Human-Imperceptible, Machine-Recognizable Images

Fusheng Hao[1,2] , **Fengxiang He**[3†] , **Yikai Wang**[4] , **Fuxiang Wu**[1,2] , **Jing Zhang**[5]
Dacheng Tao[6] , **Jun Cheng**[1,2†]
[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2]The Chinese University of Hong Kong
[3]University of Edinburgh
[4]Beijing Normal University
[5]The University of Sydney
[6]Nanyang Technological University
F.He@ed.ac.uk, jun.cheng@siat.ac.cn

Abstract

Massive human-related data is collected to train neural networks for computer vision tasks. A major conflict is exposed relating to software engineers between better developing AI systems and distancing from the sensitive training data. To reconcile this conflict, the paper proposes an efficient privacy-preserving learning paradigm, where images are encrypted to become "human-imperceptible, machine-recognizable" via one of the two encryption strategies: (1) random shuffling equally-sized patches and (2) mixing-up sub-patches. Then, minimal adaptations are made to vision transformer to enable it to learn on the encrypted images for vision tasks, including image classification and object detection. Extensive experiments on ImageNet and COCO show that the proposed paradigm achieves comparable accuracy with the competitive methods. Decrypting the encrypted images requires solving an NP-hard jigsaw puzzle or ill-posed inverse problem, which is empirically shown intractable to be recovered by various attackers, including the powerful vision transformer-based attacker. We thus show that the proposed paradigm can ensure the encrypted images have become human-imperceptible while preserving machine-recognizable information.

1 Introduction

Relying on massive personal images, the industry has shown promising capabilities for developing artificial intelligence (AI) for many computer vision tasks, e.g., image classification [He *et al.*, 2016], action recognition [Moniruzzaman *et al.*, 2022], face recognition [Mi *et al.*, 2024], etc. In this process, a major conflict has been seen relating to software engineers between better developing AI systems and distancing from the sensitive training data. To reconcile this conflict, we raise a problem in this paper: *Can we process images to be human-imperceptible and machine-recognizable?* In such a way, software engineers can use the sensitive training data to facilitate their developing AI systems, without accessing the sensitive contents.

To process images to be "human-imperceptible, machine-recognizable", we develop two encryption strategies: random shuffling (RS) and mixing-up (MI); see Figure 1. RS randomly shuffles the patch order of an image, which destroys the position configurations between patches and can thus be applied to position-insensitive scenarios. Decrypting an image encrypted by RS is to solve a jigsaw puzzle problem, which can incur considerable computational overhead since the problem to be solved is a NP-hard one [Demaine and Demaine, 2007]. MI mixes up the sub-patches in a patch, which preserves the position configurations between patches and can thus be applied to position-sensitive scenarios. Decrypting an image encrypted by MI is to solve an ill-posed inverse problem, which could be hard to solve due to the difficulty in modeling the sub-patch distribution.

Then, we make minimal adaptations to vision transformer (ViT) [Dosovitskiy *et al.*, 2020] to enable it to learn on the encrypted images; see Figure 2. By removing position encoding, ViT is made permutation-invariant and thus capable of learning on images encrypted by RS, resulting in PEViT for the image classification task. Further, we develop a reference-based positional encoding for PEViT, which can retain the permutation-invariant property and thus boost the performance by a noticeable margin. Since position information plays a key role in the low-level object detection task, MI is chosen to encrypt images for this task. By adapting the way that image patches are embedded, YOLOS [Fang *et al.*, 2021], a vanilla ViT-based model, is able to learn on images encrypted by MI, resulting in PEYOLOS for the object detection task.

We conduct extensive experiments on ImageNet [Deng *et al.*, 2009] and COCO [Lin *et al.*, 2014]. Extensive attack experiments show the security of our encryption strategies. Comparison results on large-scale benchmarks show that both PEViT and PEYOLOS achieve promising performance with highly encrypted images as input. We thus

[†]Co-corresponding authors
Code: https://github.com/FushengHao/PrivacyPreservingML

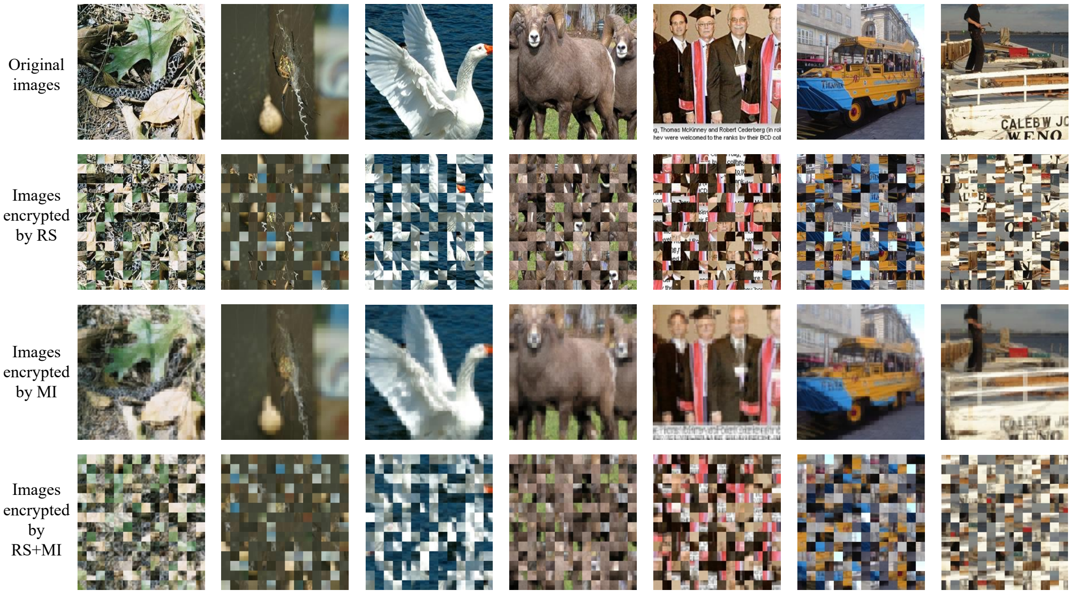

Figure 1: Illustration of images encrypted by RS, MI, and their combination. The visual contents of encrypted images are near-completely protected from recognizing by human eyes.

show that the proposed paradigm can ensure the encrypted images have become human-imperceptible while preserving machine-recognizable information. The main contributions can be summarized as follows:

- We propose an efficient privacy-preserving learning paradigm that can ensure the encrypted images have become human-imperceptible while preserving machine-recognizable information.
- RS is tailored for the standard image classification with ViT. By substituting the reference-based positional encoding for the original one, ViT is capable of learning on images encrypted by RS.
- By further designing MI, the privacy-preserving learning paradigm is extensible to position-sensitive tasks, such as object detection, for which we only need to adapt the way image patches are embedded.
- Extensive experiments demonstrate the effectiveness of the proposed privacy-preserving learning paradigm.

2 Related Work

Vision transformers. Self-attention based Transformer [Vaswani *et al.*, 2017] has achieved great success in natural language processing. To make Transformer suitable for image classification, the pioneering work of ViT [Dosovitskiy *et al.*, 2020] directly tokenizes and flattens 2D images into a sequence of tokens. Since then, researchers have been working on improving Vision Transformers and examples include DeiT [Touvron *et al.*, 2020], T2T-ViT [Yuan *et al.*, 2021], PVT [Wang *et al.*, 2021], ViTAEv2 [Zhang *et al.*, 2023], and Swin-Transformer [Liu *et al.*, 2021]. In addition, the intriguing properties of ViT are investigated in [Naseer *et al.*, 2021].

Jigsaw puzzle solver. The goal of a jigsaw puzzle solver is to reconstruct an original image from its shuffled patches. Since this problem is NP-hard [Demaine and Demaine, 2007], solving jigsaw puzzles of non-trivial size is impossible. Most of the existing works in computer vision focus on the jigsaw puzzle problem composed of equally-sized image patches and examples include the greedy algorithm proposed in [Cho *et al.*, 2010], the particle filter-based algorithm proposed in [Yang *et al.*, 2011], the fully-automatic solver proposed in [Pomeranz *et al.*, 2011], and the genetic-based solver proposed in [Sholomon *et al.*, 2013].

Privacy-preserving machine learning. The aim of privacy-preserving machine learning is to integrate privacy-preserving techniques into the machine learning pipeline. According to the phases of privacy integration, existing methods can be basically divided into four categories: data preparation, model training and evaluation, model deployment, and model inference [Xu *et al.*, 2021]. Federated learning allows multiple participants to jointly train a machine learning model while preserving their private data from being exposed [Liu *et al.*, 2022]. However, the leakage of gradients [Hatamizadeh *et al.*, 2022] or confidence information [Fredrikson *et al.*, 2015] can be utilized to recover original images such as human faces. From the perspective of data, encrypting data and then learning and inferencing on encrypted data can provide a strong privacy guarantee, called confidential-level privacy, which receives increasing attention recently. Manually anonymizing large-scale datasets faces the challenge of inefficiency and the need to develop specialized techniques [Ma *et al.*, 2023], and thus the automatic encryption approaches such as homomorphic encryption and functional encryption have been employed to encrypt data due to their nature of allowing computation over encrypted data [Xu *et al.*, 2019; Karthik *et al.*, 2019]. However, scaling them to deep net-

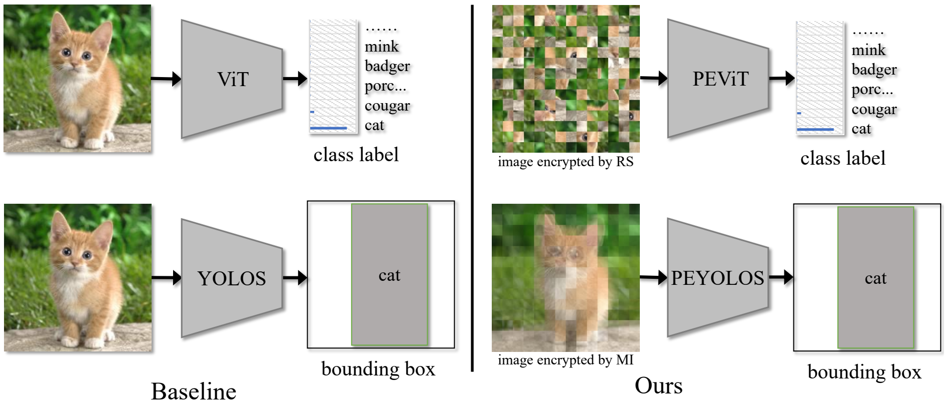

Figure 2: Images encrypted by RS and MI are still machine-learnable, by further designing architectures PEViT and PEYOLOS, based on ViT and YOLOS.

works and large datasets still faces extreme difficulties due to the high computational complexity involved for encryption. The block-wise pixel shuffling encryption strategy [Tanaka, 2018; Madono *et al.*, 2020] and the pixel-based image encryption strategy [Sirichotedumrong *et al.*, 2019] have low computational complexity for encryption, but they are vulnerable to reconstruction attacks because of the limited dimension of the key space [Sirichotedumrong *et al.*, 2019] or suffer from large performance degradation [Madono *et al.*, 2020]. In contrast, our encryption strategies are simple, efficient and easy to implement, and can be applied to position-sensitive tasks. Moreover, hardware-based deidentification methods has also been explored to improve vulnerability [Lopez *et al.*, 2024].

Differential privacy. Differential privacy is defined in terms of the application-specific concept of adjacent datasets [Abadi *et al.*, 2016], which bounds the disclosure risk of any individual participating in a dataset to guarantee data privacy. Then, researchers have been working on expanding its application scope. For example, Ferdinando et al. propose a differential-privacy mechanism for releasing hierarchical counts of individuals [Fioretto *et al.*, 2021]. Liu et al. study the interpretation robustness problem from the perspective of R*é*nyi differential privacy. We aim to process images to be "human-imperceptible, machine-recognizable", which is different from differential privacy that limits the information that attackers can learn about datasets. Similar idea is also mentioned in [Huang *et al.*, 2020] but the scheme has proved to be not private in [Carlini *et al.*, 2021].

Positions in transformers. Recent studies in the natural language processing field suggest that higher-order co-occurrence statistics of words play a major role in masked language models like BERT [Sinha *et al.*, 2021]. It has been shown that the word order contains surprisingly little information compared to that contained in the bag of words, since the understanding of syntax and the compressed world knowledge held by large models (e.g. BERT and GPT-2) are capable to infer the word order [Malkin *et al.*, 2021]. Due to the property of attention operation, when removing positional encoding, ViT is permutation-invariant w.r.t. its attentive tokens. As evaluated by our experiments, removing the positional embedding from ViT only leads to a moderate performance drop (3.1%, please see Table 1). Such a phenomenon inspires us to explore the permutation-based encryption strategy. Moreover, masked jigsaw puzzle [Ren *et al.*, 2023] has been explored to find a balance among accuracy, privacy, and consistency.

3 Method

In this section, we propose an efficient privacy-preserving learning paradigm that can ensure the encrypted images have become human-imperceptible while preserving machine-recognizable information. We first provide the encryption strategies and then detail the building blocks of ViT. Next, we describe how to learn on the encrypted images with minimal modifications to ViT. Finally, we discuss the encryption strategies in the context of cryptography.

3.1 Human-Imperceptible Images

We first consider typical vision tasks which are not quite position-sensitive, such as image classification that predicts the global category. Here, *Random Shuffling (RS)* images to a set of equally-sized image patches can destroy human-recognizable contents. This shuffle-encrypted process is simple, easy to implement, and is decoupled from the network optimization. Under the circumstance, decrypting an image is solving a jigsaw puzzle problem, which can incur a large computational overhead since the problem to be solved is an NP-hard one [Demaine and Demaine, 2007]. In particular, the dimension of the key space when applying the shuffle-encrypted strategy is the number of puzzle permutations. For an image with N patches, the dimension of the key space is given by,

$$K_S = N! \,. \tag{1}$$

For example, a 7×7 puzzle has $49! \approx 6 \times 10^{62}$ possible permutations. Based on this, it is easy to further increase the complexity of decrypting an image, by reducing the patch size of puzzles or increasing the resolution of the image. Besides, the complexity of decrypting an image can be further increased by dropping some image patches, as experimentally evaluated in Figure 6 and Figure 5.

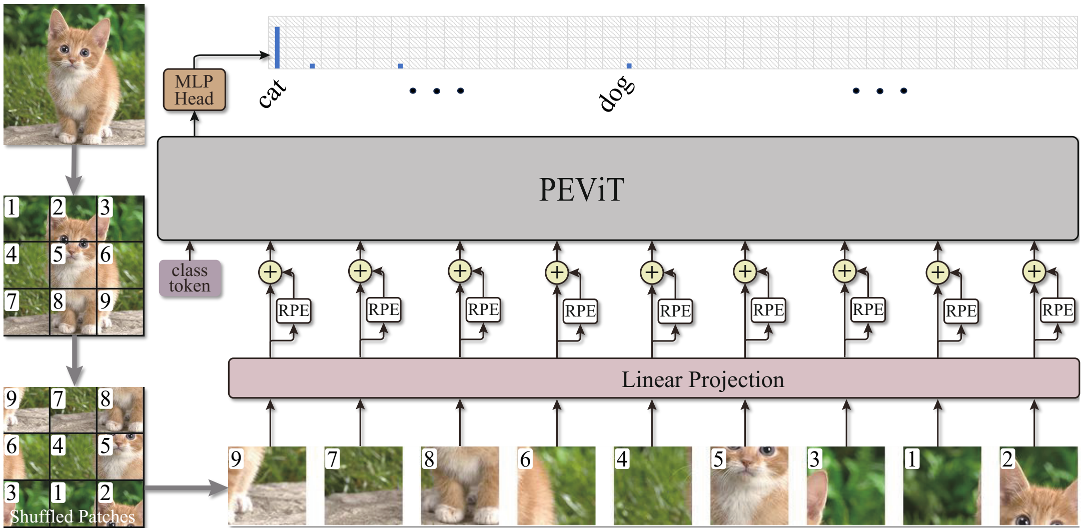

Figure 3: Architecture overview of PEViT with RPE.

However, always shuffling image patches is prone to underestimating the position information. To further make our paradigm applicable to position-sensitive tasks like object detection, where precise positions of bounding boxes are predicted, we design another encryption strategy named *Mixing (MI)*. Specifically, MI mixes sub-patches of image patches to destroy human-recognizable contents while preserving the position information, so that the encrypted data can be learned by networks that need position information. Let $\mathbf{x}^p$ denote an image patch. The mixing-encrypted strategy can be formulated as follows,

$$\mathbf{x}^p_S = \frac{1}{M}\sum_{i=1}^{M} \mathbf{s}^p_i \ , \tag{2}$$

where $\mathbf{s}^p_i$ denotes the i-th sub-patch of $\mathbf{x}^p$, M denotes the number of sub-patches, and $\mathbf{x}^p_S$ is the encrypted version of $\mathbf{x}^p$. It is to be noted that $\mathbf{x}^p_S$ has the same size as $\mathbf{s}^p_i$, which is smaller than that of $\mathbf{x}^p$. Since the sum function is permutation-invariant, MI makes an encrypted image permutation-invariant to its sub-patches. With the MI encryption process, decrypting a patch is solving the following ill-posed inverse problem,

$$\underset{\mathbf{s}^p_1,\cdots,\mathbf{s}^p_M}{\arg\min} \parallel \mathbf{x}^p_S - \sum_{i=1}^{M} \mathbf{s}^p_i \parallel^2 \ . \tag{3}$$

Both modeling the sub-patch distribution and restoring the sub-patch order make decrypting an patch a great challenge. Decrypting an image of N patches magnifies this challenge by a factor of N.

3.2 Building Blocks of ViT

In this part, we analyze how the change of input permutation affects each component of ViT.

Self-attention. The attention mechanism is a function that outputs the weighted sum of a set of k *value* vectors (packed into $V \in \mathbb{R}^{k\times d}$). The k weights are obtained by calculating the similarity between a *query* vector $q \in \mathbb{R}^d$ and a set of k *key* vectors (packed into $K \in \mathbb{R}^{k\times d}$) using inner products, which are then scaled and normalized with a softmax function. For a sequence of N query vectors (packed into $Q \in \mathbb{R}^{N\times d}$), the output matrix O (of size $N \times d$) can be computed by,

$$O = \mathrm{Attention}(Q, K, V) = \mathrm{Softmax}(QK^\top/\sqrt{d})V, \tag{4}$$

where the $\mathrm{Softmax}$ function is applied to the input matrix by rows.

In self-attention, *query*, *key*, and *value* matrices are computed from the same sequence of N input vectors (packed into $X \in \mathbb{R}^{N\times d}$) using linear transformations: $Q = XW^{\mathrm{Q}}$, $K = XW^{\mathrm{K}}$, $V = XW^{\mathrm{V}}$. Since the order of Q, K, and V is co-variant with that of X, the permutation of the input of self-attention permutes the output.

Multi-head self-attention. Multi-head Self-Attention (MSA) [Vaswani *et al.*, 2017] consists of h self-attention layers, each of which outputs a matrix of size $N \times d$. These h output matrices are then rearranged into a $N \times dh$ matrix that is reprojected by a linear layer into $N \times d$,

$$\mathrm{MSA} = \mathrm{Concat}(\mathrm{head}_1, \cdots, \mathrm{head}_h)W^{\mathrm{O}}, \tag{5}$$

where $\mathrm{head}_i = \mathrm{Attention}(Q_i, K_i, V_i)$, $Q_i = XW^{\mathrm{Q}}_i$, $K_i = XW^{\mathrm{K}}_i$, $V_i = XW^{\mathrm{V}}_i$, and $W^{\mathrm{O}} \in \mathbb{R}^{dh\times d}$. The order of each head is co-variant with that of X. Considering that the Concat operation only concatenates the vectors from different heads at the same position, MSA is co-variant to the order of X. Therefore, we conclude that the permutation of the input of MSA permutes the output.

Layer normalization. The mean and standard deviation are calculated over the features of all positions in Layer normalization [Ba *et al.*, 2016]. Then, they are used to transform features in a position-wise way. Therefore, we conclude that the permutation of the input of layer normalization permutes the output.

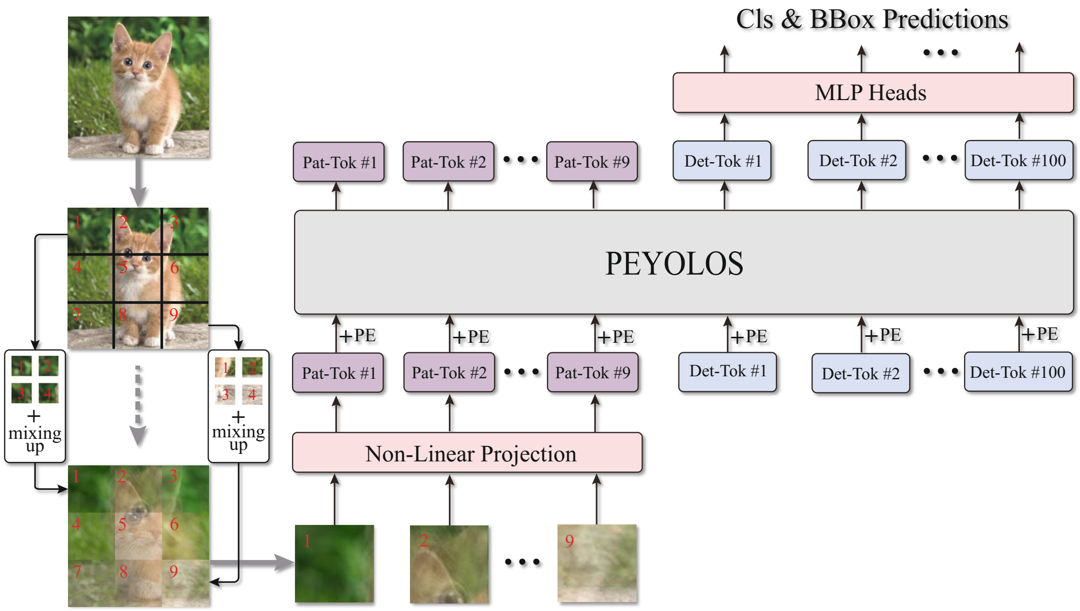

Figure 4: Architecture overview of PEYOLOS.

Residual connection. The residual connection [He *et al.*, 2016] can be formulated as $\mathcal{F}(X) + X$. If the order of the nonlinear mapping $\mathcal{F}(\cdot)$ is co-variant with that of X, the residual connection is co-variant with that of X. Therefore, we conclude that the permutation of the input of residual connection permutes the output.

Feed-forward network. Feed-Forward Network (FFN) consists of two linear layers separated by a GELU activation [Hendrycks and Gimpel, 2016]. The first linear layer expands the dimension from d to $4d$, while the second layer reduces the dimension from $4d$ back to d. Considering that FFN is applied in a position-wise way, we conclude that the permutation of the input of FFN permutes the output.

Positional encoding. To retain the positional information, position embeddings are usually added to the patch embeddings. Since there is little to no difference between different ways of encoding positional information, learnable 1D position embeddings are used in ViT [Dosovitskiy *et al.*, 2020]. It is to be noted that the positional encoding is unaware of input permutation.

3.3 Classification on Encrypted Images

As a standard method to handle images in ViT, the fixed-size input image of $H \times W \times C$ is decomposed into a batch of N patches of a fixed resolution of $P \times P$, resulting in a sequence of flattened 2D patches $X^p \in \mathbb{R}^{N \times (P^2 \cdot C)}$. For example, the sequence length $N = HW/P^2$ of ViT could be 196 for image classification on the ImageNet dataset. To destroy the human-recognizable contents, we choose RS as the encryption strategy to encrypt images. The reason is two-fold: (1) The key space of an image encrypted by RS is big enough and (2) The drop in performance is insignificant. To learn on the images encrypted by RS, we design permutation-invariant ViT (PEViT), defined as follows,

$$\{\mathbf{x}_1^p, \mathbf{x}_2^p, \cdots, \mathbf{x}_N^p\} \xrightarrow{\text{Shuffling}} \{\mathbf{x}_2^p, \mathbf{x}_N^p, \cdots, \mathbf{x}_1^p\} \tag{6}$$

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}};\, \mathbf{x}_2^p\mathbf{E};\, \mathbf{x}_N^p\mathbf{E}; \cdots ;\, \mathbf{x}_1^p\mathbf{E}],\ \ \mathbf{E} \in \mathbb{R}^{(P^2 \cdot C) \times D} \tag{7}$$

$$\mathbf{z'}_\ell = \mathrm{MSA}(\mathrm{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1},\ \ \ell = 1 \ldots L \tag{8}$$

$$\mathbf{z}_\ell = \mathrm{FFN}(\mathrm{LN}(\mathbf{z'}_\ell)) + \mathbf{z'}_\ell,\ \ \ell = 1 \ldots L \tag{9}$$

$$\mathbf{y} = \mathrm{LN}(\mathbf{z}_L^0). \tag{10}$$

where $\mathbf{E}$ denotes the linear projection that maps each vectorized image patch to the model dimension D, and $\mathbf{x}_{\text{class}}$ denotes the class token ($\mathbf{z}_0^0 = \mathbf{x}_{\text{class}}$), whose state at the output of the visual transformer ($\mathbf{z}_L^0$) serves as the image representation $\mathbf{y}$.

The differences between PEViT and vanilla ViT are two-fold: (1) Our model takes shuffled patch embeddings as input and (2) The learned positional encodings are removed. Since the permutation of the input of all the building blocks permutes the output, the order of $\mathbf{z}_L$ is co-variant with that of $\mathbf{z}_0$. It is worth noting that the class token is fixed in $\mathbf{z}_0^0$. Therefore, $\mathbf{z}_L^0$ corresponds to the image representation $\mathbf{y}$ that is invariant to the order of patch embeddings or image patches.

When directly introducing positional encoding, the permutation-invariant property of PEViT is destroyed. Inspired by relative encoding [Shaw *et al.*, 2018], we propose a reference-based positional embedding approach that can retain the permutation-invariant property,

$$\mathbf{E}_{pos}(\mathbf{x}_i^p) = \mathrm{RPE}(\mathbf{x}_i^p - \mathbf{x}^{\text{ref}}), \tag{11}$$

where $\mathbf{x}^{\text{ref}} \in \mathbb{R}^d$ denotes the learnable reference embedding and RPE denotes the reference-based positional encoding network that consists of two linear layers separated by a GELU activation [Hendrycks and Gimpel, 2016], followed

Method	Image size	#Param.	FLOPs	Throughput (img/s)	Top-1 acc.
DeiT-B [Touvron *et al.*, 2020]	224^2	86M	17.5G	292.3	81.8
ViT-B/16 [Dosovitskiy *et al.*, 2020]	384^2	86M	55.4G	85.9	77.9
ViT-L/16 [Dosovitskiy *et al.*, 2020]	384^2	307M	190.7G	27.3	76.5
DeiT-B on images encrypted by MI	224^2	86M	17.5G	291.5	78.0
DeiT-B on images encrypted by RS	224^2	86M	17.5G	291.6	4.5
DeiT-B on images encrypted by RS + MI	224^2	86M	17.6G	291.1	1.7
PEViT-B on images encrypted by MI	224^2	86M	17.5G	291.5	77.9
PEViT-B on images encrypted by RS	224^2	86M	17.5G	291.5	78.7
PEViT-B on images encrypted by RS + MI	224^2	86M	17.5G	291.1	69.5
PEViT-B with RPE on images encrypted by RS	224^2	87M	17.6G	290.8	79.7

Table 1: Image classification on ImageNet-1K. The throughput is measured as the number of images processed per second on a V100 GPU.

Method	Backbone	Image size	AP	#Params.	FLOPs (G)	FPS
YOLOS-Ti [Fang *et al.*, 2021]	DeiT-Ti	$512 \times *$	28.7	6.5M	18.8	60
PEYOLOS	DeiT-Ti	$512 \times *$	25.3	7.1M	19.0	58
DETR [Carion *et al.*, 2020]	ResNet-18-DC5	$800 \times *$	36.9	29M	129	7.4
YOLOS-S [Fang *et al.*, 2021]	DeiT-S	$800 \times *$	36.1	31M	194	5.7
PEYOLOS	DeiT-S	$800 \times *$	32.9	31.6M	194.9	5.6
DETR [Carion *et al.*, 2020]	ResNet-101-DC5	$800 \times *$	42.5	60M	253	5.3
YOLOS-B [Fang *et al.*, 2021]	DeiT-B	$800 \times *$	42.0	127M	538	2.7
PEYOLOS	DeiT-B	$800 \times *$	39.5	128.2M	539.7	2.5

Table 2: Object detection on the COCO test2017 dataset. FPS is measured with batch size 1 on a single 1080Ti GPU.

by a sigmoid function. This reference-based positional embedding relies only on the learnable reference embedding and patch embeddings, and thus its order is co-variant with that of input vectors. Accordingly, the input to PEViT with RPE can be defined as follows,

$$\mathbf{z}_0 = [\mathbf{x}_{\text{class}};\ \mathbf{x}_2^p\mathbf{E} + \mathbf{E}_{pos}(\mathbf{x}_2^p);\ \mathbf{x}_N^p\mathbf{E}+\mathbf{E}_{pos}(\mathbf{x}_N^p); \cdots ;\ \mathbf{x}_1^p\mathbf{E} + \mathbf{E}_{pos}(\mathbf{x}_1^p)], \tag{12}$$

where the permutation-invariant property of PEViT is retained. An architecture overview of PEViT with RPE is depicted in Figure 3.

Through the reference-based positional encoding, we illustrate that introducing positional embedding while retaining permutation-invariant property is feasible. Other better positional embedding approaches can also be designed, but it is beyond the scope of this paper.

3.4 Object Detection on Encrypted Images

YOLOS [Fang *et al.*, 2021] is an object detection model based on the vanilla ViT. The change from a ViT to a YOLOS involves two steps: (1) Dropping the class token and appending 100 randomly initialized learnable detection tokens to the input patch embeddings and (2) Replacing the image classification loss with the bipartite matching loss to perform object detection in a set prediction manner [Carion *et al.*, 2020].

Since position information plays a key role in the low-level object detection task, directly encrypting images with RS disrupts the patch positions and thus leads to significant performance degradation. This brings a great challenge to destroy human-recognizable contents while preserving machine-learnable information for the object detection task. We address this challenge by adapting the way image patches are embedded. For an image patch ($\mathbf{x}_i^p$) of a fixed resolution $P \times P$, we further decompose the patch into a batch of 4 sub-patches of a fixed resolution $\frac{P}{2} \times \frac{P}{2}$. Then, these sub-patches are encrypted with MI, resulting in the encrypted patch $\mathbf{x}_i^S$. Accordingly, the input to YOLOS is adapted as follows,

$$\mathbf{z}_0 = [\mathbf{x}_1^{DET}; \cdots ; \mathbf{x}_{100}^{DET};\ \mathcal{H}(\mathbf{x}_1^S); \mathcal{H}(\mathbf{x}_2^S); \cdots ; \mathcal{H}(\mathbf{x}_N^S)] + \mathbf{P}, \tag{13}$$

where $\mathcal{H}$ denotes a nonlinear mapping composed of two linear layers separated by a GELU activation [Hendrycks and Gimpel, 2016] and $\mathbf{P} \in \mathbb{R}^{(100+N)\times D}$ denotes the learnable positional embeddings. This adaptation makes YOLOS partially permutation-invariant to its sub-patches. With our PEYOLOS, we can destroy human-recognizable contents while preserving machine-learnable information for the object detection task.

The pipeline of PEYOLOS is shown in Figure 4. The input image of $H \times W \times C$ is decomposed into a batch of N patches with a fixed resolution of $P \times P$. Then, for an image patch, we further decompose the patch into a batch of 4 sub-patches with a fixed resolution $\frac{P}{2} \times \frac{P}{2}$. Finally, these sub-patches are encrypted with MI, resulting in a highly encryted image that is not human-recognizable and is tough to be decrypted. PEYOLOS takes as input the encrypted images and outputs class and bounding box predictions for the original images.

3.5 Discussion

We would like to recall a framework termed substitution-permutation network (SPN) in cryptography [Stinson and Paterson, 2019]. SPN is a series of linked mathematical op-

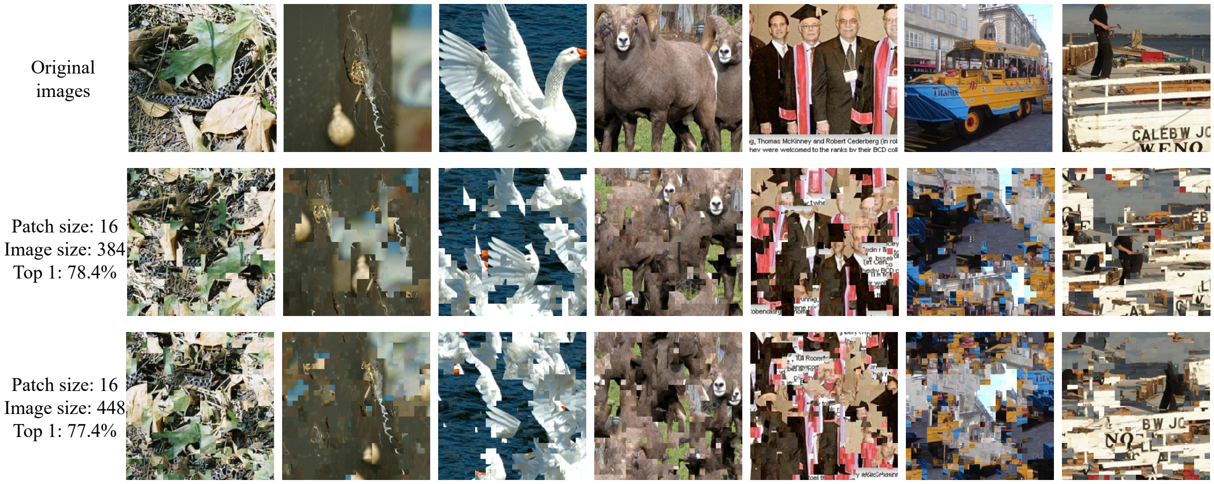

Figure 5: Reconstructed images by the jigsaw puzzle solver proposed in [Paikin and Tal, 2015]. Here, the effect of image size on image reconstruction quality and classification performance is investigated.

erations used in block cipher algorithms such as AES and DES [Stinson and Paterson, 2019]. Shannon suggests that practical and secure ciphers may be constructed by employing a mixing transformation consisting of several rounds of confusion and diffusion [C.E., 1949]; SPN is exactly an implementation of this confusion and diffusion paradigm. Such an implementation applies several alternating rounds of substitution boxes (S-boxes) and permutation boxes (P-boxes) to produce the ciphertext. An S-box substitutes a small block of bits (the input of the S-box) by another block of bits (the output of the S-box). A P-box is a permutation of bit blocks (or bits). Although a single typical S-box or a single P-box alone does not have sufficient cryptographic strength, a well-designed SPN with several alternating rounds of S-boxes and P-boxes already has a very strong proven security [Yevgeniy *et al.*, 2017].

Our encryption scheme adheres to SPN, while the basic unit of our encryption scheme is pixels instead of bits. In particular, MI can be seen as an implementation of S-box. RS can be seen as an implementation of P-box. Although both MI and RS alone does not have sufficient cryptographic strength, alternating several rounds of MI and RS can enhance the cryptographic strength to a large extent; see Figure 1. We would like to stress out that the performance on encrypted images is also a major concern of our work. Although alternating several rounds of MI and RS can enhance cryptographic strength, it also decreases the performance by a significant margin. In practice, there is a tradeoff between performance and cryptographic strength.

4 Experiments

In this section, we first provide the experimental settings, and then contrast the performance on image classification and object detection tasks. Finally, we investigate the attackers. More results and ablation studies are provided in the appendix.

4.1 Experimental Settings

Datasets. For the image classification task, we benchmark the proposed PEViT on ImageNet-1K [Deng *et al.*, 2009], which contains ∼1.28M training images and 50K validation images. For the object detection task, we benchmark the proposed PEYOLOS on COCO [Lin *et al.*, 2014], which contains 118K training, 5K validation and 20K test images.

Implementation details. The pseudocodes of RS and MI in a PyTorch-like style are shown in the appendix. We implement the proposed PEViT based on the Timm library [Wightman, 2019]. We adopt the default hyper-parameters of the DeiT training scheme [Touvron *et al.*, 2020] except setting the batch size to 192 per GPU, where 8 NVIDIA A100 GPUs are used for training. It is worth noting PEViT (w.o RPE) is equivalent to removing positional embeddings from DeiT. We implement the proposed PEYOLOS based on the publicly released code in [Fang *et al.*, 2021] and adapt the way image patches are embedded.

Baseline. We propose an efficient privacy-preserving learning paradigm with the aim of destroying human-recognizable contents while preserving machine-learnable information. The proposed PEViT and PEYOLOS are inherited from DeiT [Touvron *et al.*, 2020] and YOLOS [Fang *et al.*, 2021] respectively, which are thus selected as baselines. It is worth noting that both PEViT and PEYOLOS are not designed to be high-performance models that beats state-of-the-art image classification and object detection models, but to unveil that destroying human-recognizable contents while preserving machine-learnable information is feasible.

Measurement of privacy protection. As shown in Figure 1, the visual contents of encrypted images are nearly-completely protected from recognizing by human eyes. To measure the strength of privacy protection, we try to restore the original images with various attack algorithms, including puzzle solver attacker and gradient leakage attacker. Then, the quality of restored images can reflect the strength of privacy protection.

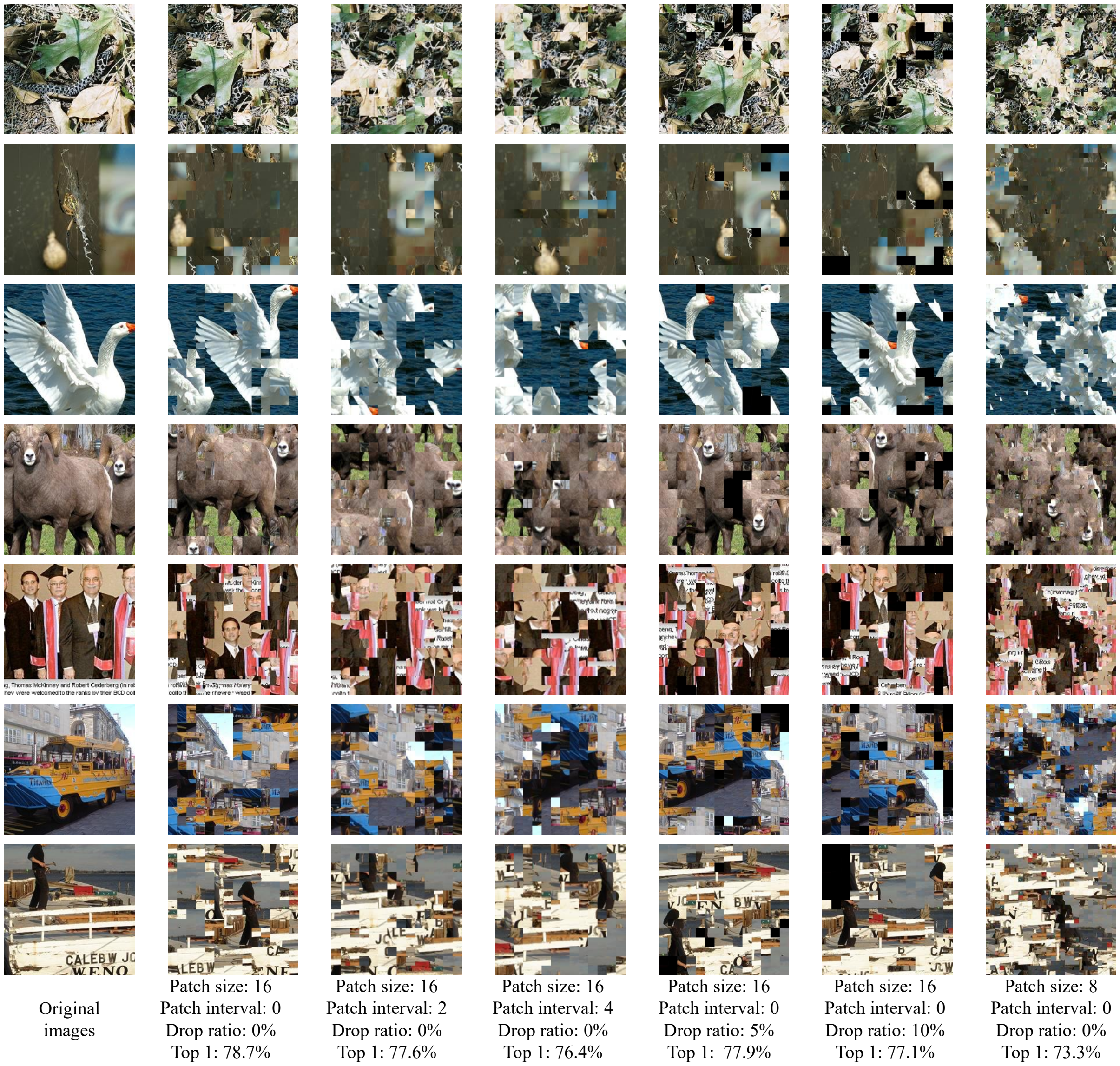
Original images
Patch size: 16 Patch interval: 0 Drop ratio: 0% Top 1: 78.7%
Patch size: 16 Patch interval: 2 Drop ratio: 0% Top 1: 77.6%
Patch size: 16 Patch interval: 4 Drop ratio: 0% Top 1: 76.4%
Patch size: 16 Patch interval: 0 Drop ratio: 5% Top 1: 77.9%
Patch size: 16 Patch interval: 0 Drop ratio: 10% Top 1: 77.1%
Patch size: 8 Patch interval: 0 Drop ratio: 0% Top 1: 73.3%

Figure 6: Reconstructed images by the jigsaw puzzle solver proposed in [Paikin and Tal, 2015], where the default image size is 224 × 224. Here, the effect of patch size, patch interval, and patch drop ratio on image reconstruction quality and classification performance is investigated.

4.2 Comparison Results

ImageNet-1K classification. The summary of the image classification results is shown in Table 1. Our PEViT-B works well on images encrypted by MI, RS and RS + MI, while DeiT-B does not. It is worth noting that although the performance of PEViT-B is not state-of-the-art, it is applied on encrypted images where the visual contents of images cannot be recognized by human eyes, while the comparison methods cannot. Moreover, RS and MI can be combined to further improve data security but at the expense of performance. Therefore, we conclude that PEViT-B achieves a trade-off between performance and visual content protection. Moreover, it is worth noting that the extra computation cost incurred by our encryption strategies is negligible.

COCO object detection. The summary of the object detection results is shown in Table 2. PEYOLOS maintains the same efficiency as YOLOS and the performance drop is only ∼3.0. Although PEYOLOS does not outperform YOLOS, it is currently the unique model to achieve a trade-off between performance and visual content protection for object detection. It is worth noting that MI is not limited to YOLOS. With minimal adaptations, other object detection frameworks based on plain ViT can also be adapted to work on images encrypted by MI. Recently, it has been shown that, with ViT backbones pre-trained as Masked Autoencoders, ViTDet [Li *et al.*, 2022] can compete with the previous leading methods that were all based on hierarchical backbones. By adapting the way the image patches are mapped like PEYOLOS, PE-

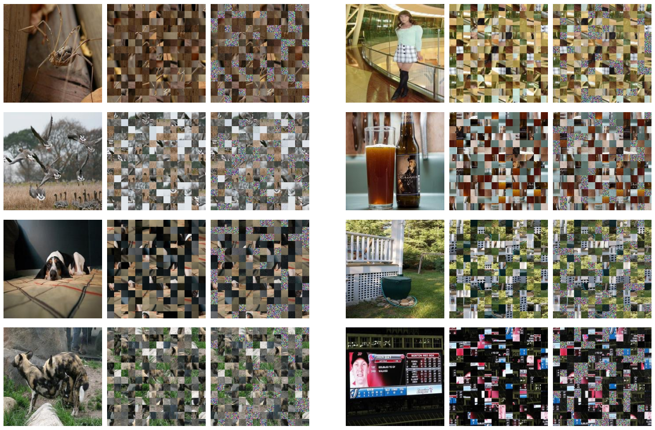

Figure 7: Impact of gradient leakage on the security of our method. Left: original images. Middle: images encrypted by RS. Right: recovered images.

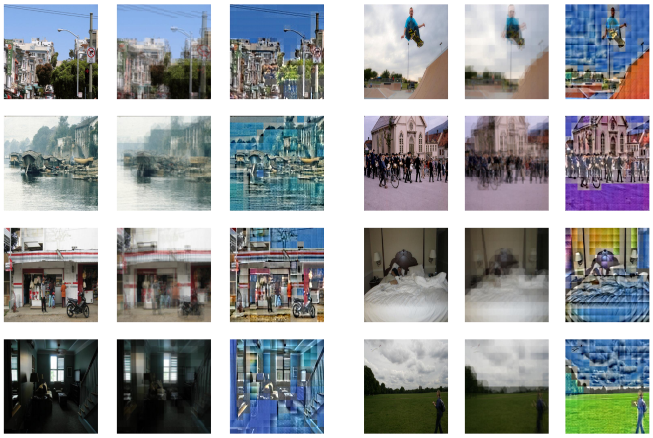

Figure 8: Reconstruction attack on MI. Left: original images. Middle: images encrypted by MI. Right: reconstructed images.

ViTDet can be readily obtained, achieving an AP of 41.2.

4.3 Attackers

Puzzle solver attacker. Jigsaw puzzle solver, i.e., reconstructing the image from the set of shuffled patches, can be used to attack the images encrypted with MS. Since this problem is NP-hard [Demaine and Demaine, 2007], solving jigsaw puzzles of non-trivial size is impossible. Most of the existing works in computer vision focus on the jigsaw puzzle problem composed of equally-sized image patches [Cho *et al.*, 2010; Pomeranz *et al.*, 2011; Sholomon *et al.*, 2013; Paikin and Tal, 2015], in which only pixels that are no more than two pixels away from the boundary of a piece are utilized. We choose the solver proposed in [Paikin and Tal, 2015] to inveterate the effect of attacks. The reason is twofold: (1) It is a fast, fully-automatic, and general solver, which assumes no prior knowledge about the original image and (2) It can handle puzzles with missing pieces.

The number of patches is the core factor determining the security level of our paradigm, i.e., the smaller the size, the more quantity, the safer. Also, there are other ways to enhance security such as adapting the way images are decomposed. Figure 5 and Figure 6 show that dropping some patches, reducing the patch size or interval, or increasing the image size can enhance the security at the cost of slight performance degradation. In particular, even if 10% of the image patches are dropped, the performance is only reduced by 1.6%, which allows users to drop patches containing sensitive information. It is to be noted that no extra computation cost is incurred by dropping some patches and reducing the interval.

Moreover, our paradigm can benefit from advances in transformers. Pixel Transformer [Nguyen *et al.*, 2024] treats an image as a set of pixels and shows the permutation-invariant property. This may facilitate pixel-level shuffling and mixing-up, further enhancing the security.

Gradient leakage attacker. It has been shown that the training set will be leaked by gradient sharing [Zhu *et al.*, 2019]. To evaluate the impact of gradient leakage on the security of our method, we use the gradient leakage attacker to recover the images encrypted by RS. The results are shown in Figure 7. It can be observed that (1) Gradient leakage attacker can restore images and (2) the restored images are encrypted. Therefore, we conclude that our paradigm does not prevent gradient leakage attacks, but can make the attacked images useless, thus protecting privacy.

Reconstruction attacker We use a recently proposed powerful Transformer-based framework, MAE [He *et al.*, 2022] (tiny), to recover the original clean images from the images encrypted by MI. We adapt MAE with two modifications: (1) Patches are not dropped and (2) The linear patch embedding is replaced by a nonlinear patch embedding, which is consistent with the patch embedding used in PEYOLOS. Here the nonlinear patch embedding is composed of two linear layers separated by a GELU activation. Corresponding results are shown in Figure 8. We observe that: (1) The style of reconstructed images is very different from original images and (2) Privacy-sensitive patches such as faces and texts are blurred, and thus the reconstruction with MAE still cannot reveal the original identity of faces or the contents of texts. These observations indicate that recovering the original clean natural images from images encrypted by MI is a great challenge, demonstrating the effectiveness of MI regarding privacy preserving.

5 Conclusion

In this paper, we propose an efficient privacy-preserving learning paradigm that can destroy human-recognizable contents while preserving machine-learnable information. The key insight of our paradigm is to decouple the encryption algorithm from the network optimization via permutation-invariance. Two encryption strategies are proposed to encrypt images: random shuffling to a set of equally-sized image patches and mixing image patches that are permutation-invariant. By adapting ViT and YOLOS with minimal adaptations, they can be made (partially) permutation-invariant and are able to handle encrypted images. Extensive experiments on ImageNet and COCO show that the proposed paradigm achieves comparable accuracy with the competitive methods, meanwhile destroying human-recognizable contents.

Acknowledgements

This work was supported by Guangdong Major Project of Basic and Applied Basic Research (2023B0303000016); National Natural Science Foundation of China (62206268, 62306163, U21A20487, 62372440, 62376261); the National Research Foundation, Singapore, under its NRF Professorship Award No. NRF-P2024-001; Natural Science Foundation of Guangdong Province (2025A1515012957, 2024A1515011754); Shenzhen High-tech Zone Development Special Plan Innovation Platform Construction Project; Shenzhen Technology Project (JCYJ20240813154723031); the proof of concept center for high precision and high resolution 4D imaging; Guangdong Technology Project (2023TX07Z126); Yunnan Science & Technology Project (202302AD080008, 202305AF150152).

References

[Abadi *et al.*, 2016] Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *ACM SIGSAC Conference on Computer and Communications Security*, pages 308–318, 2016.

[Ba *et al.*, 2016] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint:1607.06450*, 2016.

[Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229, 2020.

[Carlini *et al.*, 2021] Nicholas Carlini, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, and Florian Tramer. Neuracrypt is not private. *arXiv preprint:2108.07256*, 2021.

[C.E., 1949] Shannon C.E. Communication theory of secrecy systems. *The Bell System Technical Journal*, 28(4):656–715, 1949.

[Cho *et al.*, 2010] Taeg Sang Cho, Shai Avidan, and William T. Freeman. A probabilistic image jigsaw puzzle solver. In *CVPR*, pages 183–190, 2010.

[Demaine and Demaine, 2007] E. Demaine and M. Demaine. Jigsaw puzzles, edge matching,and polyomino packing: Connections and complexity. *Graphs and Combinatorics*, 23:195–208, 2007.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint:2010.11929*, 2020.

[Fang *et al.*, 2021] Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. You only look at one sequence: Rethinking transformer in vision through object detection. In *NeurIPS*, pages 26183–26197, 2021.

[Fioretto *et al.*, 2021] Ferdinando Fioretto, Pascal Van Hentenryck, and Keyu Zhu. Differential privacy of hierarchical census data: An optimization approach. *Artificial Intelligence*, 296:103475, 2021.

[Fredrikson *et al.*, 2015] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *ACM SIGSAC Conference on Computer and Communications Security*, page 1322–1333, 2015.

[Hatamizadeh *et al.*, 2022] Ali Hatamizadeh, Hongxu Yin, Holger Roth, Wenqi Li, Jan Kautz, Daguang Xu, and Pavlo Molchanov. Gradvit: Gradient inversion of vision transformers. In *CVPR*, pages 10021–10030, 2022.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[He *et al.*, 2022] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022.

[Hendrycks and Gimpel, 2016] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint:1606.08415*, 2016.

[Huang *et al.*, 2020] Yangsibo Huang, Zhao Song, Kai Li, and Sanjeev Arora. InstaHide: Instance-hiding schemes for private distributed learning. In *ICML*, pages 4507–4518, 2020.

[Karthik *et al.*, 2019] Nandakumar Karthik, Ratha Nalini, Pankanti Sharath, and Halevi Shai. Towards deep neural network training on encrypted data. In *CVPR workshops*, pages 40–48, 2019.

[Li *et al.*, 2022] Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. *arXiv preprint:2203.16527*, 2022.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint:2103.14030*, 2021.

[Liu *et al.*, 2022] Ziyao Liu, Jiale Guo, Kwok-Yan Lam, and Jun Zhao. Efficient dropout-resilient aggregation for privacy-preserving machine learning. *IEEE Transactions on Information Forensics and Security*, 2022.

[Lopez *et al.*, 2024] Jhon Lopez, Carlos Hinojosa, Henry Arguello, and Bernard Ghanem. Privacy-preserving optics for enhancing protection in face de-identification. In *CVPR*, pages 12120–12129, 2024.

[Ma *et al.*, 2023] Sihan Ma, Jizhizi Li, Jing Zhang, He Zhang, and Dacheng Tao. Rethinking portrait matting with privacy preserving. *IJCV*, 2023.

[Madono *et al.*, 2020] Koki Madono, Masayuki Tanaka, Masaki Onishi, and Tetsuji Ogawa. Block-wise scrambled image recognition using adaptation network. In *Workshop on Artificial Intelligence of Things*, 2020.

[Malkin *et al.*, 2021] Nikolay Malkin, Sameera Lanka, Pranav Goel, and Nebojsa Jojic. Studying word order through iterative shuffling. In *EMNLP*, page 10351–10366, 2021.

[Mi *et al.*, 2024] Yuxi Mi, Zhizhou Zhong, Yuge Huang, Jiazhen Ji, Jianqing Xu, Jun Wang, Shaoming Wang, Shouhong Ding, and Shuigeng Zhou. Privacy-preserving face recognition using trainable feature subtraction. In *CVPR*, pages 297–307, 2024.

[Moniruzzaman *et al.*, 2022] Md Moniruzzaman, Zhaozheng Yin, Zhihai He, Ruwen Qin, and Ming C Leu. Human action recognition by discriminative feature pooling and video segment attention model. *IEEE Transactions on Multimedia*, 24:689–701, 2022.

[Naseer *et al.*, 2021] Muzammal Naseer, Kanchana Ranasinghe, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. In *NeurIPS*, pages 23296–23308, 2021.

[Nguyen *et al.*, 2024] Duy-Kien Nguyen, Mahmoud Assran, Unnat Jain, Martin R. Oswald, Cees G. M. Snoek, and Xinlei Chen. An image is worth more than 16x16 patches: Exploring transformers on individual pixels. *arXiv preprint:2406.09415*, 2024.

[Paikin and Tal, 2015] Genady Paikin and Ayellet Tal. Solving multiple square jigsaw puzzles with missing pieces. In *CVPR*, pages 4832–4839, 2015.

[Pomeranz *et al.*, 2011] Dolev Pomeranz, Michal Shemesh, and Ohad Ben-Shahar. A fully automated greedy square jigsaw puzzle solver. In *CVPR*, pages 9–16, 2011.

[Ren *et al.*, 2023] Bin Ren, Yahui Liu, Yue Song, Wei Bi, Rita Cucchiara, Nicu Sebe, and Wei Wang. Masked jigsaw puzzle: A versatile position embedding for vision transformers. In *CVPR*, pages 20382–20391, 2023.

[Shaw *et al.*, 2018] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint:1803.02155*, 2018.

[Sholomon *et al.*, 2013] Dror Sholomon, Omid David, and Nathan S. Netanyahu. A genetic algorithm-based solver for very large jigsaw puzzles. In *CVPR*, pages 1767–1774, 2013.

[Sinha *et al.*, 2021] Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *EMNLP*, page 2888–2913, 2021.

[Sirichotedumrong *et al.*, 2019] Warit Sirichotedumrong, Yuma Kinoshita, and Hitoshi Kiya. Pixel-based image encryption without key management for privacy-preserving deep neural networks. *IEEE Access*, 7:177844–177855, 2019.

[Stinson and Paterson, 2019] Douglas R. Stinson and Maura B. Paterson. Cryptography: theory and practice (fourth edition). *Chapman and Hall/CRC*, 2019.

[Tanaka, 2018] Masayuki Tanaka. Learnable image encryption. In *IEEE International Conference on Consumer Electronics-Taiwan*, pages 1–2, 2018.

[Touvron *et al.*, 2020] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint:2012.12877*, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint:1706.03762*, 2017.

[Wang *et al.*, 2021] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *arXiv preprint:2102.12122*, 2021.

[Wightman, 2019] Ross Wightman. Pytorch image models. In *https://github.com/rwightman/pytorch-image-models*, 2019.

[Xu *et al.*, 2019] Runhua Xu, James B.D. Joshi, and Chao Li. Cryptonn: Training neural networks over encrypted data. In *International Conference on Distributed Computing Systems*, 2019.

[Xu *et al.*, 2021] Runhua Xu, Nathalie Baracaldo, and James Joshi. Privacy-preserving machine learning: Methods, challenges and directions. *arXiv preprint:2108.04417*, 2021.

[Yang *et al.*, 2011] Xingwei Yang, Nagesh Adluru, and Longin Jan Latecki. Particle filter with state permutations for solving image jigsaw puzzles. In *CVPR*, pages 2873–2880, 2011.

[Yevgeniy *et al.*, 2017] Dodis Yevgeniy, Katz Jonathan, Steinberger John, Thiruvengadam Aishwarya, and Zhang Zhe. Provable security of substitution-permutation networks. *Cryptology ePrint Archive*, 2017.

[Yuan *et al.*, 2021] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint:2101.11986*, 2021.

[Zhang *et al.*, 2023] Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. Vitaev2: Vision transformer advanced by exploring inductive bias for image recognition and beyond. *IJCV*, 2023.

[Zhu *et al.*, 2019] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In *NeurIPS*, 2019.