一般社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
IEICE Technical Report
EMM2024-105(2025-01)

制限ランダム置換行列を用いたVision Transformerのための プライバシー保護強化法

堀尾　虹輝†　西川　清史†　貴家　仁志†

† 東京都立大学システムデザイン学部　〒191–0065 東京都日野市旭ヶ丘 6-6
E-mail: †horio-kouki@ed.tmu.ac.jp, ††{kiyoshi,kiya}@tmu.ac.jp

あらまし　Vision Transformer (ViT) のための新しいプライバシー保護画像分類法を提案する．提案法では，制限ランダム置換行列を用いて事前学習モデルに暗号化を施し，かつその制限ランダム置換行列に基づきクエリ画像及びモデルのファインチューニング用画像に暗号化を施す．先行研究において，制限ランダム行列の使用が，画像暗号化の影響に起因するモデルの性能劣化の課題を改善できることが指摘された．しかし，同時に制限の影響によって，画像の視覚保護強度が低下するという新たな課題が明らかになった．本稿では，制限ランダム置換行列を用いた場合にも視覚保護強度の低下を完全に回避可能な方法を提案する．実験では，画像分類において分類精度と視覚保護強度の観点から提案法の有効性を示す．

キーワード　Vision Transformer，プライバシー保護，画像暗号化

A method for enhancing privacy-preserving Vision Transformer using restricted random permutation matrices

Kouki HORIO†, Kiyoshi NISHIKAWA†, and Hitoshi KIYA†

† Department of Computer Science, Tokyo Metropolitan University　6-6 Asahigaoka, Hino, Tokyo, 191–0065 Japan
E-mail: †horio-kouki@ed.tmu.ac.jp, ††{kiyoshi,kiya}@tmu.ac.jp

Abstract　A novel method using the Vision Transformer (ViT) for privacy-preserving image classification tasks is proposed. In the proposed method, a pre-trained model is encrypted with restricted random permutation matrices, and query images and images for fine-tuning a model are also encrypted on the basis of the random matrices. Previous studies pointed out that the use of restricted random permutation matrices can improve the performance degradation of the model caused by using encrypted images, but the invisibility of encrypted images is degraded by using the random matrices. Accordingly, we propose a novel framework that can perfectly maintain the invisibility of encrypted images even when restricted matrices are used.

Key words　Vision Transformer, Privacy Preserving, Image Encryption

1. まえがき

深層ニューラルネットワーク（deep neural networks: DNNs）は，多方面での応用が期待され，進展が続いている．一方，DNN モデルの学習には，一般に膨大な学習用データが必要で，かつその計算コストは大きい．そのため，クラウド環境を利用して，DNN の学習の実行やサービスを展開することが多い．しかし，クラウド環境は信頼できるとは言えず，データに含まれるプライバシー情報をいかに保護するかが課題となっている [1]〜[3]．本稿では，Vision Transformer (ViT) の使用を前提にして，画像分類タスクを例にして上述の課題に取り組む．

DNN におけるプライバシー保護の方法の一つに，知覚暗号化されたデータを利用する方法がある [1]．この暗号化は，学習可能な暗号化とも呼ばれ，視覚情報が保護された暗号化画像を DNN モデルに直接入力することができる．また，プライバシー保護に限らず，モデルの誤作動を目的に画像にノイズを付加する敵対的サンプル攻撃に対する防御においても [4]〜[6]，学習可能な暗号化の有用性が確認されている．

プライバシー保護のための他のアプローチとして，暗号化したまま計算ができる準同型暗号の利用等がある [7], [8]．しかし，計算量が大きいことに加え，広く知られたモデルに直接暗号化データを入力することはできない．さらに，敵対的サンプル攻撃への使用は検討されていない．また，複数のクライアントが各自のデータを直接公開せずに，共同でモデル学習を可能とする連合学習がある [3], [9]．しかし，このアプローチは，クエリ画像のプライバシー保護を行うことはできず，かつ敵対的

This article is a technical report without peer review, and its polished and/or extended version may be published elsewhere.
Copyright ©2025 by IEICE

サンプル攻撃への対策にも有効ではない．

以上の背景から，本稿では，事前学習された ViT を，学習可能な暗号化を施したデータによってファインチューニングする新しい方法を提案する．暗号化画像を用いた従来のファインチューニング法は，暗号化の影響によって，モデルの性能を低下させた [10],[11]．その後，モデル性能の低下の課題を解決する方法が提案されたが [12],[13]，学習用データのプライバシー保護に十分適応できず，また敵対的サンプル攻撃に効果がないことが明らかになった．

提案法は，制限ランダム行列の使用によって，モデル性能の低下と同時に，学習用データの保護と敵対的サンプル攻撃への応用を可能にする．提案法では，ViT の持つ 2 つの埋め込み構造を利用して事前学習モデルを，制限ランダム置換行列を用いて暗号化する．また，制限ランダム置換行列は，ファインチューニングのための学習画像及びクエリ画像を暗号化する際にも使用される．実験では，ランダム置換行列の選択によって，ファインチューニングの学習効率や学習されたモデルの性能を制御できることを示す．さらに、制限の選択にかかわらず、画像の視認性を常に維持できることを確認する。

2. 関 連 研 究

2.1 Vision Transformer

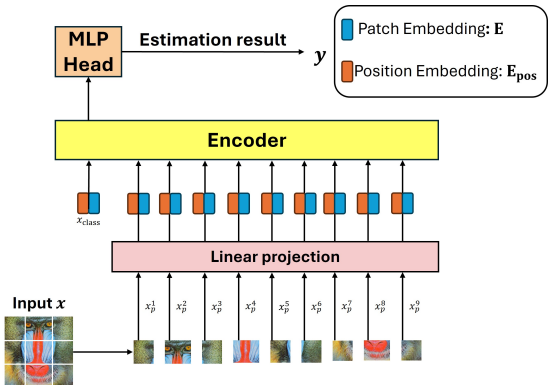


図 1: ViT のモデル構造の概要（$N = 9$）

Vision Transformer (ViT) [14] は，画像分類タスクにおいて，高い分類性能を提供することが知られている．図 1 に示すように，ViT では，入力画像 $x \in \mathbb{R}^{h \times w \times c}$ をサイズ $p \times p$ の N 個のパッチに分割する．ここで，h, w, c は入力画像の幅，高さ，チャネル数を表し，N は $N = hw/p^2$ で与えられる．次に，パッチ i $(i = 1, ..., N)$ はベクトル $x_p^i = [x_p^i(1), x_p^i(2), \cdots, x_p^i(p^2c)]$ に平坦化され，Patch Embedding が行われる．その後，クラストークンが付与され，Position Embedding が行われる．以上の処理により，パッチ数列は次のように与えられ，Encoder へ入力される．

$$z_0 = [x_{\text{class}}; x_p^1\mathbf{E}; \cdots; x_p^i\mathbf{E}; \cdots; x_p^N\mathbf{E}] + \mathbf{E}_{\mathbf{pos}} \qquad (1)$$

ここで，

$$\mathbf{E}_{\mathbf{pos}} = \{(e_{pos}^0)(e_{pos}^1)\cdots(e_{pos}^i)\cdots(e_{pos}^N)\},$$
$$\mathbf{E} \in \mathbb{R}^{L \times D}, \quad \mathbf{E}_{\mathbf{pos}} \in \mathbb{R}^{(N+1) \times D},$$
$$x_{\text{class}} \in \mathbb{R}^D, \quad L = p^2c, \quad x_p^i \in \mathbb{R}^L, \quad e_{pos}^i \in \mathbb{R}^D.$$

x_{class} はクラストークン，$\mathbf{E}$ は長さ L の平坦化された各パッチを次元 D に線形変換する Patch Embedding，$\mathbf{E}_{\mathbf{pos}}$ は各パッチに位置情報を付与する Position Embedding である．e_{pos}^0 はクラストークンの位置情報，他の e_{pos}^i は各パッチの位置情報，z_0 は Encoder への入力である．Encoder はクラストークンのみを出力し，MLP Head が x のクラスを推定する．

2.2 暗号化画像を用いた学習

学習可能な暗号化を用いたモデル学習の先行研究を要約する．乱数を用いて画像を変換（知覚暗号化）し，深層学習モデルに適応する研究は，圧縮可能な暗号化に着眼を得て 2018 年から始まった [15],[16]．図 2 は，プライバシー保護を考慮して画像分類を，暗号画像を用いて実行する方法の一例である．学習用画像は暗号化され，事前学習モデルをファインチューニングするために使用される．一方，クエリ画像は，同様に暗号化され，ファインチューニングされたモデル（暗号化モデル）に入力され，分類結果を推定される．クラウドプロバイダは，一般に信頼できるとは限らない．このフレームワークでは，プロバイダは，視覚情報及び暗号化鍵を持っておらず，クライアントの視覚情報を保護したままで処理が実行される．

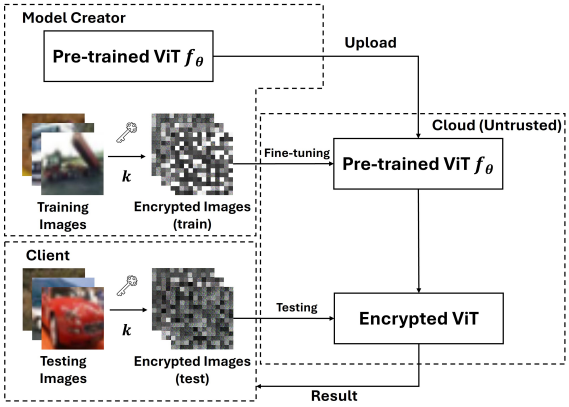


図 2: プライバシー保護を考慮した画像分類例

これまで様々の学習可能な暗号化法が，提案されてきた [1],[17],[18]．同時に，暗号化データに対する攻撃手法も研究され，暗号化法の攻撃耐性向上に貢献してきた [19]．しかし，従来の暗号化法の多くは，暗号化データの使用の影響によって，モデルの性能を低下させてしまうという課題があった．その課題を解決する手法として，ViT の埋め込み構造を利用した暗号化手法が提案された．この手法では，モデル学習を非暗号化画像で行い，その後に学習済みモデルの情報の一部をランダムに変換する [12]．従って，クエリ画像のプライバシー保護が可能にするが，クラウド上でモデル学習を，暗号化画像を用いて実行することはできない．さらに，この制約は，敵対的サンプル攻撃対策としての暗号化画像の有効性を消滅させてしまう．

一方，モデルの性能の低下という課題を解決するために，ViT において制限ランダム置換行列を用いて画像に暗号化を施す方法が提案された [20]．しかし，この方法は，モデルの性能の低

下を改善できるが，暗号化画像の非視認性の効果を低下させるという新たな課題を持つ．本稿の目的は，制限ランダム置換行列を用いて画像に暗号化を施した場合にも，暗号化画像の非視認性の効果を低下させない，新しいフレームワークを提案することである．

3. 提案方法

3.1 概　　要

提案手法の概要を図 3 に示す．モデル作成者は，事前学習モデル f_θ の埋め込み層（Patch Embedding と Position Embedding）を鍵 k_M を用いて生成されたランダム置換行列で暗号化し，暗号化された事前学習モデル $\hat{f}_\theta$ を生成する．その後，$\hat{f}_\theta$ はクラウド上にアップロードされる．モデル作成者は，鍵 k_I を用いて学習用画像に暗号化を施し，ラベルと一緒に暗号化画像をクラウドに送る．クラウドプロバイダは，暗号化された学習用画像とラベルを用いて $\hat{f}_\theta$ をファインチューニングし，モデル学習を完了する．一方，クライアントは，モデル作成者から鍵 k_I を受け取り，クエリ画像に暗号化を施す．暗号化されたクエリ画像は，クラウドに送られ，学習済みのモデルによって分類される．ユーザは，分類結果をクラウドから受け取る．

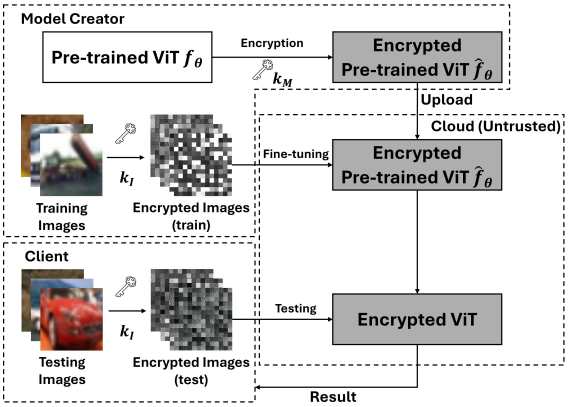


図 3: 提案法のフレームワーク

上述の手順において，クラウドが受け取る画像は視覚情報が保護されており，かつ暗号鍵はクラウドに共有されていない．以下では，各手順の詳細について述べる．本稿の貢献は，事前学習モデルと画像の暗号化において使用されるランダム置換行列の選択とその有効性の確認にある．

3.2 制限ランダム置換行列を用いた事前学習モデルの変換

ランダム置換行列を用いて，ViT の埋め込み構造を変換することにより事前学習モデルを暗号化する．秘密鍵によって 2 つのランダム置換行列 $\mathbf{E}^{\mathbf{M}}_{\mathbf{bs}} \in \mathbb{R}^{N\times N}$ と $\mathbf{E}^{\mathbf{M}}_{\mathbf{ps}} \in \mathbb{R}^{L\times L}$ を生成する．$\mathbf{E}^{\mathbf{M}}_{\mathbf{bs}}$ は Position Embedding $\mathbf{E}_{\mathbf{pos}}$ を変換するために使用され，$\mathbf{E}^{\mathbf{M}}_{\mathbf{ps}}$ は Patch Embedding $\mathbf{E}$ に対して使用される．

今，式 (1) の $\mathbf{E}_{\mathbf{pos}}$ とクラストークン x_{class} を同時に扱うために，$\mathbf{E}^{\mathbf{M}}_{\mathbf{bs}}$ を次式のように拡張する．

$$\hat{\mathbf{E}}^{\mathbf{M}}_{\mathbf{bs}} = \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{E}^{\mathbf{M}}_{\mathbf{bs}} \end{pmatrix} \in \mathbb{R}^{(N+1)\times(N+1)} \tag{2}$$

次に，提案法では，$\mathbf{E}_{\mathbf{pos}}$ と $\mathbf{E}$ を次式に従いそれぞれ変換する．

$$\hat{\mathbf{E}}_{\mathbf{pos}} = \hat{\mathbf{E}}^{\mathbf{M}}_{\mathbf{bs}}\mathbf{E}_{\mathbf{pos}} \tag{3}$$

$$\hat{\mathbf{E}} = \mathbf{E}^{\mathbf{M}}_{\mathbf{ps}}\mathbf{E} \tag{4}$$

後述するように，本稿では，$\mathbf{E}^{\mathbf{M}}_{\mathbf{bs}}$ と $\mathbf{E}^{\mathbf{M}}_{\mathbf{ps}}$ に制限ランダム置換行列を適用する．以上の $\mathbf{E}$ と $\mathbf{E}_{\mathbf{pos}}$ の暗号化によって，式 (1) の埋め込みのパッチ数列は次式となる．

$$\hat{z_0} = [x_{\text{class}}; x^1_p\hat{\mathbf{E}}; \cdots; x^i_p\hat{\mathbf{E}}; \cdots; x^N_p\hat{\mathbf{E}}] + \hat{\mathbf{E}}_{\mathbf{pos}} \tag{5}$$

3.3 ランダム置換行列を用いた画像暗号化

提案法では，ランダム置換行列を用いて画像に暗号化を施し，それらをモデルのファインチューニングとクエリ画像として用いる．以下に，暗号化の手順を要約する．

入力画像 $x \in \mathbb{R}^{h\times w\times c}$ を，大きさ $p\times p$ の N 個の非重複ブロック $B = \{B_1, \cdots, B_i, \cdots, B_N\}$ に分割する．次に，各ブロックに対して，ブロックスクランブルとピクセルシャッフルを施す．

A. ブロックスクランブル

ブロックスクランブルのためのランダム置換行列 $\mathbf{E}^{\mathbf{I}}_{\mathbf{bs}} \in \mathbb{R}^{N\times N}$ を準備し，以下の手順でブロック B を並び替える．

A-a) ブロック集合 B の各要素順番に対応するベクトル

$$\boldsymbol{c} = (1, 2, \cdots, N)^\top \tag{6}$$

を定義する．

A-b) $\mathbf{E}^{\mathbf{I}}_{\mathbf{bs}}$ を用いて次式により変換する．

$$\begin{aligned}\hat{\boldsymbol{c}} &= \mathbf{E}^{\mathbf{I}}_{\mathbf{bs}}\boldsymbol{c} \\ &= (\hat{c}(1), \hat{c}(2), \cdots, \hat{c}(N))^\top\end{aligned} \tag{7}$$

ここで，$\hat{c}(i) \in \{1, 2, \cdots, N\}$ である．

A-c) 並び替えられたブロックの集合 $\hat{B} = \{\hat{B_1}, \cdots, \hat{B_N}\}$ の要素 $\hat{B}_i$ を

$$\hat{B}_i = B_{\hat{c}(i)} \tag{8}$$

と与える．$\hat{B}$ を連結し，暗号化画像 $\hat{x}$ を生成する．この画像が，ブロックスクランブルされた画像になる．

B. ピクセルシャッフル

B-a) 各ブロック B_i を次のようにベクトル $\boldsymbol{x}^{\boldsymbol{i}}_{\boldsymbol{p}} \in \mathbb{R}^L$ に平坦化する．

$$\boldsymbol{x}^{\boldsymbol{i}}_{\boldsymbol{p}} = [x^i_p(1), \cdots, x^i_p(L)] \tag{9}$$

B-b) ピクセルシャッフルのためのランダム置換行列 $\mathbf{E}^{\mathbf{I}}_{\mathbf{ps}} \in \mathbb{R}^{L\times L}$ を準備し，次式に従い各ブロックのピクセル値を並び替える．

$$\begin{aligned}\tilde{\boldsymbol{x}}^{\boldsymbol{i}}_{\boldsymbol{p}} &= \boldsymbol{x}^{\boldsymbol{i}}_{\boldsymbol{p}}\mathbf{E}^{\mathbf{I}}_{\mathbf{ps}}{}^\top \\ &= [\tilde{x}^i_p(1), \cdots, \tilde{x}^i_p(L)]\end{aligned} \tag{10}$$

ここで，$\tilde{x}^i_p(j) \in [x^i_p(1), \cdots, x^i_p(L)]$ である．

B-c) $\tilde{\boldsymbol{x}}_{\boldsymbol{p}}^{\boldsymbol{i}}$ から各ブロック要素 $\tilde{B}_i$ を構築する．

B-d) $\tilde{B}_i$ を連結し，暗号化された画像 $\tilde{x}$ にする．この画像が，ピクセルシャッフルされた画像になる．

ブロックスクランブルとピクセルシャッフルは，独立な操作である．本稿では，実行の順番に制約を与えない．特に断らない限り，暗号化画像はこの 2 つの操作によって生成される．

3.4 制限ランダム置換行列

上述したように，提案法では，4 つのランダム置換行列（$\mathbf{E}_{\mathbf{bs}}^{\mathbf{M}}$, $\mathbf{E}_{\mathbf{ps}}^{\mathbf{M}}$, $\mathbf{E}_{\mathbf{bs}}^{\mathbf{I}}$, $\mathbf{E}_{\mathbf{ps}}^{\mathbf{I}}$）をモデルと画像の変換のために使用する．ここでは，4 つの行列の生成法と関係について述べる．

暗号化画像を用いた従来法では，まずモデルを暗号化せず，画像のみを暗号化する方法が提案された [1],[10]〜[12],[15]．しかし，それらの方法では，画像の視覚情報を保護することはできるが，暗号化無しの場合と比べ，モデル性能が劣化する．本稿では，モデル性能の低下を極力抑え，かつ暗号化画像の視覚情報の保護を維持可能な方法を提案する．以下では，提案される制限ランダム置換行列について述べる．

議論を単純化するために，$N = 5$ の $\mathbf{E}_{\mathbf{bs}}^{\mathbf{M}}$ を例にして，提案法で使用する制限ランダム置換行列を説明する．さらに，$N \times N$ の置換行列 $\mathbf{E}_{\mathbf{bs}}^{\mathbf{M}}$ において，対角要素の個数を $N_{bs} = 3$ に制限した例を示す．

$$\mathbf{E}_{\mathbf{bs}}^{\mathbf{M}} = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 1^* & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1^* & 0 \\ 0 & 0 & 0 & 0 & 1^* \end{pmatrix} = P + D$$

$$= \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1^* & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1^* & 0 \\ 0 & 0 & 0 & 0 & 1^* \end{pmatrix} \quad (11)$$

”*”のついた要素は，$N_{bs} = 3$ の制限された対角要素を表す．$\mathbf{E}_{\mathbf{bs}}^{\mathbf{M}}$ は対角要素の行列 D と対角以外を要素とする行列 P 分けて表現することができる．行列 P の要素は，与えられた行列 D の下で，$\mathbf{E}_{\mathbf{bs}}^{\mathbf{M}}$ が置換行列を維持できるようにランダムに決定される．

一方，画像暗号化に使われる置換行列 $\mathbf{E}_{\mathbf{bs}}^{\mathbf{I}}$ は，(11) に対応する形式で決定される．$\mathbf{E}_{\mathbf{bs}}^{\mathbf{M}}$ が式 (11) で表されるとき，例えば $\mathbf{E}_{\mathbf{bs}}^{\mathbf{I}}$ は以下のようになる．

$$\mathbf{E}_{\mathbf{bs}}^{\mathbf{I}} = P + \tilde{D}$$

$$= \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1^* \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 1^* & 0 & 0 & 0 \\ 0 & 0 & 0 & 1^* & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1^* \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1^* & 0 & 0 & 0 \\ 0 & 0 & 0 & 1^* & 0 \end{pmatrix} \quad (12)$$

ここで，$\tilde{D}$ は，(11) の行列 D において，対角要素に制限した 3 つの行（2, 4, 5 行）にランダムに”1”の要素を位置付けた行列である．

Position Embedding $\mathbf{E}_{\mathbf{pos}}$ の変換に用いる置換行列 $\mathbf{E}_{\mathbf{bs}}^{\mathbf{M}}$ において，制限される対角要素の個数を $N_{bs} \leq N$ とする．Patch Embedding $\mathbf{E}$ の変換に用いる置換行列 $\mathbf{E}_{\mathbf{ps}}^{\mathbf{M}}$ に対しても，同様に制限ランダム置換行列を定義することができる．行列サイズが $L \times L$ であること，固定される対角要素の個数が $N_{ps} \leq L$ であることが $\mathbf{E}_{\mathbf{bs}}^{\mathbf{M}}$ との違いである．もしモデル暗号化に使われる置換行列の並び替えが全て対角に制限されたとき（$N_{bs} = N, N_{ps} = L$），すなわち $\mathbf{E}_{\mathbf{bs}}^{\mathbf{M}}$ と $\mathbf{E}_{\mathbf{ps}}^{\mathbf{M}}$ が単位行列のとき，モデルは暗号化されず，従来法 [20] と一致する．

3.5 制限ランダム置換行列の効果

本稿では，画像の暗号化と同時に事前学習されたモデルを暗号化する方法を提案する．さらに，その暗号化は制限ランダム置換行列に基づいて実行される．以下では，制限なしの行列を用いた場合，モデル性能への暗号化の影響を完全に打ち消せることを示す．換言すると，制限によってモデル性能への影響を制御できる．

A. ピクセルシャッフルの影響

ピクセルシャッフルされた画像 $\tilde{x}$ を暗号化された ViT に入力する．このとき，Encoder へ入力 $\hat{z_0}$ は，式 (5) と (10) から，

$$\begin{aligned} \hat{z_0} &= [x_{\text{class}}; \tilde{x}_p^1 \hat{\mathbf{E}}; \cdots; \tilde{x}_p^N \hat{\mathbf{E}}] + \hat{\mathbf{E}}_{\mathbf{pos}} \\ &= [x_{\text{class}}; x_p^1 \mathbf{E}_{\mathbf{ps}}^{\mathbf{I}}{}^{\top} \mathbf{E}_{\mathbf{ps}}^{\mathbf{M}} \mathbf{E}; \cdots; x_p^N \mathbf{E}_{\mathbf{ps}}^{\mathbf{I}}{}^{\top} \mathbf{E}_{\mathbf{ps}}^{\mathbf{M}} \mathbf{E}] + \hat{\mathbf{E}}_{\mathbf{pos}} \end{aligned} \quad (13)$$

と整理される．置換行列 $\mathbf{E}_{\mathbf{ps}}^{\mathbf{M}}$ は，どの対角要素も制限されなかった場合 ($N_{ps} = 0$),

$$\mathbf{E}_{\mathbf{ps}}^{\mathbf{I}} = \mathbf{E}_{\mathbf{ps}}^{\mathbf{M}} \quad (14)$$

を満たし，さらに，$\mathbf{E}_{\mathbf{ps}}^{\mathbf{I}}$ は置換行列なので，

$$\mathbf{E}_{\mathbf{ps}}^{\mathbf{I}}{}^{\top} = \mathbf{E}_{\mathbf{ps}}^{\mathbf{I}}{}^{-1} \quad (15)$$

が成立する．よって，Encoder へ入力は

$$\hat{z_0} = [x_{\text{class}}; x_p^1 \mathbf{E}; \cdots; x_p^N \mathbf{E}] + \hat{\mathbf{E}}_{\mathbf{pos}} \quad (16)$$

となり，クエリ画像の暗号化の影響を完全に回避できる．

B. ブロックスクランブルの影響

ブロックスクランブルされた画像 $\hat{x}$ を暗号化された ViT に入力する．このとき，Encoder への入力 $\hat{z_0}$ は，式 (5) と (8) から，

$$\hat{z_0} = \hat{\mathbf{E}}_{\mathbf{bs}}^{\mathbf{I}} [x_{\text{class}}; x_p^1 \hat{\mathbf{E}}; \cdots; x_p^N \hat{\mathbf{E}}] + \hat{\mathbf{E}}_{\mathbf{pos}} \quad (17)$$

となる．ここで，$\hat{\mathbf{E}}^{\mathbf{I}}_{\mathbf{bs}}$ は，クラストークンは並び替えず，画像のブロックのみを並び替えを表す置換行列で，$\mathbf{E}^{\mathbf{I}}_{\mathbf{bs}}$ を

$$\hat{\mathbf{E}}^{\mathbf{I}}_{\mathbf{bs}} = \begin{pmatrix} 1 & 0 \\ 0 & \mathbf{E}^{\mathbf{I}}_{\mathbf{bs}} \end{pmatrix} \in \mathbb{R}^{(N+1)\times(N+1)} \tag{18}$$

と拡張した行列である．式 (3) から，式 (17) は

$$z_0 = \hat{\mathbf{E}}^{\mathbf{I}}_{\mathbf{bs}}[x_{\text{class}}; x_p^1\hat{\mathbf{E}}; \cdots; x_p^N\hat{\mathbf{E}}] + \hat{\mathbf{E}}^{\mathbf{M}}_{\mathbf{bs}}\mathbf{E}_{\mathbf{pos}} \tag{19}$$

となる．置換行列 $\mathbf{E}^{\mathbf{M}}_{\mathbf{bs}}$ において，どの対角要素も制限されなかった場合 ($N_{bs} = 0$),

$$\mathbf{E}^{\mathbf{I}}_{\mathbf{bs}} = \mathbf{E}^{\mathbf{M}}_{\mathbf{bs}} \tag{20}$$

を満たす．よって，Encoder へ入力は

$$\begin{aligned} z_0 &= \hat{\mathbf{E}}^{\mathbf{I}}_{\mathbf{bs}}[x_{\text{class}}; x_p^1\hat{\mathbf{E}}; \cdots; x_p^N\hat{\mathbf{E}}] + \hat{\mathbf{E}}^{\mathbf{M}}_{\mathbf{bs}}\mathbf{E}_{\mathbf{pos}} \\ &= \hat{\mathbf{E}}^{\mathbf{I}}_{\mathbf{bs}}([x_{\text{class}}; x_p^1\hat{\mathbf{E}}; \cdots; x_p^N\hat{\mathbf{E}}] + \mathbf{E}_{\mathbf{pos}}) \end{aligned} \tag{21}$$

となる．

上式において，$\hat{\mathbf{E}}$ の影響は先に述べたようにクエリ画像のピクセルシャッフリングによって容易に回避できる．また，$\hat{\mathbf{E}}^{\mathbf{I}}_{\mathbf{bs}}$ 影響，すなわちブロックの並び替えの影響は，各パッチ x_p^i の順番の並び替えと $\mathbf{E}_{\mathbf{pos}}$ の順番に共通に残る．この影響の共通化によって，モデルの分類性能は，$\hat{\mathbf{E}}^{\mathbf{I}}_{\mathbf{bs}}$ によって低下することはない．換言すれば，事前学習済みモデルの $\mathbf{E}$ と $\mathbf{E}_{\mathbf{pos}}$ を暗号化し，次に暗号化画像を用いてファインチューニング場合の $\mathbf{E}$ と $\mathbf{E}_{\mathbf{pos}}$ は，暗号化なしの画像を用いてファインチューニングした際の $\mathbf{E}$ と $\mathbf{E}_{\mathbf{pos}}$ をそれぞれ暗号化したものに一致する [21].

一方，制限ランダム行列 ($N_{bs} \leq 0, N_{ps} \leq 0$) を事前学習済みモデルの暗号化とファインチューニング用の画像の暗号化に適用した場合，学習されたモデルへの影響は，$\mathbf{E}$ と $\mathbf{E}_{\mathbf{pos}}$ に留まらず，他のモデルパラメータも，暗号化なしの画像を用いてファインチューニングしたモデルのパラメータとは一致しない．したがって，制限ランダム置換行列によって暗号化されたクエリ画像をモデルに入力したとき，制限が強いほどその影響は，モデル性能を低下させる．換言すると，提案法は，制限の適切な選択によって，その影響の大きさを柔軟に設計することを可能とする．さらに，提案法では，式 (12) に例示するように，制限を強めても画像暗号化に使用される置換行列に対角要素を増やす必要はなく，画像の非視認性を制限によって犠牲にしない．

4. 実　　験

事前学習済みの ViT を用いて画像分類の実験を行い，提案法の有効性を確認する．

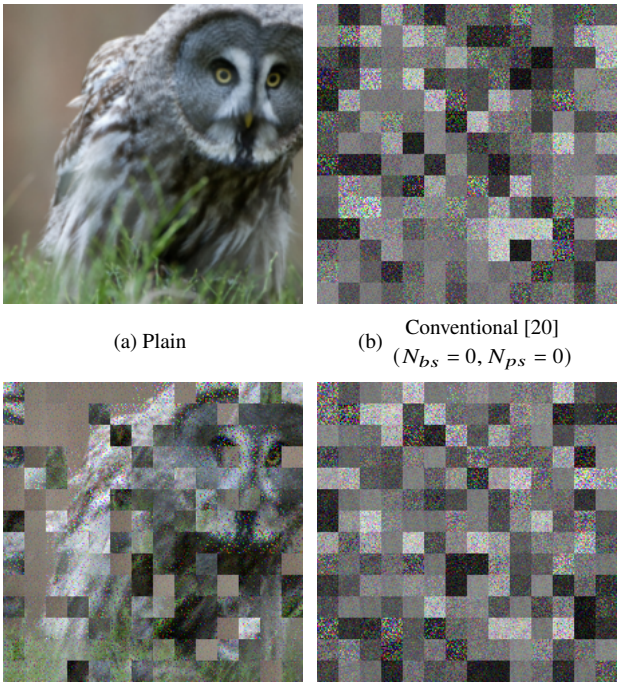
4.1 実験準備

CIFAR-10 データセットを用いて画像分類の実験を行った．CIFAR-10 は，10 クラス，計 6 万枚の画像（学習画像 5 万枚，テスト画像 1 万枚）から構成されている．学習画像は，4 万枚を訓練用データ，1 万枚を検証用データに分割して使用した．全ての画像は，$224 \times 224 \times 3$ にリサイズして使用した．バッチサイズ 64，学習率 0.0001，momentum 0.9，weight decay 0.0005 の条件で，Pytorch に実装された確率的勾配降下アルゴリズム（SGD）を用いて 50 エポック学習した．学習率の調整には，timm ライブラリの CosineLRScheduler を使用した．初期学習率は 0.00005 に設定し，最初の 10 エポックをウォームアップ期間とした．ウォームアップ後は，残りの 40 エポックで学習率をコサインカーブに従い最小 0.0001 まで減衰させた．損出関数はクロスエントロピー誤差関数を用いた．事前学習済みの ViT は，timm ライブラリの”vit_base_patch16_224”を用いた．本実験は，$N = 196$ 及び $L = 768$ の下で行われた．

4.2 暗号化画像の視覚情報保護効果

暗号化画像の視覚保護効果を，ランダム置換行列の制限条件の観点から評価する．図 4(a) は原画像であり，(b) および (c) は従来法 [20] によって暗号化された例である．一方，(d) は提案法を用いて暗号化した場合の例である．従来法では，置換の制限を厳しくするほど視覚情報保護が弱まり，対象物の詳細を非開示にすることができなくなる．一方，提案法では制限条件によって視覚情報が弱まることはない．同図 (b) と (d) の比較から，提案法は制限した場合にも，従来法の制限なしの状態と同程度の視覚情報保護効果を実現していることが確認できる．

以上のように，従来法 [20] では，制限によって視覚情報効果が低下するという課題がある．一方，提案法では，その課題を解決することができる．さらに，制限条件は暗号化鍵と同様に秘密とされるため，制限ランダム置換行列の鍵空間は，一般に制限なしランダム置換行列と同じ広さを持つことができる．



(a) Plain　(b) Conventional [20] ($N_{bs} = 0, N_{ps} = 0$)

(c) Conventional [20] ($N_{bs} = 128, N_{ps} = 512$)　(d) Proposed ($N_{bs} = 128, N_{ps} = 512$)

図 4: 暗号画像例

4.3 分類性能

表 1 に画像分類精度の実験結果を示す．事前学習済みモデルと画像の両方に対して暗号化処理を施さなかった場合（プライ

バシー保護なし）を，Baseline とした．また，従来法とは，事前学習済みモデルを暗号化せず，制限ありの置換行列を用いて画像のみ暗号化した場合の結果である [20].

制限なしの置換行列（$N_{bs} = N_{ps} = 0$）を用いて事前学習モデルと画像を暗号化した場合，提案法は従来法と同様に Baseline とほぼ同等の精度を達成できた．他の制約条件の下においても，従来法と提案法の精度はほぼ同等である．さらに，提案法では，制限の大小に関わらず画像の視覚情報を強固に保護することが可能である．また，敵対的サンプルへの応用において，制限が大きいほど敵対的サンプルに含まれるノイズの影響を強く回避できることが期待される．

表 1: モデルの分類性能

Encryption	Accuracy	
	Conventional [20]	Proposed
$N_{bs} = 0$, $N_{ps} = 0$	**0.9901**	**0.9909**
$N_{bs} = 0$, $N_{ps} = 768$	**0.9871**	**0.9876**
$N_{bs} = 49$, $N_{ps} = 192$	**0.9864**	**0.9875**
$N_{bs} = 98$, $N_{ps} = 384$	**0.9808**	**0.9822**
$N_{bs} = 196$, $N_{ps} = 0$	**0.9793**	**0.9814**
$N_{bs} = 196$, $N_{ps} = 768$	**0.8305**	**0.8456**
Plain (Baseline)	**0.9904**	

5. まとめ

本稿では，事前学習された ViT を暗号化し，暗号化した画像を用いてファインチューニングする新たな手法を提案した．提案手法では，ViT の持つ２つの埋め込み構造の特性に着目し，ランダム置換行列を用いてモデルと画像に暗号化を施す．その際，制限ランダム置換行列を導入することで，画像の視覚情報を強固に保護しつつ，画像暗号化がモデル性能に与える影響を低減することを目指した．画像分類タスクの実験において，提案法の有効性をモデル性能および画像の視覚情報保護の観点から評価した

謝　辞

本研究は，JSPS 科研費 (JP21H01327) の助成を受けたものである．

文　献

[1] H. Kiya, A. P. M. Maung, Y. Kinoshita, S. Imaizumi and S. Shiota: “An overview of compressible and learnable image transformation with secret key and its applications”, APSIPA Transactions on Signal and Information Processing, **11**, 1, pp. – (2022).

[2] T. Chuman, W. Sirichotedumrong and H. Kiya: “Encryption-then-compression systems using grayscale-based image encryption for jpeg images”, IEEE Transactions on Information Forensics and Security, **14**, 6, pp. 1515–1525 (2019).

[3] K. Wei, J. Li, C. Ma, M. Ding, W. Chen, J. Wu, M. Tao and H. V. Poor: “Personalized federated learning with differential privacy and convergence guarantee”, IEEE Transactions on Information Forensics and Security, **18**, pp. 4488–4503 (2023).

[4] M. Maung, A. Pyone and H. Kiya: “Encryption inspired adversarial defense for visual classification”, 2020 IEEE International Conference on Image Processing (ICIP), pp. 1681–1685 (2020).

[5] M. AprilPyone and H. Kiya: “Block-wise image transformation with secret key for adversarially robust defense”, IEEE Transactions on Information Forensics and Security, **16**, pp. 2709–2723 (2021).

[6] R. Iijima, S. Shiota and H. Kiya: “A random ensemble of encrypted vision transformers for adversarially robust defense”, IEEE Access, **12**, pp. 69206–69216 (2024).

[7] L. T. Phong, Y. Aono, T. Hayashi, L. Wang and S. Moriai: “Privacy-preserving deep learning via additively homomorphic encryption”, IEEE Transactions on Information Forensics and Security, **13**, 5, pp. 1333–1345 (2018).

[8] Y. Wang, J. Lin and Z. Wang: “An efficient convolution core architecture for privacy-preserving deep learning”, 2018 IEEE International Symposium on Circuits and Systems (ISCAS), pp. 1–5 (2018).

[9] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh and D. Bacon: “Federated learning: Strategies for improving communication efficiency”, arXiv (2016).

[10] M. AprilPyone and H. Kiya: “Privacy-preserving image classification using an isotropic network”, IEEE MultiMedia, **29**, 2, pp. 23–33 (2022).

[11] Z. Qi, M. AprilPyone, Y. Kinoshita and H. Kiya: “Privacy-preserving image classification using vision transformer”, EURASIP European Signal Processing Conference, Belgrade, Serbia, August 31, pp. 543–547 (2022).

[12] H. Kiya, R. Iijima, M. AprilPyone and Y. Kinoshita: “Image and model transformation with secret key for vision transformer”, IEICE Transactions on Information and Systems, **E106.D**, 1, pp. 2–11 (2023).

[13] H. Kiya, T. Nagamori, S. Imaizumi and S. Shiota: “Privacy-preserving semantic segmentation using vision transformer”, Journal of Imaging, **8**, 9 (2022).

[14] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby: “An image is worth 16x16 words: Transformers for image recognition at scale”, 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, OpenReview.net (2021).

[15] M. Tanaka: “Learnable image encryption”, 2018 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), pp. 1–2 (2018).

[16] K. Madono, M. Tanaka, O. Masaki and O. Tetsuji: “Block-wise scrambled image recognition using adaptation network”, Workshop on Artificial Intelligence of Things (AAAI WS) (2020).

[17] W. Sirichotedumrong, Y. Kinoshita and H. Kiya: “Pixel-based image encryption without key management for privacy-preserving deep neural networks”, IEEE Access, **7**, pp. 177844–177855 (2019).

[18] Z. Qi, A. MaungMaung and H. Kiya: “Privacy-preserving image classification using convmixer with adaptative permutation matrix and block-wise scrambled image encryption”, Journal of Imaging, **9**, 4 (2023).

[19] H. Ito, Y. Kinoshita, M. Aprilpyone and H. Kiya: “Image to perturbation: An image transformation network for generating visually protected images for privacy-preserving deep neural networks”, IEEE Access, **9**, pp. 64629–64638 (2021).

[20] K. Horio, K. Nishikawa and H. Kiya: “Privacy-preserving vision transformer using images encrypted with restricted random permutation matrices” (2024).

[21] T. Nagamori, S. Shiota and H. Kiya: “Efficient fine-tuning with domain adaptation for privacy-preserving vision transformer”, APSIPA Transactions on Signal and Information Processing, **13**, 1, p. e8 (2024).