Scalable Compression of Stream Cipher Encrypted Images Through Context-Adaptive Sampling

Jiantao Zhou, *Member, IEEE*, Oscar C. Au, *Fellow, IEEE*, Guangtao Zhai, *Member, IEEE*, Yuan Yan Tang, *Fellow, IEEE*, and Xianming Liu, *Member, IEEE*

***Abstract*— This paper proposes a novel scalable compression method for stream cipher encrypted images, where stream cipher is used in the standard format. The bit stream in the base layer is produced by coding a series of nonoverlapping patches of the uniformly down-sampled version of the encrypted image. An off-line learning approach can be exploited to model the reconstruction error from pixel samples of the original image patch, based on the intrinsic relationship between the local complexity and the length of the compressed bit stream. This error model leads to a greedy strategy of adaptively selecting pixels to be coded in the enhancement layer. At the decoder side, an iterative, multiscale technique is developed to reconstruct the image from all the available pixel samples. Experimental results demonstrate that the proposed scheme outperforms the state-of-the-arts in terms of both rate-distortion performance and visual quality of the reconstructed images at low and medium rate regions.**

***Index Terms*— Signal processing in encrypted domain, scalable coding, image compression, adaptive sampling.**

I. Introduction

THE standard way of providing security is to encrypt the data using cryptographic algorithms, e.g., AES and RC4 [1]. In many practical scenarios such as cloud computing, the parties who process the encrypted data are often untrusted, and hence have no access to the secret key. This has led to the challenging problem of how to achieve superior efficiency of processing the encrypted data with zero knowledge of the secret.

As one of the most common operations on multimedia data, multimedia compression over encrypted domain has received increasing attention since the last decade. Johnson *et. al* proved that coding with side information principles could be used to compress stream cipher encrypted i.i.d data, without hurting either the compression efficiency or the information-theoretic security [2]. Schonberg *et. al* extended Johnson's work by investigating information sources with memory and unknown statistics [3], [4]. Lazzeretti and Barni suggested methods for lossless compression of encrypted grayscale/color images through applying LDPC codes in various bit-planes and exploiting the inter/intra correlation [5]. Noticing the nearly i.i.d property of prediction error sequence, Kumar and Makur applied the approach of [2] to the prediction error domain and achieved improved lossless compression performance on the encrypted grayscale/color images [6]. Aided by rate-compatible punctured turbo codes, Liu *et. al* developed a lossless, progressive approach to compress stream cipher encrypted grayscale/color images [7]. Recently, Klinc *et al.* extended Johnson's framework to efficiently compress block cipher encrypted data [8].

In addition to lossless compression of encrypted images, lossy compression, which offers higher compression ratios, was also investigated [9]–[18]. Kumar and Makur proposed to use a compressive sensing (CS) mechanism to compress encrypted images resulted from linear encryption [9]. The original image could be estimated from the compressed and encrypted data via a modified basis pursuit algorithm [9]. Another CS-based approach for lossy compression of encrypted images was reported in [10]. In a pioneering work, Zhang *et. al* proposed a scalable lossy coding framework of encrypted images via a multi-resolution construction [11]. Furthermore, Zhang designed an image encryption scheme by performing permutation operations in the pixel domain, and showed that the resulting file can be efficiently compressed by discarding the excessively rough and fine information of coefficients in the transform domain [12]. Recently, Zhang *et. al* suggested a new compression approach for encrypted images through multi-layer decomposition [13]. Kang *et. al* later proposed an interpolation-based

Manuscript received March 20, 2014; revised July 24, 2014; accepted August 21, 2014. Date of publication August 27, 2014; date of current version October 3, 2014. This work was supported in part by the Macau Science and Technology Development Fund under Grant FDCT/009/2013/A1 and Grant FDCT/100/2012/A3, in part by the Research Committee, University of Macau, Macau, China, under Grant SRG023-FST13-ZJT, Grant MRG021/ZJT/2013/FST, Grant MYRG2014-00031-FST, Grant SRG010-FST11-TYY, Grant MYRG205(Y1-L4)-FST11-TYY, and Grant MYRG187 (Y1-L3)-FST11-TYY, in part by the Research Grants Council, Hong Kong, through the General Research Fund, under Grant 610112 and Grant FSGRF12EG01, and in part by the National Science Foundation of China under Grant 61371146, Grant 61402547, Grant 61300110, and Grant 61273244. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Z. Jane Wang.

J. Zhou is with the Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macau, China, and also with the UMacau Zhuhai Research Institute, Zhuhai 519080, China (e-mail: jtzhou@umac.mo).

O. C. Au is with the Department of Electronics and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong (e-mail: eeau@ust.hk).

G. Zhai is with the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai 200030, China (e-mail: zhaiguangtao@sjtu.edu.cn).

Y. Y. Tang is with the Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macau, China (e-mail: yytang@umac.mo).

X. Liu is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: xmliu@jdl.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIFS.2014.2352455

method for recovering compressed encrypted images, where compression is achieved by down-sampling and bit-plane decomposition [14]. We also devised a high-performance compression system for encrypted images, achieving nearly the same compression efficiency as a conventional image codec taking original, unencrypted images as input [15]. More recently, Zhang *et. al* developed a novel system of compressing encrypted images, in which the content owner encrypts the original uncompressed images and also generates some auxiliary information that can be used for data compression and image reconstruction [16]. Extensions to blind compression of encrypted videos were reported in [17] and [18].

In this paper, our primary focus is on the design of a scalable compression system for encrypted images, such that the end terminal can receive and recover the coded images at different resolution and quality levels. Instead of considering dedicated encryption algorithms tailored to the scenario of compressing encrypted images, we restrict ourselves to the conventional stream cipher applied in the standard format. That is, the ciphertext is generated by bitwise XORing the plaintext with the key stream. The resulting compression paradigm for encrypted images is more practically useful because of two reasons: 1) stream cipher used in the standard format (e.g., AES in the CTR mode, in abbreviation, AES-CTR), is still one of the most popular and reliable encryption tools, due to its provable security and high software/hardware implementation efficiency [19]. It may not be easy, or even infeasible, to persuade customers to adopt new encryption algorithms that have not been thoroughly evaluated e.g. [9], or apply the traditional encryption algorithms in a non-standard format e.g. [11]; 2) large number of data have already been encrypted using stream cipher in a standard way. It is practically impossible to decrypt, and encrypt them using another encryption scheme before processing the encrypted file. This work proposes a novel scalable coding scheme for encrypted images. The bit stream in the base layer is produced by coding a series of non-overlapping patches of the uniformly down-sampled version of the encrypted image. An off-line learning approach can be exploited to model the reconstruction error of original image patch based on the intrinsic relationship between the local complexity and the length of the compressed bit stream. Such error model leads to a greedy strategy of adaptively selecting pixels to be coded in the enhancement layer. At the decoder side, an iterative, multi-scale technique is developed to reconstruct the image from all the pixel samples available from the base layer and the enhancement layer. Experimental results demonstrate that the proposed scheme outperforms the state-of-the-arts in terms of both rate-distortion (RD) performance and visual quality of the reconstructed images, for bit rates below 1.2 bit per pixel (bpp), corresponding to low and medium rate regions.

The rest of this paper is organized as follows. Section II presents the proposed scalable coding scheme of stream cipher encrypted images via adaptive sampling. Section III describes the iterative, multi-scale image reconstruction algorithm for decoding the images from pixel samples. In Section IV, we give the extensive experimental results to validate the performance of our proposed scheme. We conclude in Section V.

II. Scalable Coding of Encrypted Images via Adaptive Sampling

Consider the application scenario illustrated in Fig. 1, where an image content owner Alice wants to securely and efficiently transmit an image $\mathbf{f}$ to a recipient Bob, via an untrusted channel provider Charlie. For simplicity, we assume that $\mathbf{f}$ is of square size. To protect the security and privacy of the image data, Alice directly encrypts it by using a traditional encryption algorithm. However, Alice has no motivation to compress the data, and thus, is not willing to use her limited computational resources to run a compression engine, especially when she uses a resource-deprived mobile device. In contrast, the channel provider Charlie is always interested in compressing all the network traffic in a flexible and scalable way, such that the network utilization is maximized. Also, delivering the encrypted image in a scalable manner enables the end terminal to receive and decode the image at different resolution and quality levels. Note that the compression task of Charlie has to be conducted over the encrypted domain, as he has no access to the secret key.

The current work focuses on designing a scalable compression scheme of stream cipher encrypted images, where encryption is carried out by applying a stream cipher in the standard format. More specifically, the encrypted image is obtained by

$$[[\mathbf{f}]] = \mathbf{f} \oplus \mathbf{K} \tag{1}$$

where $\mathbf{f}$ and $[[\mathbf{f}]]$ denote the original and the encrypted images of size $N \times N$, respectively, and $\mathbf{K}$ is the secret key stream. Here, $\oplus$ represents the bitwise XOR operator, and all images are assumed to be 8-bits. We also assume that the encryption of any bit is independent with that of the others. Namely, we are considering a stream cipher like AES-CTR, rather than AES in the OFB mode [1]. Clearly, the spatial relationship of pixels and the bit plane structure keep intact after encryption. Throughout the paper, we use $[[\mathbf{x}]]$ to represent the encrypted version of $\mathbf{x}$.

To efficiently and flexibly compress the encrypted image $[[\mathbf{f}]]$, we develop a novel scalable image coding framework operated over the encrypted domain. In Fig. 2, we show the schematic diagram of the proposed compression scheme, consisting of a base layer and an enhancement layer. Here, the original image is given just for better illustration purpose.

A. Base Layer Encoding

To generate the bit stream in the base layer, we first down-sample $[[\mathbf{f}]]$ into $[[\mathbf{f}_{\downarrow}]]$. Here, we stick to conventional square pixel grid by uniform spatial down sampling of $[[\mathbf{f}]]$. Out of practical considerations, we make a more compact representation of $[[\mathbf{f}]]$ by decimating every *four* rows and every *four* columns, namely, the resulting $[[\mathbf{f}_{\downarrow}]]$ is of size $N/4 \times N/4$. This simple down-sampling strategy can be readily implemented in the encrypted domain, as the spatial relationship among pixels keeps unchanged after encryption. Further, such down-sampling offers an important operational advantage: $[[\mathbf{f}_{\downarrow}]]$ still remains a uniform rectilinear grid of pixels,

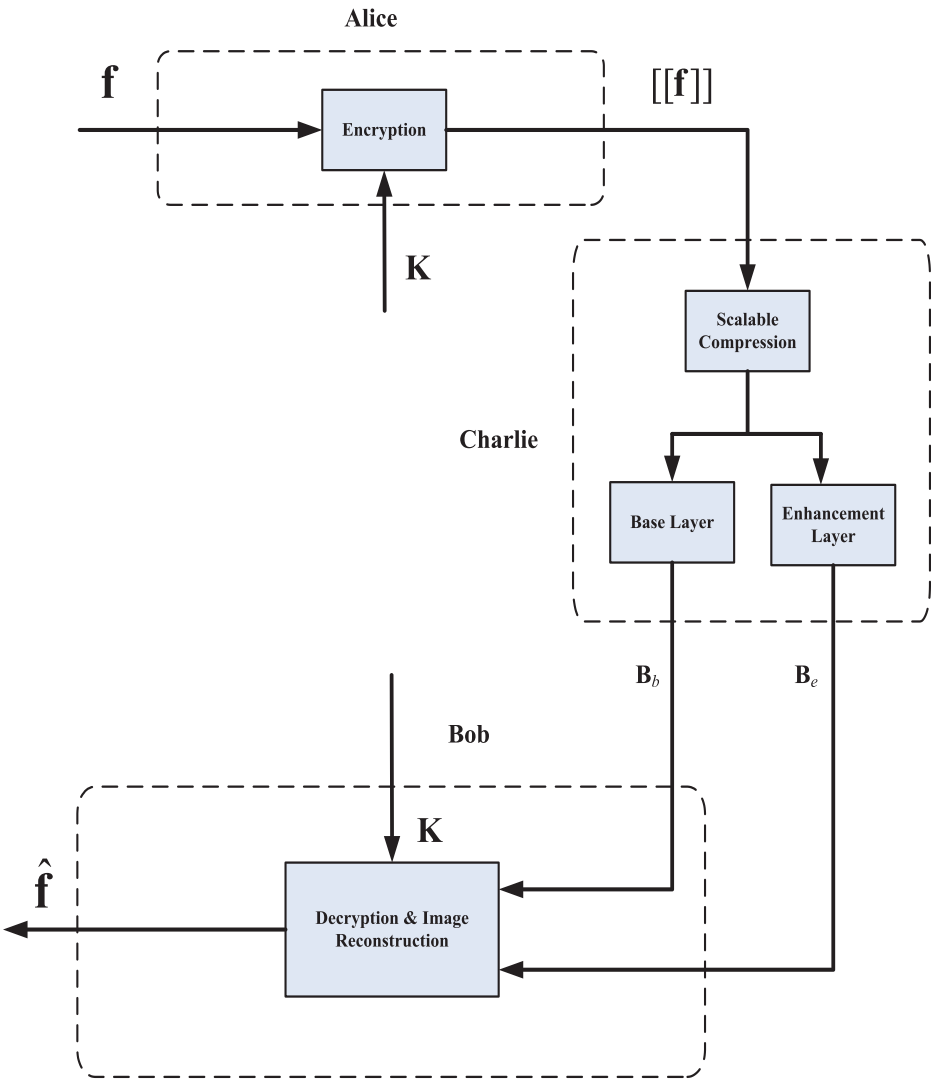


Fig. 1. The scenario of scalable coding of encrypted images.

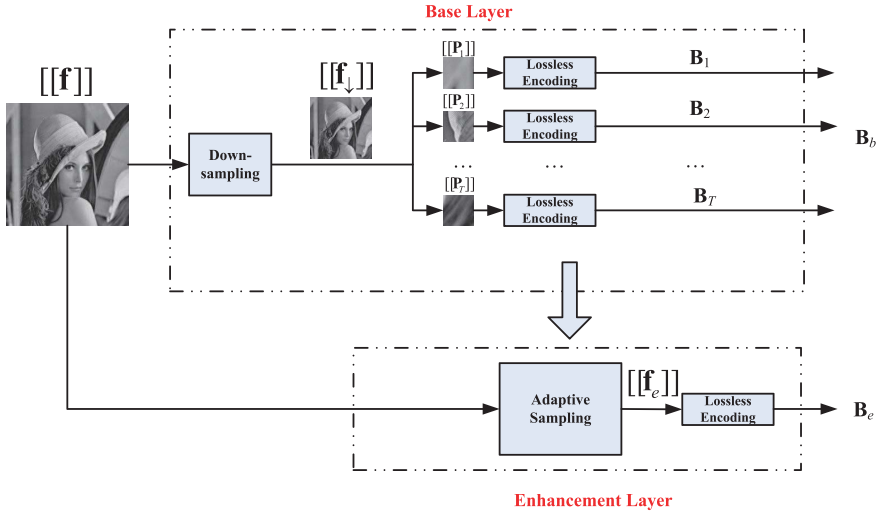


Fig. 2. The proposed scalable image coding scheme.

making it readily compressible by *any* existing techniques for compressing encrypted images.

Instead of coding $[[\mathbf{f}_\downarrow]]$ as a whole, we propose to partition it into a series of non-overlapping patches $[[\mathbf{P}_t]]$ of size $M \times M$. Each $[[\mathbf{P}_t]]$ is then treated as a subimage, and coded individually into a binary bit stream $\mathbf{B}_t$, by applying the lossless compression method of [7]. As the coding of each $[[\mathbf{P}_t]]$ is independent with that of the other patches, the

encoding operations in the base layer can be readily made parallel, achieving high throughput. The encoding complexity in the base layer is around $1/16$ of that to compress the whole encrypted image $[[\mathbf{f}]]$ using the method of [7]. The final bit stream in the base layer $\mathbf{B}_b$ is formed by concatenating all $\mathbf{B}_t$'s, i.e.

$$\mathbf{B}_b = \mathbf{B}_1\mathbf{B}_2\cdots\mathbf{B}_T \tag{2}$$

where

$$T = \frac{N^2}{16M^2} \tag{3}$$

As the lossless compression approach of [7] is based on LDPC codes, a feedback channel is assumed to be available between the compressor and the decoder. The patch-by-patch encoding in the base layer brings another advantage: the length of each $\mathbf{B}_t$, denoted by h_t, i.e., $h_t \triangleq |\mathbf{B}_t|$, is freely available as a byproduct. As also mentioned in [15], the length of the compressed file reflects the complexity of the original signal. Namely, h_t, called *local complexity indicator* (**LCI**), could provide critical information concerning the local complexity of $\mathbf{P}_t$. Smooth patches that are more compressible would produce smaller values of h_t, while those high-activity regions such as textures, edges, etc., typically generate larger h_t. Such valuable information accessible in the encrypted domain without extra cost will be shown to be useful in guiding the encoding of the enhancement layer through a context-adaptive mechanism. The way of obtaining h_t's can also be regarded as a process of on-line learning the image local characteristics over the encrypted domain. In contrast, in the existing schemes [11], [14], the base layer coding and the enhancement layer coding are independently carried out, wasting the valuable information gained in the base layer.

Before presenting the details of the enhancement layer coding, let us investigate the relationship between the patch reconstruction error from pixel samples and the local complexity. This would enable us to develop an appropriate strategy of selecting pixels to be coded in the enhancement layer.

B. Context Modeling of Patch Reconstruction Error

Let $\mathbf{P}$ be a generic $M \times M$ sized patch in $\mathbf{f}_{\downarrow}$, and $\mathbf{Q}$ be the associated $4M \times 4M$ sized patch in $\mathbf{f}$. In other words, $\mathbf{P}$ is obtained by down-sampling $\mathbf{Q}$ by a factor of 4. Denote $\mathbf{R}$ as the $4M \times 4M$ version of $\mathbf{P}$ by duplicating the pixels of $\mathbf{P}$ in their original grids, while leaving the remaining locations empty. The relationship among $\mathbf{P}$, $\mathbf{Q}$, and $\mathbf{R}$ can be illustrated in Fig. 3, where solid dots denote the available pixels, while the hollow ones denote the vacant locations.

Let $\hat{\mathbf{Q}}(n)$ be the reconstructed image patch from the pixel samples available in $\mathbf{R}$ and additional $n \times s$ samples randomly selected from $\mathbf{Q}$ corresponding to the empty locations of $\mathbf{R}$, where $s = 128$ is an empirically set, constant step size. The discussion on the image reconstruction approach from pixel samples is deferred to Section III. In the context of our proposed scalable image coding scheme, the samples in $\mathbf{R}$ are provided by the base layer, while the additional samples are made available by the enhancement layer.

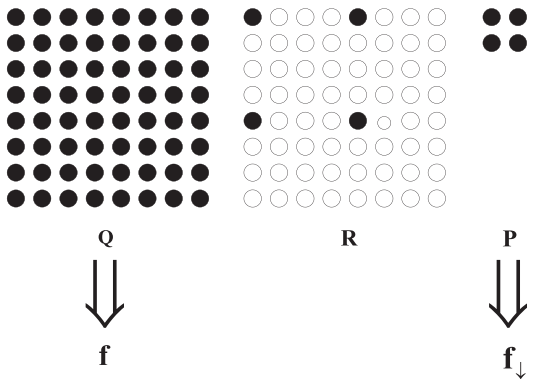


Fig. 3. The relationship among $\mathbf{P}$, $\mathbf{Q}$, and $\mathbf{R}$, for the case that $M = 2$.

From the perspective of sampling theory, additional samples allocated in the smooth regions help less in reducing the reconstruction errors, compared with those in complex regions. This is because, in smooth regions, the existing samples available through base layer might already be sufficient to guarantee faithful reconstructions, making the additional samples less useful. In contrast, high-activity regions contain rich amount of high-frequency components, which in turn require more pixel samples to achieve good reconstruction performance. Therefore, when generating the bit stream in the enhancement layer, it is advantageous to code more samples in those complex regions, so as to achieve better overall RD performance.

Given the quite limited information available in the encrypted domain, it is challenging to estimate the local complexity, as the traditional measures, e.g., local variance or local entropy will fail to provide useful information of the original image. Thanks to the patch-by-patch encoding strategy of the base layer, we can readily obtain the length of the compressed bit stream associated with each encrypted block, i.e., h_t, for $1 \le t \le T$, which could be used as a reliable estimate of the local complexity. To develop an appropriate strategy for selecting pixels to be coded in the enhancement layer, we need to have a deeper understanding about the relationship between the reconstruction error and the local complexity. In the following, we construct a context model for the patch reconstruction error based on the intrinsic relationship between h and the local complexity. This error model can be shown to be able to derive a greedy strategy for pixel sample selection in the enhancement layer. To this end, we first define the reconstruction error difference (**RED**) by

$$\begin{aligned} d(n) &\triangleq e(n-1) - e(n) \\ &= ||\hat{\mathbf{Q}}(n-1)) - \mathbf{Q}||_2^2 - ||\hat{\mathbf{Q}}(n) - \mathbf{Q}||_2^2 \end{aligned} \tag{4}$$

where $e(n) = ||\hat{\mathbf{Q}}(n) - \mathbf{Q}||_2^2$ denotes the patch reconstruction error when $n \times s$ *additional* pixel samples are given, besides those provided by the base layer. This definition also implies

$$e(n) = e(0) - \sum_{i=1}^{n} d(i) \tag{5}$$

The value of $d(n)$, which is clearly non-negative, reflects the effectiveness of increasing the number of additional samples from $(n-1)\times s$ to $n \times s$. Larger (smaller) value of $d(n)$ means

that the additional s samples can more (less) effectively reduce the reconstruction error.

As explained above and also verified experimentally, $d(n)$ strongly correlates to the local smoothness of the patch. In those regions with smooth signal waveforms, $d(n)$ tends to be small, while for those high-activity regions, it becomes large. We hence introduce the conditional probability $p(d(n)|h)$ as a statistical context model of $d(n)$. To prevent the problem of context dilution caused by too many modeling parameters, we propose to estimate the conditional expectation $\mathbb{E}(d(n)|h)$, rather than $p(d(n)|h)$ itself. Still, the space spanned by h is too large in practice. To further narrow the context space, we apply a uniform scalar quantizer with step size τ, namely,

$$\tilde{h} = \lfloor h/\tau \rfloor \tag{6}$$

where τ depends on the block size M, and is empirically set to be $\tau = 0.1M^2$. In our implementation, $M = 16$ is used.

To estimate $\mathbb{E}(d(n)|\tilde{h})$, we employ an off-line learning approach with the assistance of $U = 100$ training images of size 512×512. More specifically, for each training image $\mathbf{f}_u$, where $1 \le u \le U$, the $4M \times 4M$ sized patches $Q_{u,t}$'s can be readily extracted, which can subsequently produce the corresponding $\mathbf{R}_{u,t}$ and $\mathbf{P}_{u,t}$. Here, u is the index for the training images, and t is the patch index. For each $\mathbf{R}_{u,t}$, we increment $n \times s$ pixel samples in the vacant locations, and then get the reconstructed patch $\hat{Q}_{u,t}(n)$ through applying the technique to be presented in Section III. The corresponding $d_{u,t}(n)$ can be similarly calculated via (4). Eventually, each patch $\mathbf{R}_{u,t}$ is associated with a **RED** vector $\mathbf{d}_{u,t} = (d_{u,t}(1), d_{u,t}(2), \cdots, d_{u,t}(S))'$, where

$$S = \frac{(4M)^2 - M^2}{s} = \frac{15M^2}{s} \tag{7}$$

Meanwhile, the value of $\tilde{h}_{u,t}$ can be obtained by compressing $[[\mathbf{P}_{u,t}]]$ with the method of [7] and measuring the length of the compressed bit stream. Then, the conditional expectation $\mathbb{E}(d(n)|\tilde{h})$ can be computed by

$$\mathbb{E}(d(n)|\tilde{h} = h_o) = \frac{\sum_{(u,t)\in\Omega} d_{u,t}(n)}{|\Omega|} \tag{8}$$

where

$$\Omega = \{(u,t)|\tilde{h}_{u,t} = h_o\} \tag{9}$$

It should be noted that the above training process is completely conducted off-line, and hence the encoding complexity will not be materially increased, as all the pre-calculated conditional expectations can be implemented in a look-up table indexed by n and $\tilde{h}$.

C. Enhancement Layer Coding via Greedy Strategy

With the above context model of patch reconstruction error, we in this subsection discuss the strategy of selecting pixel samples to be coded in the enhancement layer. Let $\tilde{h} \triangleq (\tilde{h}_1, \tilde{h}_2, \ldots, \tilde{h}_T)'$, where $\tilde{h}_i$ denotes the **LCI** value of the ith $4M \times 4M$ sized patch of the encrypted image and T is given in (3). We also define the sample selection vector $\mathbf{n} \triangleq (n_1, n_2, \ldots, n_T)'$, where $n_i \times s$ is the number of samples selected from the ith $4M \times 4M$ patch. Let us assume that totally there are $F \times s$ samples to be coded in the enhancement layer, where F is a non-negative integer controlling the overall coding rate.

A natural criterion for selecting the pixel samples to be coded in the enhancement layer is to minimize the total reconstruction error. Therefore, the determination of all n_i's can be mathematically formulated as the following optimization problem

$$\begin{aligned} &\min_{\mathbf{n}} \sum_{i=1}^{T} e(n_i) \\ &\text{subject to:} \sum_{i=1}^{T} n_i = F \\ &\qquad \forall n_i \in \mathcal{Z} \\ &\qquad \forall n_i \ge 0 \end{aligned} \tag{10}$$

where $e(n_i)$ given in (5) denotes the reconstruction distortion when $n_i \times s$ additional pixel samples are provided. Noticing the fact that all n_i's are integers, the above optimization problem essentially belongs to the category of integer programming problems [20], which are unfortunately NP-hard. In other words, no polynomial-time algorithm exists for solving (10). To reduce the computational burden while still achieving reasonably good performance, we develop a greedy algorithm consisting of F stages. In each stage, s samples will be allocated, aiming at minimizing the induced reconstruction distortion at that particular stage. As will become clear shortly, the complexity of our proposed greedy heuristic is of order $O(F \times T)$, where T given in (3) is the number of $4M \times 4M$ sized patch. More specifically, the greedy algorithm for the pixel sample selection in the enhancement layer is presented as follows:

Step 1: Initialize a vector $\mathbf{E} \triangleq (E_1, E_2, \ldots, E_T)' = (\mathbb{E}(d(n_1)|\tilde{h}_1)), \mathbb{E}(d(n_2)|\tilde{h}_2)), \ldots, \mathbb{E}(d(n_T)|\tilde{h}_T))'$ by setting all n_i's to be 1, where the conditional expectations given in (8) are obtained from the off-line training process.

Step 2: Find the index q corresponding to the maximum element of $\mathbf{E}$, i.e.,

$$q = \underset{1\le i\le T}{\operatorname{argmax}} E_i \tag{11}$$

When multiple q's exist, we simply choose the minimum one. Recall the definition in (4), allocating s samples to the qth patch would result in maximum reconstruction error reduction, which is a locally optimal strategy.

Step 3: Randomly choose s encrypted pixels from $[[\mathbf{Q}]]_q$ that have never been selected, according to a *public* seed.

Step 4: Update

$$\begin{aligned} E_q &= \mathbb{E}(d(n_q + 1)|\tilde{h}_q)) \\ n_q &= n_q + 1 \end{aligned} \tag{12}$$

Step 5: Repeat Steps 2-4 F times until all $F \times s$ samples are selected.

Step 6: Reshape the selected pixel samples into a 2-D image $[[\mathbf{f}_e]]$, and encode it into a binary bit stream $\mathbf{B}_e$ using the method of [7].

As the vector $\tilde{h} = (\tilde{h}_1, \tilde{h}_2, \ldots, \tilde{h}_T)'$, and the initial $\mathbf{E}$ together with its updating rule are completely transparent to the decoder, the same operations can be performed by the decoder to determine the locations of the samples upon decoding the bit steam received from the enhancement layer. As we need to perform T times of comparison at each stage, and there are totally F stages, the complexity of the above greedy algorithm is of order $O(F \times T)$. In terms of the overall complexity of the enhancement layer, we observe that it is primarily dominated by the actual coding of those selected pixel samples, rather than the above greedy selection process. Hence, the complexity of the enhancement layer almost linearly increases with the number of pixel samples coded.

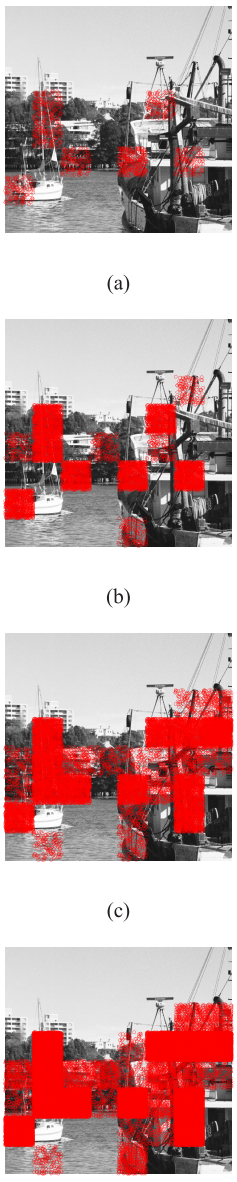
Fig. 4 demonstrates the results of sample selection when $F = 10$, $F = 50$, $F = 100$, and $F = 200$, respectively, where the original image is given just for better illustration purpose. It can be observed that most of the samples are located at the complex foreground, while the sky in the background with homogenous characteristics has slim chance of getting samples even when $F = 200$. This validates the effectiveness of our proposed sample selection algorithm operated over encrypted domain.

III. Image Reconstruction

Upon receiving the bit stream $\mathbf{B}_b$ from the base layer, the decoding algorithm of [7] can be applied to get $[[\mathbf{f}_\downarrow]]$. As the uniform down-sampling rule is public and the encryption of each bit is independent with that of the others, the encoder and the decoder can be straightforwardly synchronized. Hence, $[[\mathbf{f}_\downarrow]]$ can be appropriately decrypted into $\mathbf{f}_\downarrow$ by XORing with the corresponding key stream. In the case that the enhancement layer is not available, $\mathbf{f}_\downarrow$ can be directly up-converted to $\hat{\mathbf{f}}_b$, which is of the same size as the original image $\mathbf{f}$, using the method to be presented below. In some application scenarios, $\hat{\mathbf{f}}_b$ can be utilized as the preview version for customers.

When the bit stream of the enhancement layer arrives, the decoding algorithm of [7] can be similarly employed to obtain $[[\mathbf{f}_e]]$, where the base layer reconstruction $\hat{\mathbf{f}}_b$ serves as the side information. As the encoder and the decoder are perfectly synchronized, the key stream can be appropriately extracted to decrypt $[[\mathbf{f}_e]]$ into $\mathbf{f}_e$.

With $\mathbf{f}_\downarrow$ and $\mathbf{f}_e$, the decoder aims to go beyond and collaboratively re-estimate the original image. To this end, we propose an iterative, multi-scale technique to reconstruct the original image, as depicted in Fig. 5. Let $\mathbf{f}_0$ be the image containing all the available samples from $\mathbf{f}_\downarrow$ and $\mathbf{f}_e$, while the vacant locations are padded with 0's. In our proposed multi-scale framework, $\mathbf{f}_0$ is successively down-sampled twice by a factor of 2 respectively to form a pyramid: $\mathbf{f}_1$ and $\mathbf{f}_2$, by duplicating the corresponding pixels from the existing high-resolution images. The image reconstruction starts from the lowest level 2. We first up-convert $\mathbf{f}_2$ to a higher level $\tilde{\mathbf{f}}_1$ via a parametric-model based interpolation method. Due to the excellent interpolation performance and modest computational



(a) (b) (c) (d)

Fig. 4. Sample selection for different values of F. (a) $F = 10$. (b) $F = 50$. (c) $F = 100$. (d) $F = 200$.

complexity, the soft-decision adaptive interpolation (SAI) [21] based on 2D-piecewise autoregressive (2D-PAR) model is adopted. In our implementation, an order-8 linear AR model is used, and the soft-decision estimation is conducted on a block-by-block basis, where block size is 5×5. With $\tilde{\mathbf{f}}_1$ and $\mathbf{f}_1$, we then attempt to get a refined version by estimating the missing pixels (those padded with zeros) of $\mathbf{f}_1$. To this end, we employ a non-local mean (NLM)-based approach, in which all the weighting vectors are computed from the up-converted $\tilde{\mathbf{f}}_1$. In other words, $\tilde{\mathbf{f}}_1$ now serves as the prior knowledge for the estimation task. Specifically, the ith missing pixel in $\mathbf{f}_1$ is estimated by

$$\hat{\mathbf{f}}_1(i) = \sum_{j \in \mathcal{W}_i} w(i, j)\mathbf{f}_1(j) \tag{13}$$

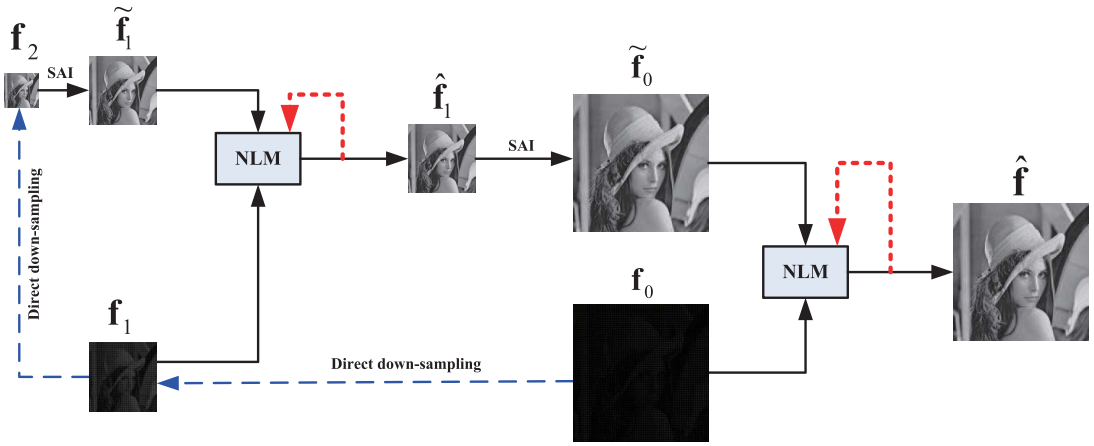


Fig. 5. Schematic diagram of image reconstruction.

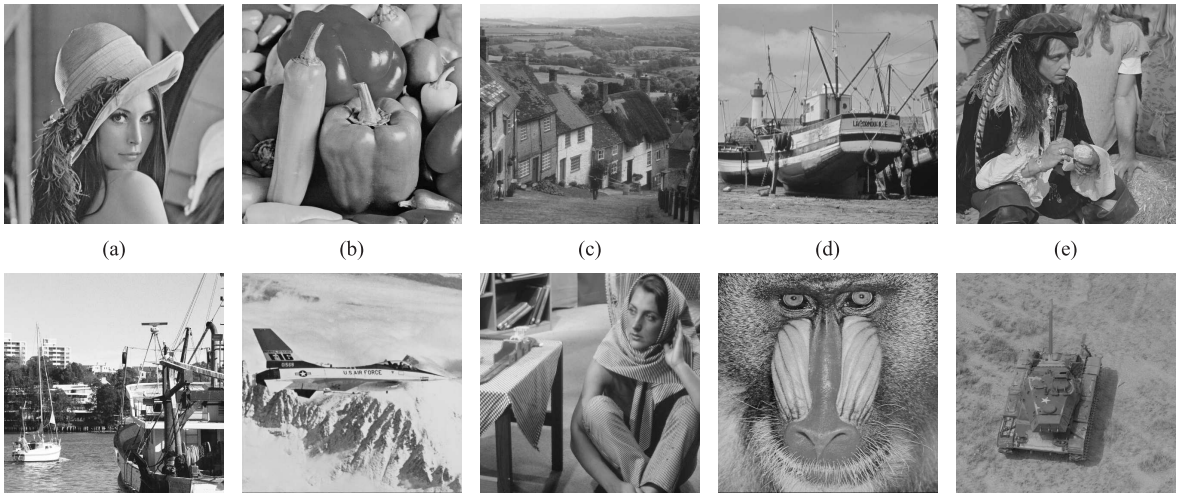


Fig. 6. 10 test images. (a) Lena. (b) Peppers. (c) Goldhill. (d) Boat. (e) Man. (f) Harbor. (g) Airplane. (h) Barbara. (i) Baboon. (j) Tank.

where $\mathcal{W}_i$ denotes a local window centered at pixel i. In the implementation, we set the window size to be 17×17. The family of weights $w(i, j)$'s depend on the similarity between the pixels i and j, and are explicitly given by

$$w(i, j) = \frac{1}{Z(i)} e^{\frac{-||\tilde{\mathbf{f}}_1(\mathcal{N}_i) - \tilde{\mathbf{f}}_1(\mathcal{N}_j)||^2_{2,\alpha}}{\delta^2}} \tag{14}$$

where the gray level vectors $\tilde{\mathbf{f}}_1(\mathcal{N}_i)$ and $\tilde{\mathbf{f}}_1(\mathcal{N}_j)$ represent two square neighborhoods of fixed size centered at pixels i and j, respectively. Here, $||\cdot||^2_{2,\alpha}$ denotes the weighted Euclidean distance, with $\alpha > 0$ being the standard deviation of the Gaussian kernel, and $Z(i)$ is a normalizing constant

$$Z(i) = \sum_j e^{\frac{-||\tilde{\mathbf{f}}_1(\mathcal{N}_i) - \tilde{\mathbf{f}}_1(\mathcal{N}_j)||^2_{2,\alpha}}{\delta^2}} \tag{15}$$

where δ controls the degree of filtering.

To further enhance the performance, we iterate the above procedure by replacing $\tilde{\mathbf{f}}_1$ with the newly refined $\hat{\mathbf{f}}_1$. Heuristically, we find that 2-3 iterations are sufficient to produce satisfactory results. The resulting $\hat{\mathbf{f}}_1$ is then up-converted to $\tilde{\mathbf{f}}_0$ via SAI. At level 0, similar iterative NLM-based refinement process can be carried out to get the final estimate $\hat{\mathbf{f}}$.

TABLE I
BASE LAYER PERFORMANCE

Image	Rate (bpp)	PSNR (dB)
Lena	0.4156	27.90
Peppers	0.4272	26.02
Goldhill	0.4298	25.80
Boat	0.4204	24.62
Man	0.4377	25.29
Harbor	0.4398	18.73
Airplane	0.4116	24.05
Barbara	0.4459	21.69
Baboon	0.4640	19.22
Tank	0.4152	28.35

In terms of the complexity of the image reconstruction process, it is higher than the counterparts of [11] and [14]. However, as to be demonstrated in the next Section, our reconstruction approach achieves much improved PSNR performance and visual quality at low and medium rate regions.

IV. EXPERIMENTAL RESULTS

In this section, we evaluate the performance of our proposed scalable compression scheme for stream cipher encrypted

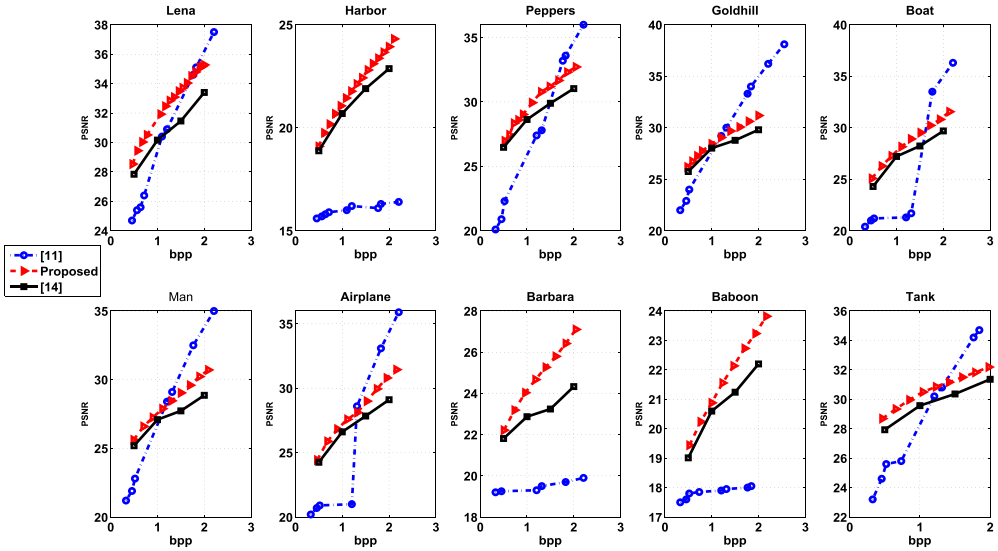
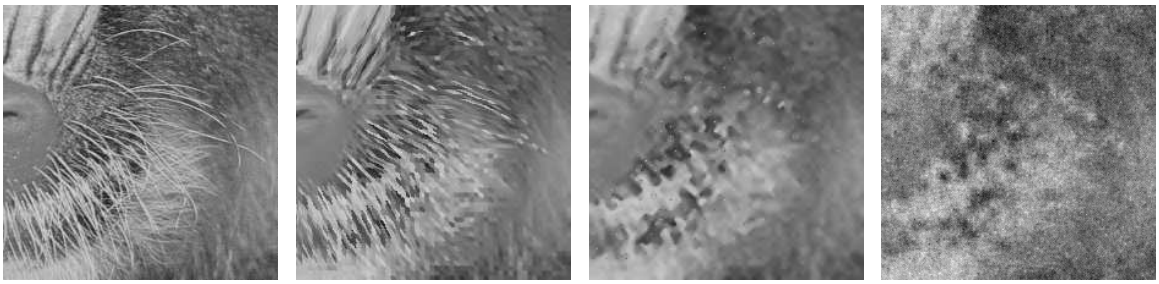
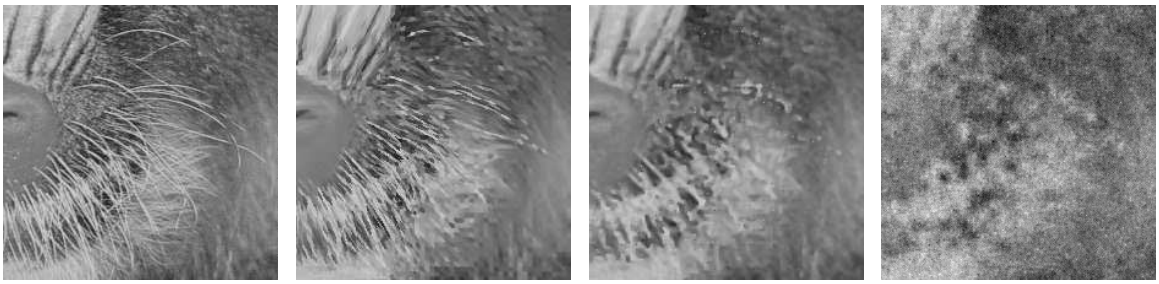


Fig. 7. Comparison of RD performance for 10 test images.



(a) Original (b) Ours (c) [14] (d) [11]

Fig. 8. Visual comparison of image baboon. (a) Original; (b) Proposed (PSNR 19.22 dB, rate 0.4640 bpp); (c) [14] (PSNR 19.03 dB, rate 0.50 bpp); (d) [11] (PSNR 17.50 dB, rate 0.4556 bpp).



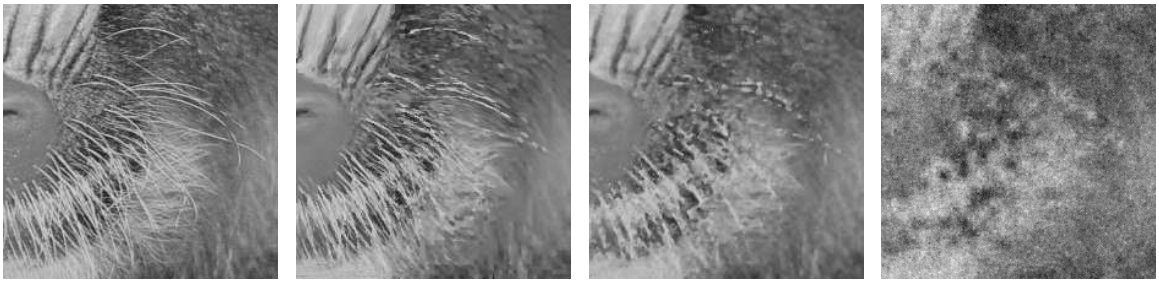
(a) Original (b) Ours (c) [14] (d) [11]

Fig. 9. Visual comparison of image baboon. (a) Original; (b) Proposed (PSNR 20.88 dB, rate 0.9995 bpp); (c) [14] (PSNR 20.6 dB, rate 1.0 bpp); (d) [11] (PSNR 17.55 dB, rate 1.2056 bpp).

images. The test set consists of 10 images of size 512×512, as shown in Fig. 6. For the sake of fairness, the test set and the training set used for obtaining the conditional expectations $\mathbb{E}(d(n)|\tilde{h})$ do not have any overlap. For each image, the encrypted version is obtained by applying AES-CTR in a pixel-by-pixel manner.

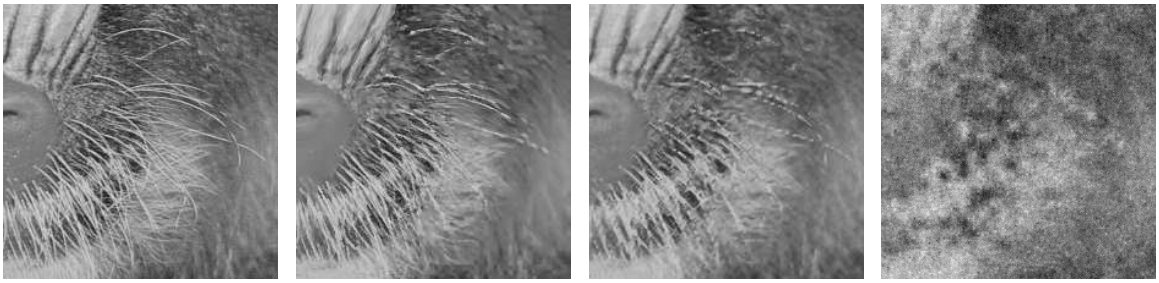
In Table I, we present the compression performance of the base layer. It can be seen that all the base rates are below 0.5 bpp, due to the block-by-block compression of the down-sampled image of factor 4.

To further demonstrate the superior performance of our proposed method, we compare it with the state-of-the-art schemes [11] and [14] in terms of ℓ_2 distortion. In Fig. 7, we show the RD curves for the 10 test images. It can be seen that for rates below 1.2 bpp, corresponding to low and medium rate regions, the proposed method offers better RD performance over the method of [11]. Especially for low rates, the gain in PSNR is quite significant. For instance, when the rate is around 0.5 bpp, the gain can be over 4 dB. However, when the bit rates increase further, our proposed method becomes inferior in PSNR performance. In addition, we observe that, for several test images such as Harbor, Barbara, and Baboon, the compression algorithm of [11] encounters the performance saturation problem, i.e., increasing the bit rates does not notably improve the PSNR values. Fortunately, our method can successfully overcome such problem.



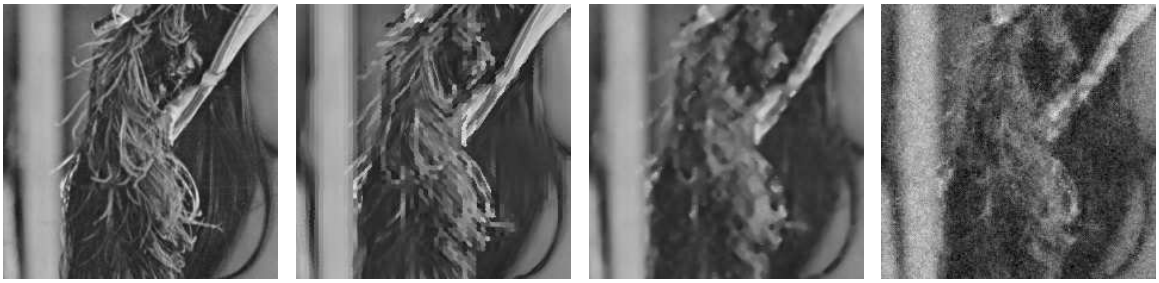
(a) Original (b) Ours (c) [14] (d) [11]

Fig. 10. Visual comparison of image baboon. (a) Original; (b) Proposed (PSNR 22.14 dB, rate 1.4674 bpp); (c) [14] (PSNR 21.24 dB, rate 1.5 bpp); (d) [11] (PSNR 17.9 dB, rate 1.7654 bpp).



(a) Original (b) Ours (c) [14] (d) [11]

Fig. 11. Visual comparison of image baboon. (a) Original; (b) Proposed (PSNR 23.26 dB, rate 1.9348 bpp); (c) [14] (PSNR 22.2 dB, rate 2.0 bpp); (d) [11] (PSNR 18.0 dB, rate 2.2079 bpp).



(a) Original (b) Ours (c) [14] (d) [11]

Fig. 12. Visual comparison of image Lena. (a) Original; (b) Proposed (PSNR 27.90 dB, rate 0.4156 bpp); (c) [14] (PSNR 27.84 dB, rate 0.5 bpp); (d) [11] (PSNR 23.10 dB, rate 0.4556 bpp).

(a) Original (b) Ours (c) [14] (d) [11]

Fig. 13. Visual comparison of image Lena. (a) Original; (b) Proposed (PSNR 31.65 dB, rate 1.0247 bpp); (c) [14] (PSNR 30.16 dB, rate 1.0 bpp); (d) [11] (PSNR 30.90 dB, rate 1.2056 bpp).

When comparing with the method of [14], our scheme consistently outperforms it in a large rate region from 0.5 bpp to 2 bpp. The gain in PSNR can be up to 2.77 dB, which is achieved by the image Barbara when the rate is 2 bpp. Also, the gain tends to increase as the bit rates get larger.

The RD behavior of our proposed compression scheme could be roughly explained as follows. When the bit budget is tight (corresponding to low and medium rate regions), appropriate pixel selection strategy in the enhancement layer could lead to much more significant distortion reduction, compared with a random or uniform selection approach. This implies that the advantages of our scheme become more apparent for low and medium rates, outperforming the method of [11]. Nevertheless, when the rate becomes high, the way

(a) Original (b) Ours (c) [14] (d) [11]

Fig. 14. Visual comparison of image Lena. (a) Original; (b) Proposed (PSNR 33.63 dB, rate 1.5030 bpp); (c) [14] (PSNR 31.46 dB, rate 1.5 bpp); (d) [11] (PSNR 34.6 dB, rate 1.7654 bpp).

(a) Original (b) Ours (c) [14] (d) [11]

Fig. 15. Visual comparison of image Lena. (a) Original; (b) Proposed (PSNR 35.42 dB, rate 2.0008 bpp); (c) [14] (PSNR 33.40 dB, rate 2.0 bpp); (d) [11] (PSNR 37.5 dB, rate 2.2079 bpp).

of selecting pixels in the enhancement layer becomes less crucial. This is because large number of pixel samples will be selected for high rates, and hence, different selection strategies only differ slightly. In the extreme case where the coding rate is even higher than the lossless rate, all the vacant pixels will be selected in the enhancement layer. Then, any pixel selection approaches will be exactly the same, and our scheme converges to a lossless solution. Therefore, in high rate regions, the RD performance of our method is outperformed by [11].

We also present the visual comparison of the image details for different methods in Figs. 8–15. Compared with the reconstructed images of [14], the proposed method can more effectively preserve shape edges and fine details, even when PSNR values are similar. For instance, as illustrated in Fig. 9, our decoded Baboon image is visually much closer to the original one, though the PSNR gain of our method is only 0.19 dB. Compared with the method of [11], our restored images are visually more pleasing at low and medium rate regions, with better preserved textural details.

V. Conclusions

In this paper, we design a novel scalable image coding scheme for stream cipher encrypted images. The base layer compresses a series of non-overlapping patches of the uniformly down-sampled version of the encrypted image. Based on the free side information offered by base layer coding, the enhancement layer strategically selects additional pixel samples to code. The decoder then applies an iterative, multi-scale technique to reconstruct the image from all the available samples. Experimental results verify the superior coding performance of our proposed scheme, especially at low and medium rate regions.

Acknowledgement

The authors would like to thank the Associate Editor Prof. Z. Jane Wang and four anonymous reviewers for their helpful comments and suggestions.

References

[1] A. J. Menezes, P. C. Van Oorschot, and S. A. Vanstone, *Handbook of Applied Cryptography*. Boca Raton, FL, USA: CRC Press, 1997.

[2] M. Johnson, P. Ishwar, V. Prabhakaran, D. Schonberg, and K. Ramchandran, "On compressing encrypted data," *IEEE Trans. Signal Process.*, vol. 52, no. 10, pp. 2992–3006, Oct. 2004.

[3] D. Schonberg, S. C. Draper, and K. Ramchandran, "On blind compression of encrypted correlated data approaching the source entropy rate," in *Proc. 43rd Annu. Allerton Conf. Commun., Control, Comput.*, 2005, pp. 1–10.

[4] D. Schonberg, S. Draper, and K. Ramchandran, "On compression of encrypted images," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2006, pp. 269–272.

[5] R. Lazzeretti and M. Barni, "Lossless compression of encrypted grey-level and color images," in *Proc. 16th Eur. Signal Process. Conf. (EUSIPCO)*, Lausanne, Switzerland, 2008, pp. 1–5.

[6] A. A. Kumar and A. Makur, "Distributed source coding based encryption and lossless compression of gray scale and color images," in *Proc. 10th Workshop MMSP*, Oct. 2008, pp. 760–764.

[7] W. Liu, W. Zeng, L. Dong, and Q. Yao, "Efficient compression of encrypted grayscale images," *IEEE Trans. Image Process.*, vol. 19, no. 4, pp. 1097–1102, Apr. 2010.

[8] D. Klinc, C. Hazay, A. Jagmohan, H. Krawczyk, and T. Rabin, "On compression of data encrypted with block ciphers," *IEEE Trans. Inf. Theory*, vol. 58, no. 11, pp. 6989–7001, Nov. 2012.

[9] A. A. Kumar and A. Makur, "Lossy compression of encrypted image by compressive sensing technique," in *Proc. IEEE Region 10th Conf.*, Jan. 2009, pp. 1–6.

[10] X. Zhang, Y. Ren, G. Feng, and Z. Qian, "Compressing encrypted image using compressive sensing," in *Proc. 7th IEEE Int. Conf. IIH-MSP*, Oct. 2011, pp. 222–2225.

[11] X. Zhang, G. Feng, Y. Ren, and Z. Qian, "Scalable coding of encrypted images," *IEEE Trans. Image Process.*, vol. 21, no. 6, pp. 3108–3114, Jun. 2012.

[12] X. Zhang, "Lossy compression and iterative reconstruction for encrypted image," *IEEE Trans. Inf. Forensics Security*, vol. 6, no. 1, pp. 53–58, Mar. 2011.
[13] X. Zhang, G. Sun, L. Shen, and C. Qin, "Compression of encrypted images with multi-layer decomposition," *Multimedia Tools Appl.*, vol. 72, no. 1, pp. 489–502, Feb. 2013.
[14] X. Kang, A. Peng, X. Xu, and X. Cao, "Performing scalable lossy compression on pixel encrypted images," *EURASIP J. Image Video Process.*, vol. 2013, no. 32, pp. 1–6, May 2013.
[15] J. Zhou, X. Liu, O. C. Au, and Y. Y. Tang, "Designing an efficient image encryption-then-compression system via prediction error clustering and random permutation," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 1, pp. 39–50, Jan. 2014.
[16] X. Zhang, Y. Ren, L. Shen, Z. Qian, and G. Feng, "Compressing encrypted images with auxiliary information," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1327–1336, Aug. 2014.
[17] D. Schonberg, S. C. Draper, C. Yeo, and K. Ramchandran, "Toward compression of encrypted images and video sequences," *IEEE Trans. Inf. Forensics Security*, vol. 3, no. 4, pp. 749–762, Dec. 2008.
[18] Q. Yao, W. Zeng, and W. Liu, "Multi-resolution based hybrid spatiotemporal compression of encrypted videos," in *Proc. IEEE ICASSP*, Apr. 2009, pp. 725–728.
[19] H. Lipmaa, P. Rogaway, and D. Wagner. *CTR-Mode Encryption*. [Online]. Available: http://csrc.nist.gov/encryption/modes/workshop1/papers/lipmaa-ctr.pdf, accessed Sept. 2000.
[20] A. Schrijver, *Theory of Linear and Integer Programming*. New York, NY, USA: Wiley, 1986.
[21] X. Zhang and X. Wu, "Image interpolation by adaptive 2-D autoregressive modeling and soft-decision estimation," *IEEE Trans. Image Process.*, vol. 17, no. 6, pp. 887–896, Jun. 2008.

Jiantao Zhou (M'11) is currently an Assistant Professor with the Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macau, China. He received the B.Eng. degree in electronic engineering from the Dalian University of Technology, Dalian, China, in 2002, the M.Phil. degree in radio engineering from Southeast University, Nanjing, China, in 2005, and the Ph.D. degree in electronic and computer engineering from the Hong Kong University of Science and Technology, Hong Kong, in 2009. He held various research positions with the University of Illinois at Urbana-Champaign, Champaign, IL, USA, the Hong Kong University of Science and Technology, Hong Kong, and the McMaster University, Hamilton, ON, Canada. His research interests include multimedia security and forensics and high-fidelity image compression. He was a co-author of a paper that received the Best Paper Award in the IEEE Pacific-Rim Conference on Multimedia in 2007.

Oscar C. Au (S'87–M'90–SM'01–F'11) received the B.A.Sc. degree from the University of Toronto, Toronto, ON, Canada, in 1986, and the M.A. and Ph.D. degrees from Princeton University, Princeton, NJ, USA, in 1988 and 1991, respectively, where he held a Post-Doctoral position with for one year. He joined Hong Kong University of Science and Technology (HKUST), Hong Kong, as an Assistant Professor, in 1992, where he is currently a Professor of the Department of Electronic and Computer Engineering, the Director of Multimedia Technology Research Center, and the Director of Computer Engineering.

His main research contributions are in video/image coding and processing, watermarking/light weight encryption, and speech/audio processing. Research topics include fast motion estimation for H.261/3/4/5, MPEG-1/2/4, and AVS, optimal and fast sub-optimal rate control, mode decision, transcoding, denoising, deinterlacing, post-processing, multi-view coding, view interpolation, depth estimation, 3DTV, scalable video coding, distributed video coding, subpixel rendering, JPEG/JPEG2000, HDR imaging, compressive sensing, halftone image data hiding, GPU-processing, and software-hardware codesign. He has authored more than 60 technical journal papers, 350 conference papers, and 70 contributions to international standards. His fast motion estimation algorithms were accepted into the ISO/IEC 14496-7 MPEG-4 international video coding standards and the China AVS-M standards. His light-weight encryption and error resilience algorithms are accepted into the China AVS standard. He was the Chair of Screen Content Coding AdHoc Group in JCTVC for HEVC. He holds more than 20 U.S. patents and is applying for more than 70 on his signal processing techniques. He has performed forensic investigation and stood as an expert witness in Hong Kong courts many times.

Dr. Au is a fellow of HKIE and a BoG Member of APSIPA. He is an Associate Editor of eight journals, including the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I, *Journal of Visual Communication and Image Representation*, *Journal of Signal Processing Systems*, the IEEE *Teacher in Service Program*, *Journal of Micromechanics and Microengineering*, and *Journal of Forensic Identification*. He is the Chair of three technical committees, including the IEEE CAS MSA TC, IEEE SPS MMSP TC, and APSIPA IVM TC. He is a member of six other TCs, including the IEEE CAS VSPC TC, DSP TC, IEEE SPS IVMSP TC, IFS TC, and IEEE ComSoc MMC TC. He served on two steering committees, including the IEEE TRANSACTIONS ON MULTIMEDIA, and INTERNATIONAL CONFERENCE ON MULTIMEDIA AND EXPOSITION. He also served on organizing committee of many conferences, including the 1997 IEEE International Symposium on Circuits and Systems, 2003 International Conference on Acoustics, Speech and Signal Processing, ISO/IEC 71st MPEG, in 2005, and 2010 International Conference on Image Processing. He was a General Chair of several conferences, including the 2007 Pacific-Rim Conference on Multimedia, the 2010 IEEE International Conference on Multimedia and Expo, 2010 Packet Video Workshop, 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, and the 2017 IEEE International Conference on Multimedia and Expo. He was recipient of the five best paper awards, including the 2007 IEEE Workshop on Signal Processing Systems, 2007 Pacific-Rim Conference on Multimedia, 2012 International Workshop on Multimedia Signal Processing, the 2013 IEEE International Conference on Image Processing, and 2013 International Workshop on Multimedia Signal Processing. He was the IEEE Distinguished Lecturer (DL) in 2009 and 2010, Asia-Pacific Signal and Information Processing Association DL in 2013 and 2014, and has been a Keynote Speaker multiple times.

Guangtao Zhai (M'10) received the B.E. and M.E. degrees from Shandong University, Shandong, China, in 2001 and 2004, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2009. He is currently a Research Professor with the Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai, China. From 2006 to 2007, he was a Student Intern with the Institute for Infocomm Research, Singapore, and a Visiting Student with the School of Computer Engineering, Nanyang Technological University, Singapore, from 2007 to 2008, a Visiting Student with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, from 2008 to 2009, where he was a Post-Doctoral Fellow from 2010 to 2012. His research interests include multimedia signal processing and perceptual signal processing.

Yuan Yan Tang (S'88–M'88–SM'96–F'04) is a Chair Professor with the Faculty of Science and Technology with the University of Macau, Macau, China, and a Professor/Adjunct Professor/Honorary Professor at several institutes, including Chongqing University, Chongqing, China, Concordia University, Montréal, QC, Canada, and Hong Kong Baptist University, Hong Kong. His current interests include wavelets, pattern recognition, image processing, and artificial intelligence. He has authored more than 400 academic papers and has authored or coauthored over 25 monographs/books/bookchapters. He is the Founder and Editor-in-Chief of *International Journal on Wavelets, Multiresolution, and Information Processing* and an Associate Editor of several international journals. He is the Founder and Chair of the Pattern Recognition Committee in the IEEE Systems, Man, and Cybernetics Society. He has served as a General Chair, Program Chair, and Committee Member for many international conferences. He is the Founder and General Chair of the series International Conferences on Wavelets Analysis and Pattern Recognition. He is the Founder and Chair of the Macau Branch of International Associate of Pattern Recognition (IAPR), where he is a fellow.

Xianming Liu is currently an Assistant Professor with the Department of Computer Science, Harbin Institute of Technology, Harbin, China. He received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2006, 2008, and 2012, respectively. In 2007, he joined the Joint Research and Development Laboratory, Chinese Academy of Sciences, Beijing, China, as a Research Assistant. From 2009, he was with the National Engineering Laboratory for Video Technology, Peking University, Beijing, as a Research Assistant. From 2012 to 2013, he was a Post-Doctoral Fellow with McMaster University, Hamilton, ON, Canada. His research interests include image/video coding, image/video processing, and machine learning.