71
XceptionNet	Brain Tumor	100	4×4	98.30	91.42	88.96
XceptionNet	COVID 19	50	6×6	96.24	94.63	93.66

TABLE 3. The F1 scores achieved by DenseNet trained with our proposed scheme encrypted images using different filter sizes on different datasets.

Dataset	Filter size	F1 score	Epochs
Brain Tumor	1×1	93.36	100
Brain Tumor	3×3	90.97	100
Brain Tumor	4×4	89.30	100
Brain Tumor	5×5	86.85	100
COVID 19	1×1	95.01	50
COVID 19	5×5	94.84	50
COVID 19	6×6	94.71	50
COVID 19	7×7	94.33	50

of the encrypted images using the SKK scheme. Despite employing some of the SKK schemes, encrypted images are recovered with the wrong color when applying the minimum difference attack, but one can still obtain the details in the recovered images.

Once encrypted with our proposed scheme, the images recovered by both the leading bit attack and the minimum difference attack contain fewer features than those recovered images encrypted by the SKK scheme. For example, the gyrification in the brain and the ribs is unable to recover for the attacks targeted to the images encrypted by our proposed method.

D. LEARNING ABILITY

Figure 11 and Figure 12 demonstrate the original images and their corresponding predictions from three models mentioned above. The region of interest of each model is similar to each other and the results show that despite the smoothing filters have been added into the encrypted images, those images are still learnable for the models.

Additionally, as can be seen in Figure 11 and Figure 12, the model trained with our proposed scheme has a larger region of interest than the other two models trained without our scheme due to the feature representation of the images being smoothed by filters.

E. ABLATION STUDIES

Since our proposed scheme is parameterizable, the filter size used on block-based statistical smoothing can be used to control the trade-off between the accuracy achieved by the model and the safety of the encrypted images.

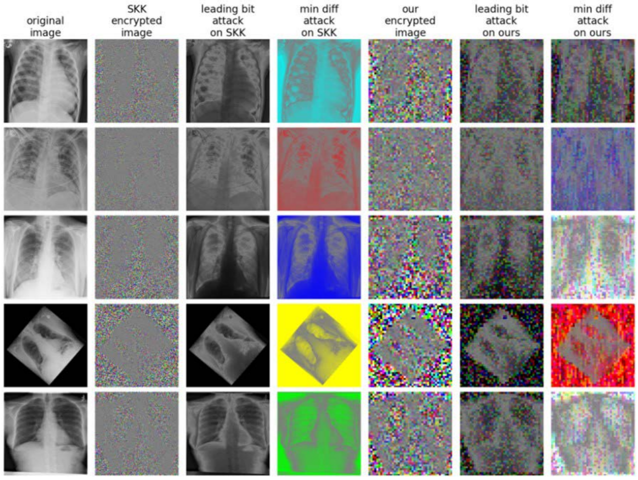

FIGURE 10. The results of applying leading bit attack and minimum difference attack on the encrypted X-ray image with SKK scheme and our proposed scheme.

As shown in Table 3, the smaller the filter size is used, the higher the F1 score will be. Note that the encrypted images are the same as those encrypted with the SKK scheme with a filter size of 1×1.

The larger the filter size is used, the better robustness against the attacks proposed in [7] can be. Figure 13, Figure 14 and Figure 15 present part of the result with

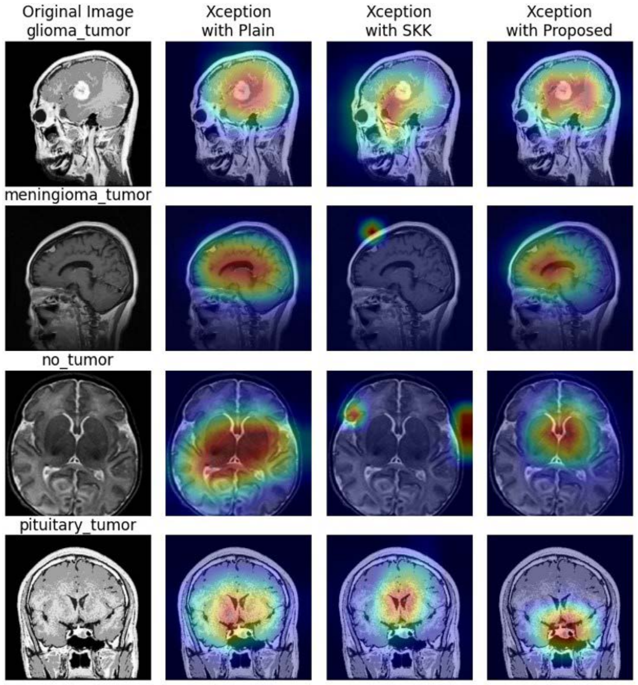

FIGURE 11. The GradCAM Visualization [8] of XceptionNet that trained with plain images, SKK scheme encrypted images, and our proposed scheme encrypted images. (MRI Brain Tumor).

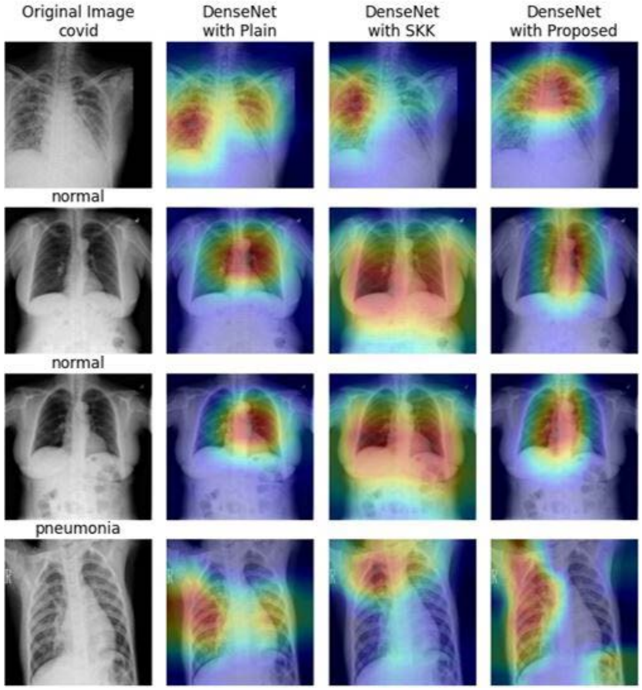

FIGURE 12. The GradCAM Visualization [8] of DenseNet that trained with plain images, SKK scheme encrypted images, and our proposed scheme encrypted images. (COVID-19).

applying different filter sizes to our proposed scheme and the recovered images that have been attacked by both leading bit and minimum difference attacks.

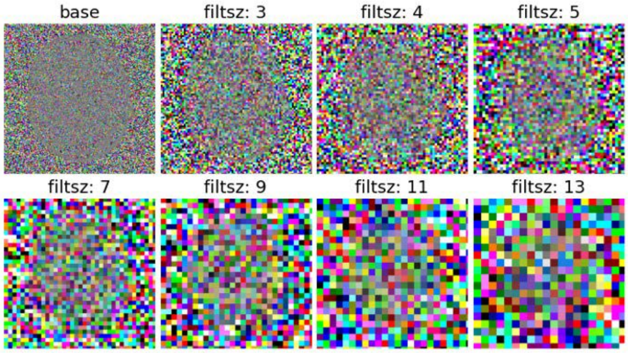

FIGURE 13. The encrypted MRI images with different filter sizes used in our proposed scheme.

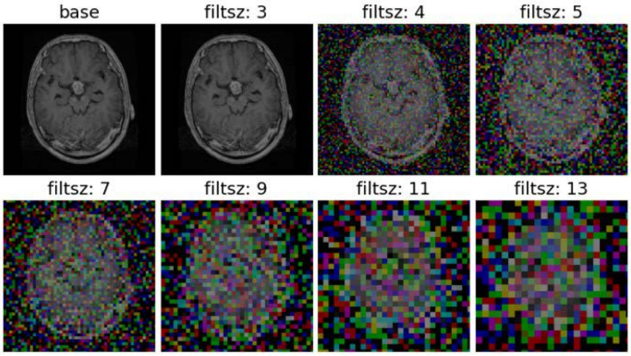

FIGURE 14. The results of applying leading bit in [7] on encrypted MRI images with different filter sizes used in our proposed scheme.

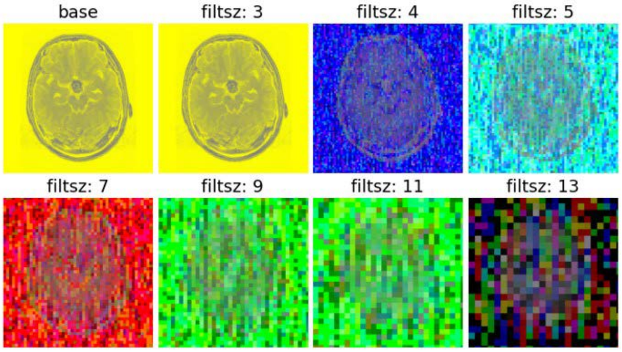

FIGURE 15. The results of applying minimum difference attacks in [7] on encrypted MRI images with different filter sizes used in our proposed scheme.

V. CONCLUSION

In this paper, we have proposed an improved version of the SKK learnable image encryption scheme that is parameterizable and thereby to control the trade-off between accuracy gained and the security of images. This scheme is beneficial for those who want to solve the privacy issue regarding sending images to the cloud for training a complex DNN model that requires large computing resources. Despite the images used in our experiment are all in a square shape, we can also change the size of the filter based on the size of the images that need to be encrypted.

In addition, we also found that this method is not suitable for images that are with low resolution since there is no many features can be smoothed while maintaining accuracy. With the work in [23]–[26], which reconstruct higher resolution images from low resolution images, we argue that these techniques can solve the weaknesses of our method.

With rapid developing in the architecture of the DNN models for various applications, the need for a learnable image encryption scheme will be predictable and increase accordingly.

VI. RECOMMENDATIONS FOR FUTURE WORK

In future works, we can include more smoothing techniques instead of using the stated statistical smoothing methods, such as Gaussian smoothing, Adaptive local noise smoothing, Alpha-trimmed smoothing, etc. In addition, we can apply our proposed method to other field of dataset if possible, such as classified the military satellite images dataset, which requires ultimate privacy while being trained online.

Many studies have centered on the attention module that helps to increase the representation power and focus on the key features of the images. For example, the SE block in [27] presents a squeeze and excitation network using global average pooling to make the DNN model focus on the relationship between channels. In addition, the CBAM block in [12] has used two sub-modules, the channel attention module, and spatial attention module to study ''what'' is meaningful to input images and ''where'' is an informative part. Thus, we suggest that since the above-mentioned attention modules are promising and beneficial in increasing the accuracy with learnable image encryption for privacy-preserving DNN models, future research may concentrate on this subject.

Moreover, we could further use these encrypted images as the base image for training GANs to generate the encrypted image, which could save more time to do encryption. Since there is not much difference in the distribution between generated encrypted images and encrypted images by filters, the result of classification would be similar as well.

ACKNOWLEDGMENT

(Qi-Xian Huang and Wai Leong Yap contributed equally to this work.)

REFERENCES

[1] J. Powles and H. Hodson, ''Google DeepMind and healthcare in an age of algorithms,'' *Health Technol.*, vol. 7, no. 4, pp. 351–367, Dec. 2017.

[2] M. Tanaka, ''Learnable image encryption,'' in *Proc. IEEE Int. Conf. Consum. Electron. Taiwan (ICCE-TW)*, May 2018, pp. 1–2.

[3] W. Sirichotedumrong, Y. Kinoshita, and H. Kiya, ''Pixel-based image encryption without key management for privacy-preserving deep neural networks,'' *IEEE Access*, vol. 7, pp. 177844–177855, 2019.

[4] W. Sirichotedumrong, T. Maekawa, Y. Kinoshita, and H. Kiya, ''Privacy-preserving deep neural networks with pixel-based image encryption considering data augmentation in the encrypted domain,'' in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 674–678.

[5] W. Sirichotedumrong, Y. Kinoshita, and H. Kiya, ''On the security of pixel-based image encryption for privacy-preserving deep neural networks,'' in *Proc. IEEE 8th Global Conf. Consum. Electron. (GCCE)*, Oct. 2019, pp. 121–124.

[6] W. Sirichotedumrong and H. Kiya, ''Visual security evaluation of learnable image encryption methods against ciphertext-only attacks,'' in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)* Dec. 2020, pp. 1304–1309.

[7] A. H. Chang and B. M. Case, ''Attacks on image encryption schemes for privacy-preserving deep neural networks,'' 2020, *arXiv:2004.13263*.

[8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, ''Grad-CAM: Visual explanations from deep networks via gradient-based localization,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

[9] Z. Wang, M. Li, H. Wang, H. Jiang, Y. Yao, H. Zhang, and J. Xin, ''Breast cancer detection using extreme learning machine based on feature fusion with CNN deep features,'' *IEEE Access*, vol. 7, pp. 105146–105158, 2019.

[10] L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, ''Deep learning to improve breast cancer detection on screening mammography,'' *Sci. Rep.*, vol. 9, no. 1, pp. 1–12, Dec. 2019.

[11] R. S. Lee, F. Gimenez, and A. D. H. Rubin, ''Curated breast imaging subset of DDSM,'' Cancer Imag. Arch, Tech. Rep., 2016.

[12] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, ''CBAM: Convolutional block attention module,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[13] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, M. Li, and A. Smola, ''ResNeSt: Split-attention networks,'' 2020, *arXiv:2004.08955*.

[14] Y. Zhang, S. Wang, H. Wu, K. Hu, and S. Ji, ''Brain tumors classification for MR images based on attention guided deep learning model,'' 2021, *arXiv:2104.02331*.

[15] T. Akilan, ''CxSE: Chest X-ray slow encoding CNN forCOVID-19 diagnosis,'' 2021, *arXiv:2106.12157*.

[16] Q. N. Natsheh, B. Li, and A. G. Gale, ''Security of multi-frame DICOM images using XOR encryption approach,'' *Proc. Comput. Sci.*, vol. 90, pp. 175–181, Jan. 2016.

[17] A. Kanso and M. Ghebleh, ''An efficient and robust image encryption scheme for medical applications,'' *Commun. Nonlinear Sci. Numer. Simul.*, vol. 24, nos. 1–3, pp. 98–116, Jul. 2015.

[18] C. Fu, W.-H. Meng, Y.-F. Zhan, Z.-L. Zhu, F. C. M. Lau, C. K. Tse, and H.-F. Ma, ''An efficient and secure medical image protection scheme based on chaotic maps,'' *Comput. Biol. Med.*, vol. 43, no. 8, pp. 1000–1010, Sep. 2013.

[19] Y. Ding, G. Wu, D. Chen, N. Zhang, L. Gong, M. Cao, and Z. Qin, ''DeepEDN: A deep-learning-based image encryption and decryption network for internet of medical things,'' *IEEE Internet Things J.*, vol. 8, no. 3, pp. 1504–1518, Feb. 2021, doi: 10.1109/JIOT.2020.3012452.

[20] W. Sirichotedumrong and H. Kiya, ''A GAN-based image transformation scheme for privacy-preserving deep neural networks,'' in *Proc. 28th Eur. Signal Process. Conf. (EUSIPCO)*, Jan. 2021, pp. 745–749.

[21] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, ''Densely connected convolutional networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[22] F. Chollet, ''Xception: Deep learning with depthwise separable convolutions,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[23] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, ''Deep Laplacian pyramid networks for fast and accurate super-resolution,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 624–632.

[24] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, ''Fast and accurate image super-resolution with deep Laplacian pyramid networks,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2599–2613, Nov. 2018.

[25] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, ''Photo-realistic single image super-resolution using a generative adversarial network,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4681–4690.

[26] Y. Guo, J. Chen, J. Wang, Q. Chen, J. Cao, Z. Deng, Y. Xu, and M. Tan, ''Closed-loop matters: Dual regression networks for single image super-resolution,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5407–5416.

[27] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

QI-XIAN HUANG is currently pursuing the Ph.D. degree with the National Tsing Hua University, Hsinchu, Taiwan. He has worked at Qualcomm Semiconductor Corporation to be a Senior Firmware Engineer, and also worked as a Researcher at the Artificial Intelligence Research Center, National Chengchi University, Taipei, Taiwan. His research interests include deep learning for computer vision algorithm and applications, beyond 5G, and 6G networks.

WAI LEONG YAP received the B.S. degree in applied mathematics from the National Chung Hsing University, Taichung, Taiwan, and the M.S. degree in computer science from the National Tsing Hua University, Hsinchu, Taiwan. His research interests include deep learning for computer vision and image processing techniques and image privacy security.

MIN-YI CHIU is currently pursuing the Ph.D. degree with the Institute of Information Systems and Applications, National Tsing Hua University, Taiwan. His research interests include information security, blockchain applications, system design, network security, and forensics.

HUNG-MIN SUN received the Ph.D. degree in computer science and information engineering from the National Chiao Tung University, Hsinchu, Taiwan, in 1995. He is currently working as a Full Professor with the Department of Computer Science, National Tsing Hua University, Hsinchu. His research interests include network security, cryptography, blockchain, and automatic trading.