

Multi-Modality Deep Network for Extreme Learned Image Compression

Xuhao Jiang¹, Weimin Tan¹, Tian Tan¹, Bo Yan^{1*}, Liquan Shen²

¹School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Shanghai Collaborative Innovation Center of Intelligent Visual Computing, Fudan University, Shanghai, China

²School of Communication, Shanghai University, Shanghai, China

20110240011@fudan.edu.cn, wmtan@fudan.edu.cn, tant21@m.fudan.edu.cn, byan@fudan.edu.cn, jsslq@163.com

Abstract

Image-based single-modality compression learning approaches have demonstrated exceptionally powerful encoding and decoding capabilities in the past few years, but suffer from blur and severe semantics loss at extremely low bitrates. To address this issue, we propose a multimodal machine learning method for text-guided image compression, in which the semantic information of text is used as prior information to guide image compression for better compression performance. We fully study the role of text description in different components of the codec, and demonstrate its effectiveness. In addition, we adopt the image-text attention module and image-request complement module to better fuse image and text features, and propose an improved multimodal semantic-consistent loss to produce semantically complete reconstructions. Extensive experiments, including a user study, prove that our method can obtain visually pleasing results at extremely low bitrates, and achieves a comparable or even better performance than state-of-the-art methods, even though these methods are at $2\times$ to $4\times$ bitrates of ours.

Introduction

During the past decades, image data on the Internet shows an explosive growth, bringing huge challenges for data storage and transmission. To meet this ever-increasing requirements, low-bitrate lossy image compression is a promising way to save storage and transmission bandwidth. Traditional image compression algorithms, e.g., Better Portable Graphics (BPG) (Bellard 2015) and Versatile Video Coding (VVC) (ITU-T and ISO/IEC 2020), are widely used in practice. However, they will cause serious blocking artifacts due to block-based processing at low bitrates. Therefore, exploring better methods for extreme image compression is urgently needed.

Recently, many single-modality learned methods (Xie, Cheng, and Chen 2021; Mentzer et al. 2020) have been proposed. However, they also fail to reconstruct satisfactory results at extremely low bitrates. Specifically, they may generate blurry results due to limited bits, or utilize Generative Adversarial Networks (GAN) to produce sharp results

*Corresponding author: Bo Yan. This work is supported by NSFC (Grant No.: U2001209, 61902076) and Natural Science Foundation of Shanghai (21ZR1406600).
Copyright © 2023, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

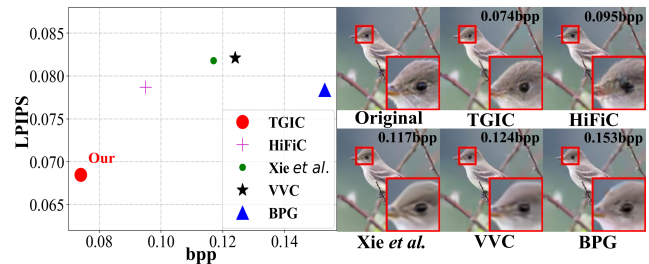


Figure 1: Visual comparisons of the proposed TGIC and some state-of-the-art methods (HiFiC (Mentzer et al. 2020), VVC (ITU-T and ISO/IEC 2020), BPG (Bellard 2015) and the work of Xie *et al.* (Xie, Cheng, and Chen 2021)).

whose textures may not be semantically consistent with the original image, as shown in Fig. 1. The text-to-image synthesis task is currently receiving a lot of attention, which generates semantically consistent images from text descriptions. Inspired by this task, multi-modality image compression may have great advantages at low bitrates. The corresponding text provides the high-level image semantic information, which can be used as prior information to assist image compression. Specifically, the text describes a rough content of the image and its local features, such as, color, location, shape, etc. This semantic information can be used to assist in reconstructing images, which can help save image bits. In addition, under the guidance of text semantic information, the multi-modality model can generate visually pleasing results that are more semantically consistent with the original image, as shown in Fig. 1. Note that the text description occupies very few bits and can be transmitted to the decoder side at marginal bandwidth cost. Even for low bitrates (0.2 bpp), text uses less than one-twentieth of the bits used by images on datasets (Wah et al. 2011; Nilsback and Zisserman 2008).

In this paper, a text-guided image compression (TGIC) generative adversarial network is proposed, in which the text description is utilized as prior information to assist in image compression. TGIC can produce better results compared with other methods, even if we use a much lower bitrate. For the image encoding, based on the image-text attention (ITA) module, text information is introduced into the codec

to guide the generation of compact feature representations. In the image decoding stage, we design an image-request complement (IRC) module to adaptive fuse the text and image information for better reconstructions. Besides, an improved multimodal semantic-consistent loss is designed to further improve the perceptual quality of reconstructions. The main contributions are as follows:

- We propose a novel codec framework for image compression, which utilizes the semantic information of the text description to improve coding performance. To the best of our knowledge, this is the first attempt that uses text semantic information as prior information to guide image compression.
- We fully study the role of text description in different components of the codec, and demonstrate its effectiveness for image compression. In particular, we adopt ITA to fuse image and text features, and propose IRC that allows the network to adaptively learn the much-needed guidance knowledge from text.
- The experiments (including a user study) show the outstanding perceptual performance of our TGIC in comparison with the existing learned image compression methods and traditional compression codecs.

Related Work

Lossy Image Compression

Lossy image compression has received significant attention from both academia and industry due to its huge practical value. Traditional compression standards, such as JPEG (Wallace 1992), JPEG2000 (Rabbani 2002), BPG (HEVC-Intra) (Bellard 2015) and VVC (ITU-T and ISO/IEC 2020) (the latest traditional codec), rely on hand-crafted module design. However, they ignore spatial correlations between image blocks, which results in image discontinuities at block boundaries.

Recently, many learned methods have been proposed to tackle the problem of image compression, and achieve promising results. Some early methods (Toderici et al. 2015, 2017) utilize the recurrent neural network to recursively compress the residual information, but they cannot directly optimize the rate in the training phase. The subsequent works are mainly based on variational autoencoder, and significant advances have been made progressively (Rippel and Bourdev 2017; Zhang et al. 2019; Chen et al. 2021; Ballé, Laparra, and Simoncelli 2017; Agustsson et al. 2017; Theis et al. 2017; Ballé et al. 2018; Mentzer et al. 2018; Lee, Cho, and Beack 2018; Hu et al. 2021; Cheng et al. 2020). Recognizing the huge potential of the hyperprior model (Ballé et al. 2018), many follow-up methods improve the entropy estimation techniques based on hyperprior design, such as coarse-to-fine model (Hu, Yang, and Liu 2020), joint model (Minnen, Ballé, and Toderici 2018) and 3D context entropy model (Guo et al. 2020). Besides, some methods (Agustsson et al. 2019; Mentzer et al. 2020; Tschannen, Agustsson, and Lucic 2019) based on GAN (Goodfellow et al. 2014) have been proposed for image compression at low bitrates. Mentzer *et al.* (Mentzer et al. 2020) investigate

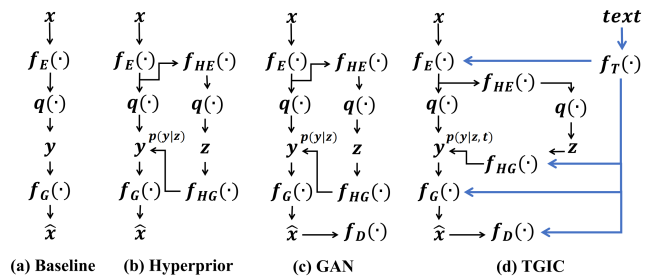


Figure 2: Overview of existing learned compression methods and the proposed TGIC. $f_E(\cdot)$ and $f_G(\cdot)$ represent the encoder and decoder of codec, respectively, and $q(\cdot)$ represents the quantization function. $f_{HE}(\cdot)$ and $f_{HG}(\cdot)$ denote the encoder and decoder of the auxiliary autoencoder (entropy model), respectively. $f_D(\cdot)$ and $f_T(\cdot)$ denote the discriminator and the text encoder, respectively.

normalization layers, generator and discriminator architectures, training strategies, as well as perceptual losses, and propose HiFiC which shows impressive performance.

However, the above-mentioned algorithms all show poor performance at extremely low bitrates, and even the excellent GAN-based method HiFiC shows general performance in this situation. The main reason is that it is impossible to faithfully reconstruct the entire content of the uncompressed image with extreme limited bits, and the GAN-based methods cannot generate images with realistic textures due to the lack of guidance from additional prior information. Therefore, benefiting from the semantic information provided by the text description, image compression based on multimodal machine learning may have a greater possibility to obtain better compression performance.

Multimodal Machine Learning

The multimodal machine learning has recently become a very hot topic due to its powerful advantages in the field of computer vision, such as text-to-image synthesis and image captioning. The text-to-image synthesis task aims to produce a high-quality image from a described text, such as (Zhang et al. 2017; Xu et al. 2018; Reed et al. 2016). For example, given the text description, AttnGAN (Xu et al. 2018) employs attention mechanism to produce images with photorealistic details. Contrary to the text-to-image synthesis task, the image captioning task (Shi et al. 2020; Anderson et al. 2018; Xu et al. 2015) is to generate a corresponding text description for a given image. Inspired by these works, we propose text-guided image compression, which uses the semantic information of text description as prior information to improve coding performance.

Method

Preliminaries

Figure. 2 provides an overview of learned image compression in the transform coding approach (Goyal 2001). As shown in Fig. 2 (a), the baseline model can be expressed

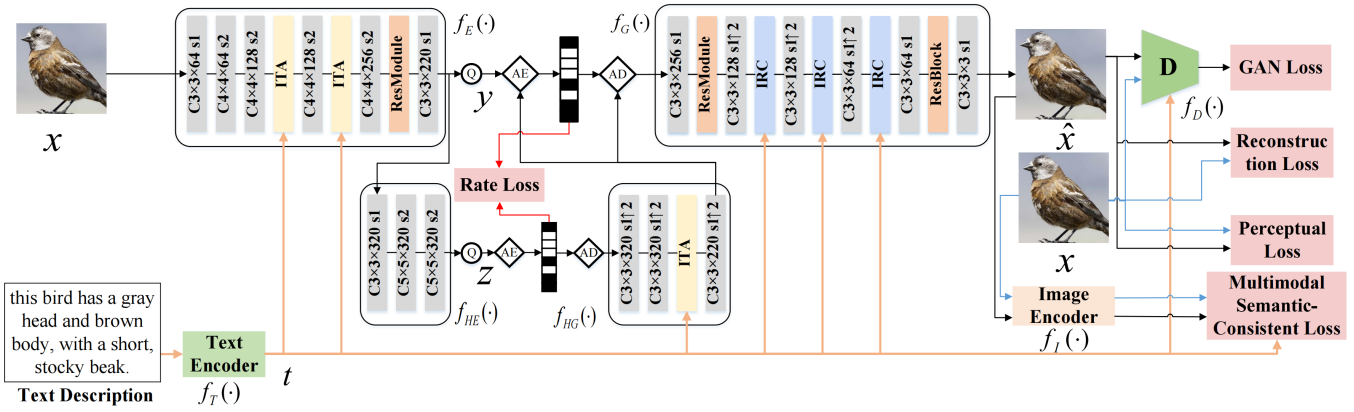


Figure 3: Architecture of our text-guided image compression (TGIC) model. $C3 \times 3 \times 64 \text{ s}2$ is a convolution with 64 channels, with 3×3 filters and stride 2. $\uparrow 2$ indicates the nearest neighbor upsampling. In addition, AE and AD are arithmetic encoding and decoding, and Q is for quantization. ITA is introduced to fuse image features and text features based on attention mechanism, and IRC is designed to adaptively use the text features for the image semantic complement. Resblock and ResModule are based on (He et al. 2016).

as

$$y = q(f_E(x)), \text{ then } \hat{x} = f_G(y), \quad (1)$$

where x , \hat{x} and y are raw images, reconstructed images and compressed codes, respectively. Specifically, x is encoded by an encoder $f_E(\cdot)$, and then quantized by $q(\cdot)$ to obtain y . y is then losslessly compressed into a bitstream by using entropy coding like arithmetic coding (Rissanen and Langdon 1981). For decoding, the decoder $f_G(\cdot)$ transforms y to obtain the \hat{x} . Here, the widely used quantization function proposed by (Ballé, Laparra, and Simoncelli 2017) is employed in our method. The corresponding rate R of y is estimated by a fully factorized density model $p_{y|\theta}$ during training, which is formulated by

$$R = \mathbb{E}[-\log_2 p_{y|\theta}(y|\theta)]. \quad (2)$$

Balle *et al.* first proposes a hyperprior model (Ballé et al. 2018), which utilizes the side information z to capture the spatial dependence of y , as shown in Fig. 2 (b). This idea is realized by introducing an additional entropy model. The side information z can be calculated by $z = q(f_{HE}(y))$, where $f_{HE}(\cdot)$ denotes the encoder of the entropy model. Then z is transformed by the decoder of the auxiliary autoencoder $f_{HD}(\cdot)$, which is used to estimate the distribution of y . Then the rate R is formulated by

$$R = \mathbb{E}[-\log_2 p_{y|z}(y|z)] + \mathbb{E}[-\log_2 p_{z|\theta}(z|\theta)]. \quad (3)$$

As shown in Fig. 2 (c), compared with the hyperprior model, the GAN-based model has one more discriminator, which generates the reconstructed images with fine-grained details through adversarial training.

Based on the previous works, we propose a new codec framework (as shown in Fig. 2 (d)), which utilizes the text information to assist in image compression. We will describe the proposed framework in detail in the following section.

Proposed Method

The architecture design of TGIC is shown in Fig. 3, which describes the details of the Fig. 2 (d). Specifically, the main

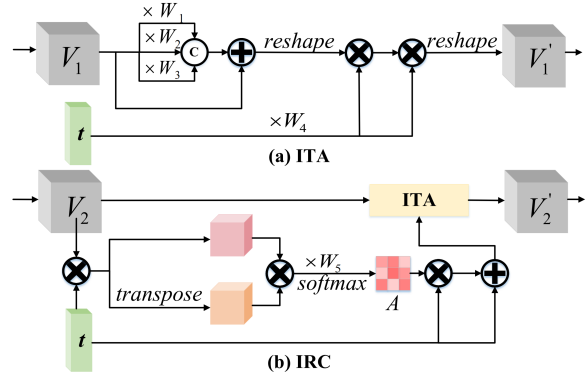


Figure 4: Architectures of ITA and IRC. C means concatenation. V_1 , V_2 , V'_1 and V'_2 are the image features.

body of TGIC is composed of four components: encoder, decoder, entropy model and discriminator. The image encoder and text encoder are employed to map the image and the text into a common semantic feature space. Firstly, the text features are extracted from the text description by using the text encoder. Then these text features are introduced into TGIC to assist image compression for better image reconstruction based on the image-text attention (ITA) module. In particular, we design an image-request complement module (IRC) in the decoder, so as to realize adaptive selection of text semantic information for the image semantic feature enhancement. Besides, we also design an improved multimodal semantic-consistency loss, which considers the semantic consistency between the reconstructions and the texts, as well as the uncompressed images.

Text-Guided Feature Representation

In our TGIC, the text encoder is a bi-direction Long Short-Term Memory (LSTM) (Schuster and Paliwal 1997), which

extracts the semantic features of texts. The corresponding calculation can be defined as

$$t = f_T(\text{text}), \quad (4)$$

where t denotes the text semantic features and $f_T(\cdot)$ represents the text encoder. Then t is input to the encoder and the entropy model respectively to achieve a compact feature representation.

Text guidance in the encoder. ITA is adopted to calculate the correlation between the text and image features, and then fuse these two features, as shown in Fig. 4 (a). Inspired by (Xu et al. 2018; Li, Tan, and Yan 2021), ITA uses multi-scale residual structure to further extract image features, and use matrix multiplication to calculate correlation, where W_1 , W_2 and W_3 represent the convolutional operations with different filter size, and W_4 is used to adjust the text feature dimension. The previous works (Agustsson et al. 2019; Duan, Chen, and Gu 2020) have verified that the semantic segmentation maps can improve coding performance. Considering that text can provide high-level semantic information of images, we utilize the high correlation between image and text features to achieve more compact image features. Therefore, y can be defined as $y = q(f_E(x, t))$, where $f_E(\cdot)$ represents the encoder of TGIC.

Text guidance in the entropy model. The hyperprior model (Ballé et al. 2018) improves the coding performance by introducing the side information, whose essence lies in predicting the distribution of latent features. The distribution of latent features is closely related to the content of the image. Considering that the text description can offer the image semantic information, the text information may help to predict the distribution of latent features. Inspired by (Li, Li, and Lu 2021), the text features are introduced to $f_{HG}(\cdot)$ of the entropy model. Similarly, this operation is also based on ITA. Then the entropy model can use the semantic relevance of text and image to predict the distribution of latent features more accurately. Under the guidance of the text features, the estimated distribution is converted from $p_{y|z}$ to $p_{y|z,t}$, which better parameterizes the distributions of latent codes and improves the entropy model performance.

Text-Guided Image Reconstruction

Due to the quantization operation, the image features in the decoder inevitably lose some information. Considering that the text description contains some semantic information of the image, the semantic features of the text is considered as prior information to enhance the image features. Then, the reconstructed image can be defined as $\hat{x} = f_G(y, t)$, where $f_G(\cdot)$ represents the decoder of TGIC. Aiming to make better use of text features for image feature enhancement, we propose the IRC to adaptively fuse the text information and image information. The architecture of IRC is shown in Fig. 4 (b). Firstly, IRC predict an attention map A based on the correlation between the input image features V_2 and text features, which is realized by the matrix multiplication of feature maps and is calculated as

$$A = \text{softmax}(W_5((V_2t)(V_2t)^T)) \quad (5)$$

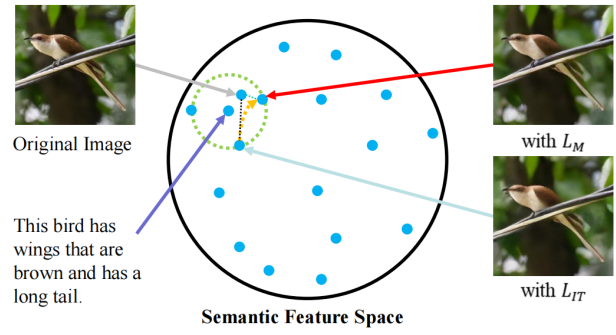


Figure 5: Projecting multi-modal embedding into the semantic feature space.

where W_5 represents the convolutional operation. Then we use the obtained A to weight and add the text features. Finally, the adaptive selected text features and the input image features are used to obtain the enhanced features V_2' by using ITA. This process can be expressed as $V_2' = ITA(V_2, At + t)$.

Text-Guided Adversarial Training

Conditional GAN (Mirza and Osindero 2014) is designed to learn a generative model of a conditional distribution, and shows excellent performance in many tasks, such as image super-resolution (Zhang et al. 2020). Considering the high correlation between the image and its corresponding text, TGIC uses text features as conditional information in the discriminator for better adversarial training. In the discriminator, the text features are first reshaped, then convolved and upsampled, so that these features can be concatenated with the image. Under the guidance of the text features, the discriminator maps an input (\hat{x}, t) to the probability $P_{x|t}$.

Multimodal Semantic-Consistent Loss

The proposed multimodal semantic-consistent loss is designed to constrain the semantic consistency of the reconstructed image and the original image as well as the text. AttnGAN (Xu et al. 2018) suggests to map the image features and text features into a common semantic space with the image encoder and text encoder, and calculates the negative log posterior probability to make the reconstructed image and the corresponding text semantically consistent. The corresponding loss is defined as

$$L_{IT} = -(\log P(t|f_I(\hat{x})) + \log P(f_I(\hat{x})|t)), \quad (6)$$

where $f_I(\cdot)$ represents the image encoder.

Although L_{IT} can constrain the semantics of the text description and reconstructions to be consistent, the images that conform to the text description are diverse. Therefore, we also add the constraint between the reconstructions and the uncompressed images, which is shown in Fig. 5. The multimodal semantic-consistent loss can be calculated as

$$L_M = L_{IT} + L_{II}, \text{ where } L_{II} = \beta \|f_I(\hat{x}) - f_I(x)\|_2, \quad (7)$$

where β is a hyper-parameter, and L_{IT} is the loss function which makes the reconstructions and the uncompressed images semantic-consistent. Specifically, as shown in Fig. 5, the text does not contain the description of the bird’s eyes, and the reconstruction quality of the eye part in the result generated only by L_{IT} is poor. However, using the improved L_M , the subjective quality of the reconstruction result is significantly improved.

In addition to L_M , we also use other four loss functions to optimize TGIC. Among them, the reconstruction loss L_R , GAN loss L_G and rate loss L_{Rate} are commonly used in the GAN-base codec. Since the text description takes up very few bits and its value is a constant, we do not consider it in the rate loss, but we take it into account when calculating the final bitrates. Here, these three loss functions are defined as

$$L_R = \|\hat{x} - x\|_2, \quad (8)$$

$$L_G = \mathbb{E}[\log f_D(x, t)] + \mathbb{E}[\log(1 - f_D(\hat{x}, t))], \quad (9)$$

$$L_{Rate} = \mathbb{E}[-\log_2 p_{y|z, t}(y|z, t)] + \mathbb{E}[-\log_2 p_{z|\theta}(z|\theta)], \quad (10)$$

where $f_D(\cdot)$ represents the discriminator. Following (Mentzer et al. 2020), the perceptual loss L_P based on a pretrained AlexNet (Krizhevsky, Sutskever, and Hinton 2012) is also adopted, which is defined as

$$L_P = \|\phi(\hat{x}) - \phi(x)\|_2, \quad (11)$$

where $\phi(\cdot)$ is the function of the pretrained AlexNet.

Finally, the global loss function of TGIC is defined as

$$L = \lambda L_{Rate} + k_1 L_R + k_2 L_G + k_3 L_P + k_4 L_M \quad (12)$$

where the k_1, k_2, k_3 , and k_4 are hyper-parameters, and the λ is a trade-off parameter to balance rate and distortion. Following (Mentzer et al. 2020), we adopt the constrained rate. The constrained rate introduces a ‘rate target’ hyper-parameter r_t and two hyper-parameter including λ_a and λ_b . When the calculated rate is larger than r_t , $\lambda = \lambda_a$. Otherwise, $\lambda = \lambda_b$. We can obtain a model with an average bitrate close to r_t by setting $\lambda_a \gg \lambda_b$

Training Implementation

Following AttnGAN (Xu et al. 2018), the image encoder and text encoder are pretrained to map the image features and text features into a common semantic space. Then the weights of the image encoder and text encoder are fixed while training the proposed TGIC. We adopt Pytorch as the training toolbox, and use the Adam optimization algorithm (Kingma and Ba 2014) with a mini-batch of 4 to optimize the model parameters. All the experiments are conducted on a NVIDIA GeForce RTX 1080 Ti. Our model is optimized for 300 epochs with a learning rate of 1×10^{-4} . The hyper-parameters k_1, k_2, k_3, k_4 and β of the global loss function are empirically set as 0.075×2^{-5} , 0.15, 5, 0.005 and 40, respectively. The λ_b is set as 2^{-4} , and the λ_a is set as $(2^3, 2^2, 2^1)$ to adapt to different bitrates.

Experiments

Datasets and Evaluation

Following the previous multimodal machine learning-based works (Zhang et al. 2017), the widely used datasets including CUB (Wah et al. 2011) and Oxford-102 (Nilsback and

Zisserman 2008) are employed for evaluation. CUB consists of 200 species of bird, with a total of 11,788 images including 8,855 images for training and 2,933 images for testing. Oxford-102 has 102 flower categories, of which 7,034 images are utilized for training and 1,155 images are utilized for testing. The images are resized and cropped into patches of size 256×256 .

The proposed TGIC aims at improving the codec performance at extremely low bitrates. Thus, we set the rate target below 0.25bpp for TGIC. Note that we use the bit sum of the image and the text to calculate the bitrates for TGIC, while we only utilize the bits of the image to calculate the bitrates for other methods. The bitrate of text are defined as $R_{text} = \frac{Size_{text} \times 8}{H \times W}$, where $Size_{text}$ denotes the file size of the text in bytes, H and W denote the height and width of the image, respectively. Following (Mentzer et al. 2020; Yang, Van Gool, and Timofte 2021), we use LPIPS (Zhang et al. 2018), KID (Bińkowski et al. 2018), FID (Heusel et al. 2017) to evaluate our TGIC and other compared methods, which are highly consistent with human perception of images. In addition, we also use Peak Signal-to-Noise Ratio (PSNR) to measure the fidelity of our results.

Comparison against SOTA Methods

In this part, TGIC is compared with state-of-the-art (SOTA) methods, including BPG (Bellard 2015), VVC (ITU-T and ISO/IEC 2020), HiFiC (Mentzer et al. 2020), and the works of Ballé et al. (Ballé et al. 2018), Minnen et al. (Minnen, Ballé, and Toderici 2018), Cheng et al. (Cheng et al. 2020) and Xie et al. (Xie, Cheng, and Chen 2021). Among these methods, BPG and VVC are the traditional methods, and others belong to the deep learning-based methods. Especially, HiFiC is a GAN-based model, aiming to produce results with better subjective quality. We use the BPG software (YCbCr 4:4:4) to test the images. For VVC, we use the VVC Official Test Model VTM 10.0 (YCbCr 4:4:4) of official version with an intra-profile configuration to test on images. We draw the rate-distortion (RD) curves to compare the coding performance of different methods.

Quantitative results. Fig. 6 shows the RD curve comparison on the CUB and Oxford-102 datasets. It can be observed that Our TGIC shows much better performance compared to other methods in terms of LPIPS, FID and KID. Especially, TGIC achieves a comparable or even better performance than other methods, even though these methods are at 2x to 4x bitrates of ours. In addition, we evaluate the fidelity of TGIC in terms of PSNR. As shown in Fig. 7, TGIC reaches competitive PSNR values compared with HiFiC. This verifies that TGIC can maintain an acceptable fidelity when compressing images towards subjective quality.

Qualitative results. Fig. 8 shows the visualization results of TGIC and other SOTA methods at similar bitrates. It can be found that the results of TGIC show better subjective quality than other algorithms. For VVC and BPG, their results have obvious blocking artifacts. The results of Xie et al. are blurry due to the MSE-based training. In addition, HiFiC utilizes GAN to obtain the hallucinated details, and produces visually pleasing reconstructions. Unfortunately, the hallucinated content may be inconsistent with the original image

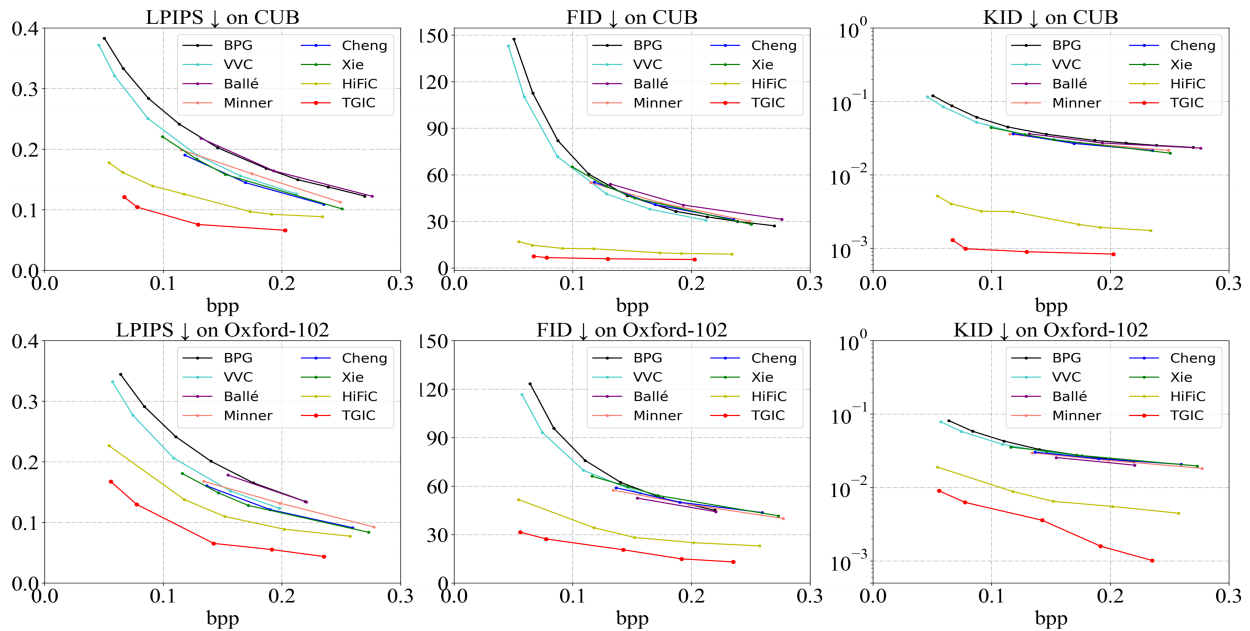


Figure 6: Performance evaluation on CUB and Oxford-102 datasets in terms of LPIPS, FID and KID.

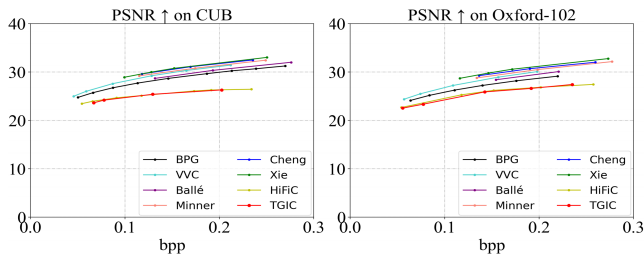


Figure 7: The PSNR results on CUB and Oxford-102.

content, resulting in obvious artifacts. Under the guidance of the text description, TGIC can produce satisfied reconstructions with photorealistic details, which are semantic-consistent with the original images.

User Study. A user study is further conducted with 32 participants. Given a pair of reconstructed images, the user is asked to judge which one owns a higher perceptual quality and is more consistent with the corresponding text description. We randomly select 50 images from CUB and 50 images from Oxford-102. Note that the bitrates of their compressed results by different methods are required to be similar. TGIC is compared with the state-of-the-art methods including VVC, HiFiC and the work of Xie *et al.*. This user study requires 9600 comparisons in total. Each participant needs to spend 40 minutes to complete the subjective test. As shown in Fig. 9, it can be found that the results of TGIC gain more preference than that of other methods, and all exceed 70%. This subjective comparisons are consistent with the quantitative results in Fig. 6, which verifies that our TGIC is superior to other algorithms.



Figure 8: Visual comparisons with SOTA approaches on CUB and Oxford-102 datasets. Above each line of the images is the corresponding text description. Better zoom in.

Ablation Study

In this part, we study and analyze the contributions of text guidance and different loss functions to our algorithm. Since the resulting bitrates do not exactly match the setting value, we try our best to compare performance of different models at the similar bitrates. We first remove L_P , L_M and all the text guidance including text-guided feature representation (TGFR), text-guided image reconstruction (TGIR) and text-guided adversarial training (TGAT) from our TGIC, and re-

Method	Baseline 0.088 bpp	Case 1			Case 2		Case 3		TGIC 0.078 bpp
		w/o TGFR 0.077 bpp	w/o TGIR 0.075 bpp	w/o TGAT 0.076 bpp	w/o L_P 0.077 bpp	w/o L_M 0.075 bpp	w/o IRC 0.076bpp	w/o L_{II} 0.078 bpp	
LPIPS↓	0.2411	0.1242	0.1289	0.1174	0.1625	0.1102	0.1297	0.1071	0.1048
FID↓	37.39	8.56	10.45	7.61	7.93	9.09	9.17	9.20	6.75
KID↓ (10^{-3})	18.78	1.68	2.93	1.24	1.46	2.06	2.23	1.77	0.99

Table 1: Performance Comparisons between variations of our TGIC on CUB dataset. The best results are boldfaced.

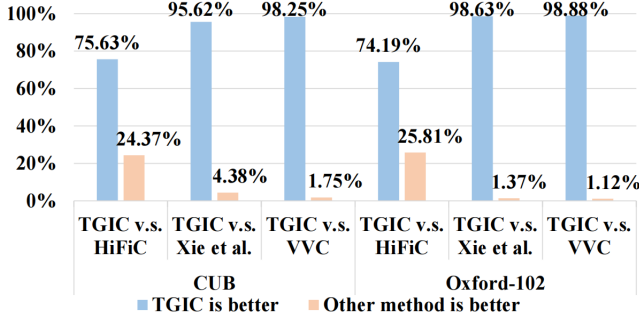


Figure 9: User study results. The reported value indicates the performance rate of the proposed TGIC against the other SOTA methods, respectively.

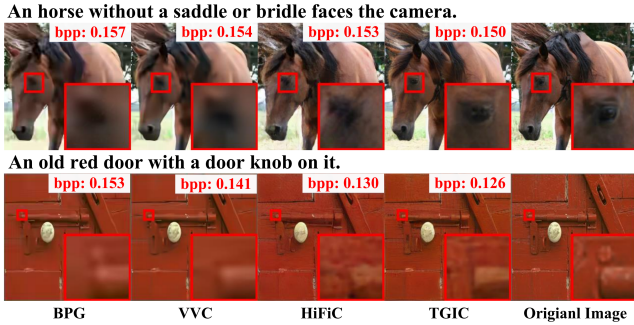


Figure 10: Visual comparisons with VVC on COCO (first row) and Kodak (second row). Better zoom in.

gard this model as the baseline model. As shown in Table 1, our TGIC obtains much better performance than baseline, which confirms the performance gain of introducing text.

Case 1: Effectiveness of text. We conduct the ablation study on the text effectiveness. We test the performances of TGIC without TGFR (w/o TGFR), TGIC without TGIR (w/o TGIR) and TGIC without TGAT (w/o TGAT). As shown in Table 1, the performances of three models all decrease compared with TGIC. Note that w/o TGIR has the largest performance drop and w/o TGAT has the smallest drop. This means that the text has the greatest effect in the decoder and the least effect in the discriminator.

Case 2: Effectiveness of L_P and L_M . Since L_R , L_{Rate} and L_G are common in GAN-based codec, we explore the effectiveness of L_P and L_M . As shown in Table 1, both have performance gains for our TGIC.

Case 3: Effectiveness of IRC and L_{II} . We also take the

Method	COCO		Kodak	
	BPP	LPIPS/FID/KID↓	BPP	LPIPS/FID/KID↓
BPG	0.153	0.289/181.5/0.0426	0.166	0.321/177.0/0.0578
VVC	0.152	0.250/175.3/0.0418	0.150	0.299/195.3/0.0621
HiFiC	0.140	0.143/110.3/0.0139	0.134	0.142/105.1/0.0198
TGIC	0.137	0.092/ 77.6 /0.0037	0.138	0.105/ 74.8 /0.0089

Table 2: Performance comparisons of our TGIC and VVC on COCO and Kodak datasets. Models are trained on training set of COCO, and tested on Kodak and 1000 images of testing set of COCO. The best results are boldfaced.

ablation study to verify the effectiveness of the proposed IRC and L_{II} . As shown in Table 1, the performances both decrease when removing IRC and L_{II} . This tells that IRC and L_{II} are beneficial to our model.

Text-guided Image Compression in Real-World Scenarios

In real-world scenarios, the image content is diverse and complex. To explore more possibilities of our TGIC, we also conduct experiments on COCO (Lin et al. 2014) with 80 types of objects. To further verify the generation performance, we also conduct the cross-dataset experiments on Kodak (Franzen 1999). Since Kodak does not have corresponding text descriptions, we consider using the image captions methods (e.g. OFA (Wang et al. 2022)) to generate texts for our experiments. The generated texts are of high quality and highly semantically with the images. We compare our TGIC with BPG, VVC and HiFiC, and the visualization results are shown in Table 2 and Fig. 10. We can find that our TGIC produces better results compared to other methods, which confirms the great potential of our TGIC.

Conclusion

In this paper, we propose a text-guided adversarial generation network for image compression (TGIC). We adopt the image-text attention module to introduce text information into the codec as prior information. Specifically, the text description can help the codec achieve a compact features representation, and can also be used for the image feature enhancement. In addition, we design an image-request complement module to adaptively learn the much-needed guidance knowledge of text information for feature enhancement. Moreover, a new multimodal semantic-consistent loss is well-designed that constrains the semantic consistency between the reconstructions, the texts and the uncompressed images. Experimental results demonstrate that TGIC outperforms the SOTA methods. Especially, even at bitrates below 0.1 bpp, the TGIC can produce appealing visual results.

References

- Agustsson, E.; Mentzer, F.; Tschannen, M.; Cavigelli, L.; Timofte, R.; Benini, L.; and Van Gool, L. J. 2017. Soft-to-Hard Vector Quantization for End-to-End Learning Compressible Representations. In *Advances in Neural Information Processing Systems*.
- Agustsson, E.; Tschannen, M.; Mentzer, F.; Timofte, R.; and Gool, L. V. 2019. Generative adversarial networks for extreme learned image compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 221–231.
- Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.
- Ballé, J.; Laparra, V.; and Simoncelli, E. P. 2017. End-to-end optimized image compression. In *5th International Conference on Learning Representations, ICLR 2017*.
- Ballé, J.; Minnen, D.; Singh, S.; Hwang, S. J.; and Johnston, N. 2018. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*.
- Bellard, F. 2015. BPG Image format. <https://bellard.org/bpg/>.
- Bińkowski, M.; Sutherland, D. J.; Arbel, M.; and Gretton, A. 2018. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*.
- Chen, T.; Liu, H.; Ma, Z.; Shen, Q.; Cao, X.; and Wang, Y. 2021. End-to-End Learnt Image Compression via Non-Local Attention Optimization and Improved Context Modeling. *IEEE Transactions on Image Processing*, 30: 3179–3191.
- Cheng, Z.; Sun, H.; Takeuchi, M.; and Katto, J. 2020. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7939–7948.
- Duan, S.; Chen, H.; and Gu, J. 2020. JPAD-SE: High-Level Semantics for Joint Perception-Accuracy-Distortion Enhancement in Image Compression. *arXiv preprint arXiv:2005.12810*.
- Franzen, R. 1999. Kodak lossless true color image suite. source: <http://r0k.us/graphics/kodak>, 4(2).
- Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.
- Goyal, V. K. 2001. Theoretical foundations of transform coding. *IEEE Signal Processing Magazine*, 18(5): 9–21.
- Guo, Z.; Wu, Y.; Feng, R.; Zhang, Z.; and Chen, Z. 2020. 3-d context entropy model for improved practical image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 116–117.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.
- Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.
- Hu, Y.; Yang, W.; and Liu, J. 2020. Coarse-to-fine hyperprior modeling for learned image compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 11013–11020.
- Hu, Y.; Yang, W.; Ma, Z.; and Liu, J. 2021. Learning end-to-end lossy image compression: A benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.
- ITU-T; and ISO/IEC. 2020. Versatile Video Coding. *ITU-T Rec. H.266 and ISO/IEC 23090-3*.
- Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. ImageNet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.
- Lee, J.; Cho, S.; and Beack, S.-K. 2018. Context-adaptive Entropy Model for End-to-end Optimized Image Compression. In *International Conference on Learning Representations*.
- Li, J.; Li, B.; and Lu, Y. 2021. Deep contextual video compression. *Advances in Neural Information Processing Systems*, 34: 18114–18125.
- Li, J.; Tan, W.; and Yan, B. 2021. Perceptual Variousness Motion Deblurring With Light Global Context Refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4116–4125.
- Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.
- Mentzer, F.; Agustsson, E.; Tschannen, M.; Timofte, R.; and Van Gool, L. 2018. Conditional Probability Models for Deep Image Compression. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4394–4402. IEEE.
- Mentzer, F.; Toderici, G. D.; Tschannen, M.; and Agustsson, E. 2020. High-Fidelity Generative Image Compression. *Advances in Neural Information Processing Systems*, 33.
- Minnen, D.; Ballé, J.; and Toderici, G. 2018. Joint Autoregressive and Hierarchical Priors for Learned Image Compression. In *NeurIPS*.
- Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.
- Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, 722–729. IEEE.

- Rabbani, M. 2002. JPEG2000: Image compression fundamentals, standards and practice. *Journal of Electronic Imaging*, 11(2): 286.
- Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*, 1060–1069. PMLR.
- Rippel, O.; and Bourdev, L. 2017. Real-time adaptive image compression. In *International Conference on Machine Learning*, 2922–2930. PMLR.
- Rissanen, J.; and Langdon, G. 1981. Universal modeling and coding. *IEEE Transactions on Information Theory*, 27(1): 12–23.
- Schuster, M.; and Paliwal, K. K. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11): 2673–2681.
- Shi, Z.; Zhou, X.; Qiu, X.; and Zhu, X. 2020. Improving Image Captioning with Better Use of Caption. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7454–7464.
- Theis, L.; Shi, W.; Cunningham, A.; and Huszár, F. 2017. Lossy image compression with compressive autoencoders. *arXiv preprint arXiv:1703.00395*.
- Toderici, G.; O’Malley, S. M.; Hwang, S. J.; Vincent, D.; Minnen, D.; Baluja, S.; Covell, M.; and Sukthankar, R. 2015. Variable rate image compression with recurrent neural networks. *arXiv preprint arXiv:1511.06085*.
- Toderici, G.; Vincent, D.; Johnston, N.; Jin Hwang, S.; Minnen, D.; Shor, J.; and Covell, M. 2017. Full resolution image compression with recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5306–5314.
- Tschannen, M.; Agustsson, E.; and Lucic, M. 2019. Deep generative models for distribution-preserving lossy compression. *Advances in Neural Information Processing Systems 31*, 5929–5940.
- Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset. *Technical Report CNS-TR-2011-001, California Institute of Technology*.
- Wallace, G. K. 1992. The JPEG still picture compression standard. *IEEE transactions on consumer electronics*, 38(1): xviii–xxxiv.
- Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *arXiv preprint arXiv:2202.03052*.
- Xie, Y.; Cheng, K. L.; and Chen, Q. 2021. Enhanced invertible encoding for learned image compression. In *Proceedings of the 29th ACM International Conference on Multimedia*, 162–170.
- Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057. PMLR.
- Xu, T.; Zhang, P.; Huang, Q.; Zhang, H.; Gan, Z.; Huang, X.; and He, X. 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.
- Yang, R.; Van Gool, L.; and Timofte, R. 2021. Perceptual Video Compression with Recurrent Conditional GAN. *arXiv e-prints*, arXiv–2109.
- Zhang, H.; Xu, T.; Li, H.; Zhang, S.; Wang, X.; Huang, X.; and Metaxas, D. N. 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.
- Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.
- Zhang, X.; Song, H.; Zhang, K.; Qiao, J.; and Liu, Q. 2020. Single image super-resolution with enhanced Laplacian pyramid network via conditional generative adversarial learning. *Neurocomputing*, 398: 531–538.
- Zhang, Y.; Li, K.; Li, K.; Zhong, B.; and Fu, Y. 2019. Residual non-local attention networks for image restoration. *arXiv preprint arXiv:1903.10082*.