EViT: Privacy-Preserving Image Retrieval via Encrypted Vision Transformer in Cloud Computing

QIHUA FENG, Jinan University, Guang Zhou, China

PEIYA LI*, Jinan University, Guang Zhou, China

ZHIXUN LU, Jinan University, Guang Zhou, China

CHAOZHUO LI, Beihang University, Beijing, China

ZEFANG WANG, ZHIQUAN LIU, Jinan University, Guang Zhou, China

CHUNHUI DUAN, Beijing Institute of Technology, Beijing, China

FEIRAN HUANG, Jinan University, Guang Zhou, China

Image retrieval systems help users to browse and search among extensive images in real-time. With the rise of cloud computing, retrieval tasks are usually outsourced to cloud servers. However, the cloud scenario brings a daunting challenge of privacy protection as cloud servers cannot be fully trusted. To this end, image-encryption-based privacy-preserving image retrieval schemes have been developed, which first extract features from cipher-images, and then build retrieval models based on these features. Yet, most existing approaches extract shallow features and design trivial retrieval models, resulting in insufficient expressiveness for the cipher-images. In this paper, we propose a novel paradigm named Encrypted Vision Transformer (EViT), which advances the discriminative representations capability of cipher-images. First, in order to capture comprehensive ruled information, we extract multi-level local length sequence and global Huffman-code frequency features from the cipher-images which are encrypted by stream cipher during JPEG compression process. Second, we design the Vision Transformer-based retrieval model to couple with the multi-level features, and propose two adaptive data augmentation methods to improve representation power of the retrieval model. Our proposal can be easily adapted to unsupervised and supervised settings via self-supervised contrastive learning manner. Extensive experiments reveal that EViT achieves both excellent encryption and retrieval performance, outperforming current schemes in terms of retrieval accuracy by large margins while protecting image privacy effectively. Code is publicly available at https://github.com/onlinehuazai/EViT.

CCS Concepts: • Information systems \rightarrow Image search; • Security and privacy \rightarrow Domain-specific security and privacy architectures; • Computing methodologies \rightarrow Image representations.

Additional Key Words and Phrases: Image retrieval, privacy-preserving, deep learning, JPEG, Vision Transformer, self-supervised learning.

Authors' addresses: Qihua Feng, fengqh2020@gmail.com, Jinan University, Guang Zhou, China, 510000; Peiya Li, lpy0303@jnu.edu.cn, Jinan University, Guang Zhou, China, 510000; Chaozhuo Li, Beihang University, Beijing, China, 100000; Zefang Wang, Zhiquan Liu, Jinan University, Guang Zhou, China, 510000; Chunhui Duan, Beijing Institute of Technology, Beijing, China, 100000; Feiran Huang, Jinan University, Guang Zhou, China, 510000.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Association for Computing Machinery.

Manuscript submitted to ACM

^{*}Corresponding author.

ACM Reference Format:

1 INTRODUCTION

Image retrieval is an important research field in the computer vision community, which draws increasing attention due to its significant research impacts and tremendous practical values [11, 29, 52]. Given a query image, the objective of image retrieval is to search similar images in an extensive image database. With the rise of cloud computing, traditional local storage mode has been changed to cloud storage, which satisfies user demand for storing massive image data on the server and enables users to access the data from any location at any time. Although cloud computing contributes to alleviating the challenge of limited local storage space and provides great convenience to users, the images are at danger of privacy leakage since the cloud server cannot be fully trusted and is vulnerable to hacking [89]. A typical strategy is to encrypt the images prior to uploading them to the cloud server, but conventional image encryption algorithms may hinder the subsequent encrypted image data retrieval operation [46, 100]. Therefore, it is urgent to develop privacy-preserving image retrieval technology that can provide both privacy protection and accurate retrieval simultaneously.

Schemes	Image er	ncryption & Featu	re extraction	Retrieval model			
	Visual security	Adaptive encryption key	Multi-level features	Unsupervised	Supervised	Deep learning	
Zhang [100]	-	✓	×	✓	×	×	
Liang [47]	1	×	×	√	×	×	
Xia [88]	1	×	×	√	×	×	
Li [46]	1	✓	×	√	×	×	
Xia [89]	1	×	×	√	×	×	
Xia [87]	1	×	×	√	×	×	
Chen [14]	1	✓	×	×	✓	×	
Feng [26]	1	×	×	×	✓	✓	
Ours							

Table 1. Our scheme and the current schemes

Existing privacy-preserving image retrieval schemes can be roughly classified into two categories: feature-encryption-based and image-encryption-based schemes [89]. In the first category, the content owner first extracts features (e.g., scale-invariant feature transform) from plain-images, then separately encrypts images and features [37, 54, 90, 91]. In order to match cipher-images, the encrypted features of similar images should keep distance similarity [86], and both encrypted images and features are uploaded to the server. This type of schemes can effectively protect privacy with standard cryptographic technologies, however, it performs image encryption and feature extraction/encryption Manuscript submitted to ACM

¹ "✓" indicates that corresponding requirements are supported; otherwise, 'X" is used.

² "↑" indicates better visual security than Zhang [100] (mentioned in Section 5).

separately, which may incur additional computational workload and inconvenience for the owner and users [46, 89, 100]. To address it, image-encryption-based schemes were proposed by extracting features from encrypted images directly [13–15, 46, 47, 87–89, 100]. In such schemes, the content owner only needs to encrypt images, and then uploads encrypted images to the server. The task of extracting features from cipher-images can be outsourced to the server, which can decrease computational workload for owner and users. There are mainly three modules in the system of the second type of schemes, namely image encryption algorithm, feature extraction method and retrieval model. The mentioned three modules are inherently correlated to each other. The image encryption algorithm is the foundation, which retains effective ruled features in cipher-images and provides desirable image security. Feature extraction method is the bridge, which offers comprehensive inputs for retrieval model and determines whether adaptive encryption key [33] is supported or not (the same image can be encrypted with different secret keys and the extracted features are unchanged before and after encryption). The focus of retrieval model is to achieve excellent retrieval accuracy, which needs to couple with the extracted features and learn discriminative representations for the cipher-images.

Our work follows the idea of image-encryption-based privacy-preserving image retrieval, which can reduce additional computational workload. Zhang et al. [100] proposed the first image-encryption-based privacy-preserving scheme, while they had poor visual security with simple permutation encryption [47]. In order to achieve better visual security, state-of-the-art schemes [14, 26, 46, 47, 87-89] introduced sophisticated encryption operations (e.g., value replacement and stream cipher) to encrypt images, but still have some deficiencies in feature extraction methods and retrieval models. For example, [14, 46, 47, 87–89, 100] just extract shallow features (e.g., histogram features) from cipher-images, which are unable to express enough information. Furthermore, [47, 87-89] fail to support the adaptive encryption key [33] because the feature spaces are randomly changed with different secret keys. Feng et al. [26] divided images into 8 × 8 blocks and proposed Vision Transformer [24] (ViT) based supervised retrieval model, which achieves better retrieval performance than convolutional neural network (CNN), but it is painstaking to assign labels to images. Schemes [46, 47, 87-89, 100] utilized trivial models (e.g., K-means [56] and Bag-of-Words (BOW) [69]) to build unsupervised retrieval models, but it is difficult for these trivial models to learn non-linear embedding of complex image datasets [93], while deep neural network (DNN) is skilled in it. To capture more plentiful information of cipher-images, our feature extraction method designs multi-level features from 8 × 8 blocks and global Huffman-code of cipher-images. We utilize the length of variable-length integer (VLI) code unchanged with stream cipher to support the adaptive encryption key. Self-supervised learning is the typical unsupervised framework [10, 18, 27, 34, 38, 72], and ViT [24] divides an image into non-overlapping blocks to learn global dependency relations with self-attention mechanism [76]. Hence, inspired by Feng [26], and ViT can couple with our multi-level features from 8 × 8 blocks of cipher-images, we build ViT-based unsupervised retrieval model in self-supervised learning manner.

In this paper, we propose a novel privacy-preserving image retrieval scheme named Encrypted Vision Transformer (EViT). First, images are encrypted during JPEG compression process, in which the VLI code of Discrete Cosine Transform (DCT) coefficient is encrypted by stream cipher. Second, EViT extracts well-designed multi-level features from cipher-images: the length sequence of DCT coefficients' VLI code in each 8 × 8 block (local features) and the global Huffman-code frequency features, which can express more plentiful information of cipher-images. The VLI code encryption with stream cipher has an inherent advantage: the length of VLI code before and after encryption are unchanged, so multi-level features remain invariable whichever secret keys are used, namely EViT supports the adaptive encryption key. Finally, EViT adopts self-supervised contrastive learning [10, 12, 34] manner to build the unsupervised retrieval model, which can reduce label overhead. In order to learn discriminative representations of the cipher-images, EViT proposes modified Vision Transformer-based retrieval model, and replaces original *Cls_Token* [24] of ViT with Manuscript submitted to ACM

learnable global Huffman-code frequency feature. Conventional data augmentations in the plain-image retrieval domain (e.g., random crop) will entirely change the multi-level features of cipher-images (Section 4.3.3), so EViT directly takes two adaptive data augmentations for multi-level features to improve representation ability of retrieval model. EViT can also achieve the supervised retrieval model by easily Fine-Tuning on the unsupervised model.

As shown in Tab. 1, EViT can satisfy all requirements (e.g., adaptive encryption key, multi-level features, deep learning) simultaneously, and our experimental results show that our EViT can improve retrieval performance by large margins. The main contributions are summarized as follows:

- 1) Ingenious multi-level features, local length sequence and global Huffman-code frequency, are extracted from cipher-images to express more abundant features of cipher-images. Our feature extraction method also enables EViT to use the adaptive encryption key since the length of VLI code is unaffected by stream cipher.
- 2) To the best of our knowledge, EViT is the first to propose self-supervised learning for privacy-preserving image retrieval. In order to enhance model's representation ability, EViT proposes two adaptive data augmentations for retrieval model, and uses learnable global Huffman-Code frequency to improve existing ViT.
- 3) Experimental results demonstrate that our EViT outperforms other state-of-the-art schemes in retrieval performance for both unsupervised and supervised learning models. EViT can not only protect image privacy but also complete image retrieval efficiently.

The rest of our paper is organized as follows. Section 2 presents the related work for privacy-preserving image retrieval. Preliminaries are introduced in Section 3. Section 4 gives our proposed scheme. Section 5 presents the experimental results. Finally, Section 6 summaries the conclusion.

2 RELATED WORK

In recent years, researchers have paid more attention on privacy-preserving image retrieval [50, 53, 54, 87, 90, 95], and applied these techniques to boost the performance of real-life applications [61, 81, 82]. The current privacy-preserving image retrieval schemes mainly can be divided into two categories [89]: feature-encryption-based and image-encryption-based schemes.

Feature-encryption-based schemes: In this type of works, content owner first extracts features from plain-images, then encrypts these features and images separately. It's noted that the features of plain-images also need to be protected because the features always contain sensitive information of plain-images [7, 86]. Lu et al. [54] proposed the first privacy-preserving image retrieval scheme, and they improved retrieval speed based on safe search indexes in [53]. Wai et al. [84] developed a new asymmetric scalar-product-preserving encryption (ASPE) that preserved a special type of scalar product which could provide k-nearest neighbor (kNN) computation on encrypted features. Xia et al. [90] utilized a stable KNN algorithm to protect image feature vectors and designed a watermark-based protocol to prevent the authorized query users from illegally distributing the retrieved images. Zhang et al. [99] used fully homomorphic encryption to encrypt the features and proposed to boost privacy-preserving search by distributed and parallel computation. Xia et al. [92] proposed a scheme with extracting features by the Scale-invariant feature transform (SIFT) feature and BOW model. They calculated the distance among image features through the Earth Mover's Distance (EMD) and adopted a linear transformation on EMD to protect parameter information. Cheng et al. [98] used hash codes to encrypt features which were learned by deep neural networks (DNN), and they utilized S-Tree to increase the search efficiency. Feature-encryption-based schemes can achieve great security of images, but it performs image encryption Manuscript submitted to ACM

and feature extraction/encryption separately, resulting in an additional computational workload and inconvenience for the content owner and users [46, 47].

Image-encryption-based schemes: This type of work extracts features directly from cipher-images. Content owner only needs to encrypt images before uploading cipher-images to the server, the task of extracting features from cipher-images can be outsourced to the server, which solves the problem of separation of image encryption and feature extraction/encryption. Zhang et al. [100] proposed a scheme with encryption of JPEG images by permuting DCT coefficients and extracting features from these coefficients' histogram invariance. Cheng et al. [13] proposed to encrypt the DC coefficients by using stream cipher, encrypt the AC coefficients by using scrambling encryption, and conduct retrieval based on the histogram of the AC coefficients. However, Cheng et al. [13] could not ensure JPEG format compliance. Xia et al. [88] encrypted DC coefficients by stream cipher on the Y component and encrypted U and V components by value replacement and permutation encryption, then extracted AC histograms features of Y component and color histograms features of U and V components. Liang et al. [47] extracted Huffman-code histograms from cipher-images which were encrypted by stream cipher and permutation encryption. The work [87] encrypted images by color value substitution and permutation encryption, and extracted local color histogram features from cipher-images, then they built unsupervised bag-of-encrypted-words (BOEW) model to achieve retrieval. Li et al. [46] proposed a new block transform encryption method using orthogonal transforms rather than 8 × 8 DCT. [89] extracted secure Local Binary Pattern (LBP) features from cipher-images, then built BOW model to conduct retrieval. These above schemes built unsupervised retrieval model, but they fail to use deep learning to learn non-linear embedding of cipher-images, and just extracted shallow features from cipher-images. Cheng et al. [14] used stream cipher and permutation encryption to encrypt JPEG images, and extracted features from cipher-images by Markov process, then built supervised support vector machine (SVM) model to conduct retrieval. But it is painstaking to assign labels to images, and SVM is a linear model which is limited to learn discriminative representations of cipher-images.

Our EViT follows the idea of image-encryption-based privacy-preserving image retrieval which can decrease additional computational workload. EViT extracts multi-level features from cipher-images and introduces DNN based unsupervised retrieval model in a self-supervised manner, which can significantly improve retrieval performance.

3 PRELIMINARIES

3.1 JPEG Compression

We encrypt images during JPEG compression process like some privacy-preserving image retrieval schemes [13, 46, 47, 86, 88, 100]. Here, we briefly introduce the JPEG compression process, whose overview is shown in Fig. 1. According to the JPEG compression standard [17, 63], images are converted to YUV color space. After 8×8 DCT and quantization, each block has a total of 64 coefficients, of which the first coefficient is the direct current (DC) coefficient, and the remaining 63 coefficients are the alternating current (AC) coefficients. The DC coefficient is encoded by differential pulse code modulation (DPCM), and the remaining 63 AC coefficients in the same block are converted into a sequence using zig-zag scanning. AC coefficients are encoded with run-length encoding (RLE), which are converted to the (r, v) pairs [17, 63].

The lossless Huffman variable-length entropy coding technology is used to further compress the DC differential (ΔDC) and AC coefficients (r, v) pairs, and all coefficients are coded to binary sequence. Specifically, each ΔDC is encoded into two parts: DC Huffman code (DCH) and DC VLI code (DCV); each (r, v) pairs is encoded as two parts: AC Huffman code (ACH) and AC VLI code (ACV). As shown in Fig. 2, we take an example for lossless Huffman

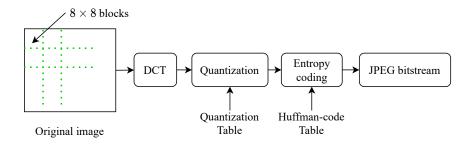


Fig. 1. Overview of JPEG compression process.

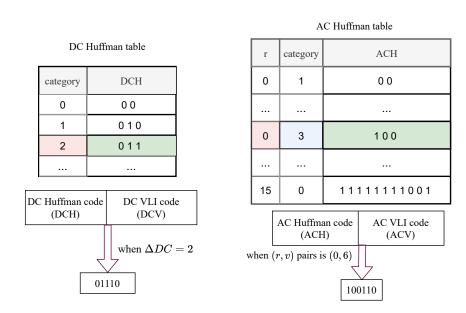


Fig. 2. An example of lossless Huffman variable-length entropy coding.

variable-length entropy coding, where category is the range of amplitudes of coefficients [63], and r corresponds to the run value in the AC Human table [63]. For example, when $\Delta DC = 2$, it's DCH = 011 and DCV = 10, so it is coded into 01110; when (r, v) = (0, 6), it's ACH = 100 and ACV = 110, therefore it is coded into 100110. Each coefficient can be coded into binary sequence through Huffman-code table mapping, more details please refer to [17, 63].

3.2 Unsupervised Contrastive Learning

Supervised learning has been widely used in many image retrieval tasks [19, 36, 49, 55, 96, 97], but it is painstaking to assign labels to images. In order to decrease the overhead with human annotations, unsupervised algorithms are explored by researchers, such as K-means [56] and BOW [69] model. Both K-means and BOW utilize the idea of clustering, but in many image databases, it is ineffective to learn images' representations because clustering with euclidean distance cannot learn non-linear embedding [93]. Deep neural networks (DNN) is competent in learning non-linear embedding which are indispensable for complex image databases. Supervised-learning based on DNN generally uses target labels Manuscript submitted to ACM

to train model, but it is painstaking to assign labels to images. Self-supervised learning can build DNN model without target labels, which generates virtually unlimited labels from existing images and uses those to learn the representations. In recent years, many self-supervised learning methods have been proposed [8, 10, 30, 34, 58, 60, 67, 72, 85]. Due to its simple and effective property, self-supervised contrastive learning [10, 34, 72, 85] has been used in many unsupervised-learning computer vision tasks [3, 18, 22, 27, 38, 94]. SimCLR [10] is a popular self-supervised contrastive learning method, whose process is roughly illustrated as follows: given an image x, we can obtain different images x_i and x_j by stochastic data augmentations such as random color distortions and random Gaussian blur; then the two images through same encoder can generate corresponding embeddings z_i and z_j , and SimCLR can learn representations of images by maximizing the similarity z_i and z_j . SimCLR obtains the positive samples of x by data augmentations, the other images are negative samples. But SimCLR [10] needs a very large batch size to build negative samples when training, this is expensive for most researchers to use GPU with large memory. Therefore MoCo [12, 34] proposed momentum contrast to solve the problem of large batch size by building a dynamic dictionary with a queue and momentum updating. In this paper, we utilize self-supervised contrastive learning based on MoCo to build our unsupervised-learning retrieval model.

3.3 Vision Transformer

Since the proposal of Google's Transformer [76] in 2017, it has almost dominated natural language processing tasks [23, 25, 66]. Transformer is a sequence model based on self-attention mechanism, which can capture the global dependencies between words. In 2021, Google proposed to apply Transformer in computer vision [24], called Vision Transformer (ViT). Then many researchers explore ViT and find ViT can obtain excellent performance in many computer vision tasks [16, 41, 44, 48, 64, 71], even surpass the CNN model in performance [4, 31, 51, 78, 94]. Vision Transformer not only makes breakthroughs in computer vision but also forms the unified model for computer vision and natural language processing.

ViT [24] divides images into non-overlapping blocks, and each block is equal to a word in natural language processing. Suppose image's size is $H \times W$, the size of each block is $P \times P$, hence the number of blocks is $\frac{H \times W}{P^2}$. Assuming that the dimension of word embedding is D, ViT uses linear projection [24] to map each block to dimension D, called block embedding. Like the [class]token in BERT [23], ViT prepends a learnable embedding Cls_Token in block embeddings. The goal of Cls_Token is to learn the representation of image, and ViT initializes Cls_Token with ones (dimension D). The results of adding the block and position embeddings [24, 76] are then used as the input to the Transformer encoder. The Transformer encoder of ViT is similar with standard Transformer [76], includes self-attention [76], layer normalization [6] and residual module [35]. Self-attention is an import part in the Transformer encoder. When calculating self-attention, three matrices are needed: query (Q), key (K), value (V). Suppose the input of self-attention is X, and the definition of self-attention is expressed as:

$$Q = W_O X, \quad K = W_K X, \quad V = W_V X \tag{1}$$

$$Attention(Q, K, V) = softmax(\frac{QK^{T}}{\sqrt{d_k}})V$$
 (2)

where W_Q , W_K , W_V are linear projection matrices, and d_k is the dimension of K. In order to capture more information, Transformer uses multi-head self-attention to learn different subspaces [76]. Multi-head self-attention uses h different linear projections to map Q, K, V, and then concatenates different results of self-attention and does a linear projection.

The definition of multi-head self-attention is as follows:

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O$$
 (3)

$$head_i = Attention(Q_i, K_i, V_i), \quad i \in [1, h]$$
 (4)

where W^O is linear projection matrices.

In this paper, we use ViT as backbone in our retrieval model, and the reasons are as follows: a) ViT is popular and makes excellent performance in recent many computer vision tasks [4, 31, 41, 44, 48, 51, 64, 71, 78, 94]. b) Our feature extraction method couples with the inputs of ViT. We extract features from cipher-images' 8×8 blocks, and ViT divides image into non-overlapping blocks. So the features of each 8×8 block are equal to word embedding of ViT. c) ViT can learn global dependency relations with self-attention mechanism [76] compared with local CNN [59]. Previous privacy-preserving image retrieval scheme [26] also used ViT as backbone and proved that ViT was more fit to encrypted images than CNN in their experiments.

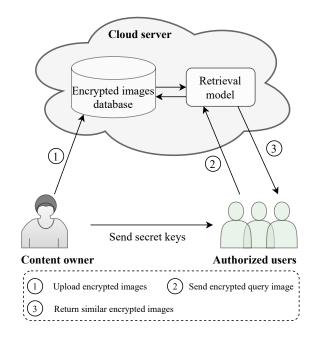


Fig. 3. System model of image-encryption-based privacy-preserving image retrieval.

4 PROPOSED SCHEME

Generally, the system of image-encryption-based privacy-preserving image retrieval includes three parts: content owner, server, and authorized users. As shown in Fig. 3, the content owner first encrypts images and uploads encrypted images to the server. Then authorized users encrypt the query image with the same encryption algorithm and upload it to the server. The server extracts features from the query image, and the retrieval model searches similar cipher-images in encrypted image database according to the features. Finally, these similar cipher-images are returned to authorized users, and authorized users decrypt images with encryption keys which are shared from the content owner through Manuscript submitted to ACM

a secure channel [14, 47]. In this system, three modules need to be designed: image encryption algorithm, feature extraction method, and image retrieval model. We specify these modules of EViT in detail below.

4.1 Image Encryption

Images are encrypted during JPEG compression process which is explained briefly in Section 3. Specifically, in the stage of entropy coding, we take stream exclusive-or operator for VLI code of DC and AC coefficients [65]. The corresponding encryption keys are k_{DC} and k_{AC} , and we encrypt DCV and ACV as follows:

$$DCV' \leftarrow DCV \oplus k_{DC}$$
 (5)

$$ACV' \leftarrow ACV \oplus k_{AC}$$
 (6)

where DCV' and ACV' are encrypted VLI codes, \oplus is exclusive-or operator. Our encryption keys are generated by the adaptive encryption key generation method [33] which uses image as input of hash function BLAKE2 [5], therefore different images have different encryption keys. During encryption process, original image I can be compressed to encrypted JPEG bitstream, and different color spaces of I have different secret keys. Stream exclusive-or with VLI code theoretically does not increase storage memory of cipher-images [17]. In Algorithm 1, the encryption algorithm of our proposed EVIT is presented. The encrypted bitstream is decodable because the file structure and Huffman-code are unchanged [65].

```
Algorithm 1: Encryption algorithm
```

```
input : I, k_{DC}, k_{AC}
output : Encrypted JPEG bitstream

1 Convert I from RGB to YUV color space;

2 Denote the width and height of the image I as W and H;

3 for I_i in I, i \in Y, U, V do

4 Divide I_i into several 8 \times 8 non-overlapping blocks B_j^i, j \in [1, \dots, blknum], where blknum = \frac{W \times H}{8 \times 8};

5 Conduct DCT, quantization and entropy coding;

6 for j = 1 to blknum do

7 Encrypt the VLI code of B_j^i;

8 DCV_{B_j^i}^i \leftarrow DCV_{B_j^i}^i \oplus k_{DC}^{B_j^i};

9 ACV_{B_j^i}^i \leftarrow ACV_{B_j^i}^i \oplus k_{AC}^{B_j^i};

10 end

11 end
```

4.2 Feature Extraction

Image-encryption-based schemes extract features directly from cipher-images. Existing schemes just extract shallow features (e.g. DCT histogram), which are unable to express plentiful information of cipher-images. EViT extracts multi-level features from our cipher-images: local length sequence and global Huffman-code frequency features.

Local length sequence features: EViT extracts each 8×8 block's local features. As shown in Fig. 4, EViT calculates the corresponding VLI code's length of DCT coefficients and builds length sequence features in each 8×8 block. For example, when $\Delta DC = 5$, it's VLI code is '101', hence the length is 3. The length sequence features are generated by zig-zag scanning [63]. If the coefficient is zero, then EViT denotes its length as zero. Because each block has three Manuscript submitted to ACM

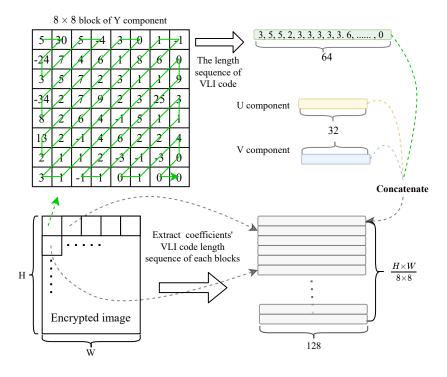


Fig. 4. Sketch of extracting local features from our encrypted images.

components (Y/U/V), EViT concatenates three length sequence features of different components. It's noted that when extracting length sequence features for U and V components, EViT just chooses top 32 coefficients in length sequence features due to there are many zeros of U and V components in the back coefficients [17].

Global Huffman-code frequency features: EViT extracts global Huffman-code frequency features from the cipherimages. We take a simple example to describe Huffman-code frequency features, as shown in Fig. 2. If the second row of DC Huffman table is used 10 times during entropy coding stage for an encrypted image, the corresponding Huffman-code frequency feature is denoted as 10. The rows of DC Huffman table is 12, and the rows of AC Huffman table is 162 [63]. Therefore, the dimension of Huffman-code frequency features is $(12+162)\times 3=522$, where 3 represents three components (YUV).

Global features and local features are extracted from cipher-images, which are employed to train our retrieval model. Here, different encryption keys can be utilized for different images since the stream exclusive-or operator does not change the length of VLI code. For example, suppose the DCV is '101', and the different encryption keys are '100' and '011', then the DCV' are '001' and '110' respectively. Hence the length of binary VLI code remains unchanged no matter which encryption key is used, namely the features that extracted from cipher-images are unchanged. This also means that we can use different encryption keys for different images, and the authorized user can encrypt query image with the same encryption algorithm but different keys.

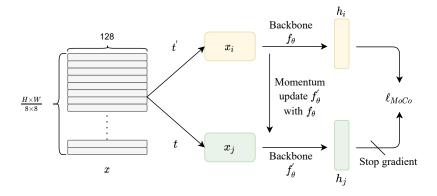


Fig. 5. Overview of our unsupervised-learning retrieval model.

4.3 Unsupervised-learning Retrieval Model

After extracting features from encrypted images, EViT uses these features to train retrieval models. EViT respectively proposes the unsupervised-learning and supervised-learning retrieval models based on deep learning. Deep image retrieval model [11] is a typical deep metric learning [40] whose aim is to learn the representations of images. Given an image, we first use learnable deep neural networks $f(\cdot)$ to learn it's representation h, and $f(\cdot)$ is generally called as backbone. Then the images' representations are used to calculate the similarity such as cosine or euclidean distances between the query and the targets.

4.3.1 Loss. The unsupervised framework uses MoCo (see Section 3) due to its simple and effective property [10, 34, 72], which does not use target labels to learn the representations of cipher-images. As shown in Fig. 5, given a cipher-image, EViT extracts features from it and denotes these features as x. Using random data augmentations t' and t, we can obtain x_i and x_j respectively. Through backbone f_θ , EViT can learn the representation h_i of x_i . For x_j , the process is like x_i , but the backbone f'_θ is stop-gradient which is updated by momentum with f_θ [34]. The process of forward propagation can be defined as:

$$x_i = t'(x), \quad x_j = t(x) \tag{7}$$

$$h_i = f_{\theta}(x_i), \quad h_j = f'_{\theta}(x_j). \tag{8}$$

The backbone $f_{\theta}^{'}$ is updated with momentum manner [34] which is described as:

$$f_{\theta}^{'} = mf_{\theta}^{'} + (1 - m)f_{\theta} \tag{9}$$

where m is momentum factor and is set to be 0.99 following MoCo [34]. The structures of f'_{θ} and f_{θ} are same, but with different parameters.

The loss function is like MoCo which is called InfoNCE [74]. MoCo proposed momentum contrast to solve the problem of large batch size by building a dynamic dictionary with a queue and momentum updating. The dynamic dictionary is a queue where the current batch enqueued and the oldest batch dequeued. For one sample x_i in the current batch, the x_j is positive sample, and other samples in the current batch and the queue are negative samples. The loss can be defined as:

$$\ell = -log \frac{exp(h_i \cdot h_j/\tau)}{exp(h_i \cdot h_j/\tau) + \sum_{k^-} exp(h_i \cdot h_{k^-}/\tau)}$$
(10)

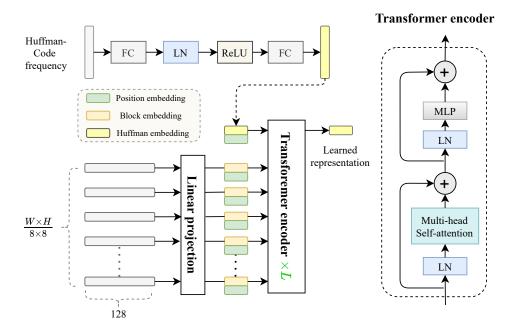


Fig. 6. Overview of our backbone.

where τ is a temperature hyper-parameter proposed by [85], which we set to be 0.1 like MoCo. h_i and h_j are representations of x_i and x_j respectively (Fig. 5). h_{k^-} are representations of negative samples, and "·" is dot product.

4.3.2 Backbone. Our backbone f_{θ} adopts the structure of ViT (see Section 3). EViT extracts two parts features from encrypted images, local length sequence features and global Huffman-code frequency features. As shown in Fig. 6, suppose the number of blocks of a cipher-image is $\frac{H \times W}{8 \times 8}$ (H is height, W is width). These local features, through linear projection [24], produce corresponding block embeddings. Original Cls_Token (mentioned in Section 3) of ViT are all ones for each image, which fails to express specific information for different images. Hence, different from standard ViT [24], EViT uses Huffman embedding to replace Cls_Token , which is helpful for retrieval performance in experiments. The Huffman embedding (He) is learned from global Huffman-code frequency features (gHff), which can be defined as:

$$He = FC(\sigma(LN(FC(gHff))))$$
(11)

where FC is fully-connected layer, LN is layer normalization [6], σ is activation function ReLU [2]. In order to keep position information, we also add position embedding [24] with block embedding and Huffman embedding. The result of these embeddings is denoted as v_0 , then through L stacked Transformer encoder [24], EViT can learn the representations v_0^1 of cipher-image. The l-th Transformer encoder can be defined as:

$$v'_{l} = MSA(LN(v_{l-1})) + v_{l-1}, \quad l = 1, 2, ... L$$
 (12)

$$v_l = MLP(LN(v_l)) + v_l, \quad l = 1, 2, \dots L$$
 (13)

where MSA is multi-head self-attention [76] (Eq. 3), MLP is multi-layer perceptron block [24]. The output of L stacked is v_L , and the representation is the first embedding v_L^0 .

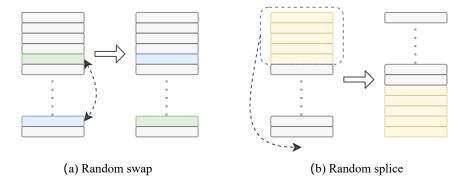


Fig. 7. Examples for our data augmentations with length sequence features.

ViT [24] uses spatial pixels to build a block of plain-images, but spatial pixels are randomly encrypted in cipher-images. EViT uses length sequence feature to replace spatial pixels in a block. Different from plain-image retrieval which can use pre-trained ViT on imagenet [42] as model's backbone, our task is privacy-preserving image retrieval and there is no pre-trained model as our backbone, so the retrieval model needs to be trained from scratch. Generally, small learning rate and warm up [35] are necessary for training a new model with the structure of ViT [73]. EViT uses cosine warm up [83] with learning rate. Our experimental results also present that cosine learning rate with warm up is helpful for retrieval performance.

4.3.3 Data augmentation. EViT uses random data augmentations t and t' for encrypted features in our unsupervised-learning retrieval model (Fig. 5). Common plain-image data augmentations, such as random crop, are not suitable for our task. For example, the plain-image is a dog, we random crop an image from plain-image. The cropped image is equal to the plain image in which the spatial structure is the dog, and just the cropped image may not have the dog's tail. But in the encrypted images, the cropped image is not a dog after decryption. Because our encryption algorithm is conducted in JPEG compression process, each adjacent 8×8 block's DCT coefficients of encrypted images are correlational. If some blocks are missing, the DCT coefficients of encrypted cropped image are all changed during decryption process. Once DCT coefficients, in particular DC coefficients, are changed, the spatial pixels and structures of decrypted image will be changed correspondingly. Other plain-image data augmentations are also unsuitable for our task, because these augmentations all change DCT coefficients in augmented image. Thus EViT directly conducts data augmentations with our length sequence features extracted from cipher-images.

In our retrieval model, we propose two kinds of adaptive data augmentations: random swap and splice for length sequence features. As shown in Fig. 7, we take examples to describe these two kinds of data augmentations. Our data augmentations are motivated by two aspects: a) the length sequence features are like different words, swapping two words is a typical augmentation in nature language process [79]; b) ViT also can achieve well performance with simple block permutation [59], and hence EViT can shuffle the length sequence features with random splice. In addition, model with random dropout [70] also plays a role in data augmentation [28]. Because an image through model with random dropout twice can obtain different embeddings in the training stage [28]. ViT [24] itself has a random dropout function, and EViT sets the dropout in our backbone as ViT [24]. Our experimental results shows that the two adaptive data augmentations are vital for model, which can significantly improve retrieval performance.

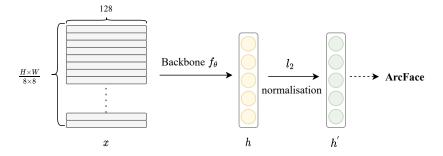


Fig. 8. Overview of supervised-learning retrieval model.

4.4 Supervised-learning Retrieval Model

Our EViT also provides a simple supervised-learning retrieval model, which uses the same structures of backbone like the unsupervised-learning model. As shown in Fig. 8, the supervised model can obtain representations h of cipher-images after backbone f_{θ} (Fig. 6). Here we can use the pre-trained backbone from unsupervised-learning model. The supervised loss function ArcFace [20] is used to train our model, which is defined as:

$$h = f_{\theta}(x), \quad h' = l_{2}(h)$$
 (14)

where l_2 is short for l_2 normalisation [20, 77].

ArcFace [20] is a common deep metric learning loss function, which has been widely used in retrieval tasks [21, 32, 62]. Compared with Triplet loss [68], ArcFace is more easy and effective [20] which gets rid of the disadvantages such as hard sample mining and combinatorial explosion in the number of triplets. ArcFace adds an additive angular margin within softmax loss [20], which can be defined as:

$$\ell_{ArcFace} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{e^{s(\cos(\theta y_i + \alpha))}}{e^{s(\cos(\theta y_i + \alpha))} + \sum_{j=1, j \neq y_i}^{n} e^{s\cos\theta_j}}$$

$$\tag{15}$$

where N and n are the batch size and the class number, and θ_j is the angle between the representation of i-th sample and j-th class center. y_i is ground-truth of i-th sample, and θ_{y_i} represents the angle between i-th sample and ground-truth class center. s and α are hyper-parameters which represent feature re-scale and angular margin parameters, respectively. More details please refer to [20].

In the inference stage, we just need to calculate cosine distances of the representations h of cipher-images, and then rank these distances and return top-K results. On the basis of unsupervised-learning model, our supervised-learning model seems straightforward, and we will further explore it in the future such as combining Triplet loss [68] and Center loss [80] to learn more discriminative representations. As shown in Algorithm 2, we present the main processes of EViT.

5 EXPERIMENTS AND ANALYSIS

In this section, experimental results of our proposed scheme are presented. We evaluate the performance on Corel10K dataset [43], which is the widely used dataset by many related researches. Corel10K dataset contains 10000 images in 100 categories, with 100 images in each category. The image sizes are 128×192 or 192×128 . Our programming language is Python. In the following section, we first describe the retrieval performance, then analyze time consumption of searching, finally present the encryption performance of our EViT.

Algorithm 2: EViT's main algorithm

```
input:plain-images, secret keys, labels=None

// first module: image encryption algorithm (Section 4.1);

cipher-images ← Image encryption (plain-images, secret keys);

// second module: feature extraction method (Section 4.2);

features ← feature extraction (cipher-images);

// third module: retrieval models;

unsupervised ← retrieval model(features) // Section 4.3;

if labels is Not None then

supervised ← Fine-Tuning (unsupervised, features, labels) // Section 4.4;

return supervised

return unsupervised

end
```

Table 2. Descriptions of Corel10K-a and Corel10K-b datasets.

Datasets	Corel10K-a	Corel10K-b
Traing set	7000	7000
Testing set	3000	3000
Classes of Training set	70	100
Classes of Testing set	30	100

5.1 Retrieval Performance

Our EViT respectively proposes the unsupervised-learning and supervised-learning retrieval models. We compare the retrieval performance of EViT with current image-encryption-based schemes whose retrieval models are divided into unsupervised and supervised model. In order to better compare retrieval performance, we split training set and testing set with two different types, and denote as Corel10K-a and Corel10K-b datasets which are described in Tab. 2. For Corel10K-a dataset, we train the retrieval model on 70 classes, so testing set has no same classes with training set (open-set classification [20]). For Corel10K-b, we train the retrieval model on 100 classes, and each class we select 70 images, and the remain 30 images of each class are in testing set (close-set classification [20]). We use stochastic gradient descent (SGD) as optimizer. The weight decay is $5e^{-5}$, and SGD momentum is 0.9. We set batch size to be 14 and 35 for the unsupervised-leaning and supervised-learning model respectively. We train the retrieval models using the PyTorch framework [1] on a machine with Nvidia RTX2080Ti 11G GPU.

The evaluation metric of retrieval performance we use is mean Average Precision [57] (mAP) which is widely used in many retrieval tasks. When returning top-*K* results, mAP is calculated as follows:

$$mAP@K = \frac{1}{Q} \sum_{q=1}^{Q} AP@K(q)$$
(16)

$$AP@K(q) = \frac{1}{R_q} \sum_{k=1}^{K} p_q(k) \ rel_q(k)$$
 (17)

Schemes		Corel10K-a	Corel10K-b	
	Xia[88]	0.378	0.230	
	Liang[47]	0.321	0.217	
	Xia[87]	0.383	0.235	
Unsupervised	Xia[89]	0.301	0.190	
1	Zhang [100]	0.396	0.269	
	Li[46]	0.410	0.269	
	Ours-unsupervised	0.466	0.295	
	Cheng[14]	_	0.407	
Supervised	Feng[26]	0.423	0.528	
	Ours-supervised	0.554	0.759	

Table 3. Comparison of retrieval performance with current schemes.

Table 4. Unsupervised-learning retrieval performance with different L values on Corel10K-a and Corel10K-b datasets.

L	4	5	6	7
Corel10K-a (mAP@100)	0.457	0.462	0.466	0.465
Corel10K-b (mAP@100)	0.289	0.291	0.295	0.292

where Q is the number of query images, R_q is the number of similar images for the query q, $p_q(k)$ is precision at rank k for the query q, and $rel_q(k)$ is 1 if the rank k result is similar to q, 0 otherwise. In this paper, we use mAP@100 to evaluate retrieval performance. The higher mAP@100, the better retrieval performance.

Here, we compare retrieval performance with current image-encryption-based schemes. All schemes are evaluated on the same testing set, and the results of comparison are shown in Tab. 3. We can see that our retrieval performance is better than other schemes. Specifically, our unsupervised-learning model can achieve 0.466 *mAP*@100, which is higher about 5.6% than state-of-the-art retrieval performance on the open-set Corel10K-a; on the closed-set Corel10K-b, our unsupervised-learning model can achieve 0.295 *mAP*@100, which is higher about 2.6% than state-of-the-art retrieval performance. For supervised-learning model, we can significantly improve retrieval performance than other supervised schemes. It is noted that Cheng [14] used classification probability as representations of cipher-images, it is unsuitable on open-set Corel10K-a due to testing set has no same classes with training set. So far the image-encryption-based supervised-learning schemes are few, our supervised-learning model can be a strong baseline to explore it. Next, we describe more experimental details about the unsupervised-leaning and supervised-learning retrieval performance respectively.

5.1.1 Unsupervised retrieval performance. Our backbone has L stacked Transformer encoders (Fig. 6), hence we use different L values to demonstrate the retrieval performances on Corel10K-a and Corel10K-b datasets. As shown in Tab. 4, we can see that L=6 is more suitable for our unsupervised-learning retrieval model.

We propose two adaptive data augmentations for EViT, and have mentioned that warm up and huffman embedding are helpful for our retrieval performance (Section 4.3.2). Here, we use ablation experiments to verify how data augmentations, Manuscript submitted to ACM

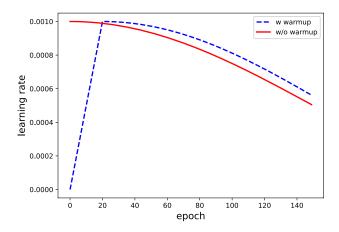


Fig. 9. Comparison of learning rate schedules.

Table 5. Ablation experiments with warm up and huffman embedding for unsupervised-learning retrieval model.

Data augmentations		✓	✓	✓
Warm up			✓	√
Huffman embedding				√
Corel10K-a (mAP@100)	0.397	0.411	0.423	0.466
Corel10K-b (mAP@100)	0.243	0.257	0.272	0.295

warm up, and huffman embedding influence the unsupervised-learning retrieval performance on Corel10K-a and Corel10K-b datasets. Our learning rate is $1e^{-3}$ with cosine warm up, as shown in Fig. 9, the "blue" line is cosine learning rate with warm up which is linearly increased to $1e^{-3}$ in the first 20 epoch; the "red" line is cosine learning rate without warm up. We add warm up and huffman embedding one by one (L=6), the results of ablation experiments are shown in Tab. 5. We can see that if we do not add data augmentations, warm up, and huffman embedding, the retrieval performance only can achieve 0.397 and 0.243 mAP@100 on Corel10K-a and Corel10K-b respectively. Our two adaptive data augmentations are helpful to enhance retrieval performance, which can improve about 1% mAP@100. Warm up improves retrieval performance with 1.2% and 1.5% on Corel10K-a and Corel10K-b respectively. Huffman embedding significantly improves retrieval performance with 4.3% and 2.3% on Corel10K-a and Corel10K-b respectively. The huffman embedding is learned from global Huffman-code frequency which is one of our multi-level features. The ablation experiments prove that multi-level features express more abundant information of cipher-images, which can directly improve retrieval performance.

5.1.2 Supervised retrieval performance. Our supervised model is Fine-Tuning on the unsupervised model. For example, when training the supervised-learning model on Corel10K-a dataset, our backbone can use unsupervised-learning model's parameters which are also trained on Corel10K-a as initial parameters. Unsupervised-learning model as pretrained model for supervised-learning is common [10, 27, 34] which can accelerate model convergence and achieve Manuscript submitted to ACM

α s	16	32	64	
0.1	0.539 / 0.690	0.554 / 0.658	0.540 / 0.588	
0.2	0.526 / 0.721	0.533 / 0.681	0.546 / 0.620	
0.3	0.529 / 0.746	0.501 / 0.725	0.534 / 0.453	
0.4	0.527 / 0.747	0.497 / 0.759	0.518 / 0.520	

Table 6. Retrieval performance with different α and s on Corel10K-a/b datasets.

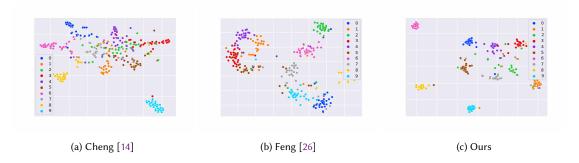


Fig. 10. Comparison of semantic space visualization with t-SNE on Corel10K-b dataset.

better performance. Due to the backbone of supervised-learning is same as unsupervised-learning model, most of strategies such as warm up and data augmentations also are used in supervised-learning model. Apart from loss function, the supervised-learning model is almost inspired by our unsupervised-learning model.

We mentioned that there are two hyper-parameters in ArcFace [20]: s and α (Eq. 15), and now we use different s and α to observe their influence on retrieval performance when L=6 on Corel10K-a and Corel10K-b (Corel10K-a/b) datasets. As shown in Tab. 6, we can see that on open-set dataset Corel10K-a, small α is more fit to supervised-learning model; while close-set dataset Corel10K-b, s should not set to be too large such as 64 which will decrease the retrieval performance. Due to there are same classes in training set and testing set on Corel10K-b dataset, supervised-learning model is more easy to learn the representations. Here, we use t-SNE [75] to visualize the semantic space of different supervised-learning schemes [14, 26] on Corel10K-b dataset. Specifically, 10 classes are chosen from testing set, with 30 instances in each class. Fig. 10 shows that the semantic space of Cheng [14] is a few disordered. Although the semantic space of Feng [26] is separated between inter-classes, it is not compact among intra-classes. Moreover, some classes are not pulled apart. For our EViT, it is not only more distinguishable in the inter-classes but also more closer in the intra-classes. In summary, our supervised-learning retrieval model can significantly improve retrieval performance, and there may be still room for improvement (Section 4.4). We also hope our supervised-learning model can be strong baseline to explore privacy-preserving image retrieval.

5.1.3 Why not CNN and non-end-to-end. Current deep plain-image retrieval works [11] are end-to-end, which directly use images as model's inputs to automatically extract features (e.g. CNN features) rather than hand-craft features. However, it is impossible for our cipher-images to extract ruled features because the spatial structure information of cipher-images (e.g., pixel values) are randomly changed and disordered by secret keys. Therefore, EViT adopts the Manuscript submitted to ACM

Table 7. Retrieval performance (*mAP*@100) with different backbone in end-to-end (ETE) and non-end-to-end (NETE) manner on Corel10K-b dataset, "failed" represents mAP@100 less than 0.1.

Backbone	ResNet50 (ETE)	ViT (ETE)	ResNet50 (NETE)	ViT (NETE)	
Unsupervised	failed	failed	0.196	0.295	
Supervised	0.102	0.156	0.598	0.759	

Table 8. Comparison of searching time and PSNR for different schemes.

Schemes	Zhang [100]	Liang [47]	Xia [88]	Li [46]	Xia [89]	Xia [87]	Feng [26]	Ours
Searching time (s)	1.01	0.11	0.12	1.01	0.36	0.35	0.15	0.09
PSNR (dB)	16.40	9.81	10.17	13.18	13.12	13.21	13.65	12.01

non-end-to-end manner due to the artificial features (e.g., local length sequence and global Huffman-code frequency) are ruled and can be used to effectively learn representations of cipher-images. In Section 3.3, we have mentioned that our backbone uses ViT rather than CNN, and here we compare retrieval performance with ResNet50 [35] (a classical CNN backbone) on the Corel10K-b dataset. As shown in Tab. 7, the non-end-to-end manner is far beyond the end-to-end manner for different backbones in retrieval performance, and ViT surpasses ResNet50 about 10% mAP@100 in the non-end-to-end unsupervised manner. Tab. 7 presents that non-end-to-end manner and ViT are vital for improving retrieval performance.

5.2 Time consumption

We test the time consumption of searching, and average the results on the entire Corel10K dataset. When searching similar cipher-images, we consider that the retrieval model is already trained. As shown in Tab. 8, our time consumption of searching is 0.09s which is acceptable. Because our unsupervised-learning and supervised-learning model use the same backbone, the search times are also same. For a fair comparison, all schemes use Python programming language, and Tab. 8 shows that EViT is more faster than other schemes in term of searching. In the future, we will deploy deep retrieval model with C++ and use Faiss [39] to further decrease searching time.

5.3 Encryption performance

In the proposed EViT, images are encrypted during JPEG compression. The adopted encryption operations do not destroy the format information of JPEG, hence our encryption scheme is format-compliant to JPEG. In Fig. 11, we take five plain-images as examples to demonstrate the encryption performance of our encryption algorithm. Here, we analyze the encryption performance from four aspects: visual security, statistical attack, differential cryptanalysis, and key security.

Visual security: According to the example shown in Fig. 11, we can find that the encrypted images are disordered enough, and do not disclose any information of plain-images. Besides the visual checking, we use Peak Signal-to-Noise Ratio (PSNR) to evaluate the visual safety. We compare our encryption algorithm with that of current privacy-preserving image retrieval schemes, and calculate the average PSNR of 10000 images, where smaller PSNR indicates better visual safety [45]. All schemes are evaluated in YUV color space. As shown in Tab. 8, we can see that some schemes [47, 88]

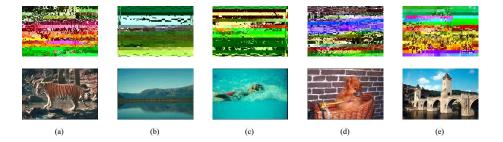


Fig. 11. Encryption examples (the first row is cipher-images, the second row is corresponding plain-images).

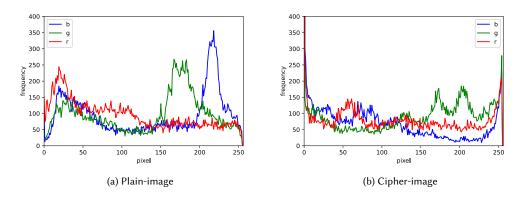


Fig. 12. Histograms of the plain and cipher images.

are with smaller PSNR than ours, because they used extra encryption steps apart from stream cipher. But our retrieval performance is significantly improved than them (Tab. 3).

Statistical attack: To make the statistical mode-based attack unavailable, the histograms of plain-images and that of cipher-images should be different. As shown in Fig. 12, taking the plain-image (Fig. 11 (e)) and its cipher-image as examples, it can be seen that there is no statistical correlation between the histogram of the encrypted image and that of the plain-image. Compared with the histogram of the plain-image, the frequency difference of different pixel values in the cipher-image is not so large, thus our scheme has a certain resistance ability against statistical attack.

Differential cryptanalysis: In order to resist differential cryptanalysis, minor change of plain-image such as modifying one single pixel should result in significant change in corresponding cipher-image [33]. For example, given plain-image P^1 , we slightly change one single pixel and obtain plain-image P^2 . Then using encryption algorithm to encrypt P^1 and P^2 , we can obtain P^1 and P^2 respectively. Differential cryptanalysis attacks cryptosystem through comparing and analyzing the differences between the cipher-images P^1 and P^2 , therefore we hope there exists significant differences between them. Generally, NPCR (number of pixels change rate) and UACI (unified average changing intensity) [9] are two commonly metrics for evaluating the ability of encryption algorithm against differential attack, which are calculated as follows:

$$D(i,j) = \begin{cases} 0, C^{1}(i,j) = C^{2}(i,j) \\ 1, C^{1}(i,j) \neq C^{2}(i,j) \end{cases}$$
 (18)

Category Church Girl Sky Architecture Africa Painting **NPCR** 95.55% 96.09% 96.51% 96.78% 96.41% 95.84% **UACI** 47.28% 48.13% 47.52% 48.00% 48.48% 48.74%

Table 9. Mean NPCR and UACI of cipher-images with one pixel changing.

$$NPCR: N\left(C^{1}, C^{2}\right) = \frac{\sum_{i,j} D(i,j)}{H \times W} \times 100 \tag{19}$$

$$NPCR: N\left(C^{1}, C^{2}\right) = \frac{\sum_{i,j} D(i,j)}{H \times W} \times 100$$

$$UACI: U\left(C^{1}, C^{2}\right) = \frac{\sum_{i,j} \frac{\left|C^{1}(i,j) - C^{2}(i,j)\right|}{255}}{H \times W} \times 100$$
(20)

where C^1 and C^2 are corresponding cipher-images of plain-images P^1 and P^2 , and C(i, j) is the pixel value at coordinates (i,j) $(1 \le i \le H, 1 \le j \le W)$. Here, we select 6 categories (church, girl, sky, architecture, painting, Africa) which contain 600 images from Corel10K, and change one single pixel of plain-image to calculate the NPCR and UACI. Each category averages the results. The closer the NPCR is to 100% and the UACI is to 33%, the more vital ability to resist differential attack [9]. As shown in Tab. 9, we can see that our encryption algorithm has a certain resistance to differential attack.

Key security: Brute-force attack is a standard ciphertext-only attack strategy in which attackers only have access to the encrypted data. Our encryption method has six encryption keys $(k_{DC*}, k_{AC*}, * \in \{Y, U, V\})$, and each component has different keys. Each of these keys are 256-bits, so the key spaces of our encryption algorithm is (2²⁵⁶)⁶. It is extremely difficult to use brute-force cracking to restore the plain images. Therefore, our scheme is safe and can be secure against ciphertext-only attack.

EViT uses adaptive encryption key generation [33] method to encrypt images, namely different images are encrypted by different keys. We have explained the reason why EViT can use adaptive encryption key generation in Section 4.2, since different keys would generate the same features of cipher-images in EViT. Some schemes such as Zhang [100] and Li [46] also can use different keys to encrypt different images, because they calculated the histograms of DCT in each frequency whose positions are unchanged, hence the features of cipher-images are unchanged. But some schemes such as Liang [47] and Xia [87] could only use the same key to encrypt different images. Liang [47] permuted the AC coefficients in each block, so different keys will change the (r, v) pairs which lead to the features extracted from cipher-images changed. Xia [87] used color value replacement to encrypt images, different keys will generate different replacement tables, so the features of cipher-images also were changed. Other schemes [26, 88, 89] have the similar situations. Once the features are changed with different keys, the feature spaces of all cipher-images are different, so it leads to compromised retrieval performance.

Our encryption algorithm cannot achieve absolute secure, and its visual safety is slightly lower than other schemes. Absolute secure encryption algorithm cannot make us extract ruled features from cipher-images, and current imageencryption-based schemes generally fail to ensure absolute security. For example, Zhang [100] leaked DCT histograms of cipher-images. Liang [47] and Xia [88] achieved smaller PSNR with visual safety than ours, but their retrieval performance is about 10% mAP@100 lower than ours. Moreover, they used the same keys to encrypt all images, so they fail to resist differential attack. Our EViT can achieve the best retrieval performance and nice visual security among different schemes. Current privacy-preserving image retrieval schemes have been seeking balances among different performances, for example slightly compromising visual security to significantly improve retrieval performance. This is also the purpose of our work.

6 CONCLUSION

In this paper, we propose a novel privacy-preserving image retrieval scheme named EViT, which can improve retrieval performance by large margins than current schemes and effectively protect security of images. First, we design multi-level features (local length sequence and global Huffman-code frequency) from cipher-images which are encrypted by stream cipher with VLI code during JPEG compression process, and EViT supports adaptive encryption keys due to the the multi-level features are unchanged with different secret keys. Second, EViT proposes the unsupervised retrieval model in a self-supervised learning manner, and adopts the structure of ViT as the model's backbone to couple with the multi-level features. To improve retrieval performance, EViT adopts two adaptive data augmentations for retrieval model, and advances ViT with learnable global Huffman-code frequency. The supervised model can be easily achieved by Fine-Tuning on the trained unsupervised retrieval model. Experimental results show that EViT not only effectively protects image privacy but also significantly improves retrieval performance than current schemes. In future work, we will try to further improve the encryption performance while keeping retrieval accuracy.

ACKNOWLEDGMENT

The authors would like to thank the anonymous reviewers for their constructive comments and suggestions.

REFERENCES

- [1] Sam Gross et al. Adam Paszke. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 8024–8035.
- [2] Abien Fred Agarap. 2018. Deep Learning using Rectified Linear Units (ReLU). CoRR abs/1803.08375 (2018). arXiv:1803.08375 http://arxiv.org/abs/1803.08375
- [3] Sara Atito Ali Ahmed, Muhammad Awais, and Josef Kittler. 2021. SiT: Self-supervised vIsion Transformer. CoRR abs/2104.03602 (2021). arXiv:2104.03602 https://arxiv.org/abs/2104.03602
- [4] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. 2021. ViViT: A Video Vision Transformer. In 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021. IEEE, 6816–6826.
- [5] Jean-Philippe Aumasson, Samuel Neves, Zooko Wilcox-O'Hearn, and Christian Winnerlein. 2013. BLAKE2: Simpler, Smaller, Fast as MD5. In Applied Cryptography and Network Security - 11th International Conference, ACNS 2013 (Lecture Notes in Computer Science, Vol. 7954). Springer, 119–135.
- [6] Lei Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. CoRR abs/1607.06450 (2016). arXiv:1607.06450 http://arxiv.org/abs/1607.06450
- [7] Fabrice Benhamouda, Shai Halevi, and Tzipora Halevi. 2019. Supporting private data on Hyperledger Fabric with secure multiparty computation. IBM Journal of Research and Development 63, 2/3 (2019), 3:1–3:8.
- [8] Piotr Bojanowski and Armand Joulin. 2017. Unsupervised Learning by Predicting Noise. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017 (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 517–526.
- [9] Guanrong Chen, Yaobin Mao, and Charles K Chui. 2004. A symmetric image encryption scheme based on 3D chaotic cat maps. Chaos, Solitons & Fractals 21, 3 (2004), 749–761.
- [10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In Proceedings of the 37th International Conference on Machine Learning, ICML 2020 (Proceedings of Machine Learning Research, Vol. 119). PMLR, 1597–1607.
- [11] Wei Chen, Yu Liu, Weiping Wang, Erwin M. Bakker, Theodoros Georgiou, Paul W. Fieguth, Li Liu, and Michael S. Lew. 2021. Deep Image Retrieval: A Survey. CoRR abs/2101.11282 (2021). arXiv:2101.11282 https://arxiv.org/abs/2101.11282
- [12] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. 2020. Improved Baselines with Momentum Contrastive Learning. CoRR abs/2003.04297 (2020). arXiv:2003.04297 https://arxiv.org/abs/2003.04297
- [13] Hang Cheng, Xinpeng Zhang, and Jiang Yu. 2016. AC-coefficient histogram-based retrieval for encrypted JPEG images. Multimedia Tools and Applications 75, 21 (2016), 13791–13803.
- [14] Hang Cheng, Xinpeng Zhang, Jiang Yu, and Fengyong Li. 2015. Markov Process Based Retrieval for Encrypted JPEG Images. In 10th International Conference on Availability, Reliability and Security, ARES 2015. IEEE Computer Society, 417–421.

- [15] Hang Cheng, Xinpeng Zhang, Jiang Yu, and Yuan Zhang. 2016. Encrypted JPEG image retrieval using block-wise feature comparison. Journal of Visual Communication and Image Representation 40 (2016), 111–117.
- [16] Zhi Cheng, Xiu Su, Xueyu Wang, Shan You, and Chang Xu. 2022. Sufficient Vision Transformer. In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (Washington DC, USA) (KDD '22). Association for Computing Machinery, New York, NY, USA, 190–200. https://doi.org/10.1145/3534678.3539322
- [17] Charilaos A. Christopoulos, Touradj Ebrahimi, and Athanassios N. Skodras. 2000. JPEG2000: the new still picture compression standard. In Proceedings of the ACM Multimedia 2000 Workshops 2000, Shahram Ghandeharizadeh, Shih-Fu Chang, Stephen Fischer, Joseph A. Konstan, and Klara Nahrstedt (Eds.). ACM Press. 45–49.
- [18] Zhiyuan Dang, Cheng Deng, Xu Yang, Kun Wei, and Heng Huang. 2021. Nearest Neighbor Matching for Deep Clustering. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021. Computer Vision Foundation / IEEE, 13693–13702.
- [19] Cheng Deng, Erkun Yang, Tongliang Liu, and Dacheng Tao. 2020. Two-Stream Deep Hashing With Class-Specific Centers for Supervised Image Search. IEEE Transactions on Neural Networks and Learning Systems 31, 6 (2020), 2189–2201.
- [20] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019. Computer Vision Foundation / IEEE, 4690–4699.
- [21] Jiankang Deng and Stefanos Zafeiriou. 2019. ArcFace for Disguised Face Recognition. In 2019 IEEE/CVF International Conference on Computer Vision Workshops, ICCV Workshops 2019. IEEE, 485–493.
- [22] Zelu Deng, Yujie Zhong, Sheng Guo, and Weilin Huang. 2021. InsCLR: Improving Instance Retrieval with Self-Supervision. CoRR abs/2112.01390 (2021). arXiv:2112.01390 https://arxiv.org/abs/2112.01390
- [23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019. Association for Computational Linguistics, 4171–4186.
- [24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In 9th International Conference on Learning Representations, ICLR 2021. OpenReview.net.
- [25] Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021. ParsBERT: Transformer-based Model for Persian Language Understanding. Neural Processing Letters 53, 6 (2021), 3831–3847.
- [26] Qihua Feng, Peiya Li, ZhiXun Lu, Guan Liu, and Feiran Huang. 2021. End-to-end Learning for Encrypted Image Retrieval. In Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2021. IEEE, 1839–1845.
- [27] Wouter Van Gansbeke, Simon Vandenhende, Stamatios Georgoulis, Marc Proesmans, and Luc Van Gool. 2020. SCAN: Learning to Classify Images Without Labels. In Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part X (Lecture Notes in Computer Science, Vol. 12355). Springer, 268–285.
- [28] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021. Association for Computational Linguistics, 6894–6910.
- [29] Xingyu Gao, Steven C. H. Hoi, Yongdong Zhang, Jianshe Zhou, Ji Wan, Zhenyu Chen, Jintao Li, and Jianke Zhu. 2017. Sparse Online Learning of Image Similarity. ACM Trans. Intell. Syst. Technol. 8, 5, Article 64 (aug 2017), 22 pages. https://doi.org/10.1145/3065950
- [30] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. 2018. Unsupervised Representation Learning by Predicting Image Rotations. In 6th International Conference on Learning Representations, ICLR 2018. OpenReview.net.
- [31] Benjamin Graham, Alaaeldin El-Nouby, Hugo Touvron, Pierre Stock, Armand Joulin, Hervé Jégou, and Matthijs Douze. 2021. LeViT: a Vision Transformer in ConvNet's Clothing for Faster Inference. In 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021. IEEE, 12239– 12249
- [32] Qishen Ha, Bo Liu, Fuxu Liu, and Peiyuan Liao. 2020. Google Landmark Recognition 2020 Competition Third Place Solution. CoRR abs/2010.05350 (2020). arXiv:2010.05350 https://arxiv.org/abs/2010.05350
- [33] Junhui He, Shuhao Huang, Shaohua Tang, and Jiwu Huang. 2018. JPEG Image Encryption With Improved Format Compatibility and File Size Preservation. IEEE Transactions on Multimedia 20, 10 (2018), 2645–2658.
- [34] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. 2020. Momentum Contrast for Unsupervised Visual Representation Learning. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020. Computer Vision Foundation / IEEE, 9726–9735.
- [35] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016. IEEE Computer Society, 770–778.
- [36] Syed Sameed Husain and Miroslaw Bober. 2019. REMAP: Multi-Layer Entropy-Guided Pooling of Dense CNN Features for Image Retrieval. IEEE Transactions on Image Processing 28, 10 (2019), 5201–5213.
- [37] T Janani and M Brindha. 2021. SEcure Similar Image Matching (SESIM): An Improved Privacy Preserving Image Retrieval Protocol over Encrypted Cloud Database. IEEE Transactions on Multimedia (2021). https://doi.org/10.1109/TMM.2021.3107681
- [38] Young Kyun Jang and Nam Ik Cho. 2021. Self-supervised Product Quantization for Deep Unsupervised Image Retrieval. In 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021. IEEE, 12065–12074.
- [39] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. IEEE Transactions on Big Data 7, 3 (2019), 535–547.
- [40] Mahmut Kaya and Hasan Sakir Bilge. 2019. Deep Metric Learning: A Survey. Symmetry 11, 9 (2019), 1066-1092.

[41] Bumsoo Kim, Junhyun Lee, Jaewoo Kang, Eun-Sol Kim, and Hyunwoo J. Kim. 2021. HOTR: End-to-End Human-Object Interaction Detection With Transformers. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021. Computer Vision Foundation / IEEE, 74–83.

- [42] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012, Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger (Eds.). 1106–1114.
- [43] Jia Li and James Ze Wang. 2003. Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach. IEEE Transactions on Pattern Analysis and Machine Intelligence 25, 9 (2003), 1075–1088.
- [44] Ke Li, Shijie Wang, Xiang Zhang, Yifan Xu, Weijian Xu, and Zhuowen Tu. 2021. Pose Recognition With Cascade Transformers. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021. Computer Vision Foundation / IEEE, 1944–1953.
- [45] Peiya Li and Kwok-Tung Lo. 2018. A Content-Adaptive Joint Image Compression and Encryption Scheme. IEEE Transactions on Multimedia 20, 8 (2018), 1960–1972.
- [46] Peiya Li and Zhenhui Situ. 2019. Encrypted JPEG image retrieval using histograms of transformed coefficients. In 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2019. IEEE, 1140–1144.
- [47] Haihua Liang, Xinpeng Zhang, and Hang Cheng. 2019. Huffman-code based retrieval for encrypted JPEG images. Journal of Visual Communication and Image Representation 61 (2019), 149–156.
- [48] Kevin Lin, Lijuan Wang, and Zicheng Liu. 2021. End-to-End Human Pose and Mesh Reconstruction with Transformers. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021. Computer Vision Foundation / IEEE, 1954–1963.
- [49] Chundi Liu, Guang Wei Yu, Maksims Volkovs, Cheng Chang, Himanshu Rai, Junwei Ma, and Satya Krishna Gorti. 2019. Guided Similarity Separation for Image Retrieval. In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019. 1554–1564.
- [50] Dandan Liu, Jian Shen, Zhihua Xia, and Xingming Sun. 2017. A Content-Based Image Retrieval Scheme Using an Encrypted Difference Histogram in Cloud Computing. Inf. 8. 3 (2017), 96–109.
- [51] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021. IEEE, 9992–10002.
- [52] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA. IEEE Computer Society, 1096–1104.
- [53] Wenjun Lu, Ashwin Swaminathan, Avinash L. Varna, and Min Wu. 2009. Enabling search over encrypted multimedia databases. In Media Forensics and Security I, part of the IS&T-SPIE Electronic Imaging Symposium, San Jose, CA, USA, January 19-21, 2009, Proceedings (SPIE Proceedings, Vol. 7254). SPIE, 725418.
- [54] Wenjun Lu, Avinash L. Varna, Ashwin Swaminathan, and Min Wu. 2009. Secure image retrieval through feature protection. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009, 19-24 April 2009, Taipei, Taiwan. IEEE, 1533–1536.
- [55] Yue Lv, Wengang Zhou, Qi Tian, Shaoyan Sun, and Houqiang Li. 2018. Retrieval Oriented Deep Feature Learning With Complementary Supervision Mining. IEEE Transactions on Image Processing 27, 10 (2018), 4945–4957.
- [56] James MacQueen et al. 1967. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Vol. 1. Oakland, CA, USA, 281–297.
- [57] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. Introduction to information retrieval. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071
- [58] Ishan Misra and Laurens van der Maaten. 2020. Self-Supervised Learning of Pretext-Invariant Representations. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020. Computer Vision Foundation / IEEE, 6706–6716.
- [59] Muzammal Naseer, Kanchana Ranasinghe, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. 2021. Intriguing Properties of Vision Transformers. CoRR abs/2105.10497 (2021). arXiv:2105.10497 https://arxiv.org/abs/2105.10497
- [60] Mehdi Noroozi and Paolo Favaro. 2016. Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. In Computer Vision ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VI (Lecture Notes in Computer Science, Vol. 9910). Springer, 69-84
- [61] Margarita Osadchy, Benny Pinkas, Ayman Jarrous, and Boaz Moskovich. 2010. SCiFI A System for Secure Face Identification. In 31st IEEE Symposium on Security and Privacy, S&P 2010, 16-19 May 2010, Berleley/Oakland, California, USA. IEEE Computer Society, 239–254.
- [62] Kohei Ozaki and Shuhei Yokoo. 2019. Large-scale Landmark Retrieval/Recognition under a Noisy and Diverse Dataset. CoRR abs/1906.04087 (2019). arXiv:1906.04087 http://arxiv.org/abs/1906.04087
- [63] William B Pennebaker and Joan L Mitchell. 1992. JPEG: Still image data compression standard. Springer Science & Business Media.
- [64] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. 2021. Multi-Modal Fusion Transformer for End-to-End Autonomous Driving. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021. Computer Vision Foundation / IEEE, 7077–7087.
- [65] Zhenxing Qian, Xinpeng Zhang, and Shuozhong Wang. 2014. Reversible Data Hiding in Encrypted JPEG Bitstream. *IEEE Transactions on Multimedia* 16, 5 (2014), 1486–1491.
- [66] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research 21 (2020), 140:1–140:67.

- [67] Zhongzheng Ren and Yong Jae Lee. 2018. Cross-Domain pervised Multi-Task Feature Learning Using Synthetic Imagery. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018. Computer Vision Foundation / IEEE Computer Society, 762–771.
- [68] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015. IEEE Computer Society, 815–823.
- [69] Josef Sivic and Andrew Zisserman. 2003. Video Google: A Text Retrieval Approach to Object Matching in Videos. In 9th IEEE International Conference on Computer Vision (ICCV 2003). IEEE Computer Society, 1470–1477.
- [70] Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research 15, 1 (2014), 1929–1958.
- [71] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. 2021. LoFTR: Detector-Free Local Feature Matching With Transformers. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021. Computer Vision Foundation / IEEE, 8922–8931.
- [72] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive Multiview Coding. In Computer Vision ECCV 2020 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XI (Lecture Notes in Computer Science, Vol. 12356). Springer, 776-794.
- [73] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine Learning Research, Vol. 139), Marina Meila and Tong Zhang (Eds.). PMLR, 10347–10357.
- [74] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. CoRR abs/1807.03748 (2018). arXiv:1807.03748 http://arxiv.org/abs/1807.03748
- [75] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing Data Using t-SNE. Journal of Machine Learning Research 9, 11 (2008), 2579-2605.
- [76] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. 5998–6008.
- [77] Feng Wang, Xiang Xiang, Jian Cheng, and Alan Loddon Yuille. 2017. NormFace: L₂ Hypersphere Embedding for Face Verification. In Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017, Qiong Liu, Rainer Lienhart, Haohong Wang, Sheng-Wei "Kuan-Ta" Chen, Susanne Boll, Yi-Ping Phoebe Chen, Gerald Friedland, Jia Li, and Shuicheng Yan (Eds.). ACM, 1041–1049.
- [78] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. 2021. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions. In 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021. IEEE, 548–558
- [79] Jason W. Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 6381–6387.
- [80] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A Discriminative Feature Learning Approach for Deep Face Recognition. In Computer Vision - ECCV 2016 - 14th European Conference, Proceedings, Part VII (Lecture Notes in Computer Science, Vol. 9911), Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer, 499–515.
- [81] Li Weng, Laurent Amsaleg, and Teddy Furon. 2016. Privacy-Preserving Outsourced Media Search. IEEE Transactions on Knowledge and Data Engineering 28, 10 (2016), 2738–2751.
- [82] Li Weng, Laurent Amsaleg, April Morton, and Stéphane Marchand-Maillet. 2015. A Privacy-Preserving Framework for Large-Scale Content-Based Information Retrieval. IEEE Transactions on Information Forensics and Security 10, 1 (2015), 152–167.
- [83] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020, Qun Liu and David Schlangen (Eds.). Association for Computational Linguistics, 38–45.
- [84] Wai Kit Wong, David Wai-Lok Cheung, Ben Kao, and Nikos Mamoulis. 2009. Secure kNN computation on encrypted databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2009, Ugur Çetintemel, Stanley B. Zdonik, Donald Kossmann, and Nesime Tatbul (Eds.). ACM. 139–152.
- [85] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. 2018. Unsupervised Feature Learning via Non-Parametric Instance Discrimination. In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018. Computer Vision Foundation / IEEE Computer Society, 3733–3742.
- [86] Zhihua Xia, Qiuju Ji, Qi Gu, Chengsheng Yuan, and Fengjun Xiao. 2022. A Format-compatible Searchable Encryption Scheme for JPEG Images Using Bag-of-words. ACM Transactions on Multimedia Computing, Communications, and Applications 18, 3 (2022), 85:1–85:18.
- [87] Zhihua Xia, Leqi Jiang, Dandan Liu, Lihua Lu, and Byeungwoo Jeon. 2022. BOEW: A Content-Based Image Retrieval Scheme Using Bag-of-Encrypted-Words in Cloud Computing. IEEE Transactions on Services Computing 15, 1 (2022), 202–214.
- [88] Zhihua Xia, Lihua Lu, Tong Qiu, HJ Shim, Xianyi Chen, and Byeungwoo Jeon. 2019. A privacy-preserving image retrieval based on AC-coefficients and color histograms in cloud environment. Computers, Materials & Continua 58, 1 (2019), 27–44.
- [89] Zhihua Xia, Lan Wang, Jian Tang, Neal N. Xiong, and Jian Weng. 2021. A Privacy-Preserving Image Retrieval Scheme Using Secure Local Binary Pattern in Cloud Computing. IEEE Transactions on Network Science and Engineering 8, 1 (2021), 318–330.
- [90] Zhihua Xia, Xinhui Wang, Liangao Zhang, Zhan Qin, Xingming Sun, and Kui Ren. 2016. A Privacy-Preserving and Copy-Deterrence Content-Based Image Retrieval Scheme in Cloud Computing. IEEE Transactions on Information Forensics and Security 11, 11 (2016), 2594–2608.

[91] Zhihua Xia, Neal N. Xiong, Athanasios V. Vasilakos, and Xingming Sun. 2017. EPCBIR: An efficient and privacy-preserving content-based image retrieval scheme in cloud computing. Information Sciences 387 (2017), 195–204.

- [92] Zhihua Xia, Yi Zhu, Xingming Sun, Zhan Qin, and Kui Ren. 2018. Towards Privacy-Preserving Content-Based Image Retrieval in Cloud Computing. IEEE Transactions on Cloud Computing 6, 1 (2018), 276–286.
- [93] Junyuan Xie, Ross B. Girshick, and Ali Farhadi. 2016. Unsupervised Deep Embedding for Clustering Analysis. In Proceedings of the 33nd International Conference on Machine Learning, ICML 2016 (JMLR Workshop and Conference Proceedings, Vol. 48), Maria-Florina Balcan and Kilian Q. Weinberger (Eds.). JMLR.org, 478–487.
- [94] Zhenda Xie, Yutong Lin, Zhuliang Yao, Zheng Zhang, Qi Dai, Yue Cao, and Han Hu. 2021. Self-Supervised Learning with Swin Transformers. CoRR abs/2105.04553 (2021). https://arxiv.org/abs/2105.04553
- [95] Yanyan Xu, Jiaying Gong, Lizhi Xiong, Zhengquan Xu, Jinwei Wang, and Yun-Qing Shi. 2017. A privacy-preserving content-based image retrieval method in cloud environment. Journal of Visual Communication and Image Representation 43 (2017), 164–172.
- [96] Jufeng Yang, Jie Liang, Hui Shen, Kai Wang, Paul L. Rosin, and Ming-Hsuan Yang. 2018. Dynamic Match Kernel With Deep Convolutional Features for Image Retrieval. IEEE Transactions on Image Processing 27, 11 (2018), 5288-5302.
- [97] Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained Transformers for Text Ranking: BERT and Beyond. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (Virtual Event, Canada) (SIGIR '21). Association for Computing Machinery, New York, NY, USA, 2666–2668. https://doi.org/10.1145/3404835.3462812
- [98] Chengyuan Zhang, Lei Zhu, Shichao Zhang, and Weiren Yu. 2020. TDHPPIR: An Efficient Deep Hashing Based Privacy-Preserving Image Retrieval Method. Neurocomputing 406 (2020), 386–398.
- [99] Lan Zhang, Taeho Jung, Kebin Liu, Xiang-Yang Li, Xuan Ding, Jiaxi Gu, and Yunhao Liu. 2017. PIC: Enable Large-Scale Privacy Preserving Content-Based Image Search on Cloud. IEEE Transactions on Parallel and Distributed Systems 28, 11 (2017), 3258–3271.
- [100] Xinpeng Zhang and Hang Cheng. 2014. Histogram-based retrieval for encrypted JPEG images. In 2014 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP). IEEE, 446–449.