# Progressive Deep Image Compression for Hybrid Contexts of Image Classification and Reconstruction

Zhongyue Lei, *Student Member, IEEE*, Peng Duan, *Student Member, IEEE*, Xuemin Hong, *Member, IEEE*, João F. C. Mota, *Member, IEEE*, Jianghong Shi, *Member, IEEE*, and Cheng-Xiang Wang, *Fellow, IEEE*

***Abstract*—Progressive deep image compression (DIC) with hybrid contexts is an under-investigated problem that aims to jointly maximize the utility of a compressed image for multiple contexts or tasks under variable rates. In this paper, we consider the contexts of image reconstruction and classification. We propose a DIC framework, called residual-enhanced mask-based progressive generative coding (RMPGC), designed for explicit control of the performance within the rate-distortion-classification-perception (RDCP) trade-off. Three independent mechanisms are introduced to yield a semantically structured latent representation that can support parameterized control of rate and context adaptation. Experimental results show that the proposed RMPGC outperforms a benchmark DIC scheme using the same GAN backbone in all six metrics related to classification, distortion, and perception. Moreover, RMPGC is a flexible framework that can be applied to different neural network backbones. Some typical implementations are given and shown to outperform the classic BPG codec and four state-of-the-art DIC schemes in classification and perception metrics, with a slight degradation in distortion metrics. Our proposal of a nonlinear-neural-coded and richly structured latent space makes the proposed DIC scheme well suited for image compression in wireless communications, multi-user broadcasting, and multi-tasking applications.**

***Index Terms*—Deep image compression, progressive compression, generative image compression, image semantics**

## I. Introduction

Image compression is vital to the digital and green society, as transmitting visual information consumes a major part of the telecommunication resources. In many visual applications, the reconstruction of a compressed image is an intermediate step within a specific, high-level task. For example, in machine-oriented tasks such as target recognition, the goal is to locate and track a target in a sequence of images with the best possible accuracy. In audiovisual production, the goal is to produce perceptually authentic and pleasing images. In essence, taking the application context into account allows one to prioritize the information that is preserved during the compression and transmission of an image. This motivates the development of task-oriented and semantic-aware image compression.

Deep image compression (DIC) [1], [2] has recently attracted significant research attention as a new paradigm to compress images using deep neural network (DNN). A major advantage of DIC over conventional schemes lies in its potential to achieve very low-rate compression under certain contexts. By "context", we mean how an image is used by its recipient for a specific purpose, e.g., visual recording/surveillance, target detection, or entertainment. Similar to conventional image compression standards [3] such as JPEG, BPG, and JPEG2000, the DIC framework is composed by structured modules such as latent representation, quantization, rate estimation, entropy codec, and reconstruction [4]. A key advantage of DIC over traditional compression schemes lies in its flexibility of optimizing latent representations according to the application context. More specifically, traditional image compression relies on expert-crafted latent representation modules, which are difficult to adapt to different contexts. In contrast, DIC can effectively learn latent features that are most important for a given task/context and, thus, preserve them better during lossy compression [5].

In real-time communications, a DIC codec is required to quickly adapt to varying channel capacities (i.e., rate-adaptation) and diverse contextual feedback (i.e., context adaptation). While rate adaptation can be implemented either within the latent representation [6]–[16] or within the quantization [17]–[22] modules in a DIC, context adaptation is mainly implemented within the former module. This paper focuses on the latent representation module for two reasons. First, rate and context adaptations can be jointly optimized; second, representation learning benefits tremendously from using deep learning techniques [23].

Compared with classic image compression schemes, a major disadvantage of DIC is the high computational power required to train a DNN-based codec, which may include millions of parameters. The computational challenge becomes particularly prominent in scenarios that require fast adaptation. The choice of the DIC architecture is thus critical to address such a challenge. Depending on how the latent representation adapts to rate and contextual change, we distinguish two types of DIC architectures in the literature: encoder-based [6]–[9] and latent-based [10]–[16]. The former encodes a different representation each time the feedback condition changes, while the latter

This work is supported by the National Natural Science Foundation of China (Grant No. 62077040), the Science and Technology Key Project of Fujian Province, China (No.2019HZ020009), and the Xiamen Special Fund for Marine and Fishery Development (21CZB011HJ02). (*Corresponding author: Xuemin Hong.*)

Zhongyue Lei, Peng Duan, Xuemin Hong, and Jianghong Shi are with the School of Informatics, Xiamen University, Xiamen 361000, China (e-mail: leizhongyue@stu.xmu.edu.cn; duanpeng@stu.xmu.edu.cn; xuemin.hong@xmu.edu.cn; shijh@xmu.edu.cn).

João F. C. Mota is with the School of Engineering and Physical Sciences, Heriot-Watt University, Edinburgh EH14 4AS, U.K. (e-mail: j.mota@hw.ac.uk).

Cheng-Xiang Wang is with the National Mobile Communications Research Laboratory, School of Information Science and Engineering, Southeast University, Nanjing 210096, China, and also with Purple Mountain Laboratories, Nanjing 211111, China (e-mail: chxwang@seu.edu.cn).

encodes a structured latent representation, which can be used later for fast adaptation to various feedback conditions. The "encode once and for all" property of the latent-based architecture avoids multiple calls to the DNN encoder function, making it much more attractive for scenarios with fading wireless channels [24], [25], multi-user broadcasting, and multi-tasking applications. In this paper, we focus on latent-based DIC.

**Problem statement.** While most works on DIC have focused on single context, usually either image reconstruction [4], [26] or image classification [5], [27], *our goal is to design compression schemes that work in hybrid contexts*. Image compression in hybrid contexts should enable the reconstructed image to obtain both high visual quality and high performance in downstream artificial intelligence (AI) tasks. The schemes should also be rate adaptive, in the sense that compression at different rates should require no retraining. Moreover, the DIC should have flexible and configurable mechanisms to realize progressive transitions between different contextual goals in a resource-limited scenario.

**Contributions.** To our best knowledge, the problem of latent-based rate adaptive DIC for hybrid contexts has not yet been studied in the literature. This is a non-trivial problem because a careful design is required to match the high-level contextual goals with low-level progressive coding mechanisms. More importantly, the encoder should be guided to learn a structured latent representation that is effective for both rate and context adaptation. Specifically, the contributions of this paper are as follows.

First, our paper is among the first to state and study the problem of learning a communication-friendly latent representation for progressive DIC. We introduce a DIC scheme called residual-enhanced mask-based progressive generative coding (RMPGC), which combines a spatial-mask mechanism, residual-based layering, and contextual importance estimators in the latent space. To our best knowledge, RMPGC yields the richest latent structure in the DIC literature to date. Such a structure can provide state-of-the-art flexibility to encode an image with variable bit rates and different contextual goals. Experiments show that using the same generative DNN backbone, RMPGC outperforms a benchmark DIC scheme called variable rate generative coding (VRGC) [16] in all tested metrics related to classification accuracy, reconstruction fidelity, and perceptual quality. Moreover, RMPGC is a flexible framework that can be applied to different DNN backbones. We illustrate this by applying RMPGC to an advanced DNN backbone. The resulting scheme outperforms the classic BPG codec and four state-of-the-art DIC schemes in classification and perception metrics, however, with a slight degradation in distortion metrics. This indicates the value of the proposed latent space design as a general and transferable technique.

Second, our paper makes an initial attempt to systematically investigate the rate-distortion-classification-perception (RDCP) trade-off, which is fundamental in hybrid context DIC. We apply a model-based approach to investigate the RDCP trade-off empirically. Given the proposed DIC model, we reduce the problem of finding optimal encoder-decoder pairs to finding optimal hyper-parameters in our model. We then propose and investigate three independent approaches for parameterized control of the RDCP trade-off: loss function weighting, spatial mask blending, and layered rate splitting. Empirical studies show that there are regions in the hyper-parameter space where distortion and classification metrics can be jointly improved. This suggests that the strict distortion-classification-perception (DCP) trade-off proved in [28] is not applicable to DIC due to a critical change of problem formulation, which further involves encoder optimization and rate constraint. Studies show that the proposed RMPGC codecs have a hyper-parameter space that is effective to maneuver the RDCP trade-off.

**Organization.** The remainder of this paper is organized as follows. Related work is introduced in Section II. Section III discusses the RDCP trade-off and provides a high-level description of our approach. Section IV introduces the proposed DIC methods and implementations, followed by experimental results and discussions in Section V. Finally, conclusions are drawn in Section VI.

## II. Related work

Table I provides an overview of existing studies on DIC and positions our paper in the literature. There is a wealth of literature for reconstruction-oriented image coding, therefore only representative and recent works are listed. Moreover, rate-adaptive coding solely for the context of classification does not attract much research attention because the code rate is already very low. A brief review of the related literature is given below.

**Single context DIC.** The most widely-studied context is image reconstruction, where DIC is optimized for pixel-level mean-squared error (MSE) [29]–[32]. This is a general-purpose context that aims to maximize the overall image fidelity and treats each pixel as equally important. For image classification (by machines), it has been shown that DIC can obtain extremely low rates by learning to encode the most discriminative features that contribute most to classification accuracy [27]. The classification-related information is often called semantics of an image [33]. For the context of image perception, low rates are shown to be attainable by DIC with generative models [26]. Popular generative models include generative adversarial nets (GANs) [34] and variational auto-encoders

TABLE I
SUMMARY OF RELATED DIC LITERATURE.

| DIC methods \ Contexts | | Reconstruction fidelity | Classification semantic | Hybrid contexts |
|---|---|---|---|---|
| **Fixed rate** | | [1], [2], [4], [26], [29]–[31], [36]–[38] | [5], [27], [42] | [43]–[45] |
| **Rate adaptive (Quantization)** | | [17]–[22] | Less attractive problem | – |
| **Rate adaptive (representation)** | Encoder-based | [6], [7] | | [8], [9] |
| | Latent-based (Spatial mask) | [16] | | Our contributions (RMPGC) |
| | Latent-based (Layers) | [10]–[15] | | |

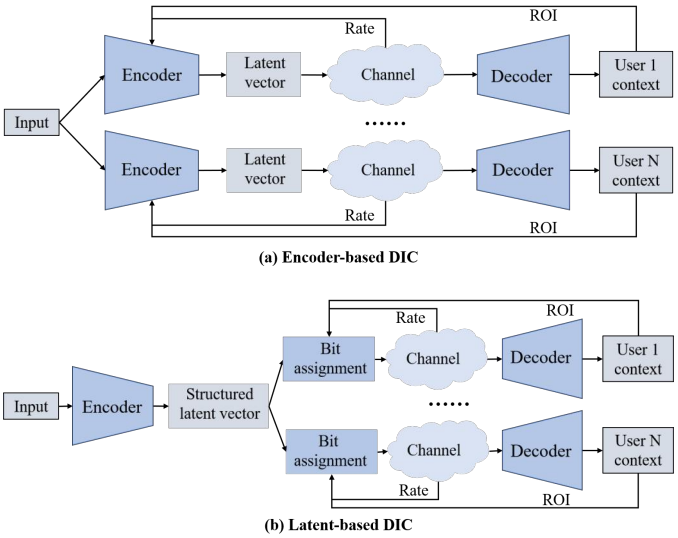



Fig. 1. Two architectures of rate adaptive DIC: (a) encoder-based architecture; (b) latent-based architecture.

(VAEs) [35]. In particular, GANs [36]–[39] have emerged as a promising framework for image compression to produce perceptually realistic images with sharp features. The potential of GAN-based image compression goes beyond improving image perception, as studies show that GANs can learn latent representations that are disentangled and semantically editable [40], [41].

**Hybrid context DIC.** DIC with hybrid contexts is an important research area that can maximize the potential utility of a compressed image. For example, in many target recognition applications, only encoding abstract features for image classification (e.g., [5], [42]) is insufficient. The receiver should also be able to reconstruct the image when necessary, so that it can be inspected by humans or used by other algorithms. Different contexts generally require different, and often competing, strategies in the lossy compression process when deciding on the priority of the information to be encoded. Particularly, a common challenge in many applications is to strike a balance between image classification and reconstruction. Some pioneering works of hybrid-context DIC were introduced in [43]–[45] to jointly optimize the classification accuracy and reconstruction fidelity. The class activation mapping (CAM) [46]–[50], which identifies class-specific image regions, was adopted in [8], [9], [45], [51] to assign more bits to semantically-salient areas. The GAN-based compression framework was initially introduced in [9] for hybrid-context DIC.

**Rate-adaptive DIC.** Extending fixed-rate DICs into rate-adaptive DICs is not a trivial task. This is because, in fixed-rate DIC, each rate is associated with a different DNN model [10]. Pipelining multiple DNN models for variable rates is cumbersome and clearly undesirable. The goal of rate-adaptive DIC is to build a compact and unified DNN model that has an explicit mechanism for rate control. Rate-adaptive DICs can be further classified into two paradigms: quantization-based and representation-based.

A straightforward approach to building rate-adaptive DIC is via quantization. Adaptive quantization techniques, widely used in conventional source coding, have been modified and adapted for DIC. The DICs proposed in [17], [18] scale the latent vector representation with different coefficients, such that the representation vectors fall into different quantization intervals. Similarly, [19], [20] proposed to manipulate the Lagrange multiplier and quantization intervals to obtain variable rates. A technique called nested quantization was introduced in [21], [22]. Quantization-based rate adaptation techniques, however, do not promote in-depth optimization of the latent presentation and cannot support context-aware adaptation in general.

**Rate-and-context adaptive DIC.** For a DIC to be jointly adaptive in terms of rate and context, the representation-based DIC paradigm is preferred. As illustrated in Fig. 1, there are two basic architectures for rate-and-context adaptive DIC: encoder-based and latent-based. The encoder-based architecture has recently been extended to address the hybrid-context problem [8], [9]. The approach therein is closely related to the well-established region-of-interest (ROI) compression methods. The drawback of this approach is that a new latent representation should be coded each time the rate or context condition changes, therefore adaption brings extra costs in terms of computational load and processing delay.

Our paper focuses on the latent-based DIC architecture [10]–[16], which encodes a structured latent representation once. Adaptation to rate and context can be achieved by manipulating data in the structured latent space. Latent-based DIC can be categorized into two types: layer-based and spatial mask-based. The first type [10]–[15] utilizes residual networks to transform the original image into a layered latent vector with ordered dependency. Rate variation is then achieved by adjusting the number of layers for transmission. The second type [16] exploits the redundancy of the latent representation in different spatial channels. Spatial masks are then used to modulate the channels given a target rate. However, all these studies on latent-based DIC are restricted to a single type of context: image reconstruction. To our best knowledge, the problem of joint rate-and-context adaptation for multiple contexts has not been studied for latent-based DIC.

## III. Goal and Intuition of Progressive DIC Design

### A. Communication-friendly latent representation for DIC

One goal of this paper is to promote a research direction called communication-friendly latent representation learning, which lies in the intersection of the fields of communication engineering and neural representation learning [52]. The term communication-friendly has two implications. First, it means the latent representations should facilitate the encapsulation of contextually-prioritized information into separate data packets, such that contextual priority and transmission priority can be aligned. This is related to the effectiveness of end-to-end communications. Second, it means the decoding of latent representations should be robust to bit errors and lossy channels.

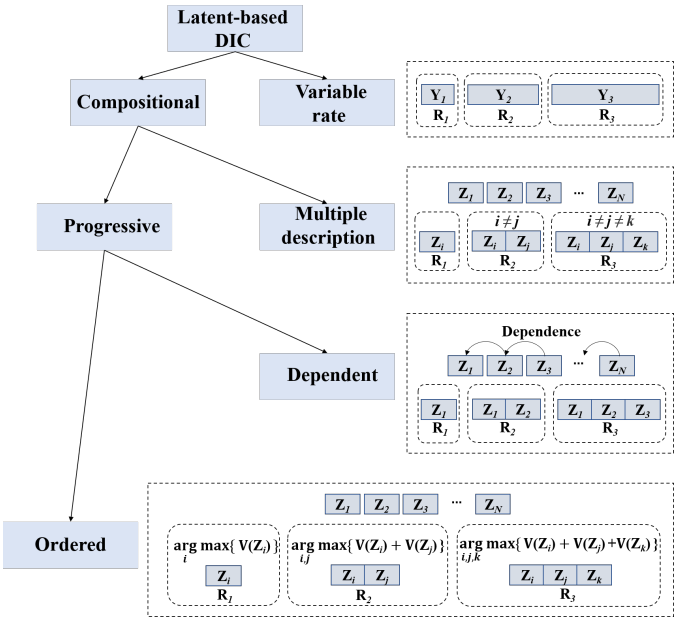



Fig. 2. Classification of latent-based variable-rate DIC schemes and illustration of their rate adaptation mechanisms ($Z_i$ denotes different pieces of data after encoding, $R_i$ denotes different rates of data packet encapsulation, $V(Z_i)$ is a function that marks the importance of data piece $Z_i$ with respect to certain context).

This is a reliability problem. In this paper, we mainly focus on the first aspect, i.e., effectiveness.

Fig. 2 illustrates the rationale of different types of communication-friendly coding. Latent-based DIC can be classified into two types: variable-rate encoding [17]–[20] and compositional encoding [13]–[16], [21], [22], [53]. In variable-rate encoding, the encoder generates latent vectors (denoted as $Y_n$ in Fig. 2) were designed to be received integrally by the decoder. Such encoding schemes [17]–[20] are mainly obtained by applying different quantization techniques to the latent vector. The emphasis of these techniques is then on quantization, not on the structure of the latent representation.

Compositional coding, in contrast, focuses on designing structured latent space vectors. A latent vector is composed of a hierarchy of subvectors $Z_n$, each of which is characterized by its own data rate, dependence on other subvectors, and relative importance to different contextual goals. In other words, an image is represented into small pieces of interrelated data. This coding paradigm typically yields incremental performance when more pieces of data are available for decoding.

Compositional encoding can be further categorized into different types. If the data pieces $Z_n$ have homogeneous importance, we have multiple description coding [53]. On the other hand, if the data pieces have ordered importance, we have progressive coding (in a wide sense). Based on whether a data piece relies on other data pieces for decoding, we can further distinguish two types of progressive coding: in layered (e.g., residual-based) image representation [13]–[15], [21], [22], there are strict dependencies among the latent subvectors, with the higher-level subvectors depending on the decoding of the lower-level ones. The second type does not have such a dependency, but still marks different data pieces for their contextual importance [16].

Our paper falls within the scope of progressive DIC, which can bring a number of benefits to context-aware communications. First, from the perspective of source coding, the latent representation can be designed to adapt to different contextual goals. This means contextually-relevant information is better preserved in lossy compression. Second, from the perspective of reliable transmission, the source and channel coding can be jointly optimized by carefully matching the source importance with different tiers of channel coding. This means contextually-relevant information is better protected against lossy channels. Finally, from the perspective of transmission delay, data pieces with higher importance have higher priority in streaming or packet scheduling. This means the perceived transmission delay can be reduced. In short, a compositional and ordered latent presentation in progressive DIC is communication-friendly.

### *B. The RDCP trade-off*

The existence of a distortion-classification/perception trade-off is well-known in image processing. A previous theoretical study in [28] established a strict DCP trade-off. We first clarify that such a conclusion does not apply to DIC because only decoder optimization was considered therein (see Eqn. (7) in [28]). In essence, [28] proved that once the latent vector is determined, there is a fundamental trade-off among DCP in the decoding process. This is because the degrees of freedom in designing the latent space are ignored.

Unlike the optimization problem formulated in [28], DIC involves joint optimization of the encoder and decoder. Furthermore, in DIC, the encoding rate is expected to have a global impact on DCP metrics. Thus, if besides distortion, classification, and perception, we take into account the encoding rate, we have an RDCP trade-off problem, which not only generalizes the DCP trade-off studied in [28], but also the classical rate-distortion (RD) trade-off in lossy compression [54].

### *C. Assumptions and high-level approach*

Navigating the optimal boundary of the RDCP trade-off entails searching for optimal encoder-decoder pairs, which is a difficult problem. To make an empirical study feasible, we make two simplifying assumptions. First, we use a model-based approach, so that the original optimization problem is reduced to searching for optimal parameters in the proposed DIC model. We will show later that our model facilitates a structured search of the feasible space. Second, we note that achieving high perceptual quality alone (measured as the divergence from the natural image distribution) does not seem to impose any rate requirements. As a result, when considering bit rate allocation policies, we will focus on the trade-off between distortion and classification. In summary, our methodology for exploring the RDCP trade-off space consists of three steps: 1) choosing an encoding rate R; 2) using some hyper-parameters to generate models with different rate splitting policies between distortion and classification; 3) training DNN models with weighted loss related to DCP.

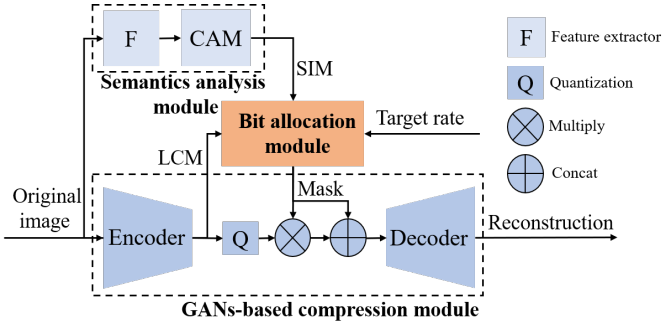



Fig. 3. Framework of the proposed progressive DIC scheme.

From a practical perspective, our goal is to navigate the RDCP trade-off via a parameterized model. Ideally, the semantically-salient and/or reconstruction-demanding areas in the image should be represented and encoded with higher priority and more bits. The remaining areas, which are contextually less important, may suffer from high distortion. A GAN network is then used to counterbalance such a distortion by producing perceptually authentic images.

### D. Framework and design intuitions

Fig. 3 shows the framework of our proposed progressive DIC scheme for hybrid contexts. The framework consists of three modules: a semantics analysis module, a GAN-based image compression module, and, within it, an adaptive bit allocation module. The GAN-based DIC enables reconstructed images to have higher perception quality, leveraging the ability of the GAN to match the statistics of the reconstructed image to the input image. The semantics analysis module classifies the input image and uses CAMs to get the input image's semantic importance heatmap corresponding to the predicted class. Then, the obtained importance heatmap is resized to the same size as the latent vector and defined as the semantic importance map (SIM). The encoder of the GAN-based image compression module transforms the input image into a latent vector, which can interact with SIM to enable semantics-aware DNN training. Moreover, during compression, a map called latent vector complexity map (LCM) is generated according to the latent vector. The LCM is calculated as the variance vector of the latent vector/tensor along the channel dimension. Based on the SIM, LCM, and target rate, the bit allocation module generates a binary mask, which is then applied to the latent vector. Finally, the latent vector is quantized, concatenated with the mask, and converted into bit streams. In the decoder, the latent vector and mask are recovered from bit streams and then jointly used to reconstruct the image.

We note that the resizing procedure to obtain SIM is essentially an approximation to an ideal function that can mark the semantic importance of each entry in the latent vector. As we will show later, this approximation is effective in practice for two reasons. First, due to the spatial invariant property of the convolutional neural network (CNN) [23], [56], the latent vector retains the spatial characteristics of the original image. Therefore, much of the semantic importance knowledge revealed by the heatmap can be transferred to the latent vector space by resizing operation. Second, during DNN training, the SIM is used as a conditional input to the generator and discriminator in GAN. This will encourage the network to learn to encode semantic salient information according to SIM.

For concreteness, we focus on two contextual goals: image classification and reconstruction. To achieve progressive behavior in terms of both bit rates and performance on these context, we introduce three different mechanisms in our design. The first mechanism consists of a suitable loss function for DNN training. This has a global effect to obtain two contextual goals. The second mechanism applies spatial masks to the image as a visual attention mechanism. This yields explicit progressive behavior when combined with different bit allocation strategies. The third mechanism relies on layered multi-scale representation to strike a balance between the two contextual goals of image reconstruction and classification. The rational for these three mechanisms are explained below.

*1) Loss function design and semantic feature matching:* Our GAN-based image compression network consists of an encoder $E$, a quantizer $Q$, a decoder $G$ (in GAN-based DIC, the generator acts as a decoder), a discriminator $D$, and a rate-distortion formulation. The encoder compresses an image $\mathbf{x} \in \mathbb{R}^{H\times W\times 3}$ into the latent vector $\mathbf{z}$, which is in turn quantized to $\hat{\mathbf{z}}$ and sent to the decoder to be reconstructed as image $\mathbf{y}$. In the training stage, $\mathbf{x}$ and $\mathbf{y}$ are input to the discriminator to calculate the discriminator loss. The input image is compressed by a factor of $K$ to a latent vector with size $\frac{H}{K} \times \frac{W}{K} \times C$, where $H$ and $W$ denote the input image's height and width, respectively, and $C$ is the number of channels of the latent vector.

The conventional loss functions used in GAN-based compression schemes include the MSE loss and the GAN loss. To encourage the encoding of semantic information into the latent representation, we add to the loss function a semantic feature matching (SFM) term $\mathcal{L}_{\text{SFM}}$, which measures the matching degree between the features extracted from the original and recovered image. Therefore, the distortion term $\mathcal{L}(\mathbf{x}, G(\hat{\mathbf{z}}))$ used for DNN model training includes three loss functions: $\mathcal{L}_{\text{SFM}}$, $\mathcal{L}_{\text{MSE}}$, and $\mathcal{L}_{\text{GAN}}$, which correspond to metrics for classification, reconstruction, and perception, respectively. The SFM loss is

$$\mathcal{L}_{\text{SFM}} = \frac{1}{H_\xi W_\xi C_\xi} \sum_{i=1}^{H_\xi} \sum_{j=1}^{W_\xi} \sum_{k=1}^{C_\xi} \parallel \xi(\mathbf{x})_{i,j,k} - \xi(\mathbf{y})_{i,j,k} \parallel^2 , \tag{1}$$

where $\xi(\cdot)$ is a pre-trained feature extractor in the semantics analysis network. Here, $H_\xi \times W_\xi \times C_\xi$ is the size of the feature extracted from $\mathbf{x}$ and $\mathbf{y}$.

*2) Spatial mask and bit allocation:* One way to achieve progressive encoding is to apply masks to the latent vector. To highlight the fact that the latent vector typically preserves the spatial structure of the original image, we call any mask applied to the latent vector a *spatial mask*. Different masks can be generated according to different contextual goals. As previously explained with Fig. 3, the semantics analysis module outputs a semantics importance map (SIM), which indicates the semantic importance of each element in the

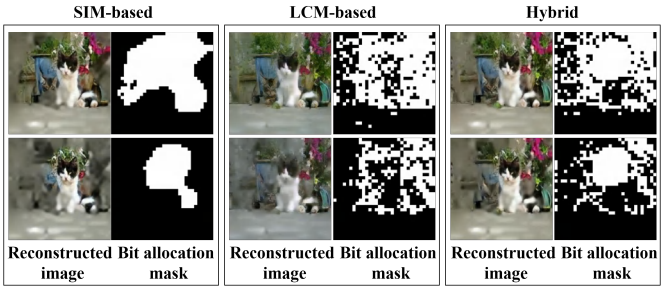



Fig. 4. Illustration of spatial mask-based bit allocation strategies. The white dots in the mask indicate areas where more bits are allocated; the first row has a higher rate than the second row.

latent vector. The SIM is essentially a spatial mask optimized for image classification. On the other hand, the LCM, which measures (or captures) the entropy or complexity of a spatial region, is essential for image reconstruction.

Fig. 4 illustrates the difference between SIM and LCM generated from the same image, as well as their impact on the reconstructed image. We can see that SIM focuses on the cat head, which contains the relevant information for classification, but LCM focuses on the flower background and texture, which are more important for reconstruction. Moreover, SIM and LCM can be blended to yield a certain trade-off between image classification and reconstruction. Once the spatial mask is complemented by a bit allocation module that assigns bits according to values in the mask, we can have a progressive encoding mechanism with explicit behavior.

*3) Residual-based layered representation:* As illustrated in Fig. 4, given a latent representation without layered structure, the spatial mask approach tends to yield tunnel-visioned images at low encoding rates. To overcome this drawback, we adopt a residual-based approach to produce a multi-scale latent vector with a layered structure. As shown in the left column of Fig. 5, multiple layers can be progressively stacked in the latent space. Each upper layer can be seen as a residual solution between a coarser and a finer scale [55]. As a result, the bottom layers are used to reconstruct an approximation

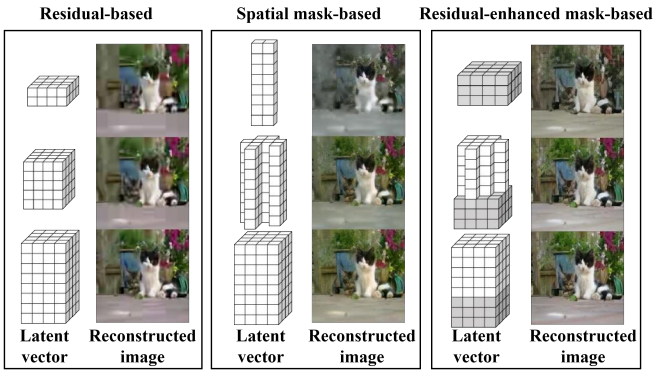



Fig. 5. Illustration of the latent representations of mask-based and/or residual-based DIC schemes. The encoding rate increases progressively from the first row to the third row. In the residual-based DIC scheme, extra available bits are used to add more layers in the latent space. In the spatial mask-based DIC scheme, extra bits are used to better encode spatially important regions.

of the original image, while the addition of each upper layer delivers gradual refinements to the image. However, the residual-based approach is applied to the entire image without spatial attention.

In this paper, we propose to unify the layered and spatial-masking approaches to yield the residual-enhanced mask-based progressive generative coding (RMPGC). The layering stacks in RMPCG are further partitioned into base layers and enhance layers. This partition is guided by a hyper-parameter $C_b$, which specifies how many lower layers are taken as the base layer. A single spatial mask is then applied to the enhance layers (upper layers) of the latent representation, while the base layers are kept intact (i.e., not subject to the spatial mask). We note that such a partition implies a two-phase rate allocation strategy: the first phase is to allocate a minimum rate to the base layer using hyper-parameter $C_b$, followed by the second phase of finer-grain rate allocation using spatial masks.

The above partition is a practical and deliberate simplification of a multi-level layer-mask blending policy, which assigns a different spatial mask to different layers. Such a multi-level policy is more flexible in theory, but suffers from two critical drawbacks. First, the DIC model is difficult to train and converge. Second, the need to transmit multiple masks as side-information will result in significant overheads. Possible extensions to a multi-level RMPGC are left for future work. Moreover, we remark that the three-dimensional latent structure of RMPGC is different from the structure illustrated in [10], which uses multi-level quantization instead of spatial attention.

*4) Parameters for rate control:* Based on the above proposed DIC design, the encoded bit rate after latent representation and quantization (and before applying lossless adaptive arithmetic coding) is controlled by a number of parameters. First, there are some hyper-parameters that jointly determine the range of encoding rates, including the downsampling rates $K$ of the original image (which decides the length $H/K$ and width $W/K$ of the 3D latent space as illustrated in Fig. 9), the number of channels $C$ in the latent vector (which decides the height of the 3D latent space), the base layer partition parameter $C_b$, and the quantization level $L$. These hyper-parameters are decided before training a DIC model to bound the range of encoding rates to be

$$R_{\min} = C_b HWK^{-2} \log_2 L, \tag{2}$$

$$R_{\max} = CHWK^{-2} \log_2 L. \tag{3}$$

Within this range, the previously introduced spatial masking techniques are used to achieve rate transitions with a minimum step of $(C - C_b) \log_2 L$ bits. We note that the above rate calculations are approximations that do not take into account the signaling overheads used to encode mask information.

Within our proposed framework, the above three progressive mechanisms can be applied separately or jointly to yield different implementations of progressive DIC schemes. We will subsequently focus on two representative DIC implementations: mask-based progressive generative compression (MPGC) and RMPGC. The latter can be seen as a generalization of the former.

## IV. Implementations of the Proposed Progressive DIC

### A. The generative compression network

In our proposed framework, a least square GAN (LSGAN) [57] is adopted as the generative compression network. During training, we use the multi-scale discriminator [58]. The multi-scale discriminator and generator are trained in an alternating manner. The discriminator is trained with the following objective function

$$\mathcal{L}_{\mathrm{D}} = \min_{D_1,D_2,D_3} \mathbb{E}_{\mathbf{x}}\{\sum_{k=1}^{3}[(D_{\mathrm{k}}(\mathbf{x})-1)^2 + (D_k(G(\hat{\mathbf{z}}))^2]\}, \quad (4)$$

where $\mathbb{E}_{\mathbf{x}}$ is the empirical mean over a batch, and $D_{\mathrm{k}}$ is the $k$th scalar discriminator.

Following the discriminator training, the encoder and decoder are trained to solve the rate-distortion optimization problem given by

$$\min_{E,G} \mathbb{E}_{\mathbf{x}}\{\mathcal{L}(\mathbf{x}, G(Q(E(\mathbf{x})))) + \lambda_{\mathrm{R}} R(\hat{\mathbf{z}})\}, \quad (5)$$

where $R(\hat{\mathbf{z}})$ is the rate of the original image's latent representation, which will be explained below. The distortion term $\mathcal{L}(\mathbf{x}, G(Q(E(\mathbf{x}))))$ is the weighted sum of the MSE loss $\mathcal{L}_{\mathrm{MSE}}$ and GAN loss $\mathcal{L}_{\mathrm{GAN}}$ [58], which are respectively defined as

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{3HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{3} \parallel \mathbf{x}_{i,j,k} - \mathbf{y}_{i,j,k} \parallel^2, \quad (6)$$

$$\mathcal{L}_{\mathrm{GAN}} = \sum_{k=1}^{3}(D_k(G(\hat{\mathbf{z}})) - 1)^2. \quad (7)$$

The final distortion loss $\mathcal{L}(\mathbf{x}, G(\hat{\mathbf{z}}))$ is given by the weighted sum of these loss terms, with the weights indicating competing priorities of different contextual goals:

$$\mathcal{L}(\mathbf{x}, G(\hat{\mathbf{z}})) = \min_{E,G} \mathbb{E}_{\mathbf{x}}\{\lambda_{\mathrm{SFM}}\mathcal{L}_{\mathrm{SFM}} + \lambda_{\mathrm{GAN}}\mathcal{L}_{\mathrm{GAN}} + \lambda_{\mathrm{MSE}}\mathcal{L}_{\mathrm{MSE}}\}, \quad (8)$$

where $\lambda_{\mathrm{SFM}}$, $\lambda_{\mathrm{GAN}}$, and $\lambda_{\mathrm{MSE}}$ are weights of SFM loss, MSE loss, and GANs' generator loss, respectively.

To adopt the GAN framework for encoding, additional treatments are required for relative rate estimation and quantization. The element of an image's latent representation is denoted as $z^l_{i,j}$, where $i \in \{1, ..., \frac{H}{K}\}$, $j \in \{1, ..., \frac{W}{K}\}$ and $l \in \{1, ..., C\}$ are the height, width, and channel index, respectively. Given $i$ and $j$, we represent the empirical variance of the latent vector along the channel dimension as

$$v_{i,j} = C^{-1}\sum_{l}(z^l_{i,j} - C^{-1}\sum_{l} z^l_{i,j})^2. \quad (9)$$

As the channel variance $\mathbf{v}$ measures the complexity of the corresponding latent vector, we will use it to determine the rate of $\hat{\mathbf{z}}$. To do so, we use the $L_1$-norm of $\mathbf{v}$, which also acts as a regularizer, as in [16].

In order to solve the non-differentiable problem caused by quantization in an end-to-end DNN training, we adopt a slight modification of the approach in [59]. Given quantization value centers $\mathcal{C} = \{q_1, ...q_L\} \subset \mathbb{R}$, and $n \in \{1, ..., L\}$, we use the nearest neighbor assignments to compute $\hat{\mathbf{z}}$ as

$$\hat{\mathbf{z}} = Q(\mathbf{z}) := \mathrm{argmin}_n \|\mathbf{z} - q_n\|. \quad (10)$$

To be able to compute gradients in the backward pass, we approximate (10) by the so-called soft-quantization

$$\hat{\mathbf{z}} = \sum_{n=1}^{L} \frac{\exp(-\sigma\|\mathbf{z}-q_n\|)}{\sum_{h=1}^{L}\exp(-\sigma\|\mathbf{z}-q_h\|)} q_n, \quad (11)$$

where $\sigma > 0$ is a parameter used to control the degree of approximation.

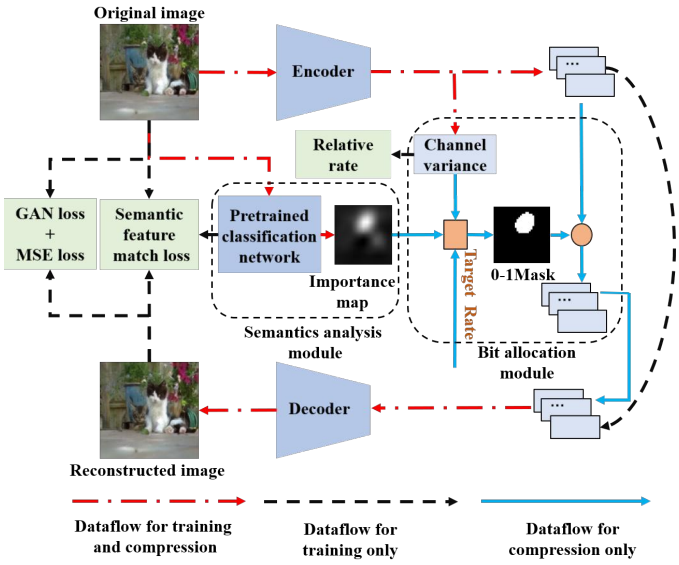



Fig. 6. The framework of the mask-based progressive compression algorithm. The rectangle orange block is a mask-generation module, which uses LCM and SIM information to generate a binary mask according to a target rate. The circle orange block means applying the binary mask to the latent vector.

### B. Implementation of mask-based progressive generative compression (MPGC)

This subsection introduces two basic MPGC schemes that adopt the spatial-mask approach for progressive encoding. The framework of MPGC is illustrated in Fig. 6, which further elaborates the modules in Fig. 3 and shows the dataflows for training and compression. We now explain the details of the adaptive bit allocation algorithm and network training procedure.

*1) Mask-based variable-rate bit allocation:* A bit allocation strategy is used to assign bits to the latent representation according to a mask that blends SIM and LCM. The LCM is the variance vector $\mathbf{v}$ of the latent vector along the channel dimension. The variance vector is defined in (9). Its elements $v_{i,j}$ estimates the entropy in position $(i, j)$ of the latent vector. A greater value of $v_{i,j}$ indicates a pixel-wise more complex spatial area that demands more bits for encoding.

Different strategies can be designed to blend the SIM and the LCM. A simple and trivial blending policy is to calculate the weighted sum between SIM and LCM. Such a strategy, however, cannot maintain the integrity of semantic

importance, which means the order of semantic importance will be interfered with by LCM. This may result in loss of semantic information at low rates. To overcome this, we propose a novel hierarchical blending strategy that maintains the original importance order. The rationale of our strategy is to first divide the latent vector space into two parts: semantic salient regions and non-salient regions. This binary partition is achieved by applying the OSTU method [60] on the SIM vector to obtain a binary vector $\mathbf{B}$, with entries '1' indicating semantically-salient positions and '0' indicating otherwise semantically unimportant positions. The salient regions are then assigned with importance values ranging from 1 to 2 by adding a constant 1. On the other hand, non-salient regions take values from a normalized LCM vector ranging from 0 to 1. In this way, the original order of SIM and LCM are preserved in the semantic salient regions and the remaining regions, respectively. The SIM is denoted as $\mathbf{e}$ and the LCM is the variance vector $\mathbf{v}$ of the latent vector along the channel dimension. Mathematically, the proposed blending strategy yields a final importance map $\mathbf{I}$ given by

$$\mathbf{I} = (1 + \mathbf{e})\mathbf{B} + \text{sigmoid}(\mathbf{v})(1 - \mathbf{B}). \tag{12}$$

Given a target rate and the importance map, the binary mask $\mathbf{M}$ can be obtained as

$$M_{i,j} = \begin{cases} 1, & \text{if} \quad I_{i,j} > T \\ 0, & \text{otherwise,} \end{cases} \tag{13}$$

where $T$ is an adaptive threshold according to the target rate. Elements of $\mathbf{M}$ and $\mathbf{I}$ are denoted as $M_{i,j}$ and $I_{i,j}$, respectively. Once the binary mask is obtained, the latent vector $\mathbf{z}$ can be transformed into $\tilde{\mathbf{z}}$ as

$$\tilde{z}^l_{i,j} = \begin{cases} z^l_{i,j}, & \text{if} \quad M_{i,j} = 1 \\ C^{-1} \sum_l z^l_{i,j}, & \text{otherwise,} \end{cases} \tag{14}$$

where $\tilde{z}^l_{i,j}$ is the element of $\tilde{\mathbf{z}}$. Therefore, a quantized latent vector $\hat{\mathbf{z}} = Q(\tilde{\mathbf{z}})$ can be obtained. For the first channel of $\hat{\mathbf{z}}$, all values are extracted. For other channels, only values at positions where $M_{i,j} = 1$ are extracted. These extracted values are assembled into a vector. The mask and the vector extracted from $\hat{\mathbf{z}}$ are further compressed by lossless adaptive arithmetic coding (AAC) before being transmitted to the receiver. Finally, the receiver recovers the latent vector $\tilde{\mathbf{z}}$ and uses it to reconstruct the image.

Based on the above scheme, the rates after compression are determined by the number of 1's in the mask $\mathbf{M}$. Letting $P$ and $S$ be the number of all elements and 1's in the mask $\mathbf{M}$, respectively, and $r$ be the target rate measured by bits per pixel (bpp), we have

$$\text{AAC}(\mathbf{M}) + S\frac{\text{AAC}(\mathbf{o})}{P} + (P - S)\frac{\text{AAC}(\mathbf{u})}{P} \leq r, \tag{15}$$

where AAC$(\cdot)$ denotes the computation of calculating the bpp of the input vector using AAC, and $\mathbf{o}$ and $\mathbf{u}$ are vectors extracted from $\hat{\mathbf{z}}$ when all elements of $\mathbf{M}$ are 1 and 0, respectively. (15) can be rewritten as

$$S \leq P\frac{r - \text{AAC}(\mathbf{u}) - \text{AAC}(\mathbf{M})}{\text{AAC}(\mathbf{o}) - \text{AAC}(\mathbf{u})}, \tag{16}$$

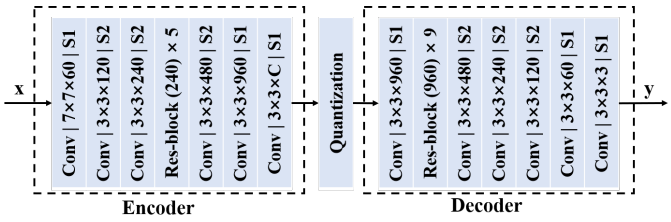



Fig. 7. The network structure of MPGC-1 model. Conv |7×7×60|S1 indicates a convolution layer using a kernel of size 7×7, 60 output channels, and stride of 1 (in the decoder, Conv indicates transposed convolution). Res-block(240)×5 indicates 5 residual blocks with 240 output channels. And after every convolution, the instance normalization is adopted here.

where $P$, $\text{AAC}(\mathbf{o})$, and $\text{AAC}(\mathbf{u})$ are fixed. Because $\text{AAC}(\mathbf{M})$ is negligible compared to $r$ and $\text{AAC}(\mathbf{u})$, it follows that $S$ can be approximated by

$$S \lesssim \lfloor P\frac{r - \text{AAC}(\mathbf{u})}{\text{AAC}(\mathbf{o}) - \text{AAC}(\mathbf{u})} \rfloor. \tag{17}$$

Thus, given a desired rate $r$, (17) gives us the number of elements $S$ to select in the importance map $\mathbf{I}$, from which we can compute the threshold $T$ in (17). That is, we select the $S$ largest elements in $\mathbf{I}$.

*2) Network structure and end-to-end training:* The training procedure of a DNN affects not only the convergence speed but also the quality of the latent representation. The training algorithm is also coupled with the network structure. Depending on whether the mask is used as side information in training, we distinguish two kinds of training procedures:

**DNN training without mask information (MPGC-1).** The network structure is shown in Fig. 7. During the training phase, the decoder obtains the latent vector and generates the reconstructed image. After training, mask-based multi-rate compression is adopted. The loss functions $\mathcal{L}_\text{D}$ and $\mathcal{L}_\text{G}$ are alternatively optimized, as described in Section IV-A. The loss term $\mathcal{L}_\text{D}$ is defined in (4), and $\mathcal{L}_\text{G}$ is given by

$$\begin{aligned} \mathcal{L}_\text{G} = \min_{E,G} \mathbb{E}_\mathbf{x}\{&\lambda_\text{SFM}\mathcal{L}_\text{SFM} + \lambda_\text{GAN}\mathcal{L}_\text{GAN} \\ &+ \lambda_\text{MSE}\mathcal{L}_\text{MSE} + \lambda_\text{R}\|\mathbf{v}\|_1\}. \end{aligned} \tag{18}$$

**DNN training with mask information (MPGC-2).** In MPGC-1, the representation learning phase and the bit allocation phase are essentially separated. The encoder does not learn to distinguish the semantic salient regions from the non-salient regions. As a consequence, the classification accuracy of the reconstructed image significantly deteriorates at low rates. To overcome this, we propose to use conditional GANs to compress the input image using the mask as a conditioned priori information. In this way, the network can learn how to use different regions of the received latent vector for image reconstruction. The network structure is shown in Fig. 8 and condition GANs' loss terms are as follows

$$\begin{aligned} \mathcal{L}_\text{D}(\mathbf{U}) = \min_{D_1,D_2,D_3} \mathbb{E}_\mathbf{x}\{\sum_{k=1}^{3}[&(D_k(\mathbf{U},\mathbf{x}) - 1)^2 \\ &+ (D_k(\mathbf{U}, G(\mathbf{U}, Q(\tilde{\mathbf{z}}))))^2]\}, \end{aligned} \tag{19}$$

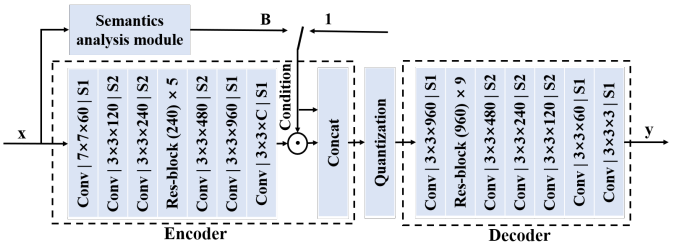



Fig. 8. Network structure of the MPGC-2 model. The vectors $\mathbf{B}$ and $\mathbf{1}$ are separate inputs as conditional vectors during training.

$$\mathcal{L}_{\text{GAN}}(\mathbf{U}) = \sum_{k=1}^{3} (D_k(\mathbf{U}, G(\mathbf{U}, Q(\tilde{\mathbf{z}}))) - 1)^2, \tag{20}$$

$$\mathcal{L}_{\text{SFM}}(\mathbf{U}) = \frac{1}{H_\xi W_\xi C_\xi} \sum_{i=1}^{H_\xi} \sum_{j=1}^{W_\xi} \sum_{k=1}^{C_\xi} \| \xi(\mathbf{x})_{i,j,k} - \xi(G(\mathbf{U}, Q(\tilde{\mathbf{z}})))_{i,j,k} \|^2 , \tag{21}$$

$$\mathcal{L}_{\text{MSE}}(\mathbf{U}) = \frac{1}{3HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \sum_{k=1}^{3} \| \mathbf{x}_{i,j,k} - G(\mathbf{U}, Q(\tilde{\mathbf{z}}))_{i,j,k} \|^2 , \tag{22}$$

where $\mathbf{U}$ is the condition vector. The multi-scale discriminator loss $D_k$ is calculated by $D_k(\mathbf{U}, \mathbf{x})$, and the concatenation of $\mathbf{U}$ and $\mathbf{x}$ is fed into the discriminator. The reconstructed image is calculated by $G(\mathbf{U}, Q(\tilde{\mathbf{z}}))$, and the input of decoder is the concatenation of $\mathbf{U}$ and $\boldsymbol{Q}(\tilde{\mathbf{z}})$.

Depending on what data is fed for DNN training, two kinds of methods can be distinguished. The first is to train the DNN with the entire image (i.e., no mask). This policy is optimized for image reconstruction. The second strategy is to train the DNN with SIM-masked data. This approach encourages the DIC to learn to better encode semantic salient areas, and is thus optimized for image classification. To serve hybrid context applications, we can strike a balance between the two objectives by applying different masks ($\mathbf{U} = \mathbf{1}$ and $\mathbf{U} = \mathbf{B}$) on the training data. The new loss functions can be given as the numerical average given by

$$\mathcal{L}_{\text{D2}} = (\mathcal{L}_{\text{D}}(\mathbf{B}) + \mathcal{L}_{\text{D}}(\mathbf{1}))/2, \tag{23}$$

$$\begin{aligned} \mathcal{L}_{\text{G2}} = \min_{E,G} \mathbb{E}_{\mathbf{x}} \{ & \lambda_{\text{SFM}}(\mathcal{L}_{\text{SFM}}(\mathbf{B}) + \mathcal{L}_{\text{SFM}}(\mathbf{1}))/2 \\ & + \lambda_{\text{R}} \|\mathbf{v}\|_1 + \lambda_{\text{GAN}}(\mathcal{L}_{\text{GAN}}(\mathbf{B}) + \mathcal{L}_{\text{GAN}}(\mathbf{1}))/2 \\ & + \lambda_{\text{MSE}}(\mathcal{L}_{\text{MSE}}(\mathbf{B}) + \mathcal{L}_{\text{MSE}}(\mathbf{1}))/2 \}. \end{aligned} \tag{24}$$

### *C. Implementation of residual-enhanced mask-based progressive generative compression (RMPGC)*

We now describe the proposed RMPGC scheme. As shown in Fig. 9, we adopt the multi-scale decomposition (MSD) [10] based on residual to decompose the latent vector into multi-layers. At the receiver end, the inverse multi-scale decomposition (IMSD) is used to integrate different layers into a latent vector. During image compression, the layers partition module divides layers into the base layers and enhance layers according to the target rates and context goal. The base layers are given the highest priority to guarantee a minimal quality

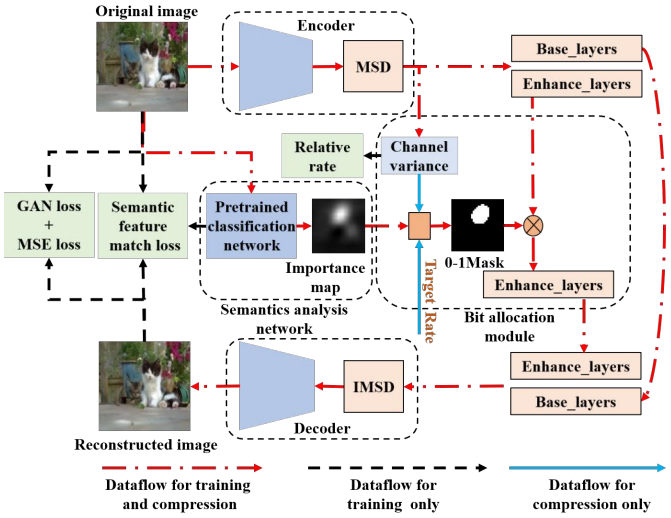



Fig. 9. The framework of the RMPGC algorithm. The rectangle orange block represents the mask-based bits allocation module, which generates a binary mask according to the target rate, $C_b$, LCM, and SIM.

of reconstruction and overall semantics, while the mask-based progressive compression is utilized to allocate additional bits when available. This structure exploits the fact that minimum-rate image reconstruction is in line with protecting the overall semantics, which benefits image classification as well.

*1) Variable-rate bit allocation:* As shown in Fig. 9, the bit allocation procedure of RMPGC is as follows.

Step 1: Bits are first assigned to the base layers. The base layers and enhance layers of the latent vector are denoted as $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, respectively. The rate of the base layers is $\text{AAC}(\boldsymbol{\alpha})$, whereas the rate of the enhance layers is

$$r^{'} = r - \text{AAC}(\boldsymbol{\alpha}). \tag{25}$$

Step 2: The importance map $\mathbf{I}$ of the latent vector is generated based on SIM and LCM, according to (12).

Step 3: The bit allocation mask $\mathbf{M}$, which is used to assign bits to the enhance layers, can be obtained by (13).

Step 4: The latent vector $\tilde{\mathbf{z}}$ is generated by

$$\tilde{\mathbf{z}} = \text{concat}(\boldsymbol{\alpha}, \boldsymbol{\beta}\mathbf{M}), \tag{26}$$

where concat$(\cdot)$ means concatenating the input vectors.

Step 5: The latent vector $\tilde{\mathbf{z}}$ is quantized to $\hat{\mathbf{z}}$. Then $\mathbf{M}$ and $\hat{\mathbf{z}}$ are packed into a code stream, which is further compressed by adaptive arithmetic coding and sent to the decoder.

The threshold $T$, which is used to generate $\mathbf{M}$, is determined by the target rate, base layers and the allocation strategy. We have

$$\text{AAC}(\mathbf{M}) + \text{AAC}(\boldsymbol{\alpha}) + S\frac{\text{AAC}(\boldsymbol{\beta})}{P} \leq r. \tag{27}$$

The equation can be rearranged as

$$S \leq P\frac{r - \text{AAC}(\boldsymbol{\alpha}) - \text{AAC}(\mathbf{M})}{\text{AAC}(\boldsymbol{\beta})}, \tag{28}$$

where $P$, $\text{AAC}(\boldsymbol{\alpha})$, and $\text{AAC}(\boldsymbol{\beta})$ are fixed values. Even though $\text{AAC}(\mathbf{M})$ changes with different $\mathbf{M}$, it is negligible compared

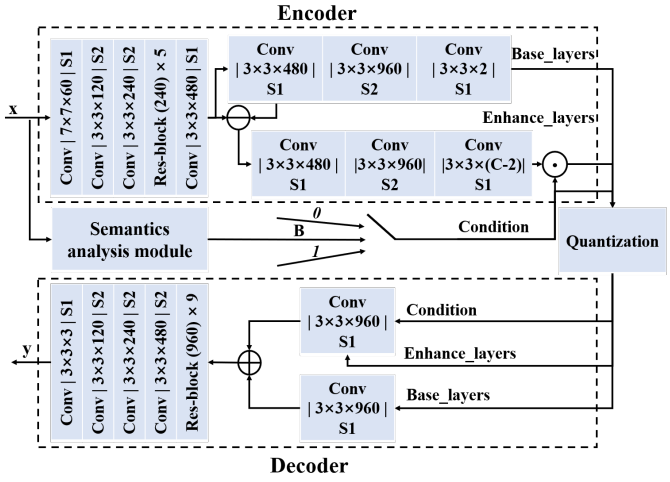



Fig. 10. Network structure of the RMPGC model. The vectors $\mathbf{B}$, $\mathbf{0}$, and $\mathbf{1}$ are separate inputs as conditional vectors during training.

to $r$ and AAC($\boldsymbol{\beta}$). Therefore, $S$ can be approximated by

$$S \lesssim \lfloor P\frac{r-\mathrm{AAC}(\boldsymbol{\alpha})}{\mathrm{AAC}(\boldsymbol{\beta})}\rfloor. \tag{29}$$

The $S$th greatest value of the importance map $\mathbf{I}$ is chosen to be the threshold $T$ to get the corresponding binary mask for the given $r$.

*2) Network structure and end-to-end training:* We realized a compression model whose base layers number is 2 and the model's network structure is shown in Fig. 10. Only the base layers are delivered at the lowest rate in the RMPGC. As the rate increases, the enhance layers are subject to a masking operation with $\mathbf{M}$. Our RMPGC implementation is based on MPGC-2, which uses conditional GANs. The loss terms are given by

$$\begin{aligned}\mathcal{L}_{\mathrm{D}}(\mathbf{U}) = \min_{D_1,D_2,D_3} \mathbb{E}_{\mathbf{x}}\{\sum_{k=1}^{3}[(D_k(\mathbf{U},\mathbf{x})-1)^2 \\ +(D_k(\mathbf{U},G(\mathbf{U},Q(\boldsymbol{\alpha},\boldsymbol{\beta}\mathbf{M}))))^2]\},\end{aligned} \tag{30}$$

$$\mathcal{L}_{\mathrm{GAN}}(\mathbf{U}) = \sum_{k=1}^{3}(D_k(\mathbf{U},G(\mathbf{U},\boldsymbol{\alpha},\boldsymbol{\beta}\mathbf{M}))-1)^2, \tag{31}$$

$$\begin{aligned}\mathcal{L}_{\mathrm{SFM}}(\mathbf{U}) = \frac{1}{H_\xi W_\xi C_\xi}\sum_{i=1}^{H_\xi}\sum_{j=1}^{W_\xi}\sum_{k=1}^{C_\xi} \parallel \xi(\mathbf{x})_{i,j,k} - \\ \xi(G(\mathbf{U},Q(\boldsymbol{\alpha},\boldsymbol{\beta}\mathbf{M})))_{i,j,k} \parallel^2,\end{aligned} \tag{32}$$

$$\begin{aligned}\mathcal{L}_{\mathrm{MSE}}(\mathbf{U}) = \frac{1}{3HW}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{3} \parallel \mathbf{x}_{i,j,k} - \\ G(\mathbf{U},Q(\boldsymbol{\alpha},\boldsymbol{\beta}\mathbf{M}))_{i,j,k} \parallel^2.\end{aligned} \tag{33}$$

Similar to MPGC-2, to achieve a balanced performance, three different conditions are utilized and trained at the same time. We set $\mathbf{U}=\mathbf{0}$, $\mathbf{U}=\mathbf{1}$, and $\mathbf{U}=\mathbf{B}$. These settings correspond to the situation that the decoder gets the base layers only, the complete latent vector, and the base layers plus masked enhance layers, respectively. The final loss functions are averaged over these conditions and given by

$$\mathcal{L}_{\mathrm{D3}} = (\mathcal{L}_{\mathrm{D}}(\mathbf{0})+\mathcal{L}_{\mathrm{D}}(\mathbf{B})+\mathcal{L}_{\mathrm{D}}(\mathbf{1}))/3, \tag{34}$$

$$\begin{aligned}\mathcal{L}_{\mathrm{G3}} = \min_{E,G}\mathbb{E}_{\mathbf{x}}\{\lambda_{\mathrm{SFM}}(\mathcal{L}_{\mathrm{SFM}}(\mathbf{0})+\mathcal{L}_{\mathrm{SFM}}(\mathbf{B}) \\ +\mathcal{L}_{\mathrm{SFM}}(\mathbf{1}))/3+\lambda_{\mathrm{GAN}}(\mathcal{L}_{\mathrm{GAN}}(\mathbf{0}) \\ +\mathcal{L}_{\mathrm{GAN}}(\mathbf{B})+\mathcal{L}_{\mathrm{GAN}}(\mathbf{1}))/3 \\ +\lambda_{\mathrm{MSE}}(\mathcal{L}_{\mathrm{MSE}}(\mathbf{0})+\mathcal{L}_{\mathrm{MSE}}(\mathbf{B}) \\ +\mathcal{L}_{\mathrm{MSE}}(\mathbf{1}))/3+\lambda_{\mathrm{R}}\|\mathbf{v}\|_1\}.\end{aligned} \tag{35}$$

## V. Experimental Results

### A. Experimental setup

Throughout our experiments, all DIC models are trained using the ImageNet [61] dataset, and tested on the ImageNet, DogvsCat [62], and Kodak [63] datasets. For ImageNet, we randomly select 20,000 images from the validation set for training, and 5,000 for testing. Also, 8,000 images from the DogvsCat dataset and 24 images from the Kodak dataset are applied to evaluate how well the models trained using ImageNet can be generalized to different datasets. All input images are resized to $256 \times 256$, and compressed by the encoder by a factor of 8. The number of channels of the latent vector is $C=8$, while the number of base layers is set to $C_b=2$. The quantization level $L$ is set to 5 with centers at $\{-2,-1,0,1,2\}$.

In our experiments, the encoding rate of the proposed models is set to cover the full range from the lowest bpp to the largest bpp by changing the spatial mask from $\mathbf{0}$ to $\mathbf{1}$. The feature extraction section of VGG16 [64] is used as the CNN feature extractor in our semantics analysis module and classification network. Throughout the paper, we adopt the pre-trained VGG16 network parameters as published on the Tensorflow official website. The VGG16 feature extraction network (i.e., excluding the last fully-connected layer) is fixed throughout the paper, such that it can be seen as a task-related, predefined semantic feature extraction module. The last fully connected layer is fine-tuned for different datasets. Once trained, the final layer is also frozen when integrating with all DIC models under test. When calculating the SFM loss during training, we first compute each SFM loss of features before four max pool layers, and then the weighted sum of those losses is taken as the final SFM loss. The weighs of each SFM loss are set as 1/32,1/16,1/8,and 1/4, respectively. GradCAM++ is adopted in the semantics analysis module to generate SIM. We use the Adam optimizer [65] with a learning rate of 2e-4 for both the generator and discriminator. During training, all the loss terms are regularized by their own historical means to the same scale of magnitude.

### B. Performance metrics

The performance of the proposed DIC and other baseline models are evaluated with a diverse range of metrics, including classification accuracy, peak signal to noise ratio (PSNR), structure similarity index measure (SSIM), feature similarity index measure (FSIM) [66], learned perceptual image patch

similarity (LPIPS) [67], and natural image quality evaluator (NIQE) [68]. Classification accuracy and PSNR are well-established metrics for classification and distortion, respectively. We discuss the meaning of the other metrics below to shed light on how they are related to the RDCP trade-off problem.

SSIM, FSIM, LPIPS, and NIQE were all proposed as perceptual metrics, but they have a different nature. SSIM, FSIM, and LPIPS evaluate the subjective perceptual similarity between two images (e.g., original and reconstructed images). SSIM measures the distortion regarding three image features: luminance, contrast, and structure. FSIM focuses on perceptually-sensitive local features such as phase congruency and gradient magnitude, while LPIPS focuses on higher level "deep features". Interestingly, it has been shown that LPIPS and semantic tasks (such as classification) attend to similar features [67]. In essence, SSIM, FSIM, and LPIPS all measure distortions in a particular feature domain. In contrast, NIQE is a non-referential metric that evaluates how much a reconstructed image's statistics deviate from natural image statistics. Recall that in the theoretical formulation of the DCP trade-off, perceptual quality is defined as the divergence between image distributions. Minimizing such a divergence is also the motivation of using GAN in the proposed DIC. In this paper, we follow the DCP trade-off framework [28] to define perceptual quality as statistical divergence. As a result, SSIM, FSIM, and LPIPS will be interpreted as distortion metrics, while NIQE will be interpreted as a perception metric. Finally, we note that the smaller the values of the LPIPS and NIQE metrics, the better; for the other metrics, the larger their values, the better the performance.

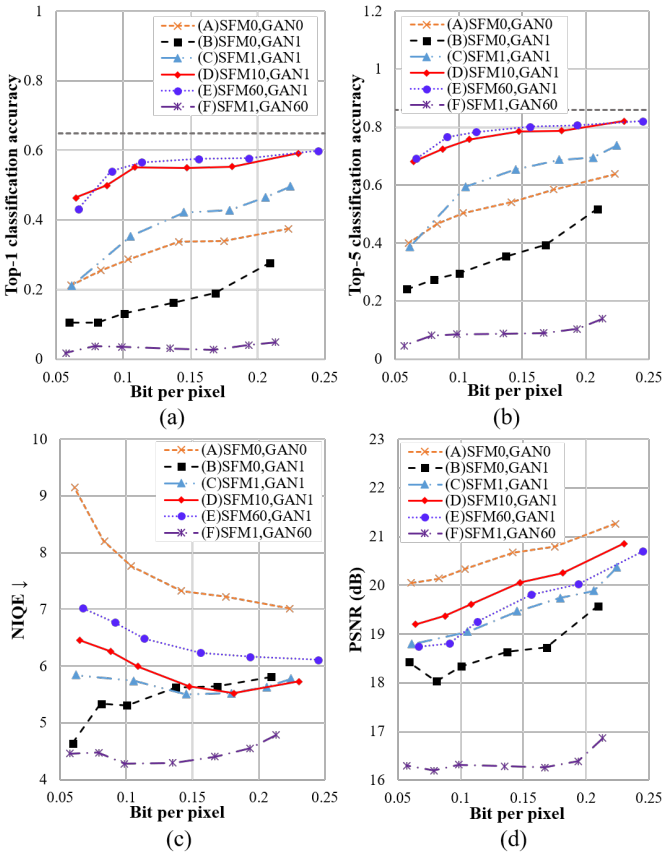



Fig. 11. Classification accuracy, PSNR, and NIQE as functions of bpp with varying weighting parameters $\lambda_{\text{SFM}}$ and $\lambda_{\text{GAN}}$. The dashed lines mark the classification accuracy using original images as inputs. A smaller value of NIQE means better performance.

### C. Impact of SFM and GAN losses

The loss function used for training the DNN has a global and implicit impact on the RDPC trade-off. Particularly, we investigate how the DPC metrics vary with changing ratios of the three loss terms: $\lambda_{\text{SFM}}$, $\lambda_{\text{GAN}}$, and $\lambda_{\text{MSE}}$. Without loss of generality, we set $\lambda_{\text{MSE}}$ and $\lambda_{\text{R}}$ to be 1. The following six cases are then investigated:

(A) $\lambda_{\text{SFM}} = 0, \lambda_{\text{GAN}} = 0$;
(B) $\lambda_{\text{SFM}} = 0, \lambda_{\text{GAN}} = 1$;
(C) $\lambda_{\text{SFM}} = 1, \lambda_{\text{GAN}} = 1$;
(D) $\lambda_{\text{SFM}} = 10, \lambda_{\text{GAN}} = 1$;
(E) $\lambda_{\text{SFM}} = 60, \lambda_{\text{GAN}} = 1$;
(F) $\lambda_{\text{SFM}} = 1, \lambda_{\text{GAN}} = 60$.

In our experiments, we use the RMPGC model and randomly select 100 categories of images from the ImageNet test set.

Fig. 11 compares the performance of all the six cases in terms of classification accuracy, PSNR, and NIQE. We make the following key observations. First, Case A sets an upper bound of the attainable PSNR as the DIC is solely optimized for reconstruction. This PSNR upper bound is largely determined by the DNN backbone we used. Second, Cases A, B, and F together show that increasing the dominance of GAN loss will improve NIQE at the cost of sacrificing other metrics. Such a DCP performance trade-off is well-expected. Third, a more interesting DCP trade-off behavior is observed when we look at Cases B, C, D, and E. In these cases, the GAN and MSE loss are kept to be 1, while the SFM loss increases from 0 to 60. We can see that the classification performance improves with diminishing returns as expected, but not at the cost of other metrics. More specifically, the PSNR improves when SFM increases from 0 to 10. Similarly, the NIQE initially improves at higher rates. It appears that all tested DCP metrics can benefit from increasing the SFM loss up to a certain limit. Finally, increasing the encoding rate (bpp) generally leads to improved classification and PSNR performances. However, this is not always the case for NIQE. Specifically, in Cases B and F, where the GAN loss dominates the SFM loss, the NIQE shows no improvement with increasing rates. This corroborates our previous argument that the perceptual performance alone (measured as statistical divergence) is not a function of rate.

Our observations above imply that RDPC does not follow a strict trade-off relationship. More importantly, it appears that the SFM loss plays an important role in exploiting the synergy among DCP performances. On one hand, SFM loss can potentially improve the distortion performance by encouraging alignments at the feature domain; on the other hand, SFM loss could help to improve the perception metrics LIPIS because both metrics focus on shareable high-level features.

### D. Comparison of mask-based bit allocation policies

The second mechanism we introduced to facilitate transitions between the two contextual goals is mask-based adaptive bit allocation, as explained in Section III-D2. To demonstrate the effectiveness of this mechanism, we run the MPGC-1 scheme, which gives intuitively understandable semantic features. Taking a typical image as an example, Fig. 12 shows the visual effect of reconstructed images under three bit allocation policies: SIM-based, LCM-based, and the hybrid policy denoted as SIM&LCM. We can see that as the rates decrease, the SIM-based policy better protects the semantic regions related to animal features, while the LCM-based policy better protects the background with complex pixel-level variations. The blended policy is able to strike a balance and preserve both the semantic salient features and the complex background features.

Fig. 13 shows the average classification accuracy and PSNR with different masking policies. We use RMPGC and set the loss weights to be $\lambda_{\text{SFM}} = 10$, $\lambda_{\text{GAN}} = \lambda_{\text{MSE}} = \lambda_{\text{R}} = 1$. As expected, SIM-based and LCM-based policies are biased towards classification and reconstruction, respectively. The hybrid policy shows promising performance in having the merits of both SIM and LCM.

SIM and LCM provide us with another angle to view the classification-reconstruction trade-off. As illustrated in Fig. 12, taking an intersection of the SIM and LCM masks will give us pixels that are important to both classification and reconstruction. The implication is that there is synergy between achieving these two goals. To gauge the potential of such a synergy, we calculate the cosine similarity of SIM and LCM vectors over 500 images from the ImageNet dataset. The outcome fits a bell-shape distribution in the [0.2 0.7] interval. This implies the universality of such a synergy in practical image datasets.

### E. Comparison of different CAM schemes

In the proposed DIC, CAM [46] is a key component in the semantics analysis module to get the SIM. Many CAM techniques have been proposed in the literature [46]–[50]. Among them, GradCAM++ is known to have precise object positioning performance [49]. Here, we compare GradCAM++ with two other latest variants of CAM schemes: XGradCAM [49] and LayerCAM [50]. Table II demonstrates that GradCAM++ achieves the best overall performance in terms of classification accuracy and PSNR. We hence use GradCAM++ in our experiments.

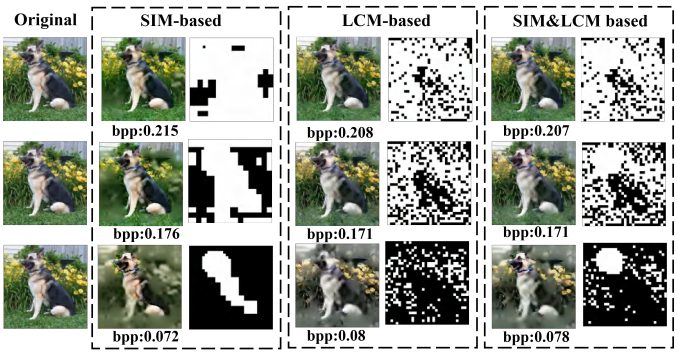



Fig. 12. Visual comparisons of reconstructed images using different bit allocation strategies.

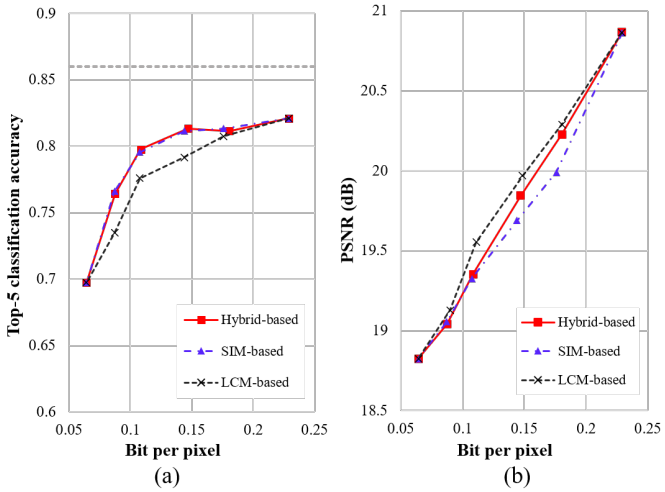



Fig. 13. Classification accuracy and PSNR as a function of bpp with different mask-based bit allocation policies.

TABLE II
CLASSIFICATION ACCURACY AND PSNR OF COMPRESSED IMAGES USING DIFFERENT CAM SCHEMES.

| Bpp | CAM | Top-1 | Top-5 | PSNR |
|---|---|---|---|---|
| 0.21 | GradCAM++ | 0.5658 | 0.8075 | **20.4984** |
| | Layer CAM | 0.5717 | **0.8134** | 20.4631 |
| | XGradCAM | **0.5776** | 0.8055 | 20.4628 |
| 0.15 | GradCAM++ | **0.5697** | **0.8134** | 19.8438 |
| | Layer CAM | 0.554 | 0.7898 | **19.8772** |
| | XGradCAM | 0.5619 | 0.7878 | 19.7546 |
| 0.11 | GradCAM++ | **0.5383** | **0.7976** | **19.3524** |
| | Layer CAM | 0.5305 | 0.7721 | 19.3501 |
| | XGradCAM | 0.5265 | 0.7701 | 19.2691 |
| 0.085 | GradCAM++ | **0.5147** | **0.7642** | 19.0403 |
| | Layer CAM | 0.499 | 0.7466 | **19.0417** |
| | XGradCAM | 0.4931 | 0.7544 | 18.9985 |

### F. Performance evaluation and comparison

This subsection aims to evaluate the performance of the proposed DICs via quantitative metrics. We set $\lambda_{\text{MSE}} = \lambda_{\text{GAN}} = \lambda_{\text{R}} = 1$, $\lambda_{\text{SFM}} = 10$, and use the hybrid SIM & LCM bit allocation strategy. Performance evaluations on the ImageNet and DogvsCat dataset are shown in Figs. 14 and 15, respectively. As a performance benchmark, the Top-1 and Top-5 classification accuracy of original images from the ImageNet test set are 64% and 85%, respectively. The Top-1 classification accuracy on the DogvsCat test set is 96%.

*1) Ablation analysis:* Three specific implementations of progressive DIC algorithms were proposed in Section IV: MPGC-1, MPGC-2, and RMPGC. Our implementations share the same generative DNN model used in VRGC2020 [16]. We briefly recall that MPGC-1 improves upon VRGC by adding semantic analysis modules, MPGC-2 refines MPGC-1 with a

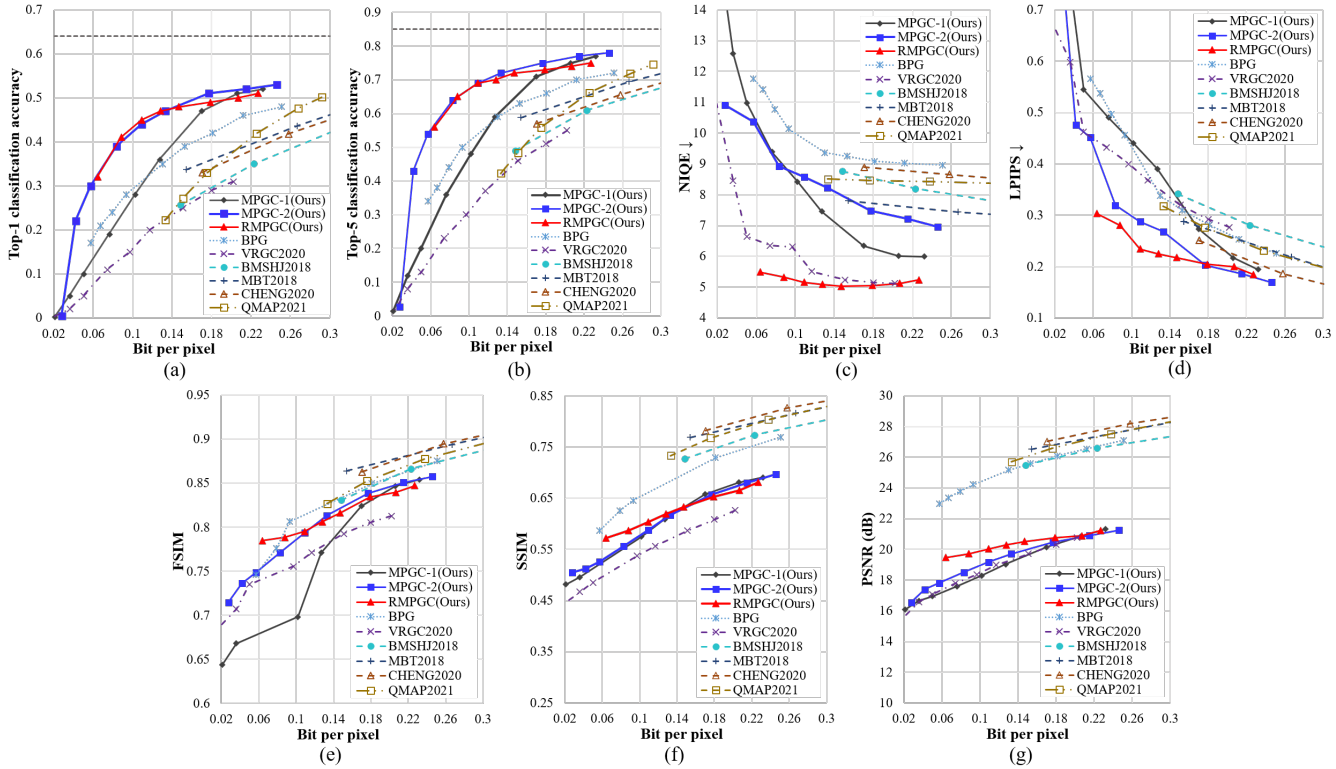



Fig. 14. Classification, perception, and distortion performances as functions of bpp for different DIC schemes using the ImageNet dataset.

new training method, and RMPGC improves upon MPGC-2 via residual-based layering. These four models are used for an ablation analysis.

MPGC-1 is shown to outperform VRGC in classification, but is worse in other metrics. MPGC-2 outperforms VRGC in all metrics apart from NIQE. Compared to MPGC-2 and RMPGC, MPGC-1 performs worse on both datasets, especially at low rates. RMPGC is shown to outperform both MPGCs in all metrics, except for a small degradation of classification accuracy at high bpps. Finally, RMPGC is shown to outperform VRGC in all metrics. Taking the ImageNet dataset results in Figs. 14 for example, RMPGC outperforms VRGC by increasing the top-1 and top-5 classification accuracy by 23% and 27%, respectively. Moreover, the PSNR is increased by 0.73 dB. For SSIM, FSIM, LPIPS, and NIQE, the relative performance enhancements (calculated as the performance difference divided by the VRGC performance) are 8%, 4%, 34%, and 6%, respectively. The above ablation studies clearly demonstrate the effectiveness of the various methods proposed in this paper and suggest that these methods are transferable techniques that can be applied to different DNN backbones.

*2) RDCP trade-off performances:* So far, we have established the proposed RMPGC as the best-performing GAN-based DIC. We will subsequently compare RMPGC with other DIC schemes. Four representative DIC schemes are chosen from the literature, including three fixed-rate DICs BMSHJ2018 [29], MTB2018 [30], and CHENG2020 [31], and an encoder-based variable-rate DIC QMAP2021 [8]. The three fixed rate DICs are single-context DICs optimized for reconstruction, while the QMAP2021 is a hybrid context DIC jointly optimized for reconstruction and classification. Apart from these DICs, the classic image codec BPG [69] will also be used for comparison.

For classification accuracy, the proposed RMPGC consistently outperforms BPG and other DIC schemes at varying rates. An average improvement of about 10% is observed for both top-1 and top-5 classification accuracy. Such an improvement is due to the inclusion of SFM loss and the use of a hybrid bit rate allocation approach, such that information of semantically-salient areas is better preserved at varying rates. The classification accuracy as a function of rate is shown to be a monotonically increasing function, but with diminishing returns as the rate increases.

For distortion performance, BPG and non-GAN-based DIC schemes significantly outperform RMPGC in PSNR by a minimum of 5 dB. There are two underlying reasons for this. First, applying the GAN loss in RMPGC tends to optimize for the overall approximation of pixel distribution rather than pixel-by-pixel fidelity. Second, the LS-GAN DNN backbone implemented in RMPGC constrains the PSNR performance, as previously discussed in Fig. 11.

For perceptual metrics, RMPGC yields unsatisfactory performance in SSIM and FSIM, both of which focus on the distortion of low-level features. However, RMPGC yields the best performance in LIPIS. This means that RMPGC is able to encode perceptually-salient high-level features with priority. Finally, for the perception metric NIQE, RMPGC consistently outperforms other models in both datasets. This validates the key advantage of GAN-based generative DIC models.

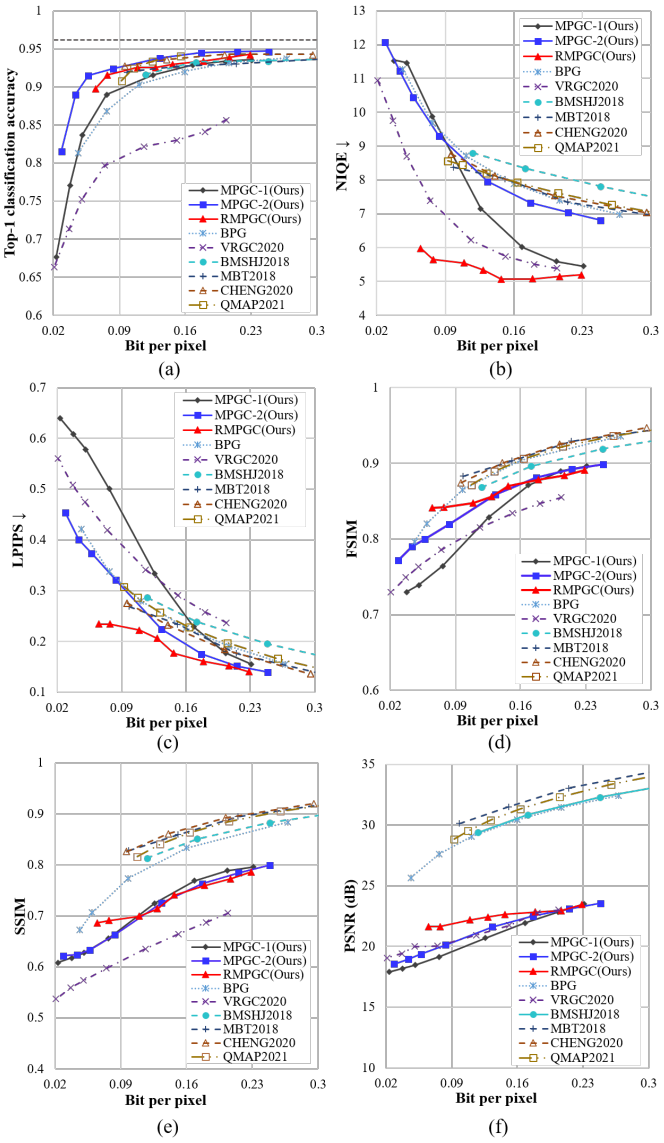



Fig. 15. Classification, perception, and distortion performances as functions of bpp for different DIC schemes using the DogvsCatdataset.

*3) Transferability of DIC models:* As above mentioned, all proposed models are trained using the ImageNet dataset. In Fig. 15, our models are tested for a different DogvsCat dataset with a predefined classifier. We can see that our models yield consistent performance gains when extended to the DogvsCat dataset. A possible explanation for this transferability is that the underlying techniques used in our model, such as SFM loss and LCM masking, attends to universally important information in image datasets. This suggests the feasibility of training a wide-purpose DIC codec using a representative dataset.

## G. Comparison of visual effects

Fig. 16 shows the visual comparisons of proposed algorithms trained on the ImageNet dataset. Compared to BPG, reconstructed images generated by RMPGC and MGPC-2 always have a better overall perception, while MPGC-1 and

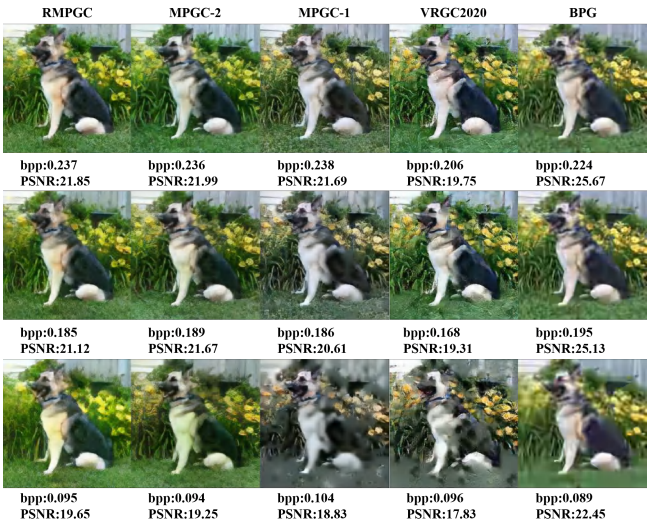



Fig. 16. Visual comparisons of different DIC algorithms.

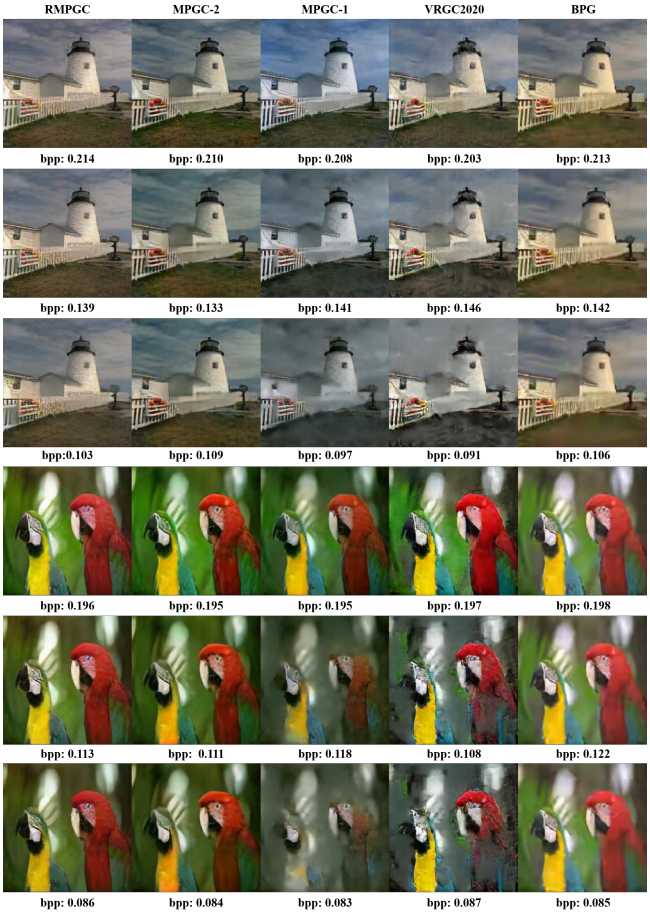



Fig. 17. Visual comparisons of different algorithms on the Kodak dataset.

VRGC perform worse at low rates. When the rates decrease, it can be observed that the BPG images get blurry and less recognizable due to the gradual loss of semantic information. The phenomenon of semantic information loss is even worse in VRGC's progressive DIC scheme because misguided protection is placed on the background region with complex

features. In contrast, the proposed schemes exhibit a desirable progressive behavior that better protects semantically relevant regions. However, because MPGC-1 does not learn to use less information to generate background, the overall perception seriously decreases when the rate is low. Because the mask is used as prior information in the MPGC-2 and RMPGC training processes, models learn how to reconstruct images using information from different regions. Therefore, a good overall perception is maintained at all rates.

Finally, Fig. 17 compares the visual effects of different compression schemes using the Kodak dataset. We show that the proposed DIC models trained on the ImageNet dataset can generalize well to the Kodak dataset. For example, semantic salient regions such as parrots' eyes are better protected compared with other algorithms. In addition, RMPGC is shown to outperform all other schemes in terms of overall visual perception.

### *H. Transferability to different DNN backbones*

The proposed latent-based progressive DIC techniques are not limited to the LS-GAN backbone, but are also applicable to different DNN backbones and different datasets. The LS-GAN backbone was used solely for benchmarking. This backbone, however, yields unsatisfactory performance regarding distortion-related metrics. To demonstrate the transferability of the proposed techniques and also to obtain better performance, we report results on using the DNN backbone in MTB2018 [30]. Apart from the proposed structural changes required to build a RMPGC codec, the following modifications were made to the original backbone during implementation. First, the multi-scale decomposition (MSD) and inverse multi-scale decomposition (IMSD) [10] were added to generate the base layers and enhance layers. Second, the autoregressive mask convolution, which is applied to predict the latent vectors entropy parameters, was removed. Third, two independent Gaussian entropy models were adopted to estimate the entropy of the latent vectors of the base layers and enhance layers. We note that the use of Gaussian entropy models means that the rate control mechanism is slightly different from the one in LS-GAN. The input images were down-sampled by a factor of 16, and the output channel numbers of the base layers and enhance layers were both set to be 96.

To investigate the performance trade-off of the new RMPGC codec, we conducted experiments using the following cases for the hyperparameters (or regularization parameters):

(A) $\lambda_{\text{SFM}} = 0, \lambda_{\text{GAN}} = 0$;
(B) $\lambda_{\text{SFM}} = 1, \lambda_{\text{GAN}} = 0$;
(C) $\lambda_{\text{SFM}} = 1, \lambda_{\text{GAN}} = 0.5$.

Case (A) means the codec is solely optimized for the MSE loss (or distortion). Case (B) considers, besides distortion, a semantic SFM loss (for classification). Case (C), in turn, further takes into account the GAN loss for perception. We note that training in case (C) is performed by fine-tuning the model obtained in case (B) by freezing the encoder and adding the additional GAN loss. We found that this progressive training strategy yields good performance in practice. In all cases, we set $\lambda_{\text{R}}$ to 1, $\lambda_{\text{MSE}}$ to 5e-4 for reconstruction with only base layers, and $\lambda_{\text{MSE}}$ to 7e-3 for reconstruction with full latent vector. In all these cases, we used the Adam algorithm with a learning rate of 1e-4. Furthermore, to test the model transferability between different datasets, we used the COCO dataset [70] for training our RMPGC models, and the ImageNet dataset for testing.

The performance of the above three RMPGC codecs with a new DNN backbone were shown in Fig. 18 and compared with BPG, BMSHJ2018 [29], MTB2018 [30], and CHENG2020 [31]. We make the following observations.

(1) Case (A) yields similar performance to BPG in all the metrics evaluated. This confirms that the new DNN backbone allows achieving a performance better than the one achieved by the LS-GAN backbone.

(2) By adding an SFM loss to Case (A), the codec in case (B) outperforms the codec in Case (A) in all metrics. This confirms our observations from Fig. 11, which showed that a properly weighted SFM loss can improve the distortion, perception, and classification performance simultaneously. Compared with BPG, the codec in case (B) improves the Top-1 classification accuracy, Top-5 classification accuracy, NIQE, LPIPS, and PSNR by 15%, 8%, 20%, 8%, and 2%, respectively. The FSIM and SSIM metrics, however, are almost identical.

(3) The codec in case (C) demonstrates a performance trade-off between distortion and perception once the GAN loss is further introduced. Specifically, when going from case (B) to case (C), the NIQE, LPIPS, and FSIM metrics improve by 17%, 35%, and 0.7%, while the SSIM and PSNR metrics are degraded by 2.5% and 2.6%, respectively. The classification performance remains almost the same.

In summary, the proposed latent-based progressive coding design is a general framework that can be applied to different DNN backbones to yield RMPGC codecs that are practically competitive.

## VI. Conclusions and Future Work

We proposed a variable-rate DIC framework that shows structured progressive behavior from the lower-rate encoding of semantically-salient information to higher-rate encoding for full image fidelity. A generative DNN backbone is used to counterbalance rate-deficient distortions with statistical approximation. Three independent mechanisms have been introduced to yield a richly structured latent representation that supports parameterized control over the RDCP performance trade-off. Ablation studies show that the three proposed mechanisms are effective in securing performance gains in all tested metrics, including classification accuracy, PSNR, SSIM, FSIM, LPIPI, and NIQE. Relative performance gains ranging from 4% to 90% are observed in these metrics when averaged over the feasible data rate range. Comparisons with the classic image codec BGP and other existing DIC methods show that the proposed DIC schemes can effectively trade the overall PSNR for better semantic awareness and perceptual quality. By designing versatile adaptability into the latent space, the proposed DIC scheme is better suited for image compression in wireless communications, multi-user broadcasting, and multi-tasking applications.

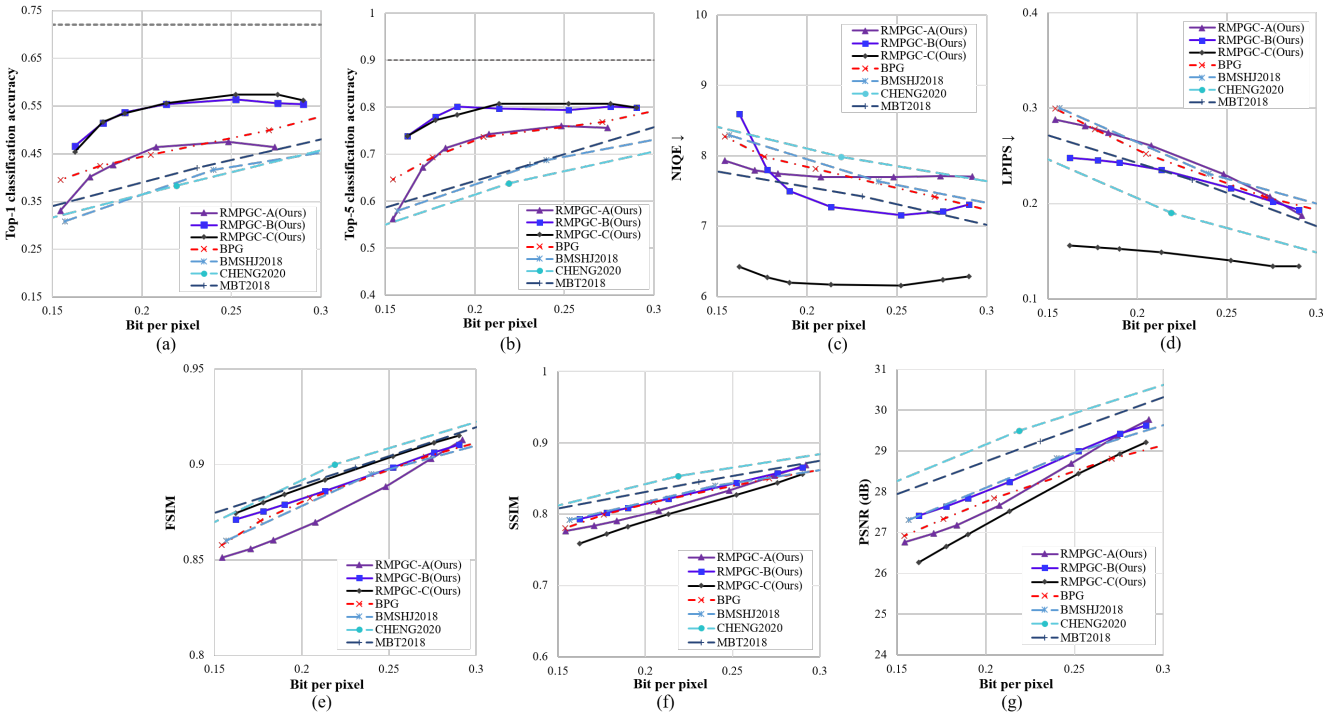



Fig. 18. Classification, perception, and distortion performances as functions of bpp for new RMPCs using the ImageNet dataset. RMPGC-A, RMPGC-B, and RMPGC-C mean training RMPGC uses case (A), case (B), and case (C), respectively.

In this paper, we are restricted to the problem of image source coding. For future work, it is worth exploring latent-exploration techniques such as contrastive learning [71] or distillation [72]. Moreover, a structured latent space makes it easier for our scheme to be extended for joint source-and-channel coding (see, e.g., the multiple description problem [53], [73]), which can inject further progressive capability to cope with opportunistic communication channels.

## REFERENCES


[1] F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L. Van Gool, "Conditional probability models for deep image compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, pp. 4394–4402, June 2018.

[2] F. Jiang, W. Tao, S. Liu, J. Ren, X. Guo, and D. Zhao, "An end-to-end compression framework based on convolutional neural networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 3007–3018, Oct. 2018.

[3] M. Ghanbari, *Standard Codecs: Image Compression to Advanced Video Coding*, 3rd ed. Edison, NJ, USA: IET, 2011.

[4] J. Balle, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, vol. 5, pp. 1–27, Apr. 2017.

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

[6] F. Yang, L. Herranz, and J. Weijer, "Variable rate deep image compression with modulated autoencoder," *IEEE Signal Process. Lett.*, vol. 27, pp. 331–335, Jan. 2020.

[7] Y. Choi et al., "Variable rate deep image compression with a conditional autoencoder," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, Korea, pp. 3146–3154, Oct. 2019.

[8] M. Song, J. Choi, and B. Han, "Variable-rate deep image compression through spatially-adaptive feature transform," in *Proc. IEEE/CVF Int. Conf. Comput. Vision (ICCV)*, Montreal, QC, CAN. pp. 2360–2369, Oct. 2021.

[9] R. Wang, Z. Sun, and S. Kamata, "Adaptive image compression using GAN based semantic-perceptual residual compensation," in *Proc. IEEE. Int. Conf. Pattern Recognit. (ICPR)*, Milan, Italy, pp. 9030–9037, Jan. 2021.

[10] C. Cai, L. Chen, X. Zhang, and Z. Gao, "Efficient variable rate image compression with multi-scale decomposition network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3687–3700, Dec. 2019.

[11] C. Cai, L. Chen, X. Zhang, and Z. Gao, "End-to-end optimized ROI image compression," *IEEE Trans. Image Process.*, vol. 29, pp. 3442–3457, Dec. 2020.

[12] G. Toderici et al., "Variable rate image compression with recurrent neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Juan, Puerto Rico, pp. 1-12, May 2016.

[13] C. Cai, L. Chen, X. Zhang, G. Lu, and Z. Gao, "A novel deep progressive image compression framework", in *Proc. Picture Coding Symp. (PCS)*, Ningbo, Zhejiang, China, pp. 1–5, Nov. 2019.

[14] G. Toderici et al., "Full resolution image compression with recurrent neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, pp. 5306–5314, July 2017.

[15] N. Johnston and D. Vincent, "Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, pp. 4385–4393, June 2018.

[16] 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, July 2020.

[17] T. Chen and Z. Ma, "Variable bitrate image compression with quality scaling factors," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Barcelona, Spain, pp. 2163–2167, May 2020.

[18] T. Chen et al., "End-to-end learnt image compression via non-local attention optimization and improved context modeling," *IEEE Trans. Image Process.*, vol. 30, pp. 3179–3191, Feb. 2021.

[19] Z. Cui, J. Wang, B. Bai, T. Guo, and Y. Feng, "G-VAE: A continuously variable rate deep image compression framework," 2020, arXiv:2003.02012 [Online]. Available: https://arxiv.org/abs/2003.02012

[20] J. Zhou, A. Nakagawa, and K. Kato, "Variable rate image compression method with dead-zone quantizer," in *Proc. IEEE/CVF Conf. Comput. Vision and Pattern Recognit. (CVPR)*, Seattle, WA, USA, pp. 624–628, June 2020.

[21] Y. Lu, Y. Zhu, Y. Yang, A. Said, and T. S. Cohen, "Progressive neural

image compression with nested quantization and latent ordering," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Anchorage, AK, USA, pp. 539–543, Sep. 2021.

[22] J. H. Lee, S. Jeon, K. P. Choi, Y. Park, and C. S. Kim, "DPIC-T: Deep progressive image compression using trit-planes," 2020, arXiv:2112.06334 [Online]. Available: https://arxiv.org/abs/2112.06334

[23] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[24] 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.

[25] L. Chen, C. Liu, X. Hong, C.-X. Wang, J. S. Thompson, and J. Shi, "Capacity and delay tradeoff of secondary cellular networks with spectrum aggregation," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3974–3987, June 2018.

[26] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L. Gool, "Generative adversarial networks for extreme learned image compression," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Seoul, Korea, pp. 221–231, Oct. 2019.

[27] L. D. Chamain et al., "End-to-end optimized image compression for machines, a study," in *Proc. IEEE Data Compress. Conf. (DCC)*, pp. 163–172, Mar. 2021.

[28] D. Liu, H. Zhang, and Z. Xiong, "On the classification-distortion-perception tradeoff," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Vancouver, CAN, pp. 1204–1213, Dec. 2019.

[29] J. Balle, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *Proc Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, CAN, Apr, 2018.

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

[31] 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, pp. 7936–7945, June 2020.

[32] 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, June 2020.

[33] 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, Aug. 2020.

[34] I. J. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Montreal, CAN, pp. 2672–2680, June 2014.

[35] F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L. Van Gool, "Practical full resolution learned lossless image compression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, pp. 10629–10638, June 2019.

[36] J. Wang et al., "Semantic perceptual image compression with a Laplacian pyramid of convolutional networks," *IEEE Trans. Image Process.*, vol. 30, pp. 4225–4237, Mar. 2021.

[37] O. Rippel and L. Bourdev, "Real-time adaptive image compression," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Sydney, Australia, vol. 34, pp. 2922–2930, Aug. 2017.

[38] L. Wu, K. Huang, and H. Shen, "A GAN-based tunable image compression system," in *IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Snowmass Village, CO, USA, pp. 2323–2331, Mar. 2020.

[39] F. Mentzer et al., "High-fidelity generative image compression," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, pp. 6–12, Dec, 2020.

[40] E. Collins, R. Bala, B. Price and S. Süsstrunk, "Editing in style: Uncovering the local semantics of GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, pp. 5770–5779, June 2020.

[41] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of GANs for semantic face editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, pp. 9240–9249, June 2020.

[42] S. Singh, S. Abu-El-Haija, N. Johnston, J. Ball, A. Shrivastava, and G. Toderici, "End-to-end learning of compressible features," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Abu Dhabi, United Arab Emirates, pp. 3349-3353, Nov. 2020.

[43] Y. Lou, L. Duan, and S. Wang, "Front-end smart visual sensing and back-end intelligent analysis: A unified infrastructure for economizing the visual system of city brain," *IEEE J. Sel. Areas Commun*., vol. 37, no. 7, pp. 1489–1503, May 2019.

[44] Z. Chen Z and T. He, "Learning based facial image compression with semantic fidelity metric," *Neurocomputing*, vol. 338, pp. 16–25, Apr. 2019.

[45] 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, June 2020.

[46] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NEV, USA, pp. 2921–2929, June 2016.

[47] 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 *Pro. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, pp. 618–626, Oct. 2017.

[48] 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, NEV, USA, pp. 839–847, Mar. 2018.

[49] R. Fu, Q. Hu, X. Dong, Y. Guo, Y. Gao, and B. Li, "Axiom-based Grad-CAM: Towards accurate visualization and explanation of CNNs," in *British Mach. Vis. Conf. (BMVC)*, UK, Sep. 2020.

[50] P. Jiang, C. Zhang, Q. Hou, M. Cheng, and Y. Wei, "LayerCAM: Exploring hierarchical class activation maps for localization," *IEEE Trans. on Image Process.*, vol. 30, pp. 5875–5888, June 2021.

[51] A. Prakash, N. Moran, S. Garber, A. DiLillo, and J. Storer, "Semantic perceptual image compression using deep convolution networks," in *Proc. IEEE Data Compress. Conf. (DCC)*, Snowbird, UT, USA, pp. 250–259, Apr. 2017.

[52] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp.1798–1828, Aug. 2013.

[53] L. Zhao, H. Bai, A. Wang, and Y. Zhao, "Multiple description convolutional neural networks for image compression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 8, pp. 2494–2508, Aug. 2019.

[54] T. M. Cover and J. A. Thomas, *Elements of information theory*, 2nd ed. Hoboken, NJ, USA: Wiley, 2006.

[55] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, pp. 770–778, June 2016.

[56] A. Bietti and J. Mairal, "Group invariance, stability to deformations, and complexity of deep convolutional representations," *J. Mach. Learn. Res.*, vol. 20, pp. 876–924, 2019.

[57] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Venice, Italy, pp. 2813–2821, Oct. 2017.

[58] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Salt Lake City, UT, USA, pp. 8798–8807, June 2018.

[59] E. Agustsson et al., "Soft-to-hard vector quantization for end-to-end learning compressible representations," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, Los Angeles, CA, USA, pp. 1141–1151, Dec. 2017.

[60] N. OSTU, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cybern.*, vol. 9, no. 1, pp. 62–66, Jan. 1979.

[61] Imagenet, [EB/OL], Available: http://www.image-net.org/

[62] Dogs vs. cats, [EB/OL], Available: https://www.kaggle.com/c/dogs-vs-cats/data

[63] Kodak photo cd, [EB/OL], Available: http://r0k.us/graphics/kodak/

[64] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, pp. 1–14, May 2015.

[65] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, pp. 1–15, May 2015.

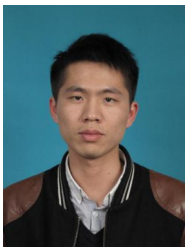
[66] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. on Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2021.

[67] R. Zhang, P. Isola, 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. (CVPR)*, Salt Lake City, UT, USA, pp. 586–595, June 2018.

[68] A. Mittal, R. Soundararajan, and A. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209-212, Feb. 2012.

[69] F. Bellard, *BPG Image Format*. Accessed: Nov. 12, 2019. [Online]. Available: https://bellard.org/bpg/

[70] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. European Conf. Comput. Vis. (ECCV)*, Zurich, Switzerland, pp. 740–755, Sep. 2014.
[71] K. He, H. Fan, Y. Wu, S. Xie, and R. B. Girshic, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, pp. 9726–9735, June 2020.
[72] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, arXiv:1503.02531 [Online]. Available: https://arxiv.org/abs/1503.02531
[73] D. B. Kurka and D. Gündüz, "Bandwidth-agile image transmission with deep joint source-channel coding," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 8081–8095, Dec. 2021.



**Zhongyue Lei** received the M.S. degree from University of Electronic Science and Technology of China, Chengdu, China, in 2014. He is currently pursuing the Ph.D. degree at the School of Informatics, Xiamen University, Xiamen, China. His current research interests include source coding and deep image compression.

**Peng Duan** received the M.S. degree from Chongqing University, Chongqing, China, in 2019. He is currently pursuing the Ph.D. degree at the School of Informatics, Xiamen University, Xiamen, China. His current research interests include semantic communications and deep learning.

**Xuemin Hong** 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 articles in refereed journals and conference proceedings. His current research interests include semantic communications, cognitive communication networks, and wireless localization systems.

**João F. C. Mota** received the M.Sc. and Ph.D. degrees in electrical and computer engineering from the Technical University of Lisbon in 2008 and 2013, respectively, and the Ph.D. degree in electrical and computer engineering from Carnegie Mellon University in 2013. He is currently an Assistant Professor of signal and image processing with Heriot-Watt University, Edinburgh. His research interests include theoretical and practical aspects of high-dimensional data processing, inverse problems, optimization theory, machine learning, data science, and distributed information processing and control. He was a recipient of the 2015 IEEE Signal Processing Society Young Author Best Paper Award.

**Jianghong Shi** 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.

**Cheng-Xiang Wang** (Fellow, IEEE) received the B.Sc. and M.Eng. degrees in communication and information systems from Shandong University, China, in 1997 and 2000, respectively, and the Ph.D. degree in wireless communications from Aalborg University, Denmark, in 2004.

He was a Research Assistant with the Hamburg University of Technology, Hamburg, Germany, from 2000 to 2001, a Visiting Researcher with Siemens AG Mobile Phones, Munich, Germany, in 2004, and a Research Fellow with the University of Agder, Grimstad, Norway, from 2001 to 2005. He has been with Heriot-Watt University, Edinburgh, U.K., since 2005, where he was promoted to a Professor in 2011. In 2018, he joined Southeast University, Nanjing, China, as a Professor. He is also a part-time Professor with Purple Mountain Laboratories, Nanjing. He has authored 4 books, 3 book chapters, and more than 460 papers in refereed journals and conference proceedings, including 25 highly cited papers. He has also delivered 23 invited keynote speeches/talks and 9 tutorials in international conferences. His current research interests include wireless channel measurements and modeling, 6G wireless communication networks, and electromagnetic information theory.

Prof. Wang is a Member of the Academia Europaea (The Academy of Europe), a Fellow of the Royal Society of Edinburgh, IEEE, IET, and China Institute of Communications (CIC), an IEEE Communications Society Distinguished Lecturer in 2019 and 2020, and a Highly-Cited Researcher recognized by Clarivate Analytics in 2017-2020. He is currently an Executive Editorial Committee Member of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He has served as an Editor for over ten international journals, including the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, from 2007 to 2009, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, from 2011 to 2017, and the IEEE TRANSACTIONS ON COMMUNICATIONS, from 2015 to 2017. He was a Guest Editor of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, Special Issue on Vehicular Communications and Networks (Lead Guest Editor), Special Issue on Spectrum and Energy Efficient Design of Wireless Communication Networks, and Special Issue on Airborne Communication Networks. He was also a Guest Editor for the IEEE TRANSACTIONS ON BIG DATA, Special Issue on Wireless Big Data, and is a Guest Editor for the IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING, Special Issue on Intelligent Resource Management for 5G and Beyond. He has served as a TPC Member, a TPC Chair, and a General Chair for more than 80 international conferences. He received 14 Best Paper Awards from IEEE GLOBECOM 2010, IEEE ICCT 2011, ITST 2012, IEEE VTC 2013Spring, IWCMC 2015, IWCMC 2016, IEEE/CIC ICCC 2016, WPMC 2016, WOCC 2019, IWCMC 2020, WCSP 2020, CSPS2021, and WCSP 2021. Also, he received the 2020–2022 "AI 2000 Most Inuential Scholar Award Honourable Mention" in recognition of his outstanding and vibrant contributions in the field of Internet of Things.