Multi-Modality Deep Network for Extreme Learned Image Compression

Abstract

Recent advancements in the realm of image-based compression utilizing singular modality techniques have showcased remarkable proficiency in the realms of data encoding and decoding. Nonetheless, these methods are plagued by deficiencies, notably the manifestation of image blurring and profound semantic degradation when operating at exceedingly diminished bitrate thresholds. In response to this predicament, our research introduces a sophisticated multimodal machine learning algorithm designed to refine image compression through the integration of textual semantic information as a supplementary directive. This auxiliary textual data acts as a preemptive cognizance to enhance the compression mechanism, thereby yielding superior results. The impact of textual descriptions has been thoroughly investigated across various segments of the compression-decompression architecture, affirming their utility in the process. The integration strategy involves the deployment of an image-text attention module along with an image-request complement module, both of which synergistically amalgamate visual and linguistic features. Furthermore, we have innovated an augmented multimodal semantic-consistent loss function, which is pivotal in facilitating the generation of semantically intact reconstructed images. Rigorous experimental analyses substantiate the efficacy of our proposed approach, confirming its capability to produce aesthetically satisfying reconstructions even at the threshold of extreme bitrate minimization.

Keywords: Image Compression, Multimodality.

1 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) [6] and Versatile Video Coding (VVC) [34], 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 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

whose textures may not be semantically consistent with the original image. 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. Note that the text description occupies very few bits and can be transmitted to the decoder side at marginal bandwidth cost.

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:

- A novel codec framework for image compression which utilizes the semantic information of the text description to improve coding performance is proposed.
- The role of text description in different components of the codec has been fully studying, and its effectiveness for image compression has been demonstrated. In particular, ITA is adopted to fuse image and text features, and IRC is proposed to allow the network to adaptively learn the much-needed guidance knowledge from text.
- The experiments (including a user study) show the outstanding perceptual performance of TGIC model in comparison with the existing learned image compression methods and traditional compression codecs.

2 Related works

2.1 Lossy Image Compression

Lossy image compression is a topic of considerable importance in both academic research and industrial applications due to its substantial practical implications. Conventional compression standards, such as JPEG [33], JPEG2000 [28], BPG (based on HEVC-Intra) [6], and the more recent VVC [34], rely on manually crafted modules. These traditional algorithms often overlook the spatial interdependencies among image segments, which can lead to visible discontinuities along block edges in the compressed output.

In contrast to these traditional approaches, contemporary research has witnessed the emergence of learned methods that address the issue of image compression more effectively, yielding notable improvements. Initial techniques [30] employed recurrent neural networks to iteratively compress residual

information; however, these methods were incapable of directly optimizing the bitrate during the training phase. Subsequent studies have predominantly utilized variational autoencoders, with progressive enhancements being reported [1, 4, 5, 7, 8, 13, 16, 19, 24, 29, 40]. Hyperprior models [5] have been recognized for their considerable potential, spurring the development of advanced entropy estimation methods that build upon this concept, including hierarchical models [12], integrated architectures [21], and three-dimensional context entropy schemes [10]. Additionally, methods [2, 20, 31] leveraging Generative Adversarial Networks (GANs) [9] have been proposed for low bitrate image compression. Mentzer et al. [20] thoroughly examined aspects such as normalization layers, generator and discriminator structures, training approaches, and perceptual loss functions, leading to the introduction of HiFiC, a model with commendable results.

Despite these advances, the aforementioned algorithms generally exhibit subpar performance at extremely low bitrates. Even the sophisticated GAN-based HiFiC approach does not escape this limitation. The underlying challenge is the inherent difficulty in accurately reconstructing an uncompressed image when only a minuscule amount of bits are available. Furthermore, GAN-based models are impeded by their inability to synthesize images with realistic textures in the absence of supplementary prior information. To surmount this challenge, there is a growing recognition that leveraging the rich semantic content embedded in textual descriptions can significantly enhance image compression outcomes. Thus, multimodal machine learning approaches that incorporate textual data hold substantial promise for achieving superior compression performance, particularly at very low bitrates.

2.2 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 [23, 36, 38]. For example, given the text description, AttnGAN [36] employs attention mechanism to produce images with photorealistic details. Contrary to the text-to-image synthesis task, the image captioning task [3, 27, 35] 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.

3 Method

3.1 Overview

The architecture design of the text-guided image compression (TGIC) is shown in Figure 1. C3 $\times$ 3 $\times$ 64 s2 is a convolution with 64 channels, with 3$\times$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 [11]. Specifically, the main body of TGIC is composed of four components:

encoder, decoder, entropy model and discriminator.

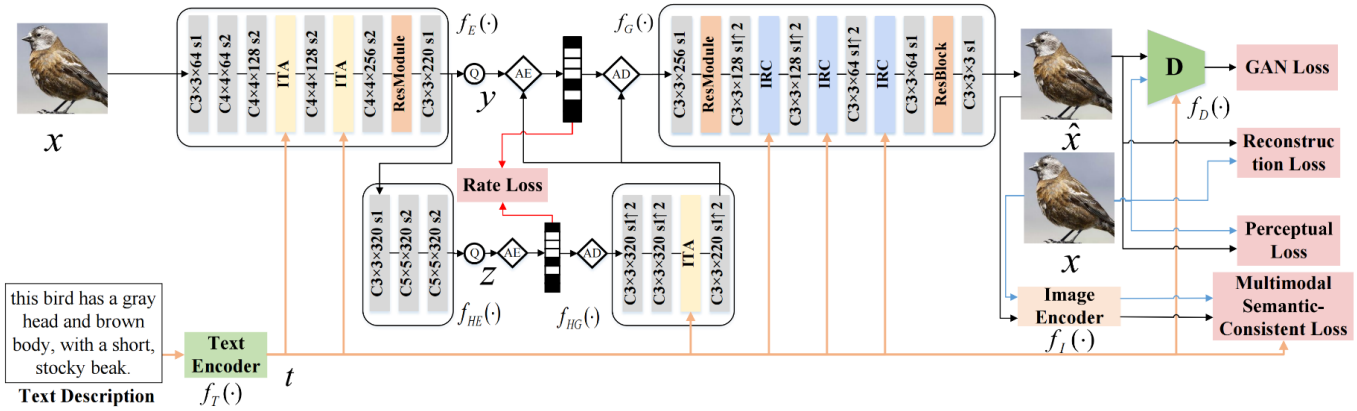

Figure 1. Architecture of TGIC model.

The ITA module uses multiscale 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 architecture of ITA is shown in Figure 2(a). The IRC module is proposed to adaptively fuse the text information and image information. The architecture of IRC is shown in Figure 2(b), where W5 represents the convolutional operation. Then the obtained A is used to weight and add the text features. Finally, the adaptive selected text features and the input image features are fused to obtain the enhanced features V_2' by using ITA.

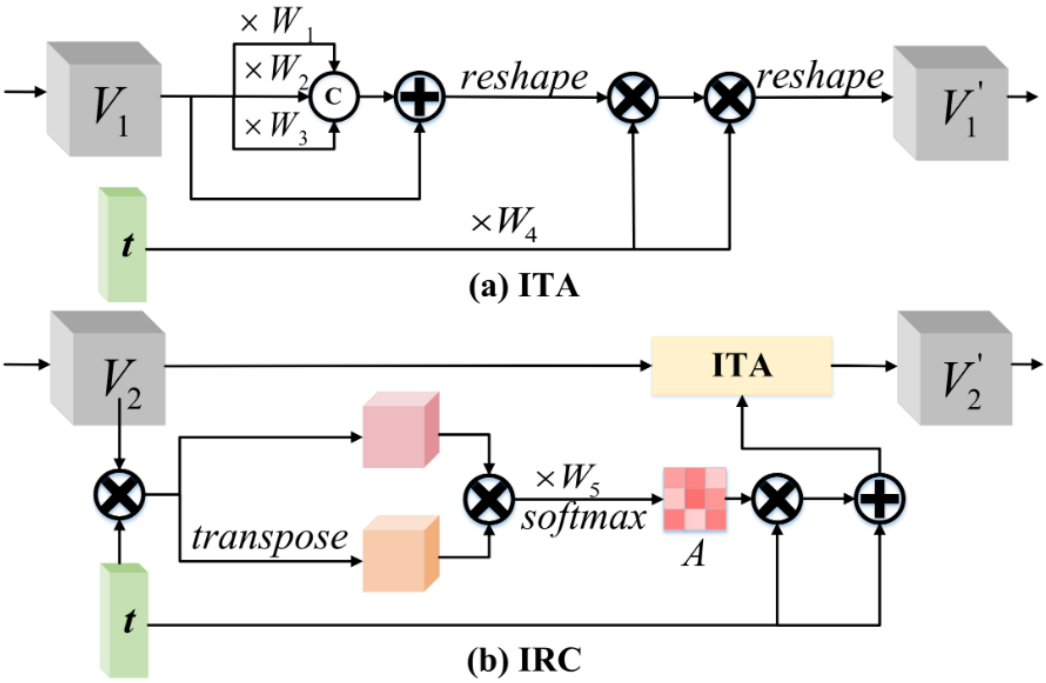

Figure 2. Architectures of ITA and IRC. C means concatenation. V_1, V_2, V_1' and V_2' are the image features.

3.2 Loss

3.2.1 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 [36] 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

$$L_{IT} = -\left(\log P(t|f_I(\hat{x})) + \log P(f_I(\hat{x})|t)\right) \tag{1}$$

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

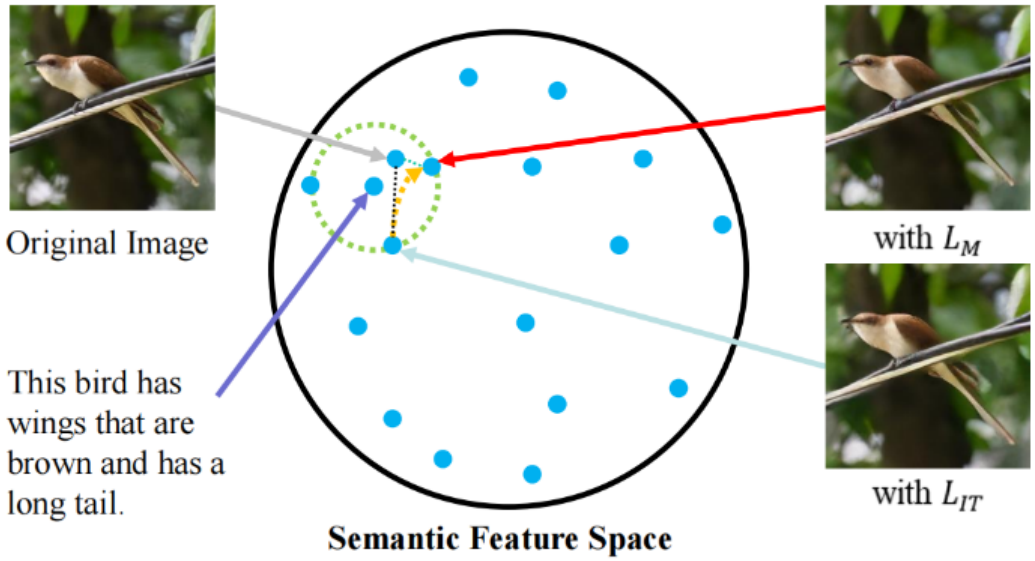

Figure 3. Projecting multi-modal embedding into the semantic feature space.

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 Figure 3. The multimodal semantic-consistent loss can be calculated as

$$L_M = L_{IT} + LII, where L_{II} = \beta||f_I(\hat{x}) - f_I(x)||_2 \tag{2}$$

where β is a hyper-parameter, and L_{II} is the loss function which makes the reconstructions and the uncompressed images semantic-consistent.

3.2.2 Reconstruction Loss

The reconstruction loss is the calculation of mean squared error between original image and generated image. The original image and the generated image are represented by x and $\hat{x}$, respectively. The reconstruction loss can be calculated as

$$L_R = ||\hat{x} - x||_2 \tag{3}$$

3.2.3 GAN Loss

The GAN loss L_G is commonly used in the GAN-based codec. In GAN, the goal of the generator is to generate a composite image that is close to the original image, while the goal of the discriminator is to distinguish between the real image and the generated image. The GAN loss between the image generated by the generator and the original image is used in the discriminator to measure the quality of the generated image, which can be calculated as

$$L_G = \mathbf{E}[\log f_D(x,t)] + \mathbf{E}[\log(1 - f_D(\hat{x},t)] \tag{4}$$

where $f_D(\cdot)$ represents the discriminator.

3.2.4 Rate Loss

The rate loss in entropy coding refers to the loss caused by the use of more efficient encoding methods during data compression, resulting in an increase in the bit rate (i.e. the compressed data size) of the output data compared to the original data. In this work, as shown in Figure 1, the rate loss represents the loss of both the output of $f_E(\cdot)$ and the output of $f_{HE}(\cdot)$ after entropy encoding, respectively. The rate loss can be defined as

$$L_{Rate} = \mathbf{E}[-\log_2 p_{y|z,t}(y|z,t)] + \mathbf{E}[-\log_2 p_{z|\theta}(z|\theta)] \tag{5}$$

3.2.5 Perceptual Loss

Following [20], the perceptual loss L_p based on a pretrained AlexNet [14] is also adopted, which is defined as

$$L_p = ||\phi(\hat{x}) - \phi(x)||_2 \tag{6}$$

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

4 Implementation details

4.1 Comparing with the released source codes

Due to the novelty of this work, there is currently no open source code available. I have built the entire model and completed all the implementation details described in the work.

In the case of non overlapping partitioning of the data preprocessing stage, the processing of each block is independent, which may lead to distortion at the edge of the block because the block boundary does not consider the information of adjacent blocks. Therefore, the solution is to choose a stride, which is the overlap size between image blocks. A step size smaller than the size of the block will result in overlapping areas. And calculate how many complete blocks can be divided for the width and height of the image. For image edges, if the remaining width or height is less than the size of the block, the following methods can be chosen for processing. The first one is to reduce the step size until the remaining portion can accommodate a complete block. Another method is to start dividing from the edges of the image, ensuring that the edge parts are also fully covered.

In the initial stage of training, insufficient complexity of the discriminator model in GAN can lead to prolonged training time. Due to the use of GAN in this work and the fact that the entire framework is equivalent to a generator, and generally speaking, a discriminator requires stronger model complexity to correctly classify the images generated by the generator, thereby improving its generation ability. Therefore, I used the weight parameters of the pretrained ResNet50 model based on ImageNet1k in this work, and modified the last layer to a binary classification layer and a sigmoid activation function. The discriminator has strong classification ability during initial training, making the generator mostly oriented towards generating high-quality images during the training process, greatly reducing model training time and improving the performance of the generator in generating images. In addition, the implementation details of the image encoder is not mentioned by this paper. Therefore, I use the transformer encoder [32] to extract the common features between the original image and the generated image, optimizing this model further.

4.2 Experimental environment setup

4.2.1 Dataset description

I trained the TGIC model based on the benchmark dataset [26], which covers various image contents, including outdor, indoor, landscape, nature, people, objects and buildings. The dataset consists of 202 high-definition original images with the size of 1920×1080 and 7878 redundancy-removed images by VVC. Each original image has 39 encoded versions with different redundancy-removed levels, which is selected by the subjective experiments.

To demonstrate the generalization capability, we also independently conduct the testing experiments of seven additional benchmark datasets, including CSIQ [15] (30 original natural-content images with the size of 512×512), KADID-10K [17] (81 original natural-content images with the size of 512×384), LIVE [25] (29 original natural-content images with various resolutions), and TID2013 [22] (25 original natural-content images with the size of 512×384).

4.2.2 Training guidelines

I randomly select image pairs from the dataset as the training set, validating set, and testing set based on the ratio of 8:1:1. In the text preprocessing stage, the dimension of a single word vector is uniformly mapped to 300. For the transformer encoder module, I have executed 3 sets of transformer blocks, where patch size is set to 4 and number of heads is set to 4. In the training stage, epoch is set to 100, learning rate is set to 1×10^{-4} and batch size is set to 3.

4.2.3 Evaluation metrics

The proposed TGIC aims at improving the codec performance at extremely low bitrates. I only utilize the bits of the image to calculate the bitrates for other methods, which is as described in this paper. Following [20, 37], I use LPIPS [39] to evaluate our TGIC and other compared methods, which are highly consistent with human perception of images. In addition, I also use Peak Signal-to-Noise Ratio (PSNR) to measure the fidelity of our results.

4.3 Interface design

1. The interface is concise and easy to understand, with three functions. The first function is to upload an image, the second function is to process the uploaded image, and the third function is to compress the processed image, as shown in Figure 4.

2. The upload image function mainly displays the original image and its BPP results.

3. The operation of image processing includes generating corresponding text data for the image and dividing the image into blocks.

4. The specific description of image compression function is as follows: calling the trained model to infer the image, completing the inference, then merging all segmented images, finally displaying a complete result image, and calculating its BPP result.

5. In addition, during processing, it has the effect of waiting in circles and displaying the current inference progress.

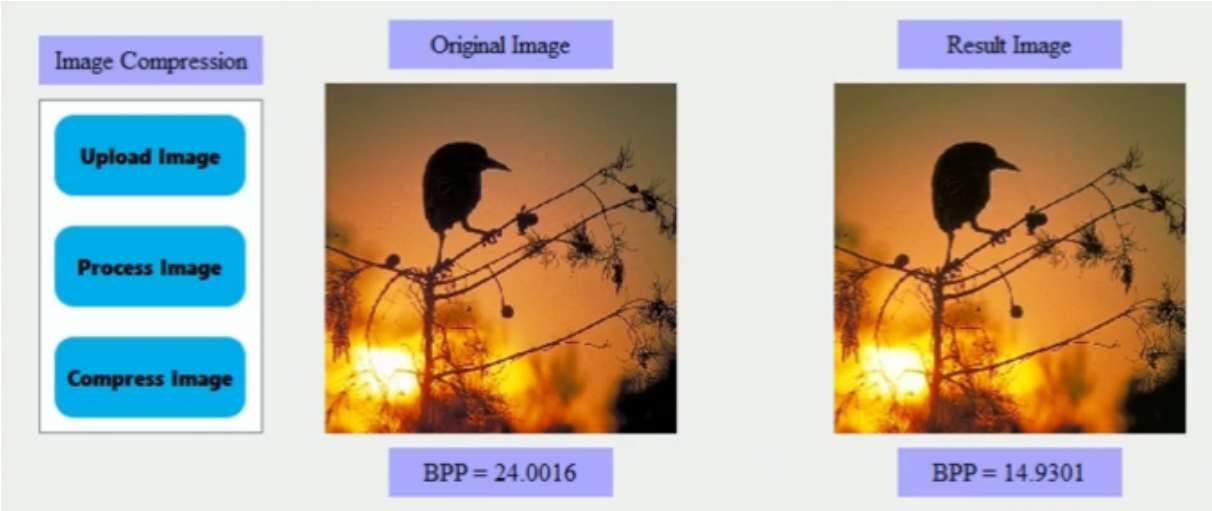

Figure 4. Interface of Image Compression.

4.4 Main contributions

1. Through a comprehensive understanding and reasoning of the key details described in this paper, I reproduced and improved the entire neural network model described in this paper without open source code.

2. The overlapping blocking method is adopted so that each block considers the local structure information of the image when processing to reduce visual distortion at block boundaries, especially during image reconstruction. But its disadvantage is that it increases the amount of training data, the number of iterations, and the training time.

3. Since the dataset I am using only has images, I use LLAVA [18] that is a small open source language model to get the text data.

4. In GAN, in order to optimize the training process and enable the model to converge quickly, I used the ResNet50 model pretrained on the ImageNet1k as a discriminator to perform strong discrimination on the images generated by the generator.

5. The implementation details of the image encoder is not mentioned by this paper. I use the transformer encoder to extract the common features between the original image and the generated image, optimizing this model further.

6. The original paper used fewer datasets to guide the experiment and present the results. Different datasets may have different data distributions, noise levels, and outliers. In order to better verify the generalization ability of the model and evaluate its performance in various situations, I have added more experiments on this model on the dataset.

7. I built a simple and easy-to-understand interface, which can upload an original image, divide the image into chunks on the backend, then call the trained model for inference, and finally merge all the chunked images to form a complete image to be displayed in the interface as the final result. Additionally, there is an option to save the resulting image at the end.

5 Results and analysis

From Figure 5, it can be observed that TGIC exhibits better performance in LPIPS compared to other methods. Especially, TGIC has achieved performance comparable to or even better than other methods. In addition, I also evaluated the fidelity of TGIC based on PSNR, as shown in Figure 6. Compared with JPEG, TGIC achieved a competitive PSNR value. This verifies that TGIC can maintain acceptable fidelity when compressing images towards subjective quality. The improved TGIC has a lower LPIPS value, which means it can save more bit rates when compressing images.

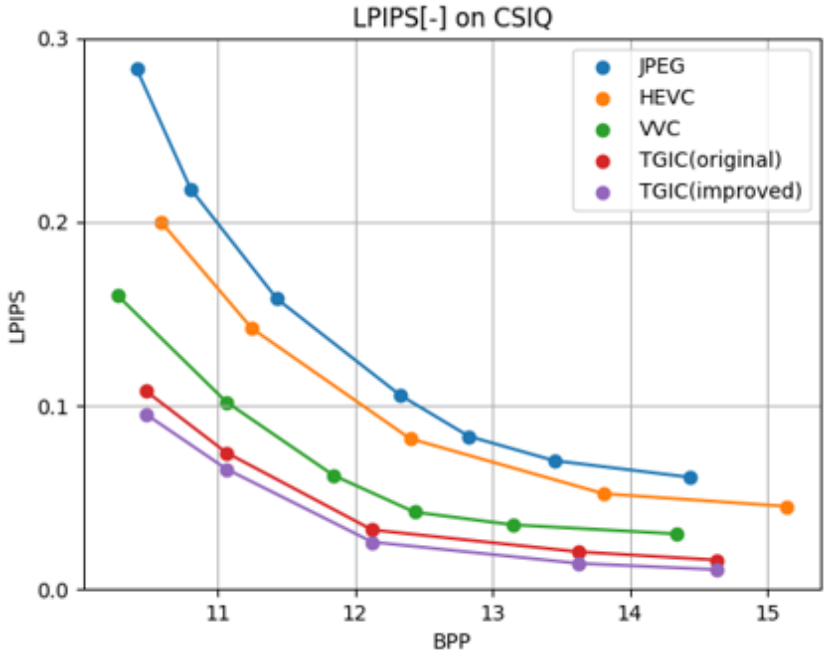

Figure 5. LPIPS [-] on CSIQ.

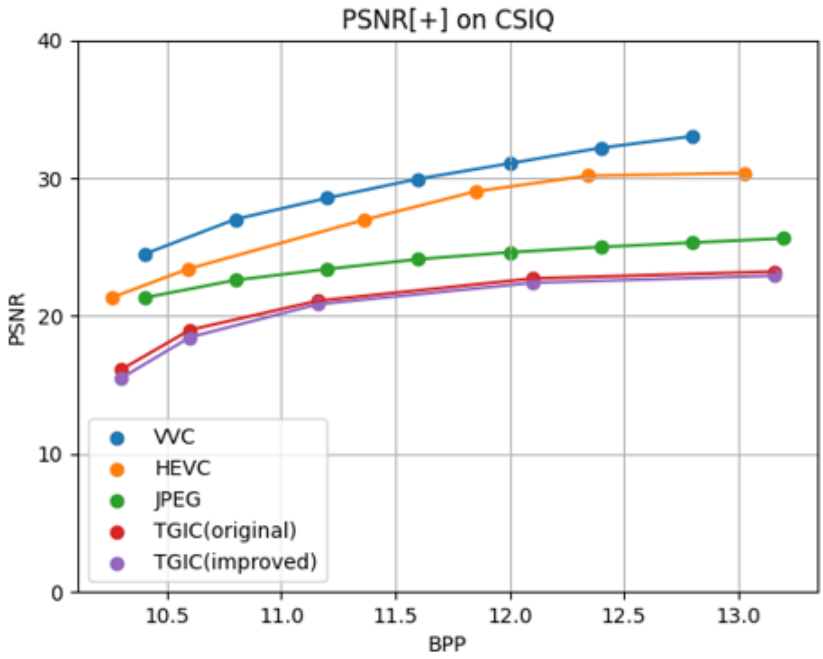

Figure 6. PSNR [+] on CSIQ.

Figure 7 shows that the use of overlapping image block strategy in TGIC significantly increases the amount of training data, resulting in an increase in model training time and slower model convergence. However, in overlapping partitioning, due to the overlap between blocks, each block takes into account the local structural information of the image within the block during processing to reduce visual distortion at the block boundaries, especially during image reconstruction. These overlapping blocks can share information, and compression algorithms can more effectively utilize these redundant information to reduce the encoded bit rate, thereby improving compression efficiency. In addition, the processing of overlapping blocks can reduce the common block effect phenomenon in image compression, as the smooth transition of overlapping areas helps mask the discontinuity between blocks. For areas with rich details, greater overlap can be used to maintain high quality, while for areas with simple textures, overlap can be reduced to save resources. Moreover, using ResNet50, which is pretrained based on ImageNet1k, as a discriminator enhances the discriminator's discriminative ability in the early stages of training, accelerates model convergence, and reduces model iteration times. Although the complexity of the model is greatly increased and the training time is prolonged, the powerful pretrained effect makes the model less likely to require training to complete the discriminative task. After ultimately combining these two improvements, the model converges faster and the number of training iterations is reduced. Due to the increased amount of data generated by the partitioning strategy, the time for each training session still increases, but the saved BPP is greater than any single improvement.

Methods \ Performance	Epochs	Time per epoch(min)	BPP
TGIC (original)	74	56	11.6798
TGIC + Overlapable blocking	↑ 13.9535%	↑ 28.5714%	↓ 0.8605%
TGIC + ResNet50 (Discriminator)	↓ 41.8919%	↓ 3.5714%	↓ 1.0430%
TGIC (improved)	↓ 29.7297%	↑ 7.1429%	↓ 1.5086%

Figure 7. Performance comparison of TGIC models before and after improvement.

Different datasets may have different data distributions, noise levels, and outliers. In order to better validate the generalization ability of the model and evaluate its performance in various scenarios, I added experiments on the model on more datasets, as shown in Figure 8.

BPP / Methods \ Datasets	LIVE2005	CSIQ	TID2013	KADID10K	SHEN2020	Average
Origin	19.7743	18.0334	16.4331	16.8663	11.1821	16.4578
JPEG	18.2156	14.5167	13.3958	15.8739	10.9306	14.5865
HEVC	16.6146	12.3290	10.1637	13.1552	10.7742	12.6073
VVC	15.8374	11.8490	9.0331	12.6226	10.4883	11.9661
TGIC (original)	15.1791	11.6834	9.0723	12.1255	10.3386	11.6798
TGIC (improved)	14.9535	11.4765	8.9282	11.9985	10.1611	11.5036

Figure 8. Experimental results of 5 methods for saving BPP on multiple datasets.

Figure 9 shows the visualization results of TGIC and other SOTA methods at similar bit rates. It can be observed that TGIC can achieve higher compression ratios while ensuring a certain level of fidelity in the image. Unfortunately, the TGIC model may produce hallucinatory changes in color that are inconsistent with the original image, leading to certain visual differences. This is due to the combined effect of the generator and discriminator in GAN. Although the generated image may have visual color differences from the original image, color readjustment can also reduce image compression by more bit rates.

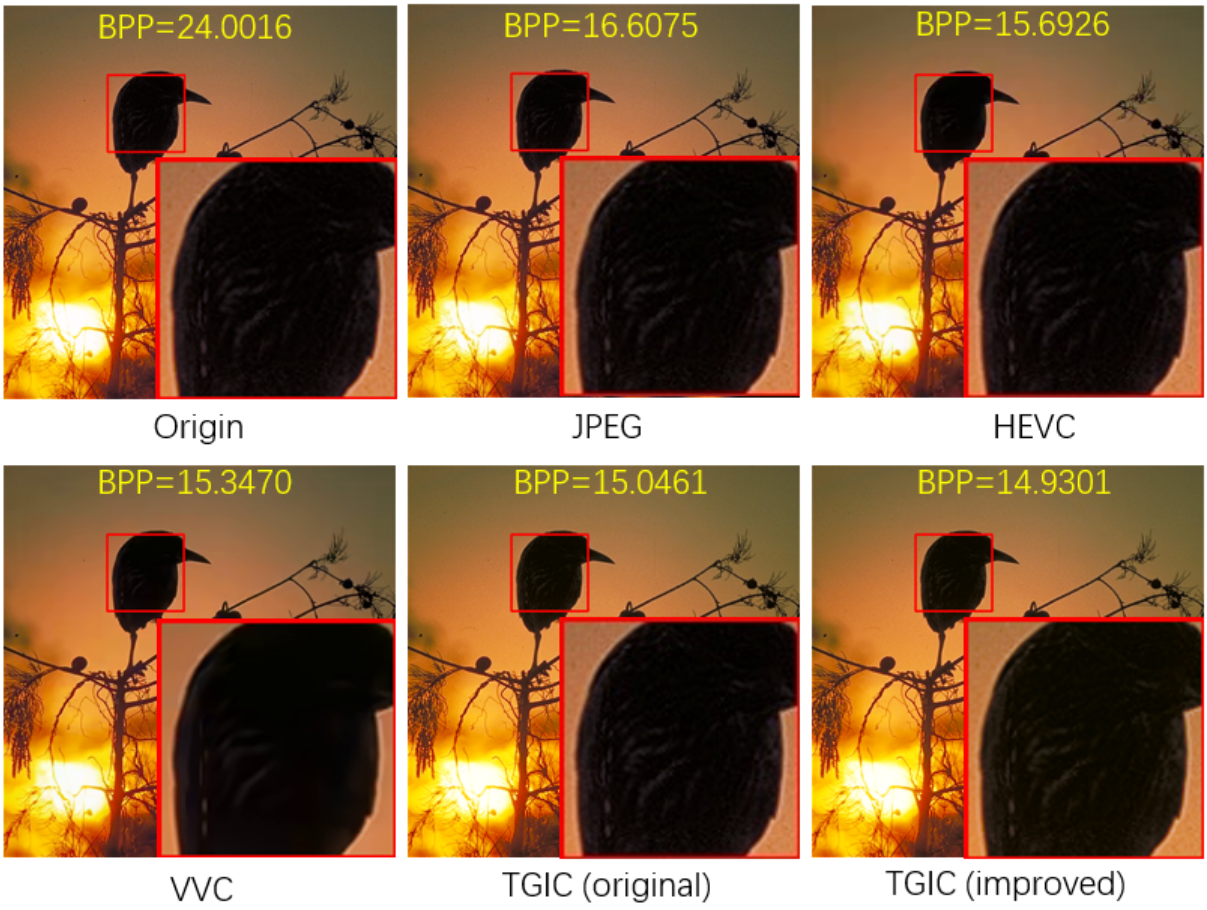

Figure 9. Visual comparisons with SOTA approaches on the CSIQ datasets.

6 Conclusion and future work

In this paper, a text-guided adversarial generation network for image compression (TGIC) is proposed. The image-text attention module is adopted 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, an image-request complement module is designed 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.

Deep learning has made great breakthroughs in the field of image processing and can be applied to the field of image compression. By using deep learning models, more efficient image compression algorithms can be achieved, improving image quality and compression ratio. Attention mechanism can help models focus more on important information when compressing images, thereby improving compression quality. By introducing attention mechanisms, more precise image compression can be achieved, reducing information loss. With the widespread application of multimodal data (i.e. images, audio, text), multimodal data compression has become an important research direction. By jointly compressing data from different modalities, more efficient data transmission and storage can be achieved.

References

[1] Eirikur Agustsson, Fabian Mentzer, Michael Tschannen, Lukas Cavigelli, Radu Timofte, Luca Benini, and Luc V Gool. Soft-to-hard vector quantization for end-to-end learning compressible representations. Advances in neural information processing systems, 30, 2017.

[2] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 221–231, 2019.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 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, pages 6077–6086, 2018.

[4] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. End-to-end optimized image compression. arXiv preprint arXiv:1611.01704, 2016.

[5] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. arXiv preprint arXiv:1802.01436, 2018.

[6] Fabrice Bellard. Bpg image format. URL https://bellard. org/bpg, 1(2):1, 2015.

[7] Tong Chen, Haojie Liu, Zhan Ma, Qiu Shen, Xun Cao, and Yao Wang. End-to-end learnt image compression via non-local attention optimization and improved context modeling. IEEE Transactions on Image Processing, 30:3179–3191, 2021.

[8] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 7939–7948, 2020.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems, 27, 2014.

[10] Zongyu Guo, Yaojun Wu, Runsen Feng, Zhizheng Zhang, and Zhibo Chen. 3-d context entropy model for improved practical image compression. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pages 116–117, 2020.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[12] Yueyu Hu, Wenhan Yang, and Jiaying Liu. Coarse-to-fine hyper-prior modeling for learned image compression. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 11013–11020, 2020.

[13] Yueyu Hu, Wenhan Yang, Zhan Ma, and Jiaying Liu. Learning end-to-end lossy image compression: A benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(8):4194–4211, 2021.

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 2012.

[15] Eric C Larson and Damon M Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. Journal of electronic imaging, 19(1):011006–011006, 2010.

[16] Jooyoung Lee, Seunghyun Cho, and Seung-Kwon Beack. Context-adaptive entropy model for end-to-end optimized image compression. arXiv preprint arXiv:1809.10452, 2018.

[17] Hanhe Lin, Vlad Hosu, and Dietmar Saupe. Kadid-10k: A large-scale artificially distorted iqa database. In 2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX), pages 1–3. IEEE, 2019.

[18] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023.

[19] Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. Conditional probability models for deep image compression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4394–4402, 2018.

[20] Fabian Mentzer, George D Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. Advances in Neural Information Processing Systems, 33:11913–11924, 2020.

[21] David Minnen, Johannes Ballé, and George D Toderici. Joint autoregressive and hierarchical priors for learned image compression. Advances in neural information processing systems, 31, 2018.

[22] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, et al. Image database tid2013: Peculiarities, results and perspectives. Signal processing: Image communication, 30:57–77, 2015.

[23] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In International conference on machine learning, pages 1060–1069. PMLR, 2016.

[24] Oren Rippel and Lubomir Bourdev. Real-time adaptive image compression. In International Conference on Machine Learning, pages 2922–2930. PMLR, 2017.

[25] H Sheikh. Live image quality assessment database release 2. http://live. ece. utexas. edu/research/quality, 2005.

[26] Xuelin Shen, Zhangkai Ni, Wenhan Yang, Xinfeng Zhang, Shiqi Wang, and Sam Kwong. Just noticeable distortion profile inference: A patch-level structural visibility learning approach. IEEE Transactions on Image Processing, 30:26–38, 2020.

[27] Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. Improving image captioning with better use of captions. arXiv preprint arXiv:2006.11807, 2020.

[28] David S Taubman, Michael W Marcellin, and Majid Rabbani. Jpeg2000: Image compression fundamentals, standards and practice. Journal of Electronic Imaging, 11(2):286–287, 2002.

[29] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy image compression with compressive autoencoders. arXiv preprint arXiv:1703.00395, 2017.

[30] George Toderici, Sean M O'Malley, Sung Jin Hwang, Damien Vincent, David Minnen, Shumeet Baluja, Michele Covell, and Rahul Sukthankar. Variable rate image compression with recurrent neural networks. arXiv preprint arXiv:1511.06085, 2015.

[31] Michael Tschannen, Eirikur Agustsson, and Mario Lucic. Deep generative models for distribution-preserving lossy compression. Advances in neural information processing systems, 31, 2018.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

[33] Gregory K Wallace. The jpeg still picture compression standard. Communications of the ACM, 34(4):30–44, 1991.

[34] Mathias Wien and Benjamin Bross. Versatile video coding–algorithms and specification. In 2020 IEEE International Conference on Visual Communications and Image Processing (VCIP), pages 1–3. IEEE, 2020.

[35] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning, pages 2048–2057. PMLR, 2015.

[36] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1316–1324, 2018.

[37] Ren Yang, Radu Timofte, and LV Gool. Perceptual video compression with recurrent conditional gan. In Messe Wien, 2022.

[38] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In Proceedings of the IEEE international conference on computer vision, pages 5907–5915, 2017.

[39] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 586–595, 2018.

[40] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. arXiv preprint arXiv:1903.10082, 2019.