ECTI Transactions on Computer and Information Technology
Journal homepage: https://ph01.tci-thaijo.org/index.php/ecticit/
Published by the ECTI Association, Thailand, ISSN: 2286-9131

Block-Wise Encryption for Reliable Vision Transformer models

Hitoshi Kiya[1], Ryota Iijima[2] and Teru Nagamori[3]

ABSTRACT

This article presents block-wise image encryption for the vision transformer and its applications. Perceptual image encryption for deep learning enables us not only to protect the visual information of plain images but to also embed unique features controlled with a key into images and models. However, when using conventional perceptual encryption methods, the performance of models is degraded due to the influence of encryption. In this paper, we focus on block-wise encryption for the vision transformer, and we introduce three applications: privacy-preserving image classification, access control, and the combined use of federated learning and encrypted images. Our scheme can have the same performance as models without any encryption, and it does not require any network modification. It also allows us to easily update the secret key. In experiments, the effectiveness of the scheme is demonstrated in terms of performance degradation and access control on the CIFAR-10 and CIFAR-100 datasets.

Article information:

Keywords: Perceptual Image Encryption, Vision Transformer, DNN, Privacy Preserving, Federated Learning, Access Control

Article history:
Received: July 6, 2022
Revised: August 13, 2023
Accepted: August 13, 2023
Published: September 2, 2023 (Online)

DOI: 10.37936/ecti-cit.2023173.253320

1. INTRODUCTION

Deep neural networks (DNNs) have been deployed in many applications including security critical ones such as biometric authentication, automated driving, and medical image analysis [1, 2]. Training successful models also requires three ingredients: a huge amount of data, GPU accelerated computing resources, and efficient algorithms, and it is not a trivial task. In fact, collecting images and labeling them is also costly and will also consume a massive amount of resources. Therefore, trained ML models have great business value. Considering the expenses necessary for the expertise, money, and time taken to train a model, a model should be regarded as a kind of intellectual property (IP). In addition, generally, data contains sensitive information, and it is difficult to train a model while preserving privacy. In particular, data with sensitive information cannot be transferred to untrusted third-party cloud environments (cloud GPUs and TPUs) even though they provide a powerful computing environment [3-9]. Accordingly, it has been challenging to train/test a DNN model with encrypted images as one way for solving these issues [10]. However, when using conventional perceptual encryption methods, the performance of models is degraded due to the influence of encryption.

In this paper, we present a block-wise encryption method for achieving reliable vision transformer (ViT) models. In the method, a model trained with plain images is transformed with a secret key to give unique features controlled with the key to the model, and encrypted images are applied to the model. In addition, three applications: privacy-preserving image classification, access control, and the combined use of federated learning and encrypted images, are presented to show the effectiveness of our method. In the method, the vision transformer (ViT) [11], which is known to have a high performance, is used to reduce the influence of block-wise encryption thanks to its architecture. It allows us not only to obtain the same performance as models trained with plain images but to also update the secret key easily. In experiments, our method is evaluated in terms of performance degradation and access control in an image classification task on the CIFAR-10 and CIFAR-100 datasets.

2. RELATED WORKS

Image encryption methods for deep learning and ViT are summarized here.

[1,2,3] The authors are with Department of Computer Science, Tokyo Metropolitan University, 6-6 Asahigaoka, Hino-shi, Tokyo 191-0065, Japan, E-mail: kiya@tmu.ac.jp, iijima-ryota@ed.tmu.ac.jp and nagamori-teru@ed.tmu.ac.jp.

2.1 Image Encryption for Deep Learning

Various image transformation methods with a secret key, often referred to as perceptual image encryption or image cryptography, have been studied so far for many applications. Figure 1 shows typical applications of image encryption with a key. Image encryption with a key allows us not only to protect the visual information of plain images but to also embed unique features controlled with the key into images. The use of visually protected images has enabled various kinds of applications. One of the origins of image transformation with a key is in block-wise image encryption schemes for encryption-then-compression (EtC) systems [12-21]. Image encryption prior to image compression is required in certain practical scenarios such as secure image transmission through an untrusted channel provider. An EtC system is used in such scenarios, although the traditional way of securely transmitting images is to use a compression-then-encryption (CtE) system. Compressible encryption methods have been applied to privacy-preserving compression, data hiding, and image retrieval [22-24] in cloud environments. In addition, visually protected images have been demonstrated to be effective in privacy-preserving learning [10, 25-30], adversarial defense [31-33], access control [34-36], and DNN watermarking [33, 37-44].

In this paper, we focus on image encryption for deep learning, called learnable encryption, under the use of ViT. In addition, it is demonstrated to be useful to privacy-preserving classification, access control, and federated learning with encrypted images while maintaining the high performance that ViT has.

2.2 Vision Transformer

The transformer architecture has been widely used in natural language processing (NLP) tasks [45]. The vision transformer (ViT) [11] has also provided excellent results compared with state-of-the-art convolutional networks. Following the success of ViT, several isotropic networks (with the same depth and resolution across different layers in the network) have been proposed such as MLP-Mixer [46], ResMLP [47], CycleMLP [48], gMLP [49], vision permutator [50], and ConvMixer [51].

Figure 2 illustrates the architecture of ViT, where ViT consists of two embedding processes (patch embedding and position embedding) and a transformer encoder. In ViT, an input image $x \in \mathbb{R}^{h\times w\times c}$ is segmented into N patches with a size of $p \times p$, where h, w, and c are the height, width, and number of channels of the image. In addition, an integer N is given as hw/p^2. After that, each patch is flattened as $x^i_{\rm p} = [x^i_{\rm p}(1), x^i_{\rm p}(2), \ldots, x^i_{\rm p}(L)]$, where $L = p^2c$. Finally, a sequence of embedded patches is given as

$$z_0 = [x_{\rm class}; x^1_{\rm p}\mathbf{E}; x^2_{\rm p}\mathbf{E}; \ldots x^i_{\rm p}\mathbf{E}; \ldots x^N_{\rm p}\mathbf{E}] + \mathbf{E}_{\rm pos}, \quad (1)$$

where

$$\mathbf{E}_{\rm pos} = ((\mathbf{e^0_{pos}})^\top (\mathbf{e^1_{pos}})^\top \ldots (\mathbf{e^i_{pos}})^\top \ldots (\mathbf{e^N_{pos}})^\top)^\top,$$
$$x_{\rm class} \in \mathbb{R}^D, x^i_{\rm p} \in \mathbb{R}^L, e^i_{\rm pos} \in \mathbb{R}^D,$$
$$\mathbf{E} \in \mathbb{R}^{\mathbf{L}\times\mathbf{D}},\ \mathbf{E}_{\rm pos} \in \mathbb{R}^{(\mathbf{N}+\mathbf{1})\times\mathbf{D}}.$$

$x_{\rm class}$ is the classification token, $\mathbf{E}$ is the embedding (patch embedding) to linearly map each patch to dimensions D, $\mathbf{E}_{\rm pos}$ is the embedding (position embedding) that gives position information to patches in the image, $e^0_{\rm pos}$ is the information of the classification token, and $e^i_{\rm pos}$, $i = 1, \ldots, N$, is the position information of each patch.

In patch embedding, patches are mapped to vectors, and the position information is embedded in position embedding. In this paper, we encrypt not only test images but also two embeddings: patch embedding $\mathbf{E}$ and position embedding $\mathbf{E}_{\rm pos}$, in a trained model. The resulting sequence of vectors is fed to a standard transformer encoder, and the output of the transformer is provided to a multi-layer perceptron (MLP) to get an estimation result.

3. IMAGE ENCRYPTION FOR VISION TRANSFORMER

An encryption method with random numbers is presented here. The method makes it possible to avoid the performance degradation of models even when using encrypted images.

3.1 Overview

Figure 3 shows the scenario of the presented scheme in a privacy-preserving image classification task, where it is assumed that the model builder is trusted, and the service provider is untrusted. The model builder trains a model by using plain images and encrypts the trained model with a key K.

The encrypted model is given to the service provider, and the key is sent to a client. The client prepares a test image encrypted with the key and sends it to the service provider. The encrypted test image is applied to the encrypted model to obtain an estimation result, and the result is sent back to the client. Note that the provider has neither a key nor plain images. The framework presented in this paper enables us to achieve this scenario without any performance degradation compared with the use of plain images.

3.2 Model Encryption

As shown in Figure 3, a trained model is encrypted by using random numbers (key). In the method, patch embedding E and position embedding $\mathbf{E}_{\rm pos}$ in Eq. (1) are encrypted by using random matrices, respectively.

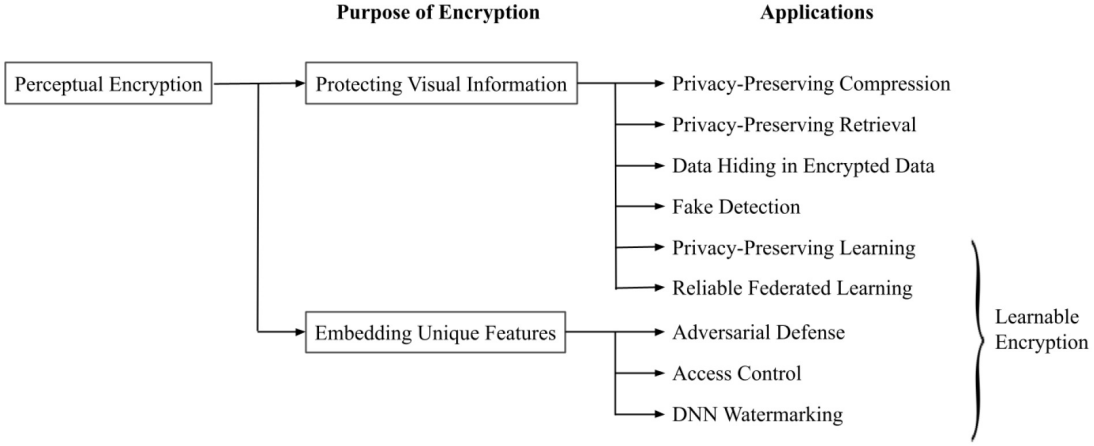

Fig.1: *Applications of perceptual image encryption.*

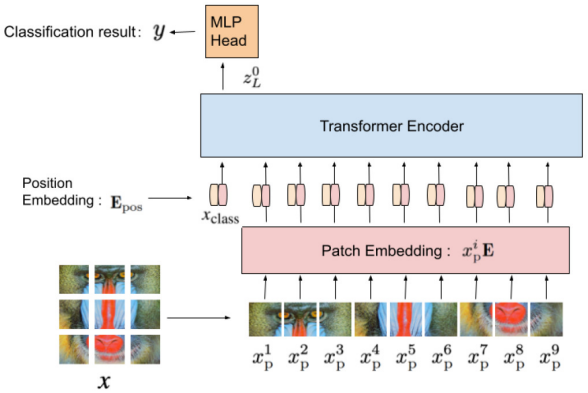

Fig.2: *Architecture of ViT [11].*

3.2.1 Patch Embedding Encryption

The following transformation matrix $\mathbf{E_a}$ is used to encrypt patch embedding $\mathbf{E}$.

$$\mathbf{E_a} = \begin{bmatrix} k_{(1,1)} & k_{(1,2)} & \cdots & k_{(1,L)} \\ k_{(2,1)} & k_{(2,2)} & \cdots & k_{(2,L)} \\ \vdots & \vdots & k_{(i,j)} & \vdots \\ k_{(L,1)} & k_{(L,2)} & \cdots & k_{(L,L)} \end{bmatrix}, \tag{2}$$

where

$$\mathbf{E_a} \in \mathbb{R}^{L\times L},\ \det \mathbf{E_a} \neq 0,$$
$$k_{(i,j)} \in \mathbb{R},\ i,j \in \{1,\ldots,L\}.$$

Note that the element values of $\mathbf{E_a}$ are randomly decided, but $\mathbf{E_a}$ has to have an inverse matrix.

Then, by multiplying $\mathbf{E}$ by $\mathbf{E_a}$, an encrypted patch embedding $\hat{\mathbf{E}}$ is given as

$$\hat{\mathbf{E}} = \mathbf{E_a}\mathbf{E}. \tag{3}$$

3.2.2 Position Embedding Encryption

Position embedding $\mathbf{E}_{\text{pos}}$ is encrypted as below.

1). Generate a random integer vector with a length of N as

$$l_t = [l_e(1), l_e(2), \ldots, l_e(i), \ldots, l_e(N)], \tag{4}$$

where

$$l_e(i) \in \{1,2,\ldots,N\},$$
$$l_e(i) \neq l_e(j) \text{ if } i \neq j,$$
$$i,j \in \{1,\ldots,N\}.$$

2). Calculate $m_{(i,j)}$ as

$$m_{(i,j)} = \begin{cases} 0 & (j \neq l_e(i)) \\ 1 & (j = l_e(i)). \end{cases} \tag{5}$$

3). Define a random matrix as

$$\mathbf{E_b} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & m_{(1,1)} & m_{(1,2)} & \cdots & m_{(1,N)} \\ 0 & m_{(2,1)} & m_{(2,2)} & \cdots & m_{(2,N)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & m_{(N,1)} & m_{(N,2)} & \cdots & m_{(N,N)} \end{bmatrix}, \tag{6}$$

where

$$\mathbf{E_b} \in \mathbb{R}^{(N+1)\times(N+1)}.$$

For instance, if $N = 3$ and $l_t = [1,3,2]$, $\mathbf{E_b}$ is given by

$$\mathbf{E_b} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \tag{7}$$

4). Transform $\mathbf{E}_{\text{pos}}$ to $\hat{\mathbf{E}}_{\text{pos}}$ as

$$\hat{\mathbf{E}}_{\text{pos}} = \mathbf{E_b}\mathbf{E}_{\text{pos}}. \tag{8}$$

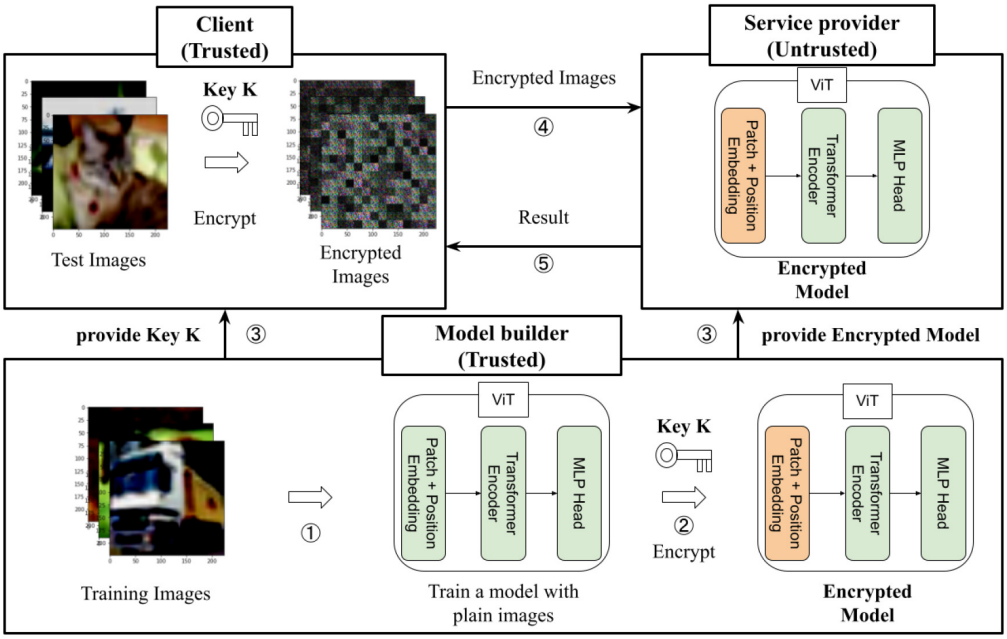

Fig.3: *Scenario of proposed scheme.*

3.3 Test Image Encryption

A test image $x \in \mathbb{R}^{h\times w\times c}$ is transformed into an encrypted image $\tilde{x} \in \mathbb{R}^{h\times w\times c}$ as below (see Figure 4).

(a) Divide x into N non-overlapped blocks with a size of $p\times p$ such that $B = \{B_1, \ldots, B_N\}$, where $p\times p$ is the same size as the patch size used in a ViT model.

(b) Generate permutated blocks $\bar{B}$ by

$$\begin{aligned}\bar{B} &= \mathbf{E_b}B \\ &= \{\bar{B}_1, \ldots, \bar{B}_N\},\end{aligned} \tag{9}$$

where $B \in \mathbb{R}^{1\times N}$.

(c) Flatten each block $\bar{B}_i \in \mathbb{R}^{p\times p\times c}$ into a vector $\bar{x}^i_{\rm p} \in \mathbb{R}^{p^2c}$ as

$$\bar{x}^i_{\rm p} = [\bar{x}^i_{\rm p}(1), \ldots, \bar{x}^i_{\rm p}(L)]. \tag{10}$$

Note that the following relation is satisfied.

$$\begin{aligned}&[x_{\rm class}; \bar{x}^1_{\rm p}; \ldots, \bar{x}^i_{\rm p}; \ldots; \bar{x}^N_{\rm p}] \\ =&\mathbf{E_b}[x_{\rm class}; x^1_{\rm p}; \ldots, x^i_{\rm p}; \ldots; x^N_{\rm p}]\end{aligned} \tag{11}$$

(d) Generate an encrypted vector $\tilde{x}^i_{\rm p}$ by multiplying vector $\bar{x}^i_{\rm p}$ by matrix $\mathbf{E_a}^{-1} \in \mathbb{R}^{L\times L}$ as

$$\tilde{x}^i_{\rm p} = \bar{x}^i_{\rm p}\mathbf{E_a}^{-1}. \tag{12}$$

(e) Rebuild vector $\tilde{x}^i_{\rm p}$ into block $\tilde{B}^i$ in the reverse order of step (c).

(f) Concatenate $\tilde{B} = \{\tilde{B}^1, \ldots \tilde{B}^N\}$ into an encrypted test image $\tilde{x}$.

Figure 5 shows an example of images encrypted with this procedure.

When replacing $\mathbf{E}$, $\mathbf{E}_{\rm pos}$, and $x^i_{\rm p}$ with $\hat{\mathbf{E}}$, $\hat{\mathbf{E}}_{\rm pos}$, and $\tilde{x}^i_{\rm p}$, respectively, the sequence in Eq. (1) is reduced to

$$\tilde{z}_0 = [x_{\rm class};, \tilde{x}^1_{\rm p}\hat{\mathbf{E}}; \ldots; \tilde{x}^i_{\rm p}\hat{\mathbf{E}}; \ldots; \tilde{x}^N_{\rm p}\hat{\mathbf{E}}] + \hat{\mathbf{E}}_{\rm pos}. \tag{13}$$

Thus, by substituting Eqs. (3),(8), and (11) with Eq. (13), we obtain:

$$\begin{aligned}\tilde{z}_0 &= [x_{\rm class}; \bar{x}^1_{\rm p}\mathbf{E_a}^{-1}\mathbf{E_a}\mathbf{E}; \ldots; \bar{x}^i_{\rm p}\mathbf{E_a}^{-1}\mathbf{E_a}\mathbf{E}; \ldots; \\ &\quad \bar{x}^N_{\rm p}\mathbf{E_a}^{-1}\mathbf{E_a}\mathbf{E}] + \mathbf{E_b}\mathbf{E}_{\rm pos} \\ &= \mathbf{E_b}[x_{\rm class}; x^1_{\rm p}\mathbf{E}; \ldots; x^i_{\rm p}\mathbf{E}; \ldots; x^N_{\rm p}\mathbf{E}] + \mathbf{E_b}\mathbf{E}_{\rm pos} \\ &= \mathbf{E_b}z_0\end{aligned} \tag{14}$$

From the equation, the influence of encryption can be avoided except for $\mathbf{E_b}$. Accordingly, the encrypted model allows us to have the same performance as that of the model trained with plain images, if test images are encrypted with the same key as that used for model encryption.

3.4 Generation of Random Matrices

A random orthogonal matrix can be generated as a random matrix $\mathbf{E_a}$ by using Gram–Schmidt orthonormalization [29,52]. The procedure for generating $\mathbf{E_a}$ with a size of $L\times L$ is given as follows.

1. Generate a real matrix $\mathbf{R}$ with a size of $L\times L$ by using a random number generator with a seed.
2. Calculate $\det(\mathbf{R})$, and proceed to step 3 if $\det(\mathbf{R}) \neq 0$. Otherwise, return to step 1.
3. Compute a random orthogonal matrix $\mathbf{E_a}$ from $\mathbf{R}$ by using Gram-Schmidt orthogonalization.

In this framework, any regular matrix can be used as $\mathbf{E_a}$ for image encryption. Several conventional

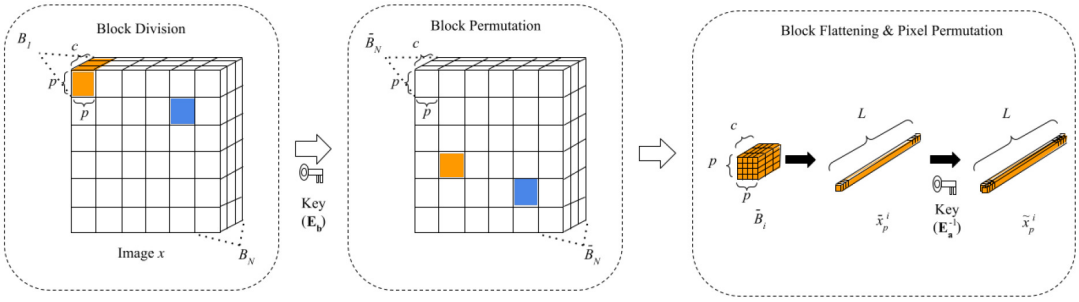

Fig.4: *Procedure of block-wise encryption.*

methods for privacy-preserving image classification use permutation matrices of pixel values [53,54], in which many elements have zero values in matrices as

$$\mathbf{E_a} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}. \tag{15}$$

In contrast, the random orthogonal matrices generated with Gram-Schmidt orthogonalization include no zero values as elements in general. The use of such matrices allows us not only to more strongly protect the visual information of plain images but to also enhance robustness against various attacks while maintaining the same performance as that of models trained with plain images. In addition, $\mathbf{E_a}^{-1}$ can easily be calculated as the transposed matrix of $\mathbf{E_a}$.

3.5 Properties of Proposed Scheme

When the block size for image encryption is the same as the patch size used in a ViT model. The proposed method has the following properties.

- The model performs well only if test images are transformed with the same key as that used for transforming the model from Eq. (14).
- The method does not cause any performance degradation in terms of the accuracy of models as in Eq. (14).
- Model training and encryption are independent (see Figure 3). Therefore, it is possible to easily update a key.

4. APPLICATIONS OF ENCRYPTED VIT

Three applications of encrypted ViT models are presented to demonstrate the usefulness of the encryption scheme.

4.1 Privacy-preserving Image Classification

One of the applications is to use encrypted ViT models for privacy-preserving image classification as shown in Figure 3, in which visually protected test images are sent to an untrusted provider.

A threat model includes a set of assumptions such as attacker's goals, knowledge, and capabilities. Users without secret key K are assumed to be the adversary. In this application, we consider the attacker's goal to be to restore visual information from encrypted test images. We assume that authorized users know key K, and the model owner securely manages both key K and the trained model without any encryption. In addition, the encryption method is also assumed to be disclosed except for key K. Thus, an adversary may perform ciphertext-only (COA) attacks via this information to restore the perceptual information from encrypted images.

Accordingly, the encryption method should satisfy the following requirements.

1. Security: No perceptual information of plain images should be reconstructed from encrypted images unless the key is exposed.
2. Model capability: Privacy-preserving methods for DNNs should maintain an approximate accuracy as when using plain images.
3. Computational requirement: Privacy-preserving DNNs should not increase the computational requirement in quantity.
4. Key update: The key should easily be updated without re-training the model.

In experiments, the effectiveness of our scheme will be evaluated in terms of the above requirements.

4.2 Access Control with Encypted Model

The second application is to protect a model from misuse when it has been stolen, referred to as access control that aims to protect the functionality of DNN models from unauthorized access. Trained models have great business value. Considering the expenses necessary for the expertise, money, and time taken to train a model, trained models should be regarded as a kind of intellectual property (IP). Accordingly, encrypted models are required not only to provide high performance to authorized users but also low performance to unauthorized users. Our scheme is effective in access control in addition to privacy-preserving deep learning as verified in an experiment.

4.3 Federated Learning in Combination with Encryption

It has been very popular for data owners to train and test deep neural network (DNN) models in cloud

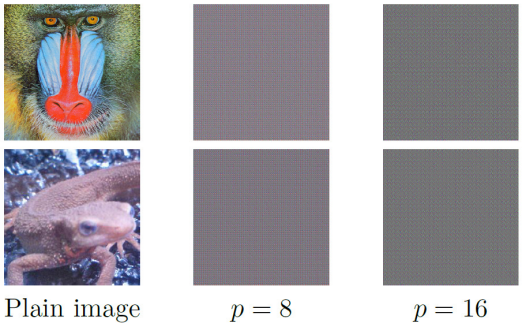

Fig.5: *Example of encrypted images.*

environments. However, data privacy such as personal medical records may be compromised in cloud environments, so privacy-preserving methods for deep learning have become an urgent problem.

One of the solutions is to use federated learning (FL) [55,56], which was proposed by Google. FL is capable of significantly preserving clients' private data from being exposed to adversaries. However, FL aims to construct models over multiple participants without directly sharing their raw data, so the privacy of test (query) images is not considered.

Another approach is to encrypt a trained model, and then encrypted test (query) images are applied to the encrypted model shown in Figure 3. However, this approach does not consider constructing models over multiple participants without directly sharing their raw data, although the visual information of test images can be protected.

For these reasons, the combined use of FL and encrypted test images is effective in privacy-preserving image classification tasks with ViT [10] (see Figure 6). The method allows us not only to train models over multiple participants without directly sharing their raw data but to also protect the privacy of test (query) images. In addition, it can maintain the same accuracy as that of models normally trained with plain images.

5. EXPERIMENT AND DISCUSSION

In experiments, the effectiveness of encrypted ViT models is verified in an image classification task.

5.1 Privacy Preservation

As the first application of encrypted ViT models, a privacy-preserving image classification task was carried out in accordance with the framework in 4.1 where visual information on test images is protected [53].

5.1.1 Experiment Setup

To confirm the effectiveness of the presented scheme, experiments were carried on the CIFAR-10 dataset (with 10 classes). The dataset consists of 60,000 color images (dimension of $3 \times 32 \times 32$), where

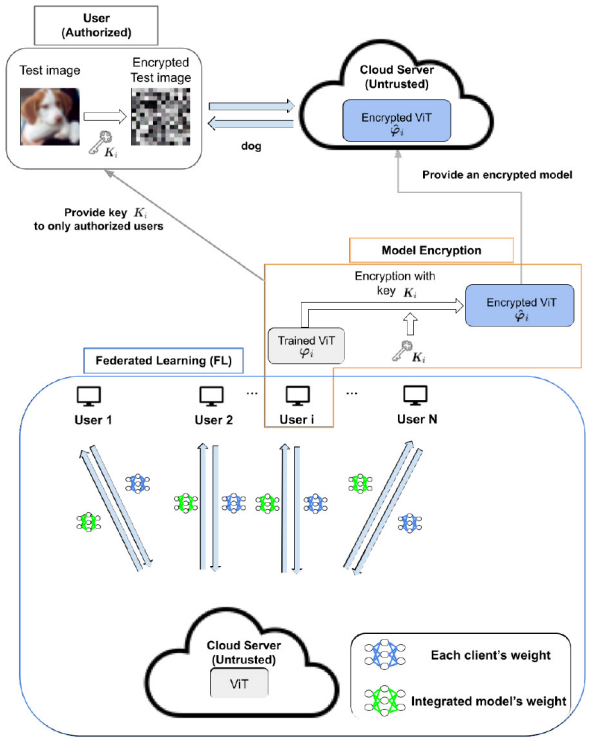

Fig.6: *Combined use of FL and encrypted test images.*

50,000 images are for training, 10,000 for testing, and each class contains 6,000 images. Images in the dataset were resized to $3 \times 224 \times 224$ to input them to ViT, before applying the proposed encryption algorithm, where the block size was 16×16. We used the PyTorch [57] implementation of ViT and fine-tuned a ViT model with a patch size $P = 16$, which was pre-trained on ImageNet-21k. The ViT model was fine-tuned for 5000 epochs. The parameters of the stochastic gradient descent (SGD) optimizer were a momentum of 0.9 and a learning rate value of 0.03.

In addition, we used three conventional visual information protection methods (Tanaka's method [58], the pixel-based encryption method [59], and the GAN-based transformation method [60]) to compare them with our method. ResNet-20 was used to validate the effectiveness of the conventional method with reference to [61]. CIFAR-10 was also used for training networks, and the networks were trained for 200 epochs by using SGD with a weight decay of 0.0005 and a momentum of 0.9. The learning rate was initially set to 0.1, and it was multiplied by 0.2 at 60, 120, and 160 epochs. The batch size was 128.

5.1.2 Performance of Encrypted ViT

First, we compared the proposed method with conventional ones in terms of the accuracy of image classification under the use of ViT and ResNet-20. As shown in Table 1, the performance of all conventional methods was degraded compared with the baselines, which were results calculated with plain images. In

Table 1: *Comparison with conventional methods in terms of classification accuracy*

Model	Method	Accuracy
ViT	Baseline	99.03
	Ours	**99.03**
ResNet-20 [61]	Baseline	91.55
	Tanaka [58]	87.02
	Pixel-based [59]	86.66
	GAN-based [60]	82.55

contrast, the proposed method did not degrade the performance at all. Accordingly, our method was verified to be able to maintain the same accuracy as that of the baselines as shown in Eq. (14).

5.1.3 Visual Protection

Figure 5 shows an example of images encrypted with the method in 3.3, where random matrices $\mathbf{E_a}$ were generated by using Gram-Schmidt orthonormalization. The images had $H \times W \times C = 512 \times 512 \times 3$ as an image size, and the block sizes used for encryption were $p = 8$ and $p = 16$. From the figures, the encrypted images have almost none of the visual information of the plain images.

In addition to visual protection, encrypted images have to be robust enough against various attacks, which aim to restore visual information from encrypted images. ViT has two embeddings: position embedding and patch embedding, so not only pixel values in every block but also the position of blocks can be changed randomly. We already confirmed that the encryption including block permutation is robust against cipher-text-only attacks (COAs) including jigsaw puzzle solver attacks [62]. In particular, the use of random matrices generated with Gram-Schmidt orthonormalization is more robust than that of simple permutation matrixes.

5.2 Application to Access Control

Next, we validated whether our method could protect models where the experimental conditions were the same as those in 5.1.1 [53]. Table 2 shows the accuracy of image classification when encrypted or plain images were input to the encrypted model. The encrypted model performed well for test images with the correct key, but its accuracy was not high when using plain test images. The CIFAR-10 dataset consists of ten classes, so 9.06 is almost the same accuracy as that when test images are randomly classified.

Next, we confirmed the performance of images encrypted with a different key from that used in the model encryption. We prepared 100 random keys, and test images encrypted with the keys were input to the encrypted model. From the box plot in Figure 7, the accuracy of the models was not high under the use of the wrong keys. Accordingly, the encrypted model was confirmed to be robust against a random key attack.

Table 2: *Robustness against use of plain images*

Model	Test Image	
	Plain	Ours
Baseline	**99.03**	-
Ours	9.06	**99.03**

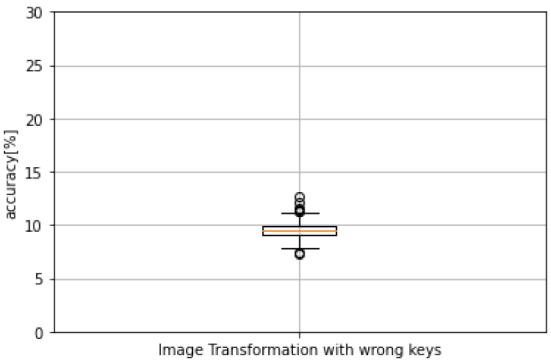

Fig. 7: *Evaluating robustness against random key attack. Boxes span from first to third quartile, referred to as $Q1$ and $Q3$, and whiskers show maximum and minimum values in range of $[Q1-1.5(Q3-Q1), Q3+1.5(Q3-Q1)]$. Band inside box indicates median. Outliers are indicated as dots.*

5.3 Combined Use of Federated Learning and Image Encryption

The effectiveness of encrypted ViT models was finally evaluated under the combined use of federated learning (FL) and image encryption as shown Figure 5 [63].

5.3.1 Setup

Experiments were conducted on the CIFAR-10 and CIFAR-100 datasets, where images were resized from $3 \times 32 \times 32$ to $3 \times 224 \times 224$ because we used ViT pre-trained with ImageNet-1K as a model. For training models with FL, 10 clients were assumed, where each client had 5, 000 training images and 1, 000 test images. Also, we used FedAVG [56] as the method of model integration. Models were trained using stochastic gradient descent (SGD) with an initial learning rate of 10^{-3}, a momentum of 0.9, and a batch size of 8. We also used the cross-entropy loss function. In addition, models were integrated every epoch, and the total number of epochs was set to 10. After the tenth integration, the integrated model was encrypted with secret keys, and every client used the secret keys to encrypt their test images.

5.3.2 Classification Performance

We evaluated the performance of the models in terms of classification accuracy. Table 3 shows the

Table 3: *Classification accuracy of proposed method*

	Integrated Model	Baseline
CIFAR-10	97.7	97.8
CIFAR-100	85.1	85.1

experimental results for the CIFAR-10 and CIFAR-100 datasets, which have 10 and 100 classes, respectively. "Integrated Model" indicates the results when the encrypted test images were applied to encrypted integrated models, and "Baseline" represents the results when plain test images were applied to the plain models normally trained with plain images.

From the results, the combined use of FL and encrypted images was verified to have the same accuracy as that of models normally trained with plain images. Accordingly, the method in Figure 6 allows us not only to train models over multiple participants without directly sharing raw data but to also protect the visual information of test images. In addition, our method enables us to easily update the key without re-training models, so each user can use an independent key to protect test images and their model.

6. CONCLUSIONS

In this paper, we presented a block-wise encryption method for ViT and its applications. The presented framework presented with the encryption method was verified to provide the same performance as that without any encryption since the embedding structure of ViT has a high similarity to block-wise encryption. In addition, three applications of the method: privacy-preserving image classification, access control, and the combined use of federated learning and the encryption were conducted to show the effectiveness of the method for reliable DNNs. In experiments, the method was demonstrated to outperform state-of-the-art methods with conventional methods for image encryption in terms of classification accuracy, and it was also verified to be effective in terms of the reliability of DNN models.

ACKNOWLEDGMENT

This study was partially supported by JSPS KAKENHI (Grant Number JP21H01327).

References

[1] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, p. 436, 2015.

[2] X. Liu, Z. Deng and Y. Yang, "Recent progress in semantic image segmentation," *Artif. Intell. Rev.*, vol. 52, no. 2, pp. 1089–1106, 2019.

[3] C.-T. Huang, L. Huang, Z. Qin, H. Yuan, L. Zhou, V. Varadharajan and C.-C. J. Kuo, "Survey on securing data storage in the cloud," *APSIPA Transactions on Signal and Information Processing*, vol. 3, p. e7, 2014.

[4] M.-R. Ra, R. Govindan and A. Ortega, "P3: Toward Privacy-Preserving photo sharing," in *10th USENIX Symposium on Networked Systems Design and Implementation (NSDI 13)*. Lombard, IL: USENIX Association, Apr. 2013, pp. 515–528.

[5] R. Lagendijk, Z. Erkin and M. Barni, "Encrypted signal processing for privacy protection: Conveying the utility of homomorphic encryption and multiparty computation," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 82–105, 2013.

[6] M. Fredrikson, S. Jha and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, ser. CCS '15. New York, NY, USA: Association for Computing Machinery, pp. 1322–1333, 2015.

[7] R. Shokri, M. Stronati, C. Song and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE symposium on security and privacy (SP). IEEE*, pp. 3–18, 2017.

[8] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, Y. Bengio and Y. LeCun, Eds., 2014.

[9] R. S. Siva Kumar, M. Nyström, J. Lambert, A. Marshall, M. Goertzel, A. Comissoneru, M. Swann, and S. Xia, "Adversarial machine learning-industry perspectives," in *2020 IEEE Security and Privacy Workshops (SPW)*, pp. 69–75, 2020.

[10] H. Kiya, A. P. M. Maung, Y. Kinoshita, S. Imaizumi, and S. Shiota, "An overview of compressible and learnable image transformation with secret key and its applications," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, e11, 2022. [Online]. Available: `http://dx.doi.org/10.1561/116.00000048`

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[12] T. Chuman, W. Sirichotedumrong, and H. Kiya, "Encryption-then-compression systems using grayscale-based image encryption for jpeg images," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 6, pp. 1515–

1525, 2019.

[13] T. Chuman, K. Kurihara, and H. Kiya, "On the security of block scrambling-based etc systems against jigsaw puzzle solver attacks," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2157–2161, 2017.

[14] J. Zhou, X. Liu, O. C. Au, and Y. Y. Tang, "Designing an efficient image encryption-thencompression system via prediction error clustering and random permutation," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 1, pp. 39–50, 2014.

[15] M. Ghonge and K. Nimbokar, "A survey based on designing an efficient image encryption-thencompression system," *International Journal of Computer Applications*, p. 8887, 2014.

[16] T. Y. Liu, K. J. Lin, and H. C. Wu, "Ecg data encryption then compression using singular value decomposition," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 3, pp. 707–713, 2018.

[17] W. Liu, W. Zeng, L. Dong, and Q. Yao, "Efficient compression of encrypted grayscale images," *IEEE Transactions on Image Processing*, vol. 19, no. 4, pp. 1097–1102, 2010.

[18] R. Hu, X. Li, and B. Yang, "A new lossy compression scheme for encrypted gray-scale images," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7387–7390, 2014.

[19] M. Johnson, P. Ishwar, V. Prabhakaran, D. Schonberg, and K. Ramchandran, "On compressing encrypted data," *IEEE Transactions on Signal Processing*, vol. 52, no. 10, pp. 2992–3006, 2004.

[20] M. Gaata and F. F. Hantoosh, "An efficient image encryption technique using chaotic logistic map and rc4 stream cipher," *International Journal of Modern Trends in Engineering and Research*, vol. 3, pp. 213–218, 2016.

[21] [W. Sirichotedumrong and H. Kiya, "Grayscale-based block scrambling image encryption using YCbCr color space for encryption-thencompression systems," *APSIPA Transactions on Signal and Information Processing*, vol. 8, p. e7, 2019.

[22] S. Imaizumi, Y. Izawa, R. Hirasawa, and H. Kiya, "A reversible data hiding method in compressible encrypted images," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E103.A, no. 12, pp. 1579–1588, 2020.

[23] K. Iida and H. Kiya, "Privacy-preserving content-based image retrieval using compressible encrypted images," *IEEE Access*, vol. 8, pp. 200 038–200 050, 2020.

[24] K. Iida and H. Kiya, "An image identification scheme of encrypted jpeg images for privacy-preserving photo sharing services," *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 4564– 4568, 2019.

[25] A. Kawamura, Y. Kinoshita, T. Nakachi, S. Shiota, and H. Kiya, "A privacy-preserving machine learning scheme using etc images," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E103.A, no. 12, pp. 1571–1578, 2020.

[26] Y. Bandoh, T. Nakachi, and H. Kiya, "Distributed secure sparse modeling based on random unitary transform," *IEEE Access*, vol. 8, pp. 211 762–211 772, 2020.

[27] T. Nakachi, Y. Bandoh, and H. Kiya, "Secure overcomplete dictionary learning for sparse representation," *IEICE Transactions on Information and Systems*, vol. E103.D, no. 1, pp. 50–58, 2020.

[28] T. Nakachi, Y. Wang, and H. Kiya, "Privacyp-reserving pattern recognition using encrypted sparse representations in l0 norm minimization," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2697–2701, 2020.

[29] I. Nakamura, Y. Tonomura, and H. Kiya, "Unitary transform-based template protection and its application to l 2 -norm minimization problems," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 1, pp. 60–68, 2016.

[30] T. Maekawa, A. Kwamura, T. Nakachi, and H. Kiya, "Privacy-preserving support vector machine computing using random unitary transformation," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E102.A, no. 12, pp. 1849–1855, 2019.

[31] M. Aprilpyone and H. Kiya, "Block-wise image transformation with secret key for adversarially robust defense," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2709–2723, 2021.

[32] A. MaungMaung and H. Kiya, "Encryption inspired adversarial defense for visual classification," in *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 1681–1685, 2020.

[33] A. MaungMaung and H. Kiya, "Ensemble of key-based models: Defense against black-box adversarial attacks," in *2021 IEEE 10th Global Conference on Consumer Electronics (GCCE)*, pp. 95–98, 2021.

[34] M. Chen and M. Wu, "Protect your deep neural networks from piracy," in *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–7, 2018.

[35] H. Chen, C. Fu, B. D. Rouhani, J. Zhao, and F. Koushanfar, "Deepattest: An end-to-end attestation framework for deep neural net-

works," *2019 ACM/IEEE 46th Annual International Symposium on Computer Architecture (ISCA)*, pp. 487–498, 2019.

[36] M. Aprilpyone and H. Kiya, "A protection method of trained cnn model with a secret key from unauthorized access," *APSIPA Transactions on Signal and Information Processing*, vol. 10, p. e10, 2021.

[37] Y. Uchida, Y. Nagai, S. Sakazawa, and S. Satoh, "Embedding watermarks into deep neural networks," in *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, ser. ICMR '17. Association for Computing Machinery, 2017, pp. 269–277.

[38] H. Chen, B. D. Rouhani, C. Fu, J. Zhao, and F. Koushanfar, "Deepmarks: A secure fingerprinting framework for digital rights management of deep learning models," in *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, ser. ICMR '19. Association for Computing Machinery, 2019, p. 105–113.

[39] B. D. Rouhani, H. Chen, and F. Koushanfar, "Deepsigns: A generic watermarking framework for ip protection of deep learning models," *arXiv preprint arXiv:1804.00750*, 2018. [Online]. Available: `https://arxiv.org/abs/1804.00750`

[40] L. Fan, K. W. Ng, C. S. Chan, and Q. Yang, "Deepip: Deep neural network intellectual property protection with passports," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6122-6139, 2021.

[41] Y. Adi, C. Baum, M. Cisse, B. Pinkas, and J. Keshet, "Turning your weakness into a strength: Watermarking deep neural networks by backdooring," in *Proceedings of the 27th USENIX Conference on Security Symposium*, ser. SEC'18. USENIX Association, 2018, pp. 1615–1631.

[42] J. Zhang, Z. Gu, J. Jang, H. Wu, M. P. Stoecklin, H. Huang, and I. Molloy, "Protecting intellectual property of deep neural networks with watermarking," in *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, ser. ASIACCS'18. Association for Computing Machinery, 2018, pp. 159–172.

[43] S. Sakazawa, E. Myodo, K. Tasaka, and H. Yanagihara, "Visual decoding of hidden watermark in trained deep neural network," in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 2019, pp. 371–374.

[44] E. Le Merrer, P. Pérez, and G. Trédan, "Adversarial frontier stitching for remote neural network watermarking," Neural Computing and Applications, vol. 32, no. 13, pp. 9233–9244, 2019.

[45] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[46] I. Tolstikhin, N. Houlsby, A. Kolesnikov, L. Beyer, X. Zhai, T. Unterthiner, J. Yung, A. P. Steiner, D. Keysers, J. Uszkoreit, M. Lucic, and A. Dosovitskiy, "MLP-mixer: An all-MLP architecture for vision," in Advances in Neural Information Processing Systems, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021.

[47] H. Touvron, P. B. andMathilde Caron, M. Cord, A. El-Nouby, E. Grave, A. Joulin, G. Synnaeve, J. Verbeek, and H. Jégou, "Resmlp: Feedforward networks for image classification with data-efficient training," CoRR, vol. abs/2105.03404, 2021.

[48] S. Chen, E. Xie, C. GE, R. Chen, D. Liang, and P. Luo, "CycleMLP: A MLP-like architecture for dense prediction," in *International Conference on Learning Representations*, 2022.

[49] H. Liu, Z. Dai, D. So, and Q. V. Le, "Pay attention to MLPs," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021.

[50] Q. Hou, Z. Jiang, L. Yuan, M.-M. Cheng, S. Yan, and J. Feng, "Vision permutator: A permutable mlp-like architecture for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 1328-1334, 2023.

[51] A. Trockman and J. Z. Kolter, "Patches are all you need?," *arXiv preprint arXiv:2201.09792*, 2022. [Online]. Available: `https://arxiv.org/abs/2201.09792`

[52] Y. Wang and K. Plataniotis, "Face based biometric authentication with changeable and privacy preservable templates," in *2007 Biometrics Symposium*, pp. 1–6, 2007.

[53] H. Kiya, R. Iijima, A. MaungMaung, and Y. Kinoshita, "Image and model transformation with secret key for vision transformer," *IEICE Transactions on Information and Systems*, vol. E106.D, no. 1, pp. 2–11, 2023.

[54] H. Kiya, T. Nagamori, S. Imaizumi, and S. Shiota, "Privacy-preserving semantic segmentation using vision transformer," *Journal of Imaging*, vol. 8, no. 9, 2022. [Online]. Available: `https://www.mdpi.com/2313-433X/8/9/233`

[55] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtarik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," in *NIPS Workshop on Private Multi-Party Machine Learning*, 2016.

[Online]. Available: https://arxiv.org/abs/1610.05492

[56] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "CommunicationEfficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Singh and J. Zhu, Eds., vol. 54. PMLR, 20–22 Apr 2017, pp. 1273–1282.

[57] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8024–8035.

[58] M. Tanaka, "Learnable image encryption," in *2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW)*, pp. 1–2, 2018.

[59] 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 *2019 IEEE International Conference on Image Processing (ICIP)*, pp. 674–678, 2019.

[60] W. Sirichotedumrong and H. Kiya, "A ganbased image transformation scheme for privacypreserving deep neural networks," in *2020 28th European Signal Processing Conference (EUSIPCO)*, pp. 745–749, 2021.

[61] H. Ito, Y. Kinoshita, M. Aprilpyone, and H. Kiya, "Image to perturbation: An image transformation network for generating visually protected images for privacy-preserving deep neural networks," IEEE Access, vol. 9, pp. 64 629–64 638, 2021.

[62] T. Chuman and H. Kiya, "A jigsaw puzzle solver-based attack on image encryption using vision transformer for privacy-preserving dnns," Information, vol. 14, no. 6, 2023. [Online]. Available: https://www.mdpi.com/2078-2489/14/6/311

[63] T. Nagamori and H. Kiya, "ombined use of federated learning and image encryption for privacy-preserving image classification with vision transformer," arXiv preprint arXiv:2301.09255, 2023. [Online]. Available: https://arxiv.org/abs/2301.09255

Hitoshi Kiya received his B.E and M.E. degrees from the Nagaoka University of Technology in 1980 and 1982, respectively, and his Dr. Eng. degree from Tokyo Metropolitan University in 1987. In 1982, he joined Tokyo Metropolitan University, where he became a Full Professor in 2000. He is currently a Professor Emeritus and Leading Professor at Tokyo Metropolitan University. From 1995 to 1996, he attended the University of Sydney, Australia as a Visiting Fellow. He is a Fellow of IEEE, IEICE, AAIA, and ITE. He served as President of APSIPA from 2019 to 2020 and as Regional Director-at-Large for Region 10 of the IEEE Signal Processing Society from 2016 to 2017. He was also President of the IEICE Engineering Sciences Society from 2011 to 2012. He has been an Editorial Board Member of eight journals, including IEEE Trans. on Signal Processing, Image Processing, and Information Forensics and Security. He has organized a lot of international conferences in such roles as TPC Chair of IEEE ICASSP 2012 and as General Co-Chair of IEEE ISCAS 2019. He has received numerous awards, including 12 best paper awards.

Ryota Iijima received his B.C.S degree from Tokyo Metropolitan University, Japan in 2022. Since 2022, he has been a Master course student at Tokyo Metropolitan University. His research interests include deep neural networks and their protection.

Teru Nagamori received his B.C.S degree from Tokyo Metropolitan University, Japan in 2022. Since 2022, he has been a Master course student at Tokyo Metropolitan University. His research interests include deep neural networks and their protection.