Wireless Networks (2024) 30:5157–5169
https://doi.org/10.1007/s11276-023-03229-4

A secure image evidence management framework using multi-bits watermark and blockchain in IoT environments

Qing Yao[1,2] · Kaiwen Xu[3] · Taotao Li[4] · Yichao Zhou[3] · Mingsheng Wang[1,2]

Accepted: 28 December 2022 / Published online: 21 January 2023

Abstract
Criminal forensics in an Internet-of-Things (IoT) environment often requires complex investigations because IoT devices usually generate a large amount of electronic data, especially image data, which brings great difficulties to traditional digital forensic methods. Therefore, designing a secure management solution for image data and making it to be image evidence available in court is a new challenge. In this paper, we present a new secure image evidence management framework based on multi-bits digital watermarking and blockchain. In this framework, we first propose a flexible and robust self-learning based watermark embedding algorithm, which can embed both image marks and binary messages into the latent space of the input images and also improve the resistance of a broad range of attacks (geometric transform, JPEG, noise, etc.) to the watermarked images. And then, we design a smart contract resided in blockchain, which can achieve safe storage and automatically authentication of embedded watermarks. The experimental results on both watermarking algorithm and smart contracts have demonstrated the feasibility and efficacy of the proposed framework.

Keywords Image evidence management · Blockchain · Digital watermarking · Self-supervised learning · Smart contract · Internet-of-Things environments

1 Introduction

Electronic forensic, as an important part of smart city, has attracted widespread interest in recent years. One of the fundamental research fields of electronic forensic is the protection and identification of electronic evidences [1–4]. With the widely used electronic imaging equipment in the Internet-of-Things (IoT) devices and the rapid development of digital image processing technology, digital data, especially digital images, becomes more and more common in our daily life, leading to the increasing importance for image evidences preservation and identification. Compared to the evidences based on text description, image evidence can provide more realistic and intuitive visualization of a crime scene. Therefore, in the field of forensic science, digital image evidence has been widely used and has played an important role in investigating various cases. However, with the recent development of deep learning technology, the threshold of image forgery is getting lower and lower, while the detection of image forgery derived by deep learning models becomes more and more difficult, which brings difficulties to the identification of digital

✉ Taotao Li
litaotao18@outlook.com

Qing Yao
yaoqing@iie.ac.cn

Kaiwen Xu
xukw1221@njust.edu.cn

Yichao Zhou
yczhou@njust.edu.cn

1 State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China

2 School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China

3 School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, Jiangsu, China

4 School of Software Engineering, Sun Yat-Sen University, Zhuhai 519082, Guangdong, China

image evidences. And thus, to establish an efficient framework for image evidence preservation and protection looms ahead.

Digital image watermarking [5] and blockchain [6–8] are two kinds of data protection technology, where the former is an initiative authentication technology and the latter is an decentralized ledger technology. Image watermarking technology first constructs a paired encoder and decoder, and then, uses the encoder to embed watermark information into the original image in a hidden or invisible form. When an image to be verified is provided, if the corresponding decoder can extract the correct watermark information from the image, the authenticity and originality of the image can be guaranteed. Otherwise, this image is at risk of being tampered with. Blockchain is a chain composed of blocks with certain information stored in each block, and these blocks are connected into a chain according to the sequential order. So, it is extremely difficult to tamper with the information in the blockchain, which means that the information recorded by the blockchain is more authentic and reliable.

Although blockchain technology has strong ability of non-tamperability and non-forgeability for data protection, its memory overhead and computational cost is also quite high, especially for some high-dimensional data such as images and videos. Therefore, if we store the entire image on the blockchain, there will be very high requirements for the hardware resources of the blockchain deployment environment. On the other hand, although image watermarking can protect the content of the image itself, the protection of the watermark signal itself against external attacks cannot be guaranteed, especially when the paired watermark encoder and decoder are accidentally leaked or cracked. Consequently, many researchers have proposed to apply watermarking technology and blockchain technology together to solve the problem of efficient and low-carbon authenticity identification and preservation of digital image evidences. For instance, Senkyire et al. [2] proposed a crime scene forensics image verification scheme based on digital watermarking and blockchain technology; and Wang et al. [9] proposed an image storage and authentication framework for copyright protection, which utilizes the advantages of zero-watermarking algorithm and blockchain.

However, there are still some deficiencies of these proposed image preservation and authentication framework. For example, both of these proposed framework use zero-watermark method, which can only protect the originality of the image by detecting whether the image contains a watermark or not, but it cannot provide extra effective information, e.g., the event information corresponding to the image content, to realize the image source attribution identification. Consequently, it is not a suitable image evidence preservation model since image evidence usually contains image itself and its corresponding event information. Another shortcoming is that most of these framework use traditional transform-domain based image watermarking methods, which is not adaptable enough to images with different content. This motivates us to design a new framework for image evidence preservation combining self-supervised learning based multi-bits watermarking method and smart contracts resided in blockchain, which can fulfill the certification, identification, and traceability of digital image evidences. The efficacy and feasibility of the proposed system can be demonstrated through the performance tests on the multi-bits watermarking method and smart contracts.

The main contributions of this work are:

- we propose a blind multi-bits image watermarking method based on self-supervised learning model, which can adaptively find a suitable region in the image's feature space for embedding the watermark signal according to the content of the image; the proposed watermarking model is not only capable of embedding multimedia information (e.g., text, image logo, etc.), but also robust against to a broad range of attacks (e.g., geometric, JPEG, and noise attacks).
- we establish a new image evidence protection framework by incorporating multi-bits image watermarking and blockchain technology, which can effectively solve the problem of safe use and efficient storage of image evidence in the forensic application field; through designing a reasonable smart contract, the proposed system can be deployed at the IoT devices or other edge devices.

The remainder of this paper is organized as follows. In Sect. 2, we review related works and background knowledge about image protection with watermarking and blockchain. In Sect. 3, we propose our digital image evidence preservation framework with self-supervised learning based multi-bits watermarking algorithm and smart contracts resided in blockchain. And then, in Sect. 4, we present the experimental results on both image watermarking algorithms and smart contracts, including functional tests and performance comparisons. Finally, some discussion and conclusion remarks are provided in Sect. 5.

2 Related work

The image evidence, as a special kind of digital forensic evidence, usually has larger data volume compared with text or audio evidence. If we storage the image data completely on the blockchain, the memory requirement and computational cost will be a big challenge for the

deployment environment of blockchain. And thus, many researchers pay attention to the combination of blockchain and digital watermarking to solve the related problems of image forensic evidence preservation and protection.

2.1 Digital image watermarking

Image watermarking technology is an information hiding technology, which can embed a large amount of watermark information in the original digital image carrier to protect copyright or verify originality of the digital image [10–12]. Image watermarking algorithms can usually be divided into spatial-domain based algorithms and transform-domain based algorithms [5, 13, 14]. In the transform-domain based algorithms, the original image is first transformed into some transform-domain to derive the representation coefficients, i.e. Fourier coefficients or Wavelet coefficients; and then, the corresponding watermark embedding algorithm is used to embed the encrypted watermark signal into the image representation coefficients, e.g., the high-frequency coefficients of Fourier domain or the coefficients in the wavelet HH-subband. After the inverse transformation of representation coefficients from transform-domain to the spatial-domain, we can get the watermarked image. The reason for embedding watermark signal into high-frequency representation coefficients is that the modification of high-frequency coefficients will not cause large distortion to the main contours of the image, which usually reflects the low-frequency information of the image [13]. However, these image transform methods use fixed transformation basis which may not suitable for images with different content. And thus, some adaptive image decomposition approaches, e.g., principle component analysis (PCA), empirical mode decomposition(EMD) and etc., have also been applied to the field of image watermark [15–17]. However, both adaptive and non-adaptive transform-domain based image watermarking methods cannot against geometric attacks, which is very common in image forensic applications.

With the rapid development of deep learning technologies, many studies have introduced deep learning into the field of image watermarking to improve the capability and robustness of image watermarking against geometric and noise attacks. For example, the steganographic generative adversarial networks (SGAN) model [18] proposed by Volkhonskiy et al. is the first image watermarking model that incorporates the generative adversarial learning technique for information steganography. In this model, a generator is used to embed the watermark information into the input image to generate a watermarked image; and then, a discriminator is trained to discriminate the original image and the watermarked image generated by the generator. Based on this model, Shi et al.[19] replaced the original GAN model to Wasserstein GAN (WGAN) to improve the training process and derived a more realistic watermarked image with better visual quality. Since the discriminator is only used for distinguishing the original image and the watermarked image without the objective of evaluating the image quality, the visual quality of the watermarked image cannot be guaranteed.

Besides GAN-based architectures, the encoder-decoder architecture is another structure used for digital image watermarking, in which the encoder embeds watermark information into image and the decoder tries to extract the watermark [20, 21]. For instance, Zhu et al. proposed the Hiding Data Deep Network (HiDDeN) which jointly trains the encoder and decoder networks with noise layers for simulating image perturbations [20]. Wei et al. introduced a robust image watermarking algorithm based on cycle variational autoencoder (VAE) networks. Fernandez et al. presented image watermarking in a self-supervised latent space, which embeds both marks and binary messages into their latent spaces, leveraging data augmentation at marking time [22]. The self-supervised learning technique is adaptive to the image content and robust to a broad range of attacks (e.g., rotations, crops, JPEG, contrast, etc).

2.2 Blockchain

A blockchain is a decentralized shared ledger that combines transactions data into a specific chain blocks in a chain manner according to the sequential order. By means of cryptography techniques, a blockchain can ensure the imperability and non-forgeability of its data. It can safely store simple, sequential and verifiable data in the blockchain system [23–25]. A schematic diagram of components of a traditional blockchain is shown in Fig. 1. As we all know, a blockchain consists of blocks in a sequential order (as described in Bitcoin [23]). Each block is composed of a block header and a block body. A block header plays an important role in a block and is used to represent the block it belongs to. A block body mainly contains transactions, in Bitcoin [23], for instance, thousands of transactions are included in the block body. To ensure the integrality, transactions in a block header are arranged by a merkle tree structure. In this case, it easily determines whether a transaction exists in a block by verifying a merkle tree verification path corresponding to it.

In the other hand, a blockchain data structure is denoted as a directed acyclic graph (DAG) of blocks [26, 27]. Different from the Bitcoin blockchain described above, the directed acyclic graph blockchain has multiple block paths from a genesis block (the root of DAG) to the last blocks (the leaf nodes of DAG). In the aspect of consensus, a variety of consensus protocols are proposed and reach agreement on transactions data. For example, to reduce the

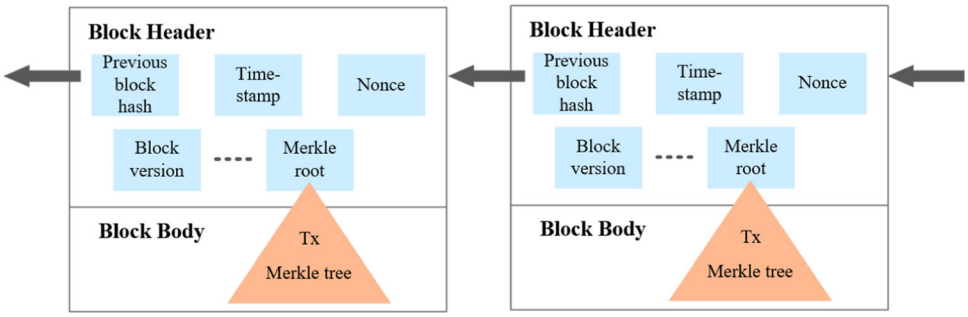

Fig. 1 A schematic diagram of components of conventional blockchain

waste of computation resource, a proof of stake consensus protocol is introduced by Kiayias *et al.* [28]; To advance the transactions throughput, various practical byzantine fault tolerance (PBFT) algorithms [29, 30] are applied in blockchain protocols. In the aspect of computation, a smart contract [31] is introduced into a blockchain system [32] to enhance autonomy. For simplicity, a smart contract is a program code running in the blockchain, which is used for executing the predefined operation according to parameters. In addition, executing a smart contract needs to consume resource, thus, *gas* is usually used to evaluate the cost of a smart contract [32].

According to the criminal forensics in an Internet-of-Things (IoT) environment, we adopt the permission blockchain with access control. This is because the digital image evidence should be managed by law enforcement authorities (i.e., court) to enhance the integrality and confidence of evidences. Compared to permissionless blockchains, in addition, permission blockchains have advantages in performance (i.e., transaction per second), satisfying the requirement of storing and verifying image evidences.

2.3 Image evidence preservation

Because of characteristics of the non-tamperability and non-forgeability of the blockchain technology, it has become a new trend for digital forensic evidence preservation and protection in recent years. After analyzing the problems of data management in Ethereum, Cao et al. [33] designed a Ethereum-based certificate framework for digital forensic evidence storage from the perspective of smart contracts. Brotsis et al. [34] proposed a blockchain-based solution in Internet-of-Things (IoT) environments, which is designed for the smart home domain, dealing with the collection and preservation of digital forensic evidence. For the image evidence, Zou *et al.* [35] proposed a blockchain based photo forensics scheme to uniformly solve photo-faking problems, photo-tracing problems and copyright dispute problems.

As mentioned above, image evidence usually has larger data volume which will be a big challenge for storage and computational cost if the image data completely stored on the blockchain. Therefore, many researchers proposed several solutions with the combination of blockchain and digital watermarking technologies for image evidence preservation and protection. In most of these solutions, a zero-watermarking algorithm is used to protect the images and a blockchain is used to ensure the watermarks storage and authentication [9, 36]. This architecture can also be extended to the other related fields, such as data provenance collection under the cloud environment [37] and crime scene forensics image verification [2].

Although these studies reflect the unique advantages of blockchain and digital watermarking in image copyright protection, they do not consider the protection of digital watermarking from external attacks, as well as the traceability of the images since most of them use zero-watermark rather than multi-bits watermark where only the later contains the event information or logos corresponding to the image. Therefore, in this paper, the multi-bits watermark algorithm, which supports rich multimedia watermark data, has been studied. Since the multi-bits watermark needs larger latent space for watermark embedding, a self-supervised learning watermark algorithm has been studied which can adaptively find the most suitable embedding area in the image's latent space according to the content of the image.

3 Image evidence protection framework

3.1 Framework architecture overview

In this paper, we propose a new image evidence protection framework, which can effectively solve the problem of

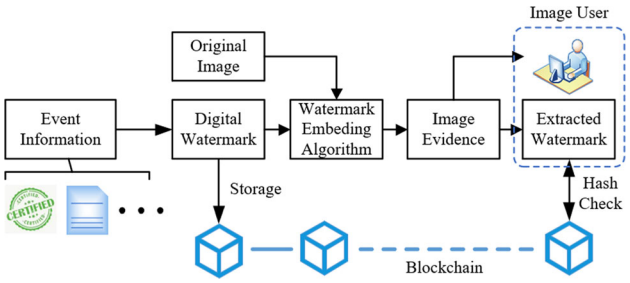

Fig. 2 The proposed overall framework for image evidence protection

efficient storage and safe use of image evidence in the field of forensic science. The overall framework is shown in Fig. 2, which consists of an image watermarking algorithm and smart contracts resided in the blockchain.

In the image forensics application scenario, the input information usually includes the original image, the identifier of the forensics institution, the event information and other related information. Compared with text content, images usually require more storage space, so directly store image data on the blockchain requires higher memory and computational resources. Therefore, in our framework, we take the key information of the event (e.g., event time, address, ID number, etc.) and the identifier of the forensics institution as the watermark information, and use the smart contract resided in the blockchain to store and protect them. At the same time, we use image watermarking algorithm to embed the watermark information to protect the authenticity and originality of the original image data. As shown in Fig. 2, we call the watermarked image as image evidence, which can be directly provided to image forensics experts or other users. When the users need to verify the authenticity and originality of the image, they only needs to use the image watermarking algorithm to extract the watermark information in the image; and then, check and verify it with the watermark stored in the smart contract. And thus, the image evidence protection framework proposed in this paper has low requirements on the computational and memory resources of the blockchain environments. By designing a reasonable smart contract, it can be deployed at the Internet-of-things(IoT) devices or other edge devices. Because the watermark information to be embedded in this framework is multimedia data, such as image marks and binary messages, we propose a new image watermark embedding algorithm based on self supervised learning, which can adaptively learn a suitable area in the latent space of the image according to its content for watermark embedding. In addition, the learned area in the latent space is more robust against to geometric and noise attacks.

3.2 Image watermarking based on self-supervised learning

Considering the powerful feature learning capability of deep learning, we propose a blind image watermarking model based entirely on deep neural networks, which means the embedding and detection of watermarks are done automatically by the deep networks. The structural of the proposed self-supervised learning-based image watermarking model is shown in Fig. 3. Similar to the mainstream image watermarking models based on deep learning [20, 38], our proposed self-supervised learning based image watermarking method also follows an encoder-and-decoder structure. For an input image, denoted as I_{in}, of arbitrary resolution and a watermark signal, denoted as S_{wm} to be embedded, the encoder generates a blind watermarked image as I_{wm}. For a secure blind watermarking encoder, it is expected that the I_{in} and I_{wm} should be as identical as possible not only for human eyes but also in distribution of deep features. Although this consistent similarity in appearance perception is able to deceive the attackers or analyzers, the corresponding decoder can extract the watermark signal from the watermarked image.

3.2.1 Detailed structure of encoder and decoder

As shown in Fig. 3, we first transform the original image I_{in} to the latent space (i.e., feature space); and then, we find a region in the latent space for embedding watermark signal S_{wm}. We expect that the image region $\mathcal{I}$ and watermark region $\mathcal{W}$ are separable with a marginal distance δ. Obviously, images with different contents will be distributed in different areas in the latent space. Consequently, we cannot guarantee that a fixed transformation model between the image space and the latent space can ensure that in the latent space, for any given image, its area $\mathcal{I}$ and watermark area $\mathcal{W}$ are non-overlap. Therefore, on the one hand, we need to train a transformation (usually done by a deep neural network, denoted as $x = \Phi(I), x \in \mathcal{I}$) so that the common geometric transformation and adjustment of pixel values (e.g. brightness, contrast, filter, and etc.) should have minimal impact on the image latent space. That is, for $I' = I_{in} + \Delta I$, $x' = \Phi(I')$ is still in or near $\mathcal{I}$. On the other hand, for a specific given image I_{in}, when we embed the watermark signal S_{wm} to the image by altering $\Phi(I_{in})$ with imperceptible perturbations, we can adopt some self-supervised learning algorithms to push the watermark area $\mathcal{W}$ far away from the image area $\mathcal{I}$.

For the watermark detection, we first transform the watermarked image I_{wm} into the latent space as $\hat{x} = \Phi(I_{wm})$, and then, extract the watermark signal by

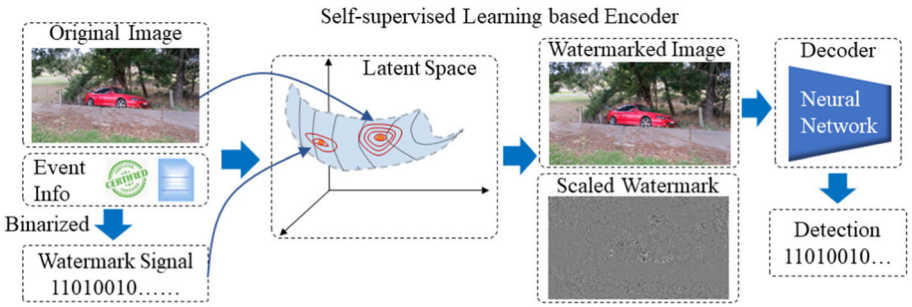

Fig. 3 The self-supervised learning-based watermarking model

projecting $\hat{x}$ to a predefined orthogonal extraction matrix. Since the extraction matrix is provided during the self-learning based watermarking embedding procedure, it can ensure that the image area and watermark area are separable with respect to the orthogonal extraction matrix. Here, we only consider the multi-bits watermark signal, denoted as $S_{wm} = (m_1, m_2, \cdots, m_k) \in \{-1, 1\}^k$, because both digital images and text messages can be binarized into a multiple 0-1 sequence. For multi-bits watermark decoding, we define a randomly sampled orthogonal family of carriers, denoted as $C = \{c_1, c_2, \cdots, c_k\} \in \mathbb{R}^d$, as the watermark extraction matrix, and then, modulate S_{wm} into the signs of the projection of the features in the latent space as $\Phi(I_{in})$ against each basis in the carrier as c_i. And thus, given a watermarked image I_{wm}, the watermark decoder is defined as

$$\hat{S}_{wm} = \left[\text{sign}\left(\hat{x}^\top c_1\right), \text{sign}\left(\hat{x}^\top c_2\right), \cdots, \text{sign}\left(\hat{x}^\top c_k\right)\right], \tag{1}$$

where $\hat{x} = \Phi(I_{wm})$.

For the deep neural network Φ, we choose a self-supervised learning (SSL) based pre-trained network because SSL based approaches do not need to assign class labels to the training samples, so the semantic collapse can be avoided. Here, we use DINO [39], proposed by Caron *et. al.*, as our pre-trained neural network, which is an end-to-end encoder-decoder structured object detector. By means of maximizing the similarity of representations spawning from augmented views of the same input image, DINO can explicitly trains invariant features for data augmentation. DINO leverages the concept of knowledge distillation, called self-distillation, which creates a student network and a teacher network, shared with the same architecture g, but with different sets of parameters θ_s and θ_t. The neural network g builds on a backbone network f (in this paper, ResNet-50 is used as the backbone network) and a multi-layer perception (MLP) projection head h, that is, $g = h \circ f$. In the initial step, two distorted views of an image are generated as x_1 and x_2, and then sent to the siamese teacher-student network. And then, the parameters of the network are updated according to a cross-entropy loss measured by the similarity of the features from both teacher and student networks. During the training procedure, only the student network is trained. After the student parameters are updated, the exponential moving average (EMA) strategy has been used to update the teacher's parameters as $\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s$ with $\lambda \leq 1$, where the parameters θ_t and θ_s denote the parameters of the teacher network and the student network, respectively.

Algorithm 1 Self-supervised Learning based Image Watermarking Algorithm

Input original image I_{in} and watermark signal S_{wm}
Output watermarked image I_{wm}

$I_{wm}^{(0)} = I_{in} \bigoplus S_{wm}$
for $i = 1, 2, \cdots, epochs$ **do**
  Data augmentation $\left\{I_{wm}^{(i)}\right\}_{t\in\mathcal{T}} \leftarrow \Gamma\left(I_{wm}^{(i)}, t\right)$
  Get feature $x = \Phi\left(I_{wm}^{(i)}\right)$
  Compute loss $\mathcal{L} = \lambda_w\mathcal{L}_w(x) + \lambda_i\mathcal{L}_i(I_{wm}^{(i)}, I_{in}) + \lambda_p\mathcal{L}_p(I_{wm}^{(i)}, I_{in})$
  Update watermarked image $I_{wm}^{(i+1)} = I_{wm}^{(i)} + \eta \times Grad(\mathcal{L})$
end for
return $I_{wm} = I_{wm}^{(i+1)}$

3.2.2 Watermark embedding algorithm

The watermark embedding algorithm embeds the watermark signal S_{wm} into the input image I_{in} and derive a visually invisible watermarked image I_{wm}. In the image latent space $\mathcal{I}$, we first add the watermark signal to the image features $x_{in} = \Phi(I_{in}) \in \mathcal{I}$ directly, and then, the gradient descent optimizer is applied with respect to the objective function to push the watermark signal into a region $\mathcal{W}$ that is visually insensitive when transform back to the image space. The objective loss function $\mathcal{L}$ should consider the following three conditions: (1) the image feature x_{in} lies far away from $\mathcal{W}$ with a marginal distance δ; (2) the watermarked image I_{wm} should be as close as possible to the input image I_{in}; and (3) the visual perceptual quality of the watermarked image I_{wm} and the input image I_{in} should be as same as possible.

For the first condition, according to the decoder defined in Eq. (1), we use a hinge loss $\mathcal{L}_w$ to constrain the image features $x = \Phi(I_{wm})$ and watermark signals $S_{wm} = (m_1, m_2, \cdots, m_k)$ with a marginal distance $\delta \geq 0$ in the latent space:

$$\mathcal{L}_w(x) = \frac{1}{k}\sum_{i=1}^{k}\max(0,\ \delta - \langle x, c_i\rangle m_i), \tag{2}$$

where c_i is the orthogonal basis belongs to the carrier matrix. For the second requirement, we use mean square error (MSE) to measure the distortion between watermarked image I_{wm} and input image I_{in}:

$$\mathcal{L}_i(I_{wm}) = \frac{1}{N}\|I_{wm} - I_{in}\|_2^2, \tag{3}$$

where N denotes the total number of pixels in the image. For the objective of visual perceptual quality, we use perceptual loss based on the features derived by the VGG network Ψ[40]:

$$\mathcal{L}_p(I_{wm}) = \frac{1}{C_j H_j W_j}\|\Psi_j(I_{wm}) - \Psi_j(I_{in})\|_2^2, \tag{4}$$

where j represents the j-th layer of the network Ψ, and C_j, H_j, W_j represents the channels and the size of the feature map of the j-th layer, respectively. Therefore, the entire loss function of the proposed watermarking algorithm can be defined as

$$\mathcal{L}(I_{wm}) = \lambda_w\mathcal{L}_w(\Phi(I_{wm})) + \lambda_i\mathcal{L}_i(I_{wm}) + \lambda_p\mathcal{L}_p(I_{wm}), \tag{5}$$

where λ_w, λ_i and λ_p are hyper-parameters.

Besides the objective function, data augmentation is still needed to ensure the robustness of the watermarking algorithm against image geometric transformation and noise attacks such as rotation, resize, compression, and etc. So, we first build a set of image transformation operators $\mathcal{T}$ including crop, resize, rotation, Gaussian blur, brightness and contrast adjustment, and JPEG compression (as shown in Table 2). And then, after the watermark signal been embedded into the image and before the optimization process, we applied the transformation operator $t \in \mathcal{T}$ to the initial watermarked image $I_{wm}^{(0)}$ to derived a transformed watermarked image as $\Gamma\left(I_{wm}^{(0)}, t\right)$. By applying different operators with different parameters, an augmented watermarked image dataset can be derived as $\left\{I_{wm}^{(0)}\right\}_{t\in\mathcal{T}}$. This image dataset is then used to generate the final watermarked image via the optimization procedure. Since the quality constraints, image transformation, and feature extractor are differentiable with respect to the pixel values, the minimization can be performed by stochastic gradient descent. The final watermarked image is the rounded version after updating k epochs. Therefore, the total self-supervised learning based image watermarking algorithm is summarized in Algortihm 1.

Algorithm 2 Save-check smart contract.

```
1: ▷ save watermarks
2: function SAVE(addr, watermark, Hash_watermark, Hash_image)
3:     W ← watermark
4:     H ← Hash_watermark
5:     H' ← Hash_image
6: end function
7: ▷ check a watermark
8: function CHECK(addr, watermark)
9:     h = Hash(watermark)
10:    n = size(H)
11:    for i ≤ n do
12:        if h = H[i] then
13:            ▷ the watermark has not been tampered
14:            return 1
15:        end if
16:    end for
17:    ▷ the watermark has been tampered
18:    return 0
19: end function
```

3.3 Smart contract

In order to conventionally save and check a digital watermark, we design a save-check smart contract denoted by Scsc. The contract is formally described in Algorithm 2. Scsc is deployed on the blockchain by a digital watermark authorizer, and is mainly used by digital watermark authorizers and digital watermark users. The main function of Scsc is to store digital watermarks and complete the verification of the digital watermark. A specific workflow with respect to Scsc is as follows:

1. A digital watermark authorizer deploys a save-check smart contract Scsc on the blockchain, where Scsc contains a function SAVE (at line 2 of Algorithm 2), which is used to save the watermark and hash value of image evidences, and a function CHECK (at line 6 of Algorithm 2), which is used for verifying digital watermarks.
2. Once Scsc is successfully deployed, a digital watermark authorizer extracts a digital watermark from a digital image evidence, and then stores the digital watermark and its hash value into the save-check smart contract by generating a special transaction to call the function SAVE of Scsc.
3. To verify the validity of a digital watermark, a digital watermark user first extracts a digital watermark from the digital image evidence it received by using the image watermarking algorithm mentioned above. Then the digital watermark user uses the hash value of the extracted digital watermark to call the function CHECK of Scsc by a special transaction. Upon received the hash value, CHECK of Scsc will match the hash value with the hash values of other digital watermarks saved

in it. If matched, it means that the digital watermark has not been tampered with; otherwise, the digital watermark has been tampered with.

Note that the watermark and hash value of image evidences saved on Scsc resided in the permission blockchain is trustworthy because the above information are maintained by the relevant law enforcement authorities running the permission blockchain protocol. On the other hand, the watermark and hash value of image evidences stored in the permission blockchain does not reveal the privacy of raw image evidences. Even if these information can be accessed by anyone, this cannot recover the raw image evidence information from them. In our framework, one purpose of the permission blockchain is pretecting the watermark of image evidences, since a malicious entity may tamper with the watermark information in the process of the watermark circulation.

4 Experimental results

In this section, we evaluate our proposed image evidence framework from the two aspects: (1) the performance of watermarking algorithm compared with other state-of-the-art (SOTA) methods; and (2) the computational cost comparisons of the designed smart contract with and without the watermarks.

4.1 Dataset and experiment preparation

In our experimental setup, two datasets are used for the performance evaluation. One includes hundreds of vehicle accident images, with the image size of 600×400 pixels, selected by ourselves used for functional validation of our proposed image evidence protection framework and the designed smart contract, because the purpose of this work is to preserve and authenticate image evidence efficiently and economically. The other is the UTKFace dataset [41], with the image size of 200×200 pixels, including faces of a wide range of ages, which is used for the metrics comparison between our proposed self-supervised learning based watermarking algorithm and other state-of-the-art watermarking methods.

For the image watermarking algorithm (as in Alg. 1), both of these two datasets are used for the performance evaluation, including the comparisons of visual quality and watermark extraction accuracy. For the computational cost analysis of smart contract, we only use the first dataset with watermarked images. The image watermarking model presented was trained on a GPU server equipped with two NVIDIA Tesla T4 in the PyTorch framework. The ResNet-50 architecture is used as the backbone model for extracting features from its last convolutional layer with dimension of 2048 ($d = 2048$). It was trained on the unlabeled ILSVRC2012 dataset using default parameters and rotation increments after 300 epochs of DINO self-supervised learning method. The save-check smart contract used in the proposed framework was deployed in a local platform equipped with Intel Core i7-12700KF CPU 3.60GHz, 32.00GB RAM and Windows 10 operating systems.

4.2 Performance comparison on image watermarking

We first compare the visual qualities of the watermarked images; and then, we compare the watermark extraction accuracy under different attacks. Because for multi-bits watermarking algorithms, the visual quality of the watermarked images is closely related to the length (i.e., payload size) of the watermark signal, we set the payload k of 32 random bits as in Eq. (1). For the hyper-parameters in Eq. (5), λ_w, λ_i and λ_p are set exactly the same as 10^4, 1.0, and 1.0 in all our following experiments, respectively. Figure 4 shows the perceptual visual qualities of the watermarked images derived by our proposed Algorithm 1, in which the top row, second row and bottom row represents the original images from UTKFace, the corresponding watermarked images, and the scaled pixel-wise differences between the original image and the watermarked image, respectively. The watermark is almost perceptual invisible to human eyes because most of the watermark information is added into the texture regions as shown in the bottom row in Fig. 4.

Then, we perform the quantitative comparisons of visual qualities of the watermarked images derived by our proposed method and other SOTA methods, including traditional transformation-based approach QDFT [13] and deep learning-based models, i.e., Hayes' method [42], HiDDen [20] and Pierre's method [22]. Two standard reference-based image quality evaluation metrics, Peak Signal to Noise Ratio (PSNR) and Fréchet Inception Distance (FID) [43], are used here, which are defined in Eqs. (6) and (7), respectively.

$$PSNR(I_{in}, I_{wm}) = 10 \log_{10} \frac{255^2 \times 3 \times N}{\|I_{in} - I_{wm}\|^2}, \tag{6}$$

where N denote the number of pixels of the input image I_{in}; and

$$FID(I_{in}, I_{wm}) = \|\mu_{in} - \mu_{wm}\|^2 + \mathrm{tr}\left(\Sigma_{in} + \Sigma_{wm} - 2(\Sigma_{in}\Sigma_{wm})^{1/2}\right), \tag{7}$$

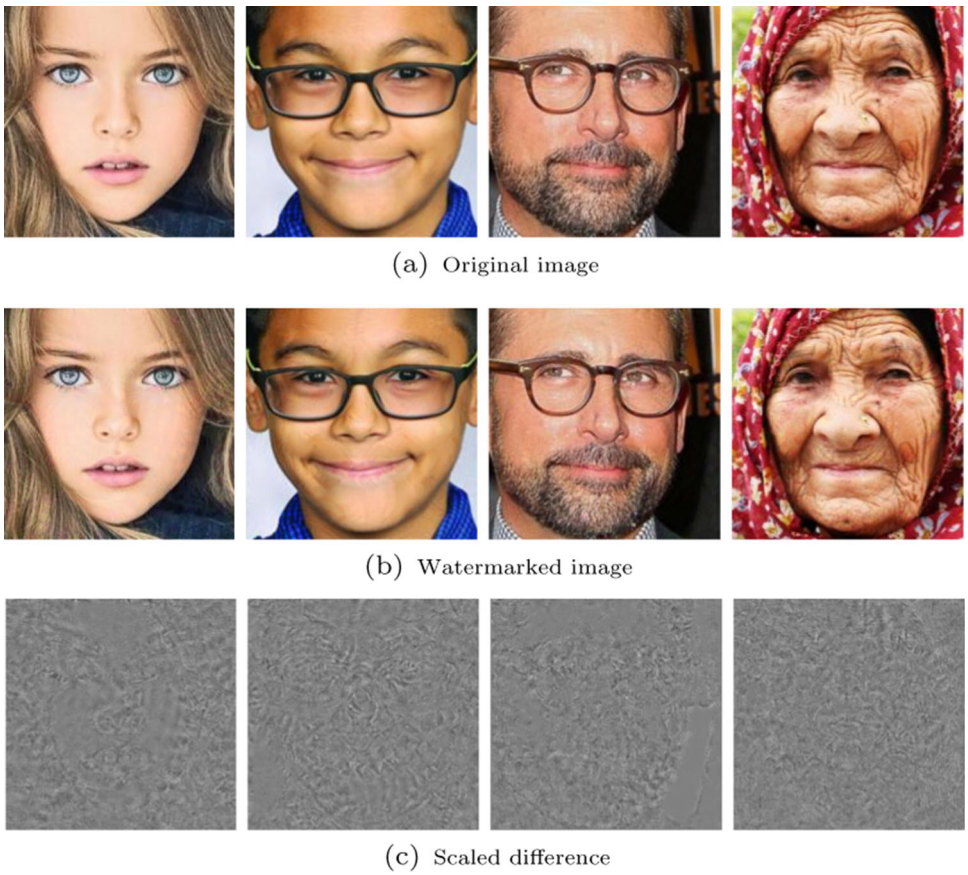

Fig. 4 Visual impact comparison before and after watermark embedding. The top row, second row and bottom row is the original images from UTKFace, the corresponding watermarked images, and the scaled pixel-wise difference w.r.t. the original images, respectively

where μ and Σ denotes the mean vector and co-variance matrix of the deep features of the image as $\Psi(I)$, respectively.

From the definitions of PSNR and FID in Eqs. (6–7), we can find that the PSNR and FID index reflects the pixel-level difference of two images and visual perceptual feature distribution distance of two images, respectively. Table 1 shows the mean values of these two metrics derived by our proposed method and the other watermark methods on the UTKFace dataset. As shown in Table 1, our proposed method outperforms other watermark methods in both of these metrics, which means that the watermarked images derived by our proposed method have the lowest distortion in pixel-level and visual perceptual level. However, since our proposed self-supervised learning-based model needs to learn the appropriate embedding area in the latent space according to the content of the given image, as described in Algorithm 1, the computational cost of our proposed watermarking model is higher than the traditional and other deep learning-based methods. In our experimental environment, as described in Sect. 4.1, it usually takes several minutes to complete the watermark embedding process, but the watermark detection process can be finished in real time.

Then, we compare the detection accuracy and robustness of different image watermarking methods under different attacks, e.g., image blur, image rotation, JPEG compression, etc. Since these attacks often occur in the procedure of image transmission and processing, we roughly divide them into two categories: one is called pixel-level noise attacks, including Gaussian blur, brightness and contrast adjustment, and JPEG compression to the image; and the other is called image geometric attacks, including image resize, image rotation and image crop. The vehicle accident image dataset was used for this experiment because it is close to the practical applications of our proposed image evidence protection framework. Figure 5 shows an example of pixel-level noise and geometric

Table 1 Distortion comparisons of different watermarks algorithms in PSNR and FID

Models	QDFT [13]	Hayes [42]	HiDDen [20]	Pierre [22]	Ours
PSNR (higher is better)	25.21	24.37	29.08	41.79	**41.86**
FID (lower is better)	31.19	30.06	23.31	17.96	**14.22**

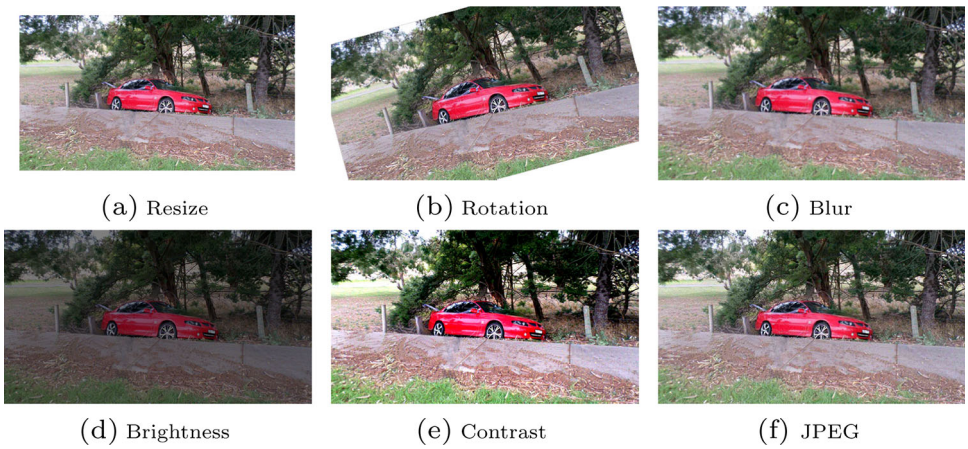

Fig. 5 An example of various attacks of watermarked image

attacks to the watermarked image. After the watermarked image been attacked by these pixel-level noise and geometric attacks with different parameters, we evaluate the detection accuracy of different methods for extracting the watermark signals from these attacked watermarked images. The accuracy rate used here is defined by true positive rate (TPR), which represents the number of watermark samples that are correctly extracted to the total number of watermark samples that are embedded to the images.

Table 2 shows the mean value of accuracy results derived by our proposed method, HiDDen model [20], and Pierre's method [22] in our dataset under different kinds of attacks with specific parameters, e.g., crop with a ratio of 0.8, rotate image with 20 degree, Gaussian blur with a kernel size of 9, and so on. As shown in Table 2, our proposed method derives almost all the best accuracy except the brightness attack, in which Pierre's method get the best result. However, under brightness attacks, the accuracy rates derived by these two methods are very close and both of them are acceptable.

Then, we evaluate the robustness of our proposed method under different attacks with different parameters. As shown the second row in Table 2, all the deep learning based methods can derive a 100% accuracy under the crop attack with a crop ratio of 0.8. Since enough image content should be provided to make a judgement in the practical application of image evidence identification in forensic science, it is meaningless when the crop ratio is too small, for instance, lower than 0.8. Therefore, we only evaluate the variation of accuracy to the other six attacks, including resize, rotation, blur, brightness, contrast, and JPEG compression, with different parameters in Fig. 6. As shown in Fig. 6, our proposed image watermarking method is very robust to the brightness and contrast adjustment. For image rotation and Gaussian blur attacks, our method also performs quite good when the degree of rotation is less than 25° or the size of blur kernel is less than 9×9. Although the watermark extraction accuracy of our method drops rapidly when the ratio of resize and the quality factor of JPEG compression is, respectively, less than 0.4 and 50, this drop is also reasonable because when the parameters of these two attacks are less than the threshold, the amount of information in the image will decrease very quickly. However, in general, the performance of our proposed image watermarking method is still better than the other deep learning based methods and can meet the requirements of practical application.

Table 2 Detection accuracy of different watermarking algorithms under different attacks

Methods		HiDDen [20]	Pierre [22]	Ours
Attacks	Crop (0.8)	1.0	1.0	**1.0**
	Resize (0.8)	0.92	1.0	**1.0**
	Rotation (∠20°)	0.71	0.83	**0.99**
	Blur (9)	0.98	0.75	**0.95**
	Brightness (0.5)	0.95	**0.99**	0.96
	Contrast (2.0)	0.94	1.0	**1.0**
	JPEG (50)	0.82	0.89	**0.92**

4.3 Performance of smart contract

In this subsection, we evaluate our save-check smart contract used in the proposed framework. The contract is designed by Solidity language, and is implemented on a local computer based on Remix. The performance of executing the save-check smart contract is shown in Table 3. The executing cost of a smart contract is mainly measured by gas, thus measuring gas is essential. The transaction gas and execution gas of each function in the save-check smart contract are the average of about 100 experimental results. The gas cost of executing the save-check smart contract is shown in Table 3. Where the gas cost of executing the function SAVE is cheap and about 0.0001 ether. The gas cost

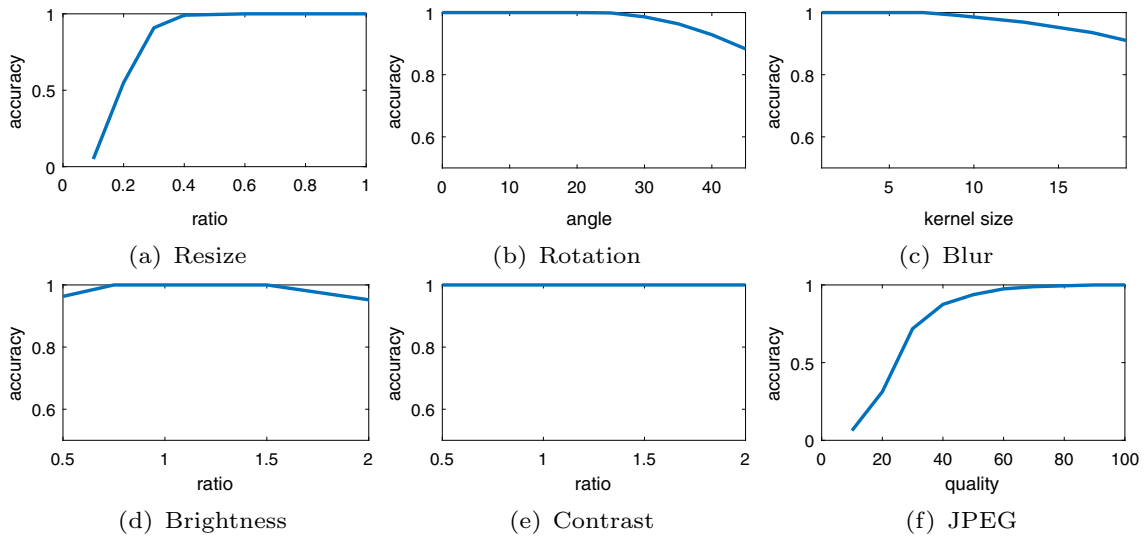

Fig. 6 Robustness of the detection accuracy against different image transformations

Table 3 Performance of smart contract

Smart contract	Save-check		
Function	DEPLOY CONTRACT	SAVE	CHECK
Transaction gas	358950	70213	116086
Execute gas	185179	57280	86326
Gas cost (ether)	0.00054413	0.00012749	0.00020241
Execute time (*ms*)	1.1	0.9	36

of running the function CHECK is higher than that of SAVE, this is because the function CHECK performs multiple matching operations for the hash value of digital watermarks to be verified. In addition, we also evaluate the executing time of save-check smart contract. Since the functions in the smart contract are triggered in the form of transactions, we use the time when a transaction is included in the blockchain to measure the executing time of each function in the smart contract. Here, the executing time of the functions DEPLOY CONTRACT and SAVE are almost identical. Compared to the above two functions, the function CHECK takes slightly longer to execute due to the matching operations for the hash value of digital watermarks. According to the experimental result above, therefore, the execution of our save-check smart contract is efficient.

5 Conclusion

In this work, we present a novel image evidence protection framework in the image forensics scenario based on digital image watermarking and blockchain technology. First, a multi-bits digital image watermarking algorithm has been proposed, which can embed the watermark signals containing additional event information related to the image. In addition, the proposed image watermarking algorithm is based on self-supervised learning algorithm, which means that the watermarked image is more robust against to geometric and noise attacks than the traditional image watermarking models with fixed transform basis. Furthermore, because the watermark signal stored in the smart contract includes the event information corresponding to the image, when the watermarked image (as the image evidence) is provided to the image forensic experts or other users, both the image and its corresponding event information can be authenticated after checking and verifying the watermark extracted from the image with the watermark stored in the smart contract. Finally, the experimental results have demonstrate the feasibility and efficacy of the proposed framework.

Albeit our proposed framework can be viewed as a firststep to the image evidence protection, further reducing the distortion of watermarked image as well as extending to the video evidence preservation are important issues that warrant further study. In addition, for multi-bits watermarking algorithms, the limitation of the payload size is not only related to the number of pixels in the image, but, more important, also to the content of the image. So, it is difficult to provide an accurate limitation of the payload size according to the number of pixels of an image, especially for self-supervised learning-based image watermarking models. Therefore, our future study also includes the exploration of the limitation of the payload size of our

proposed watermarking algorithm according to the given image.

Funding There is no funding for this study.

Data Availability The data are not publicly available due to them containing information that could compromise research participant consent.

References

1. Li, Z., Liu, Y., Hu, X., & Wang, G. (2022). A new uniform framework of source attribution in forensic science. *Humanities and Social Sciences Communications, 9*, 1–11.
2. Senkyire, I.B., & Kester, Q.-A. (2019) Validation of forensic crime scene images using watermarking and cryptographic blockchain. In *2019 international conference on computer, data science and applications (ICDSA)* (pp. 1–4).
3. Xu, X., Fang, Z., Zhang, J., He, Q., Yu, D., Qi, L., & Dou, W.-C. (2021). Edge content caching with deep spatiotemporal residual network for IOV in smart city. *ACM Transactions on Sensor Networks, 17*, 29–12933.
4. Ren, J., Li, J., Liu, H., & Qin, T. (2022). Task offloading strategy with emergency handling and blockchain security in SDN-empowered and fog-assisted healthcare IOT. *Tsinghua Science and Technology, 27*, 760–776.
5. Barni, M., Podilchuk, C. I., Bartolini, F., & Delp, E. J. (2001). Watermark embedding: Hiding a signal within a cover image. *IEEE Communications Magazine, 39*(8), 102–108. https://doi.org/10.1109/35.940048
6. Dai, H., Zheng, Z., & Zhang, Y. (2019). Blockchain for internet of things: A survey. *IEEE Internet of Things Journal, 6*, 8076–8094.
7. Yuan, L., He, Q., Chen, F., Zhang, J., Qi, L., Xu, X., Xiang, Y., & Yang, Y. (2021). Csedge: Enabling collaborative edge storage for multi-access edge computing based on blockchain. *IEEE Transactions on Parallel and Distributed Systems, 33*, 1873–1887.
8. Wang, Z., Zhang, F., Yu, Q., & Qin, T. (2021). Blockchain-envisioned unmanned aerial vehicle communications in space-air-ground integrated network: A review. *Intelligent and Converged Networks, 2*, 277–294.
9. Wang, B., Jiawei, S., Wang, W., & Zhao, P. (2022). Image copyright protection based on blockchain and zero-watermark. *IEEE Transactions on Network Science and Engineering, 9*, 2188–2199.
10. Hartung, F., & Kutter, M. (1999). Multimedia watermarking techniques. *Proceedings of the IEEE, 87*(7), 1079–1107. https://doi.org/10.1109/5.771066
11. Sandhu, A. K. (2022). Big data with cloud computing: Discussions and challenges. *Big Data Mining Analysis, 5*, 32–40.
12. Xu, X., Tian, H., Zhang, X., Qi, L., He, Q., & Dou, W. (2022). Discov: Distributed COVID-19 detection on x-ray images with edge-cloud collaboration. *IEEE Transactions on Services Computing, 15,* 1206–1219.
13. Wang, X.-Y., Wang, C.-P., Yang, H., & Niu, P.-P. (2013). A robust blind color image watermarking in quaternion Fourier transform domain. *Journal of Systems and Software, 86*, 255–277.
14. Jamal, S. S., Shah, T., Farwa, S., & Khan, M. U. (2019). A new technique of frequency domain watermarking based on a local ring. *Wireless Networks, 25*, 1491–1503.
15. Wang, X., Hu, K., Hu, J., Du, L., Ho, A. T. S., & Qin, H. (2020). Robust and blind image watermarking via circular embedding and bidimensional empirical mode decomposition. *The Visual Computer, 36*, 2201–2214.
16. Verma, V. S., Jha, R. K., & Ojha, A. (2015). Digital watermark extraction using support vector machine with principal component analysis based feature reduction. *Journal of Visual Communication and Image Representation, 31*, 75–85.
17. Hu, X., Peng, S., & Hwang, W.-L. (2012). Emd revisited: A new understanding of the envelope and resolving the mode-mixing problem in AM-FM signals. *IEEE Transactions on Signal Processing, 60*(3), 1075–1086.
18. Volkhonskiy, D., Nazarov, I., Borisenko, B., & Burnaev, E. (2017). Steganographic generative adversarial networks. In *Proceedings of NIPS 2017 workshop on adversarial training* (Vol. 3, pp. 201–208).
19. Shi, H., Dong, J., Wang, W., Qian, Y., & Zhang, X. (2017). SSGAN: Secure steganography based on generative adversarial networks. *Advances in Multimedia Information Processing: PCM*. https://doi.org/10.1007/978-3-319-77380-3_51.
20. Zhu, J., Kaplan, R., Johnson, J., & Fei-Fei, L. Hidden: Hiding data with deep networks. In *ECCV* (2018)
21. Wei, Q., Wang, H., & Zhang, G. (2020). A robust image watermarking approach using cycle variational autoencoder. *Security and Communication Networks, 2020*, 8869096–188690969.
22. Fernandez, P., Sablayrolles, A., Furon, T., J'egou, H., & Douze, M. (2022). Watermarking images in self-supervised latent spaces. In *ICASSP*
23. Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. *Decentralized Business Review* (pp. 21260).
24. Lin, S., Zhang, L., Li, J., Ji, L., & Sun, Y. (2022). A survey of application research based on blockchain smart contract. *Wireless Networks, 28*, 635–690.
25. Qi, L., Yang, Y., Zhou, X., Rafique, W., & Ma, J. (2022). Fast anomaly identification based on multiaspect data streams for intelligent intrusion detection toward secure industry 4.0. *IEEE Transactions on Industrial Informatics, 18*, 6503–6511.
26. Lewenberg, Y., Sompolinsky, Y.,& Zohar, A. (2015). Inclusive block chain protocols. In *International conference on financial cryptography and data security* (pp. 528–547). Springer
27. Bhushan, B., Sahoo, C., Sinha, P., & Khamparia, A. (2021). Unification of blockchain and internet of things (BIoT): Requirements, working model, challenges and future directions. *Wireless Networks, 27,* 55–90.
28. Kiayias, A., Russell, A., David, B., & Oliynykov, R. (2017). Ouroboros: A provably secure proof-of-stake blockchain protocol. In *Annual international cryptology conference* (pp. 357–388) . Springer
29. Yin, M., Malkhi, D., Reiter, M.K., Gueta, G.G., & Abraham, I. (2019). Hotstuff: Bft consensus with linearity and responsiveness. In *Proceedings of the 2019 ACM symposium on principles of distributed computing* (pp. 347–356).
30. Lu, Y., Lu, Z., Tang, Q., & Wang, G. (2020). Dumbo-mvba: Optimal multi-valued validated asynchronous byzantine agreement, revisited. In *Proceedings of the 39th symposium on principles of distributed computing* (pp. 129–138)
31. Szabo, N. Formalizing and securing relationships on public networks. First monday (1997).
32. Wood, G., et al. (2014). Ethereum: A secure decentralised generalised transaction ledger. *Ethereum Project Yellow Paper, 151*(2014), 1–32.

33. Cao, D., & Chen, W. (2019). Mechanism of trusted storage in Ethereum based on smart contract. *Journal of Computer Applications, 39,* 1073–1080.
34. Brotsis, S., Kolokotronis, N., Limniotis, K., Shiaeles, S.N., Kavallieros, D., Bellini, E., & Pavué, C. (2019). Blockchain solutions for forensic evidence preservation in IOT environments. In *2019 IEEE conference on network softwarization (NetSoft)* (pp.110–114).
35. Zou, R., Lv, X., & Wang, B. (2019). Blockchain-based photo forensics with permissible transformations. *Computers & Security, 87,* 101567.
36. Chen, T. et al. (2021) An image copyright protection method using zero-watermark by Blockchain and IPFs. *Journal of Information Hiding and Privacy Protection 3* (3), 131–142 . 10.32604/jihpp.2021.026606
37. Jyoti, A., & Chauhan, R. K. (2022). A blockchain and smart contract-based data provenance collection and storing in cloud environment. *Wireless Networks, 28*, 1541–1562.
38. Ahmadi, M., Norouzi, A., Karimi, N., Samavi, S., & Emami, A. (2020). Redmark: Framework for residual diffusion watermarking based on deep networks. *Expert Systems with Applications, 146,* 113157.
39. Caron, M., Touvron, H., Misra, I., J'egou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)* (pp. 9630–9640).
40. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition* (pp. 586–595).
41. Zhang, Z. Song Y. & Qi, H. (2017). Age progression/regression by conditional adversarial autoencoder. In *IEEE conference on computer vision and pattern recognition (CVPR)*. IEEE
42. Hayes, J. & Danezis, G. (2017). Generating steganographic images via adversarial training. In *NIPS*
43. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., & Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*

Qing Yao received the B.S. and M.S. degrees both in Nanjing University of Posts and Telecommunications in 2010 and 2013, respectively. Currently, he is pursuing the Ph.D. degree in the Institute of Information Engineering, Chinese Academy of Sciences and University of Chinese Academy of Sciences, China. His research interests include forensic science, blockchain, and applied cryptography.

Kaiwen Xu received the B.S. degree in the School of Computer Science and Technology, Nanjing Tech University Pujiang Institute, Nanjing, China, in 2020. He is currently working toward the Ph.D. degree in Computer Science and Technology with the Nanjing University of Science and Technology, Nanjing, China. His research interests include Deep Generative Model and Anomaly Detection.

Taotao Li received the Ph.D. degree in cyber security from Institute of Information Engineering, Chinese Academy of Sciences and University of Chinese Academy of Sciences, China, in 2022. He is currently a postdoc with the School of Software Engineering, Sun Yat-Sen University, Zhuhai, China. His main research interests include blockchain, Web3, applied cryptography.

Yichao Zhou received the B.S., M.S., and Ph.D. degrees in computer science from Nanjing University of Science and Technology in 2005, 2008, and 2018, respectively. He is currently an associate professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology. His current research interests include artificial intelligence applications in physiological signal analysis and biomedical image processing.

Mingsheng Wang received the Ph.D. degree from the School of Mathematical Sciences, Beijing Normal University, in 1994. He is currently a Research Professor with the State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences. His main research interests include symbolic computation, coding and cryptography theory, in particular, exploring algebraic attacks and the classification questions of some cryptographic functions.