Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Human-Centred Encrypted JPEG Retrieval for Substation Inspection: A Framework and an Evaluation Protocol

Shilei Ding

dingshilei@sgepri.sgcc.com.cn

State Grid Electric Power Research Institute

Bin Li

State Grid Electric Power Research Institute

Xu Yang

State Grid Electric Power Research Institute

Hui Wang

State Grid Electric Power Research Institute

Xia Guo

State Grid Electric Power Research Institute

Gang Wang

State Grid Electric Power Research Institute

Kun Hu

State Grid Electric Power Research Institute

Hongyu Zhou

State Grid Electric Power Research Institute

Da Fang

State Grid Electric Power Research Institute

Research Article

Keywords: Encrypted JPEG retrieval, Privacy-preserving image search, Human-centred evaluation, Operator trust and workload, Smart substation inspection, Deep learning

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11570767/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Human-centric Intelligent Systems
MANUSCRIPT SUBMISSION TEMPLATE

Human-Centred Encrypted JPEG Retrieval for Substation Inspection: A Framework and an Evaluation Protocol

Article type: Research Article

Author: Shilei Ding*
Affiliation: State Grid Electric Power Research Institute, Nanjing, Jiangsu, 211100, China
Email: dingshilei@sgepri.sgcc.com.cn

Author: Bin Li
Affiliation: State Grid Electric Power Research Institute, Nanjing, Jiangsu, 211100, China

Author: Xu Yang
Affiliation: State Grid Electric Power Research Institute, Nanjing, Jiangsu, 211100, China

Author: Hui Wang
Affiliation: State Grid Electric Power Research Institute, Nanjing, Jiangsu, 211100, China

Author: Xia Guo
Affiliation: State Grid Electric Power Research Institute, Nanjing, Jiangsu, 211100, China

Author: Gang Wang
Affiliation: State Grid Electric Power Research Institute, Nanjing, Jiangsu, 211100, China

Author: Kun Hu
Affiliation: State Grid Electric Power Research Institute, Nanjing, Jiangsu, 211100, China

Author: Hongyu Zhou
Affiliation: State Grid Electric Power Research Institute, Nanjing, Jiangsu, 211100, China

Author: Da Fang
Affiliation: State Grid Electric Power Research Institute, Nanjing, Jiangsu, 211100, China
* Corresponding author.

Abstract

Substation inspection produces a rapidly growing volume of images that must be stored, searched and reused, yet these images disclose sensitive grid asset information and therefore cannot be handed to an untrusted cloud in plaintext. Content-based image retrieval (CBIR) over encrypted JPEG images reconciles this conflict, but existing schemes are evaluated almost exclusively by retrieval accuracy and file-size overhead; whether an inspector can actually use, understand and trust such a system has rarely been examined. This paper presents a human-centred encrypted JPEG retrieval framework for intelligent substations. The image owner encrypts each image by stream cipher and block scrambling in a format-compatible manner, so that the JPEG bitstream remains compressible and the file size is not expanded. The cloud server extracts the variable-length integer (VLI) code

lengths from the cipher-images and learns a fixed-length representation with a convolutional neural network; similarity is measured by Euclidean distance and the top-k cipher-images are returned to the operator. We report the encryption security under the chosen-plaintext and known-background-knowledge models, the retrieval accuracy and the computational cost, and we complement them with an evaluation protocol that specifies how the usability and trust consequences of encryption are to be measured with inspection personnel, namely task completion time, perceived workload (NASA-TLX), usability (SUS) and trust in the system, together with a calibration index that relates reported trust to observed retrieval precision. No operator data are reported here: the protocol is stated in advance of data collection so that it can be reviewed and replicated. On this basis we discuss the privacy–utility–human trade-off and the deployment constraints of power-system compliance, and we argue that operator-centred evidence should complement accuracy-oriented benchmarks for encrypted retrieval.

Keywords: Encrypted JPEG retrieval; Privacy-preserving image search; Human-centred evaluation; Operator trust and workload; Smart substation inspection; Deep learning

1. Introduction

With the rapid development of imaging equipment, especially smart mobile devices, a large number of images are generated daily in substations. The images can contain various information and have been playing very important roles in all the fields of our social life, especially in substations. In order to find the wanted image from huge image databases, many image retrieval frameworks have been designed to search similar images to a query, among which the deep learning based frameworks have become the mainstream solutions [1]. However, excessive image data often results in large storage and computation overheads. Thus, it is attractive to outsource the image storage and image-relevant services to cloud servers which can provide on-demand storage and computing resources. In this way, the image owner need not store the image database locally while the image can be retrieved directly from the cloud server. For the images captured in substations, a federated learning framework has been developed to segment the terminal blocks without uploading the raw images out of the substations [2], which shows that the privacy of substation images has already become a practical concern.

When it comes to outsourcing images to the cloud server, privacy is a top concern to the image owner. Images can contain various sensitive information, such as private location, educational background, family members, and personal interests. And the cloud server is never fully secure and trustworthy to the image owner. Recent studies on cyber-physical systems, which smart substations essentially belong to, also show that the attacks against the sensing and computing units keep evolving and can hardly be eliminated by the service provider alone [3]. First, hackers are more interested in the information on the cloud server, so the cloud server is always the key target of hacker attacks. For example, Sony's PlayStation server was hacked in 2011. The service was interrupted for more than a week, and personal information of 77 million accounts on the network was stolen. Next, the cloud server cannot be fully trusted as they may also be interested in the information of the image owner. A straightforward way to protect sensitive images is to encrypt the image by standard cryptographic methods. However, in such a way the possible uses of the stored images would be strongly limited together with the services provided to users, such as image retrieval. Many privacy-preserving image retrieval frameworks have been proposed, and they can be

divided into two categories. The first category can be named as the feature-encryption based image retrieval framework. In this kind of framework, the image owner extracts features from the plaintext image. Then, the images and the extracted features are encrypted and uploaded to the server, respectively. The key point of this kind of framework is to make sure that the similarity between images can be measured by these encrypted features. The other category is the image-encryption based framework. In the frameworks of this category, the images are encrypted by the image owner with some specially-designed encryption methods. Then, the feature extraction together with the search operation will be outsourced to the cloud server so as to further relieve the burden of the image owner. Generally, the second category of framework is the more desirable one since it outsources more tasks to the cloud server. But the existing frameworks of the second category still have many limitations in terms of security, storage expansion and retrieval accuracy.

Contribution. In this paper, we propose a privacy-preserving CBIR framework, especially for JPEG images. The main contributions of our framework can be summarized as follows:

(1) A format-compatible encryption method is proposed for JPEG images without any storage expansion. Firstly, all the Discrete Cosine Transformation (DCT) coefficients are protected by encrypting the binary code in VLI coding. Secondly, the quantization tables are also protected by the stream cipher. Finally, the positions of DC coefficients are further permuted. The proposed image encryption method can protect the image content and leads to no storage expansion. The effective features can be learned from the encrypted images without knowing image content.

(2) For the first time, we applied the deep learning algorithm to learn features from the intermediate data which is parsed from the encrypted images. Then, the Euclidean distance between the features is calculated to measure the similarity of images. Due to the powerful learning capability of the neural network, the proposed framework achieves a good retrieval accuracy.

The rest of this paper is organized as follows. Section 2 introduces the related research work. Section 3 presents an overview and technical knowledge of our framework. In Section 4, we describe the process of the proposed framework in detail. We perform the security analysis of the framework in Section 5 and give a search performance result in Section 6. In the last Section 7, we make a summary of the framework.

2. Related Work

CBIR aims to find some similar images to a query image in real life [1]. Generally, the similarity between the images is judged by calculating the distance of the feature vectors specially extracted from the images. In the privacy-preserving CBIR framework, both the image and its features should be properly protected. The existing privacy-preserving CBIR frameworks can be divided into two categories, i.e., feature-encryption based image retrieval frameworks and image-encryption based image retrieval frameworks.

In the feature-encryption based frameworks, the image owner needs to extract features from the original image at first. Then, the images and the features are encrypted respectively. Lu et al. [4] propose to calculate the occurrence frequency of visual words to characterize images. The order-preserving encryption and Min-Hash algorithm are used to encrypt the occurrence frequency. Then, the Jaccard distance between the encrypted features is calculated to measure the similarity of images.

Afterward, Lu et al. [5] proposed another three feature encryption methods. Among them, the bit-plane randomization and randomized unary encoding support the calculation of Hamming distance between the encrypted features, and random projection supports the calculation of L1 distance between the encrypted features. However, the feature protection methods in [4, 5] will decrease the accuracy of image retrieval. Homomorphic encryption has also been introduced to protect image features. Wang et al. [6] employed approximate homomorphic encryption together with deep neural networks to perform content-based retrieval directly over ciphertext images, which significantly reduces the communication rounds between the user and the server. Nevertheless, homomorphic-encryption based frameworks still involve considerable overhead in terms of computation and ciphertext storage.

Hash-based and index-based techniques are widely adopted to improve the search efficiency of encrypted features. Wang et al. [7] combined deep hashing with searchable encryption, in which the deep features extracted by a residual network are quantized into hash sequences and protected by the Paillier homomorphic encryption to build a secure searchable index. Xia et al. [8] proposed a privacy-preserving image retrieval scheme using the secure local binary pattern in cloud computing. Li et al. [9] studied the similarity search for encrypted images in secure cloud computing, where the feature vectors are protected by the secure KNN algorithm and the search efficiency is improved accordingly. Furthermore, Li et al. [10] designed a secure and verifiable multikey image search scheme in cloud-assisted edge computing, which supports multiple image owners and verifies the correctness of the returned results, making it suitable for resource-constrained edge scenarios such as intelligent substations. Tian et al. [11] proposed a privacy-preserving image retrieval scheme with access control based on searchable encryption in the media cloud, in which a Bkd-tree index is constructed over the encrypted features extracted by a pre-trained CNN model. Li et al. [12] proposed a verifiable encrypted image retrieval scheme, in which the verification information is embedded into the encrypted image by reversible data hiding, so that the query user can verify the returned images and a malicious server cannot forge the retrieval results. Cao et al. [13] further stored the hash codes of the encrypted images on the blockchain, where the retrieval is performed by smart contracts, so that the reliability of the retrieval results is guaranteed by the consensus mechanism without extra verification. From the perspective of trustworthiness, Bogdanova et al. [14] presented an explainable data collaboration framework (DC-SHAP) for privacy-preserving distributed machine learning, which improves the consistency of model explanations among participants and enhances trust in privacy-preserving intelligent systems. In addition, Rawat and Rajavat [15] designed a trust-aware federated representation learning framework, in which the encrypted model updates are aggregated with trust-weighted averaging, indicating that the privacy of the participants can be preserved without a fully trusted aggregator.

In feature-encryption based frameworks, the image owner is responsible for extracting image features. It is also a heavy burden to the owner and is desirable to be outsourced to the cloud server. In image-encryption based frameworks, the image owner only needs to encrypt the images. The feature extraction and the search operation are outsourced to the cloud server. The key point is how to encrypt images and make sure that useful features can be extracted from such encrypted images. Xia et al. [16] proposed the bag-of-encrypted-words (BOEW) model, in which the image is encrypted by color value substitution, block permutation and intra-block pixel permutation, and the local histograms are calculated from the encrypted image blocks directly by the cloud server. From

the viewpoint of visual cryptography, Buckley et al. [17] investigated multi-image random grids, where secret images are encoded into maximally entropic shares and revealed by stacking the shares in different ways. Kayani et al. [18] encrypted the images with a Gauss-Square chaotic map before outsourcing and retrieved similar images in the encrypted domain with a modified Euclidean distance over the encrypted features. Unlike the noise-like encryption, thumbnail-preserving encryption keeps the thumbnail of the encrypted image visually similar to that of the original one. An et al. [19] combined thumbnail-preserving encryption with local visual features to retrieve images directly from the encrypted thumbnails, and Fan et al. [20] proposed a universal approximate thumbnail-preserving encryption method which supports the lossless recovery of the original images, indicating that the encryption of images does not necessarily destroy their usability for retrieval.

The image-encryption based frameworks above are some good solutions which outsource most tasks to the cloud server. However, the image encryption in most of these frameworks is conducted in the spatial domain. Thus, the encrypted images cannot be well compressed as the dependency between the pixels is destroyed by the encryption. Zhang and Cheng [21] proposed a privacy-preserving CBIR especially for JPEG images. In this framework, the DCT coefficients are permuted at the same location in different blocks at the image owner side. Then, the histogram of DCT coefficients is calculated to measure the similarity between images. As the encryption is compatible with the JPEG compression process, the encrypted image data can be compressed without much file size expansion. In order to enhance the security, Cheng et al. [22] conducted the encryption and feature extraction on the pairs of run-length and coefficient value, named the (r, v) pair. In this framework, the DC coefficients are encrypted by a stream cipher, and all the (r, v) pairs are permuted in each color component. Then, the histogram of coefficient values is extracted from each component for image retrieval. In order to improve the retrieval accuracy, Cheng et al. [23] slightly changed their encryption method. Specifically, the DC coefficients together with the quantization tables are encrypted by the stream cipher. Then, the (r, v) pairs are permuted within the blocks. Finally, all the DCT coefficient blocks are shuffled within each component. The histograms of AC coefficients are calculated from each block. A block-wise distance comparison method is proposed to measure the similarity, which is, however, quite time consuming. Afterwards, Yu et al. [24] combined the local Markov features extracted from encrypted big-blocks with the bag-of-words model to characterize the intra-block, inter-block, and inter-component dependencies among DCT coefficients, and achieved an improved retrieval accuracy. Liang et al. [25] proposed a JPEG encrypted image retrieval framework based on Huffman coding. They scrambled the AC coefficients, while the DC coefficient, VLI code, and quantization table are encrypted by a stream cipher. The similarity between images is measured by the Huffman coding histogram.

More recently, deep learning has been introduced into encrypted JPEG image retrieval. Lu et al. [26] proposed the Huffman-code based self-attention networks (**HBSAN**), in which a self-attention multi-layer perceptron module is designed to learn the global dependencies within the representations of cipher-images. Feng et al. [27] proposed the DCT histograms-based attention networks (DHAN), which learn non-linear cipher-image representations from the ruled DCT histogram features in a not-end-to-end manner and greatly improve the retrieval performance. Furthermore, Feng et al. [28] proposed the encrypted Vision Transformer (EViT), which extracts multi-level local length sequence and global Huffman-code frequency features from cipher-images and trains a self-supervised Vision Transformer with Huffman-embedding, outperforming existing schemes in terms of retrieval

accuracy by large margins. Very recently, Feng et al. [29] further introduced adaptive secret keys into the encrypted JPEG image retrieval, where the encryption is still aligned with the JPEG compression process and the features are learned from the cipher-images by self-supervised block-augmented pretraining, so that images encrypted with different keys can still share a stable feature space.

In this paper, we propose a secure outsourcing CBIR framework for JPEG images. Deep learning is adopted to learn useful features for image retrieval. An improved retrieval accuracy is achieved. In addition, the image encryption method in our framework makes no size expansion of the image file. Except for the image encryption, all of the other tasks, i.e., feature extraction and search operation, are outsourced to the cloud server.

3. System Overview

3.1 System Model

The proposed framework involves two entities, namely the image owner and the cloud server, as shown in Fig. 1. The image owner and the cloud server have their own tasks, performing several different steps respectively.

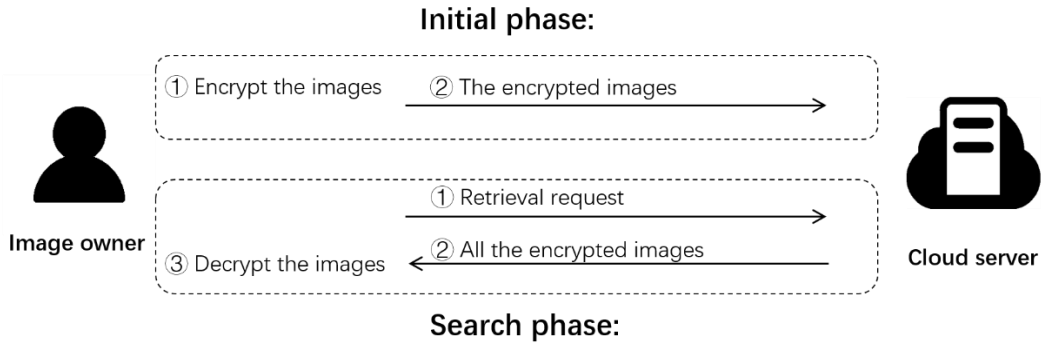


Fig. 1. The system model of the proposed framework.

The image owner has a plaintext image database $\mathbf{I}=\{I_i\}_{i=1}^{n}$, with a corresponding identity set $\mathbf{ID}=\{ID_i\}_{i=1}^{n}$. In order to protect image privacy, the image database needs to be encrypted and generated a corresponding encrypted image set $\mathbf{C}=\{C_i\}_{i=1}^{n}$. The image owner then uploads the encrypted image set **C** to the cloud server. When the image owner wants to search some similar images of an image, the query image can be encrypted in the same way as the image database **I**. The cloud server will be used to effectively retrieve the image database **C** and find some images which are similar to the query image. Finally, the corresponding decryption algorithm is used to restore the retrieved images.

The cloud server stores the encrypted image database C for the image owner. Then, the cloud server extracts the features from all encrypted images in the database C to build a secure index and provides the image retrieval service. When the cloud server receives the query request, the similarity between the query image and the images in the database C can be calculated by the distance of their features. Finally, all similar images will be returned to the image owner.

3.2 Security Model

Similar to many previous methods, we first assume that the cloud server is semi-honest. The cloud server can honestly follow the specified protocols, but will be curious to access and analyze data of sensitive information. In our framework, the image owner also has the demand to query images, which is completely trustworthy and will not leak any information to the cloud server. Although the cloud server can search images in the database which are similar to the query image, these images are encrypted and will not leak image content.

3.3 Preliminaries

3.3.1 JPEG Image Encoding

JPEG is a kind of common image compression format. In particular, JPEG images are widely stored and transmitted on the Internet because of the nature of high compression ratio. If an RGB image needs to be compressed into the JPEG format, the color space transformation is first performed. That is, the RGB color space will be converted into another color space. Then, DCT, quantization and entropy coding operations are performed on all components in the transformed color space, respectively. Finally, the encoded data are combined together to generate a JPEG image according to the JPEG format. Due to the combination of quantization and entropy coding, JPEG images have a high compression ratio. In the image compression process, although the high-frequency coefficients after DCT transformation are mostly quantized to zero, image quality can still be preserved during reconstruction because the image information is mainly concentrated in the low-frequency region.

The image encryption method we designed is format compatible. During the JPEG compression process, image data is protected by encrypting intermediate data in the process of JPEG compression. Only by understanding the process of JPEG compression can we better understand the encryption method. Therefore, we will give an introduction of JPEG compression. As shown in Fig. 2, a series of steps are performed during the process of JPEG compression. In the intermediate process, AC coefficients are encoded as the tuple of run-length and values, which expressed as (r, v). In this paper, the feature is extracted from the intermediate data and then the similar images can be retrieved by comparing these image features.

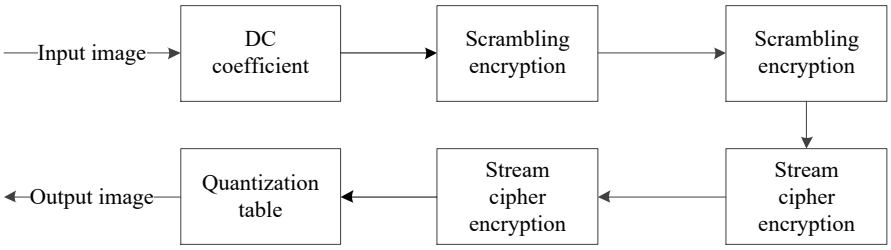


Fig. 2. JPEG coding flow chart.

(1) Color space transformation. In image processing, the RGB color space is a common image representation. Our eyes are more sensitive to brightness information than to color information, so we can take advantage of this visual characteristic to reduce the amount of image data. If the processing of brightness is detailed while the hue and saturation are only roughly processed, the compression ratio can be increased without affecting the visual effect too much. Therefore, the image represented by the RGB color space will be transformed into another color space. Each color space generates a brightness component and two chrominance components. For the computer, the

components of the color space transformation are represented by YCbCr. Among them, Y stands for brightness while Cb and Cr stand for saturation and chromaticity. The conversion relationship of the two color spaces is shown as follows:

$$\begin{bmatrix} Y \\ C_b \\ C_r \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ -0.1687 & -0.3313 & 0.500 \\ 0.500 & -0.4187 & -0.0813 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} + \begin{bmatrix} 0 \\ 128 \\ 128 \end{bmatrix} \quad (1)$$

(2) Block segmentation. After the color space is converted, all color space channels of the image are divided into some 8×8 small blocks respectively. Each image block contains 64 different pixels. These pixels and blocks are saved one by one for the calculation of the subsequent steps.

(3) DCT. There are many transformation methods in data compression, and DCT is the transformation from the spatial domain to the frequency domain. After the DCT, the 8×8 block can be transformed into the 8×8 block of DCT coefficients as

$$F(u,v) = C(u)C(v)\sum_{i=0}^{N-1}\sum_{i=0}^{N-1} f(i,j)\cos\left(\frac{(i+0.5)\pi u}{N}\right)\cos\left(\frac{(j+0.5)\pi v}{N}\right) \quad (2)$$

where $u, v \in \{0, \cdots, 7\}$ represent horizontal spatial frequency and vertical spatial frequency respectively. $C(x) = \sqrt{1/N}$ when $x = 0$ and otherwise $C(x) = \sqrt{2/N}$, which means the scale factor of a $N \times N$ block. $F(u,v)$ denotes the DCT coefficient at the position of (u,v). In which, $F(0,0)$ is called the DC coefficient and the other 63 values are called the AC coefficient. Among all 64 DCT coefficients, the DC coefficient is in the upper left corner.

(4) Quantization. For the purpose of compression, the DCT coefficients need to be quantized. The so-called quantization is a result of dividing the coefficients by the corresponding value in the quantization table. Then, these values will be rounded off to get the nearest integer. In this result, some high-frequency components are close to zero and others will become small integers. The purpose of the entire quantization is to reduce the number of non-zero coefficients and increase the number of zero-valued coefficients. In general, it reduces the number of bits and the storage space of image content. The quantization process satisfies the sensitivity requirement of the human visual system. In the JPEG algorithm, quantization tables of brightness and chrominance are respectively given. The size of the quantization table is also 8×8, which corresponds to 64 coefficients one by one. The quantization table is obtained on the basis of experiments combined with the visual characteristics of humans. Utilizing these characteristics, the short bit representation will be assigned to low-frequency components for the sake of obtaining a high compression ratio, which contain a large amount of information in the image. Since the value in the upper left corner of the quantization table is small and the value in the lower right corner is large, it can maintain the low-frequency components and suppress the high-frequency components. In general, the Y component is more important than other components. We can use a refined quantization for Y with a coarse quantization for Cb and Cr components to further increase the compression ratio.

(5) Intermediate processing. After the DCT coefficients are quantized, the generated sequence continues to be encoded by blocks into an intermediate format. The RLE (Run-Length Encoding) is efficient when the input sequence contains many zeros. A large number of zeros will appear in the quantized DCT coefficients in the JPEG encoding process, so high encoding efficiency can be achieved by using RLE. The scan order of RLE for JPEG images is not based on spatial coordinates but on the basis of frequency, which is called Zig-Zag scanning. As the example shown in Fig. 3, the

coefficients can be expressed as (−32, −6, 0, −6, −4, −9, 4, 0, 0, 0, 2, 0, 0, 0, 2, 0, 0, 0, 0, 0, 2), and the remaining zeros can be omitted. In the encoding, JPEG prefers to efficiently encode the coefficients in the low-frequency region because these coefficients are much more important than the coefficients of the high frequency, and most quantized values in the high-frequency region are zero. Then, DC coefficients perform differential encoding as DC coefficients between adjacent 8×8 blocks have a strong correlation. After that, the remaining 63 AC coefficients are quantized by using RLE. Each non-zero coefficient will be encoded into a pair of the format of run-length and value, where value is the current non-zero coefficient value and run-length is the number of consecutive zero-valued coefficients before this non-zero coefficient value, which can be denoted as (r, v). Since only 4 bits are used to represent one run-length, the maximum value is 15. When the value of run-length is greater than 15, one or more bits can be inserted into the sequence until the number of remaining zeros is less than 15. Therefore, DCT coefficients are represented as a sequence consisting of pairs, and $(0, 0)$ can be used at the end of the block to indicate that the subsequent values are all zeros. As illustrated in Fig. 3, if the DC coefficient in the previous block is −31, the differential coding will be −1. Thus, the intermediate coding of the block can be calculated and expressed as $\{(null, -1),\ (0, -6),\ (1, -6),\ (0, -4),\ (0, -9),\ (0, 4),\ (3, 2),\ (3, 2),\ (5, 2),\ (0, 0)\}$.

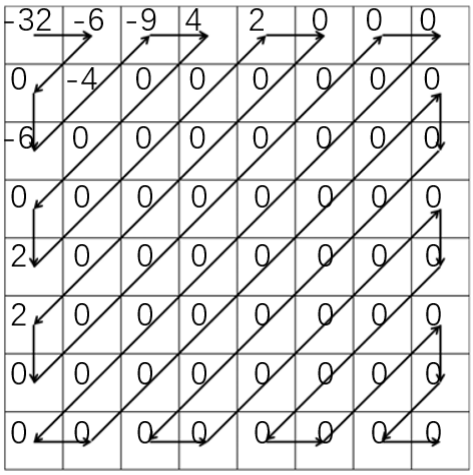


Fig. 3. The example of a DCT coefficient block and Zigzag scanning.

It is worth noting that our encryption method is also conducted on the intermediate information generated here. The image encryption and secure feature extraction are mainly conducted on the intermediate processing.

Table 1. The VLI coding table.

Value	Group index	Binary code
0	0	/
-1, 1	1	0,1
-3, -2, 2, 3	2	00, 01, 10, 11
-7, -6, -5, -4, 4, 5, 6, 7	3	000,001,010,011,100,101,110,111
-15, … , -8, 8, … , 15	4	0000, 0001, … ,1110,1111

-31, … ,-16, 16, … ,31	5	000000, … ,111111
-63, … ,-32, 32, … ,63	6	0000000, … ,1111111
-127, … ,-64, 64, … ,127	7	00000000, … ,11111111
-255, … , -128, 128, … ,255	8	000000000, … ,111111111
-511, … ,-129,129, … ,511	9	0000000000, … ,1111111111
-1023, … ,-512,512, … ,1023	10	00000000000, … ,11111111111
-2047, … ,-1024,1024, … ,2047	11	000000000000, … ,111111111111
-4095, … ,-2048,2048, … ,4095	12	0000000000000, … ,1111111111111
-8191, … ,-4096,4096, … ,8191	13	00000000000000, … ,11111111111111
-16383, … ,-8192,8192, … ,16383	14	000000000000000, … ,111111111111111
-32767, … ,-16384,16384, … ,32767	15	0000000000000000, … ,1111111111111111

(6) Entropy encoding. For the data given above, the intermediate information can continue to be encoded by VLI coding and Huffman coding. The (r, v) pair will be converted to a triple of (run-length, group index, binary code) by VLI encoding. According to Table 1, we can change the intermediate coding to the triples $\{(,1,0_B), (0,3,001_B), (1,3,001_B), (0,3,011_B), (0,4,0110_B), (0,3,100_B), (3,2,10_B), (3,2,10_B), (5,2,10_B)\}$. For Huffman coding, JPEG provides a Huffman table for DC coefficients and AC coefficients, including both brightness and chrominance values for an image. The run-length and group index in the triplet are encoded by Huffman coding. After entropy coding, the encoded data of JPEG will be obtained for output. Entropy coding indicates that a higher frequency value will be represented by fewer bits, and a lower frequency value will be represented by more bits. Thereby, the amount of bits occupied by the image data is reduced, i.e., the amount of data of the image is reduced.

3.3.2 Deep Learning

Deep learning is a method in which computers simulate the processes of human cognition and learn from experience. It is a branch of machine learning and is built on the basis of traditional neural networks with a multi-layer network structure that uses a high-level abstraction of data. It possesses multiple layers that consist of complex structures or nonlinear transformations. The first layer of input data is the input layer and the last layer of the network is the output layer. The layers between the input layer and the output layer can be called the hidden layers. If every neuron in the network is connected to all the neurons in the upper layer and the lower layer, the network can be called a fully connected network. Normally, CNN is also a common network and it changes the traditional layer composed of neurons into a layer composed of filters.

Deep learning is a method of learning data based on characterization. The output can represent the many characteristics of an image, such as the relationship between pixels, or more abstractly represent the shape and edge corners of a particular area. The advantage of deep learning is the use of unsupervised or semi-supervised feature learning and efficient hierarchical feature extraction algorithms instead of manual features. With the continuous development of existing methods and

computer technology, deep learning has been successfully applied to a wide range of image analysis and retrieval tasks [1, 30, 31], which can solve the increasingly complex practical application problems.

4. The Proposed Framework

4.1 Overview of the Proposed Framework

The algorithm of the proposed framework is mainly performed by two different entities, namely the image owner and the cloud server. These two entities have their own tasks. In addition to key generation and encrypting images, the image owner is also responsible for submitting query requests and decrypting similar images finally. The cloud server performs index construction and searches similar images. The main process of the proposed framework is shown in Fig. 4.

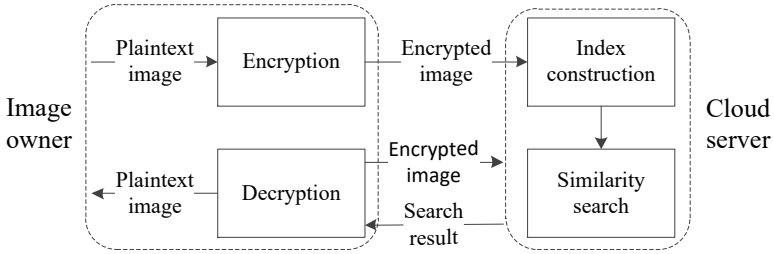


Fig. 4. Main process of the proposed framework.

In the initial phase, the image owner owns the image database $\mathbf{I}=\{I_i\}_{i=1}^{n}$ and the corresponding identification set $\mathbf{ID}=\{ID_i\}_{i=1}^{n}$. It generates the key **K** by a key generation method. Then, the image owner encrypts the JPEG image database to generate an encrypted JPEG image database $\mathbf{C}=\{C_i\}_{i=1}^{n}$, and uploads the encrypted image database **C** to the cloud server. After receiving the encrypted image database **C**, the cloud server builds the index **Idx**.

4.2 Privacy-Preserving CBIR Framework

Next, we will introduce the main steps of our framework in detail.

4.2.1 JPEG Image Encryption

Our encryption method is based on the bit-stream of a JPEG image. The main steps of our image encryption method are shown in Fig. 5.

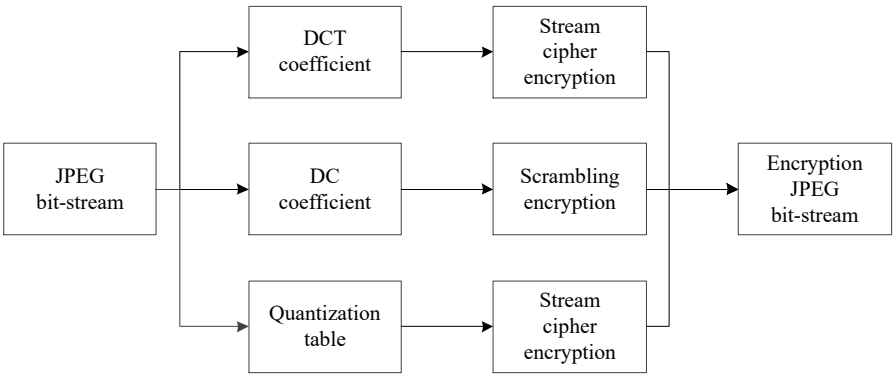


Fig. 5. Main steps of our image encryption method.

The image owner first reads the JPEG file information, and then uses the image data to parse the DCT coefficient entropy coding bit sequence of all components in the JPEG image. For example, we assume that $H_x^{(i,j)}$,$V_x^{(i,j)}(1 \le x \le 63)$ represent the Huffman coding and the VLI coding of the DCT coefficients of the (i,j)th block in the same image component. By using the Huffman table of the DC coefficient, the image owner can read and retrieve $H_0^{(i,j)}(DC)$ to obtain the group index of the DC coefficient in the VLI coding table, which is equal to the actual length of binary coding. The value of the DC coefficient before entropy coding can be obtained as $V_0^{(i,j)}(DC)$, which is equal to the value of binary coding. In a similar way, the value of the AC coefficient before entropy coding can be obtained. Then, the image owner can parse the bit sequence of all DCT coefficients. After parsing the JPEG bit-stream, the image can be encrypted by changing the value of the bit sequence.

(1) DCT coefficient encryption. Since the encryption process of the YCbCr components is in the same way, the Y component is taken as an example here to describe the encryption process. The VLI coding of the DCT coefficients for all blocks of the Y component is extracted into a bit sequence V in the original order. Assuming that the bit sequence V = $[v_1, v_2, \dots, v_n]$, randomly generate a sequence of bits E = $[e_1, e_2, \dots, e_n]$ with equal length to the sequence V. The sequences V and E are encrypted using stream cipher encryption to generate the encrypted bit v_i':

$$v_i' = v_i \oplus e_i \quad i = 1,2, \dots, n \tag{3}$$

Among them, the sequence **E** is controlled by the secret key **K**. To keeping the position of the bit sequence in the JPEG bit-stream unchanged, we replace the original bit v_i with the encrypted bit v_i'. Then, we complete the VLI coding encryption of the DCT coefficient of the **Y** component. The **Cb** and **Cr** components perform the same encryption operation while the encryption secret key of these components can be different.

(2) DC coefficient encryption. For the better security of images, we scramble the position of DC coefficients in the same component. Here, the scrambling is performed by the unit of bit sequences T = $[H_0^{(i,j)}(DC),\ V_0^{(i,j)}(DC)]$, which are encoded by the DC coefficient entropy, so that the size of the JPEG image can be kept constant. We generate a pseudo-random permutation P in the range [1…N], and the value of the i-th position can be represented as P_i. The key is generated from the private key K and N is the total number of DCT coefficient blocks in the image. Specifically, the random permutation position generated by the key is used to replace the position of T, and a new unit of bit sequences T′′ is generated as follows:

$$\mathrm{T}_i' = \mathrm{T}_j \tag{4}$$

in which the value of j is equal to the value of P_i. We can better protect image content by scrambling the position of DC coefficients by different keys in all components.

(3) Quantization table encryption. Since the image retrieval does not involve quantization tables, all quantization tables of the YCbCr components can be encrypted too. In general, the quantization tables of the Cb and Cr components are the same table. The encryption method of the quantization table also uses stream cipher encryption which is similar to DCT coefficient encryption. The quantization table of the Y component is taken as an example here to describe the encryption process. The VLI coding of the DCT coefficients for all blocks of the Y component is extracted into a bit sequence U in the original order. Assuming that the bit sequence U = $[u_1, u_2, \dots, u_n]$, randomly

generate a sequence of bits S = $[s_1, s_2, \ldots, s_n]$ with equal length to the sequence U. The sequences U and S are encrypted using stream cipher encryption to generate the encrypted bit u_j':

$$u_j' = u_j \oplus s_j \quad j = 1,2,\ldots,n \tag{5}$$

We replace the original bit u_j with the encrypted bit u_j'. Then, we complete the encryption of the quantization table of the **Y** component. The **Cb** and **Cr** components perform the same encryption operation as the **Y** component above. The encryption secret key of these components can be different which controlled by the secret key **K**.

4.2.2 Image Feature Extraction

For image retrieval, the similarity between images is often obtained by comparing image features. Therefore, feature extraction must be performed before retrieval. Here, we propose a method based on deep learning to extract features of encrypted JPEG images. The theoretical basis is that similar JPEG images have similarities in DCT coefficients. The detailed process is as follows.

First, the cloud server extracts the header file information and the Huffman coding of the DCT coefficients by parsing the encrypted JPEG bit-stream. Since the Huffman coding is unencrypted and contains the information of the VLI coding, it is easy to obtain the run-length and group index from the Huffman coding. Based on these information, it is easy to construct three matrices D_Y, D_U, and D_V which correspond to three components respectively. The elements of these matrices are related to the DCT coefficients. If the coefficient is a DC coefficient or zero-value AC coefficient, its corresponding matrix element will be set to zero. Otherwise, the matrix element value of the corresponding position can be the replaced by the group index of the non-zero AC coefficient. Then, we extract the features from matrices, for which the deep learning introduced here. Deep learning proposes a method to automatic learn features, thereby reducing the shortcomings caused by manual features.

What's more, CNN refers to neural networks that use convolution operations to replace general matrix operations in at least one layer of the network. It is a neural network designed to process data with a similar grid structure or a mainstream deep learning model. The overall structure of CNN mainly consists of three layers. The convolutional layer consists of filters and the activation function. Generally, the parameters include the number, size, stride and padding of filters. It also selects the appropriate activation function. For the pooling layer, the parameters include the choice of maximum pooling or average pooling, as well as the size of window and stride. For the fully connected layer, the parameters are the activation function and the number of neurons. CNN can be easily extended to images with large widths and heights, as well as images of variable size.

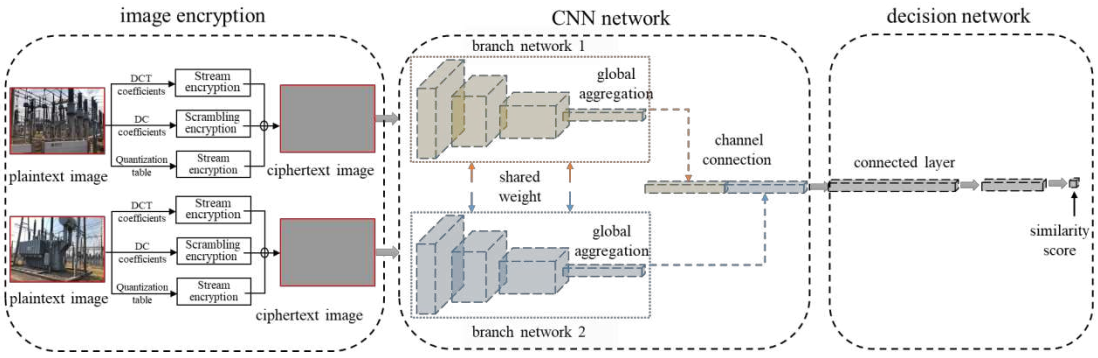


Fig. 6. An example of the structure used in our method.

The inspiration comes from the construction of CNN architecture, as shown in Fig. 6. The detailed process of our CNN will described later. Then, the value of output k is artificially set, which also represents the feature dimension of the image. So that, each image can be represented by a feature vector $\mathbf{f}= (f_i)_{i=1}^{k}$, and a linear search can be performed. The image identification and image feature vectors can be combined into an index table.

4.2.3 Image Retrieval

The image owner will encrypt the JPEG image and then uploads it to the cloud server if he wants to query similar images. After receiving the encrypted query JPEG image, the cloud server uses the same CNN model and parameters to calculate the corresponding k-dimensional feature vector as the query feature vector $\mathbf{f}_\mathrm{S} = (f_s)_{s=1}^{k}$. The cloud server uses $\mathbf{f}_\mathrm{S}$ and the index table at this stage. It can directly calculate the distance between $\mathbf{f}_\mathrm{S}$ and the feature vector $\mathbf{f}$ in the index table. In this process, each image feature of all encrypted images will be traversed and the distance between the feature vectors and $\mathbf{f}_\mathrm{S}$ will be calculated. The cloud server arranges these images by distance and selects the most similar θ image to return to the image owner while stopping the retrieval service.

4.3 Specific Architecture of Deep Learning

4.3.1 Single Input Model

In this case, we take a single encrypted image as the input in turn, and obtain the state matrix from it for training. Because the image sizes may not be the same, it may affect the effect of the deep learning network. We fixed the image size to a uniform size at first. For instance, the network scales all input images to the size of 298×298, and the inputs are scaled to [0, 1] for normalization. In the first convolutional layer, we use 30 5×5 convolution kernels with a stride of 2. Then, the output uses batch normalization. In this step, the input is normalized to solve the influence of data migration and increase, and to speed up the training. After that, the max-pooling layer is performed for the 2×2 region. The LeakyReLU is used as the activation function. It gives all negative values a slope a, and a can be a fixed parameter in (1, +∞). Here, we set a to be 0.3.

$$x' = \begin{cases} x/a\,, & x < 0 \\ x\,, & x \geq 0 \end{cases} \tag{6}$$

In order to fuse features on different scales, we next used a CNN module with three different CNN convolutional layers and finally merged them. The network is divided into three parallel parts. The first part is 16 3×3 convolution kernels with a stride of 1, followed by batch normalization, max-pooling, and the activation function. The second part is 16 5×5 convolutional layers with a stride of 2, followed by batch normalization and the activation layer. Because pooling increases the receptive field and loses some information, the pooling layer is discarded here. The third part is 16 3×3 convolutional layers with a stride of 2, followed by batch normalization and the activation layer. The pooling layer is also discarded here. These three networks are superimposed into a 48-channel feature map. After that, we use k 3×3 convolution kernels with a stride of 2, batch normalization and activation function. Finally, the global pooling is used to generate k neurons connected to the final neurons as output. The number of neurons is the number of categories in the image set, which uses Softmax for classification.

4.3.2 Pair Input Single Branch Model

For the large image database, the safest way to obtain a high-quality model is to increase the depth or width of the model, but the disadvantages of this way will appear in general as there are too many parameters in the model. If the training data set is limited, it is easy to overfit. The larger the network is, the higher the computational complexity will be. What's more, it is difficult to apply in reality. The deeper the network is, the easier the gradient will disappear and it is difficult to optimize the model. The method to solve these shortcomings is to convert connection or convolution into a sparse connection. However, the computing efficiency of the computer is poor for sparse data. We need a method that can not only keep the sparsity of the network but also make use of the high computing performance.

Inception is a classic CNN architecture. The main idea of Inception is to find out how to use dense components to approximate the optimal local sparse nodes, that is, to design a network with excellent local topology. Multiple convolution operations or pooling operations are performed on the input image in parallel, and all the results are spliced into a very deep feature image. Because different convolution and pooling operations can get different information of the input image, parallel processing of these operations and combining all the results will get better image representation. The function of Inception is to determine the filter type in the convolution layer or whether it is necessary to create a convolution layer and a pooling layer. The network will learn the parameters it needs by itself instead of manual work. The feature representation can avoid bottlenecks due to compression of data. High-dimensional information is more suitable for local processing. Increasing the nonlinear activation response in the convolution network can decouple more features and train the network faster. Image information can be embedded in a low dimension, which will not lead to a decrease in network representation ability. For example, before large-scale convolution, we can reduce the dimension of input information. If the information is easy to compress, the method of dimension reduction can even speed up the learning rate. It can balance the depth and width of the network. By balancing the number of filters in each layer of the network and the number of layers of the network, the network can achieve the best performance.

In order to integrate the features of different scales, we use the module with the CNN structure. Specifically, the plaintext image is encrypted by our method and transmitted to the CNN model, which scales all encrypted images to the size of 224×224 as input in pairs, and scales the input to [0, 1] for normalization. First, the convolution operation is performed on the input data. After that, LeakyReLU is used as the activation function to process the data. Then, we perform the max-pooling layer. The next step is the construction of the connection layer, which uses the pooling data to generate k neurons as output. Finally, the output is classified by the similarity of output data in pairs. In order to extract information from images more effectively, we can train a large image set in advance and get some more appropriate parameters to build the CNN framework. After that, we use the weights of the pre-trained model as the initial values to train the actual image set and get the output we want.

5. Security Analysis

As mentioned above, we assume that the cloud server is honest but curious. However, the cloud server may analyze the data of the image owner while obtaining compliance with the protocol. The cloud server sometimes can obtain private information of images, such as the content information

and the feature information of the encrypted image. The security of the framework is mainly analyzed in the ciphertext-only attack (COA), known-background attack (KBA), and access pattern.

5.1 Security Under the COA Model

The security of our framework under the COA model follows the rule of secure computation. Our framework involves the image owner and the cloud server. We assume that the honest but curious cloud server can be a potential opponent. We use a simulator M to simulate opponents and build the ideal experiment to show that our solution is safe.

The security of the image database. Assuming that the simulator M knows the total number of images and the size of each image, all images $\mathbf{I_M}$ and the corresponding identity set $\mathbf{ID_M}$ are simulated. The image information includes color information and texture information, which are included in the color channels. In our framework, image information will be protected by scrambling encryption and stream cipher encryption. After encryption, the original JPEG image cannot be reconstructed without the secret key. Without knowing the original content of the image, the image on the simulator can only be generated randomly and the security of the image database is guaranteed.

The security of the query image. In our framework, the query image uses the same encryption method as the image database on the cloud server, which is the method that the image owner encrypts the image. Therefore, the security strength of the query image is the same as the security strength of the image database on the cloud server.

5.2 Security Under the KBA Model

In the KBA model, in addition to knowing the information in the COA model, the attacker also knows some features about the original image. For example, pixel values in the same image set are always unevenly distributed, and adjacent pixels may have some similarities. Since our framework counts the group index of the VLI coding in the JPEG image and then generates the image features through the training of the CNN model by deep learning, the security strength of our framework does not decrease under this model with the disclosure of some plaintext information.

5.3 Access Pattern

Typically, the access pattern has a high level of privacy. In the image retrieval framework, the cloud server learns the information of those images retrieved by the image owner through the access pattern. If the cloud server knows the associated information, it is possible to group all images with similar interests without knowing what the specific interests are. The consequences of access pattern leakage depend on the application scenario, and usually it is not very harmful. Therefore, most researchers do not consider the access pattern leakage in their works.

6. Experimental Results

This section is intended to introduce the performance of our method primarily. We use Matlab 2016a to run the program. The device is configured as Intel Core(TM) i7-7700HQ, CPU 2.80 GHz, 8 GB RAM, Win10 system. We used the Corel Image Database as an experimental database and it contains 1000 color JPEG images with a required storage space of 30.3 MB. The image database can

be divided into 10 categories: Africans, beaches, buildings, buses, dinosaurs, elephants, flowers, horses, mountains, food. There are 100 images in every category, and each image has a size of 384×256 or 256×384. This database is also used in other encrypted image retrieval frameworks. In our experiments, these 1000 images are encrypted by our proposed encryption method, and these corresponding ciphertext images were used as query images. In order to evaluate the search performance of this program, we use a common search standard: precision and recall. The precision is the probability that the returned image is similar to the query image. The recall refers to the probability that an image in the image database, which is similar to the query image, is returned to the image owner. If these two parameters appear in the same coordinate system, we can get a performance curve called the Precision-Recall curve (P-R curve). Then, we can calculate the mean average precision (mAP).

6.1 The Effectiveness of Image Encryption

In our scenario of substations, the JPEG image is encoded to form the JPEG bit-stream. Then, the encrypted JPEG bit-stream can be formed by stream cipher encryption and permutation encryption. In addition, the image information becomes chaotic and disordered after the encrypted JPEG image is generated. As shown in Fig. 7, the image information is greatly changed and the encrypted image is very different from the original image, and the image information is well protected. These ensure good protection of substation information.

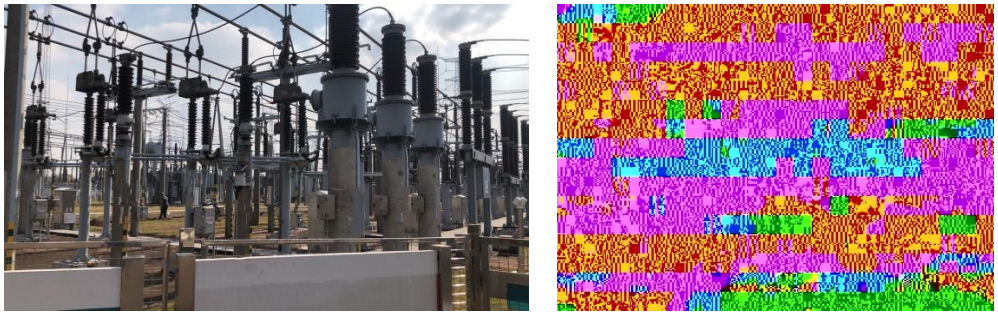


Fig. 7. The visual effects corresponding to JPEG images: (a) the original JPEG image; (b) the encrypted JPEG image.

6.2 The Precision of Linear Retrieval

In our experiment, for the image database used, first, we calculate the P-R curve of the proposed method, and then calculate mAP. The retrieval accuracy of the proposed framework mainly depends on the dimension k of the CNN output layer, that is, the dimension of the feature vector. We set different parameters to calculate the effectiveness of our solution, as shown in Table 2. The retrieval accuracy corresponding to different parameters is shown in Fig. 8.

Table 2. Retrieval accuracy with different k.

k	32	64	128	256
mAP	0.4445	0.4276	0.4754	0.4528

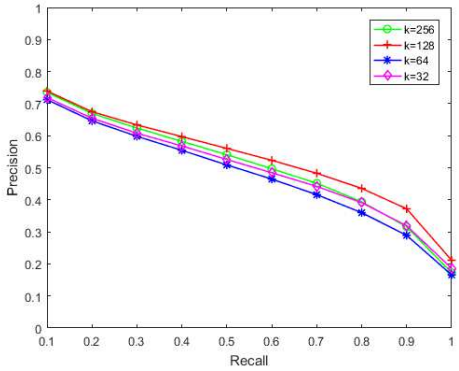


Fig. 8. P-R curve with different values of k.

The experimental results show that the change of the parameter k affects the final retrieval accuracy, and the retrieval accuracy is the highest when k = 128 among several different parameters. The results show that the dimension of the feature vector and the accuracy of the retrieval are nonlinear. Since the difference between the features may also be a factor affecting the retrieval effect, we use the activation function to reduce the difference between feature vectors. At this point, we will not change the value of k. Different activation functions are used here, but the rest of the parameters are still the same setting parameters as the previous best result. The retrieval accuracy of the experiment is shown in Table 3. The result shows that the use of the activation function reduces the difference between the feature vectors, but does not improve the retrieval accuracy.

Table 3. Retrieval accuracy with different activation functions.

Function	Round	Sigmoid	Sign	Arctan
mAP	0.4612	0.4472	0.1947	0.4424

Then, we compare the proposed framework with other frameworks and the retrieval accuracy of these frameworks is shown in Fig. 9. We can find that the retrieval efficiency of our framework is higher than [21] and [22].

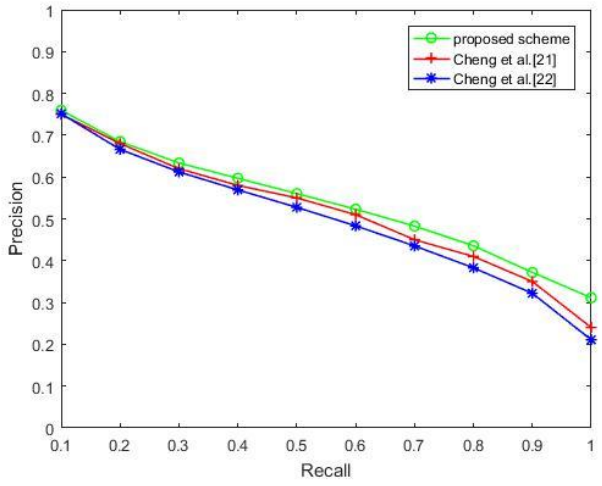


Fig. 9. Image retrieval accuracy of different image retrieval frameworks.

6.3 The Efficiency of Linear Retrieval

In the proposed framework, for the JPEG bit-stream, the DCT coefficients and the quantization tables are encrypted by stream cipher encryption and the DC coefficients are scrambled and encrypted. Image encryption is fast, and the proposed image encryption method is more efficient than some other encryption methods. The cloud server trains the image database using CNN, and each image outputs a value of a fixed dimension as the feature vector. The time required is related to the specific parameters used in the network. Linear indexes are used in our framework, so the cloud server needs to search the index table of the entire image database to find the most similar θ images. In addition, the change in the dimension k of the feature vector also affects the time consumption. The time consumption of different operations with different parameters k is shown in Table 4. Although our retrieval framework involves image encryption and deep learning, the time consumption of our framework is not high.

Table 4. Time consumption of operations corresponding to different parameters k.

k	32	64	128	256
Feature (s)	23.395	24.053	24.793	26.502
Search(s)	0.229	0.252	0.306	0.328

6.4 Evaluation Protocol for Operator-Centred Assessment

Retrieval accuracy alone does not determine whether an encrypted retrieval system is usable in inspection practice. This section therefore specifies a protocol by which that question can be answered with inspection personnel; no operator data are reported in the present article, and the protocol is stated in advance of data collection so that the design is fixed, reviewable and replicable by other groups. The protocol is exploratory rather than confirmatory: its purpose is to establish the direction and the plausible magnitude of the effects of encryption and of preview availability, and to provide the variance estimates needed to power a full-scale study. The conditions, the instruments and the outcome measures are summarised in Table 5.

Participants are inspection and maintenance personnel who currently hold field responsibility for defect identification, so that the task reflects their normal work rather than a laboratory analogue. The target sample is at least fifteen participants, following the power analysis reported below, and recruitment spans more than one operating unit so that organisational effects are not confounded with the experimental conditions. A comparison group of staff without domain expertise is drawn from the same organisation in order to examine whether domain knowledge moderates the effect of encryption on retrieval behaviour. Participation is voluntary and unpaid, written informed consent is obtained before each session, and the protocol is submitted to the institutional review board of the host utility before recruitment begins. Records are pseudonymised at the point of collection, and audio is recorded only with separate consent.

Each participant completes the same retrieval task under three conditions, presented in a counterbalanced order following a Latin square to control for learning effects. In C1 the participant searches a plaintext image database; in C2 the database is encrypted and no visual preview is available; in C3 the database is encrypted but a thumbnail preview generated in the encrypted domain is shown alongside each returned item. In every trial the participant receives a defect description, for example “broken shed of a composite insulator”, and is asked to identify the ten most similar images in a database of 2,000 images drawn from the evaluation collection described in Section 6.2. Ground truth is labelled independently by two senior inspectors and disagreements are resolved by discussion; the same ground truth serves all three conditions, so that differences in accuracy cannot be attributed to labelling.

Trust is measured with the twelve-item scale of Jian et al. [32], which yields a single score on a seven-point scale; a calibration index is computed as the absolute difference between the trust score rescaled to the unit interval and the observed Precision@10, so that over-trust and under-trust can be distinguished. Perceived workload is measured with the raw NASA-TLX score [33] and usability with the System Usability Scale [34]. Task performance is captured by completion time per trial and Precision@10, and short semi-structured comments are collected after each condition in order to explain unexpected patterns rather than to serve as an outcome.

Because the design is exploratory, the analysis reports descriptive statistics with 95% confidence intervals, and the three-condition comparison uses Friedman tests followed by post-hoc Wilcoxon signed-rank tests with Holm correction, Kendall’s W and r being reported as effect sizes; the calibration index is treated as a secondary outcome and analysed in the same way. An a priori power analysis for a repeated-measures design ($f = 0.40$, $\alpha = 0.05$, power = 0.80, three measurements) indicates a required sample of fifteen participants. The de-identified response matrix and the analysis script will be released with the results, so that every statistic reported from this protocol can be reproduced.

Table 5 sets out the design rather than an outcome. Once the data are collected, the quantities of interest are the change in completion time and in Precision@10 between C3 and C1/C2, the NASA-TLX and SUS scores, and the trust score together with the calibration index, each reported with the test statistics described above. The protocol does not presuppose the direction of these effects; the working hypothesis, to be tested rather than assumed, is that encryption without preview trades speed and calibrated trust for privacy, and that a preview generated in the encrypted domain recovers part of both at the cost of a leakage that has to be stated explicitly.

Table 5. Evaluation design of the operator-centred protocol: conditions, instruments and outcome measures.

Dimension	C1 plaintext	C2 encrypted	C3 encrypted + preview
Visual information available	full image	none	encrypted-domain thumbnail
Completion time per trial (s)	recorded	recorded	recorded
Precision@10 (%)	recorded	recorded	recorded

NASA-TLX workload (0–100)	recorded	recorded	recorded
SUS usability (0–100)	recorded	recorded	recorded
Trust, Jian 12-item scale (1–7)	recorded	recorded	recorded
Trust calibration index	recorded	recorded	recorded
Planned contrast	reference condition	C2 vs C1	C3 vs C1 and C3 vs C2

7. Discussion

7.1 From Retrieval Accuracy to Operator-Centred Evaluation

The encrypted retrieval literature, including the JPEG-oriented line of work that this paper extends [21–25, 28, 29], has converged on two indicators: mean average precision and file-size expansion. Both are necessary, but neither tells a utility whether an inspector can complete a search task within an acceptable time, nor whether the returned cipher-images are trusted to the right degree. The protocol of Section 6.4 is designed to test exactly this divergence: a scheme that preserves accuracy may still impose a workload penalty when no visual preview is available. This is consistent with the broader observation that the interpretability of a retrieval model is a precondition for its acceptance in expert practice [30, 36], and with work showing that relevance feedback supplied by the user can recover part of the lost precision at a modest annotation cost [31, 35]. For encrypted retrieval such feedback can be applied in the ciphertext domain, since re-ranking requires only linear operations on the feature vectors, which makes a human-in-the-loop variant an attractive next step.

7.2 The Privacy–Utility–Human Trade-off

The three objectives are not independent. Format-compatible encryption keeps the file size unchanged and therefore preserves storage utility, but it also preserves statistical structure that a preview may expose; thumbnail-preserving encryption deliberately trades a controlled amount of visual privacy for usability, and the visual-cryptography construction of Buckley et al. [17] makes the same trade-off in a different way. The protocol makes the preview condition (C3) the place where the three objectives meet: it is the condition most likely to be preferred by operators, and at the same time the one that requires the most careful privacy argument. We therefore suggest that reports on encrypted retrieval state the leakage implied by any usability feature explicitly, rather than treating usability as an unqualified benefit.

7.3 Deployment Constraints in Power Systems

A retrieval service for substations is deployed inside an environment governed by the security protection regulations for power monitoring systems and by the classified protection scheme for cybersecurity, both of which impose network partitioning and forbid direct connection between monitoring zones and external networks. Two consequences follow for the present framework. First,

the cloud server is best read as an intranet or dedicated-cloud service rather than a public cloud, which weakens the threat model of Section 5 but does not remove it, since insider access and compromised hosts remain in scope. Second, key management must reside in the same protected zone as the image owner, and the audit log of retrieval requests becomes a compliance artefact in its own right. A production deployment would also require a retention policy for cipher-images and a documented procedure for key rotation.

7.4 Limitations and Ethical Considerations

Four limitations qualify the present results. The operator-centred assessment is specified as a protocol only: no human-subject data are reported, so the claims about workload and trust remain hypotheses rather than findings, and the sample foreseen in Section 6.4 will support statements about direction rather than about effect size. The retrieval experiments rely on the Corel collection, whose semantic categories differ from those of substation inspection images, and the domain gap may bias the accuracy obtained on that collection and, by extension, the workload that the protocol would record in the field. The security analysis covers chosen-plaintext and known-background-knowledge adversaries but does not evaluate whether the learned representation leaks information about the plaintext, which is a recognised risk for features extracted in the encrypted domain. Finally, a system that returns plausible results with confidence may encourage over-trust; the calibration index introduced in Section 6.4 is a first step towards detecting this, but it should be monitored in any field deployment, and operators should retain the authority to reject a returned set without justification.

8. Conclusion

This paper proposes a novel privacy-preserving encrypted JPEG image retrieval method for intelligent substations that outsources the main computational overhead to the cloud server, which further reduces the burden on the image owner. The framework protects the image content by using stream cipher encryption and scrambling encryption. It extracts the VLI coding from the encrypted image to construct matrices and uses the CNN frame to generate an output value as the feature vector representing each image on the basis of deep learning. The similarity between images is measured by the Euclidean distance between the feature vectors. Experiments have shown that our solutions are proven to be safe in substations, having good performance and retrieval accuracy. Beyond accuracy, we specify a protocol by which the usability and trust consequences of encryption can be measured with inspection personnel, on the argument that operator-centred measures should complement accuracy-oriented benchmarks when encrypted retrieval is evaluated for field use. In future work, we consider building a more efficient network to improve retrieval accuracy, executing the protocol with a larger multi-site sample, and examining whether relevance feedback supplied in the ciphertext domain can recover the precision lost to encryption.

Statements and Declarations

Author Contributions

S.D. conceived the framework, designed and performed the retrieval experiments and the security analysis of the JPEG bitstream, and wrote the main manuscript text. B.L. and X.Y. contributed to the methodology design and to the development of the encryption scheme and the feature-learning

pipeline. H.W. and X.G. curated the evaluation image collection, prepared Figures 1–9 and Tables 1–5, and carried out the data analysis. G.W. contributed to the deployment analysis for power-system compliance and revised the manuscript. All authors discussed the results, commented on the manuscript, and read and approved the final version.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests

The authors declare that they have no competing interests.

Data Availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Ethics Approval

No human-subject data are reported in this article. The evaluation protocol described in Section 6.4 will be submitted to the institutional review board of the host utility before recruitment, and written informed consent will be obtained from all participants.

Consent to Participate / Consent for Publication

Not applicable. No human participants took part in the work reported here; informed consent will be obtained from all participants when the protocol of Section 6.4 is executed.

Declaration of AI-Assisted Technologies

No AI-assisted technologies were used.

References

[1] Dubey S.R. A decade survey of content based image retrieval using deep learning. IEEE Trans Circuits Syst Video Technol, 32(5): 2687-2704, 2022.

[2] Xie R., Chen Z., Cao W., Wu C., Li T. A novel federated learning framework for semantic segmentation of terminal block in smart substation. Pattern Recognit, 167: 111665, 2025. doi:10.1016/j.patcog.2025.111665

[3] Khawaji A., Martin R.J. ML-based threat severity detection in cyber-physical systems for smart environments using behavioral analysis. Hum-Cent Intell Syst, 6(1): 48-65, 2026. doi:10.1007/s44230-026-00134-9

[4] Lu W., Swaminathan A., Varna A.L., Wu M. Enabling search over encrypted multimedia databases. In IS&T/SPIE Electronic Imaging, SPIE, San Jose, CA, USA, pp. 725418, 2009.

[5] Lu W., Varna A.L., Swaminathan A., Wu M. Secure image retrieval through feature protection. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), IEEE, Taipei, Taiwan, pp. 1533-1536, 2009.

[6] Wang Y., Chen L., Wu G., Yu K., Lu T. Efficient and secure content-based image retrieval with deep neural networks in the mobile cloud computing. Comput Secur, 128: 103163, 2023. doi:10.1016/j.cose.2023.103163

[7] Wang Z., Zhang Q., Meng L., Liu Y. Secure content based image retrieval scheme based on deep hashing and searchable encryption. Comput Mater Continua, 75(3): 6161-6184, 2023.

[8] Xia Z., Wang L., Tang J., Xiong N.N., Weng J. A privacy-preserving image retrieval scheme using secure local binary pattern in cloud computing. IEEE Trans Netw Sci Eng, 8(1): 318-330, 2021.

[9] Li Y., Ma J., Miao Y., Wang Y., Liu X., Choo K.-K.R. Similarity search for encrypted images in secure cloud computing. IEEE Trans Cloud Comput, 10(2): 1142-1155, 2022.

[10] Li Y., Ma J., Miao Y., Liu L., Liu X., Choo K.-K.R. Secure and verifiable multikey image search in cloud-assisted edge computing. IEEE Trans Ind Inform, 17(8): 5348-5359, 2021. doi:10.1109/TII.2020.3032147

[11] Tian M., Zhang Y., Zhang Y., Xiao X., Wen W. A privacy-preserving image retrieval scheme with access control based on searchable encryption in media cloud. Cybersecurity, 7(1): 1-12, 2024. doi:10.1186/s42400-024-00213-z

[12] Li M., Zhu Y., Du R., Jia C. Verifiable encrypted image retrieval with reversible data hiding in cloud environment. IEEE Trans Cloud Comput, 13(1): 397-410, 2025. doi:10.1109/TCC.2025.3535937

[13] Cao Y., Meng F., Shang X., Gui J., Tang Y.Y. A privacy-preserving large-scale image retrieval framework with vision GNN hashing. IEEE Trans Big Data, 11(4): 1970-1982, 2025. doi:10.1109/TBDATA.2024.3505052

[14] Bogdanova A., Imakura A., Sakurai T. DC-SHAP method for consistent explainability in privacy-preserving distributed machine learning. Hum-Cent Intell Syst, 3(3): 197-210, 2023. doi:10.1007/s44230-023-00032-4

[15] Rawat H., Rajavat A. Trust-aware federated user representation learning for privacy-preserving and carbon-optimized personalized recommendation in large-scale social platforms. Hum-Cent Intell Syst, 6(2): 219-243, 2026. doi:10.1007/s44230-026-00145-6

[16] Xia Z., Jiang L., Liu D., Lu L., Jeon B. BOEW: A content-based image retrieval scheme using bag-of-encrypted-words in cloud computing. IEEE Trans Serv Comput, 15(1): 202-214, 2022. doi:10.1109/TSC.2019.2927215

[17] Buckley N., Nagar A.K., Secco E.L. On multi-image random grids by arbitrary transformations. Hum-Cent Intell Syst, 4: 93-110, 2024. doi:10.1007/s44230-023-00056-w

[18] Kayani M., Riaz M.M., Ghafoor A., Khan F. Privacy preserving content based image retrieval. Multimed Tools Appl, 83(15): 44955-44978, 2024.

[19] An D., Zhang X., Hao D., Zhao R., Zhang Y. Privacy-preserving image retrieval based on thumbnail-preserving visual features. IEEE Trans Circuits Syst Video Technol, 35(8): 7719-7731, 2025. doi:10.1109/TCSVT.2025.3550782

[20] Fan H., Shi S., Li M. U-TPE: A universal approximate thumbnail-preserving encryption method for lossless recovery. J Vis Commun Image Represent, 104: 104318, 2024. doi:10.1016/j.jvcir.2024.104318

[21] Zhang X., Cheng H. Histogram-based retrieval for encrypted JPEG images. In 2014 IEEE China Summit & International Conference on Signal and Information Processing (ChinaSIP), IEEE, pp. 446-449, 2014.

[22] Cheng H., Zhang X., Yu J. AC-coefficient histogram-based retrieval for encrypted JPEG images. Multimed Tools Appl, 75(21): 13791-13803, 2016.

[23] Cheng H., Zhang X., Yu J., Zhang Y. Encrypted JPEG image retrieval using block-wise feature comparison. J Vis Commun Image Represent, 40: 111-117, 2016.

[24] Yu P., Tang J., Xia Z., Li Z., Weng J. A privacy-preserving JPEG image retrieval scheme using the local Markov feature and bag-of-words model in cloud computing. IEEE Trans Cloud Comput, 11(3): 2885-2898, 2023. doi:10.1109/TCC.2022.3233421

[25] Liang H., Zhang X., Cheng H. Huffman-code based retrieval for encrypted JPEG images. J Vis Commun Image Represent, 61: 149-156, 2019. doi:10.1016/j.jvcir.2019.03.021

[26] Lu Z., Feng Q., Li P. Encrypted JPEG image retrieval via Huffman-code based self-attention networks. In 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, pp. 1-6, 2022. doi:10.23919/APSIPAASC55919.2022.9979814

[27] Feng Q., Li P., Lu Z., Zhou Z., Wu Y., Weng J., Huang F. DHAN: Encrypted JPEG image retrieval via DCT histograms-based attention networks. Appl Soft Comput, 133: 109935, 2023. doi:10.1016/j.asoc.2022.109935

[28] Feng Q., Li P., Lu Z., Li C., et al. EViT: Privacy-preserving image retrieval via encrypted vision transformer in cloud computing. IEEE Trans Circuits Syst Video Technol, 34(8): 7467-7483, 2024. doi:10.1109/TCSVT.2024.3370668

[29] Feng Q., Lu Z., Zhang L., Li C., Huang F., Weng J., Yu P.S. Privacy-preserving image retrieval in cloud computing via adaptive secret keys and self-supervised block-augmented pretraining. IEEE Trans Serv Comput, 18(4): 2310-2325, 2025. doi:10.1109/TSC.2025.3576692

[30] Wickstrøm K., Østmo E.A., Radiya K., Mikalsen K.Ø., Kampffmeyer M., Jenssen R. A clinically motivated self-supervised approach for content-based image retrieval of CT liver images. Comput Med Imaging Graph, 107: 102239, 2023. doi:10.1016/j.compmedimag.2023.102239

[31] Sun M., Zou W., Hu N., Wang J., Chi Z. Iterative brain tumor retrieval for MR images based on user's intention model. Pattern Recognit, 127: 108650, 2022. doi:10.1016/j.patcog.2022.108650

[32] Jian J.-Y., Bisantz A.M., Drury C.G. Foundations for an empirically determined scale of trust in automated systems. Int J Cogn Ergon, 4(1): 53-71, 2000. doi:10.1207/S15327566IJCE0401_04

[33] Hart S.G., Staveland L.E. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Hancock P.A., Meshkati N. (eds) Human Mental Workload, North-Holland, Amsterdam, pp. 139-183, 1988.

[34] Brooke J. SUS: A quick and dirty usability scale. In: Jordan P.W., Thomas B., Weerdmeester B.A., McClelland I.L. (eds) Usability Evaluation in Industry, Taylor & Francis, London, pp. 189-194, 1996.

[35] Delussu R., Putzu L., Fumera G. Human-in-the-loop cross-domain person re-identification. Expert Syst Appl, 226: 120216, 2023. doi:10.1016/j.eswa.2023.120216

[36] Dwivedi R., Dave D., Naik H., Singhal S., Rana O., Patel P., Qian B., Wen Z., Shah T., Morgan G., Ranjan R. Explainable AI (XAI): Core ideas, techniques and solutions. ACM Comput Surv, 55(9): 194, 2023. doi:10.1145/3561048