A Framework for Transformation Network Training in Coordination with Semi-trusted Cloud Provider for Privacy-Preserving Deep Neural Networks

Hiroki Ito*, Yuma Kinoshita† and Hitoshi Kiya‡
Tokyo Metropolitan University, Tokyo, Japan
E-mail: {* ito-hiroki2@ed. † ykinoshita@, ‡ kiya@ }tmu.ac.jp
Tel/Fax: +81-42-585-8454

***Abstract*—We propose a framework for transformation network training in coordination with a semi-trusted cloud provider for privacy-preserving DNNs. In the framework, a user trains a transformation network using a model that a cloud provider has for transforming plain images into visually protected ones. Conventional perceptual encryption methods have a weak visual-protection performance and some accuracy degradation in image classification. In contrast, the proposed framework overcomes the two issues. In an image classification experiment, the transformation network trained under the framework is demonstrated to strongly protect the visual information of plain images, without any performance degradation under the use of two typical classification networks: ResNet and VGG. In addition, it is shown that the visually protected images are robust against a DNN-based attack.**

I. Introduction

The spread of deep neural networks (DNNs) has greatly contributed to solving complex tasks for many applications [1], [2], and the performance of these tasks is significantly improved [3]. Therefore, DNNs have been applied to privacy-sensitive/security-critical applications, such as facial recognition, biometric authentication, and medical image analysis.

Recently, it has been popular for cloud providers to offer high-performance services using DNNs in cloud environments, like software as a service (SaaS). However, since cloud providers are semi-trusted, 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. In this paper, we focus on protecting visual information by transforming images before uploading them to cloud environments.

Various perceptual encryption methods have been proposed for generating images without visual information [5]–[20], although information theory-based encryption (like RSA and AES) generates a 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. However, most perceptual encryption methods cannot be applied to DNNs. There are only three methods: Tanaka's method [16], a pixel-wise encryption method [17]–[19], and a method using an image transformation network trained with a generative adversarial network [20], for privacy-preserving DNNs. However, the use of these methods causes degradation in the performance of DNNs due to the effect of visual information protection [18]. In addition, they are not robust enough against various attacks including DNN-based attacks [21]–[23].

For such reasons, we propose an image transformation network trained by both a user (client/third party) and a cloud provider, where the cloud provider provides a part of the model information to train the transformation network. The transformation network enables us not only to protect visual information on plain images but also to maintain the high performance of DNNs. The transformation network has no secret keys as a hash function, so it can be open to the public.

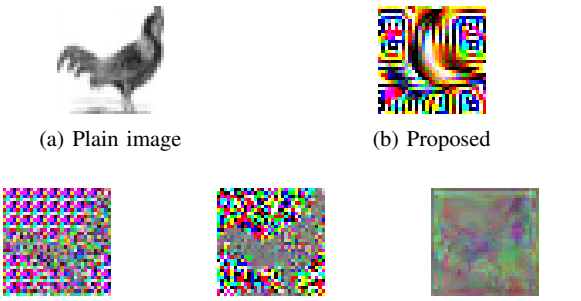
In an experiment, image classification is carried out under the use of the CIFAR-10 [24] dataset and two classification networks: ResNet-20 [25] and VGG16 [26] with batch normalization, to evaluate the effectiveness of the transformation network. From the results, the proposed transformation network is demonstrated not only to strongly protect visually information on plain images, but also to maintain high classification accuracy that using plain images achieve. In addition, the protected images are confirmed to be robust enough against attacks, even when a strong DNN-based attack is applied to the protected ones.

II. Related Work

Conventional visual protection methods are summarized here.

A. Protecting Visual Information

A transformation network performs visual protection on plain images for privacy-preserving DNNs. A lot of perceptual encryption methods have been proposed for protecting the visual information of images [5]–[20]. 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 [5], [6] 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



(a) Plain image (b) Proposed

(c) Tanaka [16] (d) Pixel-based [17] (e) GAN-based [20]

Fig. 1: Example of visually-protected images

that consider both security and efficient compression so that they can be adapted to cloud storage and network sharing [7]–[12]. Some of them [9]–[11] 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 [13], [14]. However, these methods have never been applied to DNNs.

B. Visual Protection Methods for DNNs

Three methods [16]–[20] for generating visually-protected images have been proposed for privacy-preserving DNNs. The first is Tanaka's method [16], which utilizes an adaptation network prior to DNNs to reduce the influence of image encryption. The second is a pixel-wise encryption method [17]–[19]. The third is to use an image transformation network trained with a generative adversarial network (GAN) [20]. However, these methods cause a decrease in classification accuracy. In addition, they are not robust enough against various attacks.

Figure 1 shows an example of visually-protected images. In this paper, a transformation network trained with a model is demonstrated to overcomes the above two issues: accuracy degradation and visual protection robustness that the conventional methods have.

III. Proposed Framework for Transformation Network Training

A. Overview

Figure 2 illustrates the framework for privacy-preserving DNNs using a transformation network trained with a classification model. In this framework, the cloud provider is not trusted, and cannot get any visual information on plain images. An overview of the proposed framework is given here.

① A cloud provider trains a classification model (ψ) using training images (X) and corresponding correct labels (Y).
② The cloud provider provides images ($X' \subseteq \{X\}$) to a user, i.e. a client or a third party.
③ The user trains a transformation network (h_θ) by using X' in coordination with the cloud provider. The detailed procedure will be given in Sec. III-B.

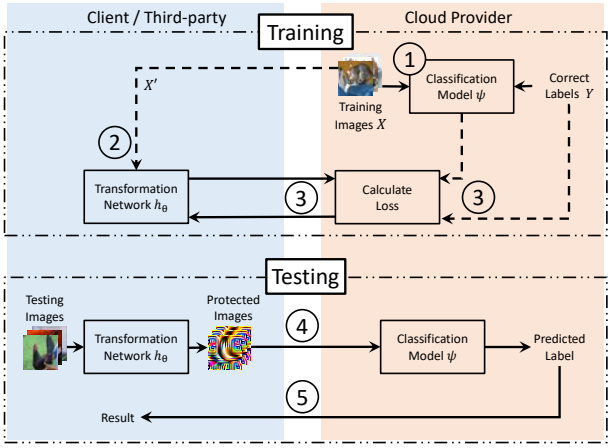


Fig. 2: Framework of proposed scheme

④ The user transforms test images into visually protected images by using h_θ. The protected images are sent to the cloud provider.
⑤ The cloud provider classifies the protected images by using ψ, and then predicted labels are returned to the user.

In this framework, cloud providers have no visual information on test images, and transformation network h_θ can be open to the public. Therefore, there is no need to manage secret keys that conventional methods have. The use of the network h_θ enables users to securely use model ψ with high performance, which cannot be prepared by themselves. In other words, cloud providers can securely provide some web-based software services like software as a service (SaaS). In addition, they can also carry out image classification without distinction between plain images and transformed ones.

B. Training Transformation Network

The training procedure of transformation network h_θ is illustrated in Fig. 3, where $X' = \{x_1, x_2, \ldots, x_m\}$ is an sub set of training images, i.e. $X' \subseteq \{X\}$, $\hat{X}' = \{\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m\}$ is an output image set from the transformation network, i.e. $\hat{x}_i = h_\theta(x_i)$, $Y = \{y_1, y_2, \ldots, y_m\}$ is a one-hot encoded target label set, and $\hat{Y} = \{\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_m\}$ is an output label set from a classification network, i.e. $\hat{y}_i = \psi(\hat{x}_i)$. One-hot encoded label $y_i = (y_i(1), y_i(2), \ldots, y_i(c))$ and output label $\hat{y}_i = (\hat{y}_i(1), \hat{y}_i(2), \ldots, \hat{y}_i(c))$ meet, respectively,

$$y_i(j) \in \{0, 1\}, \text{and} \sum_{j=1}^{c} y_i(j) = 1, \tag{1}$$

and

$$0 \leq \hat{y}_i(j) \leq 1, \text{and} \sum_{j=1}^{c} \hat{y}_i(j) = 1, \tag{2}$$

where c is the number of classes, and a user has training images X' without labels. Network h_θ is trained for reducing the loss value of classification network ψ so that visually-protected images by using h_θ are classified correctly.

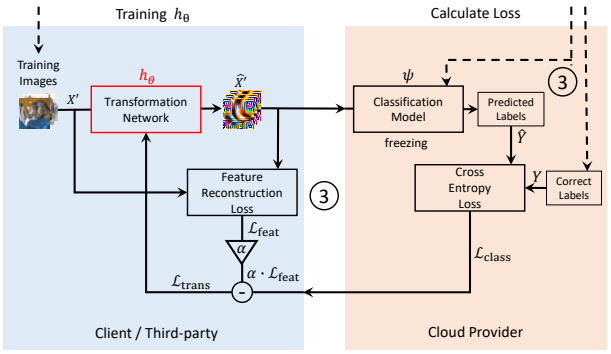


Fig. 3: Training process of transformation network h_θ

To train network h_θ with parameter θ by using input image x_i and its target label y_i, loss function $\mathcal{L}_{\text{trans}}$ is minimized as

$$\underset{\theta}{\text{minimize}} \quad \frac{1}{m}\sum_{i=1}^{m}\mathcal{L}_{\text{trans}}(x_i, h_\theta(x_i), y_i). \tag{3}$$

$\mathcal{L}_{\text{trans}}$ is defined as

$$\mathcal{L}_{\text{trans}}(x_i, \hat{x}_i, y_i) = \mathcal{L}_{\text{class}}(\hat{x}_i, y_i) - \alpha \cdot \mathcal{L}_{\text{feat}}(x_i, \hat{x}_i)), \tag{4}$$

where $\mathcal{L}_{\text{class}}$ denotes a cross entropy loss function, which is used to classify protected images $\hat{X}$ correctly, $\mathcal{L}_{\text{feat}}$ is a feature reconstruction loss function to protect visual information on plain images, and $\alpha \in \mathbb{R}$ is a weight of $\mathcal{L}_{\text{feat}}$. Note that h_θ and $\mathcal{L}_{\text{feat}}$ are calculated by the user. In contrast, $\mathcal{L}_{\text{class}}$ is calculated by the cloud provider.

$\mathcal{L}_{\text{class}}$ is calculated by using $\hat{y}_i(j)$, as

$$\mathcal{L}_{\text{class}}(\hat{x}_i, y_i) = -\sum_{j=1}^{c} y_i(j)\log \hat{y}_i(j), \tag{5}$$

$\mathcal{L}_{\text{feat}}$ is also given by

$$\mathcal{L}_{\text{feat}}(x_i, \hat{x}_i) = \frac{1}{C_k H_k W_k}\|\phi_k(\hat{x}_i) - \phi_k(x_i)\|_2^2, \tag{6}$$

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 [27].

C. Testing Protected Images with ψ

In the proposed framework, a user converts plain test images into protected images by using network h_θ, and the protected images are classified by using model ψ. Therefore, the cloud provider can also provide services to users who do not worry about the protection of visual information or cannot prepare the computing cost used for image transformation with h_θ, without any modification of the model.

D. Robustness against Inverse-Transformation Network Attack

The proposed transformation network has no security keys, so it is robust against brute-force attacks. However, other ciphertext-only attacks might be applied to images protected by using h_θ. In particular, an attack using exact pairs of plain images and the corresponding protected ones, called inverse-transformation network attack (ITN-Attack), can be applied

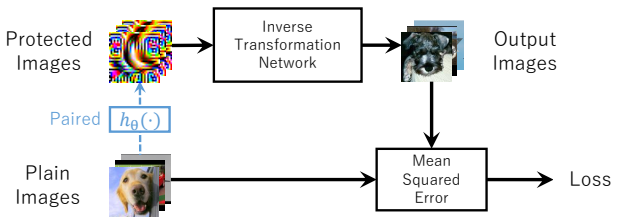


Fig. 4: Training of inverse transformation model used in this paper

to the protected one, because attackers can easily prepare a correct pair of a plain image and the protected image by using h_θ, which is open to the public. Therefore, attackers can easily create an inverse transformation network by using correct pair images, to estimate visual information on input images, as shown in Fig. 2. Protected images will be shown to be robust enough against this attack in an experiment.

IV. Simulations

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

A. Classification Accuracy

1) Experimental Setup: We used the CIFAR-10 dataset [24] to evaluate the effectiveness of the image transformation networks. ResNet-20 [25] and VGG16 [26] with batch normalization were used as classification model ψ, and we used U-Net [28] as network h_θ. Also, the second ReLU function of VGG16 without batch normalization pretrained with ImageNet was used for ϕ_k, i.e., a feature map.

The CIFAR-10 dataset [24] consists of a training set with 50,000 images and a test set with 10,000. We utilized 45,000 images in the training set to train both ψ and h_θ, and the other 5,000 images were used as validation data. The test set of CIFAR-10 was also utilized for evaluating the performance of the networks. In addition, standard data-augmentation methods, i.e., random crop and horizontal flip, were performed in the training.

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 network that provided the lowest loss value under the use of the validation set.

2) Visual-protection Performance: Figure 5 shows an example of visually protected images generated from four test images in CIFAR-10 by using h_θ trained with ResNet-20 for calculating $\mathcal{L}_{\text{class}}$, where the top row shows plain images, and the second top row to bottom row show images generated with the parameters $\alpha = 0$, 0.001, 0.005, and 0.01 in Eq. (4).

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

(a) Plain

(b) Protected ($\alpha = 0$)

(c) Protected ($\alpha = 0.001$)

(d) Protected ($\alpha = 0.005$)

(e) Protected ($\alpha = 0.01$)

Fig. 5: Visually protected images generated by proposed transformation network trained with ResNet-20

protected since $\mathcal{L}_{\text{feat}}$, i.e., a loss for visually protecting input images, did not work. From Fig. 5, it was confirmed that the visual protection was more enhanced when using a larger α value. 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 like a robust hash function. The transformation network was shown to strongly protect visual information on plain images.

3) Classification Accuracy: Table I shows the experimental results of the image classification experiment, where "ResNet-20" and "VGG16" mean that each model was used for ψ.

From the table, when $\alpha \leq 0.01$, the protected images provided a higher classification accuracy than the conventional methods. In particular, in the case of "ResNet-20," the proposed network offered higher accuracy values than in the case of using the plain images ("Plain image") because the proposed framework increases the total number of parameters due to the use of h_θ.

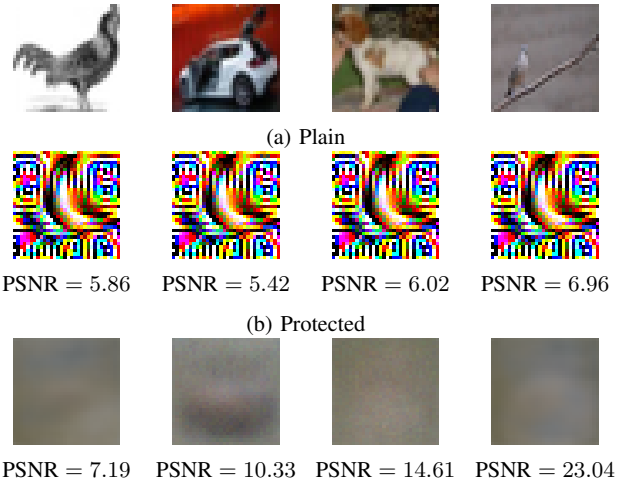
B. Evaluating Robustness against ITN-Attack

1) Experimental Setup: In this simulation, we used U-Net [28] as an inverse transformation network. The inverse transformation network was trained by using h_θ, as shown in Fig. 4, when h_θ was trained with ResNet-20 as ψ.

The CIFAR-10 dataset was used for the training and the testing. Only random horizontal flip was performed as the data augmentation in the training. The other settings were the same as in IV-A. We used the mean squared error as the loss function for training the inverse network, which was calculated by using the relation between output images from the inverse transformation network and the corresponding plain ones as shown in Fig. 4.

TABLE I: Classification accuracy (%)

Method		ψ = ResNet-20	ψ = VGG16
proposed	$\alpha = 0.005$	**91.72**	**91.94**
	$\alpha = 0.01$	**91.41**	**91.13**
	$\alpha = 0.05$	89.63	18.67
	$\alpha = 0.1$	39.92	18.99
Plain image		91.23	92.23
Tanaka [16]		85.18	85.79
Pixel-based [17], [18]		90.99	90.29



(a) Plain

PSNR = 5.86 PSNR = 5.42 PSNR = 6.02 PSNR = 6.96

(b) Protected

PSNR = 7.19 PSNR = 10.33 PSNR = 14.61 PSNR = 23.04

(c) Estimated

Fig. 6: Images estimated by inverse transformation model with h_θ trained with ResNet-20 and $\alpha = 0.005$

2) Robustness against ITN-Attack: In Fig. 6, images estimated by using the inverse transformation network are illustrated together with the corresponding plain images and the visually protected ones. The inverse transformation network was trained by using h_θ trained with $\alpha = 0.005$. Obtained in order to evaluate the quality of the estimated images, peak signal-to-noise ratio (PSNR) values between the estimated and plain images are also given in the figure. From Fig. 6, the estimated images were confirmed to have slightly higher PSNR values than those of the protected ones, but the estimated images had almost no visual information on the plain images yet.

Figure 7 illustrates PSNR values calculated by using the 10,000 images in the test set of CIFAR-10. The figure shows that the estimated images still had low PSNR values. In addition, all of the 10,000 estimated images were confirmed to have no visual information on plain images as well as in Fig. 6.

From these results, it can be seen that the visually protected images were robust against ITN-attack.

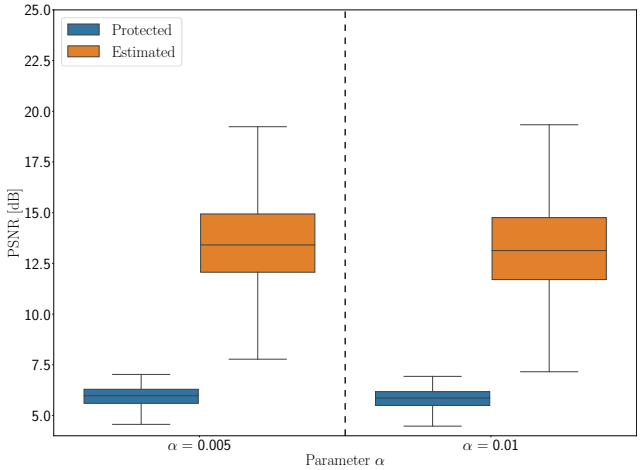


Fig. 7: PSNR values of estimated images. Boxes span from first to third quartile, referred to as Q_1 and Q_3, and whiskers show maximum and minimum values in range of $[Q_1 - 1.5(Q_3 - Q_1), Q_3 + 1.5(Q_3 - Q_1)]$. Band inside box indicates median. Outliers are not indicated.

V. Conclusion

In this paper, we proposed the framework for transformation network training in coordination with a cloud provider for privacy-preserving DNNs. The framework enables a user to securely train a transformation network with a model that a cloud provider has. In image classification experiments, visually protected images generated by the network were demonstrated to generate strongly-protected images, while maintaining high 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 ITN-Attack.

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 *Proceedings of the 31st International Conference on Machine Learning - Volume 32*, 2014, pp. 647–655.

[3] N. Tishby and N. Zaslavsky, "Deep learning and the information bottleneck principle," *2015 IEEE Information Theory Workshop (ITW)*, pp. 1–5, 2015.

[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 Transactions on Signal and Information Processing*, vol. 3, p. e7, 2014.

[5] I. Ito and H. Kiya, "One-time key based phase scrambling for phase-only correlation between visually protected images," *EURASIP Journal on Information Security*, pp. 841 045–841 056, Dec. 2009.

[6] B. Ferreira, J. Rodrigues, J. Leitao, and H. Domingos, "Privacy-preserving content-based image retrieval in the cloud," in *2015 IEEE 34th Symposium on Reliable Distributed Systems (SRDS)*, Sept. 2015, pp. 11–20.

[7] J. Zhou, X. Liu, O. Au, and Y. Tang, "Designing an efficient image encryption-then-compression system via prediction error clustering and random permutation," *Information Forensics and Security, IEEE Transactions on*, vol. 9, pp. 39–50, Jan. 2014.

[8] Y. Zhang, B. Xu, and N. Zhou, "A novel image compression-encryption hybrid algorithm based on the analysis sparse representation," *Optics Communications*, vol. 392, no. C, pp. 223–233, 2017.

[9] K. Kurihara, S. Imaizumi, S. Shiota, and H. Kiya, "An encryption-then-compression system for lossless image compression standards," *IEICE Transactions on Information and Systems*, vol. E100.D, no. 1, pp. 52–56, 2017.

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

[11] 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, June 2019.

[12] V. Itier, P. Puteaux, and W. Puech, "Recompression of jpeg crypto-compressed images without a key," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 3, pp. 646–660, 2020.

[13] T. Maekawa, A. Kawamura, 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.

[14] A. Kawamura, Y. Kinoshita, T. Nakachi, S. Shiota, and H. Kiya, "A Privacy-Preserving Machine Learning Scheme Using EtC Images," *arXiv e-prints*, p. arXiv:2007.08775, Jul. 2020.

[15] S. Beugnon, P. Puteaux, and W. Puech, "Privacy protection for social media based on a hierarchical secret image sharing scheme," in *2019 IEEE International Conference on Image Processing (ICIP)*, Sept. 2019, pp. 679–683.

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

[17] 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)*, Sept. 2019, pp. 674–678.

[18] 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. 177 844–177 855, 2019.

[19] M. T. 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.

[20] W. Sirichotedumrong and H. Kiya, "A GAN-Based Image Transformation Scheme for Privacy-Preserving Deep Neural Networks," in *28th European Signal Processing Conference (EUSIPCO)*, 2020, pp. 745–749.

[21] W. Sirichotedumrong, Y. Kinoshita, and H. Kiya, "On the security of pixel-based image encryption for privacy-preserving deep neural networks," in *2019 IEEE 8th Global Conference on Consumer Electronics (GCCE)*, 2019, pp. 121–124.

[22] A. Habeen Chang and B. M. Case, "Attacks on Image Encryption Schemes for Privacy-Preserving Deep Neural Networks," *arXiv e-prints*, p. arXiv:2004.13263, Apr. 2020.

[23] K. Madono, M. Tanaka, M. Onishi, and T. Ogawa, "An adversarial attack to learnable encrypted images," in *22nd IEICE Symposium on Image Recognition and Understanding*, July 2019.

[24] A. Krizhevsky, "Learning multiple layers of features from tiny images." Tech.Rep., 2009. [Online] Available: https://www.cs.toronto.edu/~kriz/cifar.html

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations*, 2015.

[27] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," *Lecture Notes in Computer Science*, p. 694–711, 2016.

[28] 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*, 2015, pp. 234–241.