information

Article

A Jigsaw Puzzle Solver-Based Attack on Image Encryption Using Vision Transformer for Privacy-Preserving DNNs

Tatsuya Chuman and Hitoshi Kiya *

Department of Computer Science, Tokyo Metropolitan University, 6-6 Asahigaoka, Hino-shi 191-0065, Tokyo, Japan; chuman-tatsuya1@ed.tmu.ac.jp
* Correspondence: kiya@tmu.ac.jp; Tel.: +81-42-585-8454

Abstract: In this paper, we propose a novel attack on image encryption for privacy-preserving deep neural networks (DNNs). Although several encryption schemes have been proposed for privacy-preserving DNNs, existing cipher-text-only attacks (COAs) have succeeded in restoring visual information from encrypted images. Image encryption using the Vision Transformer (ViT) is known to be robust against existing COAs due to the operations of block scrambling and pixel shuffling, which permute divided blocks and pixels in an encrypted image. However, the correlation between blocks in the encrypted image can still be exploited for reconstruction. Therefore, in this paper, a novel jigsaw puzzle solver-based attack that utilizes block correlation is proposed to restore visual information from encrypted images. In the experiments, we evaluated the security of image encryption for privacy-preserving deep neural networks using both conventional and proposed COAs. The experimental results demonstrate that the proposed attack is able to restore almost all visual information from images encrypted for being applied to ViTs.

Keywords: image encryption; jigsaw puzzle solver; vision transformer; privacy preserving

check for updates

Citation: Chuman, T.; Kiya, H. A Jigsaw Puzzle Solver-Based Attack on Image Encryption Using Vision Transformer for Privacy-Preserving DNNs. *Information* **2023**, *14*, 311. https://doi.org/10.3390/info14060311

Academic Editor: Xin Ning

Received: 3 April 2023
Revised: 7 May 2023
Accepted: 25 May 2023
Published: 29 May 2023

1. Introduction

In recent years, the rapid development of deep neural networks (DNNs) has made it possible to perform complex tasks such as speech recognition and image classification with a high accuracy [1]. Simultaneously, the utilization of cloud computing services, such as Microsoft Azure, Amazon Web Services (AWS), and Google Cloud Platform (GCP), has been on the rise because of their cost-effectiveness and simplicity of deployment. Furthermore, the appearance of cloud Automated Machine Learning (AutoML) has made it possible for users to obtain accurate results without having extensive knowledge on which algorithms would be suitable to obtain the results; for example, a user can easily get a classification result by using Cloud Vision API. However, since cloud servers are completely untrusted, an image including privacy information such as fingerprints and radiographs tends to be processed on the premises due to the risk of data leakage [2]. In light of the increasing use of cloud computing services for processing sensitive data, there is a pressing need to ensure the protection of data privacy in cloud environments. Although full encryption with provable security such as RSA and AES is the most secure option for securing multimedia data, there is a trade-off between security and other requirements such as low processing demand, bitstream compliance, and signal processing in the encrypted domain. Several perceptual encryption schemes have been developed to achieve this trade-off [3–8]. On the other hand, for protecting data privacy in a cloud server, privacy-preserving DNNs for image classification are proposed [9,10]. The use of learnable image encryption for privacy-preserving DNNs enables us to protect personally identifiable information in an image such as fingerprints and facial information. Moreover, encrypted images can be applied to machine learning algorithms in the encrypted domain. In this paper, we focus on protecting visual information in an image by encrypting it before uploading to the cloud environment.

Although various perceptual encryption schemes for privacy-preserving DNNs have been proposed to protect the visual information of images [11–13], state-of-the-art ciphertext-only attacks (COAs) succeed in reconstructing visual information from encrypted images [14–17]. Therefore, encryption schemes that are robust against various attacks are essential for privacy-preserving DNNs without degrading image classification performance.

On the other hand, image encryption using isotropic networks such as the Vision Transformer (ViT) [18], a model for image classification based on the transformer architecture, has been shown to outperform conventional methods in terms of classification [19,20]. Moreover, this scheme enhances robustness against attacks by permuting divided blocks in an encrypted image, an operation called block scrambling [19]. As a way of restoring visual information from encrypted images including permuted blocks, jigsaw puzzle solver attacks have been proposed that utilize the correlation between permuted blocks to be assembled [21]. Jigsaw puzzle solvers have succeeded in solving very large jigsaw puzzles, such as 30,745 piece puzzles, by using genetic algorithms [21]. Besides, jigsaw puzzles with the small size of pieces, such as 7×7 pixels, are partially assembled by utilizing hierarchical piece loops [22]. It has been confirmed that the use of an extended jigsaw solver enables us to restore visual information from encrypted images with permuted blocks, including permuted, rotated, inverted, negative–positive transformed, and RGB shuffled blocks [23]. Therefore, if the operation of block scrambling is included in image encryption for privacy-preserving DNNs, the security against jigsaw solver attacks need to be evaluated.

On the other hand, the latest encryption scheme for application to ViTs is known to be robust against jigsaw puzzle solver attacks because of applying not only block scrambling but also pixel shuffling, which permutes the pixels in divided blocks [20]. However, since the operation of pixel shuffling in the scheme is performed using a common secret key for all blocks, an attack that solves pixel shuffling using edge information in each block is assumed. Therefore, in this paper, we propose a novel jigsaw puzzle solver-based attack to restore visual information from images encrypted for being applied to ViTs. The proposed attack is feasible even for other block-based image encryption methods when a common key is applied to all divided blocks. A part of this work was presented in [24]. In this paper, in particular, we compare the proposed method with conventional attacks [14–17] under various conditions in terms of the restoration of visual information from encrypted images. The conventional attacks are utilized to show the attacks cannot restore visual information from images encrypted for being applied to ViTs. The contribution of this work is that we propose an attack that enables the restoration of almost all visual information from images encrypted for being applied to ViTs [20]. In experiments, the security of image encryption for privacy-preserving DNNs is evaluated by using the proposed attack and five conventional COAs: a brute force attack (BF-attack), a feature reconstruction attack (FR-attack) [14], an inverse transformation network attack (ITN-attack) [15], a jigsaw puzzle solver attack (JPS-attack) [21], and an extended jigsaw puzzle solver attack (EJPS-attack) [23].

The rest of this paper is organized as follows. Section 2 provides an overview of privacy-preserving DNNs and the encryption schemes used for them. Section 3 presents the proposed attack for image encryption for privacy-preserving DNNs in addition to the conventional attacks. The experimental results including robustness against the proposed and conventional attacks are given in Section 4. Finally, Section 5 concludes this paper.

2. Preparation

Privacy-preserving DNNs for image classification and the image encryption used for them are summarized in this section.

2.1. Privacy-Preserving DNNs

Privacy-preserving DNNs for image classification are carried out as illustrated in Figure 1.

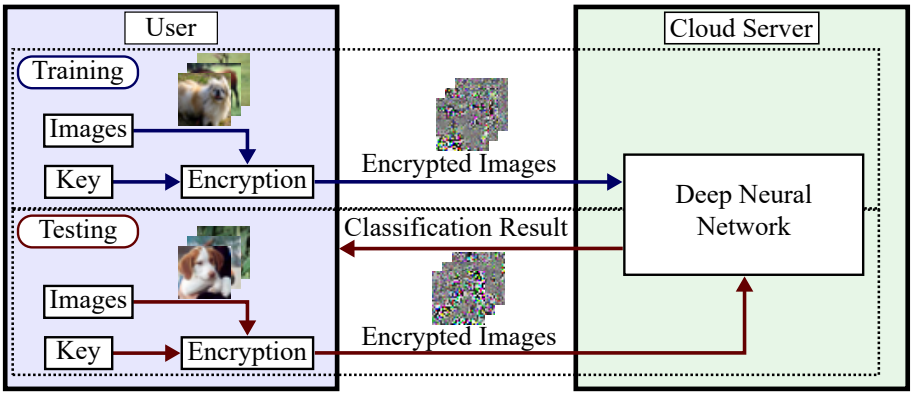


Figure 1. Privacy-preserving deep neural networks for image classification.

In training, a user encrypts training images by using image encryption for privacy-preserving DNNs [11–13] and sends them to a cloud server. The cloud server trains an image classification model using the visually protected images, that is, the privacy of the images is preserved. On the other hand, for testing, the user sends an encrypted testing image to the cloud server and gets classification results. Depending on the type of image encryption, the testing images can be encrypted by using a different secret key from the one used for training. The visually protected images are used in both training and testing, thereby protecting the privacy of images in the case of data leakage on the cloud server. However, several attacks have been proposed to reconstruct visual information from encrypted images. Therefore, the security of the encryption scheme needs to be discussed in addition to its classification performance.

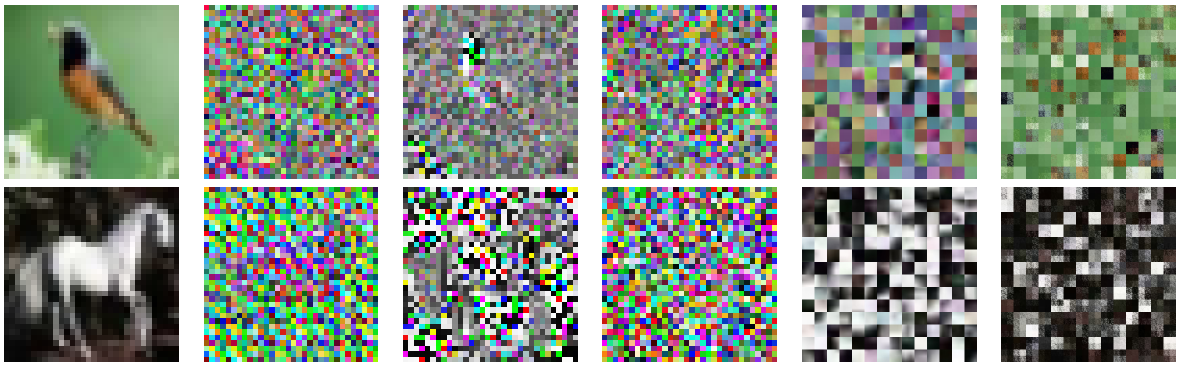
Convolutional neural networks (CNNs), such as VGG [25] and ResNet [26], are mostly used for privacy-preserving image classification. It has been known that an encryption scheme using block scrambling, which permutes divided blocks, enhances the robustness against several attacks. However, permuting the positions of divided blocks in encrypted images degrades the classification performance. Therefore, block scrambling cannot be applied to image encryption for CNN-based privacy-preserving DNNs.

On the other hand, image encryption using isotropic networks such as the ViT, a model for image classification based on the transformer architecture, has been proposed to enhance the security and classification performance [27]. The use of the ViT enables us to apply block scrambling to image encryption for privacy-preserving image classification because it includes an operation for patch embedding. For example, images from the CIFAR-10 dataset are resized from 32×32 to 224×224 or 384×384 and then divided into a 16×16 patch to fit the same patch size of pretrained models such as the ViT-B/16 and ViT-L/16. Furthermore, the latest image encryption using the ViT not only uses block scrambling but also pixel shuffling that permutes the pixels in divided blocks, which makes the encryption scheme more robust [20]. In this paper, we aim to evaluate the security of image encryption using the ViT.

2.2. Image Encryption for Privacy-Preserving DNNs

There are two types of image encryption for privacy-preserving DNNs, image encryption without permutations [11,12] and image encryption with permutations [13,20,28], as shown in Table 1.

Table 1. Summary of image encryption for privacy-preserving deep neural networks. √ indicates image encryption using block scrambling. $X \times Y$ indicates the resolution of encrypted image. n denotes number of blocks in encrypted image and M indicates block size.

	Plain	**LE [11]**	**PE [12]**	**ELE [13]**	**EtC [28]**	**VTE [20]**
Block scrambling				√	√	√
Key type		Same	Different	Different	Same	Same
Key space		$(M^2 \cdot 6)! \cdot 2^{M^2 \cdot 6}$	$2^{3 \cdot X \cdot Y} \cdot 6^{X \cdot Y}$	$((M^2 \cdot 6)! \cdot 2^{M^2 \cdot 6})^n \cdot n!$	$8^n \cdot 2^n \cdot 6^n \cdot n!$	$n! \cdot (\frac{M}{2})^2!$
Example						

This paper focuses on encryption for a 24-bit RGB color image I with $X \times Y$ pixels. Each type of encryption is carried out by dividing I into blocks with $M \times M$ pixels.

Tanaka first proposed learnable image encryption (LE), which uses an adaptation network to reduce the effect of encryption, thus maintaining a high classification performance [11]. In this scheme, a secret key is shared between training and testing; thus, proper secret key management is needed to prevent the secret key from leaking. LE images are generated by following the procedure below.

1. Divide the RGB color image I into blocks of $M \times M$ pixels.
2. Separate each pixel into upper and lower 4 bit pixel values to form six-channel blocks.
3. Reverse the intensities of the pixel values in each block randomly by a secret key.
4. Shuffle the pixel values in each block randomly by a secret key.
5. Combine six channels in each block into three channel to generate an encrypted image.

To enhance the security of LE, extended learnable image encryption (ELE) was proposed [13]. The procedure of ELE is described below.

1. Divide RGB color image I into blocks of $M \times M$ pixels.
2. Permute randomly the divided blocks by using a secret key.
3. The same procedure of LE is applied to the permuted blocks to generate an encrypted image.

This encryption scheme enables us to use different secret keys between training and testing; thus, there is no need for secret key management. Moreover, the security of this encryption scheme has been enhanced by using different keys for training and testing. Although ELE enhances security owing to the addition of block scrambling in comparison to LE, the classification performance of ELE is lower than that of LE. Furthermore, an adaptation network is required to keep the classification accuracy the same as LE.

On the other hand, pixel-wise image encryption (PE) was proposed, which combines negative–positive transformations and color component shuffling [12]. This encryption scheme enables us to use different secret keys between training and testing, the same as ELE. In this encryption method, the following steps are carried out.

1. Divide RGB color image I into $X \times Y$ pixels.
2. Apply negative–positive transformations to each pixel of the three color channels randomly by using a secret key. In this scheme, the secret key is independently used for all color components. In this step, a pixel q is transformed to q' by

$$q' = \begin{cases} q & (r(i) = 0) \\ q \oplus (2^8 - 1) & (r(i) = 1) \end{cases}, \quad (1)$$

   where $r(i)$ is a random binary integer generated by the secret key. In this paper, the value of occurrence probability $P(r(i)) = 0.5$ has been used to invert bits randomly.
3. Shuffle three color components of each pixel by using a secret key.
4. Combine $X \times Y$ pixels to generate an encrypted image.

Block-scrambling-based image encryption has been proposed for encryption-then-compression (EtC) systems transmitted over an untrusted channel provider. An image encrypted using this scheme is referred to as an EtC image [28]. The procedure of the EtC scheme is given as follows (see Figure 2).

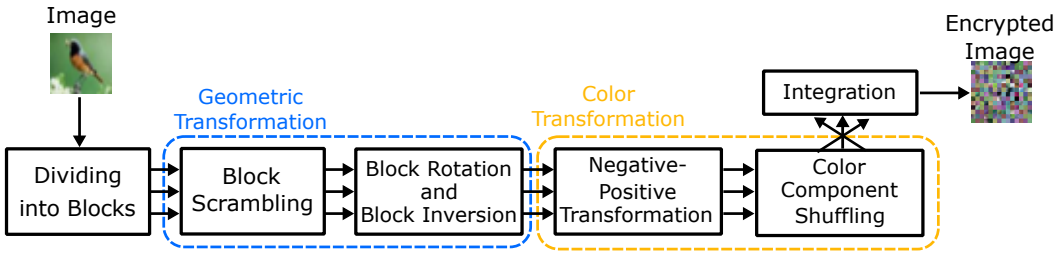


Figure 2. Block-scrambling-based image encryption (EtC) [28].

1. Divide RGB color image I into blocks of $M \times M$ pixels.
2. Permute randomly the divided blocks using a secret key.
3. Rotate and invert randomly each block by using a secret key.
4. Apply negative–positive transformations to each block by using a secret key according to Equation (1).
5. Shuffle three color components of each block by using a secret key.
6. Integrate the encrypted blocks to generate an encrypted image.

Note that the secret keys are commonly used for all color components. Although the classification performance when using EtC images is low, the use of isotropic networks such as the ViT improves it. As in the original paper [27], an EtC image is generated after resizing to 224×224 or 384×384 pixels to fit the resolution of a pretrained model.

A vision-transformer-based image encryption (VTE) has been proposed that combines block scrambling and pixel shuffling [20]. Similar to EtC images, the VTE scheme is carried out after an image is resized. Figure 3 shows the procedure of VTE. The procedure for performing this encryption scheme is given as follows.

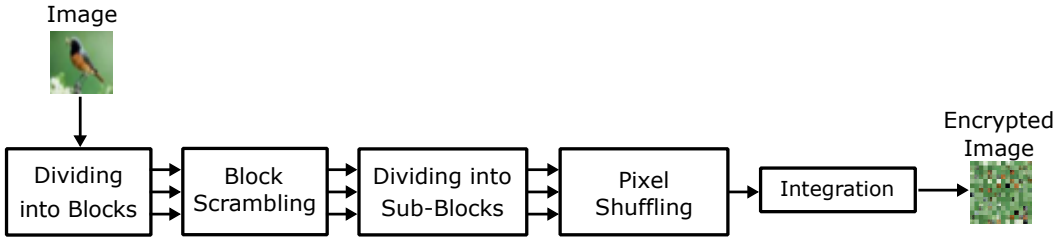


Figure 3. Vision-transformer-based image encryption (VTE) [20].

1. Divide an RGB color image I into blocks $B = \{B_1, \ldots, B_i, \ldots, B_n\}$, $i \in \{1, \ldots, n\}$ with $M \times M$ pixels, where n is the number of divided blocks calculated by

$$n = \lfloor \frac{X}{M} \rfloor \times \lfloor \frac{Y}{M} \rfloor. \quad (2)$$

2. Permute the divided blocks by using a secret key K_{VTE1}, where K_{VTE1} is commonly used for all color components. Accordingly, the scrambled blocks $B' = \{B'_1, \ldots, B'_i, \ldots, B'_n\}$ are generated.
3. Divide each scrambled block B'_i into four non-overlapping square sub-blocks $S_{ij}, j \in \{UL, UR, LL, LR\}$ with $\frac{M}{2} \times \frac{M}{2}$ pixels, where S_{iUL} is defined as the upper left position of the ith blocks, S_{iUR} as the upper right, S_{iLL} as the lower left, and S_{iLR} as the lower right. Thereby, scrambled blocks divided into sub-blocks $S = \{S_{1j}, \ldots, S_{ij}, \ldots, S_{nj}\}$ are generated. The number of sub-blocks m is described as

$$m = 4n. \quad (3)$$

4. Shuffle the pixel position within a sub-block by using a secret key K_{VTE2} to generate pixel shuffled sub-blocks $S' = \{S'_{1j}, \ldots, S'_{ij}, \ldots, S'_{nj}\}$, where K_{VTE2} is commonly used for all sub-blocks and color components. As a result, each scrambled block is divided into four encrypted sub-blocks, denoted by $S'_i = \{S'_{iUL}, S'_{iUR}, S'_{iLL}, S'_{iLR}\}$.
5. Merge all blocks to generate an encrypted image.

In this paper, $M = 4$ is used for LE and ELE images, while $M = 16$ is used for the EtC and VTE images, similarly to [11,13,20,27]. On the other hand, EtC and VTE images are generated after resizing to 224×224 pixels to fit the resolution of a pretrained model.

3. Proposed Attacks

In this section, five conventional cipher-text-only attacks (COAs), a brute force attack (BF-attack), a feature reconstruction attack (FR-attack) [14], an inverse transformation network attack (ITN-attack) [15], a jigsaw puzzle solver attack (JPS-attack) [21], and an extended jigsaw puzzle solver attack (EJPS-attack) [23], are summarized, and the novel jigsaw puzzle solver-based attack is proposed.

3.1. Threat Models

The objective of an attacker in an image classification scenario where privacy is a concern is to recover visual information from encrypted images. Encrypted images are transmitted to an untrusted cloud provider for model training and testing, as shown in Figure 1. Therefore, we assume that the attacker has access to encrypted images and the encryption algorithm but does not possess the secret key. That is to say, we assume that the attacker can only carry out a cipher-text-only attack (COA) using encrypted images.

Several COAs have been proposed to restore visual information from encrypted images, as described in Section 2.2. In this paper, we use five conventional COAs: a brute force attack (BF-attack), a feature reconstruction attack (FR-attack) [14], an inverse transformation network attack (ITN-attack) [15], a jigsaw puzzle solver attack (JPS-attack) [21], and an extended jigsaw puzzle solver attack (EJPS-attack) [23].

The robustness of an encrypted image against the brute force attack (BF-attack) has been evaluated on the basis of the size of the key space [28]. The key space of each encryption scheme is calculated on the basis of the resolution of the image $X \times Y$, the number of divided blocks n, and the block size M used for encryption, as shown in Table 1.

For example, the key space of LE depends on only the block size, while that of PE depends on the resolution of the image. The key space of the LE [11] is calculated as

$$\begin{aligned} O(LE) &= N_{ps}(M) \cdot N_{np}(M) \\ &= (M^2 \cdot 6)! \cdot 2^{M^2 \cdot 6} \\ &= (4^2 \cdot 6)! \cdot 2^{4^2 \cdot 6} \\ &= 96! \cdot 2^{96}, \end{aligned} \tag{4}$$

where $N_{ps}(M)$ and $N_{np}(M)$ are the key spaces of the pixel shuffling and intensities reversing. On the other hand, the key space of ELE is larger than that of LE because of the operation of block scrambling. Hence, the key space of ELE [13] is given by

$$\begin{aligned} O(ELE) &= N_{dps}(M, n) \cdot N_{bs}(n) \\ &= ((M^2 \cdot 6)! \cdot 2^{M^2 \cdot 6})^n \cdot n! \\ &= ((4^2 \cdot 6)! \cdot 2^{4^2 \cdot 6})^{64} \cdot 64! \\ &= (96! \cdot 2^{96})^{64} \cdot 64!, \end{aligned} \tag{5}$$

where $N_{dps}(M, n)$ is the key space of pixel shuffling and intensities reversing and $N_{bs}(n)$ is the key space of block scrambling. Note that, in this scheme, pixel shuffling and intensities reversing are carried out independently on each block. The key space of the PE [12] is represented by

$$\begin{aligned} O(PE) &= N_{npi}(X, Y) \cdot N_{psi}(X, Y) \\ &= 2^{3 \cdot X \cdot Y} \cdot 6^{X \cdot Y} \\ &= 2^{3 \cdot 32 \cdot 32} \cdot 6^{32 \cdot 32} \\ &= 2^{3072} \cdot 6^{1024}, \end{aligned} \tag{6}$$

where $N_{npi}(X, Y)$ and $N_{psi}(X, Y)$ are the key spaces of the negative–positive transformation and pixel shuffling in each pixel. The number of blocks in an EtC image is larger than ELE because of resizing. Accordingly, the key space of the EtC [28] is calculated as

$$\begin{aligned} O(EtC) &= N_{rib}(n) \cdot N_{npb}(n) \cdot N_{psb}(n) \cdot N_{bs}(n) \\ &= 8^n \cdot 2^n \cdot 6^n \cdot n! \\ &= 8^{196} \cdot 2^{196} \cdot 6^{196} \cdot 196!, \end{aligned} \tag{7}$$

where $N_{rib}(n)$ is the key space of block rotation and inversion, $N_{npb}(n)$ is the key space of the negative–positive transformation, and $N_{psb}(n)$ is the key space of color component shuffling. Similar to EtC images, the number of blocks in VTE images is expanded owing to resizing. Thus, the key space of VTE [20] is represented as

$$\begin{aligned} O(VTE) &= N_{bs}(n) \cdot N_{pss}(M) \\ &= n! \cdot (\frac{M}{2})^2! \\ &= 196! \cdot (\frac{16}{2})^2! \\ &= 196! \cdot 64!, \end{aligned} \tag{8}$$

where $N_{pss}(M)$ is the key space of pixel shuffling in each sub-block. The key spaces for the LE, PE, ELE, EtC, and VTE are represented as

$$O(ELE) \gg O(PE) \gg O(EtC) \gg O(VTE) \gg O(LE) \gg 2^{256}. \tag{9}$$

The key space of each encryption scheme is larger than a 256-bit key, so LE, PE, ELE, EtC, and VTE are robust against brute force attacks.

The feature reconstruction attack (FR-attack), which uses the edge information on images, was proposed to restore the visual information of encrypted images, as described in Algorithm 1 [14].

Algorithm 1 FR-attack [14]

Require: Encrypted input image I_e of size $X \times Y$;
number of bits L; leading bit $b \in \{0, 1\}$

for $q = (u, v) \in I_e$ **do**
  for $c \in R, G, B$ **do**
    if $\lfloor q_c/(2^L - 1) \rfloor \neq b$ **then**
      $q_c \leftarrow q_c \oplus (2^L - 1)$
    end if
  end for
end for

It is known that the FR-attack is an effective attack method against PE [12] images. This algorithm utilizes a neighborhood of surrounding pixels, so permuting pixels or blocks in an encrypted image enhance the robustness against the FR-attack.

The inverse transformation network attack (ITN-attack) is an attack where the adversary prepares exact pairs of plain and encrypted images using different keys. An ITN-attack is capable of restoring visual information from LE [11] images. A loss between a reconstructed image and the original one is utilized to train the transformation model. As described in the original paper [15], the architecture of the transformation model varies depending on the type of encryption. For example, the transformation model for PE consists of 1×1 locally connected layers each with a kernel size and a stride of (1, 1) .

The jigsaw puzzle solver attack (JPS-attack), which considers the blocks of an encrypted image as pieces of a jigsaw puzzle, was proposed to reconstruct visual information. The JPS-attack is carried out by calculating the correlation between permuted blocks. It has been shown that assembling jigsaw puzzles becomes difficult when the encrypted images have a large number of blocks and the block size is small [28]. Although it is known that the application of block scrambling to encrypted images enhances the robustness against various attacks, the use of a jigsaw puzzle solver attack enables visual information to be reconstructed [21]. Furthermore, the extended jigsaw puzzle solver attack (EJPS-attack) enables us to reconstruct EtC images, namely encrypted images including permuted, rotated, inverted, negative–positive transformed, and RGB shuffled blocks [23]. It has been confirmed that the use of an EJPS-attack successfully restores visual information from EtC [28] images. However, the jigsaw puzzle solver attack cannot restore images including pixel shuffling. Therefore, we propose a novel jigsaw puzzle solver-based attack that enables us to reconstruct an encrypted image including pixel shuffled blocks.

3.2. Jigsaw Puzzle Solver-Based Attack

In this paper, a novel jigsaw puzzle solver-based attack, which utilizes the correlation between blocks for reconstruction, is proposed. The proposed attack aims to restore visual information from images encrypted for being applied to ViTs [20]. In contrast, the proposed attack is feasible even for other block-based image encryption methods when a common key is applied to all divided blocks. The proposed attack consists of the two steps shown in Figure 4; the first step is sub-block restoration, which aims to solve pixel shuffling in an encrypted image, and the second is a jigsaw puzzle solver attack to assemble a scrambled image.

The purpose of sub-block restoration is to solve pixel shuffling in a sub-block by using the pixels at the edges of the sub-block to generate a non-shuffled pixel block, denoted by $\hat{S}ij$.

First, the positions of the upper left corner p_{ul}, upper right corner p_{ur}, lower left corner p_{ll}, and lower right corner p_{lr} in $\hat{S}_{ij}$ are determined as illustrated in Figure 5.

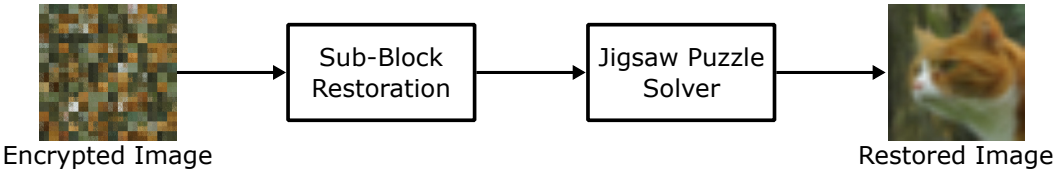


Figure 4. Proposed jigsaw puzzle solver-based attack

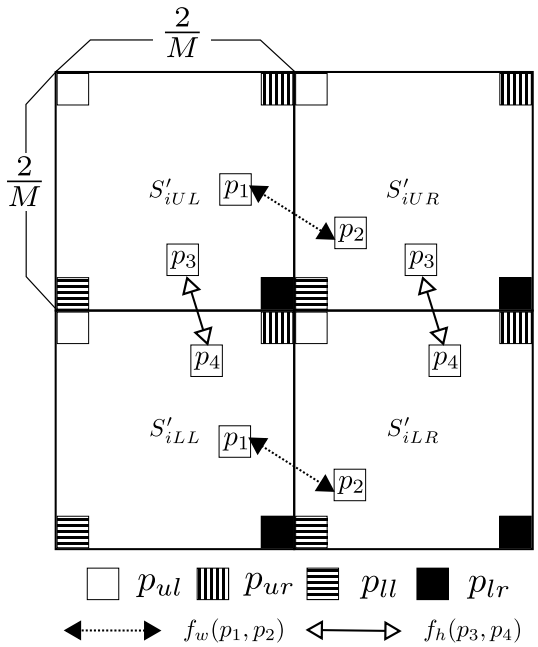


Figure 5. Positions of upper left corner p_{ul}, upper right corner p_{ur}, lower left corner p_{ll}, and lower right corner p_{lr} in sub-block $\hat{S}_{ij}$. The MSE values between p_1 and p_2 and between p_3 and p_4 are calculated, respectively.

Given the four different pixel positions $p_1 = (x_1, y_1)$, $p_2 = (x_2, y_2)$, $p_3 = (x_3, y_3)$, $p_4 = (x_4, y_4)$, $p_1 \neq p_2 \neq p_3 \neq p_4$, $x_1, y_1, x_2, y_2, x_3, y_3, x_4, y_4 \in \{1, 2, \ldots, \frac{M}{2}\}$ in S'_{ij}, the pixel intensities are defined as $S'_{ij}(p_1, c)$, $S'_{ij}(p_2, c)$, $S'_{ij}(p_3, c)$, and $S'_{ij}(p_4, c)$, $c \in \{R, G, B\}$. Thus, the sum of all mean squared error (MSE) values between left and right sub-blocks is calculated by

$$f_w(p_1, p_2) = \sum_c \sum_{i=1}^{n} (S'_{iUL}(p_1, c) - S'_{iUR}(p_2, c))^2 + (S'_{iLL}(p_1, c) - S'_{iLR}(p_2, c))^2. \tag{10}$$

On the other hand, the sum of all MSE values between upper and lower sub-blocks is calculated as

$$f_h(p_3, p_4) = \sum_c \sum_{i=1}^{n} (S'_{iUL}(p_3, c) - S'_{iLL}(p_4, c))^2 + (S'_{iUR}(p_3, c) - S'_{iLR}(p_4, c))^2. \tag{11}$$

Using Equations (10) and (11), p_{ul}, p_{ur}, p_{ll}, and p_{lr} are given as

$$p_{lr}, p_{ll}, p_{ur}, p_{ul} = \underset{p_1, p_2, p_3, p_4}{\arg\min} \{f_w(p_1, p_2) + f_h(p_1, p_3) + f_w(p_3, p_4) + f_h(p_2, p_4)\}. \tag{12}$$

Next, the positions of the restored right edge $p_r = \{p_{r1}, \ldots, p_{rk}, \ldots, p_{r\frac{M}{2}-2}\}$ and left edge $p_l = \{p_{l1}, \ldots, p_{lk}, \ldots, p_{l\frac{M}{2}-2}\}$ are calculated as

$$p_{rk}, p_{lk} = \underset{p_1, p_2}{\arg\min}\, f_w(p_1, p_2). \tag{13}$$

Similar to p_{rk} and p_{lk}, the positions of the restored upper edge $p_u = \{p_{u1}, \ldots, p_{uk}, \ldots, p_{u\frac{M}{2}-2}\}$ and lower edge $p_{dy} = \{p_{d1}, \ldots, p_{dk}, \ldots, p_{d\frac{M}{2}-2}\}$ are defined as

$$p_{uk}, p_{dk} = \underset{p_3, p_4}{\arg\min}\, f_h(p_3, p_4). \tag{14}$$

The remaining positions in $\hat{S}_{ij}$ are determined by minimizing the MSE of the surrounding pixels to generate the restored image $\hat{I}$ as illustrated in Figure 6.

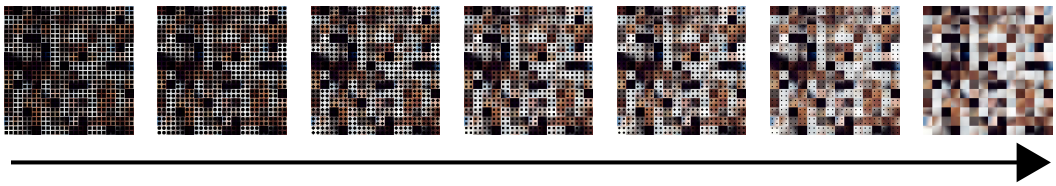


Figure 6. Procedure of sub-block restoration.

After solving the pixel shuffling encryption by using sub-block restoration, the jigsaw puzzle solver is applied to $\hat{I}$ to generate restored image $\hat{I}'$. It is known that an encrypted image that includes only shuffled blocks with 14×14 pixels can be easily restored [29]. Therefore, an image encrypted by using block-wise image encryption with $M = 16$ can be restored when sub-block restoration performs well.

4. Experiments

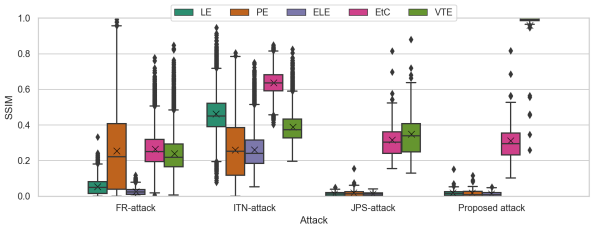
4.1. Experimental Conditions

In this section, the security of the image encryption for privacy-preserving DNNs was evaluated by using the FR-attack, the ITN-attack, the JPS-attack, the EJPS-attack, and the proposed attack. We used 10,000 testing images with 32×32 pixels from the CIFAR-10 dataset for the FR-attack and ITN-attack, and 100 of the testing images were used for the JPS-attack, EJPS-attack, and proposed attack. Before performing the EtC and VTE, each image was resized to 224×224 pixels.

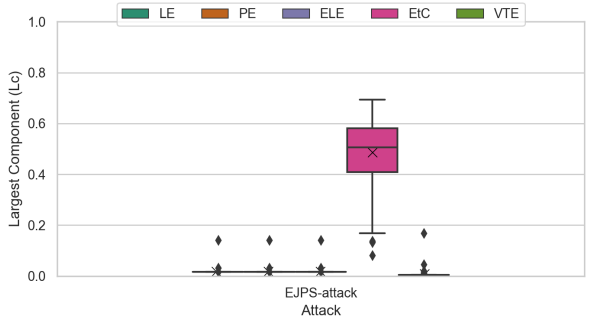
For the ITN-attack, the same architecture was used as in paper [30]. On the other hand, a jigsaw puzzle solver using genetic algorithms was utilized to restore encrypted images as the JPS-attack [21]. The average of the structural similarity index measure (SSIM) values between an original image and the restored one were calculated for the FR-attack, ITN-attack, JPS-attack, and proposed attack. The resolution of an image reconstructed by using the EJPS-attack is sometimes different from the original one because of the algorithm used for assembling encrypted blocks. Thus, the largest component $L_c \in [0, 1]$, which means the ratio of the correct pairwise adjacencies, was utilized to evaluate the robustness of the encrypted images against the EJPS-attack [31]. In this measure, a larger value means a higher compatibility, namely an adversary succeeds in reconstructing visual information from the encrypted images. For the evaluation of security against the EJPS-attack, ten different encrypted images were generated from one ordinary image by using different keys. We assembled the encrypted images by using the extended jigsaw puzzle solver and chose the image that had the highest L_c. We performed this procedure for each encrypted image independently and calculated the average L_c for the 100 images [23]. Furthermore, in order to demonstrate the effectiveness of the proposed attack, we measured the computation time using a PC with a 3.2 GHz processor and a main memory 128 Gbytes Processor:Intel Core i9-12900K 3.2GHz, OS:Ubuntu 20.04 LTS).

4.2. Experimental Results

Figure 7 shows the security of the image encryption for privacy-preserving DNNs evaluated by using the FR-attack, ITN-attack, JPS-attack, EJPS-attack, and proposed attack. As shown in Figure 7a, the LE images were almost reconstructed by using the ITN-attack, as the average values of SSIM were 0.46. Moreover, the PE images were restored by using the FR-attack. On the other hand, the EtC images were robust against the FR-attack, ITN-attack, JPS-attack, and proposed attack because of the operation of block scrambling. However, when using the EJPS-attack, half of visual information was reconstructed as the average value of L_c was 0.48, as illustrated in Figure 7b. Although the ELE images were robust against the FR-attack, ITN-attack, JPS-attack, EJPS-attack, and proposed attack, it was known that the classification performances are lower than others and an adaptation network is needed.



(**a**)



(**b**)

Figure 7. Average structural similarity index measure (SSIM) or largest component L_c values of images reconstructed by cipher-text-only attacks (COAs). 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 and cross inside boxes indicate median and average values, respectively. Dots represent outliers. (**a**) SSIM values of restored images under a feature reconstruction attack (FR-attack) [14], an inverse transformation network attack (ITN-attack) [15], a jigsaw puzzle solver attack (JPS-attack) [21] and the proposed attack. (**b**) Largest component values of extended jigsaw puzzle solver attack (EJPS-attack) [23].

Figure 8 shows examples of encrypted and reconstructed images, where Table 1 indicates the original and encrypted ones. As illustrated in Figure 8, the proposed attack succeeded in reconstructing the VTE images more clearly than the FR-attack, ITN-attack, JPS-attack, and EJPS-attack, as the average value of SSIM was 0.96.

Figure 9 illustrates the running time of the JPS-attack, EJPS-attack, and proposed attack, where the average time of 100 images from the CIFAR-10 dataset were plotted. Note that the running time of the FR-attack and ITN-attack is short compared to the other attacks. As shown in Figure 9, the running time of the proposed attack is almost the same as the other attacks. It was confirmed that the use of a jigsaw puzzle solver-based attack enables us to restore visual information, even though it is known that applying block scrambling and pixel shuffling to image encryption enhances the robustness.

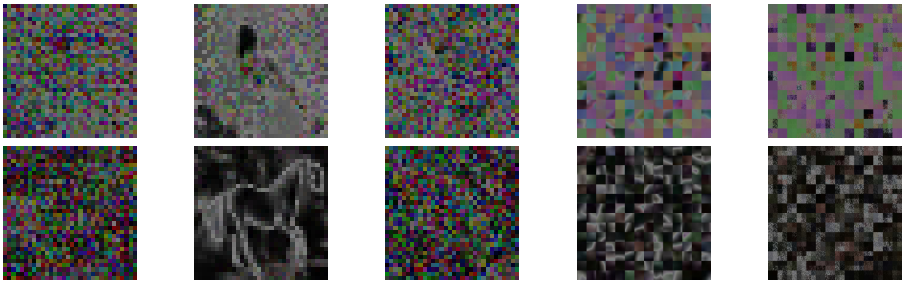
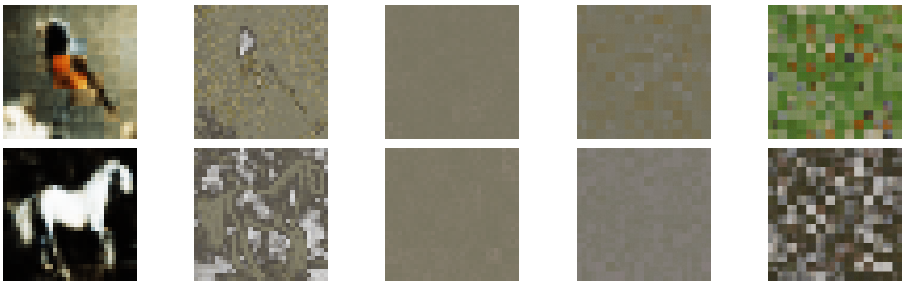
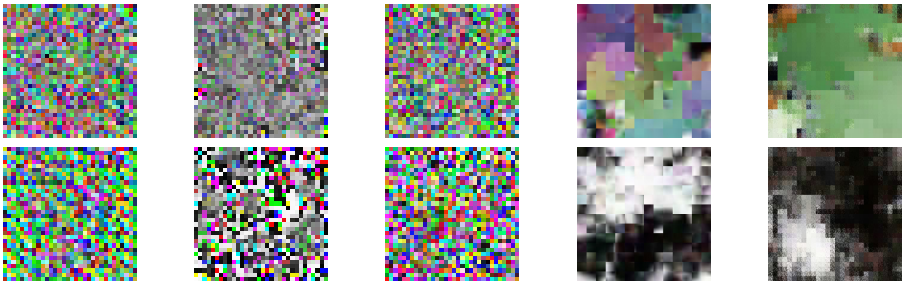
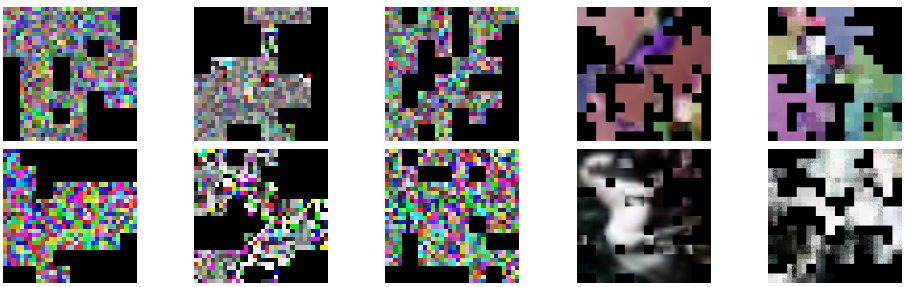
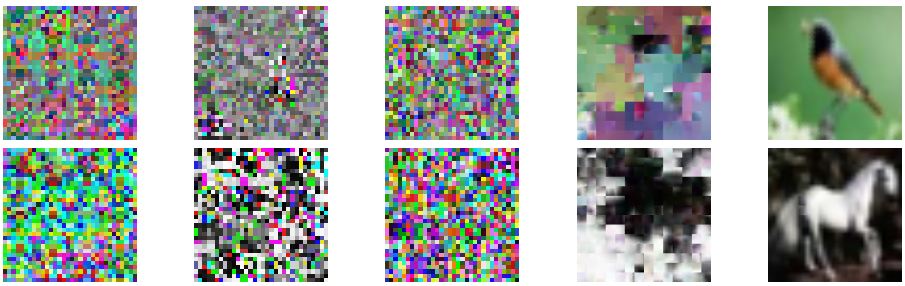
	LE	PE	ELE	EtC	VTE
Image size ($X \times Y$ pixels)	32×32	32×32	32×32	224×224	224×224
Block size ($M \times M$ pixels)	4×4	1×1	4×4	16×16	16×16
FR-attack					
ITN-attack					
JPS-attack					
EJPS-attack					
Proposed attack					

Figure 8. Examples of images reconstructed by using the feature reconstruction attack (FR-attack) [14], the inverse transformation network attack (ITN-attack) [15], the jigsaw puzzle solver attack (JPS-attack) [21], the extended jigsaw puzzle solver attack (EJPS-attack) [23], and the proposed attack. Table 1 shows the original and encrypted images.

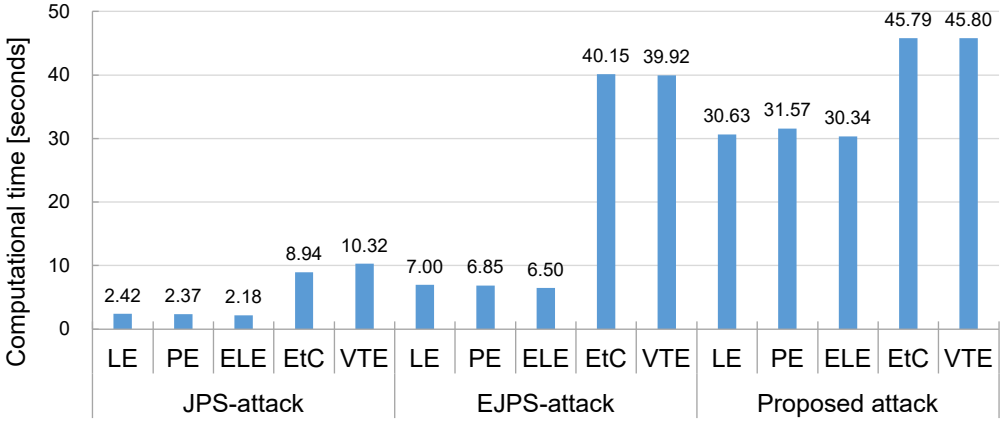


Figure 9. Running time of the jigsaw puzzle solver attack (JPS-attack) [21], the extended jigsaw puzzle solver attack (EJPS-attack) [23], and the proposed attack.

5. Conclusions

In this paper, we evaluated the security of image encryption for privacy-preserving DNNs by using the latest COAs. The BF-attack, FR-attack, ITN-attack, JPS-attack, EJPS-attack, and proposed attack were utilized as COAs. Furthermore, a novel jigsaw puzzle solver-based attack was proposed for encrypted images including scrambled blocks and shuffled pixels. In experiments, the robustness of LE, PE, ELE, EtC, and VTE images was evaluated by using the CIFAR-10 dataset. Although it is known that applying block scrambling and pixel shuffling to image encryption enhances the robustness against various attacks, the use of the proposed attack succeeded in reconstructing visual information.

Author Contributions: Conceptualization, T.C. and H.K.; methodology, T.C.; software, T.C.; validation, T.C.; formal analysis, T.C.; investigation, T.C.; resources, T.C.; data curation, T.C.; writing—original draft preparation, T.C.; writing—review and editing, H.K.; visualization, T.C.; supervision, H.K.; project administration, H.K. All authors have read and agreed to the published version of the manuscript.

Funding: This study was partially supported by JSPS KAKENHI (Grant Number JP21H01327) and the Support Center for Advanced Telecommunications Technology Research, Foundation (SCAT).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Yann, L.; Yoshua, B.; Geoffrey, H. Deep Learning. *Nature* **2015**, *521*, 436–444.
2. Huang, Q.X.; Yap, W.L.; Chiu, M.Y.; Sun, H.M. Privacy-Preserving Deep Learning With Learnable Image Encryption on Medical Images. *IEEE Access* **2022**, *10*, 66345–66355. [CrossRef]
3. Zeng, W.; Lei, S. Efficient frequency domain selective scrambling of digital video. *IEEE Trans. Multimed.* **2003**, *5*, 118–129. [CrossRef]
4. Ito, I.; Kiya, H. A new class of image registration for guaranteeing secure data management. In Proceedings of the IEEE International Conference on Image Processing, San Diego, CA, USA, 12–15 October 2008; pp. 269–272.
5. Kiya, H.; Ito, I. Image matching between scrambled images for secure data management. In Proceedings of the 16th European Signal Processing Conference, Lausanne, Switzerland, 25–29 August 2008; pp. 1–5.
6. Ito, I.; Kiya, H. One-time Key Based Phase Scrambling for Phase-only Correlation Between Visually Protected Images. *EURASIP J. Inf. Secur.* **2009**, *2009*, 841045. [CrossRef]
7. Tang, Z.; Zhang, X.; Lan, W. Efficient image encryption with block shuffling and chaotic map. *Multimed. Tools Appl.* **2015**, *74*, 5429–5448. [CrossRef]

8. Sirichotedumrong, W.; Chuman, T.; Imaizumi, S.; Kiya, H. Grayscale-Based Block Scrambling Image Encryption for Social Networking Services. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, USA, 23–27 July 2018; pp. 1–6.
9. Kiya, H.; Nagamori, T.; Imaizumi, S.; Shiota, S. Privacy-Preserving Semantic Segmentation Using Vision Transformer. *J. Imaging* **2022**, *8*, 233. [CrossRef] [PubMed]
10. Kiya, H.; Iijima, R.; AprilPyone, M.; Kinoshita, Y. Image and Model Transformation with Secret Key for Vision Transformer. *IEICE Trans. Inf. Syst.* **2023**, *106*, 2–11. [CrossRef]
11. Tanaka, M. Learnable Image Encryption. In Proceedings of the IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), Taipei, Taiwan, 6–8 July 2018; pp. 1–2.
12. Sirichotedumrong, W.; Kinoshita, Y.; Kiya, H. Pixel-Based Image Encryption Without Key Management for Privacy-Preserving Deep Neural Networks. *IEEE Access* **2019**, *7*, 177844–177855. [CrossRef]
13. Madono, K.; Tanaka, M.; Onishi, M.; Ogawa, T. Block-wise Scrambled Image Recognition Using Adaptation Network. In Proceedings of the Workshop on AAAI Conference Artificial Intellignece, New York, NY, USA, 7–12 February 2020.
14. Chang, A.H.; Case, B.M. Attacks on Image Encryption Schemes for Privacy-Preserving Deep Neural Networks. *arXiv* **2020**, arXiv:2004.13263. [CrossRef]
15. Ito, H.; Kinoshita, Y.; Aprilpyone, M.; Kiya, H. Image to Perturbation: An Image Transformation Network for Generating Visually Protected Images for Privacy-Preserving Deep Neural Networks. *IEEE Access* **2021**, *9*, 64629–64638. [CrossRef]
16. Xu, J.; Ai, B.; Chen, W.; Yang, A.; Sun, P. Image Encryption Methods in Deep Joint Source Channel Coding: A Review and Performance Evaluation. In Proceedings of the IEEE International Conference on Computer and Communications (ICCC), Xiamen, China, 28–30 July 2021; pp. 240–244.
17. Kiya, H.; AprilPyone, M.; Kinoshita, Y.; Imaizumi, S.; Shiota, S. An Overview of Compressible and Learnable Image Transformation with Secret Key and Its Applications. *APSIPA Trans. Signal Inf. Process.* **2022**, *11*, e11. [CrossRef]
18. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
19. AprilPyone, M.; Kiya, H. Block-wise Image Transformation with Secret Key for Adversarially Robust Defense. *IEEE Trans. Inf. Forensics Secur.* **2021**, *16*, 2709–2723. [CrossRef]
20. Qi, Z.; AprilPyone, M.; Kiya, H. Privacy-Preserving Image Classification Using ConvMixer with Adaptive Permutation Matrix. In Proceedings of the 30th European Signal Processing Conference(EUSIPCO), Belgrade, Serbia, 29 August–2 September 2022; pp. 543–547.
21. Sholomon, D.; David, O.E.; Netanyahu, N.S. An automatic solver for very large jigsaw puzzles using genetic algorithms. *Genet. Program. Evolvable Mach.* **2016**, *17*, 291–313. [CrossRef]
22. Son, K.; Hays, J.; Cooper, D.B. Solving Square Jigsaw Puzzle by Hierarchical Loop Constraints. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 2222–2235. [CrossRef] [PubMed]
23. Chuman, T.; Kurihara, K.; Kiya, H. Security Evaluation for Block Scrambling-based ETC Systems against Extended Jigsaw Puzzle Solver Attacks. In Proceedings of the IEEE International Conference on Multimedia and Expo (ICME), Hong Kong, China, 10–14 July 2017; pp. 229–234.
24. Chuman, T.; Kiya, H. A jigsaw puzzle solver-based attack on block-wise image encryption for privacy-preserving DNNs. In Proceedings of the International Workshop on Advanced Imaging Technology (IWAIT), Hong Kong, China, 4–6 January 2023; pp. 335–340.
25. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
26. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
27. AprilPyone, M.; Kiya, H. Privacy-Preserving Image Classification Using an Isotropic Network. *IEEE Trans. Multimed.* **2022**, *29*, 23–33. [CrossRef]
28. Chuman, T.; Sirichotedumrong, W.; Kiya, H. Encryption-then-compression systems using grayscale-based image encryption for jpeg images. *IEEE Trans. Inf. Forensics Secur.* **2019**, *14*, 1515–1525. [CrossRef]
29. Chuman, T.; Kurihara, K.; Kiya, H. On the Security of Block Scrambling-Based EtC Systems against Extended Jigsaw Puzzle Solver Attacks. *IEICE Trans. Inf. Syst.* **2018**, *101*, 37–44. [CrossRef]
30. Sirichotedumrong, W.; Kiya, H. Visual Security Evaluation of Learnable Image Encryption Methods against Ciphertext-only Attacks. In Proceedings of the 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Auckland, New Zealand, 7–10 December 2020; pp. 1304–1309.
31. Gallagher, A. Jigsaw Puzzles with Pieces of Unknown Orientation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Providence, RI, USA, 16–21 June 2012; pp. 382–389.