TERA: Screen-to-Camera Image Code With Transparency, Efficiency, Robustness and Adaptability

Han Fang, Dongdong Chen, Feng Wang, Zehua Ma, Honggu Liu, Wenbo Zhou, Weiming Zhang, and Nenghai Yu

***Abstract*—With the rapid development of digital devices, the issue of how to transmit information among different devices with multimedia carriers has drawn much attention from the research community. This paper focuses on the important user scenario of "screen-to-camera information transmission". Along this direction, image coding-based techniques have been shown to be the most popular and effective methods in the past decades. However, after careful study, we find that none of the existing methods can satisfy the four important properties simultaneously, i.e., *high transparency*, *high embedding efficiency*, *strong transmission robustness* and *high adaptability to device types*. This is mainly because these properties are contradictory with each other. In this paper, we thus propose a screen-to-camera image code dubbed "TERA" (transparency, efficiency, robustness and adaptability), which makes it possible to circumvent the contradiction among the above four properties for the first time. Generally, TERA adopts the color decomposition principle to ensure the visual quality and the superposition-based scheme to ensure embedding efficiency. BCH-coding-based information arrangement and a powerful attention-guided information decoding network are further designed to guarantee the robustness and adaptability. Through extensive experiments, the superiority and broad applications of our method are demonstrated.**

***Index Terms*—Adaptability, attention-guided, color decomposition, efficiency, robustness, screen-to-camera image code, transparency.**

I. Introduction

In RECENT decades, digital devices such as personal computers, mobile phones and AR/VR devices have rapidly developed and become more and more common. With these devices, information computation and transmission become increasingly more efficient and convenient. Due to the existence of physical gaps and some real application requirements, information transmission among different devices is also very necessary and has drawn much attention from the research community. Typical examples include screen-shooting resilient watermarking for IP protection [1]–[3] and screen-to-camera communication [4], [5]. In this paper, we focus on the specific user application scenario "screen-to-camera information transmission," which refers to the information transmission channel between screen and camera. The screen is the sender and the camera is the receiver. The information displayed on the screen can be received by camera capture and post-processing operations. Therefore, hardware isolated information transmission from screen to camera is realized. This is a classical and challenging research problem, and many different types of methods [4], [6]–[10] have been proposed in the past decades. Among them, image coding-based techniques [11]–[17] have been shown to be the most popular and effective. They often represent target information with some well-designed patterns and embed these patterns into the host frame image, which is further scanned by the end user to decode the hidden information.

As is common understanding, a perfect screen-to-camera image code should satisfy four properties: great transparency, high embedding efficiency, transmission robustness and high adaptability to device types. Transparency means ensuring that the encoding process retains the original visual quality of the host image as much as possible so that human observers cannot even notice it. Embedding efficiency aims to reduce the computation burden on the screen device side because of the limited computation ability of these devices. Different from the first two requirements which are often imposed on the screen side, the last two ensure that the information embedded in the camera-captured image can be correctly extracted out on the decoder side no matter which types of camera or screen are used.

However, after careful study, we find that none of the existing methods [18]–[22] can satisfy the above four properties simultaneously. This is because of the inherent contradiction among these four properties. In more detail, if we want to ensure the transparency of the hidden message, the embedding strength should be as weak as possible, which will inevitably result in the decrease of robustness and adaptability. Meanwhile, if we

Manuscript received October 12, 2020; revised January 19, 2021; accepted February 18, 2021. Date of publication February 24, 2021; date of current version February 8, 2022. This work was supported in part by the Natural Science Foundation of China under Grants 62072421, and 62002334, in part by the Anhui Science Foundation of China under Grant 2008085QF296, and in part by the Exploration Fund Project of University of Science and Technology of China under Grant YD3480002001. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. P. K. Atrey *(Corresponding authors: Weiming Zhang; Nenghai Yu.)*

Han Fang, Feng Wang, Zehua Ma, Honggu Liu, Wenbo Zhou, Weiming Zhang, and Nenghai Yu are with the CAS Key Laboratory of Electromagnetic Space Information, University of Science and Technology of China, Hefei 230026, China (e-mail: fanghan@mail.ustc.edu.cn; nishi@mail.ustc.edu.cn; mzh045@mail.ustc.edu.cn; lhg9754@mail.ustc.edu.cn; welbeckz@mail.ustc.edu.cn; zhangwm@ustc.edu.cn; ynh@ustc.edu.cn).

Dongdong Chen is with the Microsoft Research, Redmond, Washington 98052 USA (e-mail: cddlyf@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMM.2021.3061801.

Digital Object Identifier 10.1109/TMM.2021.3061801

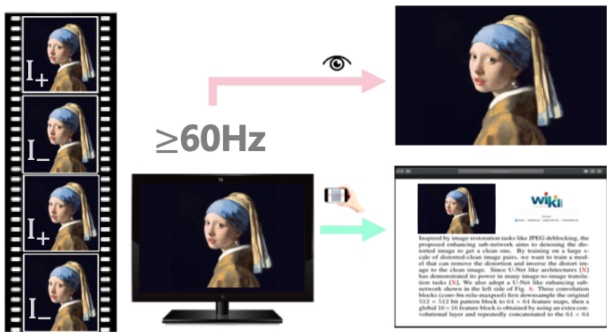


Fig. 1. The schematic diagram of the proposed "TERA" code. When alternately displaying two embedded frames on the screen at no less than 60 *Hz*, the human vision system cannot observe any visual difference, but the embedded message (e.g., website) can be extracted out by camera.

want to improve the robustness, besides enhancing the embedding strength, it is also necessary to consider the texture or the content of the image itself for designing better embedding strategy. However, such an operation will incur higher calculation cost. Essentially, satisfying these four properties at the same time is challenging.

To address this limitation, we propose a novel screen-to-camera image code scheme dubbed as "TERA" (**t**ransparency, **e**fficiency, **r**obustness and **a**daptability). To the best of our knowledge, it is the first image coding-based method that can circumvent the above contradiction problem and meet all the aforementioned requirements.

Generally, to meet the requirements of transparency, we analyze the features of the human vision system (HVS) and utilize the observation [4], [5] that the HVS will fuse two images into one if they are refreshed at a high frequency (no less than 60 *Hz*), and this refresh rate is satisfied by modern screen devices. As a result, we design a new color decomposition-based encoding scheme that encodes the information into a single host frame by creating two complementary frames. Thus, by alternately displaying two complementary frames, what can be seen by human eyes is the composition of these two frames, that is, the original image, to achieve high transparency. However, unlike the HVS, the shutter speed of modern cameras is much higher and will instead capture the decomposed frame that contains information. In this way, the visual quality observed by human beings is theoretically guaranteed and transmitting information to camera devices remains possible, as shown in Fig. 1.

For the efficiency, we designed the superposition-based scheme to significantly reduce the computation burden on the embedding side, based on which the message embedding process is carried out within a short time and in a content-independent manner.

The robustness and adaptability are satisfied with the designed attention-guided extraction network. Although the camera can effectively record the message information hidden in the image, the potential information loss (*e.g.,* light distortion, Moiré distortion and the masking effect of shutter) in the screen-to-camera process will cause enough trouble in extracting the message. To address such problems, we dedicatedly design the BCH-coding-based information arrangement scheme and leverage a new powerful attention-guided extracting network. By setting enough training datasets and designing suitable network architecture, accurate extraction which reflects robustness and adaptability can be guaranteed.

TABLE I
THE COMPARISON OF DIFFERENT ALGORITHMS IN FOUR RESPECTS: TRANSPARENCY, EMBEDDING EFFICIENCY, ROBUSTNESS AND ADAPTABILITY. COMPARED WITH THE OTHER THREE TYPES OF SCHEMES, THE PROPOSED METHOD CAN SATISFY ALL THE REQUIREMENTS

Algorithms	2D Code	Screen-camera Communications	Screen-shooting Resilient Watermarking	Proposed
Transparency	×	✓	×	✓
Efficiency	✓	✓	×	✓
Robustness	✓	×	✓	✓
Adaptability	✓	×	✓	✓

To summarize, the main contributions of this paper are threefold:

- We propose a novel image coding scheme "TERA" for screen-to-camera information transmission, based on which a complete system is further constructed. In addition, such code can simultaneously satisfy the four key properties of the image coding-based scheme.
- We design a high-efficiency superposition-based embedding scheme by BCH-coding-based arrangement and a new powerful attention-guided extraction network for excellent extraction robustness and adaptability.
- Extensive experiments have been conducted with different capturing settings, such as different distances, degrees, and camera types, which demonstrated the superior performance over existing state-of-the-art methods. Several potential applications are also attempted, which demonstrate the potential commercial value of this system.

II. RELATED WORK

For screen-to-camera information transmission, classic methods include traditional communication techniques such as cable and wireless transmission, as well as image coding-based schemes. In some specific scenarios, the former way is often more stable and reliable when using some very strict communication agreements or rules. However, this type of method is often not that flexible. With the development of smart mobile phones, image coding-based schemes have become increasingly popular in recent years. Though the underlying working principles of existing methods are very similar, they can still be roughly categorized into three different types based on their different goals. The detailed advantages and disadvantages are summarized in Table I.

Traditional Image Code: The first one is traditional 2D codes such as barcode or QR code [11]–[14], which encode '0/1' bits into specific patterns. Since the main goal of such methods is to achieve stronger robustness, their visual quality is relatively low. There also exist some works [13], [15]–[17], [23] that focus on beautifying 2D code by taking the image as the background. Liu *et al.* [23] propose a Watson's DCT-based, perceptual model-based, perceptual shaping algorithm to encode the message. By modulating the information into patterns with different angles, the encoding process is realized. Chen *et al.*

propose three other aesthetic 2D barcodes: PiCode [15], RA Code [16], and RU Code [17]. In PiCode [15], they express 0 and 1 by using two templates: inner-dark/outer-bright and inner-bright/outer-dark. On the extracting side, they utilize a 2D matched filter to demodulate the message. In RA Code [16], based on the analysis of the frequency spectrum, they design another template to express information, which can greatly guarantee the robustness of decoding. In RU Code [16], they list a series of guidelines to guide the modulation, embedding and extraction. There is a common problem with these algorithms: they cannot effectively balance the robustness and invisibility, so some obvious visual distortion can still be observed.

Screen-to-camera Communication: To achieve higher information transmission capacity and real-time communication, screen-to-camera communication schemes [4], [7], [8], [24], [25] are another important type of method. Since they are also based on the properties of HVS, their visual quality is often great. However, they highly depend on the strict collaboration between screen and camera. More strictly, high-end screen and camera devices are often needed because of the frequency requirement: otherwise, flickering artifacts will appear.

The main differences of the proposed scheme compared with screen-to-camera communication algorithms are: 1) We care more about the transmission robustness in different shooting conditions. Since the sender and the receiver always cooperatively work in pairs in screen-to-camera communication, the only transmission distortion which occurred resulted from the fixed channel of the sender and the receiver. However, for image code, the decoder equipment may not be fixed and a decoding process may occur in different shooting conditions. As a result, ensuring the transmission accuracy with various shooting conditions is more important. 2) We lift the restrictions on the equipment. To satisfy the extraction accuracy requirement, high intensity modification is needed in the traditional screen-to-camera communication schemes. Meanwhile, to cover the visual distortion, high screen refresh rate and the corresponding receiver are required. Traditional screen-to-camera communication schemes thus always rely on special equipment. However, the most commonly used screens today can support 60 *Hz* display, which is enough to realize non-visual-distortion under low embedding intensity. Thus, as long as we can ensure the extraction accuracy with low embedding intensity, such restrictions will be lifted. We have developed a powerful extraction network with adversarial training in multi-conditions, based on which the proposed scheme can achieve better extraction performance.

Screen-shooting Resilient Watermarking: The last typical type of method is screen-shooting resilient watermarking. It does not require high information capability but is more affected by quality and robustness. By analyzing the features of the image itself, information is often hidden into the texture or color components of an image with handcraft algorithms [19]–[22] or deep learning networks [?], [18], [26], [27]. Other works [19]–[21] propose to use a set of templates to represent '0/1' bits and embed them with an HVS mask in the host image to represent the message. On the extracting side, they use a fixed filter to pre-process and then extract the message by template matching. Zhu *et al.* [18] propose an auto-encoder as in data hiding networks. By joint training the encoder, decoder and the noise layer, resistance to image processing attacks (e.g., JPEG compression, cropping, filtering) can be achieved. Based on the architecture proposed by [18] on Tancik *et al.*, [27] propose a method to simulate the distortions of the camera-shooting process to achieve screen-shooting resistance. Since such methods should greatly balance the robustness and transparency, their embedding efficiency is not substantial. In addition, due to their implementation principle, transparency and robustness are still contradictory with each other to some extent. Therefore, to ensure good robustness, their visual quality is also insufficient. Compared with all the above methods, our paper is the first that can satisfy all of the four properties.

III. Method

The framework of the whole system is shown in Fig. 2. First, we encode the message sequence with BCH [28] and CRC [29], and then apply the Latin square designing (LSD) arrangement rules on the encoded message to generate the message matrix to be embedded. According to the message matrix, two complementary message templates are generated and further superimposed onto the host image to realize the embedding process. After that, by alternatively displaying the two embedded images at no less than 60 *Hz*, the complete invisibility of visual distortion can be realized. The image on the screen is then captured by cameras to conduct the extracting process. On the extraction side, we first perform perspective correction on the captured image and then feed the corrected image into an attention-guided extraction network to recover the message matrix. After that, BCH decoding and CRC error detection will be carried out on the extracted message matrix. If no CRC error is detected, the final message sequence will be extracted. Otherwise, we will recombine the sequence according to the arrangement rule and apply the same decoding process on it. The whole extraction process will be finished when no error is detected or all combinations are tried.

A. Message Matrix Generation

In the screen-to-camera process, there may be various distortions such as Moiré, light and shutter distortion which will influence the image content from different aspects. For example, Moiré and light distortion will incur information loss in one local continuous area, so embedding the complete information unit multiple times is necessary. For the shutter distortion resulting from the mismatch between the display frequency and the camera shutter speed, it will cause the image captured by the camera to be the fusion of two consecutive frames, and the message feature of some rows or columns may disappear. Therefore, we need to ensure that there is complete watermark information in the remaining columns or rows. To achieve that, we should repeatedly embed the whole message matrix into one host image many times so that even if a small region is distorted, the message can still be extracted in the remaining region.

Based on the above analysis and to meet the requirements of robustness, we design a robust message generation scheme as shown in Fig. 3. Specifically, given the original information sequence, we first generate a sequence m of length l with the error detection and correction ability by using CRC and BCH coding.

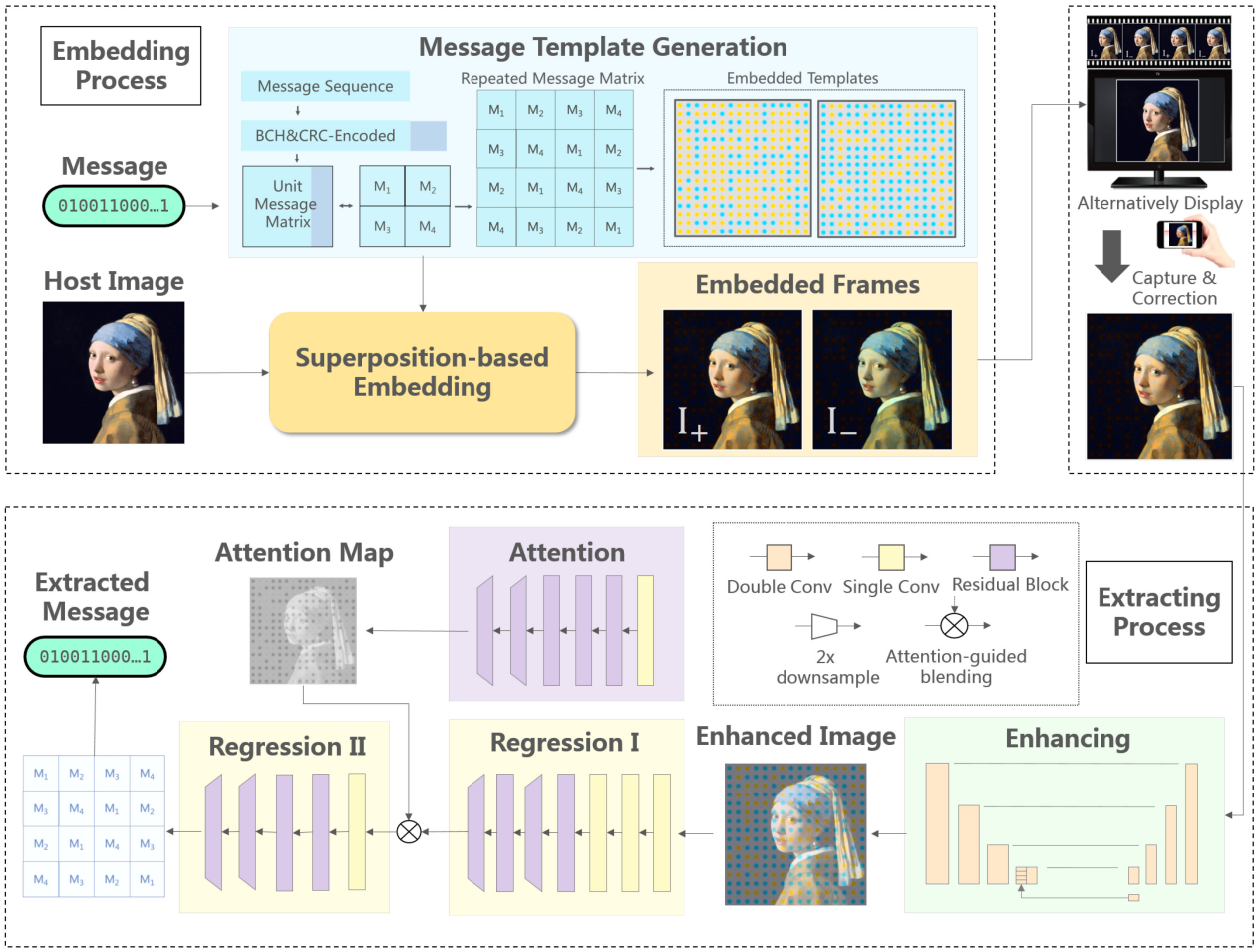


Fig. 2. The framework of the whole process, which consists of two main parts: the embedding process and the extracting process. The left part indicates the message arrangement as well as the message embedding part. After embedding the message, the embedded frames are alternatively displayed at no less than 60 Hz on the screen. Then, after capturing the image and performing perspective correction, the corrected image is fed into the extraction network in the right part, which consists of an enhancing sub-network, an attention sub-network, and a regression sub-network to realize accurate extraction.

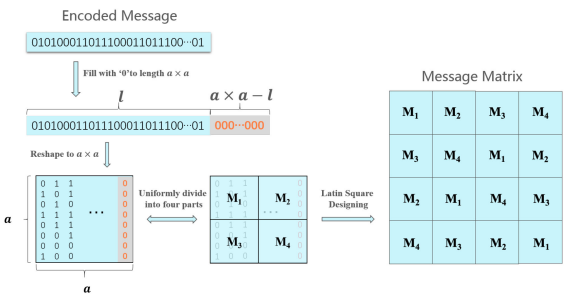


Fig. 3. The specific Latin square designing (LSD) arrangement of the message matrix. The message is first encoded with BCH & CRC coding, then the encoded message is subjected to zero padding and reshaping into size $a \times a$. One complete message matrix is uniformly divided into four parts and further repeated four times according to the LSD arrangement, as shown on the top-right part.

We then resize m into a matrix with size of $a \times a$ (zeroing the part of $a \times a - l$). After that, to repeatedly embed the message, we uniformly divide the matrix into 4 parts, M_1, M_2, M_3 and M_4 and perform the LSD arrangement in Figure 3 to generate the final message matrix to be embedded M.

Fig. 4. The influence of various distortions. The Moiré pattern distortion and light distortion will influence a continuous region of the image where the shutter distortion will cause two-frame fusion and lead to the loss of some line information. However, with the message arrangement, all information can always be combined from the clean region.

The advantage of LSD arrangement is that it successfully disperses the whole watermark unit in the image, so that even after the Moiré distortion, light distortion, and shutter distortion, there is a high probability that at least one completely clean watermark unit can be extracted, as shown in Figure 4. The corresponding advantages can be found in Section IV-D1.

Note that any combination of M_1, M_2, M_3, M_4 will contain one complete message, so ideally M contains 4 complete messages to avoid potential local information loss.

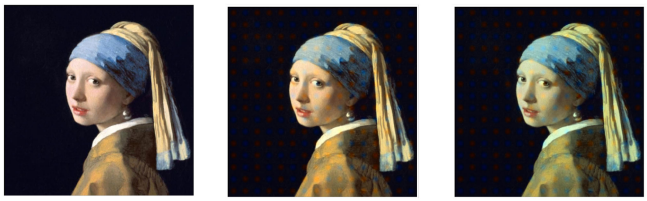


Fig. 5. One example of original image and its corresponding symmetrically embedded images. The left image is the host image, and the middle and right image are the embedded images I'_+ and I'_-.

B. Superposition-Based Message Embedding

The embedding efficiency is the key consideration to enable online message matrix embedding with a high refreshing frequency. In contrast with existing methods [21], [22] that need detailed texture analysis to find the most suitable hiding position, we adopt an extremely efficient superposition-based scheme instead. Specifically, we use an image block of size $b \times b$ to represent a 1 bit message, so the whole template is generated by concatenating all the blocks according to the message matrix. Formally, the image bit block can be generated by (1):

$$P_B(x,y) = \begin{cases} 1 - \frac{D(x,y)}{b/4}, & \text{if } D(x,y) \le b/4 \\ 0, & \text{else} \end{cases} \tag{1}$$

where

$$D(x,y) = \sqrt{(x - b/2)^2 + (y - b/2)^2} \tag{2}$$

(x, y) indicates the pixel coordinates of the image block. Considering that the human vision system is less sensitive to the red&blue components than the green component and the aforementioned color decomposition principle, we hide the information into these two components and create two complementary templates $(+, -)$:

$$P_\pm[r,g,b] = [1 \pm P_B, I_{ori}, 1 \mp P_B] \tag{3}$$

where I_{ori} indicates the G-channel of the original image blocks. The generation rules of the whole message template are illustrated as (4)

$$B_\pm(i,j) = \begin{cases} P_\mp, & \text{if } M(i,j) = 0 \\ P_\pm, & \text{else} \end{cases} \tag{4}$$

where (i, j) indicates the coordinates of information matrix M. After all the message are embedded, we can generate two templates, denoted by B_+ and B_-. The embedded image can thus be generated by (5)

$$I'_\pm = (1 - \alpha) \times I + \alpha \times B_\pm \tag{5}$$

where α indicates the embedding intensity. The two symmetrically embedded images $I'_\pm$ are shown in Figure 5. To realize the transparency, we have to alternately display two symmetric images at no less than 60 *Hz* so that human eyes can only see one still image on the screen, but the camera can effectively capture the embedding artifacts: in this way, transparency to human eyes and recordability by camera can both be achieved.

It is worth noting that alternatively displaying two frames to realize visual distortion has been widely used in the previous fusion-based screen-to-camera communication schemes. Nevertheless, due to the limitation of decoding ability, the traditional screen-to-camera communication schemes require higher embedding intensity, so the algorithm requires a higher refresh rate to compensate for the visual distortion caused by high embedding intensity. However, the proposed deep-learning-based decoder greatly improves the decoding performance, which liberates the embedding intensity limitation and meanwhile reduces the requirements of display frequency.

We also add the locating border (*e.g.,* DataMatrix) around the image for synchronization so that we can correct the perspective distortions according to the border after camera capture.

C. Attention-Guided Extraction Network

To extract the information from the captured image, we first perform perspective correction and then feed the corrected image into the following extraction network. To achieve higher extraction accuracy to meet the demands of robustness and adaptability, we design an attention-guided extraction network. As shown in Figure 2, the whole network architecture consists of four components: (1) the enhancing sub-network E with parameters θ_E, which takes the distorted image $I_d \in \mathcal{R}^{3*H*W}$ as input and generates the enhanced image $I_E \in \mathcal{R}^{3*H*W}$; (2) the attention sub-network A_t with parameter θ_{A_t}, which receives I_E and calculates the attention map $A_{I_E} \in \mathcal{R}^{64*H/4*W/4}$ of I_E; (3) the regression sub-network, which is divided into 2 parts. Regression sub-network-1 R_1 with parameter θ_{R_1} takes I_E as input and generates the feature map $F_1 \in \mathcal{R}^{64*H/4*W/4}$, which has the same size as A_{I_E}, then A_{I_E} and F_1 are multiplied channel by channel to create the attention-based feature map $F_A \in \mathcal{R}^{64*H/4*W/4}$. The regression sub-network-2 R_2 with parameter θ_{R_2} recovers the message $M \in \{0,1\}^{a*a}$; (4) provided with I_E or embedded image $I_{em} \in \mathcal{R}^{3*H*W}$, the adversary A_d with parameter θ_{A_d} evaluates the probability that the enhanced image is the clean embedded image.

1) Enhancing Sub-Network: Inspired by image restoration tasks such as JPEG deblocking, the proposed enhancing sub-network aims to recover the distorted information as much as possible for following extraction. Since U-Net-like [30] architectures have demonstrated power in many image-to-image translation tasks, we adopt the U-Net-like enhancing sub-network shown in Figure 6.

In detail, three convolution blocks (conv-bn-relu-maxpool) first progressively downsample the captured $H \times W$ image $I_d \in \mathcal{R}^{3*H*W}$ to 64×64 feature maps, then a global 16×16 feature block is obtained by using an extra convolutional layer and repeatedly concatenated to the 64×64 feature maps; finally, several convolutional blocks (upsample-conv-bn-relu) upsample the 64×64 feature maps back to the original size to produce the final enhanced image $I_E \in \mathcal{R}^{3*H*W}$ with bit patterns. To train this network in a fully supervised way, we synthesize numerous training samples by regarding the original embedded images $I_{em} \in \mathcal{R}^{3*H*W}$ as ground truth and the captured images as input, and the objective of the enhancing sub-network is to minimize the distance between I_{em} and I_E by updating parameters θ_E:

$$\mathcal{L}_E = MSE(I_{em}, I_E) = MSE(I_{em}, E(\theta_E, I_d)) \tag{6}$$

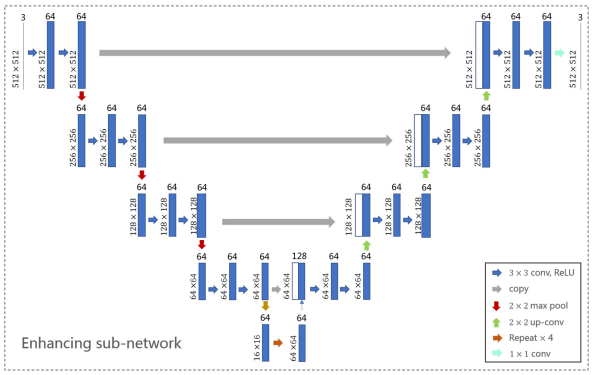


Fig. 6. The detailed information of the enhancing sub-network with input size $512 \times 512 \times 3$.

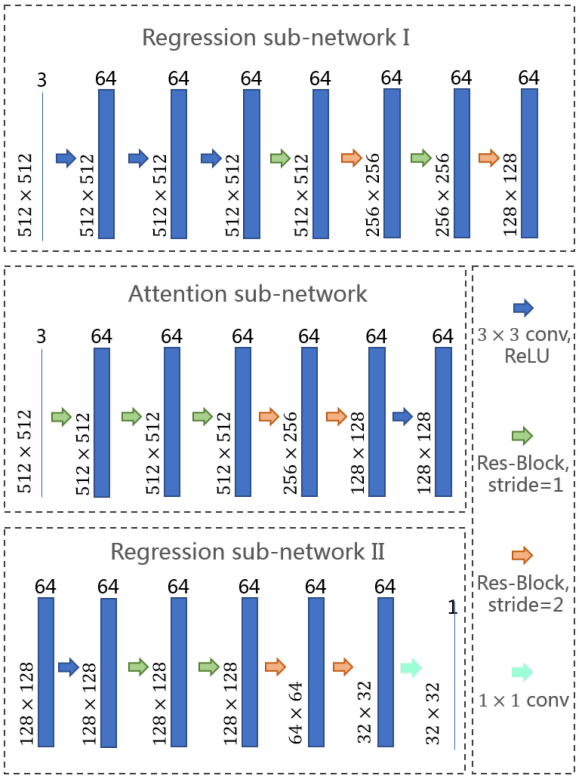


Fig. 7. The detailed information of the attention sub-network and regression sub-network with the input size $512 \times 512 \times 3$.

2) Attention Sub-Network: Since the screen-to-camera process will cause irreversible distortions on the image, the features of some regions still cannot be extracted correctly even after enhancement. Therefore, such regions should be paid less attention. Similarly, the potentially correct regions should be paid much more attention instead. On the other hand, as the '0/1' bits have different patterns, the attention network may give the regression network some visual hints and differentiate them with original texture patterns. To achieve this goal, we design an auxiliary attention sub-network A_t as the guidance, as shown in Figure 7. Given the enhanced image I_E, it will output a soft feature-level guidance map A_{I_E} and multiply it into the intermediate feature F_1 of the following regression sub-network-1. The detailed network structure consists of five residual blocks [31], where the second and fourth blocks dowsample the feature maps by $1/2$, so the size of the final attention map is $1/4$ of the original size ($A \in \mathcal{R}^{H/4*W/4}$). Note that when combining the single-channel attention map with the regression sub-network, we will expand it into the same feature channel number $A_{I_E} \in \mathcal{R}^{64*H/4*W/4}$.

3) Regression Sub-Network: The regression sub-network aims to extract the final message matrix, and we divide it into two parts so that it can effectively collaborate with the attention sub-network: the specific architecture is shown in Figure 7. Given an enhanced image, the sub-network-1 R_1 is responsible for encoding it into high-level intermediate features $F_1 \in \mathcal{R}^{64*H/4*W/4}$, which are further enhanced by the attention $A_{I_E} \in \mathcal{R}^{64*H/4*W/4}$ to generate $F_2 \in \mathcal{R}^{64*H/4*W/4}$, which is fed into the sub-network-2 R_2 to decode the final message matrix $M \in \{0,1\}^{a*a}$. In detail, the R_1 is composed of three convolution blocks (conv-bn-relu-maxpool) and two residual blocks, and the encoded feature size is also $1/4$ of the original size with 64 channels. To achieve stronger extraction ability, R_2 is composed of seven residual blocks, which progressively transform the attention-enhanced features into the message matrix. The objective of regression sub-network training is to minimize the difference between M and the original message matrix $M \in \{0,1\}^{a*a}$ by updating parameters θ_{A_t}, θ_{R_1} and θ_{R_2}:

$$\begin{aligned}\mathcal{L}_R &= MSE(M_o, M) \\ &= MSE(M_o, R_2(\theta_{R_2}, R_1(\theta_{R_1}, I_E), A_t(\theta_{A_t}, I_E)))\end{aligned} \quad (7)$$

4) Adversarial Sub-Network: To better constrain the image quality of the enhanced image, we ultilize the adversarial network. The enhancing sub-network attempts to deceive the adversary, so that the adversarial network cannot judge the correct I_{em} from I_E. To this end, $\mathcal{L}_{A_d}$ loss is used to improve the image quality of I_E by updating θ_{A_d}:

$$\mathcal{L}_{A_d} = log(1 - A_d(I_E)) = log(1 - A_d(E(\theta_E, I_d))) \quad (8)$$

In contrast, A_d should also make a correct binary classifier for I_{em} from I_E. Adversarial training is achieved by minimizing the value function and updating parameters θ_{A_d}:

$$\mathcal{L}_{A_d} = log(1 - A_d(\theta_{A_d}, I_{em})) + log(A_d(\theta_{A_d}, E(I_d))) \quad (9)$$

In this paper, we use PatchGAN [32] as A_d by default.

5) Loss Function: Thanks to the differentiability of these three sub-networks, they can be jointly trained end-to-end with different objectives. Formally, the overall objective function consists of three terms:

$$\mathcal{L} = \lambda_1 \mathcal{L}_E + \lambda_2 \mathcal{L}_R + \lambda_3 \mathcal{L}_{A_d} \quad (10)$$

where $\lambda_1, \lambda_2, \lambda_3$ are the loss weights to balance these three terms and are set as 10,1,0.001, respectively. It can be seen that we have not used any explicit attention guidance for the attention sub-network, but we find that it can be effectively automatically learned jointly.

6) Training Process: For better performance, the whole network is trained in a supervised way. We first obtain the image pairs of distorted image I_d and original embedded image I_{em}.

Then, I_d is directly fed into the enhancing sub-network to create the enhanced image I_E, which is further sent to the attention sub-network and regression sub-network-I. The output of the attention sub-network A_{I_E} and the output of the regression sub-network-I F_1 is blended to generate a high-level intermediate feature F_2 and further fed into regression sub-network-II to produce the final extracted message matrix M. The loss function in Eq. (10) is applied to train the whole network in an end-to-end manner. It can be seen that we have not used any explicit attention guidance for the attention sub-network, but we find that it can be effectively automatically learned jointly.

D. Message Decoding

After obtaining the extracted message matrix, we need to combine one complete message unit for BCH decoding and CRC error detection. Specifically, we try each possible combination of M_1, M_2, M_3 and M_4 according to the arrangement rules and decode them. If no CRC error is detected, we believe that the correct message is extracted. If not, the next combination will be continued. The whole decoding process ends when no CRC error is detected or all combinations are tried.

IV. Experiment and Analysis

A. Implementation Details

For bit sequence encoding, we use BCH(64,36) as the error correction code (ECC), where 5 bit errors can be corrected and the length of CRC bits is 7 bits. The actual message bits are 30 bits, and the message matrix size a is set as 8. The size of the block that represents 1 bit message b is set as 32. To train the extracting network, we randomly choose 1500 images from the COCO dataset [33] and scale them to 512×512 pixels. In this way, each image is embedded with 64 random bits. After displaying the embedded images on the screen and capturing them randomly at 30-60 cm and $-30^\circ - 30^\circ$, we conduct perspective correction and crop the images to generate the captured images with size 512×512. For the following experiments, the default monitor and mobile phone we used are 'AOC-G2770PF' and 'Huawei P30 Pro'. The test dataset is the classical USC-SIPI image dataset [34]. To realize the alternating display, we have written a script that can alternatingly display the specified image at the current refresh rate of the monitor with C++ and Python. It is worth noting that we do not use the video format to achieve the displaying operation because we find that when the image is written into the video, the impact of video compression will produce unnecessary artifacts, which greatly affect the visual quality. Directly alternately displaying two images can avoid the visual distortion.

B. Visual Quality Comparison

To measure the visual quality of the embedded image, we perform a mean opinion score (MOS) test. Specifically, we prepare 16 embedded images for each baseline method and show them on the screen, then ask 30 users to assign a score from 1 (bad quality) to 5 (excellent quality). From Table III, we can easily find that the MOS score of the proposed method is much better than those of other baseline methods. Since the displayed frequency is twice the frequency which can be detected by human eyes, the image human observers can see on the screen is just the same as the original image. In this sense, our method can theoretically ensure the original visual quality of the host image, while other baseline methods will affect it. We further provide some visual results in Figure 8. We can see that the visual quality of images generated with 2D image-code methods is poor, because the purpose of such methods is to generate a strong robust codeword for message transmission, and they have low requirements for visual quality. However, the visual quality of images generated by the screen-shooting resilient watermarking method is higher than that of 2D image-code methods. Thus, for fair comparison in the following robustness test, we choose three algorithms with MOS scores greater than 4.

TABLE II
THE DETAILED CONFIGURATION PARAMETERS WHEN COLLECTING THE SCREEN-SHOOTING DATASET

Process		Screen-shooting
Embedding Parameters	Image source	COCO
	Image Size	512×512 pixels
	Embedding Intensity	0.05
	Number of Images	1500
Camera Shooting Parameters	Device	AOC-G2770PF, Huawei P30 Pro
	Image Presentation	512×512 pixels in resolution of 1920×1080
	Shooting Distance	30-60 cm
	Horizontal Shooting Angle	-30°-30°
	Vertical Shooting Angle	-30°-30°

TABLE III
THE VISUAL QUALITY ASSESSMENT OF DIFFERENT SCHEMES WITH THE MEAN OPINION SCORE (MOS) TEST, WHERE THE HIGHER THE SCORE, THE BETTER THE VISUAL DISTORTION

Algorithms	PiCode [15]	RACode [16]	RUCode [17]	SSRW [22]	Nakamura [20]
MOS	2.27	2.39	2.53	3.06	4.21
Algorithms	Pramila [19]	Gugelmann [21]	Stegastemp [27]	Hidden [18]	Proposed
MOS	4.01	4.25	3.88	2.90	**4.92**

C. Robustness Test of the Proposed Method

1) Screen-Shooting Test in Different Capture Conditions: In real camera capturing scenarios, different shooting settings may be used. Therefore, we evaluate the robustness of our method under various conditions, including different shooting distances and angles. Specifically, the captured distance ranges from [30,70] cm and the shooting angles ranges from $[-40^\circ, 40^\circ]$ horizontally or vertically. For fair comparison, the bit error rate (BER) values shown in the following experiments are the results without ECC correction. Since the ECC used in the proposed scheme can correct errors of 5 bits, when BER is below $5/64 = 7.81\%$, the message can be recovered losslessly.

As we can see in Table IV, compared with the baseline methods [19]–[21], the bit error rate of the proposed method is lower in all different distances. We can thus conclude that under the same level of visual quality, the proposed algorithm performs better. In addition, even at the distance of 70 cm, which does not appear in the training set, it can realize high extraction accuracy. The changing of distance therefore affects the performance of the algorithm only minimally.

Table V shows the bit error rates of different algorithms at different capture angles. It can be seen that, when captured at the horizontal angle, the bit error rates of angles within $[-30^\circ, 30^\circ]$ are all less than 12%. However, when the shooting angle is

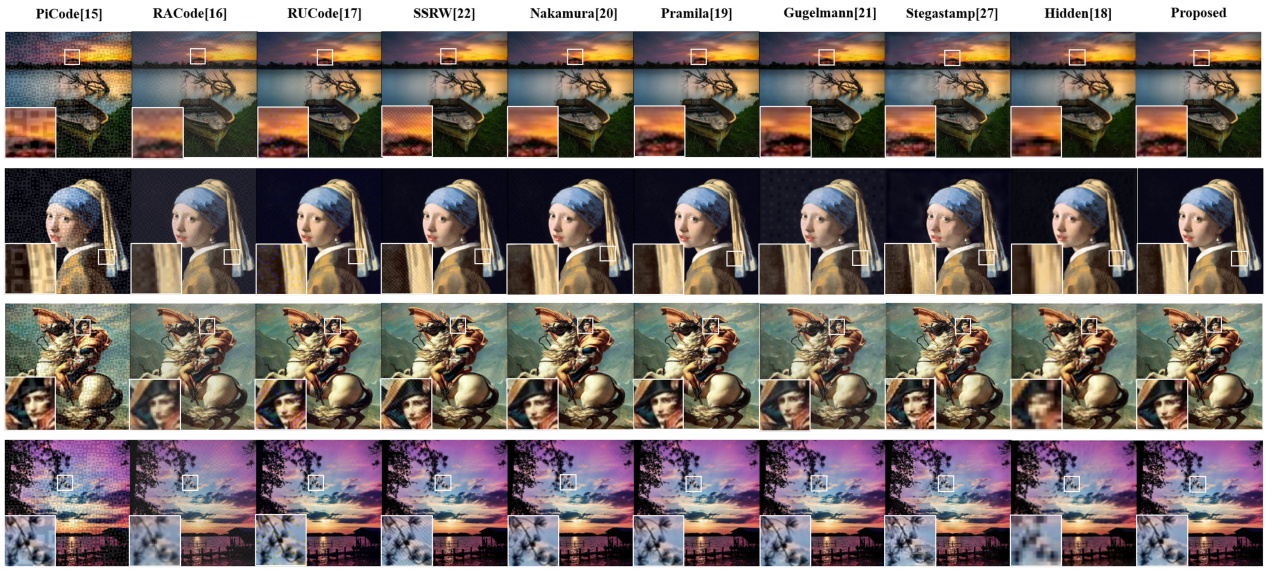


Fig. 8. Four visual samples of different methods and zoom-in to display the details. Our method can theoretically guarantee the original quality, while baseline methods will affect it.

TABLE IV
BIT ERROR RATE COMPARISON OF EXTRACTED MESSAGE WITH SCREEN-SHOOTING DISTANCE

Distance (cm)	Nakamura [20]	Primila [19]	Gugelmann [21]	Proposed
30	16.40%	23.43%	22.56%	**2.54**%
40	14.75%	23.83%	27.24%	**7.03**%
50	17.81%	20.70%	32.22%	**3.71**%
60	18.44%	20.70%	27.05%	**6.84**%
70	19.50%	19.92%	30.96%	**5.27**%

TABLE V
BIT ERROR RATE COMPARISON OF EXTRACTED MESSAGE WITH DIFFERENT SCREEN-SHOOTING ANGLES

Angle	Nakamura [20]	Primila [19]	Gugelmann [21]	Proposed
Left 40°	19.63%	19.92%	35.84%	**14.46**%
Left 30°	16.31%	16.41%	31.64%	**7.03**%
Left 15°	15.44%	19.91%	24.51%	**7.05**%
Right 15°	13.88%	20.70%	23.63%	**5.27**%
Right 30°	15.83%	20.31%	31.05%	**11.52**%
Right 40°	**22.27**%	22.34%	34.18%	23.52%
Up 40°	**16.11**%	36.17%	33.69%	23.25%
Up 30°	13.58%	22.26%	33.40%	**6.25**%
Up 15°	20.52%	17.97%	26.07%	**3.13**%
Down 15°	16.50%	22.27%	24.51%	**12.70**%
Down 30°	16.80%	21.48%	29.79%	**14.12**%
Down 40°	**26.22**%	39.06%	32.32%	29.89%

beyond the training scope, the bit error rate becomes higher than 14%. To make the algorithm more robust to larger shooting angles, adding more corresponding training datasets is further needed.

It is worth noting that in the vertical shooting angle test, when captured under the screen ("Down $15° - 45°$"), the bit error rate is much higher than that captured above the screen: the reason

TABLE VI
BIT ERROR RATE COMPARISON OF EXTRACTED MESSAGE WITH DIFFERENT SCREEN CAPTURE DEVICES

Screen \ Phone	Huawei P30 Pro	iPhone 6s	Mi 9	iPhone Xs Max
AOC-G2770PF	3.48%	0.79%	0.78%	4.17%
ViewSonic VA2261	5.47%	6.77%	8.85%	9.37%
Lenovo P22i	3.42%	4.16%	8.07%	2.60%

can be explained as that the luminous angle of the screen is not the same for all directions. Shooting under the screen readily causes significant color distortion, which leads to large image distortion even if shooting at a small angle, so the bit error rate will be larger than that of shooting at the same angle of other directions.

In addition, compared with the distance testing, the bit error rate of the angle testing is less stable. This indicates that the algorithm is sensitive to the changing of the shooting angle, because compared to distance, the change of shooting angle has a greater impact on camera shooting: this is reflected in the image with much more distortion, which thus affects the extraction.

2) Adaptability to Different Devices: As mentioned before, adaptability is a key consideration for applicability. To evaluate it, we capture the embedded image with different mobile phones ("Huawei P30 Pro," "iPhone 6 s," "Mi 9" and "iPhone Xs Max") and different screens ("AOC-G2770PF," "ViewSonic VA2261" and "Lenovo P22i") under the same conditions of "30 cm, 0°". It can be seen from Table VI that the proposed scheme can be applied to various devices and the bit error rates of all devices are less than 10%. However, since the dataset is generated with "AOC-G2270PF" and "Huawei P30 Pro," we discussed the two aspects of phone and screen of the testing results: for the screen, we can see that compared with the other two screens, the BER of "AOC-G2270PF" is slightly lower because the network is trained based on the dataset generated from "AOC-G2270PF".

TABLE VII
BIT ERROR RATE COMPARISON OF EXTRACTED MESSAGE WITH DIFFERENT MOBILE PHONE SHUTTER SPEED AND SCREEN FREQUENCY

Shutter Speed (s)	1/30	1/60	1/100	1/200
Screen Frequency (60 *Hz*)	31.25%	5.99%	2.87%	3.385%
Screen Frequency (144 *Hz*)	29.17%	28.39%	7.46%	3.02%

TABLE VIII
THE BER OF DIFFERENT EXTRACTION CONDITIONS WITH "40-CM" CAPTURED IMAGE

Conditions	Single-Image	30 fps Video	60 fps Video
BER	7.03%	1.95%	1.56%

TABLE IX
THE MOS VALUES OF DIFFERENT VIDEOS UNDER DIFFERENT REFRESH FREQUENCIES

Message Consistency				
Refresh Frequency	60 *Hz*	100 *Hz*	120 *Hz*	144 *Hz*
MOS	1	1	1	1
Message Changing				
Refresh Frequency	60 *Hz*	100 *Hz*	120 *Hz*	144 *Hz*
MOS	3	1.8	1.2	1

TABLE X
THE EXTRACTION BER UNDER DIFFERENT REFRESH FREQUENCIES

Refresh frequency	60 *Hz*	100 *Hz*	120 *Hz*	144 *Hz*
BER	4.22%	7.97%	18.28%	28.44%

For the phone, it can be seen that the performance of "Huawei P30 Pro" is comparable with those of different screens: however, the performances of the other three phones significantly vary with different screens. Based on the results in Table VI, we can draw the following two conclusions:

1) The well-trained network can work not only with the devices that are used for generating the training dataset but also with other phones and screens, which indicates high adaptability to different devices.
2) The BER with the devices used for generating the dataset is lower than that with other devices, which means that using more diverse devices to generate the training dataset is potentially beneficial to realizing higher accuracy.

3) Adaptability to Different Frequencies: In Table VII, we further provide the results for different combinations of screen and phone frequencies. We can find that faster shutter speed will produce better extraction results. For example, when displaying the embedded image at 60 *Hz*, if the phone's shutter speed is less than 1/30 s, the extraction bit error rate is less than 6%, but when shooting with 1/30 s shutter speed, the bit error rate is higher than 30%. According to Nyquist-Shannon sampling theorem, this is because the captured image will be the fusion of the two displayed continuous frames and some important information is missing in this condition. Similarly, when the refresh frequency is 144 *Hz*, the extraction can only succeed with "1/100 s" and "1/200 s" shutter speeds.

4) The Extraction Difference Between Video and Single-Image: The performances of two different extraction methods are illustrated in this section: single image capturing and video recording. Specifically, we capture the screen at "40" cm and further record for 1 s per image at 30 fps and 60 fps. We then select 5 random frames of the recorded video to extract the message. The minimum BER of the 5 images is applied as the BER of each video extraction.

Table VIII illustrates the results of the message extraction via different conditions. It can be seen that video recording instead of image capturing can greatly improve the extraction performance: the BER of the video recording extraction is less than 2%. The reason is that the video recording process can be regarded as a continuous capturing process. In a single-image capturing process, the distortion caused by frame changing may greatly influence the captured image. However, in video recording extraction, such influence can be reduced by extracting many frames in the video, which represents a spread spectrum correction. It is worth noting that theoretically, the message artifacts may not be recorded in 30 fps video since the refresh rate is 60 *Hz*. However, we find them to be extractable in practice. The main reason for this is that even if the sampling frequency of the mobile phone is twice that of the monitor in theory, in practice, the monitor is not displayed in strict accordance with 60 *Hz* as is the mobile phone recording: as a result, the information recorded by the mobile phone is not equal to the superposition of two adjacent images, so the mobile phone can still record message information.

5) The Results of Video Carrier: In this section, we mainly show and discuss two aspects of the video carrier results: the visual quality and the extraction performance. Since the screen is constantly displaying images at a refresh rate of 60 *Hz* in the proposed method, the carrier can be either an image or a video. The video we used in this paper is "Big Buck Bunny" [35], as shown in Figure 9. We embedded the same and different messages into each frame of the video and displayed them at 60 *Hz*, 100*Hz*, 120 *Hz* and 144 *Hz*. We then evaluated the visual quality with MOS and captured 10 images of the video at each refresh rate to evaluate the extraction performance. The corresponding results are shown in Table IX and Table X.

The visual quality of the embedded video is measured by MOS test. Specifically, we prepare the "Big Buck Bunny" video under the conditions of "message consistency" (each frame of the video is embedded with the same message) and "message changing" (each frame of the video is embedded with different messages). We then ask 30 volunteers to assign a score from 1 (No Flicker) to 3 (Heavy Flicker). From Table IX, it can be seen that when the message is not changing with each frame, the video will not flicker even under 60 *Hz* display, and none of the volunteers are able to sense the artifacts of the message from the embedded video. However, if the message varies with each frame, the flicker, that is, the artifacts of the message, will be observed with 60 *Hz* display, but the visual quality becomes better and better with increasing refresh frequency. We can thus

Fig. 9. The generated I_+ and I_- of 8 continuous frames in the "Big Buck Bunny" video.

conclude that the message change is a very important reason to cause visual distortion, since if the message changed frame by frame, not only the content of the video but also the message artifacts, would be different.

For message extraction, the BER is shown in Table X. Just as in the conclusion in Section IV-C3, the message remains extractable with 60 *Hz* and 100 *Hz*, but when facing 120 *Hz* and 144 Hz, the mobile phone will not be able to effectively capture the artifacts due to the Nyquist-Shannon sampling theorem, so the BER will be greatly increased when facing 120 *Hz* and 144 *Hz*.

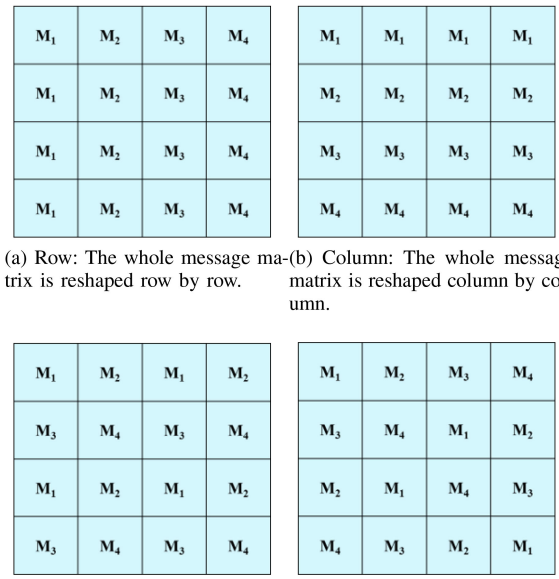
In summary, to apply the proposed scheme into video carriers, the message should remain unchanged frame to frame to satisfy 60 *Hz* display. If the video is shown with higher refresh frequency, the receiver should be adaptive to its settings.

D. Ablation Study

1) Importance of Message Matrix Arrangement: To better illustrate the importance of the proposed message matrix arrangement, we compared the proposed message arrangement with the other three arrangements shown in Figure 10 with respect to the bit error rate. Note that one whole message consists of M_1, M_2, M_3, M_4, so the bit error rate is calculated by the M_1, M_2, M_3, M_4 combination with minimum error bits. Specifically, we use the image captured from a different distance as the test image data. For each image, we divide the extracted message matrix into 4×4 parts, and then sum the error bits corresponding to each part. We assume that the message is reshaped: (a) row by row; (b) column by column; (c) square by square and (d) by scrambling. We can calculate the bit error rate of a whole message by counting each combination of M_1, M_2, M_3, M_4 and choosing the one with minimum error bits. The results are shown in Table XI.

From Table XI we can see that with the distance of $30-60$ cm, the proposed message matrix arrangement maintains a lower BER compared with other arrangements, where at 70 cm, the minimum BER is obtained from the "Square" arrangement.

In most cases, the proposed message matrix arrangement can achieve better extraction performance. We summarize the reason as: The scrambled watermark arrangement can effectively distribute the complete information in each row, column and square region so that the watermark can remain extractable as long as one row/column/square message is surviving from the screen-shooting process, which makes the method more robust.



(a) Row: The whole message matrix is reshaped row by row.

(b) Column: The whole message matrix is reshaped column by column.

(c) Square: The whole message matrix is reshaped square by square.

(d) Scramble: The whole message matrix is reshaped by the proposed arrangement.

Fig. 10. The four different message matrix arrangements.

TABLE XI
THE BER OF DIFFERENT MESSAGE MATRIX ARRANGEMENTS WITH DIFFERENT CAPTURE DISTANCE. "ROW," "COLUMN," "SQUARE", AND "SCRAMBLE" INDICATE THE DIFFERENT ARRANGEMENTS CORRESPONDING TO FIGURE 10

Distance (cm)	30	40	50	60	70
Row	2.93%	8.98%	5.08%	7.03%	6.05%
Column	7.42%	9.18%	6.05%	7.81%	7.23%
Square	3.71%	7.62%	5.08%	7.62%	**4.88%**
Scramble	**2.54%**	**7.03%**	**3.71%**	**6.84%**	5.27%

2) Experiments on Different Patterns: In this paper, we propose to use the pattern generated with (1)–(4) to represent 1 bit message. However, now there are many other pattern generation schemes [20], [21], [36] proposed to express 1 b message, so we perform the experiments in this section to test two aspects of different pattern expression schemes: visual quality and extraction accuracy.

Specifically, we apply the pattern generation method in [20], [21], [36] with the size of 32×32 pixels to compare with the

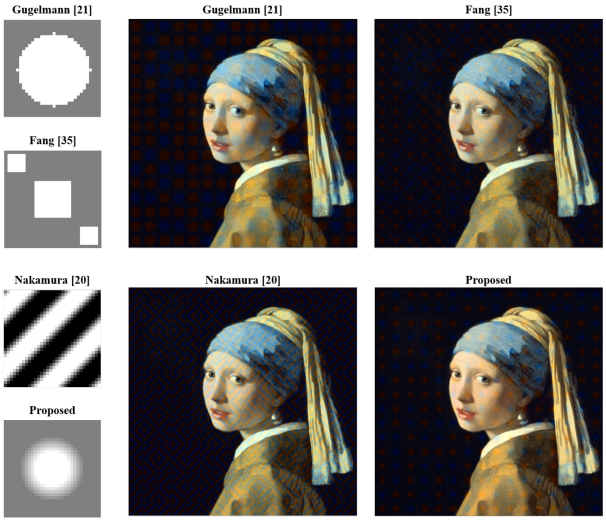


Fig. 11. The pattern as well as the encoded image appearance generated with [20], [21], [36] and the proposed scheme.

TABLE XII
THE MOS TEST SCORE OF EACH SCHEME. THE HIGHER THE SCORE, THE EASIER IT IS TO SENSE THE FLICKER

Method	Nakamura [20]	Gugelmann [21]	Fang [36]	Proposed
MOS	1.5	2.1	1.2	1

TABLE XIII
THE EXTRACTION BER OF EACH SCHEME UNDER "30" CM SCREEN SHOOTING

Method	Nakamura [20]	Gugelmann [21]	Fang [36]	Proposed
BER	3.71%	4.88%	1.13%	2.54%

proposed method. The pattern appearance as well as the encoded image are shown in Figure 11.

To better illustrate the difference between different schemes, we have generated the image dataset and trained the corresponding extraction network for each of them. The dataset generation is conducted with the settings shown in Table II. We then perform the MOS test and extraction experiments on each method with the test dataset [34], and the results are shown in Table XII and Table XIII.

In detail, we invited 30 volunteers to score the visual quality of different methods from 1 (No Flicker) to 3 (Heavy Flicker). From Table XII we can observe that the proposed pattern generation scheme maintains the best visual quality compared with other schemes. We believe the reason for this is that the brightness of the proposed pattern gradually changes from the middle to the surroundings, while the brightness of the other three schemes changes dramatically. Such a setting plays a role of visual masking to a certain extent so that the visual quality is better.

Regarding extraction accuracy, we captured the test images from "30" cm and extracted the captured images. Surprisingly, we find that the extraction network can effectively decode the message with a low bit error rate no matter the kind of pattern, which indicates the powerful ability of the proposed network.

TABLE XIV
THE EXTRACTION ACCURACY WITH/WITHOUT THE ENHANCING SUB-NETWORK AND ATTENTION SUB-NETWORK. COMPARED TO THE BASELINE "BS," ADDING ENHANCING NETWORK "ENH" AND ATTENTION SUB-NETWORK "ATT" CAN BRING SUBSTANTIAL PERFORMANCE IMPROVEMENT

Architecture	bs	bs + enh	bs + enh + att
Accuracy	93.45%	94.05%	95.66%

Fig. 12. The output image of the enhancing sub-network. *left* : The original captured image. *right* : The corresponding enhanced image.

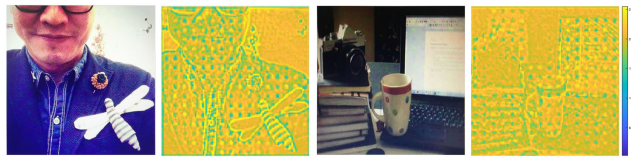


Fig. 13. The output attention map of the attention sub-network. *left* : The captured image. *right* : The corresponding attention map.

In summary, when BER is within the error correction capability, the key to choosing a pattern generation scheme is the visual quality. From this point of view, the proposed method maintains the best performance.

3) Importance of Each Sub-Network: Rather than just using a single extracting network, our method consists of three sub-networks. To demonstrate the importance of each sub-network, we have conducted two ablation experiments with/without the enhancing sub-network and attention sub-network. It can be seen from Table XIV that incorporating the enhancing network and the attention sub-network can bring approximately 0.6% and 1.61% accuracy gains, respectively. In Figure 12 and Fig. 13, we further visualize the enhancing image and the attention maps of two examples. Figure 12 indicates that even if the watermark signal is weak in the captured image, the enhancing sub-network can effectively enlarge the watermark feature, which appears as a more obvious pattern in the enhanced image. Fig. 13 shows that, though it is difficult to detect the bit pattern by human eyes, the attention network can learn where the bit patterns are placed and pay different levels of attention to each pattern.

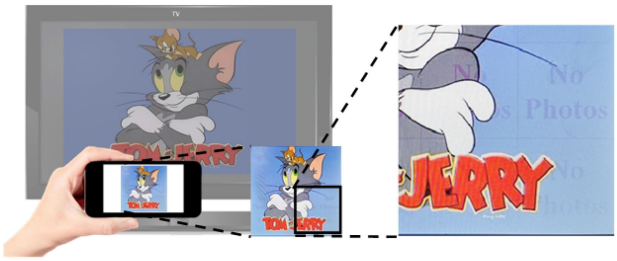
V. APPLICATIONS

In this section, we will show three typical applications of the proposed system as shown in Figure 14, which further demonstrate the broad applicability of our method.

2D image code: "TERA" code is similar to QR code and can be used as a way to realize screen-to-camera message transmission. However, in contrast with the traditional QR code, the "TERA" code will not produce any visual distortion, and thus it is more attractive to users. When users scan or capture the displayed image, the URL information can be

(a) 2D Image Code. By capturing the image displayed on the screen, the detailed information (*e.g.,* websites) can be extracted and loaded on the phone.

(b) Leak-tracking Watermark. The "TERA" code can be regarded as a watermark with the message of device information: when the secret document is sneakily shot, the leaked device can be located according to the message.



(c) Warning Watermark. By carefully designing the "TERA" code pattern, the pattern can be regarded as a sign to warn of the sneak shot behavior.

Fig. 14. The three typical applications of "TERA" Code.

extracted and transferred to the mobile phone so that a more detailed introduction of the displayed image can be loaded.

Leak-tracking watermark: "TERA" code can be regarded as a kind of screen-shooting resilient watermarking algorithm. By embedding the watermark ($e.g.$, time or device information) in confidential documents, when the confidential documents displayed on the screen are leaked out by screen-shooting, we can extract the hidden watermark from the captured photos and recover the leaking information such as leaked equipment, leaked time and employee identity, to realize accountability.

Warning watermark for IP protection: "TERA" code can also be used as a warning watermark that is only visible to the camera. When embedding the warning logo on each frame with high intensity and displaying them with an appropriate frequency, the warning logo will be invisible to human eyes and the logo will appear instead for camera devices. This can serve a warning role for IP protection in the cinema or museum.

VI. Conclusion

In this paper, we design a new screen-to-camera image code, "TERA". It is the first attempt that can satisfy the four key properties simultaneously, i.e., *great transparency*, *high embedding efficiency*, *strong transmission robustness* and *high adaptability to device types*. This method is mainly based on the inspiration of the properties of the human vision system, dedicated message embedding design and the powerful ability of a novel attention-guided extracting network. Extensive experiments also demonstrate the superiority of our method in both robustness test, visual quality and adaptability. In addition, such methods can be broadly used in many different applications such as 2D image codes, leak-tracking watermarks and warning watermarks. However, such algorithms are vulnerable to cropping attacks since the locating process might be influenced by the crop distortion, and moreover, the capacity of embeddable messages is not high. In the future, we will therefore be committed to solving these two main problems.

References

[1] Y. Huang, B. Niu, H. Guan, and S. Zhang, "Enhancing image watermarking with adaptive embedding parameter and psnr guarantee," *IEEE Trans. Multimedia*, vol. 21, no. 10, pp. 2447–2460, Oct. 2019.

[2] X. Zhong, P. Huang, S. Mastorakis, and F. Y. Shih, "An automated and robust image watermarking scheme based on deep neural networks," *IEEE Trans. Multimedia*, to be published, doi: 10.1109/TMM.2020.3006415.

[3] R. Kazemi, F. Perezgonzalez, M. A. Akhaee, and F. Behnia, "Data hiding robust to mobile communication vocoders," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2345–2357, Dec. 2016.

[4] H. Cui, H. Bian, W. Zhang, and N. Yu, "Unseencode: Invisible on-screen barcode with image-based extraction," in *Proc. IEEE INFOCOM IEEE Conf. Comput. Commun.*, 2019, pp. 1315–1323.

[5] L. Zhang *et al.*, "Kaleido: You can watch it but cannot record it," in *Proc. 21st Annu. Int. Conf. Mobile Comput. Netw.*, 2015, pp. 372–385.

[6] X. Shu and X. Wu, "Frame untangling for unobtrusive display-camera visible light communication," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 650–654.

[7] G. Woo, A. Lippman, and R. Raskar, "Vrcodes: Unobtrusive and active visual codes for interaction by exploiting rolling shutter," in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR).*, 2012, pp. 59–64.

[8] A. Wang, Z. Li, C. Peng, G. Shen, G. Fang, and B. Zeng, "Inframe : Achieve simultaneous screen-human viewing and hidden screen-camera communication," in *Proc. 13th Annu. Int. Conf. Mobile Syst., Appl., Serv.* ACM, 2015, pp. 181–195.

[9] T. Nguyen, N. Le, and Y. M. Jang, "Practical design of screen-to-camera based optical camera communication," in *Proc. Int. Conf. Inf. Netw.* , 2015, pp. 369–374.

[10] M. Izz, Z. Li, H. Liu, Y. Chen, and F. Li, "Uber-in-light: Unobtrusive visible light communication leveraging complementary color channel," in *Proc. IEEE INFOCOM 35th Annu. IEEE Int. Conf. Comput. Commun.*, 2016, pp. 1–9.

[11] S.-S. Lin, M.-C. Hu, C.-H. Lee, and T.-Y. Lee, "Efficient QR code beautification with high quality visual content," *IEEE Trans. Multimedia*, vol. 17, no. 9, pp. 1515–1524, Sep. 2015.

[12] G. Garateguy, G. R. Arce, D. L. Lau, and O. P. Villarreal, "QR images: Optimized image embedding in QR codes," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 2842–2853, Jul. 2014.

[13] Y. Lin, Y. Chang, and J. Wu, "Appearance-based QR code beautifier," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 2198–2207, Dec. 2013.

[14] M. Xu *et al.*, "Stylized aesthetic QR code," *IEEE Trans. Multimedia*, vol. 21, no. 8, pp. 1960–1970, Aug. 2019.

[15] C. Chen, W. Huang, B. Zhou, C. Liu, and W. H. Mow, "Picode: A new picture-embedding 2D barcode," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3444–3458, Aug. 2016.

[16] C. Chen, B. Zhou, and W. H. Mow, "Ra code: A robust and aesthetic code for resolution-constrained applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 11, pp. 3300–3312, Nov. 2018.

[17] C. Chen, W. Huang, L. Zhang, and W. H. Mow, "Robust and unobtrusive display-to-camera communications via blue channel embedding," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 156–169, Jan. 2019.

[18] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, "Hidden: Hiding data with deep networks," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 657–672.

[19] A. Pramila, A. Keskinarkaus, V. Takala, and T. Seppänen, "Extracting watermarks from printouts captured with wide angles using computational photography," *Multimedia Tools Appl.*, vol. 76, no. 15, pp. 16063–16084, 2017.

[20] T. Nakamura, A. Katayama, M. Yamamuro, and N. Sonehara, "Fast watermark detection scheme for camera-equipped cellular phone," in *Proc. 3rd Int. Conf. Mobile Ubiquitous Multimedia. ACM*, 2004, pp. 101–108.

[21] D. Gugelmann, D. Sommer, V. Lenders, M. Happe, and L. Vanbever, "Screen watermarking for data theft investigation and attribution," in *Proc. 10th Int. Conf. Cyber Conflict*., 2018, pp. 391–408.

[22] H. Fang, W. Zhang, H. Zhou, H. Cui, and N. Yu, "Screen-shooting resilient watermarking," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 6, pp. 1403–1418, Jun. 2019.

[23] J.-C. Liu and H.-A. Shieh, "Toward a two-dimensional barcode with visual information using perceptual shaping watermarking in mobile applications," *Opt. Eng.*, vol. 50, no. 1, 2011, Art. no. 017002.

[24] V. Nguyen *et al.*, "High-rate flicker-free screen-camera communication with spatially adaptive embedding," in *Proc. IEEE INFOCOM 35th Annu. Int. Conf. Comput. Commun.*, 2016, pp. 1–9.

[25] T. Li, C. An, X. Xiao, A. T. Campbell, and X. Zhou, "Real-time screen-camera communication behind any scene," in *Proc. 13th Annu. Int. Conf. Mobile Syst., Appl., Serv.* ACM, 2015, pp. 197–211.

[26] Y. Liu, M. Guo, J. Zhang, Y. Zhu, and X. Xie, "A novel two-stage separable deep learning framework for practical blind watermarking," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1509–1517.

[27] M. Tancik, B. Mildenhall, and R. Ng, "StegaStamp: Invisible hyperlinks in physical photographs," 2019, *arXiv:1904.05343*.

[28] R. C. Bose and D. K. Ray-Chaudhuri, "On a class of error correcting binary group codes," *Inf. Control*, vol. 3, no. 1, pp. 68–79, 1960.

[29] W. W. Peterson and D. T. Brown, "Cyclic codes for error detection," in *Proc. IRE*, vol. 49, no. 1, pp. 228–235, 1961.

[30] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*. Springer, 2015, pp. 234–241.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[32] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5967–5976.

[33] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur Conf Comput vis.* Springer, 2014, pp. 740–755.

[34] "The Usc-Sipi Image Database," Accessed: Mar. 2021. [Online]. Available: http://sipi.usc.edu/database/

[35] The "Big Buck Bunny" *Video Database*. Accessed: Mar. 2021. [Online]. Available: https://peach.blender.org/download/

[36] H. Fang *et al.*, "Deep template-based watermarking," *IEEE Trans. Circuits Syst. Video Technol.*, to be published, doi: 10.1109/TCSVT.2020.3009349.

Han Fang received the B.S. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2016. He is currently working toward the Ph.D. degree in information security with the University of Science and Technology of China, Hefei, China. His research interests include image watermarking, information hiding, and image processing.

Dongdong Chen received the Ph.D. degree under the joint Ph.D. program between the University of Science and Technology of China, Hefei, China and Microsoft Research Asia, Beijing, China. He is currently a Senior Researcher with Microsoft Research. His research interests mainly include style transfer, image generation, image restoration, low-level image processing, and general representation learning.

Feng Wang received the B.S. degree, in 2018 from the University of Science and Technology of China, Hefei, China, where he is currently working toward the master's degree in information and communication engineering. His research interests include optical watermarking, information hiding, and 3D mesh watermarking.

Zehua Ma received the B.S. degree, in 2018 in information security from the University of Science and Technology of China, Hefei, China, where he is currently working toward the M.S. degree in information security. His research interests include image watermarking, information hiding, and image processing.

Honggu Liu received the B.S. degree from Southwest Jiaotong University, Chengdu, China, in 2018. He is currently working toward the Ph.D. degree in information security with the University of Science and Technology of China, Hefei, China. His research interests include deepfake and adversarial examples.

Wenbo Zhou received the B.S. degree from the Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2014 and the Ph. D degree from the University of Science and Technology of China, Hefei, China, in 2019. He is currently a Postdoctoral Researcher with the University of Science and Technology of China. His research interests include information hiding and AI security.

Weiming Zhang received the M.S. and Ph.D. degrees from Zhengzhou Information Science and Technology Institute, Zhengzhou, China, in 2002 and 2005, respectively. He is currently a Professor with the School of Information Science and Technology, University of Science and Technology of China, Hefei, China. His research interests include information hiding and multimedia security.

Nenghai Yu received the B.S. degree, in 1987 from the Nanjing University of Posts and Telecommunications, Nanjing, China, the M.E. degree, in 1992 from Tsinghua University, Beijing, China, and the Ph.D. degree, in 2004 from the University of Science and Technology of China, Hefei, China. He is currently a Professor with the University of Science and Technology of China. His research interests include multimedia security, multimedia information retrieval, video processing, and information hiding.