This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3061801, IEEE Transactions on Multimedia

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, how to transmit information among different devices with multimedia carrier has drawn much attention from the research community. This paper focuses on the important user scenario "screen-to-camera information transmission". Along this direction, image coding based techniques have been shown to be the most popular and effective way in the past decades. However after careful study, we find none of 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*. It is mainly because these properties are contradictory with each other. So in this paper, we propose an screen-to-camera image code dubbed "TERA" with Transparency, Efficiency, Robustness and Adaptability, which makes it possible to circumvent the contradiction among the above four properties for the first time. Generally, it adopts the color decomposition principle to ensure the visual quality and the superposition-based scheme to ensure embedding efficiency. And 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*—**Screen-to-camera Image Code, Transparency, Efficiency, Robustness, Adaptability, Color Decomposition, Attention-guided

I. Introduction

In recent decades, digital devices have been developed rapidly and become more and more common, such as personal computers, mobile phones and AR/VR devices. With these devices, information computation and transmission turn to be more efficient and convenient. But due to the existence of physical gap 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], 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 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 capturing with the camera and post-processing operations. Therefore, the 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 shown to be the most popular and effective way. 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 back.

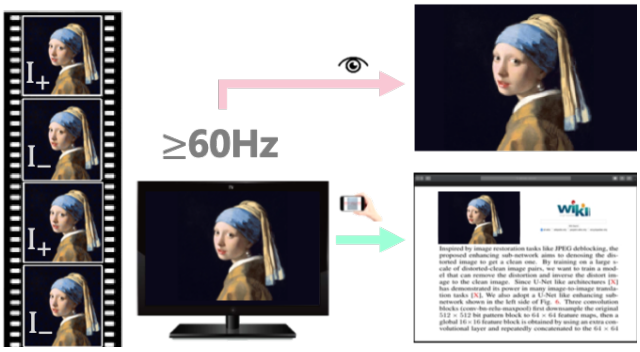


Fig. 1: The schematic diagram of the proposed "TERA" code. When alternate displaying two embedded frames on the screen with no less than $60Hz$, the human vision system cannot observe any visual difference, but the embedded message (e.g. website) can be extracted out in camera.

As the 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. For transparency, it is to ensure that the encoding process should keep the original visual quality of the host image as much as possible so that human observers cannot even notice it. And for embedding efficiency, it aims to reduce the computation burden in the screen devices side because of their limited computation ability. Different from the first two requirements which are often imposed on the screen side, the last two is to make sure the information embedded in the camera-captured image can be correctly

Han Fang, Feng Wang, Zehua Ma, Honggu Liu, Wenbo Zhou, Weiming Zhang and Nenghai Yu are all with CAS Key Laboratory of Electromagnetic Space Information, University of Science and Technology of China, Hefei, 230026, China. (e-mail: fanghan@mail.ustc.edu.cn, zhangwm@ustc.edu.cn, ynh@ustc.edu.cn), Dongdong Chen is with Microsoft Research, Redmond, 98052, USA. (e-mail: cddlyf@gmail.com). Corresponding author: Weiming Zhang and Nenghai Yu.

This work was supported in part by the Natural Science Foundation of China under Grant U1636201, Anhui Initiative in Quantum Information Technologies under Grant AHY150400, and Exploration Fund Project of University of Science and Technology of China under Grant YD3480002001.

extracted out at the decoder side no matter which types of camera or screen are used.

However, after careful study, we find none of existing methods [18]–[22] can satisfy the above four properties simultaneously. This is because of the inherent contradiction among these four properties. In more details, 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 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. But such an operation will incur higher calculation cost. So basically, 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" with **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 human vision system (HVS) and utilize the observation [4], [5] that the HVS will fuse two images into one if they are refreshed in a high frequency (no less than 60 *Hz*), and this fresh rate is satisfied in 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 in human eyes is the composition of these two frames, that is, the original image, so as to achieve high transparency. But unlike HVS, the shutter speed of modern cameras is much higher and will instead capture the decomposed frame that contains information. In this way, it theoretically guarantees the visual quality observed by human beings and remains capable of transmitting information to camera devices, as shown in Figure 1.

As 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 in a short-time and content-independent way.

The robustness and adaptability is 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, the accurately extraction which reflects the robustness and adaptability can be greatly guaranteed.

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

- We propose a novel image coding scheme "TERA" for screen-to-camera information transmission, based on which a complete system is further constructed. Besides, 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 super 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 tried which demonstrate the potential commercial value of this system.

II. Related Work

For screen-to-camera information transmission, classic methods include traditional communication techniques like cable and wireless transmission, and image coding based schemes. In some specific scenarios, the former way is often more stable and reliable by using some very strict communication agreements or rules. However, this type of methods is often not that flexible. With the development of smart mobile phones, image coding based scheme becomes more and more 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 like 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], RU Code [17]. In PiCode [15], they express 0 and 1 by using two templates: inner-dark/outer-bright and inner-bright/outer-dark. In the extracting side, they perform a 2D matched filter to demodulate the message. In RA Code [16], based on the analysis of 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. But there is a common problem with these algorithms, that is, they cannot balance the robustness and invisibility well, 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 methods. Since they are

also based on the properties of HVS, their visual quality is often great. But 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 in screen-to-camera communication, the sender and the receiver always cooperatively work in pairs, the only transmission distortion occurred resulted from the channel of the fixed the sender and the receiver. However, as for image code, the decoder equipment may not be fixed so decoding process may occur on 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, high intensity modification is needed in the traditional screen-to-camera communication schemes. Meanwhile, to cover the visual distortion, high fresh rate screen and the corresponding receiver are required. So traditional screen-to-camera communication schemes always rely on special equipment. However, the commonly used screen nowadays can support $60Hz$ display, which is enough to realize non-visual-distortion under low embedding intensity. So as long as we can ensure the extraction accuracy with low embedding intensity, such restrictions will be lifted. And 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 cares more about quality and robustness. By analyzing the features of the image itself, they often hide information into the texture or color components of an image with handcraft algorithms [19]–[22] or deep learning networks [**?**], [18], [26], [27]. Works [19]–[21] propose to use a set of templates to represent '0/1' bit and embed them with an HVS mask in the host image to represent the message. At 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 like 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 camera-shooting process to obtain screen-shooting resistance. Since such methods should greatly balance the robustness and transparency, their embedding efficiency is not that high. Besides, 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 not good enough. Compared with all the above methods, our paper is the first than can satisfy all the four properties.

TABLE I: The comparison of different algorithms in four respects: Transparency, Embedding efficiency, Robustness and Adaptability. Compared with the other three type 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	✓	×	✓	✓

III. Method

The framework of the whole system is shown in Figure 2. Firstly, 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. And 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 with no less than 60 Hz, the complete invisibility of visual distortion can be realized. Then the image on the screen is captured by cameras to conduct the extracting process. At 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, the 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 until no error is detected or all combinations are tried.

A. Message Matrix Generation

In the process of screen-to-camera, there may be various distortions such as moiré, light and shutter distortion, and different distortions will influence the image content from different aspects. For example, Moiré and light distortion will incur the information loss in one local continuous area, so embedding the complete information unit multiple times is necessary. As 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 the robustness, we design a robust message generation scheme as shown in Figure 3. Specifically, given the original information sequence, we first generate a sequence m of length

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3061801, IEEE Transactions on Multimedia

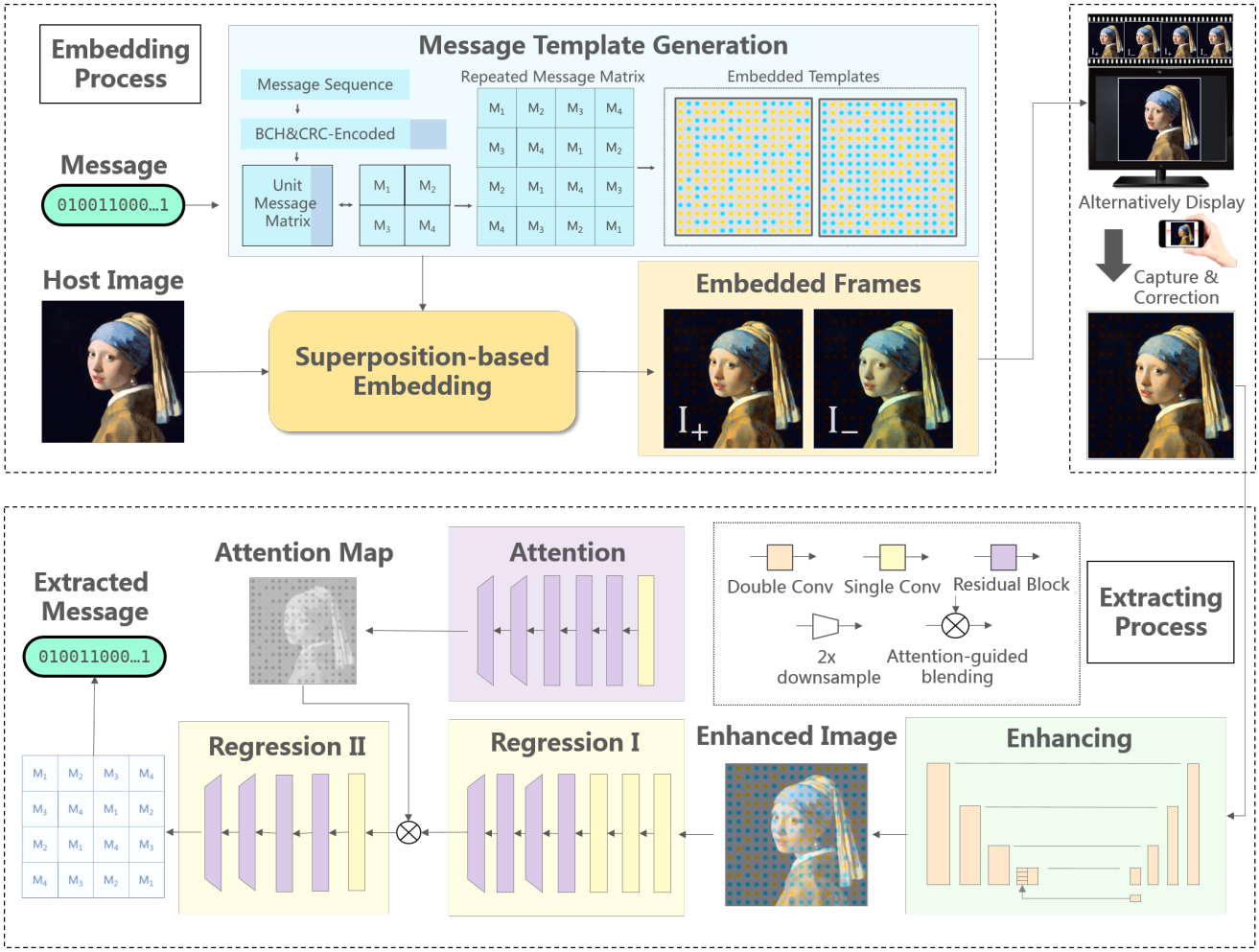


Fig. 2: The framework of the whole process. It consists of two main parts: the embedding process and the extracting process. The left part indicates the message arrangement as well as the message embedding part. After embedding the message, the embedded frames are alternatively displayed with no less than 60Hz 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 consist of an enhancing sub-network, an attention sub-network, and a regression sub-network to realize the accurate extracting.

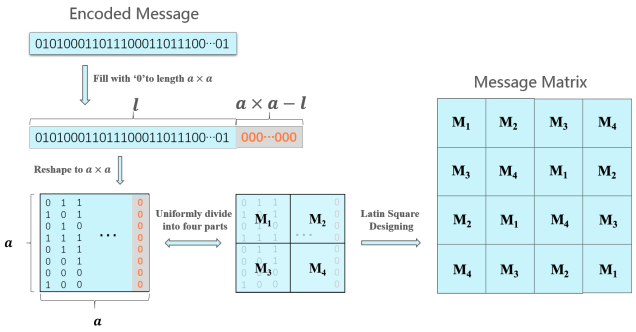


Fig. 3: The specific Latin square designing (LSD) arrangement of the message matrix. The message is first encoded with BCH & CRC coding, then the encoded message is zero padding and reshaping into size $a \times a$. Then one complete message matrix is uniformly divided into four parts and further repeated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

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. But with the message arrangement, the whole information can always be combined from the clean region.

l with the error detection and correction ability by using CRC and BCH coding. Then we resize m into a matrix with size of $a \times a$ (zeroing the part of $a \times a - l$). After that, to repeatedly embedded 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.

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.

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

B. Superposition-based Message Embedding

In order to enable online message matrix embedding in a high refreshing frequency, the embedding efficiency is the key consideration. Different with existing methods [21], [22] that need detailed texture analysis to find the 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 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 Eq. 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 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 Eq. 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_-. So the embedded image can be generated by Eq. 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 with no less than 60 *Hz*, so that human eyes can only see one still image in the screen, but the camera can effectively capture the embedding artifacts, thus the transparency in human eyes and recordable in camera can be both achieved.

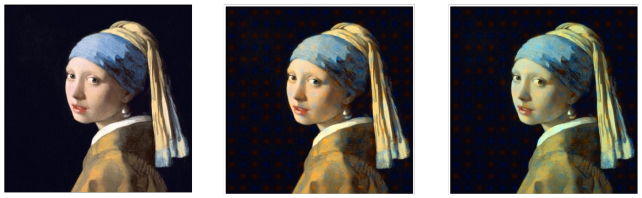


Fig. 5: One example of original image and its corresponding symmetrically embedded images. The left image is the host image, the middle and right image are the embedded image I'_+ and I'_-.

It's worth noting that alternatively displaying two frames to realize visual distortion free have 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 in order to compensate for the visual distortion caused by high embedding intensity, the algorithm requires a higher refresh rate. 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 capturing.

C. Attention-guided Extraction Network

To extract the information from the captured image, we first do perspective correction then feed the corrected image into the following extraction network. In order to achieve higher extracting 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 takes the distorted image $I_d \in \mathcal{R}^{3*H*W}$ as input and generate the enhanced image $I_E \in \mathcal{R}^{3*H*W}$; (2) The attention sub-network A_t with parameter θ_{A_t} receives I_E and calculate the attention map $A_{I_E} \in \mathcal{R}^{64*H/4*W/4}$of I_E; (3) The regression sub-network is divided into 2 parts. The regression sub-network-1 R_1 with parameter θ_{R_1} takes I_E as input and generate 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 like JPEG deblocking, the proposed enhancing sub-network aims to recover the distorted information back as much as possible for following extracting. Since U-Net [30] like architectures has demonstrated its power in many image-to-image translation tasks, we adopt a U-Net like enhancing sub-network shown in Figure 6.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3061801, IEEE Transactions on Multimedia

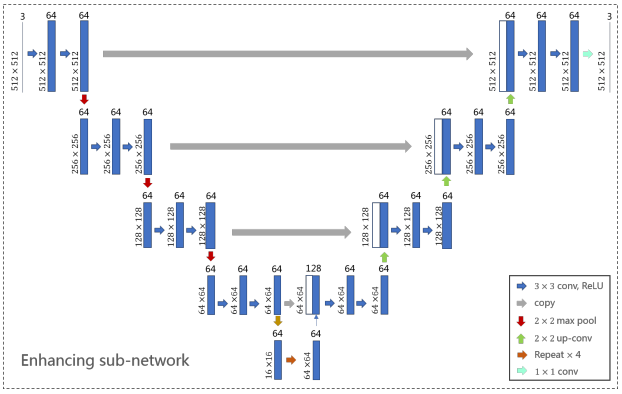


Fig. 6: The detail information of enhancing sub-network with the input size $512 \times 512 \times 3$.

In details, 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 get 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 a lot of 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, the objective of enhancing sub-network is to minimize the distance between I_{em} and I_E by updating parameters θ_E:

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

2) Attention sub-network.: Since the screen-to-camera process will cause irreversible distortions on the image, even after enhancing, the features of some regions still cannot be extracted correctly. 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. For the detailed network structure, it consists of five residual blocks [31] where the second and fourth blocks dowsample the feature maps by $1/2$, so the size of 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 collaborate with the attention sub-network well, the specific architecture are shown in Figure. 7. Given an enhanced image, the sub-network-1 R_1 is responsible to encode 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 details, 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 extracting 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 constraint the image quality of the enhanced image, we ultilize the adversarial network. The enhancing sub-network are trying 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}:

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

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

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

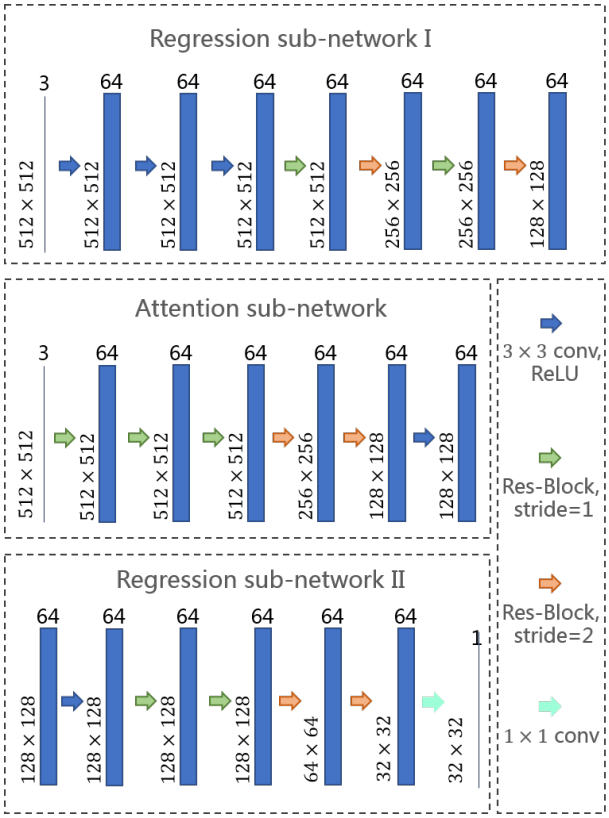


Fig. 7: The detail information of attention sub-network and regression sub-network with the input size $512 \times 512 \times 3$.

In this paper, we use the 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:

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

where $\lambda_1, \lambda_2, \lambda_3$ are the loss weights to balance these three terms and 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 it can be automatically learned very well jointly.

6) Training Process: For better performance, the whole network is trained in a supervised way. We firstly 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 enhance 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 generated a high-level intermediate features F_2 and further fed into regression sub-network-II to get the final extracted message matrix M. And the loss function in Eq. (10) is applied to train the whole network in an end-to-end way. It can be seen that we have not used any explicit attention guidance for the attention sub-network, but we find it can be automatically learned very well 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 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. So 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 30-60cm and $-30° - 30°$, 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'. And the test dataset is the classical USC-SIPI image dataset [34]. In order to realize the alternately displaying, we have written a script that can alternating display the specified image at the current fresh 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. And direct alternately displaying two images can avoid the visual distortion.

TABLE II: The detailed configuration parameters when collecting 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°

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 the MOS score of proposed method is much better than that of other baseline methods. Since the displayed frequency is twice bigger than the frequency human eyes can be detected, the image human observers can see on the screen is just 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 more or less. We further provide some visual results in Figure 8. And 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 screen-shooting resilient watermarking method is higher than that of 2D image-code methods. So for fair comparison in the following robustness test, we choose three algorithms with MOS score greater than 4.

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 $[-40°, 40°]$ 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 5 bits errors, when BER is below $5/64 = 7.81\%$, the message can be lossless recovered.

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. So we can conclude that

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3061801, IEEE Transactions on Multimedia

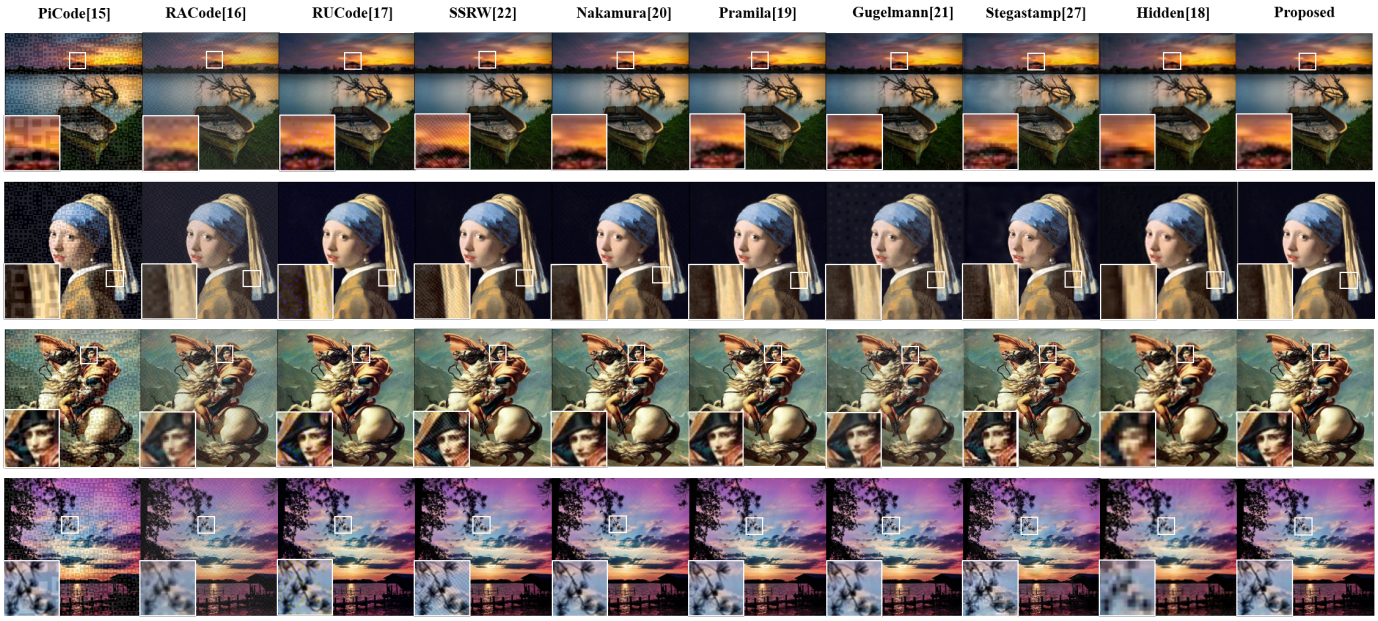


Fig. 8: Four visual samples of different methods and zoom-in to see the details. Our method can theoretically guarantee the original quality while baseline methods will affect it more or less.

TABLE III: The visual quality assessment of different schemes with the mean opinion score (MOS) test, the bigger 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**

TABLE IV: Bit error rates 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 rates 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%

under the same level of visual quality, the proposed algorithm performs better. Besides, even in the distance of 70cm, which is not appeared in the training set, it can realize high extraction accuracy. So the changing of distance affects little on the performance of the algorithm.

Table V shows the bit error rates of different algorithms at different capture angles. It can be seen that, when captured in the horizontal angle, the bit error rates of angle within $[-30°, 30°]$ are all less than 12%.But when the shooting angle is beyond the training scope, the bit error rate becomes higher than 14%. To make it more robust to larger shooting angles, adding more corresponding training datasets is further needed.

It's worth noting that in vertical shooting angle test, when captured under the screen ("Down $15° - 45°$"), the bit error rate is much bigger than that captured above the screen, the reason can be explained that the luminous angle of the screen is not the same for all directions. Shooting under the screen is easy to cause a lot of color distortion, which leads to a large image distortion even if shooting at a small angle, so the bit error rate will be larger than that of shooting in the same angle of other directions.

Besides, 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, which is reflected in the image with much more distortion, 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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3061801, IEEE Transactions on Multimedia

TABLE VI: Bit error rates comparison of extracted message with different screen shooting 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%

mobile phones ("Huawei P30 Pro", "iPhone 6s", "Mi 9" and "iPhone Xs Max") and different screens ("AOC-G2770PF", "ViewSonic VA2261" and "Lenovo P22i") under the same condition of "30cm, 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%. But since the dataset is generated with "AOC-G2270PF" and "Huawei P30 Pro", we discussed the testing results from 2 aspects of phone and screen: As for screen, we can see that compared with the other two screens, the BER of "AOC-G2270PF" is a little lower because the network is trained based on the dataset generated from "AOC-G2270PF". As for the phone, it can be seen that the performance of "Huawei P30 Pro" maintains comparable with different screens, however, the performance of the other three phones varies a lot with different screens. So based on the results in Table VI, we can draw the following two conclusion:

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 the high adaptability to different devices.

2). The BER with the devices which are used for generating the dataset is lower than that with other devices, which means using more diverse devices to generate training dataset is potentially beneficial to realize higher accuracy.

TABLE VII: Bit error rates 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 (60Hz)	31.25%	5.99%	2.87%	3.385%
Screen Frequency (144Hz)	29.17%	28.39%	7.46%	3.02%

3) Adaptability to Different Frequencies.: In Table VII, we further provide the results for different combination of screen and phone frequencies. We can find that faster shutter speed will produce better extracting results. For example, when displaying the embedded image in 60Hz, if the phone's shutter speed is less than 1/30s, the extraction bit error rate is less than 6%, but when shooting with 1/30s shutter speed, the bit error rate is higher than 30%. According to Nyquist–Shannon sampling theorem, it 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 fresh frequency is 144Hz, the extraction can only succeed with "1/100s" and "1/200s" shutter speed.

4) The Extraction Difference Between Video and Single-image: The performance of two different extraction ways: single image capturing and video recording are illustrated in this section. Specifically, we capture the screen at "40" cm and further record them for 1 second per image with 30 fps and 60 fps. Then we 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: The BER of different extraction conditions with "40-cm" captured image.

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

TABLE IX: The MOS value of different videos under different fresh frequency.

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

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. And one single-image capturing process, the distortion caused by frame changing may greatly influence the captured image. But in video recording extraction, such influence can be reduced by extracting many frames in the video, which appeared as a spread spectrum correction. It is worth noting that theoretically, the message artifacts may not be recorded in 30fps video since the fresh rate is 60Hz. However, we find it extractable in practice. The main reason is that even if the sampling frequency of the mobile phone is twice that of the monitor in theory, but in practice, the monitor is not displayed in strict accordance with 60Hz and so does the mobile phone recording, which leads to that 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 the result of video carrier in two aspects, the visual quality and the extraction performance. Since in the proposed method, the screen is constantly displaying images at a refresh rate of 60Hz, so 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 message into each frame of the video and display them with 60Hz, 100Hz, 120Hz and 144Hz. Then we evaluate the visual quality of them with MOS and capture 10 images of the video in each fresh 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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2021.3061801, IEEE Transactions on Multimedia

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

TABLE X: The extraction BER under different fresh frequency.

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

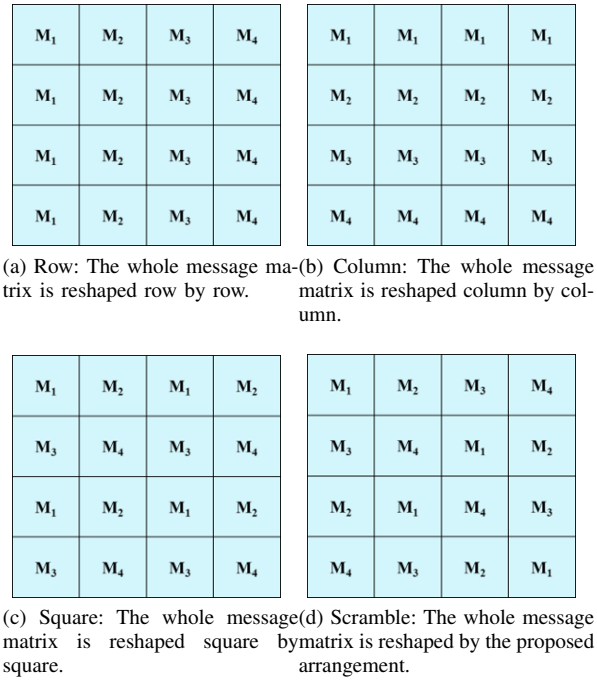
MOS test. Specifically, we prepare the "Big Buck Bunny" video under the conditions of "message consistency" (each frame of the video is embedded with the same message) and "message changing" (each frame of the video is embedded with different messages). Then we ask 30 volunteers to assign a score from 1 (No Flicker) to 3 (Heavy Flicker). From Table IX, it can be seen that when message is not changing with each frame, the video will not flicker even under 60 *Hz* display, all the volunteers are not 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 with the fresh frequency increase, the visual quality becomes better and better. So we can conclude that the message changing 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 will be different.

As for message extraction, the BER is shown in Table X. Same as 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 able to effectively capture the artifacts due to the Nyquist–Shannon sampling theorem, so the BER will be greatly increased in 120 *Hz* and 144 *Hz*.

In summary, to apply the proposed scheme into video carrier, the message should remain unchanged frame to frame in order 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 arrangement shown in Figure. 10 in the aspect of 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.



(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 arrangement.

Specifically, we use the image captured from different distance as the test image data. For each image, we divide the extracted message matrix into 4×4 part, and then sum the error bits corresponding to each part. Assuming that the message is reshaped : (a) row by row; (b) column by column; (c) square by square and (d) scrambling. We can calculate the bit error rate of a whole message by counting each combinations 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 in 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.

So in most cases, the proposed message matrix arrangement can achieve better extraction performance. We summarize the reason as: The scrambled watermark arrangement can effectively make the complete information distributed in each row, column and square region, so that the watermark can maintain extractable as long as one row/column/square mes-

TABLE XI: The BER of different message matrix arrangement with different capture distance. “Row”, “Column”, “Square” and “Scramble” indicate the different arrangement 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%

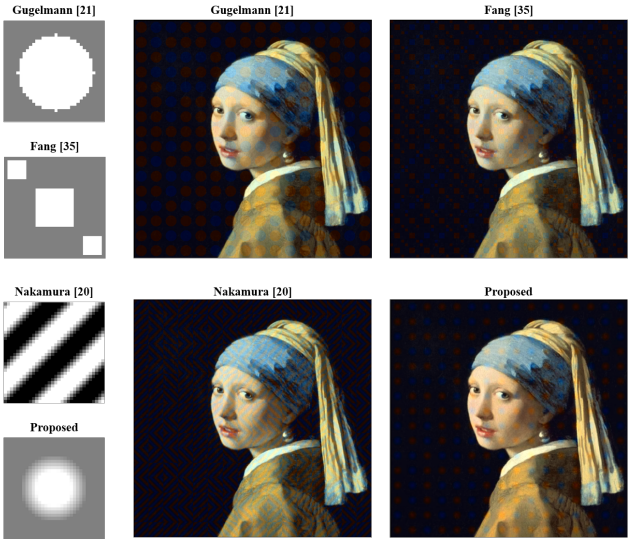


Fig. 11: The pattern as well as the encoded image appearance generated with [20], [21], [36] and the proposed scheme.

sage surviving from the screen-shooting process, which makes the method more robust.

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

Specifically, we apply the pattern generation method in [20], [21], [36] with the size of 32×32 pixels to compared with the 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. Then we perform the MOS test and extraction experiments on each methods with the test dataset [34], the results are shown in Table XII and Table XIII.

In detail, we invite 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 is that the brightness of the proposed pattern gradually changes from the middle to the surrounding, 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.

TABLE XII: The MOS test score of each schemes. The bigger 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 schemes under “30”cm screen shooting.

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

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

In summary, when BER is within the error correction capability, the key to choose 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 about 0.6% and 1.61% accuracy gain respectively. In Figure 12 and Figure 13, we further visualize the enhancing image and the attention

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

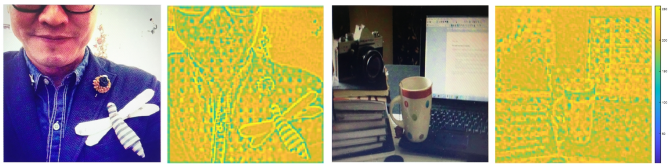


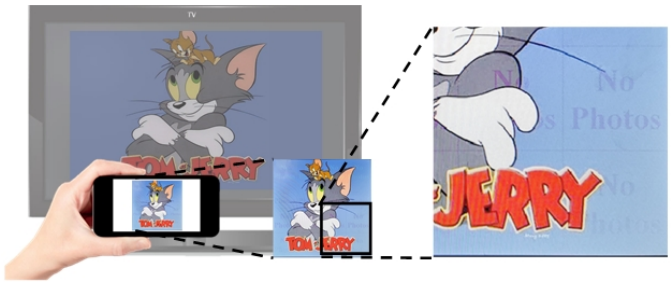
Fig. 13: The output attention map of the attention sub-network. *left:* The captured image. *right:* The corresponding attention map.

TABLE XIV: The extracting accuracy with/without the enhancing sub-network and attention sub-network. Compared to the baseline "bs", adding enhance 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%

(a) 2D Image Code. By capturing the image displayed on the screen, the detail 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 sneaky 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 kind of sign to warn the behavior of sneak shot.

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

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. While in Figure 13, it shows that, though it is hard to see the bit pattern by human eyes, the attention network can learn where the bit patterns are placed and pay different attentions 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. But different from the traditional QR code, the "TERA" code will not produce any visual distortion, thus it is more attractive to users. When users scan or capture the displayed image, the URL information can be extracted and transferred to the mobile phone, so that a more detailed introduction of the displayed image can be load.

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 in 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, so as to realize the accountability process.

Warning watermark for IP protection. "TERA" code can also be used as a warning watermark that is only visible for 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, but for camera devices, the logo will appear instead. This can play a warning role for IP protection in 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*. It is mainly based on the inspiration of human vision system's property, 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. Besides, such methods can be broadly used in many different applications such as 2D image code, leak tracking watermark and warning watermark. 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. So in the future, we will 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 Transactions on Multimedia*, vol. 21, no. 10, pp. 2447–2460, 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 Transactions on Multimedia*, pp. 1–1, 2020.

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

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

[5] L. Zhang, C. Bo, J. Hou, X.-Y. Li, Y. Wang, K. Liu, and Y. Liu, "Kaleido: You can watch it but cannot record it," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, 2015, pp. 372–385.

[6] X. Shu and X. Wu, "Frame untangling for unobtrusive display-camera visible light communication," in *Proceedings of the 24th ACM international conference on 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 *2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 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 *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*. ACM, 2015, pp. 181–195.

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

[10] M. Izz, Z. Li, H. Liu, Y. Chen, and F. Li, "Uber-in-light: Unobtrusive visible light communication leveraging complementary color channel," in *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, 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 Transactions on Multimedia*, vol. 17, no. 9, pp. 1515–1524, 2015.

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

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

[14] M. Xu, H. Su, Y. Li, X. Li, J. Liao, J. Niu, P. Lv, and B. Zhou, "Stylized aesthetic qr code," *IEEE Transactions on Multimedia*, vol. 21, no. 8, pp. 1960–1970, 2019.

[15] C. Chen, W. Huang, B. Zhou, C. Liu, and W. H. Mow, "Picode: A new picture-embedding 2d barcode," *ieee transactions on image processing*, vol. 25, no. 8, pp. 3444–3458, 2016.

[16] C. Chen, B. Zhou, and W. H. Mow, "Ra code: A robust and aesthetic code for resolution-constrained applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 11, pp. 3300–3312, 2017.

[17] C. Chen, W. Huang, L. Zhang, and W. H. Mow, "Robust and unobtrusive display-to-camera communications via blue channel embedding," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 156–169, 2018.

[18] J. Zhu, R. Kaplan, J. Johnson, and L. Fei-Fei, "Hidden: Hiding data with deep networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 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 and Applications*, vol. 76, no. 15, pp. 16 063–16 084, 2017.

[20] T. Nakamura, A. Katayama, M. Yamamuro, and N. Sonehara, "Fast watermark detection scheme for camera-equipped cellular phone," in *Proceedings of the 3rd international conference on Mobile and 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 *2018 10th International Conference on Cyber Conflict (CyCon)*. IEEE, 2018, pp. 391–408.

[22] H. Fang, W. Zhang, H. Zhou, H. Cui, and N. Yu, "Screen-shooting resilient watermarking," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 6, pp. 1403–1418, 2018.

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

[24] V. Nguyen, Y. Tang, A. Ashok, M. Gruteser, K. Dana, W. Hu, E. Wengrowski, and N. Mandayam, "High-rate flicker-free screen-camera communication with spatially adaptive embedding," in *IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on Computer Communications*. IEEE, 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 *Proceedings of the 13th Annual International Conference on Mobile Systems, Applications, and Services*. 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 *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1509–1517.

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

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

[29] W. W. Peterson and D. T. Brown, "Cyclic codes for error detection," *Proceedings of the 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 *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[31] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[32] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," pp. 5967–5976, 2017.

[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[34] A. http://sipi.usc.edu/database/ [Online], "The usc-sipi image database," 2020.

[35] The "Big Buck Bunny" Video Database. Accessed: Aug. 2020. [Online]. Available: https://peach.blender.org/download/.

[36] H. Fang, D. Chen, Q. Huang, J. Zhang, Z. Ma, W. Zhang, and N. Yu, "Deep template-based watermarking," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2020.