Hindawi
Security and Communication Networks
Volume 2023, Article ID 6295486, 11 pages
https://doi.org/10.1155/2023/6295486

Research Article

Exploring Frame Difference to Enhance Robustness for Video Steganography on Social Networks

Pingan Fan,[1,2] **Hong Zhang**,[1,2] **and Xianfeng Zhao**[1,2]

[1]*State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100085, China*
[2]*School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100085, China*

Correspondence should be addressed to Xianfeng Zhao; zhaoxianfeng@iie.ac.cn

Received 3 January 2023; Revised 1 May 2023; Accepted 28 July 2023; Published 16 August 2023

Academic Editor: Zhili Zhou

The proliferation of video sharing on social networks has created a novel avenue for covert communication. Since most social networking channels are lossy, researchers have devoted efforts to robust video steganography to establish covert communication on social networks. Although there are various methods available, they often overlook the differences between frames in a video and are designed for a single frame. In this paper, we explore the general process of video recompression and present the frame quantization step (FQS) and interframe mutual information (IFMI) to measure the frame difference in the process of video recompression. Based on the two proposed metrics, we design a heuristic frame selection strategy and then propose a new robust video steganographic method in the DWT-SVD domain. Extensive experiments demonstrate that heuristic frame selection can effectively improve the robustness and reduce the computational complexity of video steganography. Our proposed method provides better robustness and higher efficiency than existing methods for building reliable covert communication on social networks, such as YouTube and Vimeo.

1. Introduction

Steganography is the science of concealing secret messages within digital media without detection. Due to the widespread use of social media platforms and advancements in video coding technology, video sharing on the Internet has become an increasingly prevalent trend. It has spurred the development and research of both high-level and low-level tasks based on videos, such as object recognition and tracking [1, 2], video denoising [3], and video compression [4]. The behavior of video sharing serves as an ideal cloak for covert communication [5]. In recent years, video steganography has emerged as a prominent research topic within the field of information hiding.

Through the utilization of popular social networking channels such as YouTube and Vimeo, hidden communication can be accomplished by disseminating stego videos that carry secret messages. Covert messages can be disseminated to numerous recipients without being detected by regular users. The concealed sender-receiver relationship ensures complete protection of secret communication. Nonetheless, most video-sharing platforms employ lossy processing techniques on the uploaded multimedia data. The utilization of lossy processing can introduce errors in message extraction, which include but are not limited to video recompression, geometric attacks, and visual enhancement techniques. Video recompression is a widely used method on the Internet, which can highly reduce the transmission bandwidth and save the storage space. Generally, video recompression mechanisms are unknown on social networks, and only the input and output can be obtained. To construct reliable hidden communication, researchers should pay more attention to robust video steganography under lossy black-box channels.

The current focus of research on robust video data hiding has been predominantly centered on spatial-transform domains in recent years. There are many robust watermarking methods but relatively few robust steganographic methods. A robust video watermarking method was presented by Huan et al. in [6], which implanted a watermark in each video frame by changing the coefficients in the joint subbands of the

DTCWT (dual tree-complex wavelet transform) domain. Although this method is robust against a range of geometric attacks, its embedding capacity is limited due to the repetitive watermark embedding. Sadek et al. [7] introduced a video steganographic technique that utilized human skin regions to hide secret messages. Nonetheless, the skin detection algorithm utilized in this approach exhibits a notable decline in accuracy following video recompression. A robust video steganographic technique designed by Fan et al. [8] aimed to mitigate social networking transcoding by embedding messages in the DWT-SVD (discrete wavelet transform-singular value decomposition) domain. However, this method is both time-consuming and inefficient, requiring previous embedding and extraction during the message embedding process.

Although some advancements have been made in current robust video data-hiding techniques, the majority still treat videos as a sequence of consecutive frames, embedding messages based on each individual frame. In the process of message embedding, there is no distinction between frames in a video. However, the difference between frames does exist and cannot be ignored. Actually, video compression and recompression encode frames in a video into bitstreams of different lengths. The used codec tries to get an optimal trade-off between visual distortion and bitstream length, known as rate-distortion optimization. The frames within a video are superimposed with different levels of video encoding noise. Even though the same hiding method is used to deal with these frames, there is still a robustness difference between frames in a video.

In this paper, we try to explore the robustness difference between frames within the same video. By analyzing the process of general video coding, we provide two evaluation metrics to quantify the difference between frames, called the frame quantization step and interframe mutual information. And then, we present a heuristic frame selection method based on the two metrics. The optimal frames are selected to carry secret messages. In our approach, we make use of the luminance (Y) component and extract coefficients within the DWT-SVD domain for a large embedding capacity. To decrease the BER of the transmitted message, ECC (error correction code) is also integrated into our approach. Experimental results demonstrate that employing the proposed heuristic frame selection enhances both the robustness and efficiency of message embedding. Compared with some existing methods, our proposed method is more robust and reliable in constructing hidden communication over social networks, such as YouTube and Vimeo.

The rest of this paper is organized as follows. In Section 2, some related work is described. In Section 3, we explore the difference between frames in a video. Section 4 explains the details of heuristic frame selection, message embedding, and message extraction. Section 5 presents the experimental outcomes, while Section 6 concludes the paper and examines potential future research.

2. Related Work

Traditional multimedia steganography utilizes signal processing techniques and manually modulates features for message embedding. However, there are also attempts to utilize deep learning to achieve end-to-end steganography [9, 10]. As deep learning-based steganography is unable to resist JPEG or video compression, robust steganographic methods are generally designed based on signal processing techniques. Embedding domain construction and coefficient modulation are two main parts of robust steganography.

2.1. Embedding Domain. In robust steganography, spatial and spatial-transform domains are two common types of embedding domains. Commonly used spatial domains include RGB (red, green, and blue) color components [11], YUV (luminance, chrominance) components [12], or color histograms [13, 14]. In comparison to spatial domains, spatial-transform domains provide superior imperceptibility and robustness. These domains include the DCT (discrete cosine transform) domain [15], the DWT domain [16, 17], the SVD domain [18], and the DTCWT domain, among others. Furthermore, it is a common practice to combine various transformations to form joint embedding domains, such as the DWT-DCT domain [19], DWT-SVD domain [8, 20], and DTCWT-SVD domain [6, 21].

2.2. Coefficient Modulation. The technique of coefficient modulation is implemented to alter the input cover elements for message embedding. There are broadly three kinds of schemes for coefficient modulation, such as least significant bit modulation, spread spectrum modulation [22, 23], QIM (quantization index modulation) [24–26], and coefficient correlation-based modulation [6, 27]. Considering the difference between coefficients to be modified, some researchers also design coefficient selection strategies to further enhance robustness. Fan et al. [8] proposed a frame selection strategy to improve robustness. Huan et al. [6] selected some reasonable coefficients in the process of message extraction.

3. Differences between Frames

In this section, we try to explore the robustness difference between frames in the process of video compression and recompression. First, we explain the general procedure of video encoding and decoding in common video codecs. Second, we analyze the noise source during lossy compression or recompression. Third, two evaluation metrics are proposed to measure the robustness difference between frames, called the frame quantization step and interframe mutual information.

3.1. General Procedures for Video Encoding and Decoding. The processes of video encoding and decoding are broadly similar in common video codecs, including H.246/AVC [28] and H.265/HEVC [29]. To explore the difference between frames, we briefly introduce the general procedures of video encoding and decoding.

As shown in Figure 1, video encoding includes intraframe and interframe prediction, DCT, quantization, and entropy encoding. Intraframe and interframe prediction

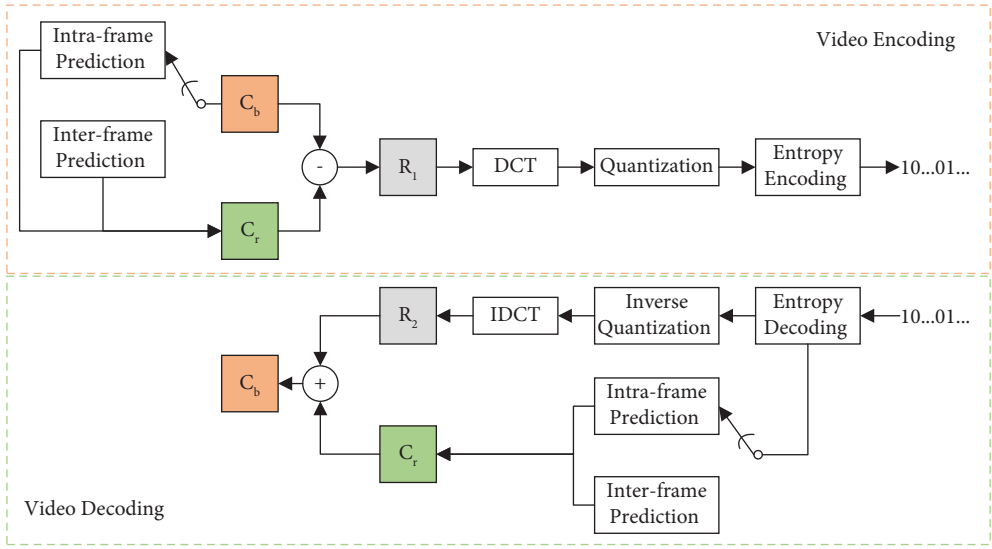

FIGURE 1: Schematic diagram of the general video encoding and decoding procedures. $\mathbf{C}_b$ and $\mathbf{C}_r$ are the current macroblock and the reference macroblock, respectively. Their residual is denoted as $\mathbf{R}_1$. $\mathbf{R}_2$ is the quantized version of $\mathbf{R}_1$.

helps determine the optimal reference macroblock. Then, the residual between the current macroblock and reference macroblock is transformed to get the DCT coefficients, which are then quantized based on the preset quantization parameter. Due to the difference in the preset rate control parameter, the assigned quantization parameter of the current macroblock is different in different video codecs. Video decoding aims to restore the current macroblock by parsing the corresponding bitstream, which is the inverse process of video encoding.

3.2. Video Compression Noise. In the process of video encoding, quantization is the only procedure that introduces video compression noise. Empirically, the level of video compression noise is mainly related to the size of the assigned quantization step.

Let us denote the residual of the current macroblock as $\mathbf{R}$. The whole lossy process is defined as follows:

$$\begin{aligned}\mathbf{R}_D &= \mathrm{DCT}(\mathbf{R}),\\ \mathbf{R}_Q &= \left[\frac{\mathbf{R}_D}{q_{\mathrm{step}}}\right] \times q_{\mathrm{step}},\\ \mathbf{R}^* &= \mathrm{IDCT}\left(\mathbf{R}_Q\right),\end{aligned} \tag{1}$$

where $\mathbf{R}_D$ and $\mathbf{R}_Q$ refer to the DCT coefficients and quantized DCT coefficients, respectively, q_{step} is the assigned quantization step size, and the quantized form of $\mathbf{R}$ is denoted as $\mathbf{R}^*$. According to the principle of quantization, the quantization noise $\mathbf{n}$ of the current macroblock can be roughly estimated, where $\mathbf{n} = |\mathbf{R}_Q - \mathbf{R}_D| \le 1/2q_{\mathrm{step}}$. The quantization noise of a frame that contains multiple macroblocks is directly proportional to its average quantization step size.

Even under the same quantization step size, interframe correlation also changes the compression noise added to each frame. Suppose that $\mathbf{R} = \mathbf{C}_b - \mathbf{C}_r$ is close to 0 before message embedding. After coefficient modulation, their new residual is defined as $\mathbf{R}' = \mathbf{C}_b' - \mathbf{C}_r'$. If $\mathrm{DCT}(\mathbf{R}') \le 1/2q_{\mathrm{step}}$, $\mathbf{R}_Q = 0$ after quantization. Then, $\mathbf{R}^*$ is equal to 0, which means the current macroblock is the same as its reference macroblock after quantization, and the message bit embedded into the current macroblock is erased. If there exists a strong correlation between the current frame and its reference frame, the probability of the above phenomenon occurring would be high, and the BER of this frame would increase accordingly.

3.3. Two Frame Difference Metrics. Section 3.2 analyzes the source of video compression noise in units of macroblocks. The codec's quantization step size and the correlation between the current macroblock and its reference macroblock are only correlated with macroblock robustness with high probability but cannot directly determine the robustness of the current macroblock. Thus, this section designs two evaluation metrics, called frame quantization step and interframe mutual information, to quantify the robustness of each frame. A frame quantization step is given to measure the frame quantization noise under a given video codec. Interframe mutual information aims to quantify the correlation between two consecutive frames.

3.3.1. Frame Quantization Step. Most robust video data-hiding methods are constructed based on spatial domains. Generated videos are then compressed and recompressed on social networks, where quantization noise is bound to be introduced. As stated in Section 3.2, the level of quantization noise is found to be positively correlated with the quantization step size employed in a given video codec. Thus, we

define the frame quantization step as one frame difference metric to measure the frame quantization noise. The frame quantization step is calculated as

$$Q_{\text{step}}(i) = \frac{1}{n}\sum_{j=1}^{n} q_{\text{step}}(i,j),$$
$$q_{\text{step}}(i,j) = \text{MAP}(qp(i,j)), \tag{2}$$

where $Q_{\text{step}}(i)$ is the quantization step size of the i-th frame, $q_{\text{step}}(i,j)$ and $qp(i,j)$ are the assigned quantization steps and quantization parameters of the j-th macroblock in the i-th frame, respectively, and MAP is the quantization table building a mapping from $qp(i,j)$ to $q_{\text{step}}(i,j)$. The quantization table is specified in the video coding standard, and different video coding standards set different quantization tables.

3.3.2. Interframe Mutual Information. To reduce interframe redundancy, video encoders usually encode the residual between the current macroblock and the reference macroblock instead of the current macroblock itself. From another perspective, the macroblock residual reflects the information increment of the current macroblock. Since the reference macroblocks of macroblocks in the same frame may be distributed in different frames, for simplicity, we just calculate interframe mutual information between two consecutive frames to quantify the information increment. Here, interframe mutual information is defined as the cross entropy between the current frame and its previous frame. It is formulated as

$$\text{MI}_i = \begin{cases} 0, i = 1, \\ I(X_{i-1}; X_i), i = 2, 3, \cdots, n, \end{cases}$$
$$I(X_{i-1}; X_i) = H(X_{i-1}) + H(X_i) - H(X_{i-1}, X_i), \tag{3}$$

where X_i refers to the i-th frame, $H(\circ)$ represents the information entropy of $\circ$, and n is the number of frames in a video. MI measures the mutual dependence between two consecutive frames. The smaller the value of MI_i, the weaker the correlation between the i-th frames and its previous frame. Due to rate-distortion optimization, the frame with a large MI tends to be assigned a large quantization step. Even under the same quantization step, messages embedded in these frames are possibly erased after quantization. Thus, interframe mutual information also effectively measures the robustness difference between frames in a video.

4. Proposed Method

In Section 3, we explore the difference between frames and propose two evaluation metrics to measure the robustness of each frame. Based on the two presented metrics, we introduce a novel approach for robust video steganography in the DWT-SVD domain. First, robustness features are calculated by combining the frame quantization step with the interframe mutual information. Then, we adaptively determine the optimal frames by setting a fixed threshold. Third, we perform preprocessing on available frames to obtain candidate coefficients to be modulated. Finally, secret messages are encoded using RS (Reed-Solomon) codes [30] and then embedded into candidate coefficients based on QIM. The proposed embedding framework is illustrated in Figure 2.

4.1. Heuristic Frame Selection. In this section, we provide a heuristic frame selection method. This method adaptively selects available frames based on the cover video itself, regardless of the used steganographic method. The idea of selecting available frames was first proposed in [8]. Their method is time-consuming and inefficient because the previous message embedding and extraction are necessary to choose available frames. Contrary to their method, our frame selection method is independent of the used steganographic scheme. Section 3.2 gives the frame quantization step and interframe mutual information as metrics to measure video compression noise. The stronger the noise added to a frame, the weaker the robustness of this frame. Thus, we combine the two presented metrics to make a comprehensive evaluation of the frame robustness and design a heuristic strategy to select those available frames. The details of heuristic frame selection are explained as follows:

(i) Perform transport channel matching [31] on the input video to generate the cover video. Then, extract the quantization parameter of each macroblock in the video stream.

(ii) Look up the corresponding quantization table of the video codec. According to the quantization parameter, extract the quantization step of each macroblock. And then calculate the FQSs of all frames in a video, denoted as $\mathbf{Q}_{\text{step}} = (Q_{\text{step}}(1), Q_{\text{step}}(2), \ldots, Q_{\text{step}}(n))$.

(iii) Decode the cover video and generate the YUV component sequence. Calculate the IFMI of the Y component, defined as $\mathbf{MI} = (MI_1, MI_2, \ldots, MI_n)$.

(iv) Both $\mathbf{Q}_{\text{step}}$ and $\mathbf{MI}$ are negatively correlated with the frame robustness. Let us denote the scale factor and threshold as η and ρ. The robustness feature is calculated as $\mathbf{r} = \mathbf{MI} + \eta\mathbf{Q}_{\text{step}}$, which measures the robustness of the current frame.

(v) Distinguish between available and unavailable frames in a video. If there exists a value of i that satisfies $r_i \le \rho, i = 1, 2, \ldots, n$, the i-th frame is determined as an available frame; otherwise, it is determined as an unavailable frame.

After transport channel matching, we can capture the quantization parameter of a given channel. Even for a black-box channel, our heuristic frame selection method can also select the optimal frames for message embedding. Besides, $\mathbf{r}$ is calculated based on the cover video itself without considering the actual message embedding. Thus, our proposed frame selection can be used as a preprocessing method to improve the robustness performance of some existing robust video steganographic methods.

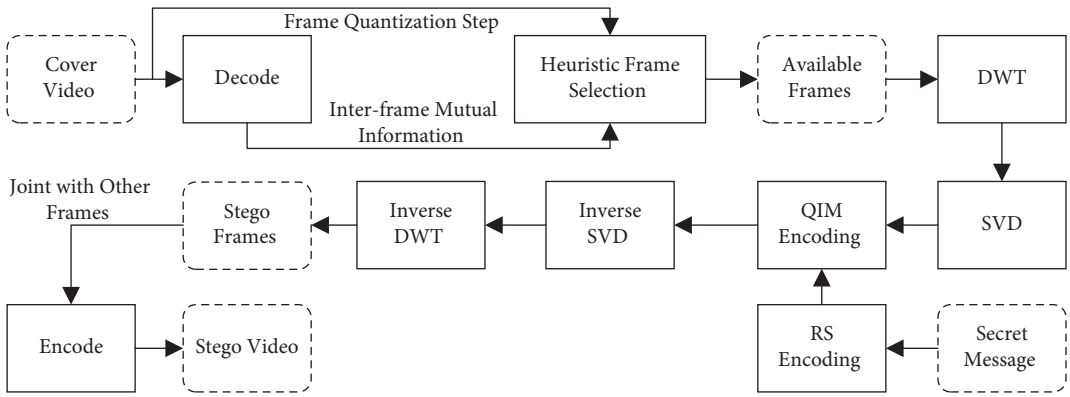

Figure 2: The steganographic framework by exploring differences between frames.

4.2. Preprocessing. A set of cover coefficients can be produced through preprocessing for subsequent message embedding and extraction. In this paper, the cover coefficients belong to the DWT-SVD domain.

To get a large embedding capacity, we select the Y component and divide it into nonoverlapping 16×16 blocks. For each block, DWT and SVD are successively conducted to generate the candidate coefficient to be modified. DWT is performed using high-pass and low-pass filters, which divide each block into four frequency subbands, namely, LL, LH, HL, and HH. The LL subband of the i-th pixel block decomposed by DWT is denoted as $\mathbf{X}_i^L$. SVD is conducted as

$$\mathbf{X}_i^L = \mathbf{U}\Lambda_i^L\mathbf{V}^T, \tag{4}$$

where the matrices $\mathbf{U}$ and $\mathbf{V}$ are both unitary, and $\Lambda_i^L = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r, 0, \ldots, 0)$ denotes the resulting singular value matrix, where r refers to the rank of Λ_i^L. The foremost singular value λ_1 is used as the candidate coefficient of the current block. Suppose that a video contains m available frames, and each frame consists of n pixel blocks. We define the sequence of extracted coefficients as $\mathbf{c} = (\lambda_1^{1,1}, \lambda_1^{1,2}, \ldots, \lambda_1^{1,n}, \lambda_1^{2,1}, \ldots, \lambda_1^{2,n}, \ldots, \lambda_1^{m,n})$.

4.3. Message Embedding. This section presents the message embedding procedure of our method in the DWT-SVD domain, which consists of three parts: preprocessing, RS encoding, and QIM encoding. The following subsections provide a detailed description of each step of the message embedding process.

(i) For a video, make a heuristic frame selection to determine the available frames in a video. Then, the tags $\mathbf{L}$ of all frames are generated for the receiver to distinguish available frames from unavailable frames.

(ii) Make preprocessing on the first k frames to get the coefficient sequence $\mathbf{c}_1$. And make preprocessing on the remaining available frames to generate the coefficient sequence $\mathbf{c}_2$.

(iii) Encode the secret message $\mathbf{m}$ based on RS code to improve the success rate of hidden communication. In addition, to ensure the security of the encoded data, we apply scrambling to obtain a secure version denoted as $\mathbf{m}_e$, which makes it difficult for potential attackers to recover the original message. Following the same process, encode and scramble $\mathbf{L}$ to obtain $\mathbf{L}_e$.

(iv) Conduct QIM encoding and modulate the coefficient sequence $\mathbf{c}_1$ to embed the data $\mathbf{L}_e$ and $\mathbf{c}_2$ to embed the data $\mathbf{m}_e$. Suppose that the quantization step sizes of QIM are set as Δ_1 and Δ_2. The stego coefficient sequences are calculated as $\mathbf{s}_1 = (\lfloor \mathbf{c}_1/\Delta_1 \rfloor + \lfloor \mathbf{c}_1/\Delta_1 + \mathbf{L}_e \rfloor \mod 2) \times \Delta_1$ and $\mathbf{s}_2 = (\lfloor \mathbf{c}_2/\Delta_2 \rfloor + \lfloor \mathbf{c}_2/\Delta_2 + \mathbf{m}_e \rfloor \mod 2) \times \Delta_2$.

(v) Replace $\mathbf{c}_1$ and $\mathbf{c}_2$ with the stego coefficient sequences $\mathbf{s}_1$ and $\mathbf{s}_2$. Perform the inverse SVD on the LL subband $\mathbf{X}^L$ and then conduct inverse DWT on the resulting 16×16 blocks to yield the new pixel blocks. Finally, merging all these blocks produces the modulated Y component.

(vi) All modulated frames are joined with other frames to get the whole Y component sequence. Setting the value of CRF (constant rate factor) to 0, we can encode the YUV component in a lossless way and then produce a new stego video.

It is worth mentioning that the first k frames carrying the tags $\mathbf{L}$ may not all be available frames. As the length of $\mathbf{L}$ is equal to the number of frames in the entire video, it is recommended to use short videos for message embedding and to embed the tags $\mathbf{L}$ in the first frame.

4.4. Message Extraction. Message extraction aims to recover the embedded message from the stego video. The quantization step sizes of QIM, Δ_1 and Δ_2, are taken as the key parameters to be shared with the receiver for blind extraction. The specific steps of message extraction are described as follows:

(i) Decode the stego video to generate the YUV component sequence. Obtain the first kY components to extract the stego coefficient sequence $\mathbf{s}_1$.

(ii) Conduct QIM decoding and extract the hidden data from $\mathbf{s}_1$. Suppose that the quantization step size of

QIM is set as Δ_1. The hidden data is calculated as $\mathbf{L}_e = [\mathbf{s}_1/\Delta_1] \mod 2$. Antiscramble and decode it to get the tags $\mathbf{L}$ of frames.

(iii) Based on $\mathbf{L}$, select the remaining available Y components and then perform preprocessing to generate the stego coefficient sequence $\mathbf{s}_2$. Conduct QIM decoding and extract the hidden data $\mathbf{m}_e = [\mathbf{s}_2/\Delta_2] \mod 2$.

(iv) Antiscramble $\mathbf{m}_e$ and utilize the corresponding RS code to decode it. Then, the final message $\mathbf{m}$ is extracted.

5. Experiments

In this section, extensive experiments are conducted to test the effectiveness of our proposed method. The concrete experimental settings, such as source video, lossy channel, and experimental platform, are described in detail. We perform the robustness experiment, the security experiment, and the computational complexity experiment to evaluate the overall performance of our method. Besides, the practical performance is also verified on social networks, such as YouTube and Vimeo.

5.1. Experimental Settings

5.1.1. Source Video. The dataset consists of 100 public videos downloaded from YouTube, covering diverse domains such as sports, news, advertising, and films, among others, with the resolution of 1080p (1920×1080) and varying durations ranging from 30 seconds to 10 minutes. These videos are stored in the 4 : 2 : 0 chroma sampling format. To create the cover video dataset for the experimental validation, we randomly selected 100 video clips, each comprising 300 frames, from the original videos. We then cropped these clips to generate videos with resolutions of 480p (640×480) and 720p (1080×720) at a frame rate of 30 frames per second using the H264 encoder.

5.1.2. Lossy Channel. To test the algorithm's robustness in lossy channels, we set up six lossy channels, including four local channels and two social networking channels. The coding rate of each local channel is controlled through the parameter CRF or QP (quantization parameter), with a value of 20 or 26. YouTube and Vimeo are social networking channels; their coding parameters are unknown.

5.1.3. Experimental Platform. Our experiments are based on the MATLAB platform. Video encoding and decoding are based on FFMPEG instructions. In the experiments, we utilize the MATLAB instruction "system" to invoke an FFMPEG executable for video compression and recompression. Besides, DWT and SVD are implemented by MATLAB functions. The used CPU is Intel Xeon Bronze 3106 Processor with a base frequency of 1.7 GHz.

5.2. Ablation Experiment. To verify the effectiveness of heuristic frame selection, we perform an ablation test in this section. Since heuristic frame selection is based on FQS and IFMI, we design two other strategies as supplementary: FQS-based frame selection and IFMI-based frame selection, respectively. We also implement a random frame selection scheme not based on any evaluation metric as a baseline. For fair comparison, all of the above methods select the same number of frames and follow the same process to conduct message embedding.

The experimental results are shown in Table 1. The average BERs of these four methods are 2.99%, 2.19%, 1.89%, and 1.63%, respectively. The proposed frame difference metrics, FQS and IFMI, are effective in enhancing the robustness of video steganography. The FQS-based method has stronger robustness than the IFMI-based method. Our heuristic frame selection provides the best overall performance in most local channels, the average BER of which is about half of that of the baseline. Under the QP26 channel, the average BER of the FQS-based method is lower than that of heuristic frame selection. It is because some frames with small IFMI are wrongly selected increasing BER. In general, the combination of FQS and IFMI is more effective than a single metric, and our proposed heuristic frame selection can highly improve the robustness of existing video steganographic methods.

5.3. Comparison with Other Methods. To make a comprehensive assessment of our proposed method, we carry out comprehensive experiments covering four areas: embedding capacity, robustness, security, and computational complexity. Besides, we conducted a supplementary experiment to evaluate the practicability of our method on social networks, such as YouTube and Vimeo. For comparison, both Huan's method [6] and Fan's method [8] are utilized. It should be noted that Huan's method is specifically aimed at robust watermarking based on DTCWT and SVD, while Fan's method is intended for robust steganography using DWT and SVD. For fairness, all cover videos are generated utilizing transport channel matching, and the same video samples are used for these methods to conduct message embedding and extraction.

5.3.1. Embedding Capacity. In this section, we evaluate the embedding capacity of Huan's method, Fan's method, and our proposed method. Their embedding rates are the same and set to 1, where each message bit is embedded into a single candidate coefficient. Since certain techniques employ ECC and add error correction bits, the effective embedding capacity is computed as

$$C_a = \% \left\lfloor \frac{eC_m}{z} \% \right\rfloor, \tag{5}$$

where the metric used to quantify C_a is bpf (bits per frame), C_m refers to the maximum embedding capacity, z refers to the overall frame count in a video, and e is the code rate of ECC. Assuming a method does not use ECC, e is equal to 1.

The experimental results are shown in Figure 3. C_a of Huan's method varies from 15 bpf to 50 bpf, Fan's method varies from 250 bpf to 800 bpf, and our method varies from 150 bpf to 390 bpf. Compared to two other methods, Huan's

Table 1: Average BER of the method using frame selection based on different metrics.

FQS	IFMI	480p (%)				720p (%)			
		crf26	crf32	qp26	qp32	crf26	crf32	qp26	qp32
×	×	3.03	3.25	2.84	4.84	2.38	2.35	2.51	2.55
×	√	2.75	2.45	2.32	2.82	2.29	1.40	2.19	2.31
√	×	2.51	2.04	**1.15**	2.77	1.92	1.41	**0.87**	2.45
√	√	**1.50**	**1.54**	1.63	**2.61**	**1.29**	**1.13**	1.27	**2.05**

The bold values are the minimum value of the column. The smaller the value, the stronger the robustness of the corresponding algorithm under the same compression channel. Therefore, the bold values indicate that the corresponding method provides the strongest robustness under the current compressed channel.

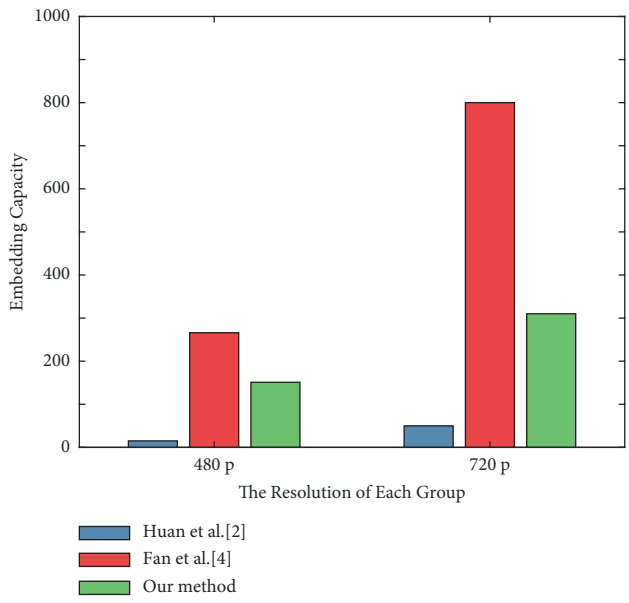

Figure 3: The average embedding capacity of videos under a resolution of 480p or 720p.

method provides the least embedding capacity even without introducing ECC. Thus, robust watermarking is often not suitable for covert communication scenarios due to its low embedding capacity. C_a of our method is lower than Fan's method. It is because heuristic frame selection excludes many nonrobust frames and leads to a decrease in C_a.

5.3.2. Robustness. This section presents a robustness evaluation based on four local lossy channels. For fairness, we fine-tune the length of messages embedded into videos and keep the average embedding capacity of Fan's method the same as our method. Besides, the embedding strength of Huan's method is carefully determined for the best robustness, and the quantization steps of Fan's method and our method are set to be equal and adaptively adjusted according to the given channels. It is worth mentioning that the ECC used is different in these three methods. Huan's method does not utilize any ECC, Fan's method uses BCH (15, 5), and our method introduces RS (127, 63).

In order to accurately assess the level of robustness, we perform a calculation of the BER for the generated message both prior to and subsequent to RS decoding. These are denoted as $R_{e,1}$ and $R_{e,2}$, correspondingly. We utilized the average BER denoted as $\overline{R}_e$ and the success rate denoted as Rs to evaluate the robustness of each group of videos. $\overline{R}_e$ and R_s are calculated as

$$\overline{R}_e = \frac{\left(\sum_{i=1}^{v} R_{e,1}(i)\right)}{v},$$
$$R_s = \frac{\left(\sum_{i=1}^{v}\left(R_{e,2}(i) == 0\right)\right)}{v}, \tag{6}$$

where the BER for the generated message prior to and subsequent to RS decoding from the i-th video is represented as $R_{e,1}(i)$ and $R_{e,2}(i)$, respectively, and v denotes the number of videos in each group. A value of $R_{e,2}(i) = 0$ indicates a successful completion of covert communication between the sender and receiver. As a result, R_s serves as a record of the communication success rate.

Table 2 presents the experimental results, demonstrating that the average BER across Huan's method, Fan's method, and our method is 19.70%, 2.18%, and 1.65%, correspondingly. It could be observed that the BER of Huan's method increases sharply when expanding its embedding capacity and applying it to covert communication scenarios. Compared with Huan's and Fan's methods, our method provides better robustness against video recompression. In addition, our method achieves higher success rates in covert communication than two other methods, ranging from 53% to 92%. Under the qp32 channel, Fan's method has a lower $\overline{R}_e$ and R_s because the effect of message embedding on frame selection cannot be ignored under the qp32 channel. Overall, our method provides superior robustness to construct reliable covert communication on lossy channels.

5.3.3. Feasibility on Social Networks. In this section, we test the feasibility of our method on social networks, such as YouTube and Vimeo. Thus, the recompression channel is changed to YouTube or Vimeo instead of the previous local channels. For simplicity, we randomly selected 30 480p videos and 30 720p videos for validation. The used ECC is changed to RS (127, 31) for better error correction performance. Other experimental settings are the same as those of the robustness experiment.

Table 3 presents the corresponding experimental results. It can be observed that the average $\overline{R}_e$ of Huan's method is 20.87%, the average $\overline{R}_e$ of Fan's method is 11.12%, and the average $\overline{R}_e$ of our method is 6.18%. The average $\overline{R}_e$ of our method is about 5% lower than that of Fan's method. Since

Table 2: Average BER $\overline{R}_e$ and success rate R_s under the local channel of crf26, crf32, qp26, or qp32.

Resolution	Algorithm	crf26 (%)		crf32 (%)		qp26 (%)		qp32 (%)	
		$\overline{R}_e$	R_s	$\overline{R}_e$	R_s	$\overline{R}_e$	R_s	$\overline{R}_e$	R_s
480p	Huan et al. [6]	15.34	—	20.17	—	12.98	—	18.33	—
	Fan et al. [8]	2.31	63	2.16	69	2.24	68	3.87	16
	Our method	**1.41**	**88**	**1.39**	**88**	**1.63**	**78**	**2.94**	**53**
720p	Huan et al. [6]	21.43	—	24.87	—	20.93	—	23.51	—
	Fan et al. [8]	1.64	87	1.47	90	2.03	70	**1.69**	**86**
	Our method	**1.20**	**89**	**1.09**	**92**	**1.40**	**83**	2.13	68

The bold values are the minimum R_e or maximum R_s of the column at a certain resolution. The smaller the R_e, the stronger the robustness of the corresponding algorithm, and the larger the R_s, the stronger the robustness of the corresponding algorithm. The bold values indicate that the corresponding algorithm has the strongest robustness at the current resolution.

Table 3: Average BER $\overline{R}_e$ and success rate R_s on social networks.

Resolution	Algorithm	YouTube (%)		Vimeo (%)	
		$\overline{R}_e$	R_s	$\overline{R}_e$	R_s
480p	Huan et al. [6]	20.06	—	17.32	—
	Fan et al. [8]	14.02	0	9.55	13
	Our method	**7.75**	**10**	**6.35**	**30**
720p	Huan et al. [6]	24.41	—	21.67	—
	Fan et al. [8]	8.21	20	12.68	10
	Our method	**3.82**	**70**	**6.78**	**27**

The bold values are the minimum R_e or maximum R_s of the column at a certain resolution. The smaller the R_e, the stronger the robustness of the corresponding algorithm, and the larger the R_s, the stronger the robustness of the corresponding algorithm. The bold values indicate that the corresponding algorithm has the strongest robustness at the current resolution.

the transcoding mechanisms are unknown, the average BERs of these three methods increase on social networks compared with local channels. Although $\overline{R}_e$ increases on social networks, our method still outperforms Huan's and Fan's methods. On YouTube, R_s of our method is up to 70%. Under the Vimeo channel, R_s of our method can reach 30%. Our method demonstrates better robustness than the two other methods even on social networks. Our method is more practical for reliable covert communication on social networks, such as YouTube and Vimeo.

5.3.4. Security. This section aims to assess the security performance of our method against video steganalysis, in comparison with Fan's method. Huan's method is a watermarking method and is not directly relevant to this security evaluation. Besides, we conduct message embedding without frame selection in the DWT-SVD domain, which is used as the baseline for the security evaluation. Steganalysis based on SPAM (subtractive pixel adjacency matrix) features [32] is realized. Meanwhile, a steganalysis algorithm to detect DCT-based data-hiding methods for H.264/AVC videos [33] is realized, and VDCTR (Video DCT Residuals) features are extracted. The ensemble classifier [34] is used as the classifier to train SPAM features and VDCTR features, respectively. We generate cover and stego samples at embedding rates of 0, 0.1, 0.2, and 0.3. Half of the cover-stego pairs are randomly selected for training purposes, whereas the other half is reserved for the test set.

To evaluate the security of the steganalysis classifier, we consider the OOB (out of bag) error denoted as E_{OOB} as the evaluation criterion. It is worth noting that E_{OOB} serves as an unbiased estimate of the minimum overall detection error rate, denoted as P_E. The latter can be defined as

$$P_E = \min_{P_{\mathrm{FA}}} \frac{1}{2}\left(P_{\mathrm{FA}} + P_{\mathrm{MD}}\right), \tag{7}$$

where P_{FA} and P_{MD} refer to the false alarm rate and the missed detection rate, respectively. A larger value for E_{OOB} indicates better security performance against steganalysis.

The experimental results are given in Figure 4. Under SPAM-based steganalysis, the average detection error rates of the baseline, Fan's method, and our method are 0.3740, 0.4729, and 0.4740, respectively. Under VDCTR-based steganalysis, their average P_E is 0.4481, 0.4912, and 0.4807, respectively. The average P_E of the baseline is lower than Fan's and our methods. The main reason is that a stego video contains both frames modified and frames not modified due to frame selection. In the training phase, the used classifier is confused by the steganalysis features that are extracted from unmodulated frames in a stego video. Moreover, our and Fan's methods are secure enough to resist SPAM-based and VDCTR-based steganalysis, and the steganalysis performance of SPAM is even stronger than that of VDCTR. It is because the stego videos undergo video recompression in lossy channels, and video recompression noise erases the modulation noise to a certain extent. Thus, in a real covert communication scenario, our method can provide a satisfactory level of security performance thanks to video recompression and frame selection.

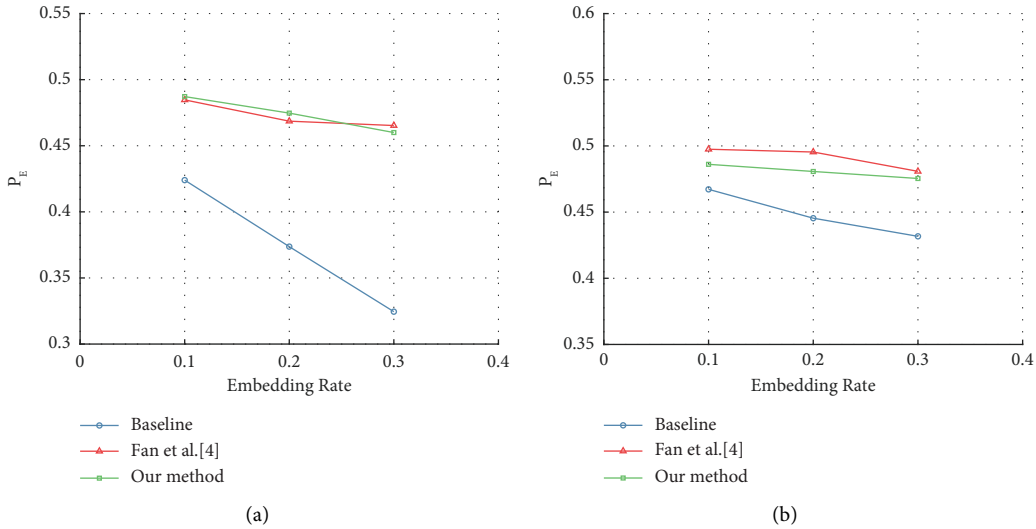

Figure 4: P_E of video steganalysis based on (a) SPAM or (b) VDCTR.

Table 4: Computational time of each frame in each phase under a resolution of 480p or 720p.

Method	480p (spf)				720p (spf)			
	t_{sf}	t_{em}	t_{ex}	t_{tol}	t_{sf}	t_{em}	t_{ex}	t_{tol}
Fan et al. [8]	0.9980	**0.6999**	0.2257	1.9236	2.2374	**1.2673**	**0.4540**	3.9587
Our method	**0.0288**	0.7016	**0.2250**	**0.9554**	**0.0867**	1.2738	0.4599	**1.8204**

The bold values are the minimum value of the column at a certain resolution. The smaller the value, the shorter the computational time. The bold values indicate that the corresponding algorithm has lower computational complexity at the current resolution.

5.3.5. Computational Complexity. This section is designed to test the computational complexity. For fair comparison, each frame is loaded with the same number of messages. Huan's method is not considered due to its low embedding capacity, and we just calculate the spent time of our proposed method and Fan's method.

We measure the computational complexity by calculating the total time of frame selection, embedding, and extraction. Since the number of frames differs in different videos, the time spent per frame is used as an evaluation metric. It is defined as

$$t_{tol} = t_{sf} + t_{em} + t_{ex},$$
$$t_{sf}, t_{em}, t_{ex} = \frac{T}{z}, \tag{8}$$

where t_{tol}, t_{sf}, t_{em}, and t_{ex} are measured by spf (seconds per frame) and refer to the total time consumed per frame and the time consumed per frame in each stage, respectively, T refers to the execution time in each phase, and z is the number of frames in a video.

Table 4 shows the computational time at each stage. The average time of our method is 0.9877 spf in the message embedding phase and 0.3425 in the message extraction phase, both of which are close to the computational time of Fan's method. However, the average time of our method is 0.0578 spf in the frame selection phase, which is about 1/28 of that of Fan's method. The total time spent per frame is about half of Fan's method. Therefore, our method is twice as efficient as Fan's method for covert communication due to the low computational complexity of heuristic frame selection.

6. Conclusion and Future Work

In this paper, we explore the robustness difference between frames in the process of video compression or recompression. Two evaluation metrics are given to measure the difference between frames, called the frame quantization step and interframe mutual information. A new robust video steganographic method is proposed to resist video recompression based on the two metrics. Experimental results demonstrate that our proposed method greatly reduces computational complexity and improves robustness in lossy channels. Besides, our method is still effective in resisting video recompression on social networks, such as YouTube and Vimeo.

There are some limitations in our current work. Though our approach shows good robustness against video recompression, it falls short in the face of geometric attacks. Furthermore, the success rate of covert communication on social networks still has room for improvement. In the future, we will explore other sources of video recompression noise. The improvement of the modulation algorithm is also a direction worthy of further research.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the NSFC under 61972390 and 62272456 and the National Key Technology Research and Development Program under 2022QY0101.

References

[1] Z. Zhou, Y. Li, J. Li et al., "GAN-siamese network for cross-domain vehicle re-identification in intelligent transport systems," *IEEE Transactions on Network Science and Engineering*, vol. 15, pp. 1–12, 2022.

[2] Z. Zhou, C. Ding, J. Li et al., "Sequential order-aware coding-based robust subspace clustering for human action recognition in untrimmed videos," *IEEE Transactions on Image Processing*, vol. 32, pp. 13–28, 2023.

[3] P. K. Ostrowski, E. Katsaros, D. Wesierski, and A. Jezierska, "BP-EVD: forward block-output propagation for efficient video denoising," *IEEE Transactions on Image Processing*, vol. 31, pp. 3809–3824, 2022.

[4] Z. Hu, D. Xu, G. Lu, W. Jiang, W. Wang, and S. Liu, "FVC: an end-to-end framework towards deep video compression in feature space," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 4, pp. 4569–4585, 2023.

[5] P. Fan, H. Zhang, and X. Zhao, "Adaptive QIM with minimum embedding cost for robust video steganography on social networks," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3801–3815, 2022.

[6] W. Huan, L. Sheng, Q. Zhenxing, and Z. Xinpeng, "Exploring stable coefficients on joint sub-bands for robust video watermarking in DT CWT domain," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 4, pp. 1955–1965, 2021.

[7] M. M. Sadek, A. S. Khalifa, and M. G. Mostafa, "Robust video steganography algorithm using adaptive skin-tone detection," *Multimedia Tools and Applications*, vol. 76, no. 2, pp. 3065–3085, 2017.

[8] P. Fan, Z. Hong, C. Yifan, X. Pei, and Z. Xianfeng, "A robust video steganographic method against social networking transcoding based on steganographic side channel," in *ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec)*, pp. 127–137, ACM, New York, NY, USA, 2020.

[9] Y. Cao, Z. Zhou, C. Chakraborty et al., "Generative steganography based on long readable text generation," *IEEE Transactions on Computational Social Systems*, vol. 16, pp. 1–11, 2022.

[10] Z. Guan, J. Jing, X. Deng et al., "DeepMIH: deep invertible network for multiple image hiding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 372–390, 2023.

[11] S. K. Das, S. K. Bhutia, T. P. Kegelman et al., "MDA-9/syntenin: a positive gatekeeper of melanoma metastasis," *Frontiers in Bioscience*, vol. 17, no. 1, pp. 1–15, 2012.

[12] B. Ines, J. S. Ben, and Z. Ezzeddine, "Online multi-sprites based video watermarking robust to collusion and transcoding attacks for emerging applications," *Multimedia Tools and Applications*, vol. 77, no. 11, pp. 14–379, 2018.

[13] O. Cetin and A. T. Ozcerit, "A new steganography algorithm based on color histograms for data embedding into raw video streams," *Computers & Security*, vol. 28, no. 7, pp. 670–682, 2009.

[14] H. M. Kelash et al, O. F. A. Wahab, O. A. Elshakankiry, and H. S. El-sayed, "Utilization of steganographic techniques in video sequences," *International Journal of Computing & Network Technology*, vol. 02, no. 1, pp. 17–24, 2014.

[15] J. Zhang, A. T. S. Ho, G. Qiu, and P. Marziliano, "Robust video watermarking of H.264/AVC," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 54, no. 2, pp. 205–209, 2007.

[16] C. Xu and X. Ping, "A steganographic algorithm in uncompressed video sequence based on difference between adjacent frames," in *International Conference on Image and Graphics (ICIG)*, pp. 97–302, IEEE, Chengdu, China, 2007.

[17] R. J. Mstafa and K. M. Elleithy, "A high payload video steganography algorithm in DWT domain based on BCH codes (15, 11)," in *Wireless Telecommunications Symposium (WTS)*, pp. 1–8, IEEE, New York, NY, USA, 2015.

[18] D. Wang, S. Liu, X. Luo, and S. Li, "A transcoding-resistant video watermarking algorithm based on corners and singular value decomposition," *Telecommunication Systems*, vol. 54, no. 3, pp. 359–371, 2013.

[19] G. Shubham, G. Mohit, A. Pranshu, and S. Ranabir, "Combined DWT–DCT-based video watermarking algorithm using arnold transform technique," in *International Conference on Data Engineering and Communication Technology*, pp. 455–463, Springer, Berlin, Germany, 2017.

[20] S. Ponni alias Sathya and S. Ramakrishnan, "Fibonacci based key frame selection and scrambling for video watermarking in DWT-SVD domain," *Wireless Personal Communications*, vol. 102, no. 2, pp. 2011–2031, 2018.

[21] A. M. Jahangir, "A blind and robust video watermarking scheme in the DT CWT and SVD domain," in *Picture Coding Symposium (PCS)*, pp. 277–281, IEEE, Cairns, QLD, Australia, 2015.

[22] I. J. Cox, J. Kilian, F. T. Leighton, and T. Shamoon, "Secure spread spectrum watermarking for multimedia," *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1673–1687, 1997.

[23] H. Ito and T. Kasezawa, "Permutation-based signature generation for spread-spectrum video watermarking," *IEICE Transactions on Info and Systems*, vol. 102, no. 1, pp. 31–40, 2019.

[24] B. Chen and G. W. Wornell, "Quantization index modulation: a class of provably good methods for digital watermarking and information embedding," *IEEE Transactions on Information Theory*, vol. 47, no. 4, pp. 1423–1443, 2001.

[25] H.-Y. Huang, C.-H. Yang, and W.-H. Hsu, "A video watermarking technique based on pseudo-3-D DCT and quantization index modulation," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 4, pp. 625–637, 2010.

[26] N. I. Yassin, N. M. Salem, and M. I. El Adawy, "QIM blind video watermarking scheme based on wavelet transform and principal component analysis," *Alexandria Engineering Journal*, vol. 53, no. 4, pp. 833–842, 2014.

[27] Y. Zhang, X. Luo, C. Yang, D. Ye, and F. Liu, "A framework of adaptive steganography resisting JPEG compression and detection," *Security and Communication Networks*, vol. 9, no. 15, pp. 2957–2971, 2016.

[28] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[29] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[30] I. S. Reed and G. Solomon, "Polynomial codes over certain finite fields," *Journal of the Society for Industrial and Applied Mathematics*, vol. 8, no. 2, pp. 300–304, 1960.

[31] Z. Zhao, Q. Guan, H. Zhang, and X. Zhao, "Improving the robustness of adaptive steganographic algorithms based on transport channel matching," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 7, pp. 1843–1856, 2019.

[32] T. Pevny, P. Bas, and J. Fridrich, "Steganalysis by subtractive pixel adjacency matrix," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 2, pp. 215–224, 2010.

[33] P. Wang, Y. Cao, X. Zhao, and M. Zhu, "A steganalytic algorithm to detect DCT-based data hiding methods for H.264/AVC videos," in *ACM Workshop on Information Hiding and Multimedia Security (IH&MMSec)*, pp. 123–133, ACM, New York, NY, USA, 2017.

[34] J. Kodovsky, J. Fridrich, and V. Holub, "Ensemble classifiers for steganalysis of digital media," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 2, pp. 432–444, 2012.