Hindawi
International Journal of Intelligent Systems
Volume 2023, Article ID 8931092, 16 pages
https://doi.org/10.1155/2023/8931092

Research Article

Privacy-Preserving Image Retrieval Based on Disordered Local Histograms and Vision Transformer in Cloud Computing

Zhangdong Wang, Jiaohua Qin, Xuyu Xiang, and Yun Tan

College of Computer Science and Information Technology, Central South University of Forestry and Technology, Changsha 410004, China

Correspondence should be addressed to Jiaohua Qin; qinjiaohua@163.com

Received 30 December 2022; Revised 5 September 2023; Accepted 12 September 2023; Published 21 September 2023

Academic Editor: Vasudevan Rajamohan

Frequent data breaches in the cloud environment have seriously affected cloud subscribers and providers. Privacy-preserving image retrieval methods can improve the security of cloud image retrieval; however, existing methods have limited accuracy on dynamically updated image databases and mobile lightweight devices. In this study, we propose a privacy-preserving image retrieval method based on disordered local histograms and vision transformer in cloud computing, by designing a multiple encryption method and transformer-based feature model to better mine the local feature value of encrypted images. Specifically, the user performs different value substitution, position substitution, and color substitution on the subblocks of the image to protect the image information. The cloud server extracts the unordered local histogram from the encrypted image and generates retrievable features using transformer. Experiments show that compared with similar CNN schemes, the retrieval accuracy of this method is improved by 8.5%, and the retrieval efficiency is improved by 54.8%.

1. Introduction

The rapid growth of global data and the COVID-19 outbreak are driving more and more individuals and businesses to use remote offices and cloud servers [1]. The cloud environment reduces the cost of storing and managing massive amounts of private data [2], but it significantly increases the risk of data leakage [3]. The average cost of a data breach reached a record high of $4.35 million in 2022. 45% of these data breaches occur in the cloud [4]. Although data encryption can reduce the loss caused by data leakage [5], it increases the difficulty of retrieving encrypted data [6, 7].

Image is one of the important component media of cloud data. Privacy-preserving content-based image retrieval (PPCBIR) [8] transmits the image to the cloud in encrypted form, while maintaining its searchability [9], which better plays the value of the cloud. In addition, more and more users are using the cloud to store or backup images on mobile terminal (resource-limited) devices such as mobile phones [10], sensor networks [11, 12], and vehicle-mounted equipment [13, 14]. In 2021, the number of active users of Huawei's terminal cloud in the world will exceed 730 million. Therefore, it is of great value to study PPCBIR to support real-time image updating on mobile (restricted) terminal devices [15].

Existing PPCBIR schemes have limitations in dynamic update and mobile terminal scenarios. PPCBIR scheme [16] uses homomorphic encryption to protect images. Although it has high security and retrieval accuracy, the calculation cost is too high. PPCBIR scheme [17] extracts features from images and builds a security index, and then, traditional encryption protects images [18]. Although it has high security and retrieval accuracy, users are required to extract features, build, and update indexes in real time, but mobile devices cannot bear the computational complexity under large-scale and dynamic image updating. PPCBIR scheme [19] uses special encryption algorithms and feature models to directly extract features from encrypted images for retrieval. Although the client only carries out encryption operations, feature processing and retrieval calculations are

completed in the cloud, which can better support dynamic updates and mobile devices, but the retrieval accuracy is not high. Therefore, this paper studies how to design encryption algorithms and feature models to improve the retrieval accuracy of PPCBIR under dynamic update images and mobile terminals.

Image feature representation ability is the key factor in determining the retrieval accuracy of PPCBIR. Some existing PPCBIR schemes use global features such as histogram saturation value (HSV) [20] and local binary patterns (LBP) [19] to characterize encrypted images, which are relatively simple to compute but less accurate. Some schemes use local features such as scale-invariant feature transform (SIFT) [21, 22] and accelerated robust features (SURF) [23] to characterize encrypted images with better accuracy, but the relative position information between local features is easily leaked. In addition, some PPCBIR schemes use CNN for feature enhancement of global features [24, 25], and the accuracy is better than that of PPCBIR [26] based on global features. However, the rough representation of global features limits the retrieval accuracy to some extent. Although the disordered local feature has richer details than the global feature, the relative location information of the local feature will not be revealed. However, CNN cannot model the out-of-order local features due to the spatial invariance of its inductive bias and regional limitations.

The vision transformer (ViT) [27] demonstrated superior semantic features and long-range feature capture capabilities over CNN in image classification tasks. The disturbance invariance of the self-attention mechanism in transformer enables it to learn the relationship between disordered local features, which has more advantages than CNN features in learning global features [28]. Therefore, this paper proposes a privacy-preserving image retrieval method based on disordered local histograms and vision transformers in cloud computing, which can satisfy the dynamic update of the image and mobile (limited) user's device efficient retrieval. The main contributions of this work are summarized as follows:

(1) A secure multiple encryption method is proposed, where local, global, color, and texture multiple encryption methods not only protect image information but also support models to extract unordered histogram features from the encrypted image for secure retrieval. Security analysis shows that our encrypted images are resistant to brute force and statistical attacks.

(2) A transformer-based feature extraction model is proposed, which uses a transformer to fully excavate the correlation between local features and improve the retrieval accuracy without reducing the security. The retrieval experiments on benchmark datasets show that the retrieval precision of our scheme is better than that of some existing similar schemes.

The rest of this paper is structured as follows. Section 2 describes the related work, Section 3 provides an introduction to the system and security model, Section 4 details the proposed scheme, Section 5 provides a security analysis of the proposed scheme, Section 6 presents the retrieval experimental results and analysis, and Section 7 concludes the work.

2. Related Work

Existing privacy-preserving image retrieval schemes can be classified as homomorphic encryption, feature index encryption, and pure image encryption image retrieval schemes.

2.1. Homomorphic Encryption-Based Image Retrieval. Hsu et al. [29] proposed a SIFT extraction method based on homomorphic encryption. Zhang et al. [30] performed multilevel homomorphic encryption on the histogram of image visual words. Bellafqira et al. [31] extracted SIFT and discrete wavelet transform features for retrieval in homomorphically encrypted domains. Guo et al. [32] extracted CNN features in homomorphically encrypted domains for retrieval. Lu et al. [33] compared homomorphic encryption methods with feature and index encryption methods. Homomorphic encryption can achieve the same retrieval accuracy as the plaintext domain but is computationally intensive and complex to communicate.

2.2. Feature Index Encryption-Based Image Retrieval. Cheng et al. [34] studied inverse index generation using visual words of images. Zou et al. [35] used tree structure and Euclidean distance for secondary search to improve the efficiency and accuracy of the fuzzy search. Li et al. [36] used feature descriptors extracted by CNN models to improve search accuracy and designed a hierarchical index tree for K-means clustering based on affinity propagation clustering to enhance efficiency. In addition, a limited key leakage mechanism is constructed based on the KNN [37] algorithm to support untrusted image users to generate trapdoors locally without the image owner online. Li et al. [38] used CNN to extract features, carried out K-means clustering to build index trees, and finally used dynamic trees to verify the correctness of the results. Huang [39] uses the ViT model to improve retrieval accuracy and design a secure multi-indexed hash structure to filter datasets to improve retrieval efficiency. Feature and index encryption methods achieve better security and retrieval performance in fixed databases, but the computational effort of index construction increases dramatically with the frequency of image library updates.

2.3. Pure Image Encryption-Based Image Retrieval. Lu et al. [40] proposed the first privacy-preserving content-based image retrieval (CBIR) scheme based on an encrypted image database. The global HSV histogram is extracted to construct visual word sets, and the similarity is measured by the Jaccard distance between visual word sets. Ferreira et al. [20] separated the color and texture information in the image, protected the color values by random permutations, retrieved the HSV histogram of the encrypted image, and

measured the similarity by the Hamming distance of the features. Liu et al. [41] encrypted the images by value substitution and positional dislocation, extracted the encrypted difference histogram as features, and measured the similarity by calculating the Euclidean distance between the features. Xia et al. [42] constructed features from a combination of AC coefficient histograms and color histograms extracted from encrypted Y-component histograms and measured image similarity by calculating the Manhattan distance between features. Xia et al. [43] encrypted the image by block transform, intrablock transform, and order-preserving pixel replacement to extract secure LBP descriptors for retrieval. Xia et al. [44] extracted local features from encrypted DCT blocks and then constructed a word bag model for local features to represent the encrypted images. Wang et al. [25] encrypted images using block-internal scrambling, interblock scrambling, and channel scrambling and used hashing and reversible information hiding for leak tracking. The global HSV is enhanced by CNN to improve retrieval accuracy. Ma et al. [45] used an improved DenseNet network to extract semantic features from encrypted images. Zhou et al. [46] considered encrypted image retrieval under a distributed environment and extracted color histograms of encrypted images for retrieval. Yu et al. [47] first characterize the image by encrypting the DCT coefficient blocks, then extracting the local Markov features of the encrypted image, and finally constructing the feature vector of local features by the BOW model. The pure image encryption method is computationally simpler than the homomorphic encryption and feature index encryption methods and meets the requirements for real-world application.

3. Problem Formulation

This section will introduce the system model and the threat model. We summarize the global symbols of the article as detailed in Table 1, with method-specific local symbols declared at their first occurrence.

3.1. System Model. The system in this paper can be divided into three parts: data owner, cloud server, and query user. The system framework is shown in Figure 1.

3.1.1. Data Owner. It is the owner of the image data. First, the image database $I = \{i_1, i_2, \cdots, i_n\}$ consisting of n images is encrypted using the randomly generated key $K = \{k_1, k_2, \cdots, k_n\}$, and then, the encrypted image database $E = \{e_1, e_2, \cdots, e_n\}$ is used to train the feature extraction model Ψ. Finally, we upload E and Ψ to the cloud server.

3.1.2. Cloud Server. It is the provider of storage space and computing power. First, E and Ψ from the data owner are stored and deployed. When the query occurs, the cloud server receives the encrypted query image EQ from the query user and sends E and EQ to Ψ to get the encrypted image database feature $F_E = \{f_{e1}, f_{e2}, \cdots, f_{en}\}$ and the query image feature F_{EQ}, and similarity is measured by calculating the Euclidean distance between the F_E and F_{EQ} features. Finally, the k most similar encrypted query result images $\mathrm{ER} = \{\mathrm{er}_{\mathrm{ID1}}, \mathrm{er}_{\mathrm{ID2}}\mathrm{er}_{\mathrm{IDk}}\mathrm{er}_{\mathrm{IDk}}\}$ are returned to the query user.

3.1.3. Query User. Query user is the user who retrieves the images. First, the query image Q is encrypted to get EQ. Second, the query user uploads the EQ to the cloud server after the cloud server retrieves the encrypted query result image set ER. Then, the query user sends the query result image ID $\mathrm{ID}_R = \{\mathrm{ID}_{R1}, \mathrm{ID}_{R2}, \cdots, \mathrm{ID}_{\mathrm{Rk}}\}$ to the data owner and obtains the corresponding key $\mathrm{RK} = \{\mathrm{rk}_{\mathrm{IDR1}}, \mathrm{rk}_{\mathrm{IDR2}}, \cdots, \mathrm{rk}_{\mathrm{IDRk}}\}$ to decrypt the query result image $R = \{r_{\mathrm{IDR1}}, r_{\mathrm{IDR2}}, \ldots, r_{\mathrm{IDR}k}\}$.

The deployment and retrieval process of the system is as follows.

Step 1. The data owner encrypts the image database I using the key K to get the encrypted image database E. Meanwhile, the image database is used to train the feature extraction model Ψ.

Step 2. The data owner uploads the encrypted image database E and feature extraction model Ψ to the cloud server, which performs database storage and model deployment.

Step 3. The querying user sends authentication information to the data owner to obtain query authorization and encryptor.

Step 4. The query user feeds the query image to the encryptor to obtain the encrypted query image.

Step 5. Query user sends encrypted query image to the cloud server.

Step 6. The cloud server uses a feature extraction model to extract features F_E, F_{EQ} from the encrypted image database and encrypted query images.

Step 7. The cloud server performs a similarity measure on the features to obtain similar encrypted query results $\mathrm{ER} = \{\mathrm{er}_{\mathrm{IDR1}}, \mathrm{er}_{\mathrm{IDR2}}, \cdots, \mathrm{er}_{\mathrm{IDR}k}\}$.

Step 8. The cloud server returns the encrypted query results $\mathrm{ER} = \{\mathrm{er}_{\mathrm{IDR1}}, \mathrm{er}_{\mathrm{IDR2}}, \cdots, \mathrm{er}_{\mathrm{IDR}k}\}$ to the querying user.

Step 9. The query user sends the ID $\mathrm{ID}_R = \{\mathrm{ID}_{R1}, \mathrm{ID}_{R2}, \cdots, \mathrm{ID}_{Rk}\}$ of the encrypted query result to the data owner and obtains the corresponding key $\mathrm{RK} = \{\mathrm{rk}_{\mathrm{IDR1}}, \mathrm{rk}_{\mathrm{IDR2}}, \cdots, \mathrm{rk}_{\mathrm{IDR}k}\}$.

Step 10. The query user decrypts the encrypted image $\mathrm{ER} = \{\mathrm{er}_{\mathrm{IDR1}}, \mathrm{er}_{\mathrm{IDR2}}, \cdots, \mathrm{er}_{\mathrm{IDR}k}\}$ using the key $\mathrm{RK} = \{\mathrm{rk}_{\mathrm{IDR1}}, \mathrm{rk}_{\mathrm{IDR2}}, \cdots, \mathrm{rk}_{\mathrm{IDR}k}\}$ to obtain the query results $R = \{r_{\mathrm{IDR1}}, r_{\mathrm{IDR2}}, \cdots, r_{\mathrm{IDR}k}\}$.

Compared with the schemes of homomorphic encryption and feature index encryption, the query user in our system only needs to perform encryption operations on images without extracting features, which significantly reduces the device requirements on the user side and supports the lightweight mobile user side. The feature extraction

Table 1: The summary of notations.

Notations	Definitions
n	The size of the image dataset
m^2	Number of blocks
$I = \{i_1, i_2, \cdots, i_n\}$	The plaintext image dataset
$E = \{e_1, e_2, \cdots, e_n\}$	The encrypted image dataset
$K = \{k_1, k_2, \ldots, k_n\}$	The set of security keys
$F_E = \{f_{e1}, f_{e2}, \cdots, f_{en}\}$	The encrypted image feature dataset
Q	The plaintext query image
EQ	The encrypted query image
F_{EQ}	The encrypted query image feature
Ψ	The feature extraction model
$\mathrm{ER} = \{\mathrm{er}_{\mathrm{ID1}}, \mathrm{er}_{\mathrm{ID2}}, \cdots, \mathrm{er}_{\mathrm{IDk}}\}$	The encrypted query result images
$\mathrm{ID}_R = \{\mathrm{ID}_{R1}, \mathrm{ID}_{R2}, \cdots, \mathrm{ID}_{Rk}\}$	The query result image ID
$\mathrm{RK} = \{\mathrm{rk}_{\mathrm{IDR1}}, \mathrm{rk}_{\mathrm{IDR2}}, \cdots, \mathrm{rk}_{\mathrm{IDRk}}\}$	The key corresponding to the resulting image
$R = \{r_{\mathrm{IDR1}}, r_{\mathrm{IDR2}}, \cdots, r_{\mathrm{IDRk}}\}$	The query result image
$\mathrm{sub}I = \{\mathrm{sub}I_1, \mathrm{sub}I_2, \cdots, \mathrm{sub}I_{m^2}\}$	Image subblock
$\mathrm{sub}E = \{\mathrm{sub}E_1, \mathrm{sub}E_2, \cdots, \mathrm{sub}E_{m^2}\}$	Encrypted image subblock

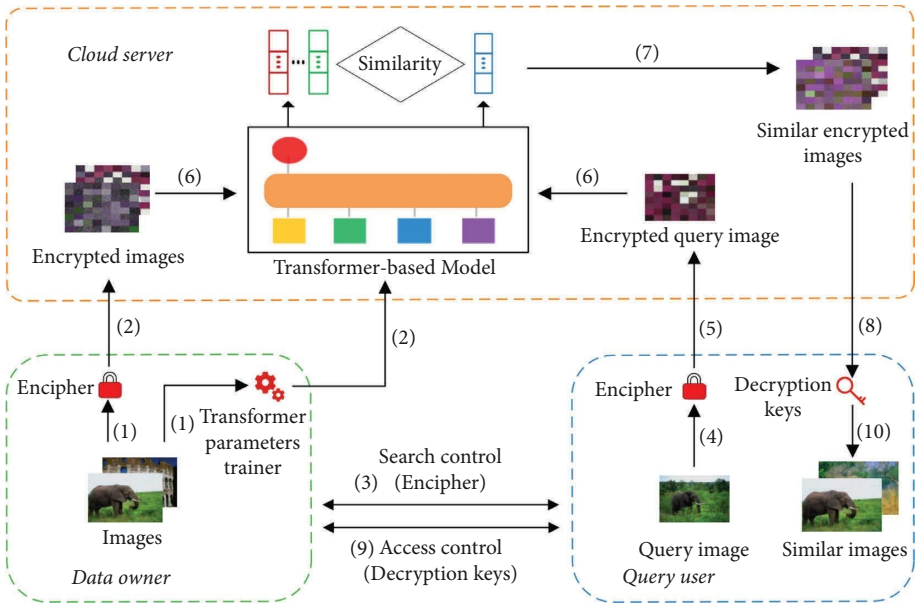


Figure 1: Encrypted image retrieval system in the cloud environment.

model in the cloud server does not need to be updated in real time with the database, which can better support the retrieval under a real-time update database.

3.2. Threat Model. Our method focuses on protecting the privacy of images used in a cloud environment, mainly against security issues caused by cloud servers. Similar to many existing works, cloud servers are assumed to be "honest-but-curious" [19]. The cloud server can execute commands correctly according to the protocol, but it may record and analyze retrieved information and encrypted image databases. The data owner and the querying user are fully trusted and secure objects [45], and there is no leakage or illegal distribution of plaintext images, and attackers cannot get plaintext data. In addition, entity objects use secure channels to communicate without compromising any data. Therefore, we mainly consider the ciphertext-only attack (COA) and known-background attack (KBA) and do not consider the selective plaintext attacks such as differential attacks [48].

4. The Proposed Method

This section describes the proposed scheme in detail, via image encryption, feature extraction, and retrieval.

4.1. Image Encryption. To achieve the availability of local histogram features and the security of global information in image encryption, we propose a secure step-by-step image encryption method, processing image local and global

information in steps and processing image texture and color information separately. The encryption process is shown in Figure 2.

Our encryption scheme is divided into the following steps: first, the image is divided into nonoverlapping image subblocks. Next, the positions of the RGB channel values within the subblocks are randomly scrambled to protect the local texture information. Then, the global texture information is protected by randomly scrambling the positions between subblocks. Finally, the RGB channel values of the encrypted image subblocks are replaced, the RGB channels are swapped to protect the global color information, and the value substitution of different subblocks is fixedly related to the position of the encrypted subblocks. A random function generates the encryption key. The detailed image encryption algorithm is shown in Algorithms 1–3.

4.2. Transformer-Based Feature Extractor. This study also proposes a transformer-based feature extraction model that can extract features directly from encrypted images to support ciphertext retrieval, as shown in Figure 3.

The feature extraction model consists of three parts: image block, histogram extraction, and vision transformer encoder.

4.2.1. Image Block. We divide an encrypted image $E \in \mathbb{R}^{H\times W\times 3}$ into m^2 nonoverlapping encrypted subblocks $\left\{\left\{\mathrm{sub}E_{(i,j)}\right\}_{i=0}^{m-1}\right\}_{j=0}^{m-1} \in \mathbb{R}^{W/m\times H/m\times 3}$, where (H, W) is the resolution of the original encrypted image, $(H/m, W/m)$ is the resolution of each encrypted subblock, (i, j) represents the position of the encrypted subblock in the original encrypted image, and the values of i and j are in the range $[0, m-1]$.

4.2.2. Histogram Extraction. First, the histograms of RGB three channels $\mathrm{sub}H_{(i,j)} = (\mathrm{sub}^R_{(i,j)}, \mathrm{sub}^G_{(i,j)}, \mathrm{sub}^B_{(i,j)})$ are extracted separately for the encrypted subblock $\left\{\left\{\mathrm{sub}E_{(i,j)}\right\}_{i=0}^{m-1}\right\}_{j=0}^{m-1}$, where $\mathrm{sub}H^R_{(i,j)} = \left\{\mathrm{bin}^R_u\right\}_{u=0}^{255}$, $\mathrm{sub}H^G_{(i,j)} = \left\{\mathrm{bin}^G_u\right\}_{u=0}^{255}$, $\mathrm{sub}H^B_{(i,j)} = \left\{\mathrm{bin}^B_u\right\}_{u=0}^{255}$, and bin^R_u represent the number of histograms with the value u in the R channel. Then, the RGB histogram $(\mathrm{sub}H^R_{(i,j)}, \mathrm{sub}H^G_{(i,j)}, \mathrm{sub}H^B_{(i,j)})$ is subjected to channel transformation and histogram translation operations (equations (1) and (2)) to obtain a three-channel histogram $\mathrm{sub}H^{'}_{(i,j)} = (\mathrm{sub}H^{'R}_{(i,j)}, \mathrm{sub}H^{'G}_{(i,j)}, \mathrm{sub}H^{'B}_{(i,j)})$, $\mathrm{sub}H^{'}_{(i,j)} \in \mathbb{R}^{256\times 3}$, where $\mathrm{sub}H^{'R}_{(i,j)} = \left\{\mathrm{bin}^{'R}_u\right\}_{u=0}^{255}$, $\mathrm{sub}H^{'G}_{(i,j)} = \left\{\mathrm{bin}^{'G}_u\right\}_{u=0}^{255}$, $\mathrm{sub}H^{'B}_{(i,j)} = \left\{\mathrm{bin}^{'B}_u\right\}_{u=0}^{255}$. Finally, the RGB three channels of the subblock are stitched to obtain the local features $\left\{\mathrm{sub}F_i\right\}_{i=1}^{m^2}$, $\mathrm{sub}F_i \in \mathbb{R}^{768}$. Histogram extraction outputs local features $\left\{\mathrm{sub}F_i\right\}_{i=1}^{m^2}$, $\mathrm{sub}F_i \in \mathbb{R}^{768}$.

$$\mathrm{sub}H^{'}_{(i,j)} = \begin{cases} \mathrm{sub}H^{'G}_{(i,j)} = \vartheta\left(\mathrm{sub}H^{R}_{(i,j)}, \dfrac{256i}{m}\right), \\ \mathrm{sub}H^{'B}_{(i,j)} = \vartheta\left(\mathrm{sub}H^{G}_{(i,j)}, \dfrac{256j}{m}\right), \\ \mathrm{sub}H^{'R}_{(i,j)} = \vartheta\left(\mathrm{sub}H^{B}_{(i,j)}, \dfrac{128\,(i+j)}{m}\right), \end{cases} \tag{1}$$

$$\vartheta\left(\left\{\mathrm{bin}_u\right\}_{u=0}^{255}, \Delta\mathrm{bin}\right) = \begin{cases} \mathrm{bin}_u \longrightarrow \mathrm{bin}^{'}_{u-\Delta\mathrm{bin}}, \Delta\mathrm{bin} \le u, \\ \mathrm{bin}_u \longrightarrow \mathrm{bin}^{'}_{u-\Delta\mathrm{bin}+256}, \Delta\mathrm{bin} > u, \end{cases} \tag{2}$$

where bin_u is the input histogram, ϑ represents the histogram translation operation, $\Delta\mathrm{bin}$ is the displacement length of the histogram translation, and bin'_u is the output histogram.

4.2.3. Vision Transformer Encoder. We use a transformer encoder [27] to learn local features $\mathrm{sub}F = \left\{\mathrm{sub}F_i\right\}_{i=1}^{m^2}, \mathrm{sub}F \in \mathbb{R}^{m^2\times 768}$. We add a special sequence $[S]$ that can be learned before the local feature sequence and feed the new sequence $Z = [[S], \mathrm{sub}F_1, \ldots, \mathrm{sub}F_{m^2}]$ into the transformer encoder. Our transformer encoder stacks 12 identical blocks, each consisting of a multihead self-attention and a two-layer multilayer perceptron. Layer normalization is performed before each module, and residual concatenation is performed after the block to prevent gradient disappearance and speed up convergence $[S]^{'}$ in the sequence $Z^{'} = [[S]^{'}, \mathrm{sub}F'_1, \ldots, \mathrm{sub}F'_{m^2}]$ output by transformer encoder which is used as the representational feature f_e of the encrypted image.

The superiority of our feature extraction model comes from the following three points. First, local features within subblocks can be extracted stably and with stable feature dimensionality. Second, the perturbation invariance of the self-attentive mechanism in the transformer makes it possible to learn disordered features. Third, the transformer architecture has better long-range feature capture capability and better feature characterization capability and robustness.

4.3. Similar Image Search. In this paper, feature extraction and similarity metrics are performed in a cloud environment, which dramatically reduces the computational burden on data owners and query users. As with existing methods [25, 45], the Euclidean distance D_{Q-E} between query image features F_{EQ} and image database features $F_E = \left\{f_{e_1}, f_{e_2}, \ldots, f_{e_n}\right\}$ is calculated in the cloud environment to measure the similarity of encrypted images (equation (3)), and the retrieval results are returned according to the similarity ranking.

$$D_{Q-E} = \left\|F_{\mathrm{EQ}} - f_{e_i}\right\|, \text{where } i = 1, 2, \ldots, n. \tag{3}$$

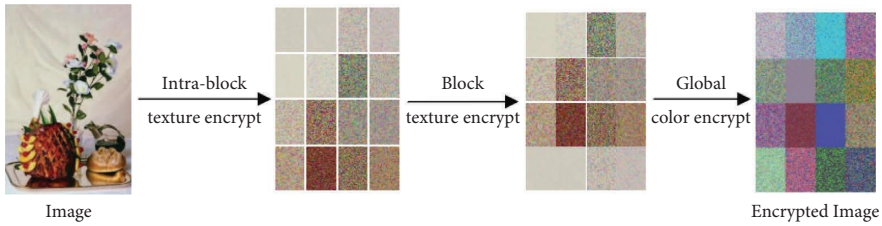


Figure 2: Image encryption process.

Input: Original image I of size $w \times h \times 3$.
Output: Encrypted image E, Image key K.
(1) Divide Original image $I \longrightarrow$ subblocks sub$I = \{\text{sub}I_1, \text{sub}I_2 \ldots, \text{sub}I_{m^2}\}$, size of $w/m \times h/m \times 3$.
(2) **for** $\forall$ sub$I_i \in$ sub$I$ **do**
(3) subE_i, subK_i = Intrablock_texture(subI_i)
(4) **end for**
(5) **for** i in reversed(range(m^2)) **do**
(6) random$(0, i) \longrightarrow$ num $\longrightarrow$ img_sub_key$[i]$
(7) sub$E_{[\text{num}]}$, sub$E_{[i]} \longrightarrow$ sub$E'_{[i]}$, sub$E_{[\text{num}]}$
(8) sub$K_{[\text{num}]}$, sub$K_{[i]} \longrightarrow$ sub$K_{[i]}$, sub$K_{[\text{num}]}$
(9) **end for**
(10) **for** $\forall$ sub$E'_i \in$ sub$I$ **do**
(11) subE''_i = ColorEncryption(subE'_i)
(12) **end for**
(13) Combine the subblocks subE''_i to get an Encrypted image E.
(14) Combine subK' and img_sub_key to get Image key K.

Algorithm 1: Image encryption.

Input: Subblock sub$I$ of size $w/m \times h/m \times 3$.
Output: Encrypted subblock sub$E$, subblock key sub$K$.
(1) Divide RGB channel of sub$I \longrightarrow (l_r, l_g, l_b)$ of size $(w/m \times h/m)$
(2) **for** i in reversed(range($w/m \times h/m$)) **do**
(3) random$(0, i) \longrightarrow r \longrightarrow$ sub_img_r_key$[i]$
(4) random$(0, i) \longrightarrow g \longrightarrow$ sub_img_g_key$[i]$
(5) random$(0, i) \longrightarrow b \longrightarrow$ sub_img_b_key$[i]$
(6) $l_{r[\text{r}]}, l_{r[i]} \longrightarrow l_{r[i]}, l_{r[\text{r}]}$
(7) $l_{g[\text{g}]}, l_{g[i]} \longrightarrow l_{g[i]}, l_{g[\text{g}]}$
(8) $l_{b[\text{b}]}, l_{b[i]} \longrightarrow l_{b[i]}, l_{b[\text{b}]}$
(9) **end for**
(10) (l_r, l_g, l_b) resize and put into RGB channel of sub$E$.
(11) (sub_img_g_key, sub_img_b_key, sub_img_r_key) put into sub$K$.

Algorithm 2: Intrablock texture encryption.

Input: Subblock sub$E'_{[i]}$ of size $w/m \times h/m \times 3$.
Output: Encrypted subblock sub$E''_{[i]}$.
(1) Divide RGB channel of sub$E'_{[i]} \longrightarrow (e_r, e_g, e_b)$ of size $(w/m \times h/m)$
(2) $e'_b = (e_r + i/m \times 256/m)\%256$
(3) $e'_r = (e_g + (i\%m) \times 256/m)\%256$
(4) $e'_g = (e_b + (i/m + i\%m) \times 128/m)\%256$
(5) (e'_r, e'_g, e'_b) resize and put into RGB channel of sub$E''_{[i]}$.

Algorithm 3: Color encryption.

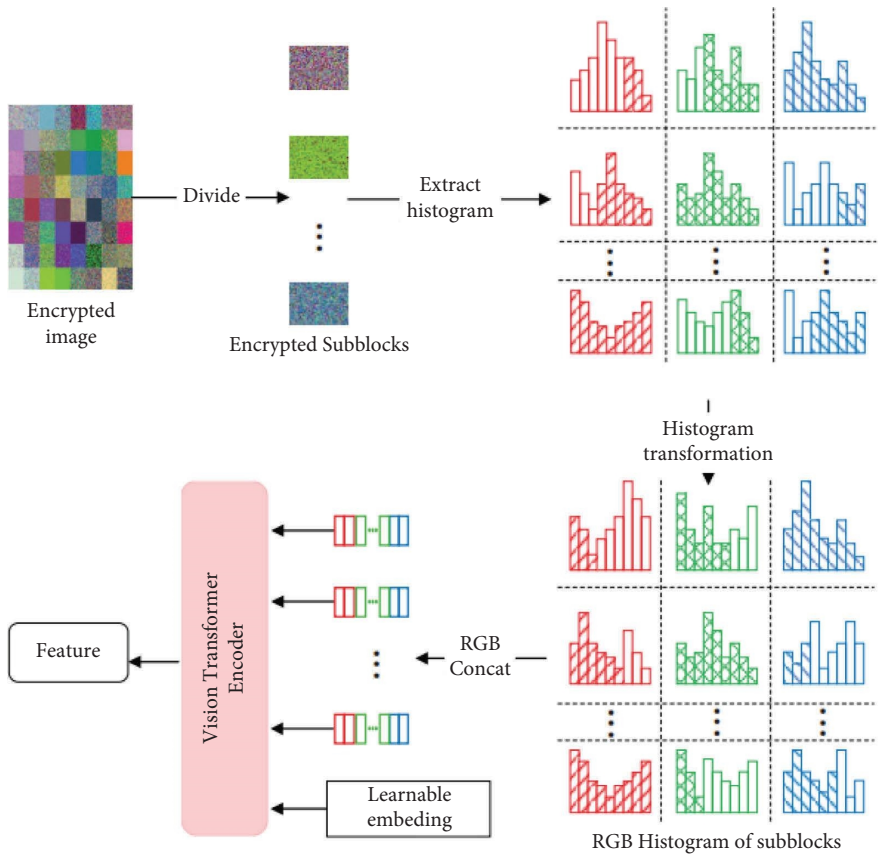


Figure 3: Transformer-based feature extraction model.

5. Security Analysis

In this section, we will analyze the security of the proposed scheme through image security and image feature security. Algorithm 4 formalizes the desired functionality and leakage information of this scheme.

5.1. The Privacy Security of Image Content. We will analyze the security of the image content in terms of both visual security and security strength. We encrypt the position and value of the image with random dislocation and difference substitution to protect the texture and color information of the image. A visual comparison of encrypted images of different schemes is shown in Figure 4.

Compared with the original image (a), pixel encryption within ordered subblocks (c) can protect local information but still leaks global texture and color information because the global texture information is not encrypted. Global pixel encryption (b) and unordered subblock pixel encryption (d) can protect global texture information better but still leak global color statistics. Our scheme (e) can better protect texture and color information. The more the number of chunks in our encryption scheme, the better the visual encryption effect. The visual effect of different numbers of blocks in our encryption schemes is shown in Figure 5.

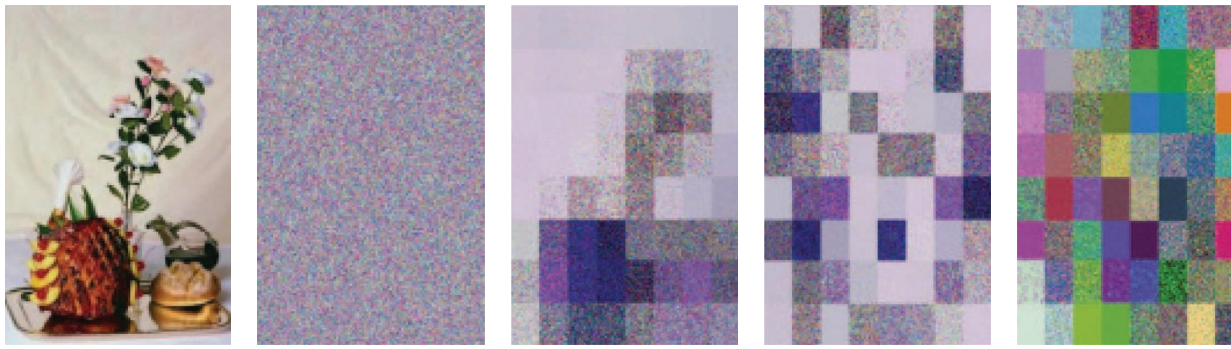
Under the COA, the attacker can only obtain E. For an $E \in \mathbb{R}^{H\times W\times 3}$ with m^2 subblocks, the security strength for intrablock RGB channel value alignment is $\log_2 (\mathrm{HW}/m^2!)^3$ bits, the security strength for interblock alignment is $\log_2 m^2!$ bits, and the security strength for RGB channel value replacement is $H \times W \times 3 \times \log_2 256$ bits. Therefore, the security strength of the encrypted image in our scheme is $(\log_2 (\mathrm{HW}/m^2!)^3 + \log_2 m^2! + 48\mathrm{HW})$ bits under the COA model.

Under the KBA, the attacker can obtain some statistical information about the natural image in addition to E. The color values of natural images do not occur uniformly, and there is a probabilistic distribution of differences. Our scheme performs value substitution for the RGB three-channel values within a subblock. The value substitution is closely related to the position of the subblock in which it is located, which transfers the randomness of the position to the RGB channel values in the form of differential rotation to a certain extent, thus making the color histogram of the image homogeneous. The color histogram of the encrypted image with different blocks is shown in Figure 6.

In terms of visual security and security strength, the higher the number of blocks, the higher the security of the image content. However, there is a trade-off between security, retrieval accuracy, and retrieval efficiency in ciphertext image retrieval. The scheme in this paper focuses more on extracting features directly from encrypted images for efficient retrieval, thus reducing computational consumption on the user side to support mobile devices and real-time updated databases.

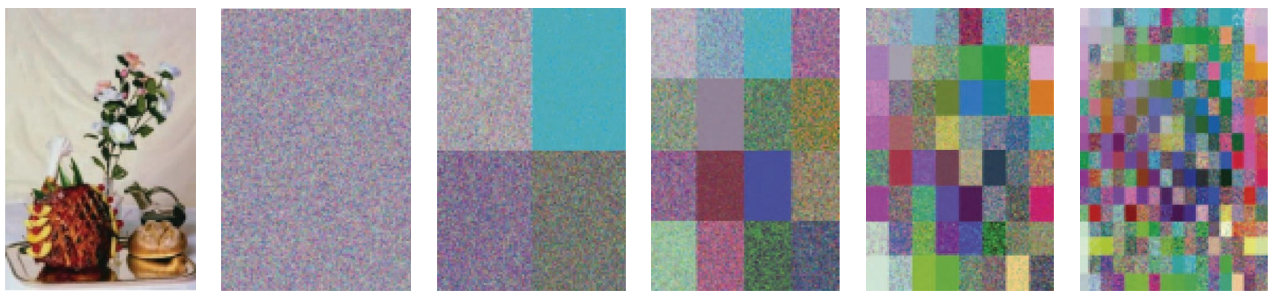
(1) **Γ. StoreImg (E):**
Functionality. The data owner uploads E to the cloud server.
Leakage. The total number of images, resolution of each image, and encrypted images E.
(2) **Γ. Feature ($\mathbf{F_E}$, $\mathbf{F_{EQ}}$):**
Functionality. Cloud server extracts F_E, F_{EQ} from E, EQ.
Leakage. Encrypted global histogram feature, F_E, F_{EQ}, and the similarities between F_E, F_{EQ} and the frequent distribution information of the encrypted global histogram.
(3) **Γ. Search (EQ, ER):**
Functionality. Query user sends EQ to cloud server and gets ER.
Leakage. EQ, ER.

Algorithm 4: The ideal functionality Γ and all leakage information of the proposed scheme.



(a) (b) (c) (d) (e)

Figure 4: Visual comparison of encrypted images of different schemes. (a) The original image, (b) the global pixel encrypted image, (c) pixel encryption within ordered subblocks, (d) unordered subblock pixel encryption, and (e) the scheme proposed in this paper.



(a) (b) (c) (d) (e) (f)

Figure 5: Visual comparison of our encryption scheme with different numbers of subblocks. (a) The original image and (b–f) the encrypted images in 1, 2^2, 4^2, 8^2, and 16^2 subblocks, respectively.

5.2. The Privacy Security of Image Features. In our scheme, the global histogram of the encrypted images tends to be average (Figure 6). Due to the deep fitting between the encryption algorithm and the model, it is difficult for the untrained model to characterize the encrypted data. However, the weak interpretability of deep learning makes the trained transformer model unable to be reversed parsed, so the attacker cannot learn the connection between histogram features and retrieval features through encrypted image retrieval features and model reverse. In addition, the attacker may analyze the correlation between the ciphertext feature representations, and the feature dimension of ViT-EIR is 768, so the computational complexity required for analysis is (768!), and the security strength of the feature is $\log_2(768!) > 512$, so the safety strength is greater than 512 bits.

6. Performance Evaluation

In this section, the proposed scheme in this paper is experimentally evaluated in terms of retrieval accuracy and efficiency. Experiments on the data owner and cloud server

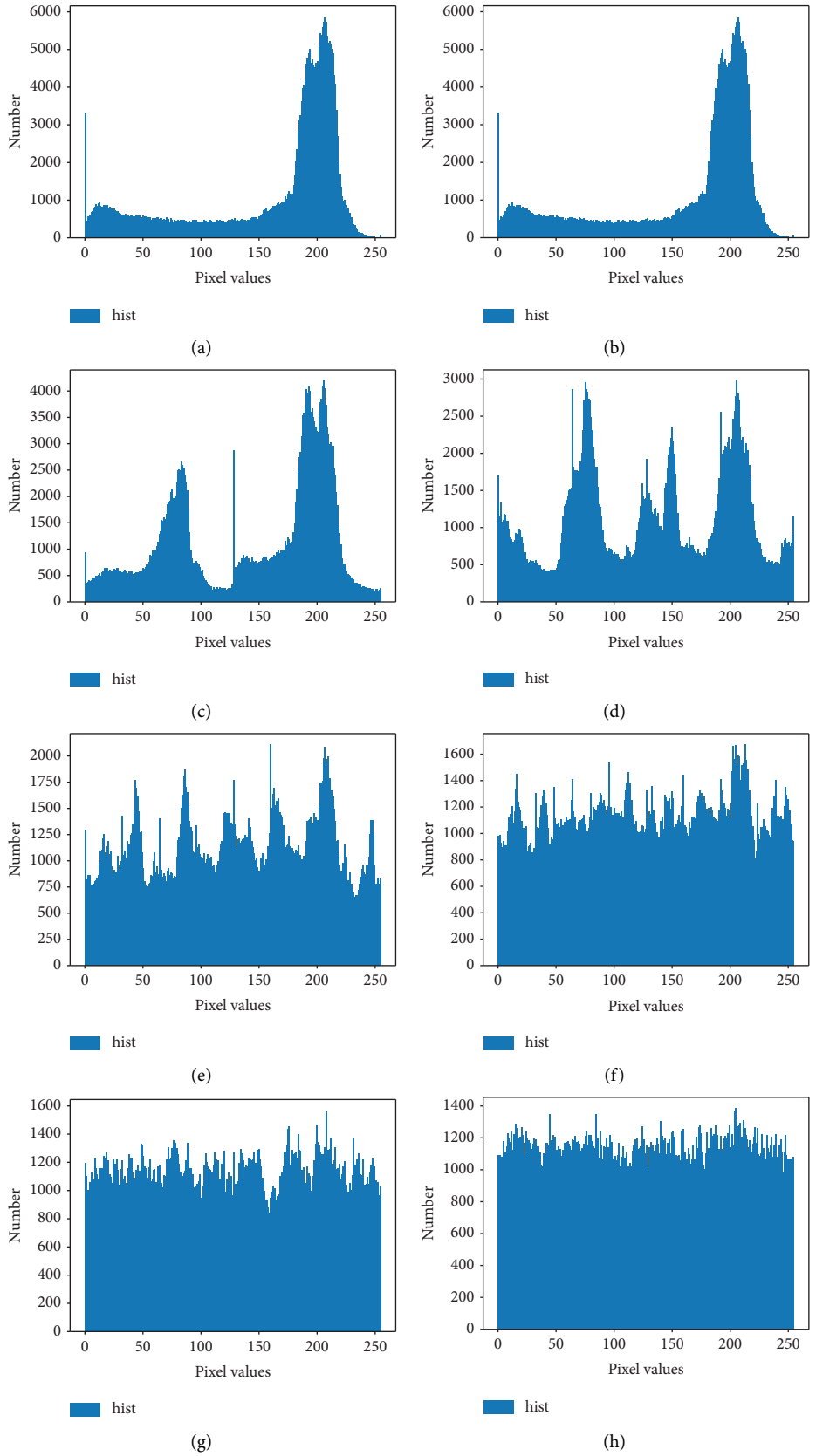


Figure 6: Color histogram comparison of encrypted images with different subblocks. (a) The original image and (b–h) the encrypted images in 1, 2^2, 4^2, 8^2, 16^2, 32^2, and 64^2 subblocks, respectively.

side were conducted on a Tesla V100 (32 G) GPU, and experiments on the query user side were conducted on an Intel i7-7700HQ (2.80 GHz) CPU. The transformer part of the feature extraction model is fine-tuned using the pretraining parameters of ViT [27] on the ImageNet dataset. The batch size is 64, and the initial learning rate is 0.001. The code link will be https://github.com/one-zd/PPIR-DLHVT.

6.1. Datasets and Metrics. The proposed scheme is experimentally evaluated on the following two mainstream datasets to facilitate comparison with similar schemes:

Corel 10k [49] consists of 100 categories, each with 100 color images with a resolution of 384×256 or 256×384. We randomly select 20 images from each category as query images (2000 images in total) and the remaining 8000 images as database images, respectively.

INRIA Holidays [50] consists of 500 groups of 1491 color images with image resolutions ranging from 480×640 to 3888×2592, with high-resolution images making up the majority. We select the first one of each group as the query image (500 images in total) and the remaining 991 images as the database images.

The evaluation metrics are consistent with similar schemes: the precision evaluation metric on the Corel dataset is the precision of returning k query results, and the precision evaluation metric on INRIA Holidays is MAP. The steps to calculate MAP include sorting the retrieval results for each query image, calculating precision and average precision, and calculating the average precision of all query images as MAP values. Besides, the efficiency evaluation metrics include encryption time, feature extraction time, feature matching time, and total retrieval time.

6.2. Experiments on Corel 10k

6.2.1. Retrieval Accuracy. Similar pure image encryption retrieval schemes evaluated under the Corel 10k dataset are Xia et al.'s [51] CDCBIR, Xia et al.'s [52] EPCBIR, Qin et al.'s [22] Har-EIR, and Wang et al.'s [25] TTCBIR. We performed a retrieval accuracy comparison, and the experimental results are shown in Figure 7, where Hist represents the scheme that uses only global RGB histogram features for retrieval.

The retrieval accuracy of our proposed scheme is nearly 40 points higher than the global histogram feature (Hist), about 42 points higher than MPEG-7 CLD-based EPCBIR [52], about 30 points higher than MPEG-7 CSD-based CDCBIR [51], about 11 points higher than SURF and Harris-based Har-EIR [22], and nearly 5 points higher than CNN-based TTCBIR [25]. With the increase of Top-k, the accuracy of our scheme decreases slower than other schemes. The retrieval accuracy of our scheme outperforms existing pure image encryption retrieval schemes that use local features, global, and CNN features.

Existing CNN and transformer schemes cannot learn disordered local features to demonstrate the advantage of disordered local features in this scheme. We further compare with schemes [25] that use ResNet [53], ViT [27], and PVTV2 [54] networks to enhance the retrieval after global histogram features, and the results are shown in Table 2, where ViT-EIR stands for the proposed methods that abbreviate ViT-based encrypted image retrieval.

It can be seen that ViT-EIR outperforms schemes that directly augment global histogram features using ResNet, ViT, and PVTV2. In the global histogram feature enhancement scheme, the accuracy increases from ResNet18 to ResNet50 but decreases from ResNet50 to ResNet152. The reason is that the global statistical feature dimension is small and network degradation occurs when the network is too deep. ViT is about 4 points higher than ResNet, and PVTV2 with multiple feature reuses is 8 points higher than ViT, indicating that the normal transformer network is stronger than CNN for global histogram feature mining. The accuracy of ViT-EIR is 4 points higher than that of PVTV2, indicating that the unordered local feature enhancement scheme in this paper can better characterize the encrypted image than the global feature enhancement scheme.

In this scheme, the images are divided into subblocks of the same size without overlapping. When the number of subblocks increases, the number of local histogram features increases, the number of pixel points within each subblock decreases, and the amount of information in the local histogram decreases. Comparison of retrieval accuracy for different numbers of blocks of ViT-EIR on Corel 10k is shown in Table 3 where ViT-EIR-4, ViT-EIR-16, ViT-EIR-64, and ViT-EIR-256 represent the schemes with the different numbers of blocks proposed in this paper.

As the number of image blocks increases, the number of unordered local sequences input to the transformer increases, and the amount of information contained in each sequence decreases. When the number of blocks increases from 4 to 64, the transformer network can learn more sequences, and the best retrieval accuracy is achieved when the number of blocks in ViT-EIR is 64. However, when there are 256 blocks, the amount of information in each sequence is very small, which leads to a loss of accuracy.

6.2.2. Retrieval Efficiency. Similar pure image encryption retrieval schemes evaluated on Corel 10k are Xia et al.'s [51] CDCBIR, Xia et al.'s [52] EPCBIR, Qin et al.'s [22] Har-EIR, and Wang et al.'s [25] TTCBIR, with which the scheme in this paper is compared for retrieval efficiency, and the results are shown in Figure 8.

Compared with some schemes using traditional local features (e.g., CDCBIR [51], EPCBIR [52], and Har-EIR [22]), our scheme achieves similar single-image search times. It is worth noting that our scheme extracts feature directly from encrypted images under cloud servers, without index generation and construction. The efficiency advantage of our scheme is more obvious when the database is updated faster. Compared to TTCBIR [25], which uses a CNN model, the parallel processing of local features of images and the ViT architecture is more efficient, so the retrieval is more efficient.

We replace the used CNN and sum transformer to test the efficiency of the global feature enhancement scheme,

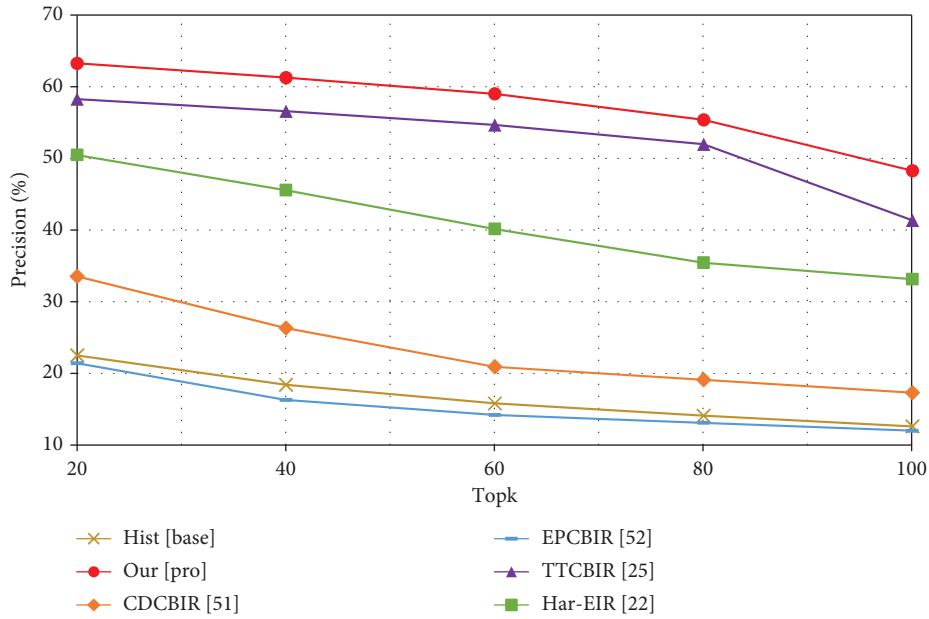


Figure 7: Retrieval accuracy comparison with different pure image encryption retrieval schemes on Corel 10k.

Table 2: Retrieval accuracy comparison with global histogram feature enhancement scheme on Corel 10K.

Network	Top-k (precision)						
	$k=1$	$k=5$	$k=10$	$k=20$	$k=40$	$k=60$	$k=80$
ResNet18	0.506	0.478	0.457	0.427	0.384	0.347	0.311
ResNet50	0.518	0.491	0.477	0.456	0.416	0.375	0.332
ResNet101	0.512	0.494	0.473	0.449	0.419	0.393	0.356
ResNet152	0.488	0.487	0.475	0.453	0.420	0.389	0.354
ViT-B	0.555	0.523	0.501	0.474	0.435	0.401	0.368
PVTV2	0.633	0.607	0.594	0.581	0.557	0.533	0.513
ViT-EIR (ours)	**0.675**	**0.656**	**0.645**	**0.631**	**0.611**	**0.589**	**0.553**

Bold values indicate the optimal performance in the comparison scheme.

Table 3: Retrieval accuracy comparison under different numbers of blocks of ViT-EIR on Corel 10k.

Number of blocks	Top-k (precision)						
	$k=1$	$k=5$	$k=10$	$k=20$	$k=40$	$k=60$	$k=80$
ViT-EIR-4	0.584	0.556	0.543	0.521	0.490	0.459	0.421
ViT-EIR-16	0.621	0.596	0.582	0.559	0.527	0.497	0.458
ViT-EIR-64	**0.675**	**0.656**	**0.645**	**0.631**	**0.611**	**0.589**	**0.553**
ViT-EIR-256	0.668	0.643	0.628	0.611	0.586	0.562	0.525

Bold values indicate the optimal performance in the comparison scheme.

where the time is the total time for 2000 retrievals, and the results are shown in Figure 9.

Compared to ResNet and ViT networks, ViT-EIR-4 and ViT-EIR-16 have shorter feature enhancement times. The method in this paper increases the complexity of the transformer as the number of chunks increases, so the feature extraction time also increases. The output feature dimension is the same for different chunk number schemes, so the feature matching time is almost unchanged. Combined with the retrieval accuracy, the ViT-EIR-64 scheme has the best retrieval accuracy and better retrieval efficiency on Corel 10k.

6.2.3. Feature Visualization. We compared 2D and 3D visualization of features including global histogram features (Hist), features enhanced by CNN for global histogram (CNN), and features enhanced by a transformer for unordered chunked histogram (Our), and the results are shown in Figure 10.

Hist features of different categories are basically clustered together, and it is impossible to distinguish the categories with better security. The intraclass distance of CNN-enhanced features decreases, so the retrieval accuracy is higher than that of Hist features, but there is an overlap of features from different classes. Our scheme features smaller

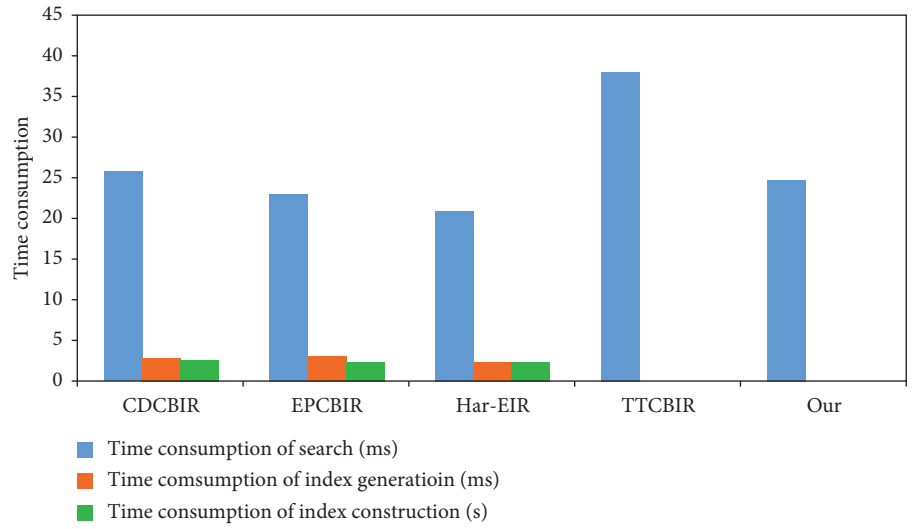


Figure 8: Retrieval efficiency comparison to state-of-the-art schemes on Corel 10k.

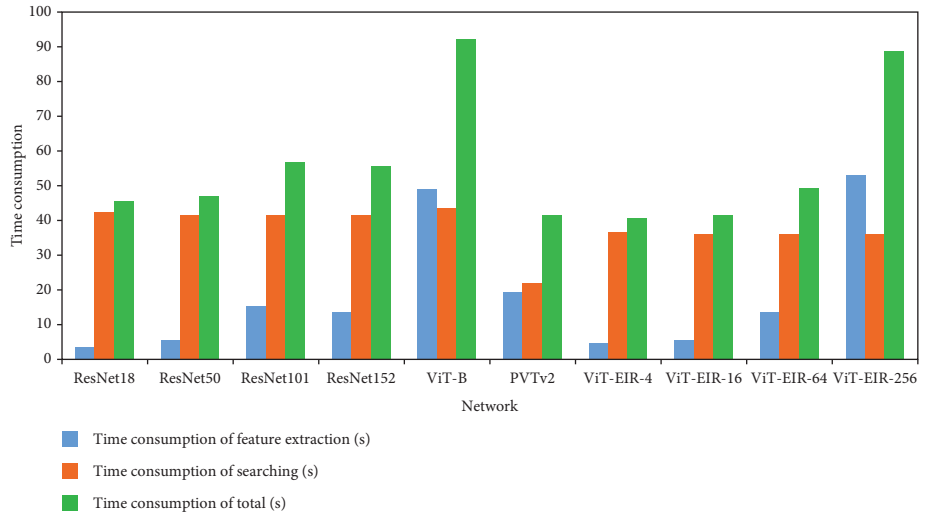


Figure 9: Retrieval time comparison of different networks on Corel 10k.

intraclass distances and larger interclass distances and thus has better retrieval accuracy.

6.3. Experiments on INRIA Holidays

6.3.1. Retrieval Accuracy. The pure image encryption retrieval methods experimented on INRIA Holidays are the methods of Lu et al. [55], Ferreira et al. [20], Liu et al. [41], Xia et al. [42], and Xia et al. [19], and our dataset partitioning is the same as these schemes. However, we did not have access to the authors' source code. To ensure fairness, we only give the retrieval accuracy for reference. The results are shown in Table 4.

We compare the proposed scheme with the scheme that uses ResNet to enhance global features. The results are shown in Table 5.

It can be seen that on INRIA Holidays, the accuracy of our scheme is nearly 10 points higher than that of the scheme using ResNet-enhanced global features. We also test the retrieval accuracy under different numbers of blocks, and the results are shown in Table 6.

From 4 to 256 blocks, increasing the number of sequences in the transformer improves the retrieval accuracy as the number of blocks increases. At a block number of 256, the ViT-EIR scheme has the best retrieval accuracy. When the number of blocks is 1024, there is too little information in the local features, which leads to a decreasing trend of accuracy.

6.3.2. Retrieval Efficiency. We test the retrieval efficiency of the proposed scheme. The results are shown in Figure 11.

On INRIA Holidays, the time consumption of feature enhancement for ViT-EIR-4 and ViT-EIR-16 is similar to the global feature enhancement method using ResNet, due to the former having more subblock sequences and a more complex transformer section.

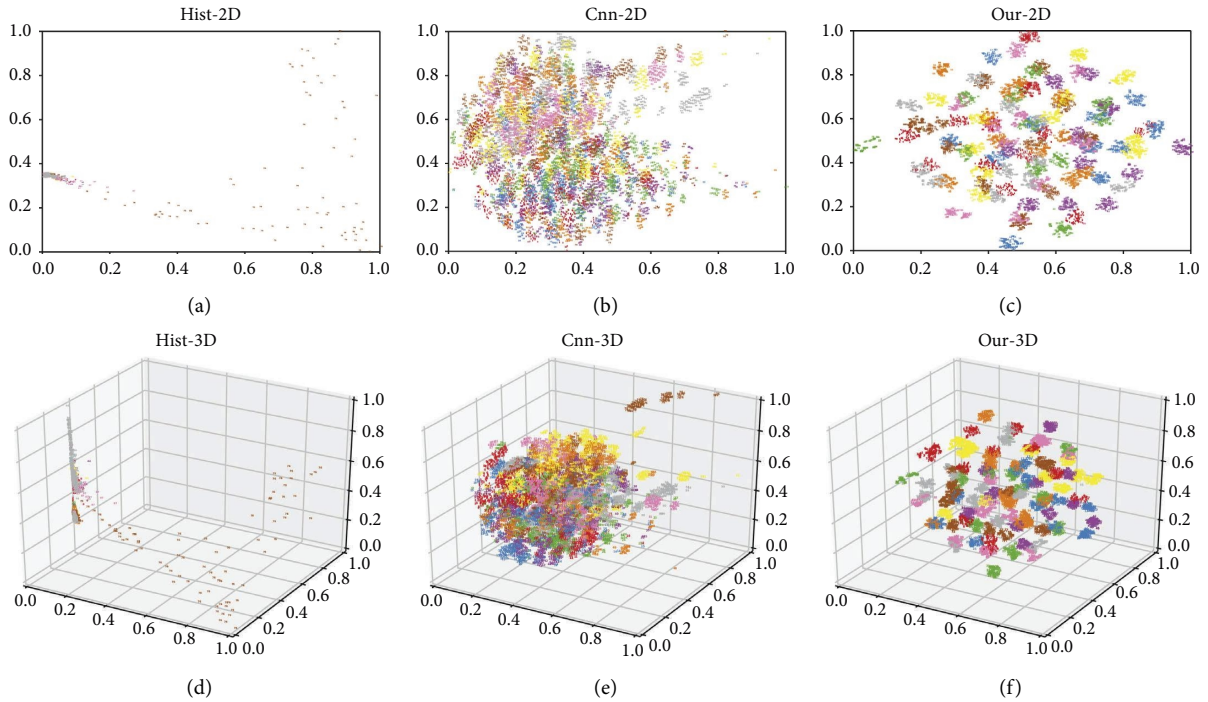


Figure 10: Feature visualization of the different schemes. (a–c) A 2D visualization plot and (d–f) a 3D visualization plot. (a, d) The visualization of Hist features, (b, e) the visualization of CNN features, and (c, f) the visualization of our features, and the numbers in the figure represent the labels of different categories.

Table 4: Retrieval accuracy comparison to state-of-the-art schemes on INRIA Holidays.

Scheme	Lu et al. [55]	Ferreira et al. [20]	Liu et al. [41]	Xia et al. [42]	Xia et al. [19]	Our
MAP	0.491	0.504	0.464	0.529	0.515	**0.578**

Bold values indicate the optimal performance in the comparison scheme.

Table 5: Retrieval accuracy comparison with global histogram feature enhancement scheme on INRIA Holidays.

Scheme	ResNet18	ResNet50	ResNet101	ResNet150	Our
MAP	0.485	0.408	0.372	0.403	**0.578**

Bold values indicate the optimal performance in the comparison scheme.

Table 6: Retrieval accuracy comparison under different numbers of blocks on INRIA Holidays.

Scheme	ViT-EIR-4	ViT-EIR-16	ViT-EIR-64	ViT-EIR-256	ViT-EIR-1024
MAP	0.473	0.487	0.495	**0.578**	0.568

Bold values indicate the optimal performance in the comparison scheme.

6.4. Time Consumption of Image Encryption. We test encryption time with different numbers of blocks and different image resolutions, and the results are shown in Figure 12.

For large-resolution images (2560×1920), the encryption time can be shortened by increasing the number of blocks (when the number of blocks is 256, the encryption time is nearly 7 seconds shorter than that of global encryption). For small-resolution images (384×256), the advantage of block encryption is not obvious. Because the encryption time is too short, computer performance fluctuations mask encryption time variations.

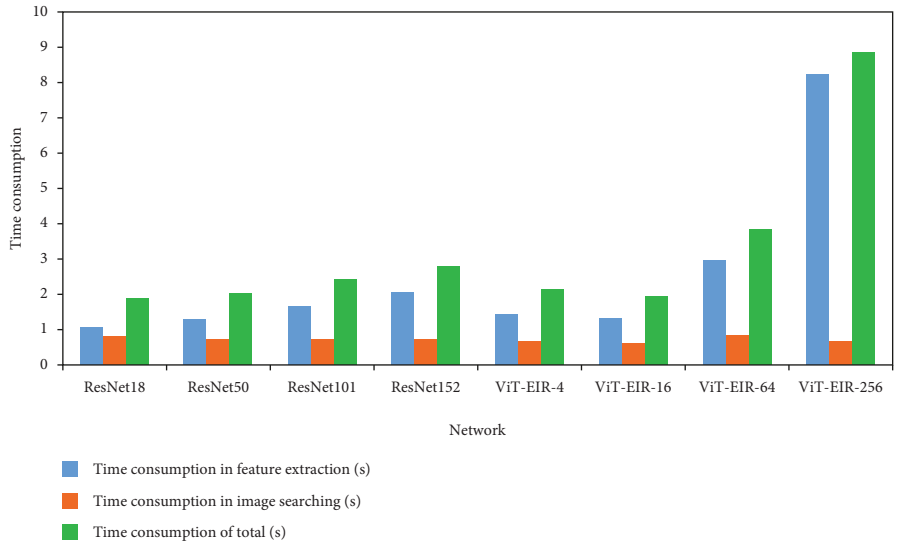


Figure 11: Retrieval time comparison of different networks on INRIA Holidays.

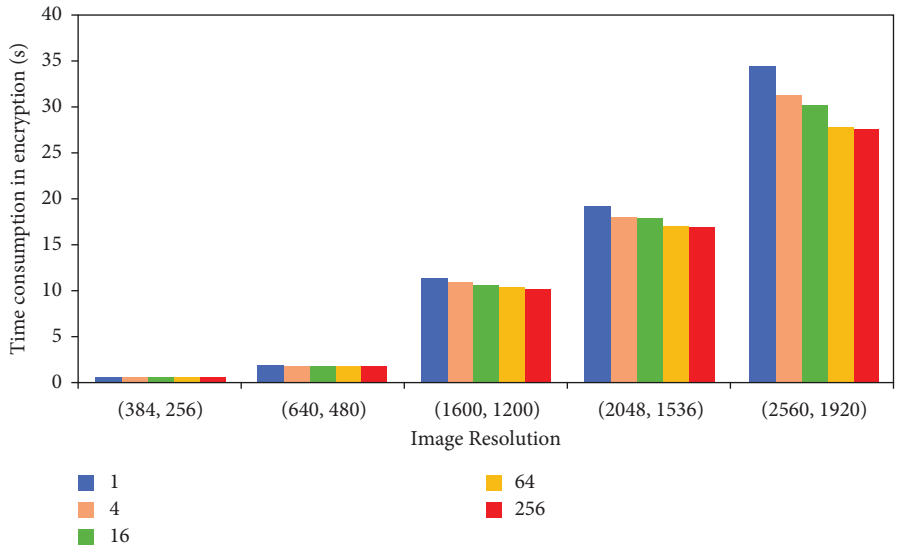


Figure 12: Encryption time comparison under different numbers of blocks and image resolutions.

7. Conclusion

This paper proposes a transformer-based encrypted image retrieval method for the cloud environment, shifting the complex computational operations to the cloud by taking full advantage of the cloud environment. It mainly uses secure step-by-step encryption to protect the texture and color information of images and build feature extraction models based on transformers to learn disordered local features from encrypted images. Experimental results demonstrate the superiority of this scheme in terms of retrieval accuracy and efficiency in realistic real-time update datasets and lightweight user-side environments.

For future work, we consider expanding the solution to cross-media encrypted retrieval to better leverage the value of multimedia data [56]. We will consider more encrypted retrieval in industrial scenarios such as IoT and sensors to improve the conversion rate of industry-academia research.

Data Availability

The Corel 10K data used to support the findings of this study have been deposited in the "https://www-db.stanford.edu/~wangz/image.vary.jpg.tar."

Conflicts of Interest

The authors declare that they have no conflicts of interest to report regarding the present study.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62372478 and in part by the Natural Science Foundation of Hunan Province under Grant 2022JJ31019.

References

[1] S. Meng, "Security-aware dynamic scheduling for real-time optimization in cloud-based industrial applications," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 6, pp. 4219–4228, 2021.

[2] L. Qi, "Privacy-aware data fusion and prediction with spatial-temporal context for smart city industrial environment," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 6, pp. 4159–4167, 2021.

[3] X. Xu, "PDM: privacy-aware deployment of machine-learning applications for industrial cyber–physical cloud systems," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5819–5828, 2021.

[4] I. Security, "Cost of a data breach report 2022," 2022, https://www.ibm.com/security/data-breach.

[5] D. Xia, "Research on the detection of privacy information sharing behaviour of ecommerce users based on big data," *International Journal of Autonomous and Adaptive Communications Systems*, vol. 15, no. 3, pp. 249–265, 2022.

[6] J. Li, "Searchable symmetric encryption with forward search privacy," *IEEE Transactions on Dependable and Secure Computing*, vol. 18, no. 1, pp. 460–474, 2021.

[7] F. Song, "Privacy-preserving keyword similarity search over encrypted spatial data in cloud computing," *IEEE Internet of Things Journal*, vol. 9, no. 8, pp. 6184–6198, 2022.

[8] Z. Xia, Y. Zhu, X. Sun, Z. Qin, and K. Ren, "Towards privacy-preserving content-based image retrieval in cloud computing," *IEEE Transactions on Cloud Computing*, vol. 6, no. 1, pp. 276–286, 2018.

[9] H. Zhai, S. Lai, H. Jin, X. Qian, and T. Mei, "Deep transfer hashing for image retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 742–753, 2021.

[10] F. Li and K. Wei, "Study on social network recommendation service method based on mobile cloud computing," *International Journal of Autonomous and Adaptive Communications Systems*, vol. 14, no. 4, pp. 398–408, 2021.

[11] Y. Yao and N. Xiong, "Privacy-preserving max/min query in two-tiered wireless sensor networks," *Computers & Mathematics with Applications*, vol. 65, no. 9, pp. 1318–1325, 2013.

[12] H. Cheng and Z. Xie, "Multi-step data prediction in wireless sensor networks based on one-dimensional CNN and bi-directional LSTM," *IEEE Access*, vol. 7, pp. 117883–117896, 2019.

[13] X. Xu, "Service offloading with deep Q-network for digital twinning-empowered internet of vehicles in edge computing," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 2, pp. 1414–1423, 2022.

[14] F. Song, "Privacy-preserving task matching with threshold similarity search via vehicular crowdsourcing," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 7, pp. 7161–7175, 2021.

[15] X. Ren, X. Zheng, H. Zhou, W. Liu, and X. Dong, "Contrastive hashing with vision transformer for image retrieval," *International Journal of Intelligent Systems*, vol. 37, no. 12, pp. 12192–12211, 2022.

[16] J. Zhou, M. Zheng, Z. Cao, and X. Dong, "Pvidm: privacy-preserving verifiable shape context based image denoising and matching with efficient outsourcing in the malicious setting," *Computers & Security*, vol. 88, Article ID 101631, 2020.

[17] C. Zhang, L. Zhu, S. Zhang, and W. Yu, "Tdhppir: an efficient deep hashing based privacy-preserving image retrieval method," *Neurocomputing*, vol. 406, pp. 386–398, 2020.

[18] M. Wang, W. Zhao, and Y. Li, "Encryption of ciphertext data in Internet of Things based on HECRT key management," *International Journal of Autonomous and Adaptive Communications Systems*, vol. 15, no. 2, pp. 166–183, 2022.

[19] Z. Xia, L. Wang, J. Tang, N. Xiong, and J. Weng, "A privacy-preserving image retrieval scheme using secure local binary pattern in cloud computing," *IEEE Transactions on Network Science and Engineering*, vol. 8, no. 1, pp. 318–330, 2021.

[20] B. Ferreira, J. Rodrigues, J. Leitao, and H. Domingos, "Practical privacy-preserving content-based retrieval in cloud image repositories," *IEEE Transactions on Cloud Computing*, vol. 7, no. 3, pp. 784–798, 2019.

[21] S. Hu, Q. Wang, J. Wang, Z. Qin, and K. Ren, "Securing SIFT: privacy-preserving outsourcing computation of feature extractions over encrypted image data," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3411–3425, 2016.

[22] J. Qin, H. Li, and X. Xiang, "An encrypted image retrieval method based on Harris corner optimization and lsh in cloud computing," *IEEE Access*, vol. 7, pp. 24626–24633, 2019.

[23] Z. Abduljabbar, H. Jin, and A. Ibrahim, "Privacy-preserving image retrieval in iot-cloud," in *Proceedings of the 2016 IEEE Trustcom/Bigdatase/Ispa*, pp. 799–806, Tianjin, China, August 2016.

[24] Y. Li, J. Ma, and Y. Miao, "Secure and verifiable multi-key image search in cloud-assisted edge computing," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 8, pp. 5348–5359, 2021.

[25] Z. Wang, J. Qin, X. Xiang, and Y. Tan, "A privacy-preserving and traitor tracking content-based image retrieval scheme in cloud computing," *Multimedia Systems*, vol. 27, no. 3, pp. 403–415, 2021.

[26] W. Ma, J. Qin, X. Xiang, Y. Tan, and Z. He, "Searchable encrypted image retrieval based on multi-feature adaptive late-fusion," *Mathematics*, vol. 8, no. 6, p. 1019, 2020.

[27] A. Dosovitskiy, L. Beyer, and A. Kolesnikov, "An image is worth 16x16 words: transformers for image recognition at scale," 2021, https://arxiv.org/abs/2010.11929.

[28] Y. Bai, J. Mei, A. L. Yuille, and C. Xie, "Are transformers more robust than cnns?," 2021, https://arxiv.org/abs/2111.05464.

[29] C.-Y. Hsu, C.-S. Lu, and S.-C. Pei, "Image feature extraction in encrypted domain with privacy-preserving sift," *IEEE Transactions on Image Processing*, vol. 21, no. 11, pp. 4593–4607, 2012.

[30] L. Zhang, T. Jung, and K. Liu, "Pic: enable large-scale privacy preserving content-based image search on cloud," *IEEE Transactions on Parallel and Distributed Systems*, vol. 28, no. 11, pp. 3258–3271, 2017.

[31] R. Bellafqira, G. Coatrieux, D. Bouslimi, and G. Quellec, "Content based image retrieval in homomorphic encryption domain," in *Proceedings of the 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2944–2947, Milan, Italy, November 2015.

[32] C. Guo, J. Jia, K.-K. R. Choo, and Y. Jie, "Privacy-preserving image search (ppis): secure classification and searching using convolutional neural network over large-scale encrypted medical images," *Computers & Security*, vol. 99, Article ID 102021, 2020.

[33] W. Lu, A. Varna, and M. Wu, "Confidentiality-preserving image search: a comparative study between homomorphic encryption and distance preserving randomization," *IEEE Access*, vol. 2, pp. 125–141, 2014.

[34] B. Cheng, L. Zhuo, Y. Bai, Y. Peng, and J. Zhang, "Secure index construction for privacy-preserving large-scale image retrieval," in *Proceedings of the 2014 IEEE Fourth International Conference on Big Data and Cloud Computing*, pp. 116–120, Sydney, NSW, Australia, February 2014.

[35] Q. Zou, J. Wang, J. Ye, J. Shen, and X. Chen, "Efficient and secure encrypted image search in mobile cloud computing," *Soft Computing*, vol. 21, no. 11, pp. 2959–2969, 2017.

[36] Y. Li, J. Ma, and Y. Miao, "Similarity search for encrypted images in secure cloud computing," *IEEE Transactions on Cloud Computing*, vol. 7161, p. 1, 2020.

[37] Y. Chen and L. Zhou, "KNN-BLOCK dbscan: fast clustering for large-scale data," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 6, pp. 3939–3953, 2021.

[38] Y. Li, J. Ma, and Y. Miao, "Dvrei: dynamic verifiable retrieval over encrypted images," *IEEE Transactions on Computers*, vol. 71, pp. 1–1769, 2021.

[39] J. Huang, "Accelerating privacy-preserving image retrieval with multi-index hashing," in *Proceedings of the 2022 IEEE/ACM 7th Symposium on Edge Computing (SEC)*, pp. 492–497, Seattle, WA, USA, December 2022.

[40] W. Lu, A. Swaminathan, A. Varna, and M. Wu, "Enabling search over encrypted multimedia databases," *Electronic Imaging*, vol. 7254, 2009.

[41] D. Liu, J. Shen, Z. Xia, and X. Sun, "A content-based image retrieval scheme using an encrypted difference histogram in cloud computing," *Information*, vol. 8, no. 3, pp. 96–13, 2017.

[42] Z. Xia, L. Lu, and T. Qiu, "A privacy-preserving image retrieval based on ac-coefficients and color histograms in cloud environment," *Computers, Materials & Continua*, vol. 58, no. 1, pp. 27–43, 2019.

[43] Z. Xia, L. Jiang, and X. Ma, "A privacy preserving outsourcing scheme for image local binary pattern in secure industrial internet of things," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 629–638, 2020.

[44] Z. Xia, Q. Ji, Q. Gu, C. Yuan, and F. Xiao, "A format-compatible searchable encryption scheme for jpeg images using bag-of-words," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 18, no. 3, pp. 1–18, 2022.

[45] W. Ma, T. Zhou, and J. Qin, "A privacy preserving content-based image retrieval method based on deep learning in cloud computing," *Expert Systems with Applications*, vol. 203, Article ID 117508, 2022.

[46] F. Zhou, S. Qin, R. Hou, and Z. Zhang, "Privacy preserving image retrieval in a distributed environment," *International Journal of Intelligent Systems*, vol. 37, no. 10, pp. 7478–7501, 2022.

[47] P. Yu, J. Tang, Z. Xia, Z. Li, and J. Weng, "A privacy-preserving JPEG image retrieval scheme using the local Markov feature and bag-of-words model in cloud computing," *IEEE Transactions on Cloud Computing*, vol. 11, pp. 1–12, 2023.

[48] J. Li, "Efficient and secure outsourcing of differentially private data publishing with multiple evaluators," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 1, pp. 67–76, 2022.

[49] J. Wang, J. Li, and G. Wiederhold, "Simplicity: semantics-sensitive integrated matching for picture libraries," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 9, pp. 947–963, 2001.

[50] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," *European Conference on Computer Vision*, vol. 5302, pp. 304–317, 2008.

[51] Z. Xia, X. Wang, and L. Zhang, "A privacy preserving and copy-deterrence content-based image retrieval scheme in cloud computing," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 11, pp. 2594–2608, 2016.

[52] Z. Xia, N. Xiong, A. Vasilakos, and S. Xingming, "Epcbir: an efficient and privacy-preserving content-based image retrieval scheme in cloud computing," *Information Sciences*, vol. 387, pp. 195–204, 2017.

[53] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, June 2016.

[54] W. Wang, E. Xie, X. Li, T. Lu, P. Luo, and L. Shao, "Pvtv2: improved baselines with pyramid vision transformer," 2021, https://arxiv.org/abs/2106.13797.

[55] W. Lu, A. Swaminathan, A. Varna, and M. Wu, "Enabling search over encrypted multimedia databases," *Media Forensics and Security*, vol. 7254, 2009.

[56] Z. Wang, J. Qin, X. Xiang, Y. Tan, and J. Peng, "A privacy-preserving cross-media retrieval on encrypted data in cloud computing," *Journal of Information Security and Applications*, vol. 73, Article ID 103440, 2023.