IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 August 2023, accepted 18 October 2023, date of publication 23 October 2023, date of current version 30 October 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3326718

RESEARCH ARTICLE

Quantization-Based Adaptive Deep Image Compression Using Semantic Information

ZHONGYUE LEI[1], (Student Member, IEEE), XUEMIN HONG[1], (Member, IEEE), JIANGHONG SHI[1], (Member, IEEE), MINXIAN SU[2], CHAOHENG LIN[3], AND WEI XIA[4]

[1]School of Informatics, Xiamen University, Xiamen 361005, China
[2]Xiamen Satellite Positioning Application Company Ltd., Xiamen 361008, China
[3]Xiamen Beidou Key Laboratory of Applied Technology, Xiamen 361008, China
[4]Fujian Centerm Information Company Ltd., Fuzhou 350028, China

Corresponding author: Xuemin Hong (xuemin.hong@xmu.edu.cn)

This work was supported by the Science and Technology Key Project of Xiamen under Grant 3502Z20221027.

ABSTRACT Deep image coding (DIC) for hybrid application contexts has recently attracted significant research interest because of its potential to support both human and machine visual tasks. Since the regions of interest (ROI) are different for different application contexts, it is important to design an adaptive image coding mechanism in practical DIC. In this paper, we propose the first quantization-based adaptive DIC framework for hybrid contexts of image reconstruction and classification. This framework can be applied to upgrade existing fixed-rate DIC models into adaptive DIC for hybrid contexts. It consists of two key modules: a semantics-based ROI mask generation module and a module for generating ROI gain and inverse gain matrices. These matrices are used to control the quantization accuracy of different latent vector elements, thereby achieving encoding at different rates while prioritizing the reconstruction quality of the ROI. Moreover, we propose a five-stage training method for the quantization-based adaptive DIC model to optimize the rate-distortion-classification-perception (RDCP) tradeoff. Experiments over a wide rate range show that our method achieves superior RDCP tradeoff performance. Compared to the benchmark scheme BM-CHENG, the proposed algorithm improves the classification accuracy by an average of 15%. The average relative improvements on various metrics, such as natural image quality evaluator (NIQE), learned perceptual image patch similarity (LPIPS), and feature similarity index measure (FSIM), are about 22%, 47%, and 1%, respectively. The proposed algorithm is a promising candidate for fast adaptive coding with low-complexity constraints.

INDEX TERMS Deep image compression, semantic importance, adaptive coding, hybrid contexts.

I. INTRODUCTION

Image compression aims to reduce the number of bits needed to represent an image while retaining essential information [1], [2]. Conventional image compression algorithms, such as JPEG [3], JPEG 2000 [4], and BPG [5], include modules such as transform coding, entropy coding, and quantization. Typically, these modules are manually designed and optimized by experts to enhance encoding efficiency. In contrast, deep image coding (DIC) utilizes deep neural networks (DNN) to construct the image codec and optimize its modules through end-to-end learning [6], [7]. DIC has demonstrated superior performance over traditional methods.

A key advantage of DIC is its ability to effectively learn the most important features in specific contexts, thereby better preserving these features in lossy compression [8], [9]. The term ''context'' here refers to the external environment of image coding. In this paper, different DIC contexts are distinguished based on application types and objectives. The most common context is image reconstruction, which can serve both human visual perception and downstream computer vision tasks [10], [11], [12]. However, image

The associate editor coordinating the review of this manuscript and approving it for publication was Zhaoqing Pan.

reconstruction is semantically indifferent and treats each pixel as equally important [13]. On the contrary, high-level computer vision tasks usually have focused features or regions on the image, depending on their specific application context [14], [15]. The DIC that is jointly optimized for the context of reconstruction and high-level computer vision is called hybrid context DIC [16], [17], [18]. The high-level task context can be further categorized according to specific tasks such as image classification [19], semantic segmentation [20], and object detection [21], etc. Features that are relevant to the task are called semantic features and should be prioritized during lossy compression [22].

DIC can be optimized for single context of image construction or task-relevant feature extraction. For image reconstruction, the codec is optimized for the highest fidelity of the reconstructed image at the receiver end [23]. Although the reconstructed images can be used as task inputs, the task performance is poor at high compression rates [24], [25], [26]. This is because reconstruction-oriented DIC cannot preserve task-relevant features with high priority. On the contrary, in DIC optimized for high-level task context, the receiver directly obtains task results from the latent vector (i.e., features) without generating the reconstructed image [15]. This can yield better task performance at lower rates, but the latent vector cannot be used for human perception or other tasks.

In hybrid contexts DIC, the model is optimized for both image reconstruction and high-level tasks, resulting in high-quality reconstructed images and superior task performance [27]. In hybrid contexts DIC, capturing the semantics of the specific task target and selectively preserving the corresponding information during lossy coding is essential to ensuring optimized performance. In this paper, we study DIC for hybrid contexts of image reconstruction and classification, where the semantics vary for the same image depending on the specific classification task. Semantics is mainly related to the region where the target is located. For example, in traffic surveillance, DNN models should focus on different regions in the image when detecting pedestrian and vehicle violations. Therefore, DIC requires flexible mechanisms to adapt to different regions of interest (ROI) in different contexts. Moreover, in communication scenarios, bandwidth resources are usually limited and vary, especially in wireless communications [28]. Therefore, rate-adaptive image coding is essential for DIC used for real-time communications. The conventional approach of creating multiple DIC models for rate-adaptive coding requires significant computational and storage resources. To address this drawback, our goal is to design a single DIC model that can flexibly adapt to variable rates and hybrid contexts of image reconstruction and classification.

As shown in Figure 1, we propose a novel quantization-based adaptive ROI deep image coding scheme in this paper. The proposed codec can achieve rate and context adaption using a single DNN model, and the reconstructed images have good visual quality and high accuracy in downstream image classification tasks, supporting both human visual perception and image classification. The proposed design includes three features. First, the scheme considers the importance of different regions of the image for classification semantics and prioritizes the reconstruction quality of semantic salient regions within the rate range. To this end, a pair of ROI gain matrix (GM) and inverse-gain matrix (IGM) generation modules are proposed. These modules are deployed at the encoding and decoding ends to generate ROI GM and IGM for scaling the latent vector and controlling the quantization accuracy of the latent vector at different code rates. Moreover, the quantization-based adaptive ROI DIC mechanism can quantize elements of the latent vector corresponding to semantic salient regions at a finer grain, resulting in better reconstruction quality in these regions. Second, a semantics-based ROI mask generation module is proposed to obtain the ROI mask that can identify semantic salient regions. The issue of DIC model training is formulated as a rate-distortion-classification-perception (RDCP) trade-off problem, which is initially defined in work [29]. The weighted sum of overall image mean squared error (MSE) loss, ROI MSE loss, semantic feature matching (SFM) loss, and generative adversarial network (GAN) loss is adopted as the objective functions to optimize the distortion, classification, and perception objectives simultaneously. Third, a DIC model training method with incremental optimization of each objective is designed to obtain the quantization-based adaptive ROI DIC model.

Our contributions can be summarized as follows: (1) The proposed quantization-based adaptive ROI DIC for hybrid contexts of image reconstruction and classification can achieve superior RDCP trade-off performance. (2) The proposed ROI mask generation module and ROI GM and IGM generation modules can be directly embedded into existing DIC models to achieve arbitrary bit-rate coding and prioritize the reconstruction quality of semantic salient regions. (3) The proposed quantization-based adaptation mechanism reduces the complexity and processing latency of adaptive coding by avoiding repeated calls to the encoder and is suitable for realizing fast adaptive coding in hybrid contexts with low complexity constraints.

II. RELATED WORK

Tables 1 and 2 provide an overview of existing studies on DIC and position our paper in the literature. A brief review of the related literature is given below.

A. FIXED-RATE DIC WITH HYBRID CONTEXT

The goal of hybrid context DIC is to obtain reconstructed images that have both high visual quality and good performance in high-level tasks. There are two main approaches to improving the performance of hybrid context DIC: one is to add loss terms related to specific high-level tasks during DIC model training, and the other is to use ROI coding to improve the fidelity of task-relevant pixels.

TABLE 1. Fixed-rate hybrid context DIC.

Task loss optimization	ROI coding	Combination of two methods
[27], [30]	[33]	[26], [34]

TABLE 2. Adaptive DIC.

Adaptive method	Single context	Hybrid context
Encoder-based	[35], [36]	[41], [42]
Latent-based	[39], [40]	[18]
Quantization-based	[37], [38]	**Our contributions (QVRC)**

In the first approach, DIC algorithms incorporate task-relevant loss terms into the rate-distortion-optimized loss function for DNN model training. This can directly improve the accuracy of the reconstructed image in downstream tasks. Typical works of such were proposed in [27] and [30]. In [30], the proposed DIC algorithm used pre-trained object detection and instance segmentation models to compute the object detection and instance segmentation task losses, respectively. This DIC codec outperforms the state-of-the-art Versatile Video Coding (VVC) standard on the object detection and instance segmentation tasks, achieving -37.87% and -32.90% of BD-rate gain, respectively. In [27], the identity preserving loss was taken as the task loss term, saving about 38% of bpp at the same detection rate compared to the HEVC standard.

As the second approach, ROI coding is widely used to improve the high-level task performance of the reconstruction image obtained from traditional image/video codecs. In [31] and [32], ROI video coding methods were proposed to improve task accuracy based on the HEVC and VVC standards, respectively. These works used machine learning models to detect salient regions in the images and subsequently assign higher bit rates to the salient regions. The bit rate can be significantly reduced while maintaining the same detection accuracy as standard video coding algorithms. In [33], a ROI coding method that adapts the image quality for prioritized or non-prioritized parts for DIC was proposed to improve the accuracy of road damage detection. This method can reduce the bpp by 31% compared to the original method.

The above two approaches can be combined to further enhance the reconstructed images' visual quality and task performance. In [34], the proposed DIC algorithm added semantic-aware loss terms to DIC model training and designed an attention module based on semantic prior information to implement ROI coding in the compressed feature domain. Compared to DIC algorithms for the context of image reconstruction, the implemented algorithm has comparable visual quality performance, but better classification accuracy. In [26], the proposed DIC algorithm added SSIM loss and distortion terms in the task feature domain to improve the perceptual quality and computer vision task performance of the reconstructed images. Moreover, this algorithm designed a latent vector spatial mask generating network, where the generated mask was multiplied with the latent vector, and task-irrelevant elements were set to zero to obtain a higher compression rate.

The DIC algorithms introduced above can be categorized as fixed-rate DIC and have obvious limitations. First, because they need to train different DNN models for different bit rates and specific tasks, the training and storage of different DNN models may consume severe computation and storage resources. Second, at low bit rates, the reconstructed images may suffer from blurring due to excessive information loss, which affects the overall image perception.

B. ADAPTIVE DIC WITH SINGLE CONTEXT

Unlike fixed-rate DICs introduced above, adaptive DIC can encode an image into varying rates based on a single DNN model. Adaptive DIC can be classified into three types: encoder-based, quantization-based, and latent-based. Encoder-based adaptive DIC manipulates control conditions upon the encoder to generate different latent vectors at different rates. For example, the DIC proposed in [35] conditioned the encoder and decoder on λ, which balances distortion and rate. The encoder scaled the output features by λ at each layer to achieve discrete multi-rate coding. For continuous multi-rate coding, the authors in [36] extended [35] by adding a λ-interpolation mechanism. Quantization-based adaptive DIC assigns different quantization intervals to the latent vector elements by scaling them with different coefficients. In [37], a high-rate DIC model was pre-trained, and then several pairs of GM and IGM were trained to scale each channel of the latent vector at different rates. A more flexible method was proposed in [38], which trained the GM and IGM generation modules at the codec sides. Latent-based rate-adaptive DIC, on the other hand, generates a single latent vector that can be encoded to produce a bitstream for various bitrates. For example, the DIC proposed in [39] transformed the original image into layered latent vectors with ordered dependencies using a residual network, while [40] used a spatial mask to modulate the channel and achieve the desired rate.

C. ADAPTIVE DIC WITH HYBRID CONTEXTS

The above-mentioned studies about adaptive DIC are all optimized for a single context/goal of image construction. Efforts were made in the following literature to extend the adaptive capability of DIC from single context to multiple contexts. In [41], an encoder-based adaptive DIC

for hybrid contexts of image reconstruction and classification was proposed. It prioritized the reconstruction quality of semantic salient regions in variable-rate coding. However, during DIC model training, only the distortion objective was optimized, resulting in an essentially reconstruction-oriented ROI coding method with limited classification accuracy. To further improve the classification accuracy, the authors in [42] proposed a RDCP joint optimization framework to train the neural network for hybrid contexts. Averaged over the tested rate range, it outperforms [41] in classification accuracy, NIQE, LPIPS, and FSIM by 11%, 12.4%, 32%, and 1.3%, respectively. In [18], the RDCP joint optimization framework was also used to train a latent-based adaptive DIC. Compared with the benchmark algorithm [40], it achieves relative performance gains ranging from 4% to 90% in various metrics corresponding to distortion, perception, and classification.

The adaptive DIC methods introduced above used either encoder-based or latent-based architecture. These two types of DIC require significant modifications to the original DNN architectures and are difficult to design. Compared with these two types, quantization-based adaptive DIC has two advantages. First, it does not need to modify the backbone DNN network and can be easily generalized to existing DNN models. Second, during adaptive coding, the encoder is executed only once to generate the latent vector. This can save computational resources and reduce the processing delay. Existing quantization-based methods [37], [38], [43] are restricted to the single context of image reconstruction. To our best knowledge, this paper makes the first effort in proposing a quantization-based adaptive DIC for multiple contexts.

New image coding standards are under development based on the DIC approach, notably the JPEG-AI [44] and video for machine (VCM) standard [45]. The JPEG-AI standard mainly aims for high-fidelity reconstruction of images. In other words, it mainly concerns the single application context of image reconstruction. In contrast, the VCM standard aims for a broader scope incorporating both human and machine visual tasks. To put our work into perspective, the proposed hybrid context DIC can be seen as a candidate scheme for the VCM standard and a complementary extension to the JPEG-AI standard.

III. PROPOSED QUANTIZATION-BASED ADAPTIVE DIC FOR HYBRID CONTEXTS

Adaptive DIC for hybrid contexts of image reconstruction and classification aims to produce reconstructed images with high visual quality and classification accuracy at different rates. High-quality reconstructed images have low distortion and high perceptual quality. Therefore, the proposed adaptive hybrid context DIC model in this paper should be optimized for multiple objectives, including distortion, classification, and perception. Moreover, the classification accuracy is related to the reconstruction quality of the region where the classification target is located, which is defined as the semantic salient region. Thus, classification accuracy can be enhanced by prioritizing the reconstruction quality of the semantic salient region during adaptive coding. In this paper, we extend the quantization-based rate-adaptive DIC to multiple contexts in two ways. First, by using the RDCP optimization framework for DIC model training. Second, by using the quantization-based ROI adaptive DIC.

A. THE FRAMEWORK OF PROPOSED ADAPTIVE DIC

Figure 1 shows the framework of our proposed quantization-based adaptive DIC for hybrid contexts. The framework comprises three main modules: ROI GM and IGM generation, a semantics-based ROI mask generation, and a GAN-based DIC network. ROI GM and IGM generation modules are deployed at the encoder and decoder sides, respectively. They produce ROI GM and IGM for scaling the latent vectors at various bit rates, depending on the quality factor and ROI mask. The quality factor is proportional to the bit rate. Unlike other quantization-based rate-adaptive DIC methods, our method assigns larger values to the ROI positions in each channel of the ROI GM and IGM. This enables finer quantization of the latent vector elements corresponding to the ROI, resulting in better reconstruction quality for the ROI at various bit rates. The ROI mask generation module produces a binary mask that indicates the ROI, depending on the image encoding context. The implementation of this module varies according to the context. In this paper, we design a semantics-based ROI mask generation module for the hybrid context of image reconstruction and classification. The size of the ROI mask area is dynamically adjusted according to the quality factor. As the bit rate increases, our method prioritizes maintaining classification accuracy while also improving overall image reconstruction quality. A GAN generator acts as the decoder for image reconstruction, ensuring that the distribution of the reconstructed images is close to that of the original images and achieves a higher perceptual quality.

The encoding and decoding processes of our framework are as follows. Given an input image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$, the encoder generates a latent vector $\mathbf{y} \in \mathbb{R}^{\frac{H}{K}\times\frac{W}{K}\times C}$ by downsampling the input image by a factor of K and extracting features. The latent vector represents the compressed representation of the input image. The ROI mask generation module generates a binary ROI mask $\mathbf{m} \in \{0,1\}^{H\times W\times 1}$ corresponding to the target bit rate based on the input image and the quality factor q, $q \in [0, 1]$. The elements with a value of 1 in the ROI mask indicate that the corresponding pixels belong to regions of interest (ROI) that should have better reconstruction quality, while the rest are non-ROI pixels. The ROI GM generation module generates the corresponding GM $\mathbf{g} \in \mathbb{R}^{\frac{H}{K}\times\frac{W}{K}\times C}$ based on $\mathbf{m}$ and q. The latent vector $\mathbf{y}$ is element-wise multiplied with $\mathbf{g}$ to obtain $\tilde{\mathbf{y}}$, which is then quantized to $\bar{\mathbf{y}}$ using a unified scalar quantizer. The entropy model estimates the distribution of $\bar{\mathbf{y}}$ for entropy coding and decoding, using the contextual information extracted from $\tilde{\mathbf{y}}$. The entropy

encoder encodes $\bar{\mathbf{y}}$ into a bitstream, which is transmitted along with $\mathbf{m}$ and q. The entropy decoder decodes the bitstream into $\bar{\mathbf{y}}$. The ROI IGM generation module at the decoder side generates $\mathbf{g}' \in \mathbb{R}^{\frac{H}{K} \times \frac{W}{K} \times C}$ based on $\mathbf{m}$ and q. The reconstructed latent vector $\hat{\mathbf{y}}$ is obtained by element-wise multiplying $\bar{\mathbf{y}}$ with $\mathbf{g}'$. The deep decoder upsamples $\hat{\mathbf{y}}$ by a factor of K to obtain the reconstructed image $\hat{\mathbf{x}}$. The encoding and decoding processes can be expressed by equations (1) and (2), respectively.

$$\begin{cases} \mathbf{y} = f_{\mathrm{e}}(\mathbf{x}) \\ \mathbf{m} = f_{\mathrm{m}}(\mathbf{x}, q) \\ \mathbf{g} = f_{\mathrm{g}}(\mathbf{m}, q) \\ \tilde{\mathbf{y}} = \mathbf{y} \circ \mathbf{g} \\ \bar{\mathbf{y}} = \mathrm{Round}(\tilde{\mathbf{y}}), \end{cases} \tag{1}$$

$$\begin{cases} \mathbf{g}' = f_{\mathrm{g}'}(\mathbf{m}, q) \\ \hat{\mathbf{y}} = \bar{\mathbf{y}} \circ \mathbf{g}' \\ \hat{\mathbf{x}} = f_{\mathrm{d}}(\hat{\mathbf{y}}), \end{cases} \tag{2}$$

where $f_{\mathrm{e}}, f_{\mathrm{d}}, f_{\mathrm{m}}, f_{\mathrm{g}}$ and $f_{\mathrm{g}'}$ denote the encoder, decoder, ROI mask generation module, ROI GM generation, and ROI IGM generation module, respectively.

According to the selected quality factor, the above process compresses the image at a certain bit rate. When the target bit rate or compression scenario of the same image changes, there is no need to regenerate the latent vector. Instead, we can simply adjust the quality factor or modify the region of interest (ROI), regenerate the corresponding GM and IGM, and then apply them to the scaled latent vector to achieve adaptive DIC.

B. ROI GM AND IGM GENERATION MODULES

The ROI GM and IGM generation modules are based on the attention mechanism [46]. They have the same process, as shown in Figure 2. The quality factor is first transformed into a vector $\mathbf{u}$ with the same number of channels as the latent vector by two fully connected layers:

$$\mathbf{u} = (\mathbf{H}_2(\mathrm{Relu}(\mathbf{H}_1 q + \mathbf{b}_1)) + \mathbf{b}_2), \tag{3}$$

where $\mathbf{H}_1$, $\mathbf{H}_2$, $\mathbf{b}_1$, and $\mathbf{b}_2$ are the parameters of two fully connected layers. The vector $\mathbf{u}$ is expanded into a vector $\bar{\mathbf{u}}$ with the same dimensions as the latent vector $\mathbf{y}$ by setting $\bar{u}_{l,i,j} = u_l$, where l, i, and j are the indices corresponding to the channel, row, and column of $\mathbf{y}$, respectively.

The ROI mask m is downsampled to a vector with the same weight and height as $\bar{\mathbf{u}}$ by adaptive max-pooling. The downsampled vector is concatenated with $\bar{\mathbf{u}}$ and passed through a CNN to extract the feature attention map $\hat{\mathbf{u}}$:

$$\hat{\mathbf{u}} = \mathrm{Conv}(\mathrm{Relu}(\mathrm{Conv}(\mathrm{Concat}(\mathrm{maxpool}(\mathbf{m}), \bar{\mathbf{u}})))), \tag{4}$$

where $\mathrm{Conv}(\cdot)$, $\mathrm{maxpool}(\cdot)$, and $\mathrm{Concat}(\cdot, \cdot)$ denote convolution, adaptive max-pooling, and concatenation of two vectors along the channel dimension, respectively. The ROI IGM or GM is computed by applying an element-wise exponential function to the sum of $\bar{\mathbf{u}}$ and the product of $\bar{\mathbf{u}}$ and $\hat{\mathbf{u}}$:

$$\mathbf{g} = \exp(\bar{\mathbf{u}} + \bar{\mathbf{u}} \circ \hat{\mathbf{u}}), \tag{5}$$

where the exponential function ensures that the IGM or GM values are positive and have a larger scaling range.

C. SEMANTICS-BASED ROI MASK GENERATION MODULE

The proposed method uses a ROI mask to identify the regions of interest which need higher-quality reconstruction. The ROI mask is then fed into the ROI GM and IGM generation modules to guide the generation of the GM and IGM, respectively. These modules aim to achieve a finer quantization of the latent vector elements that correspond to the ROI. In this paper, ROI coding is adopted to improve the classification accuracy of the reconstructed images in the downstream classifier. Therefore, the ROI consists of regions that are related to the classification target, and a semantics-based ROI mask generation module is designed to produce the ROI mask. The ROI mask adapts dynamically to the target bit rate. When the rate is low, elements of the latent vector that correspond to pixels with higher relevance to the target category have higher quantization accuracy, which results in better reconstruction quality of the classification-related regions. As the rate increases, the ROI expands accordingly to improve the reconstruction quality of more pixels that are relevant to the classification. The network structure of the semantics-based ROI mask generation module is shown in Figure 2, and the detailed process is described as follows:

First, the class activation mapping (CAM) [47], [48] is used to generate the CAM map of the original image. The CAM map is a two-dimensional vector of the same size as the image, with each element corresponding to a pixel and describing its importance to the classification result. The elements of the CAM map are then normalized to [0,1], and the resulting map is defined as a semantic importance map, denoted by a vector $\mathbf{s}$.

Next, based on $\mathbf{s}$ and the quality factor q, a semantic importance energy map is generated. To ensure that the energy map is not a zero vector when $q = 0$, q is normalized to $\bar{q} = \frac{\exp(q)}{\exp(1)}$ before computing the energy map. The energy map is calculated as $\bar{\mathbf{s}} = (\bar{q}\mathbf{s})^2$.

Finally, based on $\bar{\mathbf{s}}$, equation (6) is used to calculate the ROI mask $\mathbf{m}$. The semantic importance map is first processed by two convolutional layers for feature extraction and then binarized using a sign function to obtain $\mathbf{m}$.

$$\mathbf{m} = (\mathrm{sign}(\mathrm{Conv}(\mathrm{Relu}(\mathrm{Conv}(\bar{\mathbf{s}})))) + 1)/2. \tag{6}$$

Since the sign function is not differentiable, it is replaced by the sigmoid function during training.

D. LOSS FUNCTION DESIGN

The quality of the reconstructed images in DIC depends on the choice of the training objective function. The proposed method uses a combination of five loss terms to optimize the image reconstruction process: an image distortion term

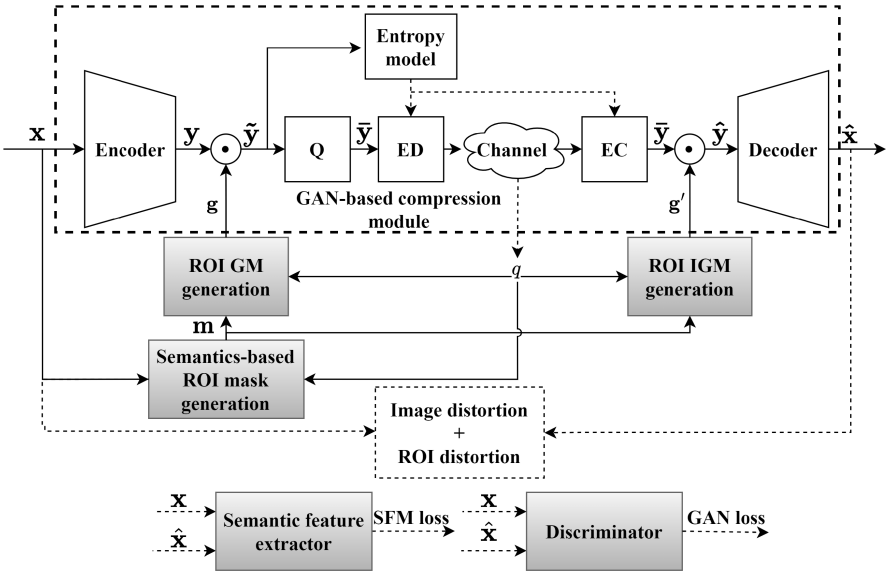


FIGURE 1. Framework of the proposed quantization-based adaptive DIC for hybrid contexts. The notations Q, EC, and ED represent quantization, entropy coding, and entropy decoding, respectively.

$\mathcal{L}_{\text{IM}}$, an ROI distortion term $\mathcal{L}_{\text{ROI}}$, an SFM loss term $\mathcal{L}_{\text{SFM}}$, a GAN term $\mathcal{L}_{\text{SFM}}$, and a rate term. The image distortion term measures the MSE between the original image and the reconstructed image, as defined by:

$$\mathcal{L}_{\text{IM}} = \frac{1}{3HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{k=1}^{3} \| x_{i,j,k} - \hat{x}_{i,j,k} \|^2 , \quad (7)$$

where $x_{i,j,k}$ and $\hat{x}_{i,j,k}$ denote the pixel values of the original and reconstructed images, respectively. The ROI distortion term focuses on the MSE between the pixels that belong to the ROI in both images, which is expressed by:

$$\mathcal{L}_{\text{ROI}} = \frac{1}{3HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{k=1}^{3} \| m_{i,j}(x_{i,j,k} - \hat{x}_{i,j,k}) \|^2 , \quad (8)$$

where $m_{i,j}$ is the element of the ROI mask vector $\mathbf{m}$. The SFM loss term evaluates the similarity between the feature vectors extracted from the original image $\mathbf{x}$ and the reconstructed image $\hat{\mathbf{x}}$ using a pre-trained feature extractor. The SFM loss is calculated by the mean absolute error (MAE) between the feature vectors, as given by:

$$\mathcal{L}_{\text{SFM}} = \frac{1}{H_{\text{F}} W_{\text{F}} C_{\text{F}}} \sum_{i=1}^{H_{\text{F}}} \sum_{j=1}^{W_{\text{F}}} \sum_{k=1}^{C_{\text{F}}} \| \text{F}(\mathbf{x})_{i,j,k} - \text{F}(\hat{\mathbf{x}})_{i,j,k} \| , \quad (9)$$

where $H_{\text{F}} \times W_{\text{F}} \times C_{\text{F}}$ represents dimensions of the feature vectors, and $\text{F}(\cdot)$ denotes the feature extracting function.

To enhance the perceptual quality and reduce the distortion of the reconstructed images, the GAN term is based on the unsaturated conditional GAN proposed by [49]. Moreover, the scaled latent vector $\hat{\mathbf{y}}$ and the ROI mask $\mathbf{m}$ are used as conditional inputs to the generator and discriminator to improve the reconstruction of the ROI. The generator and discriminator loss terms are defined as follows:

$$\mathcal{L}_{\text{GAN}} = -\log(D(G(\hat{\mathbf{y}}), \hat{\mathbf{y}}, \mathbf{m})), \quad (10)$$

$$\mathcal{L}_{\text{D}} = -\log(1 - D(G(\hat{\mathbf{y}}), \hat{\mathbf{y}}, \mathbf{m})) - \log(D(\mathbf{x}, \hat{\mathbf{x}}, \mathbf{m})), \quad (11)$$

where D and G denote the discriminator and the generator, respectively.

The rate term R is the number of bits to encode $\mathbf{y}$, which is constrained by the entropy bounds. However, the true distribution $p_{\mathbf{y}}$ of $\mathbf{y}$, is unknown, so we cannot compute the exact entropy. Therefore, we use an entropy model $\rho_{\mathbf{y}}$ to estimate $p_{\mathbf{y}}$ for entropy coding. Thus, the rate term is the cross entropy of $p_{\mathbf{y}}$ and $\rho_{\mathbf{y}}$ as follow:

$$R = \mathbb{E}_{\mathbf{y} \sim p_{\mathbf{y}}}[-\log \rho_{\mathbf{y}}(\mathbf{y})]. \quad (12)$$

E. FIVE-STAGE TRAINING

We propose a five-stage training algorithm to ensure the GAN-based codec, ROI GM and IGM generation, and ROI mask generation model function properly. The main steps of the training algorithm are summarized in Algorithm 1.

Algorithm 1 Algorithm for Five-stage Training

Input: Training dataset; number of training steps for each stage st_i; batch size b

Output: Trained GAN-based codec, ROI GM and IGM generation modules, and ROI mask generation module

//First stage training

1: **for** step =1 to st_1 **do**
2: Sample a batch of images $X^{(b)}$
3: Encode images, compute rates of quantized latent vectors, and obtain reconstructed images
4: Calculate $\mathcal{L}_{\text{S1}}$ via (13)

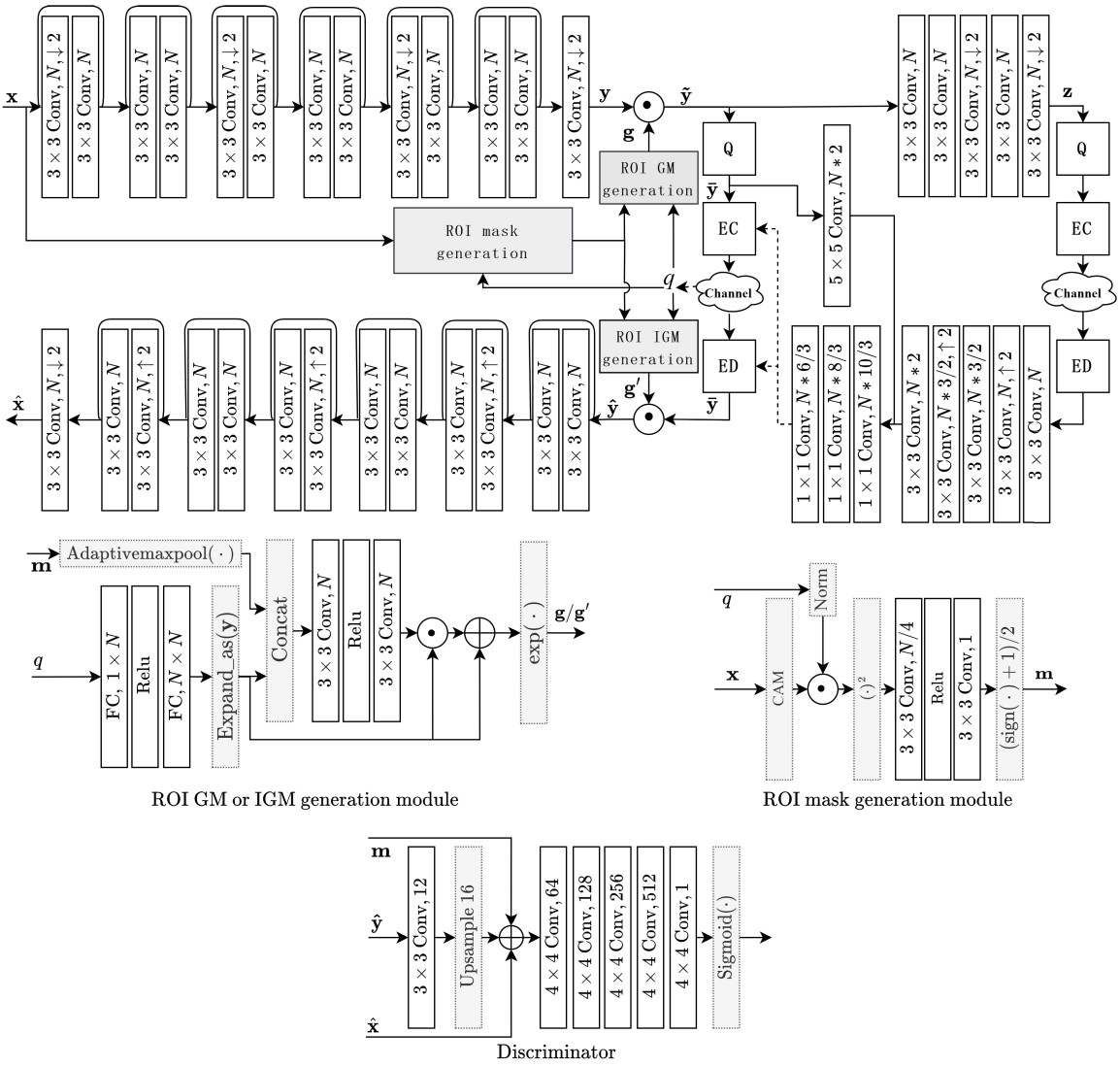


FIGURE 2. The network structure of the implementation based on the DIC model in [53]. The notations 3 × 3 *Conv*, *N*, ↓ 2 and 3 × 3 *Conv*, *N*, ↑ 2 refer to the convolutional and deconvolutional layers with a kernel size of 3 × 3 and a stride of 2, respectively. The default stride of convolutional layers is 1.

Update the parameters of the encoder, decoder, and entropy model by applying their stochastic gradient descent
end for
//Second stage training
Add ROI IGM and GM generation modules
for step = $st_1 + 1$ to st_2 **do**
Sample a batch of images $X^{(b)}$ and quality factors $q^{(b)}$
Obtain the CAM map of each image and binarize them with random thresholds to get ROI masks $m^{(b)}$
Encode the images depending on $q^{(b)}$ and $m^{(b)}$, estimate the rate of $\bar{\mathbf{y}}^{(b)}$, and get reconstructed images
Calculate $\mathcal{L}_{S2}$ via (14)
Update parameters of the encoder, decoder, entropy model, ROI IGM, and GM generation modules by descending their stochastic gradient
end for
//Third stage training
for step = $st_2 + 1$ to st_3 **do**
Sample a batch of images $X^{(b)}$ and quality factors $q^{(b)}$
Obtain the CAM map of each image and binarize them with random thresholds to get ROI masks $m^{(b)}$
Encode the images depending on $q^{(b)}$ and $m^{(b)}$, estimate the rate of $\bar{\mathbf{y}}^{(b)}$, and get reconstructed images.
Calculate $\mathcal{L}_{S3}$ via (15)
Update parameters of the encoder, decoder, entropy model, ROI IGM, and GM generation modules by descending their stochastic gradient
end for
//Fourth stage training
Add discriminator

for step = $st_3 + 1$ to st_4 **do**
  Sample a batch of images $X^{(b)}$ and quality factors $q^{(b)}$
  Obtain the CAM map of each image and binarize them with random threshholds to get ROI masks $m^{(b)}$
  Encode the images depending on $q^{(b)}$ and $m^{(b)}$, get reconstructed images, and estimate the rate of $\bar{\mathbf{y}}^{(b)}$
  Calculate $\mathcal{L}_{S4}$ via (16)
  Update parameters of the decoder by descending its stochastic gradient
  Calculate $\mathcal{L}_{D}$ via (11)
  Update parameters of the discriminator by descending its stochastic gradient
end for
//Fifth stage training
Add ROI mask generation module
for step = $st_4 + 1$ to st_5 **do**
  Sample a batch of images $X^{(b)}$ and quality factors $q^{(b)}$
  Obtain ROI masks $m^{(b)}$ using ROI mask generation module
  Encode the images depending on $q^{(b)}$ and $m^{(b)}$, get reconstructed images, estimate the rate of $\bar{\mathbf{y}}^{(b)}$
  Calculate $\mathcal{L}_{S5}$ via (17)
  Update parameters of the ROI mask generation module by descending its stochastic gradient
end for

In the first stage, we train a fixed-rate DIC model with a rate higher than the target rate range. This DIC model consists of the encoder, decoder, and entropy model. The objective function of this stage is as follows:

$$\mathcal{L}_{S1} = R_1 + \lambda_1 \mathcal{L}_{IM}, \tag{13}$$

where R_1 is the estimated entropy of the quantized latent vector. The Lagrangian coefficient λ_1 determines the bit rate of the compressed image, which is set according to [50]. To overcome the quantization indifference problem during training, we use uniform noise instead of the round operation.

In the second stage, we add the ROI GM and IGM generation modules to the pre-trained DIC model from the first stage. The latent vector is scaled by the GM and IGM before scalar quantization and decoding, respectively. The objective function of this stage is

$$\mathcal{L}_{S2} = R_2 + \lambda_2(\mathcal{L}_{IM} + \lambda_{ROI}\mathcal{L}_{ROI}), \tag{14}$$

where λ_{ROI} is the weight to balance the ROI distortion and R_2 is the estimated entropy of the quantized latent vector after being scaled by GM. During training, we obtain the ROI mask $\mathbf{m}$ by random binarization of the CAM map of the original image. To achieve a continuous rate adaptation, we randomly sample $q \in [0, 1]$ as the quality factor and map it to the Lagrangian coefficient by $\lambda_2 = \alpha(e^{\beta q})$ during training. The hyper-parameters α and β are set according to the target rate range, where α and β are determined by the lowest and highest target rates, respectively.

In the third stage, we fine-tune the adaptive DIC model from the second stage by adding the SFM loss $\mathcal{L}_{SFM}$. This stage helps the codec and ROI GM and IGM generation modules to better extract semantically relevant information from the original images. The objective function of this stage is

$$\mathcal{L}_{S3} = \mathcal{L}_{S2} + \lambda_{SFM}\mathcal{L}_{SFM}, \tag{15}$$

where λ_{SFM} is the weight of the SFM loss.

In the fourth stage, we freeze the parameters of the encoder, ROI GM and IGM generation modules, and entropy model, and fine-tune only the decoder's parameters by adding the GAN loss term. This stage aims to make the distribution of the reconstructed image closer to that of the original image, thus improving the perceptual quality of the reconstructed image at different rates. This stage requires alternating iterations to train the decoder and discriminator. The loss functions of the decoder and discriminator are equations (16) and (11), respectively.

$$\mathcal{L}_{S4} = \mathcal{L}_{S3} + \lambda_{GAN}\mathcal{L}_{GAN}, \tag{16}$$

where λ_{GAN} is the weight of GAN term.

In the fifth stage, we add the ROI mask generation module to produce the ROI mask. During this stage, only the parameters of this module are updated. The objective function of this stage is

$$\mathcal{L}_{S5} = R_2 + \lambda_{CE}\mathcal{L}_{CE}, \tag{17}$$

where $\mathcal{L}_{CE}$ is the cross-entropy between the true and predicted classification labels of the reconstructed image and λ_{CE} is the weight of the cross-entropy term. The pre-trained VGG16 model [51] is used for both the calculation of the cross-entropy loss and the ROI generation module. Furthermore, we use Grad-CAM++ [52] to obtain the CAM map.

IV. EXPERIMENTS AND RESULTS ANALYSIS

A. IMPLEMENTATION ON EXISTING COMPRESSION MODELS

The proposed design (ROI GM and IGM generation modules, ROI mask generation module, and five-stage training method) can be applied to upgrade existing fixed-rate DIC models into adaptive DIC for hybrid contexts. We adopt models in [12] and [53] as our backbone networks and apply our proposed design to them. The resulting new models are trained using the method introduced in Section III-E and are called HQVRC-MBT and HQVRC-CHENG, respectively.

Figure 2 shows the network structure of HQVRC-CHENG. The conditional Gaussian model is used to estimate the distribution of the latent vector. Mean and variance of the conditional Gaussian model are estimated by the side information extracted from the latent vector and the decoded symbols. HQVRC-MBT has the same framework as HQVRC-CHENG but two differences: 1) the encoder and decoder are without residual blocks; 2) the parameters of the entropy model are only estimated by the side information extracted from the latent vector.

The COCO dataset [54] with data augmentation by randomly cropping 256 × 256 images and the Adam optimization algorithm [55] are adopted during models' training. The hyper-parameter setting is shown in Table 3 and this setting approximately results in the bpp range of [0.08,0.35] on the Kodak dataset.

TABLE 3. Default hyper-parameters setting.

Number of channels N	192
Batch size	8
Training steps of ith stage st_i	$st_1 = 3$M, $st_2 = st_3 = st_4 = st_5 = 300{,}000$
Learning rate setting	Initial 2M steps is 10^{-4} and the rest steps is 10^{-5}
Weight of loss term λ_i	$\lambda_1 = 0.025$, $\lambda_{\text{ROI}} = 1$, $\lambda_{\text{SFM}} = 0.01$, $\lambda_{\text{GAN}} = 1$, and $\lambda_{\text{CE}} = 0.7$
bpp range control parameter α, β	$\alpha = 0.0004$, $\beta = 3.2$

To evaluate the performance of our proposed models, we use the ImageNet [56] and Kodak [57] datasets for metrics related to classification, distortion, and perception. The evaluation metrics include Top-1 and Top-5 classification accuracy of the VGG16 classifier, PSNR, SSIM [58], FSIM [59], LPIPS [60], and NIQE [61]. The classification accuracy and PSNR correspond to the classification and distortion objectives, respectively. SSIM, FSIM, and LPIPS measure the similarity of the reconstructed image to the original image on different feature domains, and NIQE is a measure of perceptual degree [18], [42]. For model performance evaluation, 500 test images are randomly selected from 100 categories of the ImageNet dataset, with 5 images per category. We adopt the pre-trained VGG16 network parameters as published on the TensorFlow official website to compute the SFM loss, obtain CAM map, and evaluate the classification performance.

B. IMPACT OF LOSS WEIGHT SETTINGS

The weights on loss terms have a global and implicit impact on the RDCP tradeoff. Previous works [18], [42] have investigated how the DCP metrics vary with changing λ_1, λ_{SFM}, and λ_{GAN}. The distortion, classification, and perception objectives are mainly determined by the weights of λ_1, λ_{SFM}, and λ_{GAN}. It has been proved that perception-distortion is a strict trade-off, while classification-perception and classification-distortion are not strict trade-offs. Moreover, setting the values of λ_{SFM} and λ_{GAN} can balance the RDCP trade-off. We set λ_{SFM} and λ_{GAN} based on the experiment results in [18] and [42]. Controlling the reconstruction quality of ROI is another way to balance the classification and distortion objectives, which is affected by the λ_{ROI} and λ_{CE}. In this section, we discuss the impact of different weight settings of λ_{ROI} and λ_{CE}.

TABLE 4. Classification accuracy and PSNR as functions of bpp with varying λ_{ROI}.

bpp	Top-1 classification accuracy				PSNR (dB)			
	$\lambda_{\text{ROI}} = 1.6$	$\lambda_{\text{ROI}} = 1$	$\lambda_{\text{ROI}} = 0.5$	$\lambda_{\text{ROI}} = 0$	$\lambda_{\text{ROI}} = 1.6$	$\lambda_{\text{ROI}} = 1$	$\lambda_{\text{ROI}} = 0.5$	$\lambda_{\text{ROI}} = 0$
0.07	24%	25%	23%	20%	25.07	25.65	25.63	25.71
0.13	36%	36%	35%	31%	27.33	27.38	27.49	27.59
0.23	47%	46%	44%	42%	29.42	29.82	30.01	30.22
0.32	54%	53%	51%	49%	31.08	31.18	31.19	31.23

1) IMPACT OF λ_{ROI}

To analyze the impact of λ_{ROI} on the tradeoff performance between image reconstruction and classification and guide the DIC model training, we train the adaptive DIC models with different λ_{ROI} using $\mathcal{L}_{S2}$ as the optimization objective function. Specifically, we set λ_{ROI} to 0, 0.5, 1, and 1.6, respectively, while keeping other hyperparameters at their default values to obtain four DIC models with different ROI coding quality. After training convergence, we evaluate the performance of the reconstructed images in terms of classification accuracy and PSNR metrics using the ImageNet test set.

Table 4 shows that increasing λ_{ROI} improves the classification accuracy. However, this improvement diminishes as the value of λ_{ROI} increases, resulting in a lower PSNR. When $\lambda_{\text{ROI}} = 0$, all image pixels have the same reconstruction quality. However, increasing the value of λ_{ROI} leads to finer quantization of the latent vector elements corresponding to ROI and better reconstruction quality of semantic salient regions, resulting in higher classification accuracy. Thus, λ_{ROI} can control the tradeoff between image reconstruction and classification. When we compare $\lambda_{\text{ROI}} = 0$ and $\lambda_{\text{ROI}} = 1$, we observe that the average degradation of PSNR is less than 0.1 dB within the rate range, while the average improvement in classification accuracy is about 5%. Thus, setting a proper λ_{ROI} can provide the proposed DIC model with a good classification and reconstruction tradeoff performance, and we set λ_{ROI} to 1 in the subsequent experiments.

2) IMPACT OF λ_{CE}

When λ_{ROI} is set, adjusting λ_{CE} will affect the size of the ROI at a certain bit rate, which in turn affects the trade-off between classification and distortion tradeoff. To analyze the impacts of λ_{CE}, we train the ROI mask generation module with different λ_{CE}. Specifically, we set λ_{CE} to 0.2, 0.5, 0.7, and 1, respectively, to obtain four different ROI mask generation models. We evaluate the classification accuracy and PSNR of the reconstructed images obtained by different ROI mask generation models using the ImageNet dataset.

Table 5 shows that the reconstructed images achieve the best classification accuracy when λ_{CE} is set to 0.7, with an improvement of up to 5% compared to $\lambda_{\text{CE}} = 0.2$. At the middle part of the rate range, the PSNR performance of λ_{CE} settings of 0.2 and 1 is approximately 0.1 dB

TABLE 5. Classification accuracy and PSNR as functions of bpp with varying λ_{CE}.

bpp	Top-1 classification accuracy				PSNR (dB)			
	$\lambda_{CE}=1$	$\lambda_{CE}=0.7$	$\lambda_{CE}=0.5$	$\lambda_{CE}=0.2$	$\lambda_{CE}=1$	$\lambda_{CE=0.7}$	$\lambda_{CE}=0.5$	$\lambda_{CE}=0.2$
0.08	29%	30%	30%	28%	25.07	25.65	25.63	25.71
0.13	43%	48%	46%	43%	27.26	27.75	27.48	27.89
0.22	58%	61%	59%	57%	29.64	29.54	29.53	29.62
0.33	62%	64%	63%	62%	31.37	31.35	31.15	31.30

higher than the other two settings, while at other rates, these four λ_{CE} settings have comparable PSNR performance. Setting λ_{CE} too small or too large leads to an ROI that is too small or too large, which can result in relatively insignificant improvements in the reconstruction quality of semantic salient regions, leading to lower classification accuracy. The proposed ROI coding mechanism achieves higher classification accuracy and lower PSNR when λ_{CE} is set to 0.7, due to the tradeoff between classification and reconstruction performance. These results demonstrate that the proposed quantization-based ROI adaptive DIC can provide good classification and reconstruction tradeoff performance with proper λ_{CE} settings. Therefore, we set λ_{CE} to 0.7 in the subsequent experiments.

C. ABLATION STUDY

In this paper, we propose two mechanisms to enable adaptive DIC for hybrid contexts based on quantization and fixed-rate DIC codecs. The first mechanism is to design the ROI mask generation module and the ROI GM and IGM generation modules to produce the GM and IGM for adjusting the quantization accuracy of the latent vector. The second mechanism is to design a five-stage training method that trains a quantization-based adaptive ROI DIC model for hybrid contexts. We use the CHENG2020 model as our backbone network and train three variants of adaptive DIC models with different optimization functions: $\mathcal{L}_{S4}$, $\mathcal{L}_{S3}$, and $\mathcal{L}_{S2}$. We denote these models as HQVRC-CHENG(A), HQVRC-CHENG(B), and HQVRC-CHENG(C), respectively. The HQVRC-CHENG(A) model optimizes for classification, perception, and distortion objectives; the HQVRC-CHENG(B) model optimizes for both classification and distortion objectives; and the HQVRC-CHENG(C) model optimizes for only distortion objectives. Moreover, we implement a baseline quantization-based rate-adaptive DIC algorithm based on [38] using the CHENG2020 model and call it BM-CHENG. We compare both BM-CHENG and CHENG2020 with our proposed models on classification, perception, and distortion metrics on the ImageNet dataset. Figure 3 shows the performance comparison of the five models. The comparison reveals that:

For classification accuracy, our four proposed quantization-based adaptive DIC algorithms outperform CHENG2020. This is because we fine-tune the higher-rate backbone model and employ the quantization-based adaptive mechanism that preserves high-level semantic information better. All three variants of HQVRC-CHENG surpass BM-CHENG, which

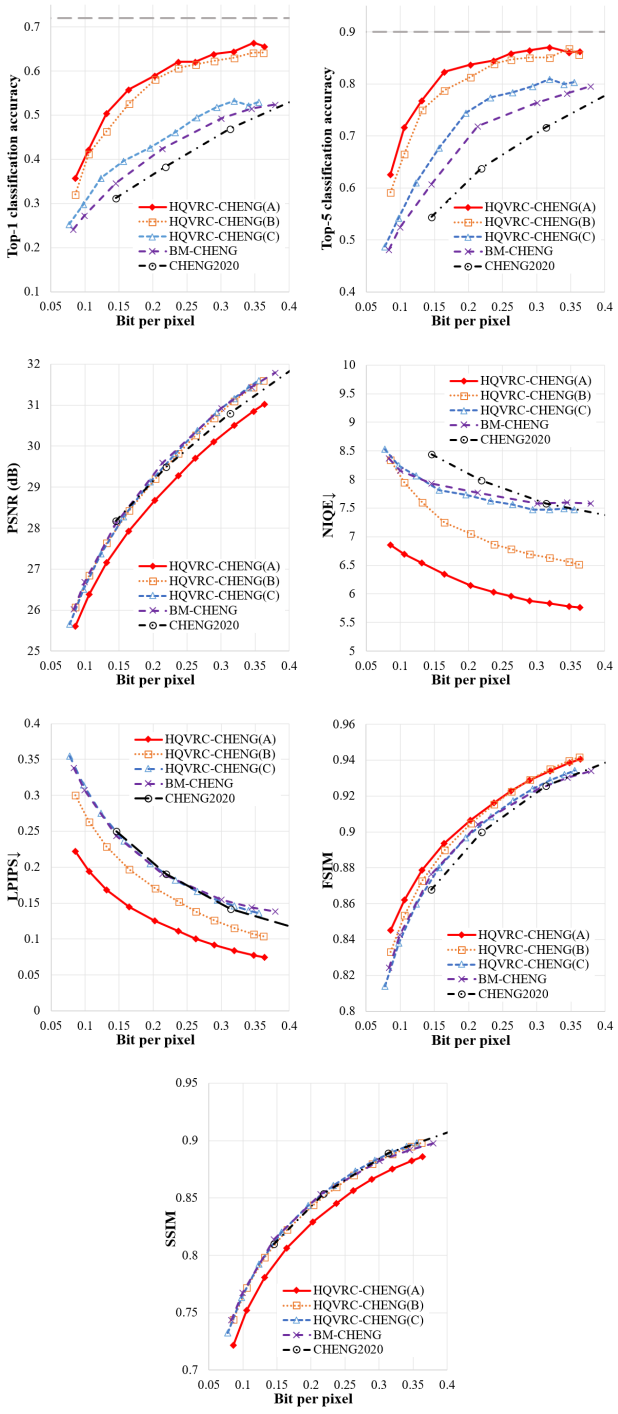


FIGURE 3. Ablation study results.

only optimizes for image reconstruction. Specifically, HQVRC-CHENG(C) shows about 3% and 5% improvements in Top-1 and Top-5 classification accuracy on average within the rate range, respectively. This improvement stems from the ROI coding mechanism that leverages the ROI GM and IGM generation modules to provide better reconstruction quality in semantic salient regions at different rates. On the other hand, HQVRC-CHENG(A) and HQVRC-CHENG(B)

attain comparable classification accuracy and exceed both HQVRC-CHENG(C) and BM-CHENG. The reason is that HQVRC-CHENG(A) and HQVRC-CHENG(B) utilize SFM loss to optimize the classification objective in DIC model training. Compared to BM-CHENG, they exhibit an average improvement of about 15% in Top-1 and Top-5 classification accuracy within the rate range.

For the distortion metric PSNR, BM-CHENG exhibits the best performance within its rate range, as it only optimizes for image reconstruction. However, HQVRC-CHENG(C) and HQVRC-CHENG(B) have comparable PSNR with a slight degradation. This is because of two factors: first, the trade-off between local and global reconstruction optimality in ROI coding; and second, the use of SFM loss to optimize the classification objective along with the ROI mechanism in HQVRC-CHENG(B), which slightly affects the PSNR performance. However, this trade-off results in a significant improvement in classification accuracy. For instance, the PSNR degradation of HQVRC-CHENG(B) by about 0.1 dB leads to an average performance improvement of about 15% in classification accuracy. Moreover, the PSNR degradation of HQVRC-CHENG(A) is 0.45 dB, mainly because it uses a GAN generator as a decoder. The GAN loss optimizes perception, which is known to be in conflict with distortion according to [18]. Therefore, the PSNR performance of HQVRC-CHENG(A) is inferior to that of BM-CHENG.

For the perception metric, we use NIQE in this paper, where a lower NIQE value indicates better performance. HQVRC-CHENG(A) has the lowest NIQE value, with an average relative improvement of about 13%, 21%, and 22%, over HQVRC-CHENG(B), HQVRC-CHENG(C), and BM-CHENG, respectively. This is because HQVRC-CHENG(A) uses a GAN generator as a decoder to obtain a reconstructed image with a distribution closer to that of a natural image. We also observe that HQVRC-CHENG(B) has a lower NIQE value than HQVRC-CHENG(C) and BM-CHENG, as it preserves the high-level semantic features better due to the use of SFM loss during DIC model training. Thus, we can infer that there is some consistency between perception and classification.

For the performance metrics LPIPS, FSIM, and SSIM, they evaluate the similarity between the reconstructed image and the original image on different feature domains. In terms of LPIPS (a lower value is better), HQVRC-CHENG(A) outperforms the other three adaptive DIC models with an average relative improvement of about 27%, 40%, and 47%, respectively. This is because the SFM loss has some similarity with LPIPS and is utilized during HQVRC-CHENG(A) and HQVRC-CHENG(B) training. Moreover, HQVRC-CHENG(A) further improves LPIPS after adding GAN loss, which shows the ability of GAN to generate higher-level features in images. Regarding FSIM, HQVRC-CHENG(A) and HQVRC-CHENG(B) have comparable performance, with an average relative improvement of about 1% and 1.6% over BM-CHENG and HQVRC-CHENG(C), respectively. For SSIM, which measures the structural similarity of images,

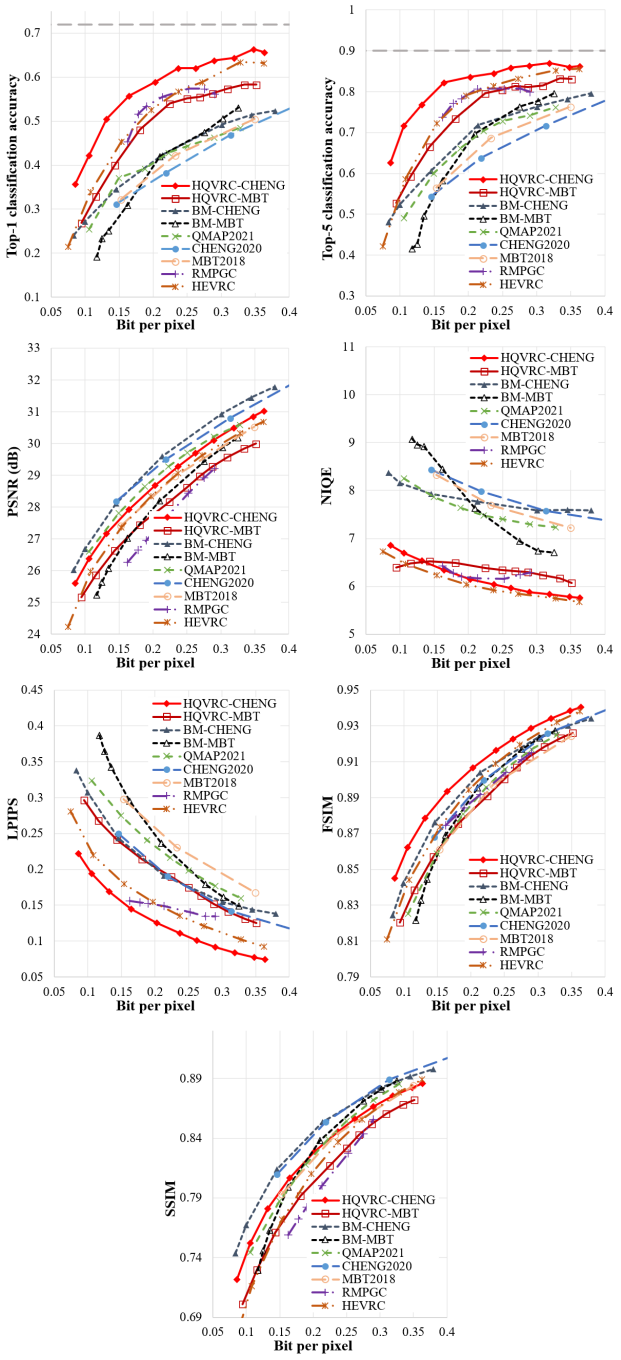


FIGURE 4. Classification, perception, and distortion performances as functions of bpp for different image coding schemes using the ImageNet dataset.

HQVRC-CHENG(A) performs poorly due to the trade-off between perception and distortion, with an average relative performance worse than BM-CHENG by about 3%.

D. PERFORMANCE EVALUATION AND COMPARISON

In this section, we compare the performance of the proposed DIC algorithms with state-of-the-art adaptive DIC algorithms. The comparison algorithms include BM-CHENG,

TABLE 6. BD-rate to other adaptive DICs of different metrics on the ImageNet dataset.

Metrics	Top-1	Top-5	PSNR	NIQE	LPIPS	FSIM	SSIM
BD-rate to BM-CHENG	-52.74	-50.86	18.13	-95.36	-49.72	-11.91	19.32
BD-rate to QMAP2021	-52.47	-52.41	5.11	-88.22	-58.28	-25.96	-2.13
BD-rate to HEVRC	-25.15	-26.04	-8.05	23.58	-22.08	-15.04	-13.64
BD-rate to RMPGC	-23.36	-29.19	-29.66	-25.48	-27.63	-23.63	-24.54

QMAP2021 [41], HEVRC [42], and RMPGC [18]. BM-CHENG is an adaptive DIC for image reconstruction context, while the rest are adaptive DIC algorithms for hybrid contexts. QMAP2021 and HEVRC are encoder-based adaptive DICs, while RMPGC is a latent-based adaptive DIC. We use the hyper-parameter setting as shown in Table 3. All these algorithms are evaluated on the ImageNet and Kodak datasets, and results are shown in Figures 4 and 5, respectively.

1) RDCP TRADE-OFF PERFORMANCES

We conduct a comparative analysis of HQVRC-CHENG with QMAP2021, HEVRC, and RMPGC. Results show that HQVRC-CHENG has similar performance on PSNR and SSIM compared with QMAP2021, but has about 16.7% and 14.5% average improvements in Top-1 and Top-5 classification accuracy, respectively, and about 18.8% average relative improvement in terms of NIQE. We note that both HQVRC-CHENG and QMAP2021 can assign higher rates to semantic salient regions. However, because HQVRC-CHENG further incorporates SFM loss and GAN loss during training, it can attain higher classification accuracy and perceptual quality. HEVRC and RMPGC also use SFM loss and GAN loss to optimize their DIC models, but HQVRC-CHENG still outperforms them on classification and distortion metrics thanks to a better entropy model. As a performance benchmark, the Top-1 and Top-5 classification accuracy for the original images from the ImageNet test set are 72% and 90%, respectively.

To better demonstrate the superior performance of HQVRC-CHENG, we summarize the BD-rate [62] with respect to other adaptive DICs in terms of different metrics in Table 6. BD-rate indicates the rate increase of the proposed algorithm compared to the baseline algorithm for the same performance metric. A negative BD-rate indicates a better coding performance. Results show that HQVRC-CHENG is only inferior to HEVRC on NIQE. This is because HEVRC has a more complex decoder, which enhances the generative ability of GANs.

2) GENERALIZATION TO DIFFERENT DNN BACKBONES

Figure 4 shows that HQVRC-CHENG and HQVRC-MBT can implement adaptive DIC based on a single DNN model and achieve better performances on classification and perception than their backbone models CHENG2020 and MBT2018, respectively. Moreover, their performances on classification and perception are also better than those of BM-CHENG and BM-MBT, which are quantization-based adaptive DIC for image reconstruction context. Due to the superior backbone network and entropy model, HQVRC-CHENG surpasses HQVRC-MBT in all performance metrics. These results indicate that the proposed method is a general mechanism that can be successfully applied to different fixed-rate DIC models, such that the original

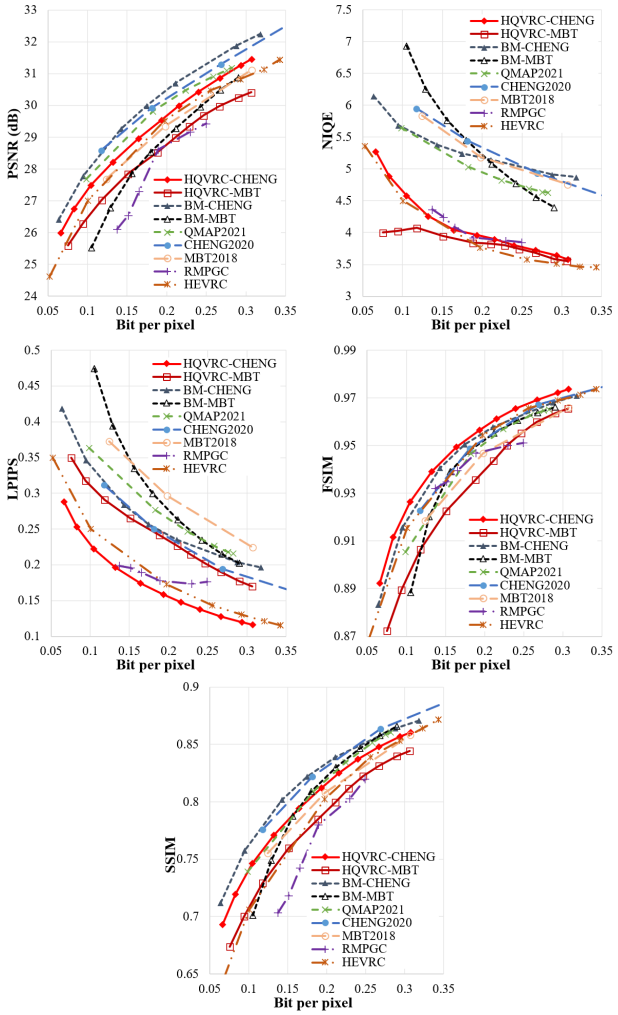


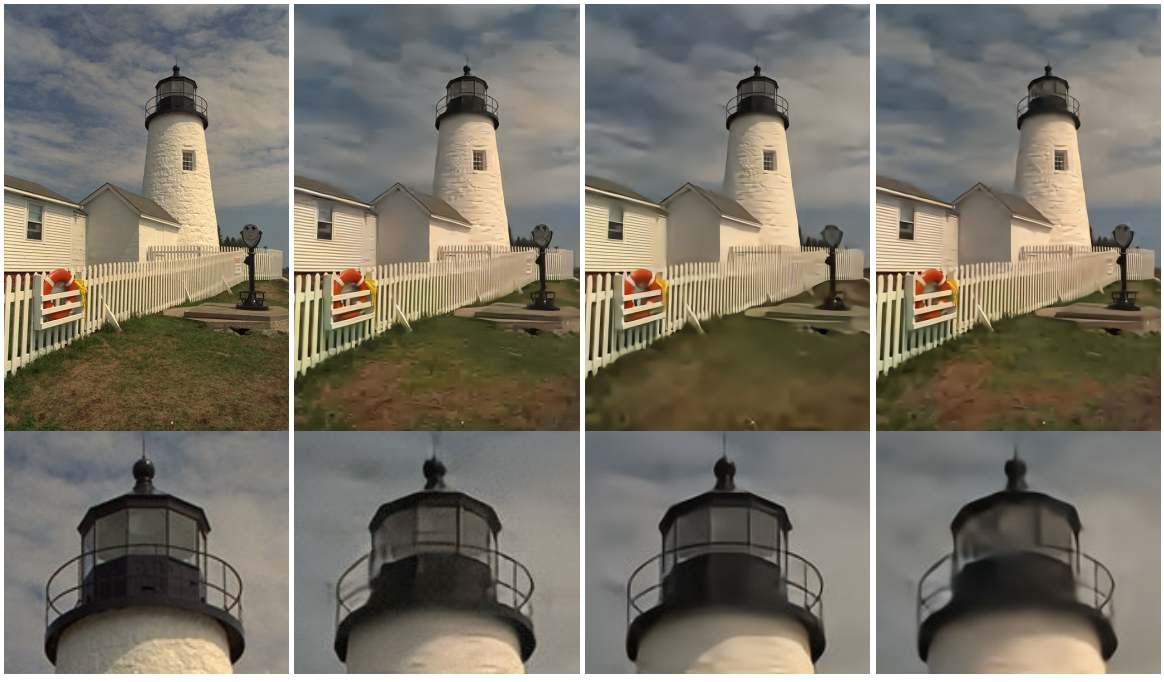
FIGURE 5. Perception, and distortion performances as functions of bpp for different image coding schemes using the Kodak dataset.

TABLE 7. Encoding time of quantization-based and encoder-based adaptive DIC.

Adaptive DIC	HQVRC-CHENG	BM-CHENG	HEVRC	QMAP2021
Time for encoding 1 rate	0.2047 s	0.1928 s	0.2796 s	0.2713 s
Time for encoding 10 rates	1.4862 s	0.8706 s	2.8897 s	3.0171 s

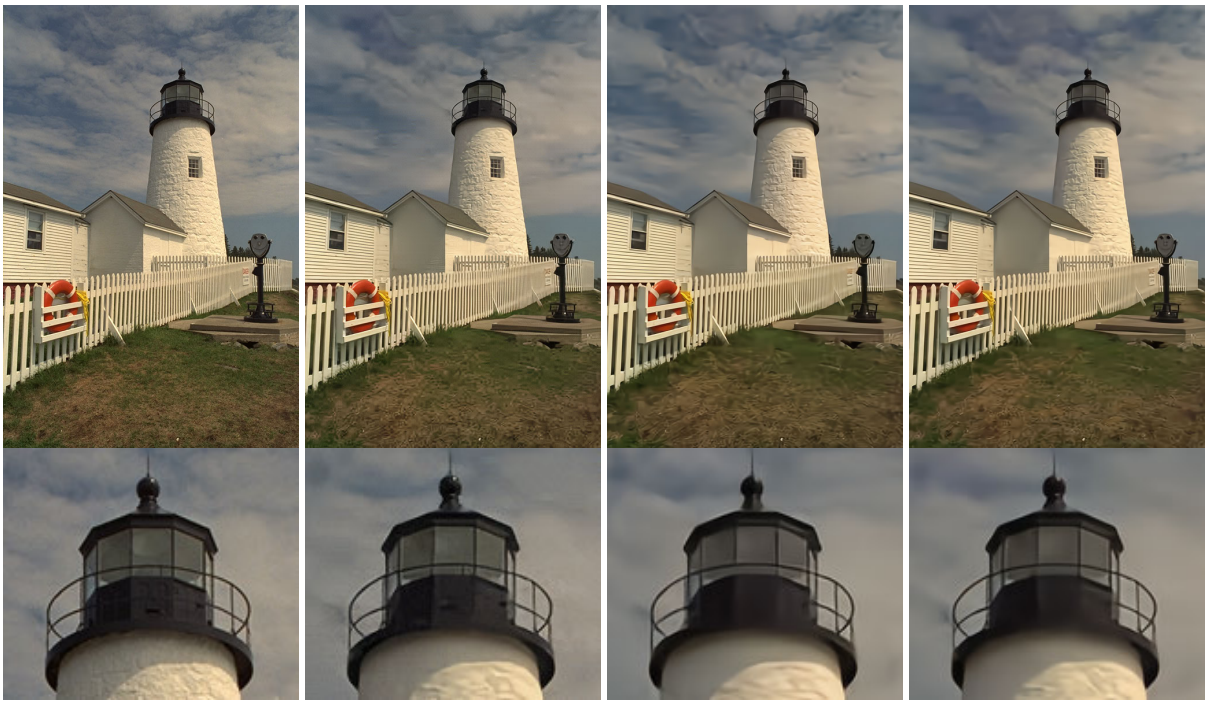
(a) Reconstructed images of Kodak-23 using different coding algorithms at 0.07 bpp

(b) Reconstructed images of Kodak-23 using different coding algorithms at 0.13 bpp



(c) Reconstructed images of Kodak-19 using different coding algorithms at 0.06 bpp

FIGURE 6. **Visual comparisons of different algorithms on the Kodak dataset.**



(d) Reconstructed images of Kodak-19 using different coding algorithms at 0.19 bpp

FIGURE 6. ***(Continued.)*** **Visual comparisons of different algorithms on the Kodak dataset.**

non-adaptive DICs can be upgraded into adaptive DICs for hybrid contexts.

3) GENERALIZATION TO DIFFERENT DATASETS

In performance comparison with other algorithms, the proposed models are trained using the COCO dataset and tested over the ImageNet dataset. Moreover, to better illustrate the visual effects, we test our models on the Kodak dataset as shown in Figure 5. It is shown that the proposed models yield consistent performance when immigrating to different datasets. This indicates the model's capability to generalize to different datasets.

4) RUN TIME PERFORMANCE

Unlike the encoder-based adaptive DIC, the quantization-based adaptive DIC can generate the latent vector only once for a given image and achieve context or rate adaptation by adjusting the quantization of the latent vector. We evaluate the time efficiency of our method by compressing 24 images from the Kodak dataset with a single rate and 10 different rates using different adaptive DIC methods. The encoding processes are executed on the same computational platform using one NVIDIA A40 GPU. The entropy-encoding time is excluded from the time measurement.

Table 7 compares the time consumption of HQVRC-CHENG, BM-CHENG, HEVRC, and QMAP2021. For encoder-based adaptive DICs HEVRC and QMAP2021, the time consumption to encode 10 rates is about 10 times the time consumption to encode a single rate, while it is only 5 times for HQVRC-CHENG and BM-CHENG. This result demonstrates that quantization-based adaptive DIC can save computational resources and reduce processing delays. Moreover, HQVRC-CHENG consumes more time than BM-CHENG due to the extra task of generating the ROI mask.

E. COMPARISON OF VISUAL EFFECTS

The above experimental results show that the implemented adaptive DIC algorithms for hybrid contexts have superior performance in terms of classification, distortion, and perception metrics. In this section, we select two typical images from the Kodak dataset, as shown in Figure 6, with clear classification targets of parrot and lighthouse, respectively. We compare the visual results of reconstructed images at different rates using HQVRC-CHENG, QMAP2021, and BM-CHENG to showcase the algorithm's superior performance.

When we compare the reconstruction quality of semantic salient regions in images coded at the same rate, HQVRC-CHENG and QMAP2021 outperform BM-CHENG, especially at lower rates. While BM-CHENG optimizes the overall image reconstruction quality, HQVRC-CHENG and QMAP2021 prioritize the reconstruction quality of semantic salient regions. By giving better reconstruction quality to these regions, they can improve the classification accuracy and provide better support for related tasks that require

confirmation of classification results by human eyes. Meanwhile, when we compare the reconstruction quality of semantic non-salient regions in reconstructed images using different coding algorithms, QMAP2021 has the worst visual effect at low rates, with large areas appearing blurred. This is because QMAP2021 lowers the coding quality of semantic non-salient regions due to its emphasis on coding quality based on semantic importance. In contrast, although HQVRC-CHENG prioritizes semantic salient regions, it uses a GAN-based generator as the decoder, which can mitigate the degradation of reconstruction quality in semantic non-salient regions. Finally, when we compare the overall perceptual quality of reconstructed images using different coding algorithms, both non-GAN coding algorithms show varying degrees of blurring at different rates, while HQVRC-CHENG always produces clear reconstructed images. This demonstrates the excellent ability of GAN in improving image perceptual quality.

V. CONCLUSION

In this paper, we propose a quantization-based adaptive DIC framework, which can be used to upgrade existing fixed-rate DIC models into rate-adaptive DIC for hybrid contexts. In this framework, the encoder generates the latent vector only once during the adaptive coding, and the rate and context adaptation is achieved by adjusting the quantization accuracy of the latent vector. The proposed framework can control the RDCP performance tradeoff via two mechanisms. First, a weighted sum of the overall image MSE loss, ROI MSE loss, SFM loss, and GAN loss is used to optimize the distortion, classification, and perception objectives simultaneously. Second, finer quantization is applied to semantic salient regions to prioritize the reconstruction quality of these regions. Experiments have shown that the proposed HQVRC-CHENG algorithm can improve both the classification and perception performance. Compared with the benchmark algorithm BM-CHENG, the average improvement in classification accuracy and NIQE are about 15% and 22%, respectively. Due to the tradeoff between perception and distortion, the average reduction of PSNR over BM-CHENG is about 1.5%. The overall results demonstrate that the proposed DIC scheme can effectively upgrade a fixed-rate DIC to be more adaptive, i.e., to achieve a desirable RDCP trade-off performance at different rates.

The proposed quantization-based adaption framework also has two limitations. First, because the latent vector is generated prior to rate-adaptive coding, the information loss introduced in the quantization phase may have undesirable impacts on complex high-level tasks such as image retrieval, object detection, and semantic segmentation. Moreover, the proposed ROI mask generation module is suitable for the single-label classification task but not multi-labels ones. Therefore, the latent representation learning and ROI mask generation process can be further improved for high-level tasks. For example, a promising direction for future work is to utilize ROI loss in both the image domain and feature domain for overall performance improvement.

REFERENCES

[1] Y. Hu, W. Yang, Z. Ma, and J. Liu, "Learning end-to-end lossy image compression: A benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 4194–4211, Aug. 2022.

[2] A. Habibi, "Hybrid coding of pictorial data," *IEEE Trans. Commun.*, vol. COM-22, no. 5, pp. 614–624, May 1974.

[3] G. K. Wallace, "The JPEG still picture compression standard," *Commun. ACM*, vol. 34, no. 4, pp. 30–44, Apr. 1991.

[4] C. Christopoulos, A. Skodras, and T. Ebrahimi, "The JPEG 2000 still image coding system: An overview," *IEEE Trans. Consum. Electron.*, vol. 46, no. 4, pp. 1103–1127, Nov. 2000.

[5] F. Bellard, *BPG Image Format*, document Release 0.9.8, Apr. 2018. [Online]. Available: https://bellard.org/bpg/

[6] S. Ma, X. Zhang, C. Jia, Z. Zhao, S. Wang, and S. Wang, "Image and video compression with neural networks: A review," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 6, pp. 1683–1698, Jun. 2020.

[7] D. Mishra, S. K. Singh, and R. K. Singh, "Deep architectures for image compression: A critical review," *Signal Process.*, vol. 191, Feb. 2022, Art. no. 108346.

[8] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 5–41, Jan. 2023.

[9] Y. Matsubara, R. Yang, M. Levorato, and S. Mandt, "Supervised compression for resource-constrained edge computing systems," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Waikoloa, HI, USA, Jan. 2022, pp. 923–933.

[10] J. Balle, V. Laparra, and E. P. Simoncelli, "End-to-end optimization of nonlinear transform codes for perceptual quality," in *Proc. Picture Coding Symp. (PCS)*, Nuremberg, Germany, Dec. 2016, pp. 1–5.

[11] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," 2018, *arXiv:1802.01436*.

[12] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Montreal, QC, Canada, Dec. 2018, pp. 10794–10803.

[13] X. Li and S. Ji, "Neural image compression and explanation," *IEEE Access*, vol. 8, pp. 214605–214615, 2020.

[14] R. Torfason, F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L. Van Gool, "Towards image understanding from deep compression without decoding," 2018, *arXiv:1803.06131*.

[15] S. Singh, S. Abu-El-Haija, N. Johnston, J. Ballé, A. Shrivastava, and G. Toderici, "End-to-end learning of compressible features," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 3349–3353.

[16] J. Cao, X. Yao, H. Zhang, J. Jin, Y. Zhang, and B. W. Ling, "Slimmable multi-task image compression for human and machine vision," *IEEE Access*, vol. 11, pp. 29946–29958, 2023.

[17] L. D. Chamain, F. Racapé, J. Bégaint, A. Pushparaja, and S. Feltman, "End-to-end optimized image compression for machines, a study," in *Proc. Data Compress. Conf. (DCC)*, Snowbird, UT, USA, Mar. 2021, pp. 163–172.

[18] Z. Lei, P. Duan, X. Hong, J. F. C. Mota, J. Shi, and C.-X. Wang, "Progressive deep image compression for hybrid contexts of image classification and reconstruction," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 72–89, Jan. 2023.

[19] Z. Chen, K. Fan, S. Wang, L. Duan, W. Lin, and A. C. Kot, "Toward intelligent sensing: Intermediate deep feature compression," *IEEE Trans. Image Process.*, vol. 29, pp. 2230–2243, 2020.

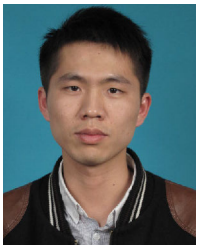
[20] J. Liu, H. Sun, and J. Katto, "Improving multiple machine vision tasks in the compressed domain," in *Proc. 26th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2022, pp. 331–337.

[21] Z. Yuan, S. Rawlekar, S. Garg, E. Erkip, and Y. Wang, "Feature compression for rate constrained object detection on the edge," 2022, *arXiv:2204.07314*.

[22] D. Huang, F. Gao, X. Tao, Q. Du, and J. Lu, "Toward semantic communications: Deep learning-based image semantic coding," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 55–71, Jan. 2023.

[23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2012.

[24] K. Grm, V. Štruc, A. Artiges, M. Caron, and H. K. Ekenel, "Strengths and weaknesses of deep learning models for face recognition against image degradations," *IET Biometrics*, vol. 7, no. 1, pp. 81–89, Jan. 2018.
[25] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," in *Proc. 8th Int. Conf. Quality Multimedia Exper. (QoMEX)*, Lisbon, Portugal, Jun. 2016, pp. 1–6.
[26] B. Li, L. Ye, J. Liang, Y. Wang, and J. Han, "Region-of-interest and channel attention-based joint optimization of image compression and computer vision," *Neurocomputing*, vol. 500, pp. 13–25, Aug. 2022.
[27] J. Xiao, L. Aggarwal, P. Banerjee, M. Aggarwal, and G. Medioni, "Identity preserving loss for learned image compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, New Orleans, LA, USA, Jun. 2022, pp. 516–525.
[28] X. Hong, J. Jiao, A. Peng, J. Shi, and C.-X. Wang, "Cost optimization for on-demand content streaming in IoV networks with two service tiers," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 38–49, Feb. 2019.
[29] D. Liu, H. Zhang, and Z. Xiong, "On the classification-distortion-perception tradeoff," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, BC, Canada, Dec. 2019, pp. 1204–1213.
[30] N. Le, H. Zhang, F. Cricri, R. Ghaznavi-Youvalari, and E. Rahtu, "Image coding for machines: An end-to-end learned approach," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Toronto, ON, Canada, Jun. 2021, pp. 1590–1594.
[31] H. Choi and I. V. Bajic, "High efficiency compression for object detection, in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Calgary, AB, Canada, Apr. 2018, pp. 1792–1796.
[32] S. Kim, Y. Lee, and K. Yoon, "Versatile video coding-based coding tree unit level image compression with dual quantization parameters for hybrid vision," *IEEE Access*, vol. 11, pp. 34498–34509, 2023.
[33] H. Akutsu and T. Naruko, "End-to-end learned ROI image compression," in *Proc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, Jun. 2019, pp. 4321–4325.
[34] Q. Wang, L. Shen, and Y. Shi, "Recognition-driven compressed image generation using semantic-prior information," *IEEE Signal Process. Lett.*, vol. 27, pp. 1150–1154, 2020.
[35] Y. Choi, M. El-Khamy, and J. Lee, "Variable rate deep image compression with a conditional autoencoder," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 3146–3154.
[36] Z. Sun, Z. Tan, X. Sun, F. Zhang, Y. Qian, D. Li, and H. Li, "Interpolation variable rate image compression," in *Proc. 29th ACM Int. Conf. Multimedia*, Virtual Event, China, Oct. 2021, pp. 5574–5582.
[37] Z. Cui, J. Wang, S. Gao, T. Guo, Y. Feng, and B. Bai, "Asymmetric gained deep image compression with continuous rate adaptation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10527–10536.
[38] S. Yin, C. Li, Y. Bao, Y. Liang, F. Meng, and W. Liu, "Universal efficient variable-rate neural image compression," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, May 2022, pp. 2025–2029.
[39] R. Su, Z. Cheng, H. Sun, and J. Katto, "Scalable learned image compression with a recurrent neural networks-based hyperprior," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 3369–3373.
[40] C. Han, Y. Duan, X. Tao, M. Xu, and J. Lu, "Toward variable-rate generative compression by reducing the channel redundancy," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 7, pp. 1789–1802, Jul. 2020.
[41] M. Song, J. Choi, and B. Han, "Variable-rate deep image compression through spatially-adaptive feature transform," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 2360–2369.
[42] Z. Lei, W. Zhang, X. Hong, J. Shi, M. Su, and C. Lin, "Conditional encoder-based adaptive deep image compression with classification-driven semantic awareness," *Electronics*, vol. 12, no. 13, p. 2781, Jun. 2023.
[43] T. Chen, H. Liu, Z. Ma, Q. Shen, X. Cao, and Y. Wang, "End-to-end learnt image compression via non-local attention optimization and improved context modeling," *IEEE Trans. Image Process.*, vol. 30, pp. 3179–3191, 2021.
[44] J. Ascenso, E. Alshina, and T. Ebrahimi, "The JPEG AI standard: Providing efficient human and machine visual data consumption," *IEEE MultimediaMag.*, vol. 30, no. 1, pp. 100–111, Jan. 2023.
[45] L. Duan, J. Liu, W. Yang, T. Huang, and W. Gao, "Video coding for machines: A paradigm of collaborative compression and intelligent analytics," *IEEE Trans. Image Process.*, vol. 29, pp. 8680–8695, 2020.
[46] A. de Santana Correia and E. L. Colombini, "Attention, please! A survey of neural attention models in deep learning," *Artif. Intell. Rev.*, vol. 55, no. 8, pp. 6037–6124, Mar. 2022.
[47] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 2921–2929.
[48] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, Oct. 2017, pp. 618–626.
[49] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Montreal, QC, Canada, Dec. 2014, pp. 2672–2680.
[50] J. Bégaint, F. Racapé, S. Feltman, and A. Pushparaja, "CompressAI: A PyTorch library and evaluation platform for end-to-end compression research," 2020, *arXiv:2011.03029*.
[51] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[52] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Lake Tahoe, NV, USA, Mar. 2018, pp. 839–847.
[53] Z. Cheng, H. Sun, M. Takeuchi, and J. Katto, "Learned image compression with discretized Gaussian mixture likelihoods and attention modules," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 7936–7945.
[54] 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 *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Zurich, Switzerland, Sep. 2014, pp. 740–755.
[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[56] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 248–255.
[57] R. Franzen. *Kodak Lossless True Color Image Suite*. Accessed: Mar. 3, 2023. [Online]. Available: http://r0k.us/graphics/kodak/
[58] A. Horé and D. Ziou, "Image quality metrics: PSNR vs. SSIM," in *Proc. 20th Int. Conf. Pattern Recognit.*, Istanbul, Turkey, Aug. 2010, pp. 2366–2369.
[59] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.
[60] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 586–595.
[61] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a," completely blind" image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.
[62] G. Bjontegaard, *Calculation Average PSNR Differences Between RDCurves*, document ITU-T VCEG ISO/IEC MPEG document VCEGMM33, Apr. 2001.



ZHONGYUE LEI (Student Member, IEEE) received the M.S. degree from the University of Electronic Science and Technology of China, Chengdu, China, in 2014. He is currently pursuing the Ph.D. degree with the School of Informatics, Xiamen University, Xiamen, China. His current research interests include source coding, semantic communication, and machine learning.

XUEMIN HONG (Member, IEEE) received the Ph.D. degree from Heriot-Watt University, Edinburgh, U.K., in 2008. He is currently a Professor with the School of Informatics, Xiamen University, China. He has published over 60 papers in refereed journals and conference proceedings. His current research interests include semantic communications, cognitive communication networks, and wireless localization systems.

JIANGHONG SHI (Member, IEEE) received the Ph.D. degree from Xiamen University, Xiamen, China, in 2002. He is currently a Professor with the School of Informatics, Xiamen University. He is also the Director of the West Straits Communications Engineering Center, Zhangzhou, China. His current research interests include wireless communication networks and satellite navigation systems.

MINXIAN SU is currently a Senior Engineer and the Leader of Research and Development at his company, where he works on information technology applications, such as big data analysis and artificial intelligence. His research interests and expertise include transportation big data processing, storage, and management, transportation analysis and forecasting, and intelligent algorithms. He is skilled at solving technical problems and inventing solutions. He has led the development of key systems, such as the transportation big data analysis application platform and the smart transportation cloud platform. He has participated in 28 scientific and technological information projects at the national, provincial, and municipal levels, obtained six national patents, received three provincial science and technology progress awards, five Xiamen science and technology progress awards, and several industry science and technology progress awards. He has also been recognized as a Xiamen Top Talent and a Xiamen Local Leading Talent.

CHAOHENG LIN is currently an Expert of software architecture design and development and transportation big data mining technology and applications. He is skilled in various development languages, databases, middleware, and distributed technology, and has excellent problem analysis and solving abilities. He also has some knowledge of blockchain, AI, and other fundamental principles. He obtained the PMP project management certification, in 2013. In 2018, his article "Multi-Vehicle Space-Time Conflict Warning System for Large Events" was accepted by the 13th China Intelligent Transportation Association.

WEI XIA was born in October 1984. He received the master's degree in software engineering from Xiamen University.

He is currently the Director and a Senior Researcher with the Cloud Desktop Research and Development Department, Fujian Centerm Information Company Ltd. He also represents the company as a member of the National Security Cloud Terminal Industry Alliance. He has been engaged in technical research and product development in the fields of cloud computing, virtualization, and cloud desktop for a long time. He has held various positions, such as the Technical Team Leader, a Researcher, and the Research and Development Director within the company. He is dedicated to creating highly available and secure cloud products. He has led teams to overcome core cloud platform technologies, such as intrinsic cloud high availability, host resource slicing, virtual network isolation protection, storage resource tampering prevention, and diversified situational awareness. This has formed a complete bottom-up controllable cloud technology system, realizing the full lifecycle management and protection of cloud resources. His unique image cache management mechanism (X-ICM), hybrid image format encoding (X-HIC), and resource dynamic adaptive algorithm have been applied for 11 patents.