Towards Task-Generic Image Compression: A Study of Semantics-Oriented Metrics

Changsheng Gao, Dong Liu, *Senior Member, IEEE*, Li Li, *Member, IEEE*, and Feng Wu, *Fellow, IEEE*

***Abstract*—Instead of being observed by human, multimedia data are now more and more fed into machines to perform different kinds of semantic analysis. One image may be analyzed multiple times by different machine vision algorithms for different purposes. While machine vision-oriented image compression has been studied, the existing methods are usually driven by a specific machine vision task, and may not be applicable for the other tasks. We address the *task-generic* image compression, in the hope that an image is compressed once but used multiple times for different tasks, all with satisfactory performance. Our study is based on the end-to-end learned image compression. We focus ourselves on the distortion metric, i.e., finding out a task-agnostic metric to estimate the quality of reconstructed images. On the one hand, we study deep feature distance as the metric, which transforms images into a latent space by a pretrained convolutional network—the latent space is believed to be more aligned to semantics—and calculates distance in the latent space. On the other hand, inspired by the saliency mechanism, we study an importance-weighted pixel distance as the metric, where the weights are generated to reflect the importance of the pixels to semantics. Moreover, we combine the two distances into one metric to investigate their complementary nature. An extensive set of experiments are performed to evaluate these metrics. Experimental results show that using the combined metric performs the best, and leads to 20.79%∼42.69% bits saving under the same semantic analysis performance, compared to using signal fidelity metrics. Interestingly, we observe that using the combined metric also improves the visual quality of the reconstructed images.**

***Index Terms*—Deep feature distance, importance-weighted pixel distance, learned image compression, machine vision, task generic.**

I. Introduction

In RECENT years, the vigorous development of intelligent applications such as security surveillance and autonomous driving leads to an explosive growth of image and video data. The rapid growth of image data not only puts tremendous pressure on transmission and storage but also poses huge challenges for data analysis. In typical intelligent applications, as shown in Fig. 1, the acquisition, compression, and analysis processes of images constitute the complete pipeline, and more and more image data are fed into machines and analyzed by intelligent analysis algorithms to liberate manpower. In order to adapt to analysis algorithms, the compression methods shall take semantic fidelity into consideration during the optimization process. An image is usually analyzed multiple times for different semantic analysis tasks to support a variety of applications in actual deployment. Therefore, how to design a task-generic image compression method becomes an important problem.

A series of coding standards (*e.g.,* JPEG [1], JPEG2000 [2], and HEVC-Intra [3]) have been built to significantly improve the image compression efficiency. However, there are two drawbacks to traditional image compression methods. First, traditional image compression methods are optimized for signal fidelity, such as peak signal-to-noise ratio (PSNR), and do not take semantics into consideration. These methods completely isolate the compression and analysis processes, thereby prevents them from being better jointly optimized in actual intelligent applications. Besides, it is reported in [4] that high signal fidelity does not necessarily lead to better semantic fidelity. Second, traditional image compression methods are usually block-based methods, which divide the image into many sub-blocks and optimize each sub-block separately. On the other hand, most semantic metrics are not block-wise, e.g. object detection accuracy is defined on the whole object rather than each block of the object. This makes it difficult to optimize semantic fidelity in traditional image compression methods.

Recently, end-to-end image compression has been extensively studied. The end-to-end optimization approach makes it easy to optimize for a given objective. By taking advantage of this benefit, methods in [5], [6] achieve better perceptual quality at the same bit rate by selecting an appropriate perceptual metric. Inspired by this, some researchers introduce loss functions of semantic analysis tasks into compression and optimize compression and semantic analysis tasks jointly. For example, Luo *et al.* [7] embed the classification network as a semantic metric into the compression network. The parameters of the compression network are optimized under the constraints of bit rate, signal distortion, and classification accuracy during the training. Compared with conventional compression methods, the image reconstructed by this method achieves higher accuracy in the classification task at the same bit rate. This method demonstrates the advantages of task-driven image compression methods. However, there are two limitations with task-driven image

Manuscript received 13 June 2021; revised 21 October 2021; accepted 22 November 2021. Date of publication 25 November 2021; date of current version 9 March 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFA0701603, and in part by the Natural Science Foundation of China under Grants 62022075, 62036005, and 62021001. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Sanjeev Mehrotra. *(Corresponding author: Dong Liu.)*

The authors are with the CAS Key Laboratory of Technology in Geo-Spatial Information Processing and Application System, University of Science and Technology of China, Hefei 230027, China (e-mail: gcs@mail.ustc.edu.cn; dongeliu@ustc.edu.cn; lil1@ustc.edu.cn; fengwu@ustc.edu.cn).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TMM.2021.3130754.

Digital Object Identifier 10.1109/TMM.2021.3130754

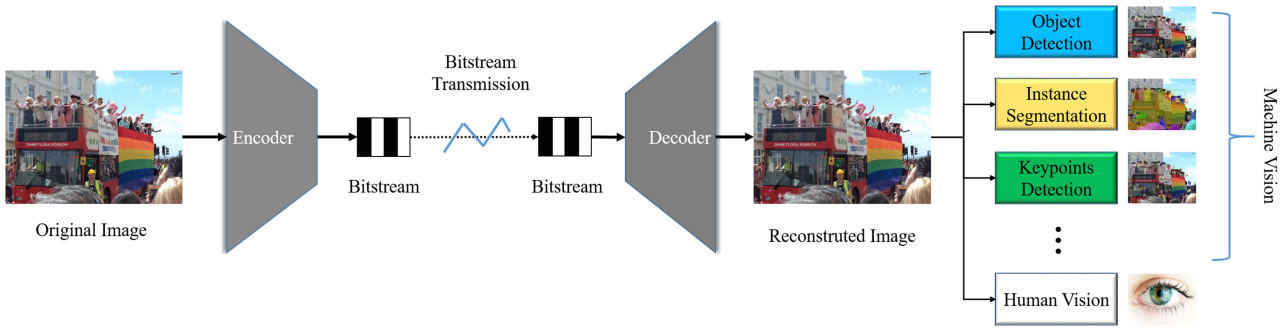


Fig. 1. The considered scenario that requires task-generic image compression. One image is compressed once but used multiple times for different machine vision tasks, and for human observation optionally.

compression methods. First, it can be hardly guaranteed that an encoder trained for a particular semantic analysis task is effective for other semantic analysis tasks. Second, it is difficult to ensure that this approach can be applied to different algorithms for the same semantic analysis task. Task-driven image compression methods consider a specific semantic analysis task or even a specific semantic analysis network, which may limit the generalization ability of the method.

Although end-to-end image compression has the advantage of being easy to optimize for a given target, there has been no research on task-generic image compression to our best knowledge. In this paper, we study the task-generic image compression method on the basis of end-to-end image compression. The key problem we are facing is to find out an appropriate metric to train the image compression networks. Based on our previous studies, if a metric was tied to a specific task or several specific tasks, the metric would have difficulty in generalizing well to the other tasks. Thus, to pursue task-generic image compression, we think it may be better to keep the metric agnostic to the tasks. We consider the task-agnostic semantic metric from two perspectives. First, we measure the distortion from a latent space that is expected to align to semantics better, and second, we expect the distortion calculated can describe the relative importance of pixels to semantics. For the first idea, we use the nonlinear transforms in VGG-16 [8] to transform images into deep features. Then we take the distance between the features transformed from the original image and those transformed from the reconstructed image as the deep feature distance. For the second idea, we reuse nonlinear transforms in VGG-16 to generate a semantic importance map that indicates the importance of pixels to semantics. And then we take the importance values as weights to calculate the weighted mean-squared-error (MSE) of images as the importance-weighted pixel distance. We then combine the deep feature distance and importance-weighted pixel distance into a single metric, and use the metric together with coding rate to optimize an end-to-end image compression network.

In summary, in this paper we make the following contributions:

- To the best of our knowledge, we are the first to investigate the research problem of task-generic image compression. In practical intelligent applications, reconstructed images need to support high-precision analysis for a variety of semantic analysis tasks. However, the existing methods only focus on a specific semantic analysis task, ignoring the generic requirement of semantics-oriented image compression.
- We provide a feasible solution to the task-generic image compression problem that considering the problem from the perspective of semantic distortion measurement. We explore the influence of various metrics on semantic fidelity and propose a task-agnostic metric.
- We apply the proposed task-agnostic metric to an end-to-end image compression method [9], and the experimental results show that the end-to-end image compression method optimized with the proposed semantic metric achieves $20.79\% \sim 42.69\%$ bit-rate reduction under the same semantic distortion in various semantic analysis tasks.

The reminder of this paper is organized as follows. Section II gives a brief review of related work. In Section III, we propose two semantic distances and apply them in an end-to-end image compression method. Section IV introduce the experimental settings, and the experimental results and ablation studies are presented in Section V. Section VI concludes this paper.

Our code and models are available online.[1]

II. Related Work

In recent years, deep learning-based semantics-oriented compression has attracted more and more attention. In this section, we introduce semantics-oriented compression and learned image compression.

A. Semantics-Oriented Compression

According to the compressed object, we classify semantics-oriented compression methods into two categories: one is semantics-oriented image compression, and the other is feature compression. In this part, we first introduce semantics-oriented image compression. After that, feature compression is introduced.

1) Semantics-Oriented Image Compression: Semantics-oriented image compression is not only studied in traditional image compression but also widely studied in deep learning-based image compression. Traditional image compression methods mainly use bit allocation to achieve high

[1][Online]. Available: https://github.com/chansongoal/semantic_image_compression.

semantic fidelity, while the methods used in deep learning-based image compression are diverse. In addition to bit allocation methods, semantic metric-based methods and generative adversarial network (GAN) based methods are also used in deep learning-based image compression.

Bit allocation-based compression methods are proposed in [10]–[17] to solve the semantics-oriented image compression problem. In [10], they first use the gradient-weighted class activation mapping (Grad-CAM) [18] to produce a localization map highlighting the important regions for predicting the concept. Then based on the map, the reinforcement learning method is used to determine the quantization parameter of each coding block. Compared with the HEVC encoder HM version 16.9, the method in [10] achieves great bit rate savings in various tasks such as classification, object detection, and semantic segmentation. However, the proposed method needs to customize a reward for every semantic analysis task, which failed to guarantee its generalization ability. In [11]–[17], the quantization module of JPEG is modified to make the compressed images more applicable to semantic analysis algorithms. For example, Choi *et al.* [13] propose an approach that learns a quantization network to estimate image-specific semantic quantization tables fully compatible with the standard JPEG codec. The learned semantic quantization tables are applied in the image compression process. However, due to the limitation of JPEG compression efficiency, a high compression ratio can hardly be achieved.

In [7], [19], [20], the authors propose a task-driven end-to-end image compression methods. In [20], the object detection network faster region-based convolutional neural networks (Faster R-CNN) [21] is integrated into to a compression network to evaluate a part of the loss function. First, they fix the analysis network and finetune the compression network. The compressed images from the finetuned compression network can achieve higher recognition accuracy under the same bit rate. Then, this work carries out joint training on the compression network and the analysis network. The recognition accuracy of images reconstructed by the jointly trained compression network is further improved. However, the efficiency of this method is only verified on Faster R-CNN and its generalization ability cannot be guaranteed. A GAN-based method proposed in [22] generates images with semantics close to the original image at an extremely low bit rate by utilizing the powerful generation ability of GAN. But the extremely low bit rate application scenario limits this method to be widely used. In addition, methods such as [23]–[25] also take GAN as an effective means to explore semantics-oriented image compression. In [26]–[28], researchers consider semantic compression problem from the perspective of image quality assessment. For example, a semantic metric is designed for optical character recognition in [27]. The encoder optimized by the proposed semantic metric can achieve a higher compression ratio for specific semantic analysis tasks.

The above methods either optimize for one semantic analysis task or multiple tasks individually in the same way. Although these methods can improve accuracy, they do not consider generalization ability in the design, which results in the failure of these methods to meet the requirement of being generic.

2) Feature Compression: Deep learning-based analysis algorithms have been widely deployed in many intelligent analysis application scenarios. Due to the analysis that is actually conducted on features extracted from images, some studies [29]–[36] propose to directly encode the features instead of the images. These methods compress the semantic information more directly and reduces the complexity of feature extraction at the decoder side. In [29], the authors focus on collaborative object detection and study the impact of both near-lossless and lossy compression of feature data on its accuracy. They also propose a strategy for improving the accuracy under lossy feature compression. In [33], [34], the authors propose to compactly represent and convey the intermediate-layer deep features with high generalization capability. This strategy enables a good balance among the computational load, transmission load, and the generalization ability for cloud servers when deploying the deep neural networks for large-scale cloud-based visual analysis. Singh *et al.* [35] propose a learning-based method that jointly optimizes for compressibility along with the task objective for learning the features to be extracted. They present results on multiple benchmarks and demonstrate that their method produces features that are an order of magnitude more compressible while having a regularization effect that leads to a consistent improvement in accuracy. Duan *et al.* [36] carry out exploration in a new area, video coding for machines (VCM), attempting to bridge the gap between feature coding for machine vision and video coding for human vision. The feature compression methods usually cannot guarantee satisfactory reconstruction of the images. In this paper, we focus on task-generic image compression, which obtains better semantic quality for different vision tasks.

B. Learned Image Compression

According to the network structure, learned image compression can be divided into two categories: recurrent neural network (RNN) based and convolutional neural network (CNN) based image compression. Toderici *et al.* propose the first end-to-end image compression method based on RNN [37]. After that, the RNN-based end-to-end image compression methods [38], [39] surpass BPG in the multi-scale structural similarity (MS-SSIM). Balle *et al.* first propose a fully convolutional end-to-end image compression structure [9]. By successively introducing the hyper-prior model [40] and the autoregressive model [41], the CNN-based end-to-end compression methods surpass BPG in PSNR. CNN-based network structure is adopted by most subsequent end-to-end image compression methods [42]–[46].

Compared with traditional image coding, the most important feature of the end-to-end solution is the joint optimization through gradient back propagation. This enables the learned image compression method to further improve its performance [47]–[49]. More importantly, it makes it easier to extend the optimization objectives of end-to-end image compression, such as perception-oriented compression [5], [50], [51], machine algorithm-oriented compression [52], [25], scalable compression [53], [54], variable rate compression [55]–[57], etc.

III. Proposed Image Compression Method

In this section, we propose to solve the task-generic image compression problem based on end-to-end image compression.

We first analyze the compression problems in intelligent applications. And then we introduce the inspirations for semantic distortion measurement. Next, we establish two semantic distance measurement methods and combine them into a single semantic metric. Finally, we apply the proposed semantic metric into an end-to-end image compression optimization procedure.

A. Requirement Analysis

In typical intelligent application scenarios, as shown in Fig. 1, cameras capture images and compress them into bitstreams. The compressed bitstreams are then transmitted to the cloud data center. The images are reconstructed from bitstreams and then fed into various semantic analysis algorithms or provided to human. In intelligent applications, the acquisition, compression, and analysis processes of images constitute a complete pipeline. However, traditional image compression methods (such as JPEG, HEVC, etc.) completely isolate image compression and semantic analysis. The downstream semantic analysis tasks are not considered in the compression stage. As a result, high analysis accuracy is hard to be achieved even though high signal fidelity is achieved. Thus, semantic fidelity shall be considered in image compression in intelligent applications.

Generally, image data in the cloud data center need to support a variety of intelligent applications. Therefore, an image may be analyzed multiple times for different semantic analysis tasks. For example, the reconstructed image should not only be able to achieve high object detection accuracy to help the traffic police to schedule traffic, but also be able to perform well on person keypoints detection task to help the police detect possible criminal behavior of pedestrians. Moreover, in some application scenarios, such as the identification of criminals, reconstructed images also need to be of high subjective quality for human confirmation. There may be many misjudgments and missed judgments if the judgment is totally up to analysis algorithms. This puts forward further requirements for semantics-oriented image compression: task-generic image compression. If the task-generic characteristic cannot be guaranteed, then an image needs to be compressed multiple times for different applications. This will greatly increase the cost of data storage and transmission, resulting in the increase of intelligent application deployment cost. Therefore, the task-generic image compression method is of great significance to the actual deployment of intelligent applications.

B. Rate-Distortion Optimization

According to Shannon's rate-distortion theory [58], the essence of compression is the trade-off between rate and distortion. The objective of compression is to minimize the distortion under the rate constraint:

$$\min D, \text{ s.t. } R \leq R_C \tag{1}$$

where R_C is the bit rate budget. D and R are the distortion and bit rate of compressed images, respectively. We can see from (1) that the distortion D determines the optimization direction. In the previous work of image compression, MSE is widely adopted as the distortion metric. Several previous methods adopted MS-SSIM as the distortion metric to train image compression models, where the resulting models were said to obtain higher visual quality than those trained by MSE [40]. In this paper, we want to find out an appropriate metric so that the compression methods are optimized for semantic fidelity.

Different from traditional image compression, all modules of end-to-end image compression are connected and can be optimized jointly by gradient back propagation. This end-to-end optimization method facilitates the network to optimize for a given objective. By taking advantage of this characteristic, the method in [9] achieves higher signal fidelity by optimizing the compression network with MSE. And better perceptual fidelity is achieved in [6] by using the GAN loss. However, the task-generic image compression has not been studied yet.

We can conclude that the difference between semantic fidelity-oriented image compression and signal fidelity-oriented image compression is the distortion measurement D. Therefore, we study the task-generic image compression problem by investigating the distortion measurement. To make sure that the distortion D can describe semantic distortion, we consider it in a task-agnostic fashion. First, inspired by a straightforward idea that semantic distortion can be better measured in a space that is more aligned to semantics, we propose to compute semantic distortion in a transformed latent space that is more aligned to semantics. The second idea is that we directly compute the semantic distortion in the pixel domain, but in a semantics-oriented way. Different from MSE, in which all pixels are treated equally, we assign each pixel a weight according to its semantic importance.

Based on the above two ideas, we design two semantic distortion measurement methods. We describe the technical details of these two methods in the next subsections.

C. Distortion Metrics

1) Deep Feature Distance: The key of the first method is to find a transform that can transform the image into a latent space that is more aligned to semantics. Inspired by deep learning-based semantic analysis algorithms, in which images are first transformed into deep features and then analysis is conducted on these features, we believe that the feature space is more aligned to semantics required for analysis than image space. Taking the classification task as an example, the VGG-16-based classification algorithm [8] can be divided into two steps: feature extraction and feature classification. In the first step, the 13 convolutional layers transform the $224\times224\times3$ image into $14\times14\times512$ feature maps. In the second step, the extracted feature maps are fed into a classifier (three fully-connected layers) to conduct classification. The actual classification process is performed on the transformed feature maps. This shows that the transformed feature maps are easier to be classified than the original images. Therefore, we believe that the transformed feature space is more aligned to semantics needed for analysis algorithms than the original image space. We take the nonlinear transforms in VGG-16 as our proposed transforms due to its wide usage in various semantic analysis tasks, such as object detection, image captioning, and so on.

The VGG-16 consists of 13 convolutional layers, 5 max-pooling layers, and three fully-connected layers. In

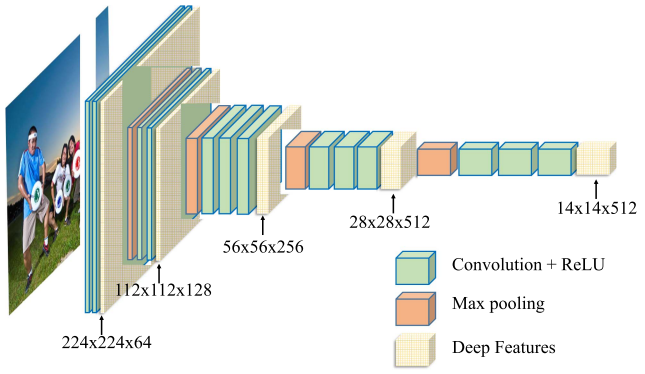


Fig. 2. Illustration of the deep features. We divide the convolutional layers of VGG-16 into 5 stages. For each stage, we use the intermediate output (indicated by arrows) as a set of deep features. The proposed deep feature distance is computed on the concatenation of all the 5 sets of deep features.

the following, the VGG-16 is considered as a feature extractor with 13 convolutional layers. As shown in Fig. 2, we divide the 13 convolutional layers of the VGG-16 into 5 stages. Then the VGG-16 can be regarded as 5 cascaded transforms. Each one transforms the input space into another space. It is generally believed that different transformed spaces contain semantic information with different granularity. In order to take a variety of semantic analysis tasks into account, we propose a multi-scale semantic distortion measurement method. The proposed distortion measurement method first calculates the Euclidean distance of the transformed original and reconstructed images in each transformed space. Then, we average the five distances as the deep feature distance L_f:

$$L_f(I,\bar{I}) = \text{M-MSE}(G(I),G(\bar{I})) = \frac{1}{K}\sum_k \left(\frac{1}{C_k M_k N_k}\sum_{c,m,n}(S_{cmn}-S'_{cmn})^2\right) \quad (2)$$

where k denotes the transform space order and K is equal to 5. S_{cmn} and S'_{cmn} are the values at the position (m,n) on the c-th channel of the transformed features of the original image and reconstructed image, respectively.

Note that we simply average the five distances in (2). For different tasks, the distances may have different importance levels, *e.g.*, high-level tasks would require rich features in deep layers, and vice versa. It suggests one use a weighted average instead of (2). However, in this paper, we want to design a task-agnostic metric, so we keep equal weights. The issue of different weights may be considered in the future work.

Although the VGG-based feature loss has been used as a *perceptual loss* in several tasks [59]–[61], the intention to use VGG-based feature loss in this paper is different. First, we want to measure the semantic distortion in a latent space that is more aligned to semantics. Second, we adopt the VGG network for the reason that VGG has been successfully used as a backbone for various semantic analysis tasks.

2) Importance-Weighted Pixel Distance: The key of the second method is to weight the distortion of all pixels in an image according to their importance to semantics. In other words, this method calculates the semantic distortion according to a semantic importance map. Therefore, the problem now is to generate an importance map that can indicate the semantic importance of pixels in an image. However, it is obvious that the required importance map is not the same for different semantic analysis tasks. We take the images shown in Fig. 3 as an example to demonstrate this point. The object detection and instance segmentation tasks only focus on some parts of an image. The person keypoints detection only cares about persons in an image while the image captioning task performs words generation based on the entire image. For example, the grass is crucial for the algorithms to generate the word *park* in the image captioning task. In this paper, we do not design an individual importance map for each analysis task, but instead we want to design a task-agnostic importance map by referring to a general backbone such as VGG16.

We propose a VGG-16-based importance map generation method in this work. VGG-16 uses the Rectified Linear Unit (ReLU) as its nonlinear activation function. The ReLU is a piecewise linear function that outputs the input directly if it is positive, otherwise, it outputs zero. Accordingly, the values of the transformed features in VGG-16 are greater than or equal to zero. It is believed that values greater than 0 are more important for semantic analysis tasks. Inspired by this, we propose to generate the importance map based on the feature maps of the VGG-16. The procedure of the importance map generation is shown in Fig. 4. First, we feed the image into the network and obtain 512 16× down-sampled feature maps from the last convolutional layer. The last convolutional layer has the largest number of convolution kernels, which makes it contain the most abundant high-level semantic information. We conduct point-wise addition operations on these 512 feature maps to fuse all high-level semantic information together. We then normalize the values of the merged feature map to the range between 0 and 255. Second, we up-sample the down-sampled feature map to the original image size using bicubic interpolation. Third, we binarize the resized feature maps to get a binary importance map. Specifically, we set the values greater than the average of the feature map (foreground) to 1, and the others (background) to a small value, such as 0.3. The binarization makes the importance map more locally consistent. Finally, we dilate the binary importance map to obtain the final importance map. The dilation operation increases the size of the region of interest and further improves the local consistency of the importance map. With the importance map, we compute the importance-weighted pixel distance L_p as:

$$L_p(I,\bar{I}) = \text{MSE}(P(I),P(\bar{I})) = \frac{1}{MN}\sum_{m,n}((I_{mn}-I'_{mn})\times P_{mn})^2 \quad (3)$$

where I_{mn} and I'_{mn} represent the pixel values of the m-th row and n-th column of the original image and reconstructed image, respectively. $P_{m,n}$ denotes the value in the m-th row and n-th column of the importance map.

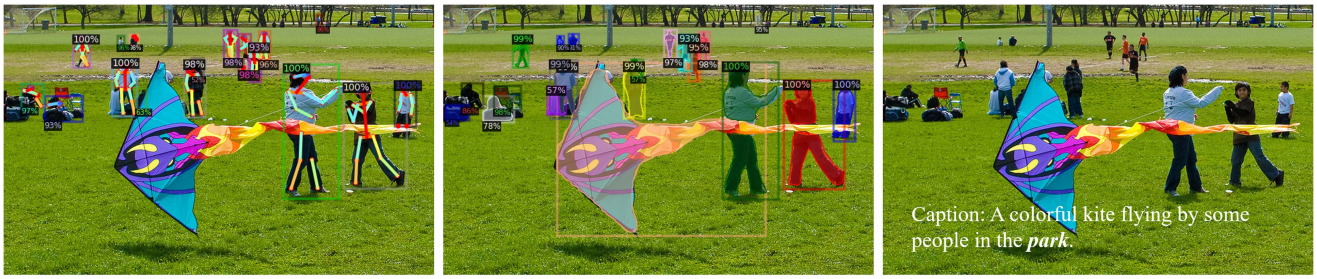


Fig. 3. Visualization of semantic analysis results of different tasks. (a) Results of person keypoints detection; (b) Results of object detection and instance segmentation; (c) Results of image captioning. Usually, the foreground objects are more important for semantic analysis, especially for detection and segmentation tasks. But image captioning also concerns the background. For example in (c), the grass in the image is crucial to infer the concept *park*.

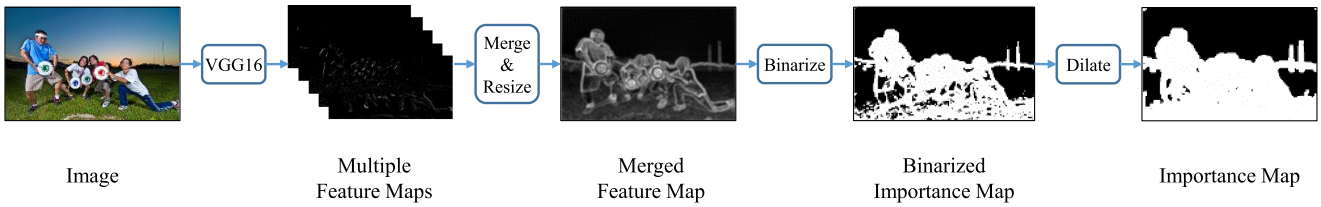


Fig. 4. Illustration of the proposed importance map generation method.

Note that we set the background value to 0.3 instead of 0 due to the following reasons. First, totally ignoring background cannot handle the semantic analysis tasks that require background. Second, it cannot ensure that the encoder reconstructs the entire image instead of only the foreground. Therefore, we set importance values of the background to 0.3 rather than 0 in the training phase. This modification introduces another benefit. It weakens the influence introduced by an inaccurate importance map.

3) Proposed Metric: Although the two semantic distances proposed above are both based on VGG-16, their characteristics are essentially different. The deep feature distance is a globally equal distance while the importance-weighted pixel distance is a locally equal distance. Take the feature maps extracted by the first convolutional layer as an example, each element on it corresponds to a 3×3 area of the original image. The constraint on an element of the feature map is equivalent to the constraint on a 3×3 area in the original image. Therefore, the MSE of all transformed feature maps can constrain the global semantic information of the original image equally. On the other hand, since we set importance values of background to 0.3, the importance-weighted pixel distance pays more attention to a local region (foreground) of the image. We regard the importance-weighted pixel distance as a locally equal distance.

The proposed two semantic distances have their own shortcomings. Although the VGG-16 can be used as a backbone in various semantic analysis tasks, it cannot guarantee that the transformed features are linear with respect to semantics. As far as we know, there is no work that can define a semantically linear space. So we try the deep feature distance on the basis of VGG-16 in this work. On the other hand, the effectiveness of importance-weighted pixel distance depends on the accuracy of the importance map. However, it is difficult to define an importance map that perfectly matches the semantic information, let alone generating such an importance map. So we try the VGG-16-based importance map in this work to preserve the semantic information of images.

In view of the characteristics and shortcomings of the above two semantic distances, we propose to combine the globally equal distance and locally equal distance as one semantic metric L_s:

$$L_s(I, \bar{I}) = L_p(I, \bar{I}) + \lambda_{fp} \times L_f(I, \bar{I}) \tag{4}$$

where λ_{fp} is the ratio of deep feature distance to importance-weighted pixel distance.

D. Proposed Image Compression Method

The focus of this work is to propose a semantic metric for analysis algorithms rather than a new end-to-end image compression framework. Therefore, instead of designing specific compression modules or algorithms, we make some modifications to an existing compression network [9]. As shown in Fig. 5, we retain the main structures of the proposed compression network in [9] and make two modifications to it.

First, we modify the input of the nonlinear transform. We concatenate the original image, importance map, and importance image together and feed them into the nonlinear transform. The importance image is obtained by multiplying the original image and the importance map pixel by pixel. The importance map equips the proposed compression network with the capability of bit allocation. Different from traditional codecs, in which the bit allocation is implemented by adjusting the quantization step during compression, the importance map-based method can achieve it in a simple way. More importantly, we can achieve bit

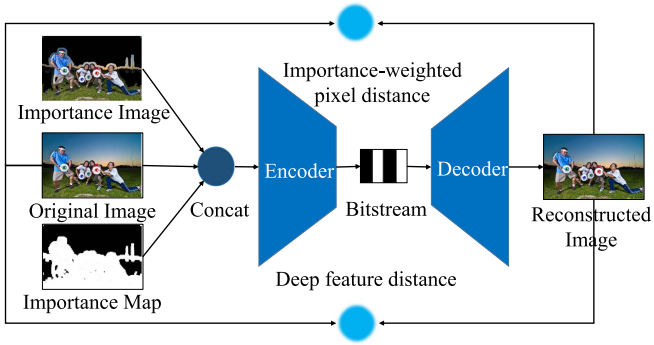


Fig. 5. The proposed framework for learned image compression. We concatenate the original image, the generated importance map, and the importance image as input. The training loss consists of deep feature distance, importance-weighted pixel distance, and bit rate.

allocation without modifying the structure of the compression network. Second, we modify the distortion metric of the proposed compression network. We substitute the MSE with the proposed semantic metric during the training phase. And the final loss function of the proposed compression network is:

$$L(I, \bar{I}) = L_s(I, \bar{I}) + \lambda_r \times R \tag{5}$$

where R denotes the bit rate of the compressed image. λ_r is the weight of bit rate.

IV. Experimental Settings

In this section, we first present the dataset used in this work. Then the configuration of the proposed encoder and the setting of training parameters are introduced in detail. Next, we clarify the network structures and pre-trained models used for each semantic analysis task.

A. Semantic Analysis Tasks and Dataset

There are so many different kinds of semantic analysis tasks and new tasks are still emerging. It is almost impossible to enumerate every task and to verify whether an image compression method performs well on every task. There have been some preliminary studies about the relations between different vision tasks, revealing that different tasks still have some commonality [62]. In this paper, we consider four semantic analysis tasks: object detection, instance segmentation, person keypoints detection, and image captioning. We select these four tasks for two reasons. First, the selected tasks are fundamental and widely studied in the recent years. For example, objects are required in many semantic analysis tasks such as semantic segmentation, object localization, and so on. Besides, the most fundamental image classification task is implicitly performed in object detection. Second, the selected tasks include both low-level and high-level tasks. The image captioning task is often regarded a high-level task, while detection and segmentation are regarded low-level tasks. Through these tasks, we want to examine the generalization ability of the proposed image compression method to some extent. However, it is still difficult to judge whether an image compression method is task-generic or not.

The Common Objects in Context (COCO) dataset is one of the most popular publicly available object recognition databases. In this work, we take COCO 2014 dataset [63] to verify the performance of the proposed algorithm. The four selected semantic analysis tasks are supported by COCO dataset. For the purpose of simplicity, we regard object detection, instance segmentation, and person keypoints detection as detection tasks in this paper. We report the standard COCO metrics AP (average precision) for detection tasks. The BLEU_1, MENTOR, ROUGE_L and CIDEr metrics are reported for the image captioning task. We train the compression networks on COCO 2014 training set (80 K images) and report the evaluation results on minVal2014 (5 K images).

B. Parameter Settings

We train the compression networks for 20 epochs using Adam optimization [64] for all bit rates. We initialize the learning rates to 0.0001 and 0.001 for the main and entropy optimizer, respectively. The other parameters are set to the same values as the default Adam configuration [64]. The batch size is set to 16. The VGG-16 networks are initialized with the pre-trained model proposed in [8] and are not optimized during the training phase. The other parts of the compression networks are initialized with the uniform initializer. We scale the training images to 256×256 to facilitate the training process. In the inference phase, we crop the validation images to a multiple of 16 and feed them to the encoder. For example, a 500×374 image is cropped to 496×368. Since the cropped pixels do not exceed 15, we believe that this operation has little impact on the accuracy of semantic analysis.

C. Network Structures

In recent years, semantic analysis tasks have been extensively studied, and many learning-based algorithms [21], [65]–[68] have been proposed. We use the network in [66] to verify the effectiveness of the proposed algorithm on the image captioning task. It first extracts the features through VGG-16, and then further processes the extracted features to obtain short sentences describing the image. In order to demonstrate that the proposed algorithm does not depend on a specific backbone, we also verify other semantic analysis tasks whose backbone is not VGG-16. Specifically, we conduct experiments on the open-source model Detectron2[2] provided by Facebook, which implements Faster R-CNN [21], Mask R-CNN [65] and Keypoints R-CNN [65] for object detection, instance segmentation, and person keypoints detection, respectively. We evaluate two different backbones for all these three tasks. Specifically, ResNet50-C4 and ResNet50-FPN for object detection and instance segmentation, and ResNet50-FPN and ResNet101-FPN for person keypoints detection.

V. Experimental Results

In this section, we first compare the proposed compression method with existing methods and then conduct ablation studies to show improvement with each of the proposed components.

[2][Online]. Available: https://github.com/facebookresearch/detectron2.

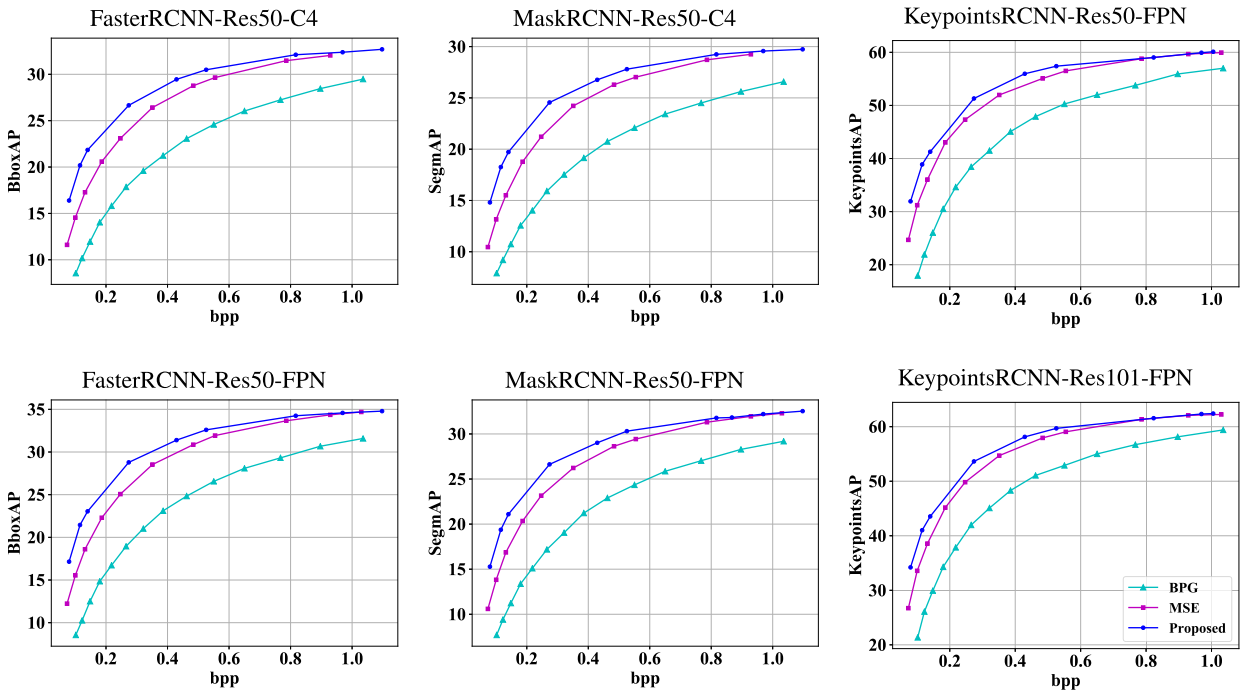


Fig. 6. Comparison of BPG and two learned image compression models. The two models are trained with MSE and the proposed distortion metric, respectively. The results from left to right correspond to object detection, instance segmentation, and person keypoints detection. For each task we evaluate two different algorithms shown in the top row and the bottom row, respectively.

As we are the first to propose an image compression method for a variety of semantic analysis tasks, there is no existing baseline for comparison. Therefore, we take the method trained with MSE in [9] as our baseline and compare our method with the baseline and traditional codec BPG. First, we compare the rate-accuracy performance of the proposed method with existing methods for various semantic analysis tasks. Second, we compare the visual quality performance of images reconstructed by the proposed method with other compression methods. Third, we analyze the linear relationships of accuracy and different metrics.

Four ablation analyses are conducted in this section. First, we explore the impact of different metrics on analysis accuracy and perceptual quality. Second, we study the influence of different weights on importance-weighted pixel distance. Third, the ratio of deep feature distance to importance-weighted pixel distance is studied. Finally, we study the influence of the input importance map on bit allocation.

A. Rate-Accuracy Results

1) Detection Tasks: For each task, we provide the evaluation results of two analysis network structures in Fig. 6. The plots, from left to right, are the evaluation results of object detection, instance segmentation, and person keypoints detection tasks. We compare our method with the baseline and the BPG codec. We report the evaluation results at the bit rates from 0.05 bits per pixel (bpp) to 1.2 bpp. The proposed method always outperforms the baseline in all tasks, especially at low bit rates. We also provide the PSNR and MS-SSIM in Fig. 7. The BPG codec achieves higher PSNR while fails to achieve high detection accuracy. The reason for it is that the BPG codec is optimized for signal fidelity and semantic fidelity is not considered during its optimization. The method in [9] outperforms our method in both PSNR and MS-SSIM, but it fails to achieve higher recognition accuracy. The results demonstrate that higher PSNR or MS-SSIM does not mean higher semantic quality. The proposed distortion metric is designed directly towards semantic fidelity and thus leads to high analysis accuracy. Compared with the baseline, we can achieve 20.79% $\sim$ 32.10% bit rate reduction under the same accuracy, as shown in Table I.

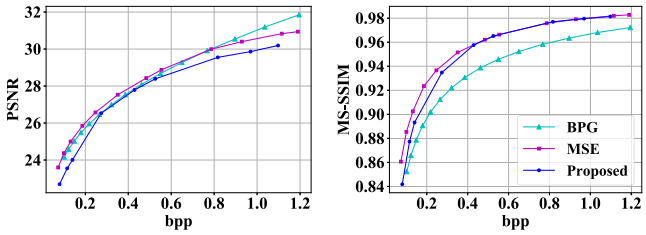


Fig. 7. Comparison of BPG and two learned image compression models in terms of rate and PSNR/MS-SSIM.

2) Image Captioning Task: Fig. 8 presents the performance of the proposed algorithm on the image captioning task. We also compare our method with the baseline and the BPG codec. Because a higher bit rate can not improve the accuracy, we only report the evaluation results at the bit rates from 0.05 bpp to 0.6 bpp. Our method also outperforms the baseline and BPG codec at all bit rates in the image captioning task. Compared with the

TABLE I
COMPARISON BETWEEN THE TWO LEARNED IMAGE COMPRESSION MODELS TRAINED WITH MSE AND THE PROPOSED METRIC (BITRATE IN BPP AND ACCURACY IN MAP)

	Faster_Res50_C4		Faster_Res50_FPN		Mask_Res50_C4		Mask_Res50_FPN		Keypoints_Res50_FPN		Keypoints_Res101_FPN	
Bitrate	MSE	Proposed	MSE	Proposed	MSE	Proposed	MSE	Proposed	MSE	Proposed	MSE	Proposed
0.93	32.04	32.38	34.36	34.58	29.25	29.57	31.95	32.20	59.67	59.93	62.08	62.33
0.48	28.77	30.49	30.86	32.60	26.29	27.80	28.63	30.31	55.08	57.39	57.95	59.69
0.25	23.09	26.65	25.07	28.79	21.21	24.55	23.16	26.63	47.32	51.31	49.82	53.63
0.07	11.61	9.35	12.23	9.03	10.46	8.47	10.60	8.04	24.68	16.75	26.72	17.66
BD-rate	**−29.99%**		**−30.14%**		**−32.10%**		**−30.23%**		**−22.21%**		**−20.79%**	

TABLE II
COMPARISON BETWEEN THE TWO LEARNED MODELS TRAINED WITH MSE AND THE PROPOSED METRIC (BITRATE IN BPP)

	BLEU_1		MENTOR		ROUGE_L		CIDEr	
Bitrate	MSE	Proposed	MSE	Proposed	MSE	Proposed	MSE	Proposed
0.55	0.687	0.691	0.228	0.23	0.505	0.507	0.824	0.83
0.25	0.676	0.688	0.222	0.228	0.497	0.505	0.784	0.825
0.13	0.656	0.679	0.213	0.224	0.484	0.499	0.723	0.794
0.03	0.556	0.605	0.161	0.186	0.407	0.445	0.41	0.56
BD-rate	**−42.69%**		**−41.58%**		**−40.69%**		**−41.92%**	

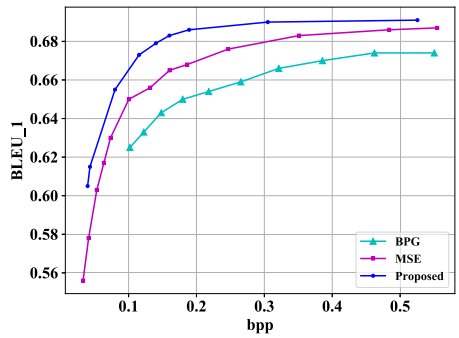


Fig. 8. Comparison of BPG and two learned image compression models on the image captioning task.

baseline, we can achieve more than 40% bit rate reduction under the same accuracy, as shown in Table II. Note that the bit rate reduction of image captioning task is more than that of detection tasks, because the image captioning task has less requirement on pixel fidelity than detection tasks, and then more image details can be ignored for image captioning task.

B. Visual Quality Evaluation

The visualization examples of our method, compared with other methods, are shown in Fig. 9. Compared with the baseline, the proposed method can preserve more texture details in the reconstructed images. For example, the stripes on the zebras in the last row of Fig. 9 are clearer than the stripes in the third row. It shows that deep feature distance is beneficial to the restoration of texture details. Compared with the BPG codec, there is no blocking artifacts in the images reconstructed by the proposed method. This is because the end-to-end method does not divide the image into blocks. Although grids appear in the images reconstructed at low bit rates, this phenomenon gradually disappears as the bit rate increases. And the grids do not affect the ability of the proposed method to preserve more details.

We also evaluate the reconstructed images using the perceptual metric LPIPS [59]. The evaluation result is shown in Fig. 10. We can see from Fig. 10 that the proposed metric performs much better than the baseline and BPG codec. It can be concluded that the proposed method achieves better perceptual quality while maintaining the analysis accuracy for various semantic analysis tasks. Our method achieves better perceptual quality due to two reasons. First, deep feature distance can help preserve texture details. Second, importance-weighted pixel distance can improve the reconstruction quality of semantically important regions through bit rate allocation, and the semantically important regions are usually also the regions that human vision system pays more attention to. We need to emphasize that our distortion measure was not designed for visual quality, so the results appear interesting. We will come back to this issue in the following subsection.

C. Correlation Between Accuracy and Metrics

We further explore the linear correlation between accuracy and different metrics. Taking detection tasks as an example, we explore the linear correlations between accuracy and MSE and the proposed semantic metric.

Specifically, we divide the models into two groups according to the ratio of deep feature distance to importance-weighted pixel distance λ_{fp}. We train 10 b rate points for each group and calculate the average metric values and accuracy. Then a linear function is used to fit the metric values and accuracy and coefficients of determination can be calculated based on the derived linear fitting. The metric value for MSE is the mean of MSE of 5000 images and the metric value for the proposed metric value is computed as the mean of $L_p + \lambda_{fp} \times L_f$.

Fig. 9. Visualization of original and reconstructed images. The images, from top to bottom, are original images, compressed images by BPG, compressed images by deep models trained with different distortion metrics: MSE, MS-SSIM, L_f, L_p, L_f, and MSE, and the proposed metric. The bit rates, from left to right, are 0.06bpp, 0.2bpp, 0.4bpp, and 0.74bpp, respectively.

TABLE III
RESULTS OF COEFFICIENT OF DETERMINATION (R^2) BETWEEN METRIC AND ACCURACY

λ_{tp}	Metric	Faster_R50_C4	Faster_R50_FPN	Mask_R50_C4	Mask_R50_FPN	Keypoints_R50_FPN	Keypoints_R101_FPN
10	MSE	0.9056	0.9002	0.9059	0.8929	0.9198	0.9231
	Proposed	0.9412	0.9364	0.9413	0.9306	0.9500	0.9518
100	MSE	0.9323	0.9270	0.9321	0.9217	0.9568	0.9596
	Proposed	0.9932	0.9916	0.9935	0.9937	0.9861	0.9825

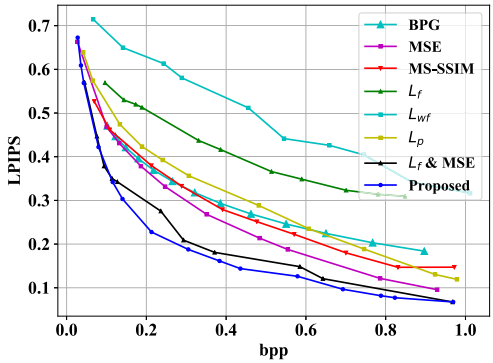


Fig. 10. Comparison of BPG and learned image compression models in terms of rate and LPIPS. Lower LPIPS is believed to correspond to better perceptual quality.

The coefficients of determination are shown in Table III. It can be seen that the linear correlation between the proposed metric and accuracy is better than that between MSE and accuracy. It partially explains why the proposed metric can achieve better accuracy on semantic analysis tasks.

D. Comparison of Different Metrics

We conduct comparison experiments on various distortion metrics, including MSE, MS-SSIM, deep feature distance L_f, importance-weighted deep feature distance L_{wf}, importance-weighted pixel distance L_p, deep feature distance and MSE $L_f\&MSE$, and the proposed metric $L_f\&L_p$. We denote the corresponding models as MSE model, MS-SSIM model, L_f model, L_{wf} model, L_p model, $L_f\&MSE$ model, and $L_f\&L_p$ model (proposed). The weighted deep feature distance is defined as:

$$L_{wf}(I,\bar{I}) = \text{M-MSE}(P(G(I)), P(G(\bar{I}))) = \frac{1}{K}\sum_k \left(\frac{1}{C_k M_k N_k} \sum_{c,m,n} ((S_{cmn} - S'_{cmn}) \times P_{cmn})^2 \right) \tag{6}$$

where $P_{c,m,n}$ is the weight at the position (m,n) on the c-th channel of the transformed features. $P_{c,m,n}$ is derived from the proposed importance map $P_{m,n}$ in (3), by resizing $P_{m,n}$ to the resolution of the transformed features and duplicating c channels.

Fig. 11(a) presents the evaluation results of object detection. The results of the other tasks have similar trends and have been provided in the supplementary material. The MS-SSIM model achieves better accuracy than L_f model, L_{wf} model, and L_p model while fails to outperform other models. L_f model performs poorly in recognition accuracy as it only focuses on deep feature distance. We find that the MSE continues to increase during the training phase of L_f model. In addition, there are regular grids and obvious color shift in the images reconstructed by this model, which is shown in the fifth row in Fig. 9. The grids change the distribution of the original image, so the analysis accuracy drops a lot. L_{wf} model outperforms L_f model. L_p model performs much better than L_f model and L_{wf} model. In order to prove the necessity of deep feature distance, we conduct a comparative experiment on $L_f\&L_p$. The higher recognition accuracy of $L_f\&L_p$ model demonstrates that the deep feature distance is helpful for semantic analysis. Besides, we also conduct experiments on $L_f\&MSE$. Compared with the $L_f\&L_p$ model, we can conclude that the importance-weighted pixel distance is better than MSE. The importance-weighted pixel distance makes the model focus on the semantically important regions and pays less attention to semantically unimportant regions. In this way, more bits are allocated to semantically important regions and the quality of this region is higher.

The visualization examples of all metrics are shown in Fig. 9. The proposed metric can preserve more texture details than L_p metric and better suppress the grids than L_f and $L_f\&MSE$ metrics. The LPIPS scores of all metrics are shown in Fig. 10. We can see from Fig. 10 that the L_f and L_p metrics can not improve the perceptual quality alone. This demonstrates that it is reasonable to combine L_f and L_p metrics. Besides, the proposed metric outperforms the $L_f\&MSE$ metric, which verifies the effectiveness of the importance-weighted pixel distance. Our method also outperforms the MS-SSIM model.

E. Ablation Study on Importance-Weighted Pixel Distance

We further explore the impact of importance weights of the importance-weighted pixel distance on analysis accuracy. Specifically, we fix the weight of importance area (foreground) to 1 and set the weights of background area (λ_{pb}) to 0, 0.1, 0.3, and 0.5, respectively.

Fig. 11(b) shows the evaluation results of the object detection task. On the whole, $\lambda_{pb} = 0.1$ outperforms $\lambda_{pb} = 0$ and $\lambda_{pb} = 0.5$ while $\lambda_{pb} = 0.3$ achieves the best accuracy. The λ_{pb} controls how much the encoder pays attention to the background region. When the encoder completely ignores the background ($\lambda_{pb} = 0$), the reconstructed image loses too much background information, resulting in a decline in the recognition accuracy. When the λ_{pb} is slightly increased, for example, $\lambda_{pb} = 0.1$, the recognition rate can be greatly improved. If the background λ_{pb} is too large ($\lambda_{pb} = 0.5$), the encoder pays too much attention

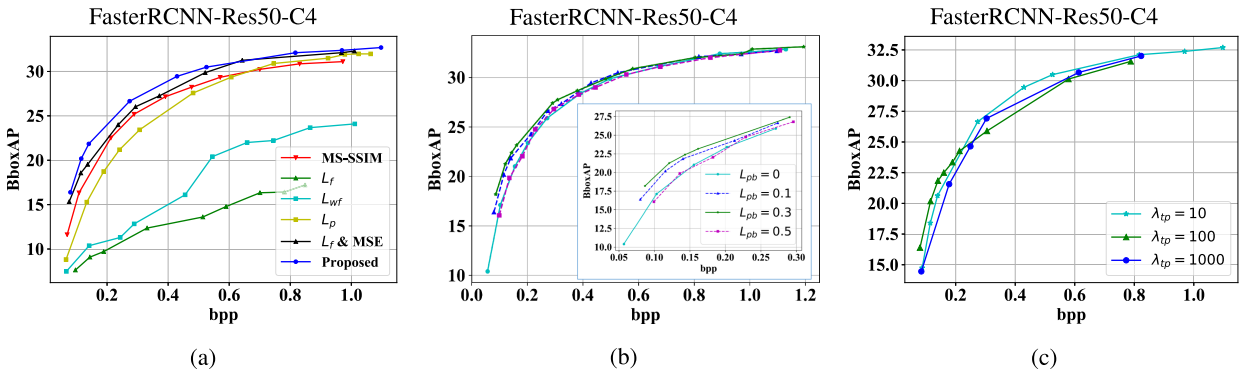


Fig. 11. Results of the ablation studies on the object detection task. (a) Ablation study on different distortion metrics. (b) Ablation study on the weight in the importance-weighted pixel distance. (c) Ablation study on the ratio of the deep feature distance to the importance-weighted pixel distance. More ablation study results are provided in the supplementary material.

to the background area, which leads to performance degradation. Reasonable control of the λ_{pb} is critical to the accuracy of semantic analysis.

On the other hand, we find that the difference between each weight at the low bit rate is more obvious. As the bit rate increases, the difference between them gradually shrinks. This is because the encoder cannot allocate enough bits to the background at low bit rates. At high bit rates, the foreground has been well reconstructed, and more bits can be allocated to the background, thereby eliminating the difference between weights. This also shows that the proposed encoder has the ability to adaptively allocate bit rates.

F. Ablation Study on λ_{fp}

We conduct another ablation study on the ratios of deep feature distance to importance-weighted pixel distance. We set the ratios λ_{fp} to 10, 100 and 1000, respectively.

Fig. 11(c) shows the results of the object detection task. The $\lambda_{fp} = 100$ model performs the best when the bit rate is less than 0.2 bpp and after that, the $\lambda_{fp} = 10$ model achieves the best accuracy. The weight of deep feature distance controls the encoder's attention to global semantics (or in other words, foreground and background). The $\lambda_{fp} = 1000$ model focuses too much on the background, which makes it fail to surpass the other two models. As the bit rate increase, the $\lambda_{fp} = 10$ model surpasses the $\lambda_{fp} = 100$ model.

G. Ablation Study on Network Input

We conduct two ablation studies about network inputs. In the first one, we verify the bit allocation capability of the importance information (importance map and importance image). In the second one, the contributions of importance map and importance image are studied.

1) Bit Allocation: Since the importance information is applied in pixel domain, we train two models with the same importance-weighted pixel distance. The input of the first model, denoted as $Model_{Img}$, is only the original image. The input of the second model, denoted $Model_{Concat}$, are the original image, segmentation ground-truth mask, and the product of them. We present the reconstructed images of these two models in

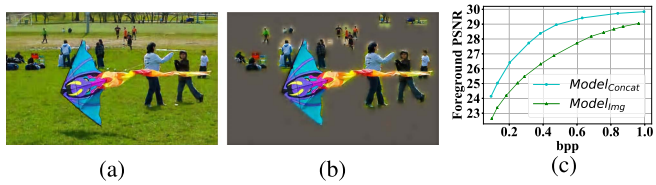


Fig. 12. Results of ablation study on bit allocation. $Model_{Img}$ inputs the original image only. $Model_{Concat}$ inputs the concatenation of the original image, the importance map, and the importance-weighted image. Both models are trained with L_{pl}. (a) A reconstructed image of $Model_{Img}$. (b) A reconstructed image of $Model_{Concat}$. (c) Comparison of the two models in terms of bitrate and foreground PSNR.

TABLE IV
DIFFERENT CONFIGURATIONS OF INPUTTING IMPORTANCE INFORMATION

Configuration	Proposed_M	Proposed_I	Proposed	Proposed_L
Original image	√	√	√	√
Importance map	√		√	
Importance image		√	√	
Label map				√
Label image				√

Fig. 12. $Model_{Img}$ reconstructs the entire image. Although the details of the background are better preserved, it is difficult for $Model_{Img}$ to allocate most bits to the foreground objects. In contrast, $Model_{Concat}$ almost completely ignores the background and only reconstructs the foreground objects, which further improves the quality of the foreground objects. We further compute the PSNR of the foreground objects for $Model_{Img}$ and $Model_{Concat}$, and compare them in Fig. 12. We can see that the foreground PSNR of $Model_{Concat}$ is much better than that of $Model_{Img}$.

2) Importance Information: The importance information is contained in both the importance map and the importance image, so we conduct an ablation study on the contributions of importance map and importance image. Specifically, we consider four kinds of inputs that are summarized in Table IV. The label map in Table IV denotes the ground-truth foreground mask provided in COCO. The label image is obtained in the same way as the importance image but the importance map is replaced by the label map. Fig. 13 shows the ablation study results for the object

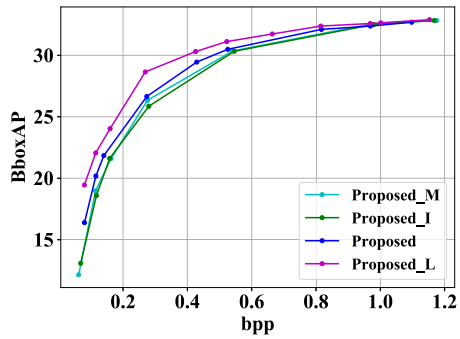


Fig. 13. Results of the ablation study regarding importance information for the object detection task.

detection task. The proposed_M model and proposed_I model achieve similar results. The proposed model outperforms both the proposed_M model and the proposed_I model. We also provide the result of the proposed_L model, which outperforms the proposed model significantly because it uses the ground-truth foreground mask.

VI. Conclusion

In this paper, we propose an image compression method for intelligent applications. Our key idea is to find out an appropriate, task-agnostic metric and use it as the target of a learned image compression method. First, inspired by the feature loss, we propose to transform images into a latent space by VGG-16 network, where the latent space are better aligned to semantics, and then calculate the Euclidean distance in the latent space as the deep feature distance. Second, inspired by the saliency mechanism, we propose to calculate the weighted distortion in the pixel domain as the importance-weighted pixel distance, where the weight is generated according to semantic importance of pixels. We then combine the deep feature distance and importance-weighted pixel distance into a single metric, and use the metric together with coding rate to optimize an end-to-end image compression network. We conduct experiments on a variety of semantic analysis tasks. Our experimental results show that the proposed metric leads to 20.79% $\sim$ 42.69% bit-rate reduction under the same analysis accuracy levels, compared to the same image compression network optimized for signal fidelity. In addition, our metric leads to reconstructed images with better visual quality.

In the future, we plan to extend this work in several directions. First, we will design a more accurate and light-weight importance map generation method to improve the analysis accuracy. Second, we want to explore other transforms, such as ResNet50 in replacement of VGG16. Third, we consider extending the proposed image compression method to video compression.

References

[1] W. B. Pennebaker and J. L. Mitchell, *JPEG Still Image Data Compression Standard*. Norwell, MA, USA: Kluwer Academic, 1992.

[2] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The JPEG 2000 still image compression standard," *IEEE Signal Process. Mag.*, vol. 18, no. 5, pp. 36–58, Sep. 2001.

[3] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[4] J. Löhdefink *et al.*, "GAN-vs. JPEG2000 image compression for distributed automotive perception: Higher peak SNR does not mean better semantic segmentation," 2019, *arXiv:1902.04311*.

[5] J. Lee *et al.*, "A training method for image compression networks to improve perceptual quality of reconstructions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 144–145.

[6] F. Mentzer, G. Toderici, M. Tschannen, and E. Agustsson, "High-fidelity generative image compression," 2020, *arXiv:2006.09965*.

[7] S. Luo *et al.*, "DeepSIC: Deep semantic image compression," in *Proc. Int. Conf. Neural Inf. Process.*, Cham, Switzerland: Springer, 2018, pp. 96–106.

[8] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[9] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," 2016, *arXiv:1611.01704*.

[10] J. Shi and Z. Chen, "Reinforced Bit allocation under task-driven semantic distortion metrics," in *Proc. Int. Symp. Circuits Syst.*, 2020, pp. 1–5.

[11] S. Suzuki, M. Takagi, K. Hayase, T. Onishi, and A. Shimizu, "Image pre-transformation for recognition-aware image compression," in *Proc. Int. Conf. Image Process.*, 2019, pp. 2686–2690.

[12] Z. Liu *et al.*, "DeepN-JPEG: A deep neural network favorable JPEG-based image compression framework," in *Proc. 55th Annu. Des. Automat. Conf.*, 2018, pp. 1–6.

[13] J. Choi and B. Han, "Task-aware quantization network for JPEG image compression," in *Proc. IEEE Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2020, pp. 309–324.

[14] L. D. Chamain, S.-C. S. Cheung, and Z. Ding, "Quannet: joint image compression and classification over channels with limited bandwidth," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2019, pp. 338–343.

[15] Z. Li, C. De Sa, and A. Sampson, "Optimizing JPEG quantization for classification networks," 2020, *arXiv:2003.02874*.

[16] L.-Y. Duan, X. Liu, J. Chen, T. Huang, and W. Gao, "Optimizing JPEG quantization table for low bit rate mobile visual search," in *Proc. IEEE Vis. Commun. Image Process.*, 2012, pp. 1–6.

[17] X. Luo, H. Talebi, F. Yang, M. Elad, and P. Milanfar, "The rate-distortion-accuracy tradeoff: JPEG case study," 2020, *arXiv:2008.00605*.

[18] R. R. Selvaraju *et al.*, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 618–626.

[19] N. Patwa *et al.*, "Semantic-preserving image compression," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 1281–1285.

[20] L. D. Chamain, F. Racapé, J. Bégaint, A. Pushparaja, and S. Feltman, "End-to-end optimized image compression for machines, a study," 2020, *arXiv:2011.06409*.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

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

[23] M. Akbari, J. Liang, and J. Han, "DSSLIC: Deep semantic segmentation-based layered image compression," in *Proc. Int. Conf. Acoust., Speech Signal Process.*, 2019, pp. 2042–2046.

[24] T. Man Hoang, J. Zhou, and Y. Fan, "Image compression with encoder-decoder matched semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 160–161.

[25] S. Duan, H. Chen, and J. Gu, "JPAD-SE: High-level semantics for joint perception-accuracy-distortion enhancement in image compression," 2020, *arXiv:2005.12810*.

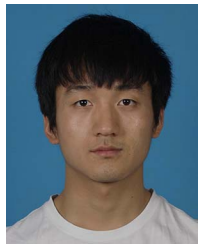
[26] Z. Chen and T. He, "Learning based facial image compression with semantic fidelity metric," *Neurocomputing*, vol. 338, pp. 16–25, 2019.

[27] D. Liu, D. Wang, and H. Li, "Recognizable or not: Towards image semantic quality assessment for compression," *Sens. Imag.*, vol. 18, no. 1, pp. 1–20, 2017.

[28] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 1, pp. 539–546.

[29] H. Choi and I. V. Bajić, "Deep feature compression for collaborative object detection," in *Proc. IEEE Int. Conf. Image Process.*, 2018, pp. 3743–3747.

[30] A. E. Eshratifar, A. Esmaili, and M. Pedram, "BottleNet: A. deep learning architecture for intelligent mobile cloud computing services," in *Proc. IEEE/ACM Int. Symp. Low Power Electron. Des*., 2019, pp. 1–6.
[31] J. Shao and J. Zhang, "BottleNet: An end-to-end approach for feature compression in device-edge co-inference systems," in *Proc. IEEE Int. Conf. Commun. Workshops*, 2020, pp. 1–6.
[32] S. Suzuki, M. Takagi, S. Takeda, R. Tanida, and H. Kimata, "Deep feature compression with spatio-temporal arranging for collaborative intelligence," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 3099–3103.
[33] Z. Chen *et al.*, "Toward intelligent sensing: Intermediate deep feature compression," *IEEE Trans. Image Process.*, vol. 29, pp. 2230–2243, 2019.
[34] Z. Chen *et al.*, "Lossy intermediate deep learning feature compression and evaluation," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 2414–2422.
[35] S. Singh *et al.*, "End-to-end learning of compressible features," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 3349–3353.
[36] L. Duan, J. Liu, W. Yang, T. Huang, and W. Gao, "Video coding for machines: A paradigm of collaborative compression and intelligent analytics," *IEEE Trans. Image Process.*, vol. 29, pp. 8680–8695, 2020.
[37] G. Toderici *et al.*, "Variable rate image compression with recurrent neural networks," 2015, *arXiv:1511.06085*,
[38] G. Toderici *et al.*, "Full resolution image compression with recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5306–5314.
[39] N. Johnston *et al.*, "Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4385–4393.
[40] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," 2018, *arXiv:1802.01436*.
[41] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 10 771–10780.
[42] J. Lee, S. Cho, and S.-K. Beack, "Context-adaptive entropy model for end-to-end optimized image compression," 2018, *arXiv:1809.10452*.
[43] M. Li, K. Zhang, W. Zuo, R. Timofte, and D. Zhang, "Learning context-based non-local entropy modeling for image compression," 2020, *arXiv:2005.04661*.
[44] J. Lee, S. Cho, and M. Kim, "An end-to-end joint learning scheme of image compression and quality enhancement with improved entropy minimization," 2019, *arXiv:1912.12817*.
[45] H. Ma, D. Liu, N. Yan, H. Li, and F. Wu, "End-to-end optimized versatile image compression with wavelet-like transform," *IEEE Trans. Pattern Anal. Mach. Intell.*, doi: 10.1109/TPAMI.2020.3026003, 2020.
[46] X. Yuan and R. Haimi-Cohen, "Image compression based on compressive sensing: End-to-end comparison with JPEG," *IEEE Trans. Multimedia*, vol. 22, no. 11, pp. 2889–2904, Nov. 2020.
[47] M. Li, W. Zuo, S. Gu, D. Zhao, and D. Zhang, "Learning convolutional networks for content-weighted image compression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3214–3223.
[48] J. Liu, G. Lu, Z. Hu, and D. Xu, "A unified end-to-end framework for efficient deep image compression," 2020, *arXiv:2002.03370*.
[49] Z. Zhong, H. Akutsu, and K. Aizawa, "Channel-level variable quantization network for deep image compression," 2020, *arXiv:2007.12619*.
[50] O. Rippel and L. Bourdev, "Real-time adaptive image compression," 2017, *arXiv:1705.05823*.
[51] L.-H. Chen, C. G. Bampis, Z. Li, A. Norkin, and A. C. Bovik, "Perceptually optimizing deep image compression," 2020, *arXiv:2007.02711*
[52] Y. Hu, S. Yang, W. Yang, L.-Y. Duan, and J. Liu, "Towards coding for human and machine vision: A scalable image coding approach," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2020, pp. 1–6.
[53] Z. Guo, Z. Zhang, and Z. Chen, "Deep scalable image compression via hierarchical feature decorrelation," in *Proc. IEEE Picture Coding Symp*., 2019, pp. 1–5.
[54] M. Akbari, J. Liang, J. Han, and C. Tu, "Learned variable-rate image compression with residual divisive normalization," in *Proc. IEEE Int. Conf. Multimedia Expo*., 2020, pp. 1–6.
[55] Y. Choi, M. El-Khamy, and J. Lee, "Variable rate deep image compression with a conditional autoencoder," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 3146–3154.
[56] F. Yang *et al.*, "Variable rate deep image compression with modulated autoencoder," *IEEE Signal Process. Lett.*, vol. 27, pp. 331–335, 2020.
[57] T. Guo *et al.*, "Variable rate image compression with content adaptive optimization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 122–123.
[58] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Hoboken, NJ, USA: Wiley, 2006.
[59] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.
[60] X. Wang *et al.*, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. IEEE Eur. Conf. Comput. Vis. Workshops*, 2018, pp. 63–79.
[61] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. IEEE Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2016, pp. 694–711.
[62] A. R. Zamir *et al.*, "Taskonomy: Disentangling task transfer learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3712–3722.
[63] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. IEEE Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2014, pp. 740–755.
[64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[65] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.
[66] K. Xu *et al.*, "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.
[67] J. Lei *et al.*, "A universal framework for salient object detection," *IEEE Trans. Multimedia*, vol. 18, no. 9, pp. 1783–1795, 2016.
[68] L. Li, S. Tang, Y. Zhang, L. Deng, and Q. Tian, "GLA: Global-local attention for image description," *IEEE Trans. Multimedia*, vol. 20, no. 3, pp. 726–737, Mar. 2017.



Changsheng Gao received the B.S. degree in electrical information engineering from Anhui University, Hefei, China, in 2017. He is currently working toward the Ph.D. degree with the School of Big Data, University of Science and Technology of China, Hefei, China. His research interests include image/video coding, signal processing, and machine learning.

Dong Liu (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electrical engineering from the University of Science and Technology of China (USTC), Hefei, China, in 2004 and 2009, respectively. He was a Member of Research Staff with Nokia Research Center, Beijing, China, from 2009 to 2012. He joined USTC in 2012 and became a Professor in 2020. His research interests include image and video processing, coding, analysis, and data mining. He has authored or coauthored more than 100 papers in international journals and conferences. He has 20 granted patents. He has several technical proposals adopted by international and domestic standardization groups. He was the recipient of the 2009 IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY Best Paper Award and the VCIP 2016 Best 10% Paper Award. He and his students were winners of several technical challenges held in ICCV 2019, ACM MM 2019, ACM MM 2018, ECCV 2018, CVPR 2018, and ICME 2016. He is a Senior Member of CCF and CSIG, an elected Member of MSA-TC of IEEE CAS Society. He is or was the Chair of IEEE Future Video Coding Study Group, a Publicity Co-Chair for ICME 2021, and a Registration Co-Chair for ICME 2019.

Li Li (Member, IEEE) received the B.S. and Ph.D. degrees in electronic engineering from University of Science and Technology of China (USTC), Hefei, Anhui, China, in 2011 and 2016, respectively. He is a Research Fellow with the Department of Electronic Engineering and Information Science, USTC. He was a Visiting Assistant Professor with the University of Missouri-Kansas City, Kansas City, MO, USA, from 2016 to 2020. His research interests include image/video coding and processing. He was the recipient of the best 10% paper awards at the 2016 IEEE VISUAL COMMUNICATIONS AND IMAGE PROCESSING (VCIP) and the 2019 IEEE INTERNATIONAL CONFERENCE ON IMAGE PROCESSING (ICIP).

Feng Wu (Fellow, IEEE) received the B.S. degree in electrical engineering from Xidian University, Xi'an, China, in 1992, and received the M.S. and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1996 and 1999, respectively. He is a Professor and the Assistant to the President with the University of Science and Technology of China, Hefei, China. Previously, he was a Principle Researcher and Research Manager with Microsoft Research Asia, Beijing, China.

His research interests include various aspects of video technology and artificial intelligence. He has authored or coauthored two books, more than 100 journal papers (including several dozens of IEEE TRANSACTIONS papers), and top-conference papers on MOBICOM, SIGIR, CVPR, and ACM MM. He has more than 150 granted patents. His 15 techniques have been adopted into international video coding standards. He is or was the Editor-in-Chief of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEM FOR VIDEO TECHNOLOGY and as an Associate Editor for IEEE TRANSACTIONS ON IMAGE PROCESSING and IEEE TRANSACTIONS ON MULTIMEDIA. He also is the General Chair in ICME 2019, TPC Chair in MMSP 2011, VCIP 2010, and PCM 2009. He is the Chair of IEEE Data Compression Standard Committee. He was the recipient of the IEEE CAS Mac Van Valkerburg Award in 2021. He was the recipient of the best paper awards in IEEE TCSVT 2009, VCIP 2016, PCM 2008, and VCIP 2007, and Best Associate Editor Award of IEEE TRANSACTIONS ON IMAGE PROCESSING in 2018.