

Received March 24, 2021, accepted April 17, 2021, date of publication April 22, 2021, date of current version May 4, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3074968

# Image to Perturbation: An Image Transformation Network for Generating Visually Protected Images for Privacy-Preserving Deep Neural Networks

HIROKI ITO<sup>1</sup>, YUMA KINOSHITA<sup>1</sup>, (Member, IEEE),  
MAUNGMAUNG APRILPYONE<sup>1</sup>, (Graduate Student Member, IEEE),  
AND HITOSHI KIYA<sup>1</sup>, (Fellow, IEEE)

Department of Computer Science, Tokyo Metropolitan University, Tokyo 191-0065, Japan

Corresponding author: Hitoshi Kiya (kiya@tmu.ac.jp)

This work was supported in part by JSPS KAKENHI under Grant JP21H01327, and in part by the Support Center for Advanced Telecommunications Technology Research Foundation (SCAT).

**ABSTRACT** We propose a novel image transformation network for generating visually protected images for privacy-preserving deep neural networks (DNNs). The proposed transformation network is trained by using a plain image dataset so that plain images are converted into visually protected ones. Conventional perceptual encryption methods cause some accuracy degradation in image classification and are not robust enough against state-of-the-art attacks. In contrast, the proposed network not only enables us to maintain the image classification accuracy that using plain images achieves but is also strongly robust against attacks including DNN-based ones. Furthermore, there is no need to manage any security keys as the conventional methods require. In an image classification experiment, the proposed network is demonstrated to strongly protect the visual information of plain images while maintaining a high classification accuracy under the use of two typical classification networks: ResNet and VGG. In addition, it is shown that the visually protected images are robust enough against various attacks in an experiment in which we tried to restore the visual information of plain images.

**INDEX TERMS** Adversarial example, deep neural network, privacy preserving, visual protection.

## I. INTRODUCTION

The spread of deep neural networks (DNNs) has greatly contributed to solving complex tasks for many applications [1], [2], such as computer vision, biomedical systems, and information technology. Deep learning utilizes a large amount of data to extract representations of relevant features, so performance is significantly improved [3]. Therefore, DNNs have been deployed in privacy-sensitive/security-critical applications, such as facial recognition, biometric authentication, and medical image analysis.

Recently, with the development of cloud services, DNNs are often carried in cloud environments. One of the

The associate editor coordinating the review of this manuscript and approving it for publication was Junhua Li<sup>1</sup>.

advantages of cloud environments is that cloud providers can provide various web-based software services like software as a service (SaaS). However, since cloud providers are not assumed to be trusted in general, private data, such as personal information and medical records, may be revealed in cloud computing [4]. Therefore, it is necessary to protect data privacy in cloud environments, and privacy-preserving DNNs have become an urgent challenge. Proposals in some studies on privacy-preserving machine learning were made on the basis of a differential privacy strategy [5], [6]. In [5], a computation-efficient decentralized stochastic gradient algorithm was proposed, and it can mask the privacy of each constituent function. In [6], a differentially private-distributed stochastic subgradient-push algorithm was also proposed to effectively mask differential

privacy. These studies have made a great contribution to the research field of privacy-preserving machine learning.

In contrast, various perceptual encryption methods have been proposed for generating images without visual information [7]–[21] in accordance with a visual information-protection strategy, although information theory-based encryption (like RSA and AES) generates ciphertext. In contrast to information theory-based encryption, images encrypted by perceptual encryption methods can be directly applied to various image processing algorithms. Perceptual encryption aims to generate images without visual information on plain images on the basis of a visual information-protection strategy since visual information includes sensitive personal information such as time, place, and personally identifiable information. However, most perceptual encryption methods cannot be applied to DNNs. There are only three methods, Tanaka's method [18], a pixel-based encryption method [19], [20], and a GAN (generative adversarial network)-based transformation method [22], for privacy-preserving DNNs. However, with these methods, performance degrades in DNNs, compared with the use of plain images. In addition, they are not robust against various attacks.

For such reasons, in this paper, we propose a novel transformation network for generating visually protected images for privacy-preserving DNNs under the visual information-protection strategy. The proposed network, which is inspired by the idea of adversarial examples, converts a plain image into a visually protected one. It is trained so that the generated images reduce the loss value of a classification network. Therefore, it enables us not only to protect visual information on plain images but also to maintain the performance of DNNs. In addition, the proposed framework has no security keys unlike the conventional methods because the proposed network irreversibly transforms images into visually protected ones with features used for classifying images, like a robust hashing function.

In an experiment, image classification is carried out under the use of the CIFAR datasets [23] and two classification networks, ResNet-20 [24] and VGG16 [25] with batch normalization, to evaluate the effectiveness of the proposed transformation network. From the results, visually protected images generated by the proposed network are demonstrated to have less visual information than those generated by using conventional methods while maintaining the classification accuracy that using plain images achieves. In addition, an experiment is conducted to evaluate robustness against various attacks including DNN-based ones.

The rest of this paper is structured as follows. Section II presents background information and related work on perceptual encryption and adversarial examples. Regarding the proposed transformation network, Section III includes an overview, a training procedure, image classification with the proposed network, and robustness against attacks. Experiments on the proposed method in terms of classification

accuracy and robustness are presented in Section IV, and Section V concludes this paper.

## II. RELATED WORK

In this section, we briefly summarize existing visual protection methods for images and their problems. Also, we explain the adversarial examples that inspired the proposed transformation network.

### A. PROTECTING VISUAL INFORMATION

This paper focuses on protecting visual information for privacy-preserving DNNs. A lot of perceptual encryption methods have been proposed for protecting the visual information of images [7]–[21]. Perceptual image encryption generates visually protected images that are described as bitmap images. Therefore, the encrypted images can be directly applied to some image processing algorithms. For example, encryption methods [7], [8] have been proposed not only for visually protecting privacy and security but also for matching and searching for images in the encrypted domain.

Compressible encryption methods have been also proposed that consider both security and efficient compression so that they can be adapted to cloud storage and photo sharing services [9]–[14]. Some of them [11]–[13] can be applied to traditional machine learning algorithms, such as support vector machine, k-nearest neighbors, and random forest, even under the use of the kernel trick [15], [16]. However, when these methods are applied to DNNs, the performance of the DNNs heavily degrades.

### B. VISUAL PROTECTION METHODS FOR DNNs

There are three visual protection methods [18]–[22] that generate visually protected images for privacy-preserving DNNs (see Fig. 1). The first is Tanaka's method [18], which utilizes an adaptation network prior to DNNs to reduce the influence of image encryption. The second is a pixel-based encryption method [19], [21]. The third is a GAN-based transformation method [22].

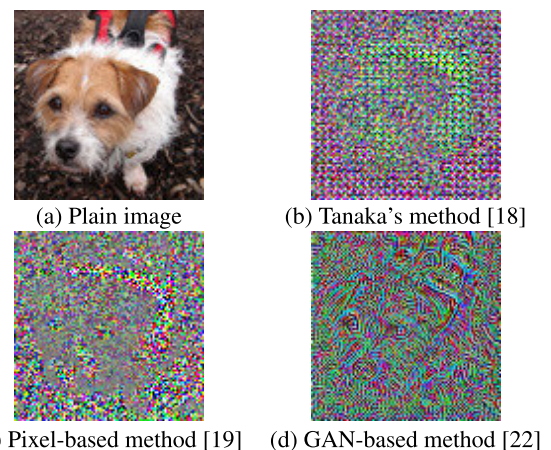


FIGURE 1. Images visually protected with conventional methods.

However, these methods cause a decrease in the classification accuracy [20]. In addition, images encrypted by using these methods are not robust against some attacks, as described later. Therefore, visual information on plain images is reconstructed by using attack methods. These methods also encrypt images by using a security key, so it is necessary to manage the key, except for the method in [22].

In this paper, our proposed novel transformation network for generating visually protected images for privacy-preserving DNNs overcomes these issues that the conventional methods have. Most conventional visual protection methods generate visually protected images without any information on a classification model. Although the GAN-based protection method [22] uses information on a classification model for training a transformation network, the model used for training a transformation network is not equal to that used in the process of image classification. In contrast, in the proposed method, a transformation network is trained by using the same model as that used in the process of image classification. Under this condition, a transformation network is trained with a loss that consists of two loss functions, classification accuracy and visual protection, as described in Eq. 2. Therefore, the proposed method does not cause accuracy to degrade as much as the conventional methods.

The contributions of this paper are summarized as below.

- We propose a novel transformation network for generating visually protected test images for privacy-preserving DNNs. The proposed transformation network is trained by using a model used for classifying test images under two loss functions, classification accuracy and visual protection, for the first time.
- We conduct experiments on image classification using the CIFAR datasets to confirm the effectiveness of the proposed method. The results show that the proposed method outperformed all conventional methods in terms of both classification accuracy and robustness against attacks. In particular, it was demonstrated to be robust enough against various attacks including state-of-the-art ones.

### C. ADVERSARIAL EXAMPLES

The proposed transformation network is inspired by the approach of adversarial examples. Adversarial examples are known as images to which specific imperceptible perturbations are added. Attacks using them are well-known as attack methods that machine learning suffers from. Neural networks, including convolutional ones, have already been demonstrated to be vulnerable to adversarial examples [26], [27]. These examples can cause neural networks to misclassify images with high confidence or force them to classify a target class. There are many studies on attacks that use adversarial examples and the defenses against them [27]–[33], and one of them is an attack

method that uses fully convolutional networks (FCNs) [28], including U-Net [34]. It trains an FCN to convert clean inputs into adversarial examples. We consider whether this method can be applied to generate visually protected images by using an FCN from clean images.

Adversarial examples were also applied to access restrictions for controlling trained networks in [35], in which input images were converted so that the output images could be classified correctly only when specific DNNs were used. However, visual protection for images has not been considered yet. In this paper, we propose a transformation network for generating visually protected images that can be correctly classified for the first time.

## III. PROPOSED TRANSFORMATION NETWORK

### A. OVERVIEW AND THREAT MODELS

Figure 2 illustrates the framework used in this paper. Transformation network  $h_\theta$  and image classification model  $\psi$  are prepared by a third party. The third party provides  $h_\theta$  and  $\psi$  to a client and a cloud provider, respectively. On the client side, visually protected test images are generated from testing plain ones by using  $h_\theta$ , and the protected test images are then sent to the cloud provider. In the cloud, the protected images are classified by using  $\psi$ , and the results are sent back to the client.

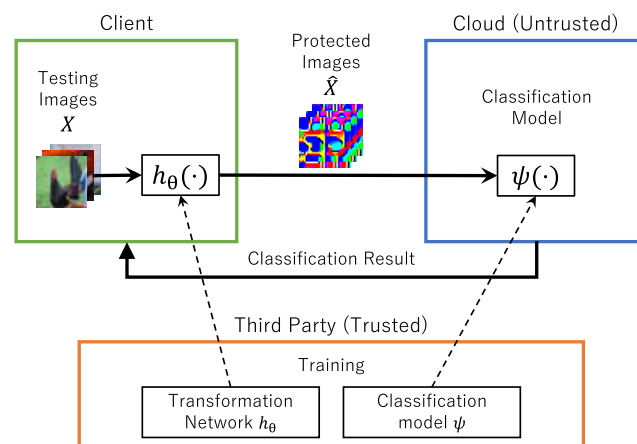


FIGURE 2. Framework of proposed scheme.

In this framework, it is assumed that a cloud provider is untrusted or semi-trusted, so visual information of test images may be leaked or illegally used in cloud computing. Therefore, in our framework, this information is not provided to the cloud provider. In contrast, a third party does not need any test images for training a transformation network, so there is no possibility of visual information leakage. The third party has to be trusted by public users because a transformation network trained by the third party is required to be used by users with confidence. Therefore, the third party should be an organization that has been evaluated and that is independent of the cloud provider.

Image classification is carried out under the use of visually protected images. The cloud provider cannot reconstruct visual information on plain images from protected images, even when transformation network  $h_\theta$  is open to the public. In addition, the proposed framework has no security keys unlike the conventional methods because the proposed network irreversibly transforms images into visually protected ones with features used for classifying images, like a robust hashing function. Accordingly, the cloud provider can offer various web-based software services like software as a service (SaaS) to many clients by using model  $\psi$  with high performance while preserving the privacy of the clients. The use of network  $h_\theta$  enables clients to securely use model  $\psi$ , which they cannot prepare themselves, with high performance even when all clients use a common transformation network.

Model  $\psi$  is also available for plain test images, so the cloud provider can provide services to clients who do not worry about protecting visual information or cannot prepare the computing cost needed for image transformation with  $h_\theta$ .

**B. TRAINING TRANSFORMATION NETWORK**

The training procedure of the proposed transformation network is illustrated in Fig. 3, where  $X = \{x_1, x_2, \dots, x_m\}$  is an input plain image set,  $\hat{X} = \{\hat{x}_1, \hat{x}_2, \dots, \hat{x}_m\}$  is an image set output from the transformation network  $\hat{x}_i = h_\theta(x_i)$ ,  $Y = \{y_1, y_2, \dots, y_m\}$  is a one-hot encoded target label set, and  $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_m\}$  is a label set output from a classification network  $\hat{y}_i = \psi(\hat{x}_i)$ . A one-hot encoded label, e.g.,  $y_i \in \{y_1, y_2, \dots, y_m\}$ , is described as  $y_i = (y_i(1), y_i(2), \dots, y_i(c)), y_i(j) \in \{0, 1\}, \sum_{j=1}^c y_i(j) = 1$ , and an output label, e.g.,  $\hat{y}_i \in \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_m\}$ , is described as  $\hat{y}_i = (\hat{y}_i(1), \hat{y}_i(2), \dots, \hat{y}_i(c)), 0 \leq \hat{y}_i(j) \leq 1, \sum_{j=1}^c \hat{y}_i(j) = 1$ , where  $c$  is the number of classes. The proposed network converts images to visually protected ones. Network  $h_\theta$  is trained so that generated images reduce the loss value of a classification network ( $\psi$ ).

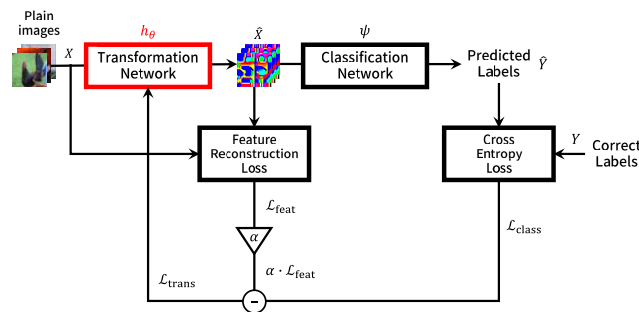


FIGURE 3. Training process of transformation network  $h_\theta$ .

To train the proposed network  $h_\theta$  with parameter  $\theta$  by using a plain input image ( $x_i$ ) and its one-hot encoded target label ( $y_i$ ), loss function  $\mathcal{L}_{trans}$  is minimized as

$$\underset{\theta}{\text{minimize}} \quad \frac{1}{m} \sum_{i=1}^m \mathcal{L}_{trans}(x_i, h_\theta(x_i), y_i), \quad (1)$$

with

$$\mathcal{L}_{trans}(x_i, \hat{x}_i, y_i) = \mathcal{L}_{class}(\hat{x}_i, y_i) - \alpha \cdot \mathcal{L}_{feat}(x_i, \hat{x}_i), \quad (2)$$

where  $\mathcal{L}_{class}$  denotes a classification loss function, which is used to classify visually protected images correctly,  $\mathcal{L}_{feat}$  is a feature reconstruction loss function to be used for visually protecting input images, and  $\alpha \in \mathbb{R}$  is a weight of  $\mathcal{L}_{feat}$ . In this paper, we used the stochastic gradient descent (SGD) optimizer, which is a well-known optimizer, to solve the minimization problem in Eq. (1). Note that, for adversarial examples,  $\alpha = 0$  is chosen in Eq. (2), and  $\mathcal{L}_{class}$  is maximized.

In this paper,  $\mathcal{L}_{class}$  is given by the cross-entropy loss as in adversarial examples. Therefore,  $\mathcal{L}_{class}$  is calculated by using  $\hat{y}_i(j)$  as

$$\mathcal{L}_{class}(\hat{x}_i, y_i) = - \sum_{j=1}^c y_i(j) \log \hat{y}_i(j), \quad (3)$$

where  $\hat{y}_i$  is an output of classification network  $\psi$  trained with plain images.  $\mathcal{L}_{feat}$  is also given by

$$\mathcal{L}_{feat}(x_i, \hat{x}_i) = \frac{1}{C_k H_k W_k} \|\phi_k(\hat{x}_i) - \phi_k(x_i)\|_2^2, \quad (4)$$

where  $\phi_k(x)$  is a feature map with a size of  $C_k \times H_k \times W_k$  obtained by the  $k$ -th layer of a network when image  $x$  is fed [36].

In simulations, we utilized U-Net [34] for transformation network  $h_\theta$ , and the feature map of the second ReLU function of VGG16 [25] without batch normalization pretrained with ImageNet was used for  $\phi_k$ .

**C. PRIVACY-PRESERVING IMAGE CLASSIFICATION**

Under the use of the proposed network pretrained with classification network  $\psi$ , image classification is performed in accordance with the following procedure.

- 1) A client inputs a plain test image ( $x$ ) to transformation network  $h_\theta$  in order to obtain a visually protected image ( $\hat{x}$ ).
- 2) The client sends  $\hat{x}$  to the cloud provider.
- 3)  $\hat{x}$  is classified by using classification network  $\psi$ , and the result ( $\hat{y}$ ) is returned to the client.

**D. ROBUSTNESS AGAINST ATTACKS AND THREAT MODELS**

In this framework, the cloud provider has no visual information of plain images, but they have visually protected images and transformation network  $h_\theta$ . Therefore, they might try to estimate the visual information of plain images from the visually protected ones. The proposed transformation network is designed not only to achieve a high classification performance but also to be robust enough against such attacks. A cloud provider might try to estimate the visual information on test images. Therefore, the proposed transformation network should be evaluated in terms of robustness against various attacks, although the visual protection is carried out without any security keys. In particular, DNN-based attacks

have been demonstrated to be able to reconstruct visual information on plain images from encrypted ones as one of the ciphertext-only attacks [20], [37], [38], where the state-of-the-art is a GAN (generative adversarial network)-based attack [38].

The GAN-based attack may enable us to estimate visual information on plain images even when a correct pair set of plain images and protected images is not prepared, as shown in Fig. 4. In the proposed scheme in Fig. 2, attackers can easily prepare a correct set because  $h_\theta$  is open to the public. Therefore, attackers may be able to more easily create an inverse transformation model by using the correct pair set to estimate visual information on plain images, as shown in Fig. 5. In this paper, even when the DNN-based attack in Fig. 5 referenced to as ITN-Attack is applied to the proposed transformation network, protected images will be shown to be robust enough against the attack in an experiment.

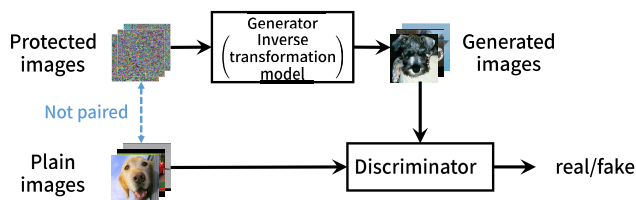


FIGURE 4. Training of GAN-based attack model.

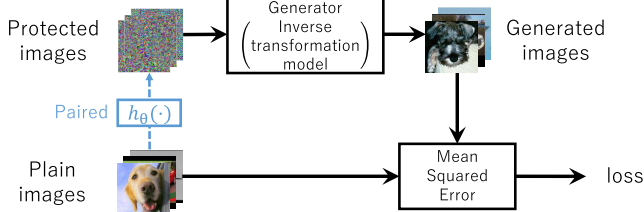


FIGURE 5. Training of inverse transformation model used in this paper.

#### IV. EXPERIMENTS

We evaluated the proposed transformation network in terms of classification accuracy and visual protection.

##### A. CLASSIFICATION ACCURACY

###### 1) EXPERIMENTAL SETUP

In this simulation, two classification networks, ResNet-20 [24] and VGG16 [25] with batch normalization as  $\psi$ , were used to evaluate the effectiveness of the proposed method. We also used two datasets, the CIFAR-10 and 100 datasets [23], which consist of a training set with 50,000 images and a test set with 10,000. To train both the classification networks and transformation network  $h_\theta$ , 45,000 and 47,500 images in the training sets of CIFAR-10 and 100, respectively, were utilized, and the other images were used as validation data. Also, the test sets of CIFAR-10 and 100 were utilized for evaluating the performance of the networks. In addition,

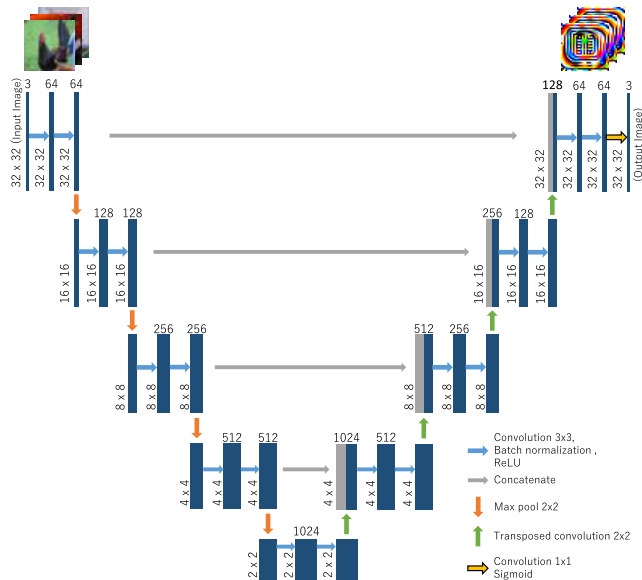


FIGURE 6. Structure of transformation network (U-Net). Each box denotes multi-channel feature map produced by each layer. Number of channels is denoted above each box. Feature map resolutions are denoted to left of boxes.

standard data-augmentation methods, i.e., random crop and horizontal flip, were performed in the training. The transformation network, based on U-Net, had the structure shown in Fig. 6.

All networks were trained for 200 epochs by using stochastic gradient descent (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. After the training, we selected the model that provided the lowest loss value under the use of the validation set.

###### 2) VISUAL-PROTECTION PERFORMANCE

Figure 7 shows an example of visually protected images generated from ten test images in CIFAR-10 by using  $h_\theta$  trained with ResNet-20 for calculating  $\mathcal{L}_{class}$ , where the top row shows plain images, and the next row to the bottom row show images generated with the parameters  $\alpha = 0, 0.005, 0.01, \text{ and } 0.05$  in Eq. (2).

From the figure, the generated images had almost no visual information on the plain images when  $\alpha \geq 0.005$ . Also, in the case of  $\alpha = 0$ , the generated images were not visually protected since  $\mathcal{L}_{feat}$ , i.e., the loss for visually protecting input images, did not work.

From Fig. 7, it was confirmed that the visual protection was more enhanced when using larger alpha values. In addition, when  $\alpha \geq 0.005$ , the protected images were very similar. The reason that the generated images were similar is that the transformation network extracts features required for image classification from plain images, and this is a positive property for visual protection.

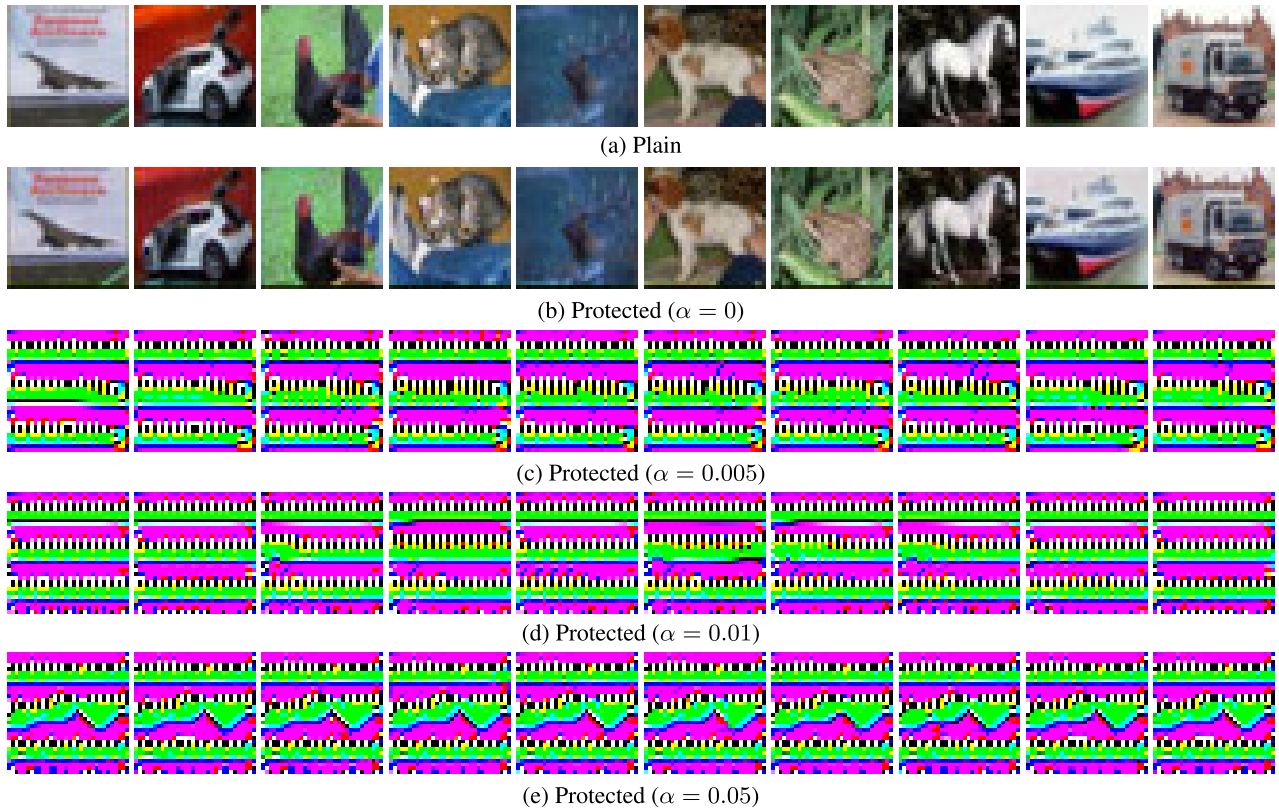


FIGURE 7. Visually protected images generated by proposed transformation network trained with ResNet-20.

From the comparison of Figs. 1 and 7, the encrypted images in Fig. 1 still had some visual information on the plain images. Therefore, the proposed transformation network was shown to strongly protect visual information on plain images.

In this framework, the output of the transformation network is expressed in an image format, so it is suitable for using well-known image classifier models without any modification. Accordingly, the cloud side can also carry out image classification without distinction between plain images and transformed ones. Model  $\psi$  is also available for plain test images, so the cloud provider can provide services to clients who do not worry about protecting visual information.

### 3) CLASSIFICATION ACCURACY

We evaluate the classification accuracy under the use of visually protected images in this section. Tables 1 and 2 show the experimental results of using CIFAR-10 and 100, respectively, where “ResNet-20” and “VGG16” mean that each network was used for  $\psi$ .

As shown in Table 1, the proposed method achieved a higher classification accuracy than the conventional methods under both ResNet-20 and VGG16. The classification accuracy was also confirmed to depend on the value of  $\alpha$ . When a value of  $\alpha = 0.01$  was used, the proposed method provided

TABLE 1. Classification accuracy with CIFAR-10.

$\alpha$	Accuracy (%)	
	$\psi = \text{ResNet-20}$	$\psi = \text{VGG16}$
0.005	90.46	81.82
0.01	91.56	91.59
0.05	91.98	44.30
0.1	68.91	10.00
Plain image	91.55	92.49
Tanaka [18]	87.02	86.96
Pixel-based [19]	86.66	88.00
GAN-based [22]	82.55	81.57

a high classification accuracy, which was almost the same as that when using plain images.

Similarly, the proposed method outperformed the conventional ones in terms of classification accuracy even with CIFAR-100 under both ResNet-20 and VGG16 with an appropriate value for  $\alpha$  selected, as shown in Table 2. When  $\alpha = 0.005$  was selected, it achieved the highest accuracy values under both ResNet-20 and VGG16.

The reason the proposed method does not cause accuracy to degrade that much is that a transformation network is trained by using the same model as that used in the process of image classification. In contrast, most conventional visual protection methods generate visually protected images without any information on a classification model. Although the GAN-based protection method [22] uses information on a

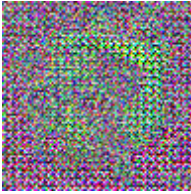

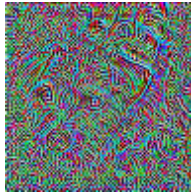


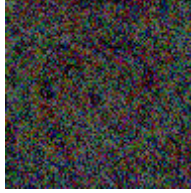
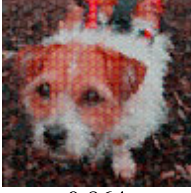

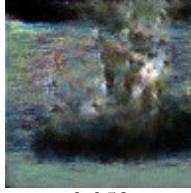



Attack method	Tanaka's [18]	Pixel-based [19]	GAN-based [22]
Protected	 0.050	 0.041	 0.164
FR-Attack [39]	 0.101	 0.303	 0.091
GAN-Attack [38]	 0.864	 0.109	 0.058
ITN-Attack [40]	 0.949	 0.017	 0.664

FIGURE 8. Images restored with three attack methods. SSIM values are given under images.

classification model for training a transformation network, the model used for training a transformation network is not equal to that used in the process of test image classification. In contrast, in the proposed method, a transformation network is trained by using the same model as that used in image classification. Therefore, the proposed method does not cause accuracy to degrade compared with the conventional methods.

TABLE 2. Classification accuracy with CIFAR-100.

$\alpha$	Accuracy (%)	
	$\psi = \text{ResNet-20}$	$\psi = \text{VGG16}$
0.005	66.08	71.01
0.01	46.44	66.28
0.05	62.86	67.86
0.1	1.43	1.00
Plain image	67.72	71.90
Tanaka [18]	61.30	60.17
Pixel-based [19]	60.89	62.41
GAN-based [22]	50.02	53.93

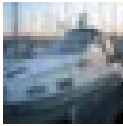
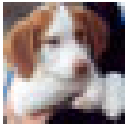




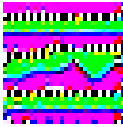
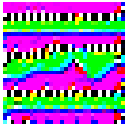
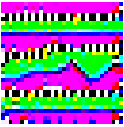
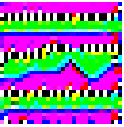
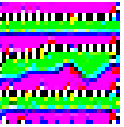
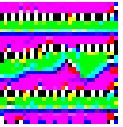
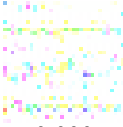

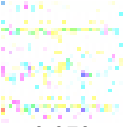
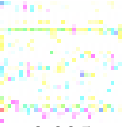
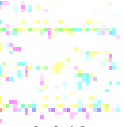
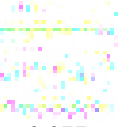

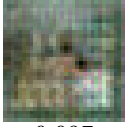
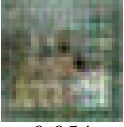



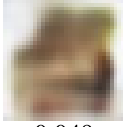
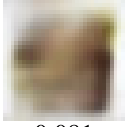
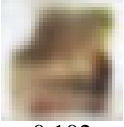
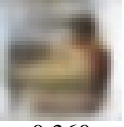
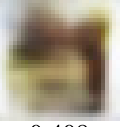
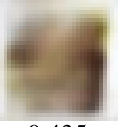
## B. EVALUATING ROBUSTNESS AGAINST ATTACKS

### 1) EXPERIMENTAL SETUP

In this simulation, transformed images were evaluated in terms of robustness against three state-of-the-art attacks: the feature reconstruction attack (FR-Attack) [39], the GAN-based attack (GAN-Attack) [38], and the inverse transformation network attack (ITN-Attack) [40]. For FR-Attack and GAN-Attack, we used the same settings as in [41]. For ITN-Attack, U-Net was used as an inverse transformation network, and the inverse transformation model was trained with  $h_\theta$  under the use of the CIFAR-10 dataset as shown in Fig. 6, where  $h_\theta$  was trained with ResNet-20 as  $\psi$ , and mean squared errors were used. Only random horizontal flip was performed as the data augmentation in the training. The other settings were the same as the training of transformation networks in IV-A.

### 2) ESTIMATING VISUAL INFORMATION WITH ATTACKS

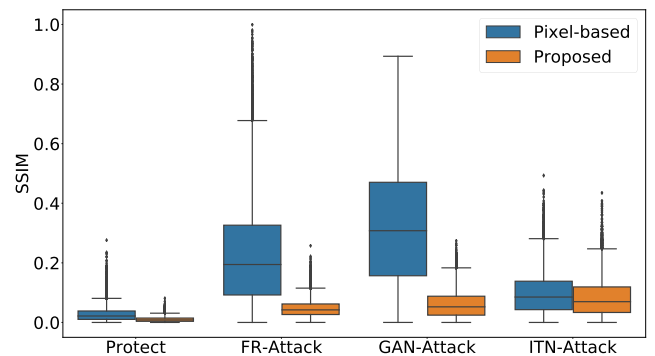
Figure 8 shows images restored from visually protected images generated by using three conventional protection methods: Tanaka's, pixel-based, and GAN-based methods

Attack method	Image 1	Image 2	Image 3	Image 4	Image 5	Image 6
Original						
Protected	 0.019	 0.009	 0.002	 0.014	 0.011	 0.007
FR-Attack [39]	 0.032	 0.048	 0.072	 0.025	 0.048	 0.077
GAN-Attack [38]	 0.152	 0.097	 0.054	 0.021	 0.026	 0.023
ITN-Attack [40]	 0.049	 0.081	 0.102	 0.360	 0.408	 0.435

**FIGURE 9.** Example of images restored from protected images generated by proposed transformation network (ResNet-20,  $\alpha = 0.05$ ). SSIM values are given under images.

[18], [19], [22], where structural similarity index measure (SSIM) [42] values between a plain image and the protected/estimated ones are also given under the images in order to evaluate the quality of the restored images. SSIM is a method for predicting the perceived quality of images, and it is used for measuring structural similarity between two images. The difference with other techniques such as mean squared error (MSE) or peak signal-to-noise ratio (PSNR) is that these approaches estimate absolute errors. The SSIM index is a real value between zero and one, where a value of one means that two images are identical, and a value of zero indicates no structural similarity.

From the reconstructed images, Tanaka’s method [18] was not robust against two attacks: GAN-attack and ITN-attack. For the pixel-based method [19], some visual information on the plain images was reconstructed by using FR-attack or GAN-attack. Similarly, the GAN-based method [22] was not robust against ITN-Attack. The GAN-based method is a model-based visual protection method as well as the proposed one, but it adopts the CycleGAN architecture [43], in which there is a process to convert a protected domain to a plain one. Therefore, ITN-Attack can reconstruct visual information on plain images from protected images. From the results,



**FIGURE 10.** SSIM values of estimated images. Boxes span from first to third quartile, referred to as  $Q_1$  and  $Q_3$ , and whiskers show maximum and minimum values in range of  $[Q_1 - 1.5(Q_3 - Q_1), Q_3 + 1.5(Q_3 - Q_1)]$ . Band inside box indicates median. Outliers are indicated as dots.

the conventional visual protection methods were not robust enough against the attacks.

In Fig. 9, visually protected images generated by using the proposed transformation network are shown together with the corresponding plain and estimated ones, where the transformation network was trained with ResNet-20 and  $\alpha = 0.05$ . From the figure, all estimated images were confirmed to have



almost no visual information on the plain images, even if the estimated images had slightly high SSIM values.

Figure 10 also shows scores calculated for the 10,000 images in the test set of CIFAR-10. The estimated images still had low SSIM values, so it was confirmed that the visual information of the plain images could not be restored from the protected images even when the state-of-the-art attacks were applied. Although some of the restored images had slightly high values, they also had almost no visual information on the plain images (see Fig. 9). We also confirmed that all of the restored images had no visual information on the plain images. Accordingly, the proposed transformation network was more robust against these attacks than the conventional methods.

## V. CONCLUSION

In this paper, we proposed a novel transformation network for generating visually protected images for privacy-preserving DNNs for the first time. The proposed network enables us not only to protect visual information on plain images but also to apply visually protected images to DNNs directly. In addition, there is no need to manage any security keys as the conventional methods require. In image classification experiments, visually protected images generated by the proposed network were demonstrated to have less visual information than those generated using the conventional methods, while maintaining the classification accuracy that using plain images achieves, under the use of the CIFAR-10 dataset and two classification networks: ResNet-20 and VGG16. We also confirmed that the visually protected images were robust against state-of-the-art attacks. As future work, we will consider applying visual protection methods to other tasks such as semantic segmentation and object detection.

## REFERENCES

- [1] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
- [2] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *Proc. 31st Int. Conf. Mach. Learn.*, vol. 32, 2014, pp. 647–655.
- [3] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Apr. 2015, pp. 1–5.
- [4] 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 Trans. Signal Inf. Process.*, vol. 3, pp. 1–17, Mar. 2014.
- [5] Q. Lu, X. Liao, H. Li, and T. Huang, "A computation-efficient decentralized algorithm for composite constrained optimization," *IEEE Trans. Signal Inf. Process. Over Netw.*, vol. 6, pp. 774–789, 2020.
- [6] Q. Lu, X. Liao, T. Xiang, H. Li, and T. Huang, "Privacy masking stochastic subgradient-push algorithm for distributed online optimization," *IEEE Trans. Cybern.*, early access, Mar. 9, 2020, doi: 10.1109/TCYB.2020.2973221.
- [7] I. Ito and H. Kiya, "One-time key based phase scrambling for phase-only correlation between visually protected images," *EURASIP J. Inf. Secur.*, vol. 2009, pp. 841045–841056, Dec. 2009.
- [8] B. Ferreira, J. Rodrigues, J. Leitao, and H. Domingos, "Privacy-preserving content-based image retrieval in the cloud," in *Proc. IEEE 34th Symp. Reliable Distrib. Syst. (SRDS)*, Sep. 2015, pp. 11–20.
- [9] J. Zhou, X. Liu, O. C. Au, and Y. Y. Tang, "Designing an efficient image encryption-then-compression system via prediction error clustering and random permutation," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 1, pp. 39–50, Jan. 2014.
- [10] Y. Zhang, B. Xu, and N. Zhou, "A novel image compression-encryption hybrid algorithm based on the analysis sparse representation," *Opt. Commun.*, vol. 392, pp. 223–233, Jun. 2017.
- [11] K. Kurihara, S. Imaizumi, S. Shiota, and H. Kiya, "An encryption-then-compression system for lossless image compression standards," *IEICE Trans. Inf. Syst.*, vol. 100, no. 1, pp. 52–56, 2017.
- [12] W. Sirichotedumrong and H. Kiya, "Grayscale-based block scrambling image encryption using YCbCr color space for encryption-then-compression systems," *APSIPA Trans. Signal Inf. Process.*, vol. 8, p. e7, Jan. 2019.
- [13] T. Chuman, W. Sirichotedumrong, and H. Kiya, "Encryption-then-compression systems using grayscale-based image encryption for JPEG images," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 6, pp. 1515–1525, Jun. 2019.
- [14] V. Itier, P. Puteaux, and W. Puech, "Recompression of JPEG crypto-compressed images without a key," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 3, pp. 646–660, Mar. 2020.
- [15] A. Kawamura, Y. Kinoshita, T. Nakachi, S. Shiota, and H. Kiya, "A privacy-preserving machine learning scheme using EtC images," *IEICE Trans. Fundamentals Electron., Commun. Comput. Sci.*, vol. 103, no. 12, pp. 1571–1578, 2020.
- [16] T. Maekawa, A. Kawamura, T. Nakachi, and H. Kiya, "Privacy-preserving support vector machine computing using random unitary transformation," *IEICE Trans. Fundamentals Electron., Commun. Comput. Sci.*, vol. 102, no. 12, pp. 1849–1855, 2019.
- [17] S. Beugnon, P. Puteaux, and W. Puech, "Privacy protection for social media based on a hierarchical secret image sharing scheme," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 679–683.
- [18] M. Tanaka, "Learnable image encryption," in *Proc. IEEE Int. Conf. Consum. Electron.-Taiwan (ICCE-TW)*, May 2018, pp. 1–2.
- [19] 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.
- [20] 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.
- [21] M. T. Gaata and F. F. Hantoosh, "An efficient image encryption technique using chaotic logistic map and RC4 stream cipher," *Int. J. Mod. Trends Eng. Res.*, vol. 3, no. 9, pp. 213–218, 2016.
- [22] 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.
- [23] A. Krizhevsky, "Learning multiple layers of features from tiny images," Univ. Toronto, Toronto, ON, Canada, Tech. Rep. TR-2009, 2009.
- [24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–14.
- [26] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–10.
- [27] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–11.
- [28] S. Baluja and I. Fischer, "Learning to attack: Adversarial transformation networks," in *Proc. AAAI*, 2018, pp. 2687–2695.
- [29] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroS&P)*, Mar. 2016, pp. 372–387.
- [30] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proc. ACM Asia Conf. Comput. Commun. Secur.*, Apr. 2017, pp. 506–519.
- [31] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–23.

- [32] M. Aprilpyone, Y. Kinoshita, and H. Kiya, "Adversarial robustness by one bit double quantization for visual classification," *IEEE Access*, vol. 7, pp. 177932–177943, 2019.
- [33] M. Maung, A. Pyone, and H. Kiya, "Encryption inspired adversarial defense for visual classification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 1681–1685.
- [34] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention—(MICCAI)*, 2015, pp. 234–241.
- [35] M. Chen and M. Wu, "Protect your deep neural networks from piracy," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2018, pp. 1–7.
- [36] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis. (Lecture Notes in Computer Science)*. Cham, Switzerland: Springer, 2016, pp. 694–711.
- [37] 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.
- [38] K. Madono, M. Tanaka, M. Onishi, and T. Ogawa, "An adversarial attack to learnable encrypted images," in *Proc. 22nd IEICE Symp. Image Recognit. Understand.*, 2019, pp. 1–4.
- [39] A. Habeen Chang and B. M. Case, "Attacks on image encryption schemes for privacy-preserving deep neural networks," 2020, *arXiv:2004.13263*. [Online]. Available: <http://arxiv.org/abs/2004.13263>
- [40] H. Ito, Y. Kinoshita, and H. Kiya, "Image transformation network for privacy-preserving deep neural networks and its security evaluation," in *Proc. IEEE 9th Global Conf. Consum. Electron. (GCCE)*, Oct. 2020, pp. 537–540.
- [41] 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.
- [42] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
- [43] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.



**HIROKI ITO** received the B.Eng. degree from Tokyo Metropolitan University, Japan, in 2020, where he is currently pursuing the master's degree. His research interest includes deep neural networks and their protection.



**YUMA KINOSHITA** (Member, IEEE) received the B.Eng., M.Eng., and Ph.D. degrees from Tokyo Metropolitan University, Japan, in 2016, 2018, and 2020, respectively. From April 2020, he worked as a Project Assistant Professor with Tokyo Metropolitan University. His research interests include signal processing, image processing, and machine learning. He is a member of APSIPA and IEICE. He received the IEEE ISPACS Best Paper Award, in 2016, the IEEE Signal Processing Society Japan Student Conference Paper Award, in 2018, the IEEE Signal Processing Society Tokyo Joint Chapter Student Award, in 2018, the IEEE GCCE Excellent Paper Award (Gold Prize), in 2019, and the IWAIT Best Paper Award, in 2020.



**MAUNGMAUNG APRILPYONE** (Graduate Student Member, IEEE) received the B.C.S. degree from the International Islamic University Malaysia, in 2013, under the Albukhary Foundation Scholarship, and the M.C.S. degree from the University of Malaya, in 2018, under the International Graduate Research Assistantship Scheme. He is currently pursuing the Ph.D. degree with Tokyo Metropolitan University, under the Tokyo Human Resources Fund for City Diplomacy Scholarship. His research interest includes neural networks and security. He received the IEEE ICCE-TW Best Paper Award, in 2016.



**HITOSHI KIYA** (Fellow, IEEE) received the B.E. and M.E. degrees from the Nagaoka University of Technology, in 1980 and 1982, respectively, and the Dr.Eng. degree from Tokyo Metropolitan University, in 1987.

...