

Cross Modal Compression: Towards Human-comprehensible Semantic Compression

Jiguo Li*

jiguo.li@vipl.ict.ac.cn

Institute of Computing Technology,
Chinese Academy of Sciences
Beijing, China

Chuanmin Jia[†]

cmjia@pku.edu.cn

Peking University
Beijing, China

Xinfeng Zhang

xfzheng@ucas.ac.cn

University of Chinese Academy of
Sciences
Beijing, China

Siwei Ma[‡]

swma@pku.edu.cn

Peking University
Beijing, China

Wen Gao

wgao@pku.edu.cn

Peking University
Beijing, China

ABSTRACT

Traditional image/video compression aims to reduce the transmission/storage cost with signal fidelity as high as possible. However, with the increasing demand for machine analysis and semantic monitoring in recent years, semantic fidelity rather than signal fidelity is becoming another emerging concern in image/video compression. With the recent advances in cross modal translation and generation, in this paper, we propose the cross modal compression (CMC), a semantic compression framework for visual data, to transform the high redundant visual data (such as image, video, etc.) into a compact, human-comprehensible domain (such as text, sketch, semantic map, attributions, etc.), while preserving the semantic. Specifically, we first formulate the CMC problem as a rate-distortion optimization problem. Secondly, we investigate the relationship with the traditional image/video compression and the recent feature compression frameworks, showing the difference between our CMC and these prior frameworks. Then we propose a novel paradigm for CMC to demonstrate its effectiveness. The qualitative and quantitative results show that our proposed CMC can achieve encouraging reconstructed results with an ultrahigh compression ratio, showing better compression performance than the widely used JPEG baseline.

CCS CONCEPTS

- **Computing methodologies** → *Supervised learning by regression*;
- **Mathematics of computing** → *Coding theory*.

*Also with University of Chinese Academy of Sciences, Beijing 100049, China.

[†]Chuanmin Jia is the corresponding author.

[‡]Also with Information Technology R&D Innovation Center of Peking University, Shaoying 312000, China.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '21, October 20–24, 2021, Virtual Event, China

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8651-7/21/10...\$15.00

<https://doi.org/10.1145/3474085.3475558>

KEYWORDS

Image/Video Compression, Deep Neural Networks, Multimedia

ACM Reference Format:

Jiguo Li, Chuanmin Jia, Xinfeng Zhang, Siwei Ma, and Wen Gao. 2021. Cross Modal Compression: Towards Human-comprehensible Semantic Compression. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3474085.3475558>

1 INTRODUCTION

Data explosion makes more and more data be accessed by machines instead of humans, especially for image and video data. Therefore, the semantic fidelity [8], rather than signal fidelity, is becoming a more important metric in image/video compression. Because signal fidelity aims to the human visual system and is widely used in traditional compression. Besides, monitoring the semantic information, such as the identification, human traffic or car traffic, rather than the raw signal, is becoming the main concern of most applications, which is named *semantic monitoring* in this paper. However, the traditional block-based image/video compression frameworks [25, 34, 39] mainly optimize the signal fidelity under certain rate constrain, cannot meet the emerging demand of machine analysis and semantic monitoring. Recent feature compression frameworks [7, 29] encode the ultimate/intermediate features of deep neural networks into bitstream via quantization and entropy coding, to concentrate on the semantic fidelity for machine analysis. However, feature compression has three limits: (1) it is mostly task-specific so the feature is difficult for multi-task analysis; (2) it is not human-comprehensible, so further analysis is necessary for semantic monitoring; (3) the evidence is not enough for these features to reconstruct the data on the semantic level. Therefore in this paper, we propose the cross modal compression framework for human-comprehensible semantic compression to conquer these limits.

The traditional image/video compression framework is a block-based hybrid architecture, including the following submodules: block splitting, prediction, transformation, quantization, entropy encoding [25, 34, 39]. It simultaneously optimizes the pixel level fidelity with the metric of peak signal to noise ratio (PSNR), and the

transmission or storage cost (bitrate). The traditional compression framework assumes that we need to reconstruct the original signal from the compression bitstream every time it is accessed, no matter for what we access the data. However, in the machine analysis, such as the retrieval for large scale surveillance video, reconstruction may be unnecessary if we can analyze the data in the compression domain. But the compression domain in traditional frameworks is the bitstream, which cannot be analyzed easily.

Feature compression is proposed to compress the semantic features so that we can analyze the data in the compression domain, without the need to reconstruct the signals. The ultimate feature compression [29] compress the task-specific ultimate features into a bitstream, then these features are stored or transmitted for the future intelligent analysis. The raw data reconstruction is unnecessary because the task-specific semantic representation is accessible for the following analysis tasks. However, the features here are mostly task-specific, and a new feature is needed if a new task is added in our intelligent analysis, as illustrated in Table 1. To overcome this limit, the intermediate feature compression [7] is proposed to extract the intermediate features from the intermediate layers of the deep model, rather than the ultimate layer, making the features more common for multi-tasks analysis. However, the features here are not human-comprehensible, so further processing is needed for semantic monitoring. Besides, the evidence for reconstructing the raw data from these intermediate features in the semantic level is not enough.

Motivated by the new demands for machine analysis and semantic monitoring, we propose the cross modal compression (CMC) to compress the high redundant data (such as images, videos. etc.) into a compact, common, and human-comprehensible compression domain (such as text, sketch, semantic map, attributions. etc.). With this compression domain, CMC has these advantages: (1) we can compress the raw data with ultra-high compression ratio while preserving the semantic; (2) this common representation can be used for multiple machine analysis tasks; (3) our compact representation is human-comprehensible, so it can be used for semantic monitoring without further processing; (4) image/video reconstructions, especially image reconstruction from text/sketch/attributions, video reconstruction from semantic maps, have been well-studied, providing enough evidence to reconstruct raw data from the human-comprehensible compression domains.

In general, our contributions in this paper can be summarized as follows:

- (1) We propose a new framework, cross modal compression (CMC), for human-comprehensible semantic compression to meet the emerging demands and formulate the semantic compression as a rate-distortion optimization problem.
- (2) We propose a novel paradigm for cross modal compression by compressing the images into the text because the text representation is compact, common, and human-comprehensible. Recent works about text-to-image generation also provide evidence to reconstruct images from the text on the semantic level.
- (3) Qualitative and quantitative results demonstrate the effectiveness of our proposed CMC, showing better compression performance than the widely used JPEG baseline.

2 RELATED WORKS

2.1 Traditional Image/Video Compression

Traditional image/video compression technologies have been widely applied in our daily life for several decades, which has brought up a series of industry standards, such as JPEG [39], JPEG2000 [25], TPG [45] for image and AVC [43], HEVC [34], AVS2 [12] for video. JPEG and JPEG2000 are image compression standards for general purposes, including variable resolutions, various color spaces, and different transmission bandwidths, taking the efficiency of both the software and hardware into consideration. It consists of well-known technologies, including 8×8 discrete cosine transformation (DCT) [1], quantization, and Huffman coding. In addition to the image compression standards, several video compression standards have also been set, including the widely applied H.264 [43], AVS2 [12] and H.265 [34]. They are all block-based hybrid coding frameworks, in which quad-split, intra/inter prediction, DCT, quantization, entropy coding [17] are used to reduce the statistical redundancy, spatial/temporal redundancy, and perceptual redundancy. Recently, deep-based frameworks have also been proposed to compress the images via end-to-end optimization [5, 6, 24], and these methods have shown the potential that surpasses widely used state-of-the-art image codecs, such as JPEG, JPEG2000 and HEVC intra, although they have the difficulties to be standardized due to the non-uniqueness of the model's parameters and structures. Also, the submodules in video compression frameworks are investigated by embedding the DNNs in the them, including intra prediction [20], inter prediction [48], loop filter [18], quantization [2] and entropy encoding [28].

2.2 Feature Compression

To alleviate the network load and the computation in the back-end, Redondi *et al.* [29] proposed an analyze-then-compress paradigm instead of the traditional compress-then-analyze framework. But this work only investigated the hand-crafted feature, such as SIFT [23]. Choi *et al.* [9] examined to lossy compress the deep ultimate features for the object detection. However, the ultimate feature is usually task-specific, making it difficult to process increasing intelligent analysis tasks. To conquer this problem, Chen *et al.* [7] presented to transmit the intermediate features, which are derived from the intermediate activation layers of the DNNs, rather than the ultimate layer to enable a good balance among the transmission load, computing load, and the generalization ability for different intelligent analysis tasks. The features from the lower layers of DNNs are less abstract and less task-specific because a DNN can be viewed as a cascaded feature extractor [7]. Therefore the intermediate feature can be adopted by various tasks and is with better generalization ability than the ultimate feature. However, a deficiency of the intermediate features is that it cannot be understood by the human, so the raw data are needed when the semantic monitoring is necessary for the human.

2.3 Cross Modal Translation

Cross modal translation aims to convert the data from one modality to another with semantic consistency, which is an emerging topic in recent years, including image-to-text [16, 19, 31, 37, 40],

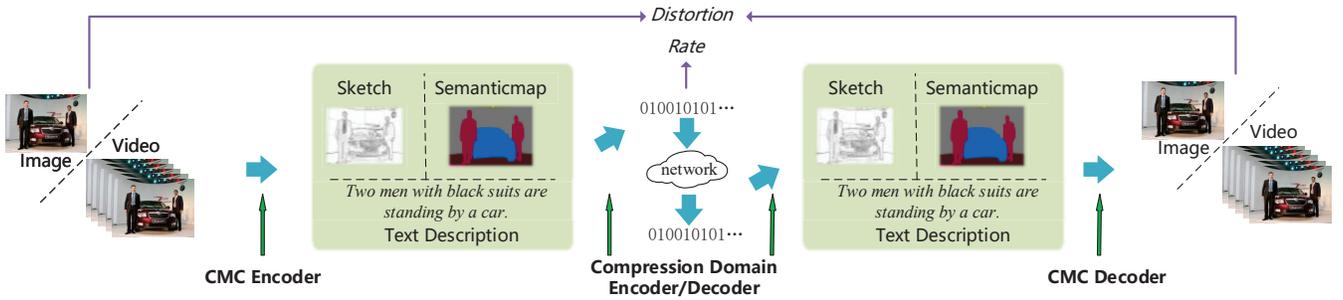


Figure 1: Illustration of our proposed Cross Modal Compression (CMC) framework. The compressed representation in the compression domain is a compact, common, and human-comprehensible feature (such as text, sketch, semantic map, attributions, etc.) which can be losslessly encoded into a bitstream. The whole framework consists of four parts: CMC encoder, CMC decoder, entropy encoder, and entropy decoder.

text-to-image [30, 44, 47], video-to-text [26, 41], text-to-video [27], sketch-to-image [22], image-to-sketch [42], etc. In this paper we mainly review the works on image-to-text and text-to-image translation. *Image-to-text* translation (I2T), also known as image caption, represents the images as a syntactically and semantically correct sentence, which is a compact, human-comprehensible form. Before DNNs were adopted in I2T, template-based methods [11] and retrieval-based methods [13] are two main approaches for I2T [16]. Once deep based methods were applied in I2T, it showed incomparable performance via an encoder-decoder framework [35, 37]. *Text-to-Image* translation (T2I), also known as image generation from the text, aims to synthesize fine-grained images from the text descriptions with semantic consistency. Reed *et al.* [30] demonstrated it feasible to synthesize images with semantic consistency from the text consistency via generative adversarial networks (GANs) [14], although the resolution of generated images is only 64×64 . Based on the advances in image generation [3, 32], the following T2I works [44, 46, 47] succeeded to synthesize images with higher resolution and more details based on a progressive framework.

3 CROSS MODEL COMPRESSION (CMC)

3.1 Problem Formulation

Data compression aims to reduce the transmission or storage cost with certain fidelity, which can be formulated as follows:

$$g = D + \lambda R, \quad (1)$$

where R is the bitrate, D denotes the distortion, which is evaluated in pixel level in traditional image/video compression. Cross modal compression (CMC) aims to compress the high redundant data into compact, common, human-comprehensible representation, which can be adopted by various machine analysis applications. In CMC, a compression domain \mathbb{Y} , where the representation is compact, common, and human-comprehensible, is defined. In this domain, the compressed representation can be losslessly encoded as a bitstream. As illustrated in Fig. 1, the framework consists of four submodules: CMC encoder, CMC decoder, entropy encoder, and entropy decoder. CMC encoder compresses the raw signal into a compact and human-comprehensible representation, which can be decoded by the CMC decoder to reconstruct the signal with semantic consistency. The bitrate is optimized by finding a compact compression domain,

while the distortion is optimized by preserving the semantic in CMC encoder and decoder.

3.2 Comparison with Related Frameworks

In this section, we compare our proposed CMC with the related frameworks and show the difference with these prior works, as shown in Table 1.

- (1) *Traditional Signal Compression.* Traditional codecs optimize the pixel level fidelity for human visual perception by minimizing the pixel level metrics. When more and more intelligent analysis applications come to our traditional signal compression systems, the data must be reconstructed so that we can process the images/videos (recognized, detected, or enhanced) to extract the semantic information. However, the explosive growth of images/videos has made that most of the visual data are not watched by the human but by the intelligent analysis applications. So pixel level optimization may waste the storage and bandwidth due to the information redundant for these intelligent applications.
- (2) *Ultimate Feature Compression.* Due to the inefficiency of traditional signal compression in some scenarios where only the analysis results are needed rather than the pixel level representation, ultimate feature compression [29] were proposed based on the tenet that most of the visual analysis tasks can be carried with the ultimate feature. In this *Analyse-then-Compress* paradigm, ultimate features are extracted on the front-end devices and then delivered to the back-end server to enable the intelligent analysis tasks. Although this paradigm can alleviate the load of the cloud center, this framework may suffer from an obstacle that the ultimate features are usually task-specific so it is difficult to adopt them in the tasks except for the specific one. Moreover, deploying various deep models in front-end devices will make the systems bloated.
- (3) *Intermediate Feature Compression.* The intermediate features, extracted from the intermediate layers of the DNNs, are less abstract and can be applied in various tasks. By conveying the intermediate features, the computational load on front-end devices and back-end cloud servers can be well balanced, indicating the flexibility to deploy this kind of framework.

Table 1: Comparison with related compression frameworks

| Methods | Compression Ratio | Multi-task Analysis | Human Comprehensible | Frondend Load | Backend Load | Data Reconstruction |
|----------------------------------|-------------------|---------------------|----------------------|---------------|--------------|---------------------|
| Traditional Compression* | Middle | ✓ | × | Middle | High | ✓ |
| Ultimate Feature Compression | High | × | × | High | Low | - |
| Intermediate Feature Compression | High | ✓ | × | Middle | Middle | - |
| Cross Modal Compression | High | ✓ | ✓ | Middle | Middle | ✓ |

* Such as JPEG [39], H.264 [43], HEVC [34], etc.

- The evidence for this task is not enough.

Besides, lossy compression for the intermediate features is also conducted to encode the feature into bitstream to be stored or transmitted. However, the intermediate features cannot be understood by humans so they cannot be used for human semantic monitoring directly, and further processing and analysis are necessary.

- (4) *Cross Modal Compression*. To accomplish a compact, common, and human-comprehensible representation for image/video data, we propose the cross modal compression (CMC), as illustrated in Fig. 1. The image/video $x, x \in \mathbb{X}$, is firstly transformed into $y, y \in \mathbb{Y}$. y is a compact and human-comprehensible representation, such as text, sketch, semantic map, or attributions, which can be adopted in semantic monitoring directly. Besides, y is a common feature and can be adopted for various intelligent analysis applications. When transmission, y can be compressed further into a bitstream by an entropy encoder. If the raw images/videos are needed, we can also reconstruct the raw data from y with semantic consistency. Our proposed CMC is different from all the above previous frameworks, as illustrated in Table 1.

CMC is a novel framework for visual data compression, such as image and video, which is a human-comprehensible semantic compression framework. We can design the specific submodules for CMC encoder/decoder, entropy encoder/decoder, when the source data domain, compression domain are determined. In the following, we will introduce a paradigm to compress the images into the text domain, which is compact, common, and human-comprehensible.

4 IMAGE-TEXT-IMAGE: A PARADIGM OF CMC

In this section, we introduce a paradigm for our proposed CMC. With the advances of image-to-text translation, understanding an image and describing it with natural languages is possible. Meanwhile, prior works on text-to-image translation have shown enough evidence for reconstructing the image from the text on the semantic level. We define the compression domain \mathbb{Y} as the text domain in our CMC framework and propose the Image-Text-Image (ITI) compression framework for cross modal image compression. As illustrated in Fig. 1, there are mainly three submodules in the framework: (1) CMC encoder to compress the raw image into a text description; (2) Lossless compression in compression domain to encode the text description into a bitstream and decode the bitstream into language; (3) CMC decoder to reconstruct images from the text description.

4.1 CMC Encoder

CMC encoder aims to compress the data from the image/video domain into a compact domain, which is the text domain in this paradigm. With the advance of image caption in recent years, a CNN-RNN with attention model referring to [37] is used in our CMC encoder, as shown in Fig. 2. Given an input image $v_i, i = 1, 2, 3, \dots, N$, N is the number of the samples, a downsampled feature f_i can be obtained by a CNN based feature extractor:

$$f_i = \text{CNN}(v_i), \quad (2)$$

where the feature $f_i = \{f_{i,1}, f_{i,2}, f_{i,3}, \dots, f_{i,L}\}$ contains L vectors that correspond different positions on the image. Following the CNN feature extractor, an RNN with attention is leveraged to decode the predicted words step by step. As introduced in [4] and [37], at time step s_t , the attention mechanism can be formulated as follows (described as “soft attention” in [37]):

$$\mathbb{E}_{p(s_t|a)}[\hat{z}_t] = \sum_{j=0}^L \alpha_{t,j} a_j, \quad (3)$$

where \hat{z}_t is the context vector at time step s_t , dynamically representing the relevant part of the image, $a_j, j = 1, 2, 3, \dots, L$, denotes the annotation vector that corresponds to the extracted feature $f_{i,j}$ at different image locations. $\alpha_{t,j}$ is a weight for each annotation vector a_j , which can be computed by an attention model f_{attn} :

$$e_{t,j} = f_{attn}(a_j, h_{t-1}) \quad (4)$$

$$\alpha_{t,j} = \frac{\exp(e_{t,j})}{\sum_{k=1}^L \exp(e_{t,k})}, \quad (5)$$

where h_t is the hidden output of the RNN at time step t , $e_{t,j}$ is the attention output before the softmax layer.

At the first time step, the starting word “<start>” is feed into the RNN to compute the attention map, build the context and decode the following words. The RNN output one word at each step until the ending word “<end>” is outputted.

4.2 Lossless Compression in Compression Domain

We think that the compression domain is more semantically compact, compared with the source data domain. Conversion from the source data domain to the compression domain reduces the semantic redundancy. However, there is still statistical redundancy in the compression domain. According to Shannon’s information theory [33], the optimal code length for a symbol is $\log_2 P$, where

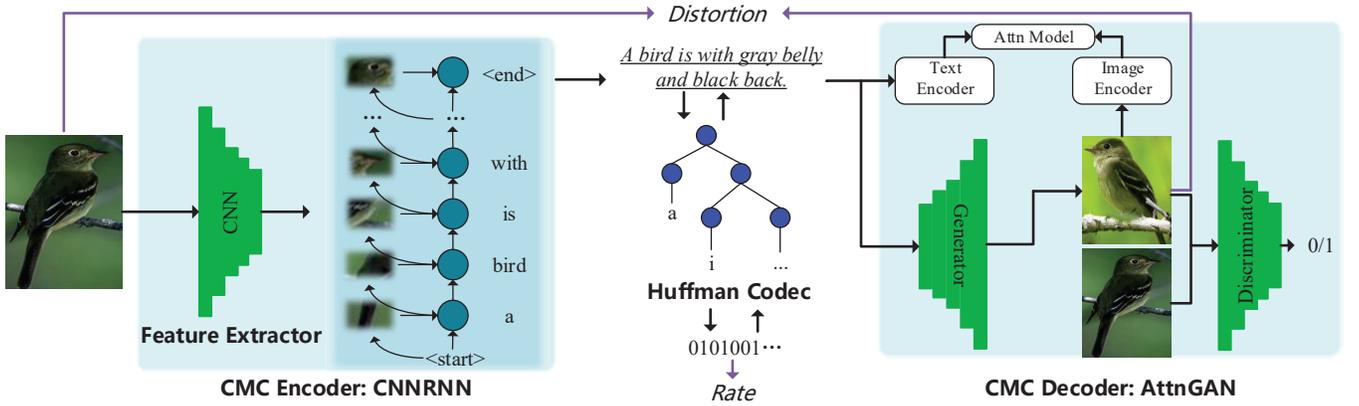


Figure 2: Illustration for a paradigm of CMC: Image-Text-Image (ITI)

P is the probability of the symbol. In our compression domain, the distribution of the text characters is a prior probability, which can be counted from the training set. In our paradigm, Huffman coding [17] can be used to reduce the statistical redundancy. The statistical probability of the symbols can be obtained from the training set under the assumption that the training set and the testing set have the same distribution. With the symbol probability, we can construct the Huffman tree and design the Huffman encoder based on the Huffman tree. The paired Huffman decoder uses the same Huffman tree with the encoder's. It is worth mentioning that Huffman coding is lossless, so we can reconstruct the text without any information loss.

4.3 CMC Decoder

CMC decoder aims to reconstruct the data from the compression domain. In our paradigm, we need to reconstruct the image from the text description with semantic consistency. With the recent advances in text-to-image generation [44, 46, 47], we use AttnGAN [44] in our CMC decoder to reconstruct images from the text due to its promising performance on text-to-image generation. AttnGAN integrates the attention mechanism into the generator by pretraining a text encoder and an image encoder to extract position-sensitive features. In the implementation, we use the pretrained Inception-v3 model [36], which is trained on Imagenet [10], as the image encoder and train the text encoder on our own dataset, following [44]. Given the text/image positive-sensitive feature e/v , the attention matching score can be calculated as follows:

$$c = \text{softmax}(\gamma_1 e^T v, \text{dim} = 0) v^T \quad (6)$$

$$R(c_j, e_j) = \cos(c_j, e_j), \quad (7)$$

where c_j is a region-context vector dynamically representing the image's subregion related to the j^{th} word, γ_1 is a factor to control the attention to its relevant sub-regions, $\cos(x, y) = x^T y / \|x\| \|y\|$ is the cosine similarity. The *attention-driven image-text matching score* between the image and the text description is defined as:

$$R(v, e) = \log \left(\sum_{j=1}^J \exp(\gamma_2 R(c_j, e_j)) \right)^{\frac{1}{2}}, \quad (8)$$

where J is the word number of the text description, γ_2 is a factor to control the importance of different word-to-subregion pairs.

To synthesize images with a resolution 256×256 , AttnGAN firstly generates images with a resolution 64×64 , then upsamples the generated images with ratio 2 and adds the details, until the image with resolution 256×256 is generated. It has been demonstrated that generating images with increasing resolutions is a promising method to synthesize high-resolution realistic images [46, 47]. More details about AttnGAN can be found in [44].

5 DATASETS AND METRICS

5.1 Datasets

We use MS COCO [21] and CUB-200-2011 [38] to evaluate our proposed model's effectiveness. MS COCO contains 82783/40504 images for training/testing set, and no less than 5 captions for each image. Images are resized with the resolution of 256×256 for MS COCO. CUB-200-2011 contains 8855/2933 images for training/testing set, and 10 captions for each image. Total 200 classes for CUB-200-2011 are split into 160/40 classes for training/testing. The images are cropped with the annotated bounding box and resized into 256×256 for the following experiments.

5.2 Metrics

We use four metrics to evaluate our proposed method on different levels, which are introduced as follows:

Peak Signal-to-Noise Ratio (PSNR) is defined as:

$$\text{PSNR} = 10 \log_{10} \left[\frac{(2^{\text{bits}} - 1)^2}{\text{MSE}} \right], \quad (9)$$

where bits denotes the bit number for a pixel, MSE denotes mean square error between the source data and the reconstructed data. PSNR is widely used in traditional image/video compression algorithm evaluation.

Inception Score (IS) [32] measures the naturalness and the diversity of the generated images, which is defined as:

$$\text{IS} = \exp(\mathbb{E}_{\mathbf{x}} \text{KL}(p(y|\mathbf{x})|p(y))), \quad (10)$$

where $p(y|\mathbf{x})$ denotes the conditional label distribution, $p(y)$ denotes the marginal distribution, \mathbf{x} denotes the perceptual features.



Figure 3: Qualitative results of our ITI framework on CUB-200-2011 (left) and MS COCO (right). For each sample, we show the raw image, the text representation, and the reconstructed image, subsequently. We also show the *bitrate* and the *compression ratio* under each text.

Fréchet Inception Distance (FID) [15] measure the distribution distance between the source data and the reconstructed data, which is formulated as:

$$FID = \|m_1 - m_2\|_2^2 + \text{Tr}(C_1 + C_2 - 2(C_1 C_2)^{\frac{1}{2}}), \quad (11)$$

where m/C is the mean/variance of the perceptual features on the testing set.

Instance Perceptual Distance (IPD) is used to measure the instance level perceptual distance, because both IS and FID are the set level metrics, neither can evaluate the instance level distortion. IPD is defined as:

$$IPD = \mathbb{E}_x \|x_{re} - x\|_2^2, \quad (12)$$

where x/x_{re} denotes the perceptual features for source/reconstructed data. For IS, FID, and IPD in our experiments, we use the pre-trained Inception-v3 model which pretrained on Imagenet [10] to

extract the perceptual features for MS COCO, and use the fine-tuned Inception-v3 model for CUB-200-2011, following [44]. In our experiments, we use the implementation of IS and FID in [44]¹. We use the same pretrained Inception-v3 model as FID's in the implementation of IPD.

6 EXPERIMENTAL RESULTS

To demonstrate the effectiveness of our proposed paradigm: image-text-image framework for image compression, we conduct experiments to show qualitative and quantitative results on both MS COCO and CUB-200-2011 datasets.

¹IS for MS COCO: <https://github.com/hanzhanggit/StackGAN-inception-model>, IS for CUB-200-2011: <https://github.com/openai/improved-gan>, FID for MS COCO and CUB-200-2011: <https://github.com/bioinf-jku/TTUR>

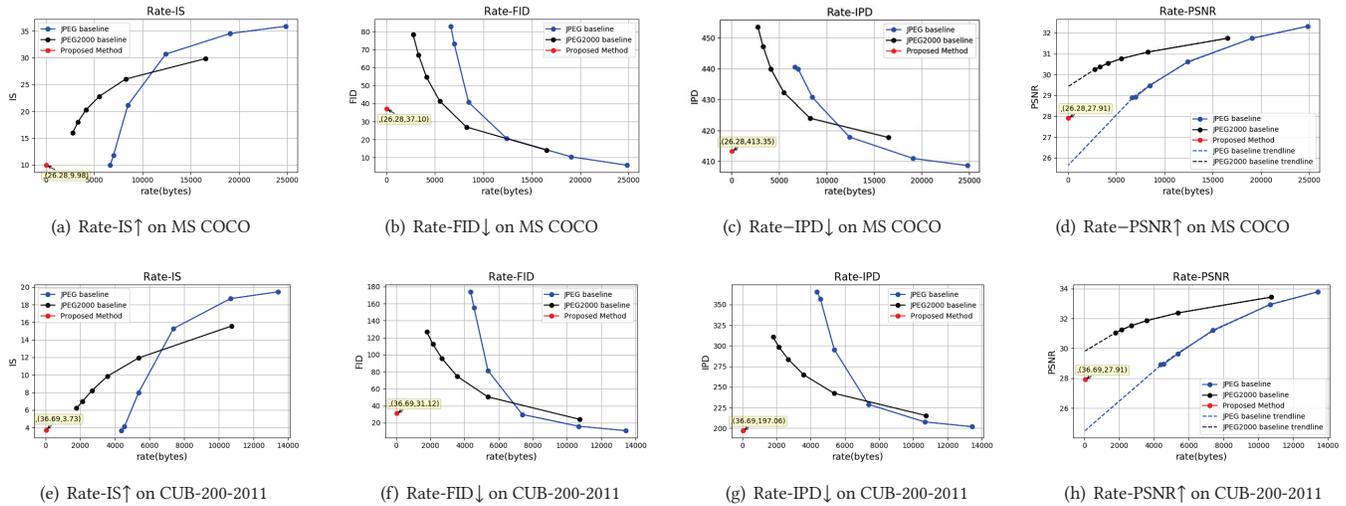


Figure 4: Quantitative results: comparison with the JPEG and JPEG2000 baselines (best view in color).

6.1 Qualitative Results

The qualitative results, as well as the compression ratio², on CUB-200-2011 and MS COCO are illustrated in Fig. 3. Some conclusions can be drawn from the results:

Our proposed ITI can reconstruct the images well on CUB-200-2011 dataset on instance level with ultrahigh compression ratio. As shown on the left part of Fig. 3, most of the reconstructed images are sharp and fine-grained, with a similar appearance as the raw images, although the background may be blurry in the reconstructed ones. Meanwhile, most of the text descriptions are semantically consistent with the raw images and the compression ratio is as high as 4000-7000 times.

Our proposed ITI can partially reconstruct the images on MS COCO dataset on instance level with ultrahigh compression ratio. On MS COCO dataset, our ITI can only partially reconstructed the images on the instance level, does not show as good performance as that on CUB-200-2011. This is because MS COCO is a more diverse dataset and CUB-200-2011 is a class-specific data for birds. However, as shown on the right of Fig. 3, our ITI can still partially reconstructed the scene and some key objects in the scene, although the objects may be incomplete and blurry.

Overall, the qualitative results show promising performance on class-specific datasets (such as CUB-200-2011) and encouraging potential on diverse datasets (such as MS COCO), demonstrating the effectiveness of our proposed ITI (a paradigm of CMC). It is worth mentioning that we can use a more powerful image generation model to improve the reconstruction performance, but the main aim in this work is to demonstrate the effectiveness of cross modal compression, so improving the submodules of CMC is beyond the scope of this work.

²The compression ratio is calculated under the assumption that the input images are resized to 256×256 with 3 channels.

6.2 Quantitative Results

To evaluate the compression performance of our proposed framework quantitatively, we compare our proposed ITI with the widely used JPEG [39] and JPEG2000 [25] standards. We compress all the images in the testing set with different quality factors, then plot the R-D curve, just like the R-D curve in the traditional image/video compression, as shown in Fig. 4. Among the four metrics, IS and FID are set level metrics, IPD is an instance level metric, PSNR is a pixel level metric. Some conclusions can be drawn from the quantitative comparison results:

ITI surpasses the JPEG baseline and is comparable with JPEG2000 baseline in set level and instance level. As illustrated in Fig. 4 (a), (e), our proposed ITI achieves a similar IS score with a rather lower bitrate when compared with JPEG baseline on both MS COCO and CUB-200-2011 datasets. Also, our proposed method is comparable with JPEG2000 if the trend of the curve is taken into consideration. From Fig. 4 (b) and (f), better FID scores are obtained by our proposed method than both JPEG and JPEG2000 baselines. IS and FID are set level metrics, so our ITI model shows better performance than the JPEG baseline and comparable performance with JPEG2000 when they are evaluated with set level metrics. IPD, as defined in Eq. 12, is an instance level metric, measuring the perceptual distance between two samples. As illustrated in Fig. 4 (c) and (g), our ITI model can even achieve better comparison results than both JPEG and JPEG2000 baselines. So our ITI proposed model (a paradigm of CMC) shows better performance than JPEG and comparable performance with JPEG2000 when evaluated with instance level perceptual metrics.

Our proposed ITI can achieve ultrahigh compression ratio with set/instance level reconstruction. As illustrated in Fig. 4, the bitrate is as low as 26.28/36.69 bytes on MS COCO/CUB-200-2011 dataset, which is rather lower than that in JPEG and JPEG2000 baselines. Our ITI model’s set/instance level reconstruction performance (IS, FID, and IPD) is comparable with the JPEG and JPEG2000 baselines at the time of low bit rate. but our compression ratio is ultrahigh

due to the low bitrate. So our ITI model (a paradigm of CMC) has the potential for the applications which require an ultrahigh compression ratio and only instance level or set level reconstruction, such as image/video data transmission for machine analysis with low bandwidth.

7 CONCLUSION AND FUTURE WORKS

In this paper, we proposed *Cross Model Compression (CMC)*, a novel compression framework towards human-comprehensible semantic compression. Compared with prior related frameworks, including traditional compression, ultimate feature compression, and intermediate feature compression, CMC is human-comprehensible and can be directly used in semantic monitoring. In particular, a paradigm of CMC, *Image-Text-Image (ITI)* is implemented for image compression. Qualitative and quantitative results showed that ITI can achieve an ultrahigh compression ratio and outperformed the JPEG baseline in set level and instance level. Our experimental results demonstrated that CMC has the potential for the applications which require an ultrahigh compression ratio and only set level or instance level reconstruction is required, such as image/video transmission for machine analysis with low bandwidth. Although encouraging results have been achieved, there are still several open problems related to CMC that we will continue exploring:

- (1) *End-to-end Optimization*. In this work, CMC encoder, and CMC decoder are optimized subsequently, which will result in suboptimal performance. End-to-end optimization can improve the performance further.
- (2) *CMC for Video Compression*. Compared with image data, video data are rather more redundant because of their time sampling. The video compression paradigm with higher compression performance is urgent due to the explosive growth of the video data on the Internet. CMC for video compression is more meaningful and necessary for the practical applications, but training deep models on video data is more difficult than that on image data.
- (3) *Scalable Compression*. In most image/video transmission scenarios, the network is fluctuant. Therefore rate control is necessary for the compression paradigm. Scalable CMC will make our paradigm more practical and bridge the gap between our experiments and practical applications.
- (4) *Semantic-based Metrics*. Although some semantic-based metrics have been proposed, the semantic-based compression frameworks have not been well evaluated so far. Better semantic-based metrics, especially differentiable metrics, are required for optimizing the compression model.

ACKNOWLEDGMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62025101 and No. 62088102, National Postdoctoral Program for Innovative Talents, China under Grant No. BX2021009, China Postdoctoral Science Foundation under Grant No. 2020M680238, and High-performance Computing Platform of Peking University, which are gratefully acknowledged.

REFERENCES

- [1] Nasir Ahmed, T. Natarajan, and Kamisetty R Rao. 1974. Discrete cosine transform. *IEEE transactions on Computers* 100, 1 (1974), 90–93.
- [2] Md Mushfiqul Alam, Tuan D Nguyen, Martin T Hagan, and Damon M Chandler. 2015. A perceptual quantization strategy for HEVC based on a convolutional neural network trained on natural images. In *Applications of Digital Image Processing XXXVIII*, Vol. 9599. International Society for Optics and Photonics, 959918.
- [3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning*. 214–223.
- [4] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations*.
- [5] Johannes Ballé, Valero Laparra, and Eero Simoncelli. 2017. End-to-end optimized image compression. In *5th International Conference on Learning Representations, ICLR 2017*.
- [6] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. 2018. Variational image compression with a scale hyperprior. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. <https://openreview.net/forum?id=rkcQFMZRB>
- [7] Zhuo Chen, Kui Fan, Shiqi Wang, Ling-Yu Duan, Weisi Lin, and Alex Kot. 2019. Lossy Intermediate Deep Learning Feature Compression and Evaluation. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2414–2422.
- [8] Zhibo Chen and Tianyu He. 2019. Learning based facial image compression with semantic fidelity metric. *Neurocomputing* 338 (2019), 16–25.
- [9] Hyomin Choi and Ivan V Bajic. 2018. Deep feature compression for collaborative object detection. In *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 3743–3747.
- [10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.
- [11] Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*. Springer, 15–29.
- [12] Wen Gao and Siwei Ma. 2014. An overview of AVS2 standard. In *Advanced Video Coding Systems*. Springer, 35–49.
- [13] Yunhao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014. Improving image-sentence embeddings using large weakly annotated photo collections. In *European conference on computer vision*. Springer, 529–545.
- [14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.
- [15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*. 6626–6637.
- [16] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)* 51, 6 (2019), 1–36.
- [17] David A Huffman. 1952. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE* 40, 9 (1952), 1098–1101.
- [18] Chuanmin Jia, Shiqi Wang, Xinfeng Zhang, Shanshe Wang, Jiaying Liu, Shiliang Pu, and Siwei Ma. 2019. Content-aware convolutional neural network for in-loop filtering in high efficiency video coding. *IEEE Transactions on Image Processing* 28, 7 (2019), 3343–3356.
- [19] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. 2015. Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE international conference on computer vision*. 2407–2415.
- [20] Jiahao Li, Bin Li, Jizheng Xu, Ruiqin Xiong, and Wen Gao. 2018. Fully connected network-based intra prediction for image coding. *IEEE Transactions on Image Processing* 27, 7 (2018), 3236–3247.
- [21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.
- [22] Yifan Liu, Zengchang Qin, Tao Wan, and Zhenbo Luo. 2018. Auto-painter: Cartoon image generation from sketch by using conditional Wasserstein generative adversarial networks. *Neurocomputing* 311 (2018), 78–87.
- [23] David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision* 60, 2 (2004), 91–110.
- [24] Siwei Ma, Xinfeng Zhang, Chuanmin Jia, Zhenghui Zhao, Shiqi Wang, and Shanshe Wang. 2019. Image and video compression with neural networks: A review. *IEEE Transactions on Circuits and Systems for Video Technology* (2019).
- [25] Michael W Marcellin, Michael J Gormish, Ali Bilgin, and Martin P Boliek. 2000. An overview of JPEG-2000. In *Proceedings DCC 2000. Data Compression Conference*.

- IEEE, 523–541.
- [26] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. 2016. Jointly modeling embedding and translation to bridge video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4594–4602.
- [27] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. 2017. To create what you tell: Generating videos from captions. In *Proceedings of the 25th ACM international conference on Multimedia*. 1789–1798.
- [28] Saurabh Puri, Sébastien Lasserre, and Patrick Le Callet. 2017. CNN-based transform index prediction in multiple transforms framework to assist entropy coding. In *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 798–802.
- [29] Alessandro Redondi, Luca Baroffio, Lucio Bianchi, Matteo Cesana, and Marco Tagliasacchi. 2016. Compress-then-analyze versus analyze-then-compress: What is best in visual sensor networks? *IEEE Transactions on Mobile Computing* 15, 12 (2016), 3000–3013.
- [30] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative Adversarial Text to Image Synthesis. In *33rd International Conference on Machine Learning*. 1060–1069.
- [31] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. 2017. Deep reinforcement learning-based image captioning with embedding reward. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 290–298.
- [32] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *Advances in neural information processing systems*. 2234–2242.
- [33] Claude E Shannon. 1948. A mathematical theory of communication. *Bell system technical journal* 27, 3 (1948), 379–423.
- [34] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. 2012. Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on circuits and systems for video technology* 22, 12 (2012), 1649–1668.
- [35] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.
- [36] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2818–2826.
- [37] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3156–3164.
- [38] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. *The Caltech-UCSD Birds-200-2011 Dataset*. Technical Report CNS-TR-2011-001. California Institute of Technology.
- [39] Gregory K Wallace. 1990. Overview of the JPEG (ISO/CCITT) still image compression standard. In *Image Processing Algorithms and Techniques*, Vol. 1244. International Society for Optics and Photonics, 220–233.
- [40] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. 2016. Image captioning with deep bidirectional LSTMs. In *Proceedings of the 24th ACM international conference on Multimedia*. 988–997.
- [41] Jing Wang, Jianlong Fu, Jinhui Tang, Zechao Li, and Tao Mei. 2018. Show, reward and tell: Automatic generation of narrative paragraph from photo stream by adversarial training. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- [42] N. Wang, X. Gao, L. Sun, and J. Li. 2017. Bayesian Face Sketch Synthesis. *IEEE Transactions on Image Processing* 26, 3 (2017), 1264–1274.
- [43] Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra. 2003. Overview of the H. 264/AVC video coding standard. *IEEE Transactions on circuits and systems for video technology* 13, 7 (2003), 560–576.
- [44] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1316–1324.
- [45] Yuan Yuan, Jingtao Sun, and Miaohui Wang. 2018. Dilated Deep Residual Network for Post-processing in TPG Based Image Coding. In *Internet and Distributed Computing Systems*, Yang Xiang, Jingtao Sun, Giancarlo Fortino, Antonio Guerrieri, and Jason J. Jung (Eds.). Springer International Publishing, Cham, 293–297.
- [46] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 5907–5915.
- [47] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2018. StackGAN++: Realistic image synthesis with stacked generative adversarial networks. *IEEE transactions on pattern analysis and machine intelligence* 41, 8 (2018), 1947–1962.
- [48] Zhenghui Zhao, Shiqi Wang, Shanshe Wang, Xinfeng Zhang, Siwei Ma, and Jian-sheng Yang. 2018. Enhanced Bi-Prediction With Convolutional Neural Network for High-Efficiency Video Coding. *IEEE Transactions on Circuits and Systems for Video Technology* 29, 11 (2018), 3291–3301.