Multi-Modality Deep Network for JPEG Artifacts Reduction

Xuhao Jiang[1], **Weimin Tan**[1*], **Qing Lin**[1], **Chenxi Ma**[1], **Bo Yan**[1*] and **Liquan Shen**[2]

[1]School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Shanghai Collaborative Innovation Center of Intelligent Visual Computing, Fudan University, Shanghai, China
[2]School of Communication, Shanghai University, Shanghai, China
{20110240011, wmtan, 18210240028, 17210240039, byan}@fudan.edu.cn, jsslq@163.com *

Abstract

In recent years, many convolutional neural network-based models are designed for JPEG artifacts reduction, and have achieved notable progress. However, few methods are suitable for extreme low-bitrate image compression artifacts reduction. The main challenge is that the highly compressed image loses too much information, resulting in reconstructing high-quality image difficultly. To address this issue, we propose a multimodal fusion learning method for text-guided JPEG artifacts reduction, in which the corresponding text description not only provides the potential prior information of the highly compressed image, but also serves as supplementary information to assist in image deblocking. We fuse image features and text semantic features from the global and local perspectives respectively, and design a contrastive loss built upon contrastive learning to produce visually pleasing results. Extensive experiments, including a user study, prove that our method can obtain better deblocking results compared to the SOTA methods.

1 Introduction

Lossy image compression algorithms are widely used in image storage and transmission. However, due to the loss of information, complex compression noise is inevitably introduced into the compressed image, such as blocking artifacts [Dong *et al.*, 2015], resulting in degradation in both the visual quality of the compressed image and the performance of the subsequent computer vision tasks. Therefore, exploring methods for compressed image artifacts reduction is urgently needed, especially for the widely used JPEG format.

To cope with JPEG compression artifacts, many methods [Zhang *et al.*, 2017; Kim *et al.*, 2019; Jiang *et al.*, 2021] have been proposed. However, in some occasions with limited bandwidth, the images are usually highly compressed for transmission, and the previous algorithms fail to effectively enhance the compressed image, as shown in Fig. 1. The main reason is that the highly compressed images lose too much information leading to the difficulty in image restoration. Faced with this extreme situation, JPEG artifacts reduction based on multimodal machine learning may have great advantages. The corresponding text provides the high-level image semantic information, which can be used as prior information to assist in image deblocking. Specifically, the text describes the main object of the image and some of its details, such as shape, color, location, etc. Under the guidance of the prior information, the multi-modality deep model can effectively remove the compression artifacts and reconstruct better deblocking results, as shown in Fig. 1.

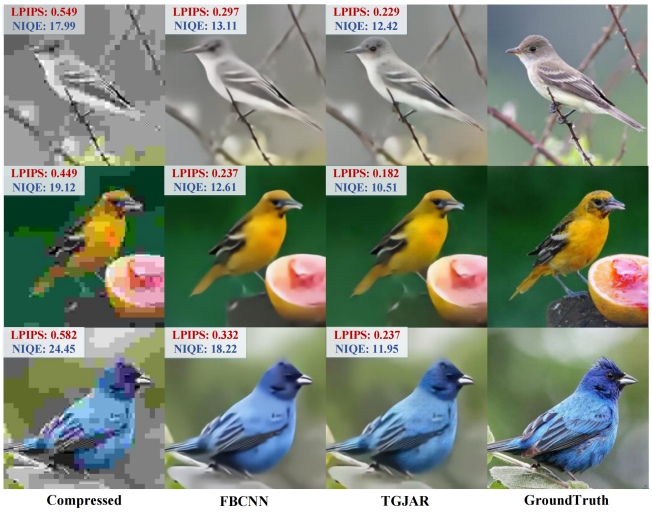

Figure 1: Visual comparisons of the proposed TGJAR and the state-of-the-art (SOTA) method FBCNN at quality factor 1.

In this paper, a text-guided JPEG artifacts reduction (TGJAR) generative adversarial network is proposed, in which text features including the word and sentence features are used to assist in image deblocking. This is an interesting attempt for multimodal machine learning. Considering that word and sentence features represent the local and global information of the text respectively, TGJAR proposes two kinds of image-text fusion modules to fuse image and text features from the global and local perspectives. For better local feature fusion, we adopt a multi-scale design to perform fusion

*Corresponding authors: Weimin Tan and Bo Yan. This work is supported by NSFC (Grant No.: U2001209, 61902076) and Natural Science Foundation of Shanghai (21ZR1406600).

on three different scales. Inspired by the contrast learning [He *et al.*, 2020], a contrastive loss is proposed to restrict the restored image from being pulled closer to the uncompressed image and away from the compressed image in the perceptual quality space. The main contributions are as follows:

- We build a text-guided JPEG artifacts reduction generative adversarial network, and design a multimodal fusion method to fuse compressed image features and the corresponding text features. To the best of our knowledge, we are the first to apply multimodal machine learning to image deblocking, and demonstrate the effectiveness of the text description guidance for JPEG artifacts reduction, especially for highly compressed images.
- Two kinds of fusion modules are employed to better fuse the features of text and images. The image-text global feature fusion can remove the global artifacts of the compressed image, and the local fusion modules employ an attention mechanism to obtain attention regions of word features for providing prior information.
- A well-designed contrastive loss is built upon contrastive learning to produce more realistic images.
- The experiments (including user study) show the outstanding perceptual performance of our TGJAR in comparison with the existing image deblocking methods.

2 Related Work

JPEG Artifacts Reduction. Recently, notable progress [Fu *et al.*, 2019; Jin *et al.*, 2020; Liang *et al.*, 2021; Zhang *et al.*, 2021b; Chen *et al.*, 2021; Fu *et al.*, 2021a; Fu *et al.*, 2021b; Zheng *et al.*, 2019; Kim *et al.*, 2020; Zini *et al.*, 2020; Li *et al.*, 2020a; Wang *et al.*, 2022b; Jiang *et al.*, 2022; Wang *et al.*, 2021] has been made for JPEG artifacts reduction by utilizing deep convolutional neural networks. Dong *et al.* [Dong *et al.*, 2015] first propose the famous ARCNN to solve this problem, which is a relatively shallow network. RNAN [Zhang *et al.*, 2019] designs local and non-local attention learning to further enhance the representation ability, and obtains good results in image restoration tasks, including image denoising, compression artifacts reduction, and image super-resolution. Li *et al.* propose the QGCN [Li *et al.*, 2020a] to handle a wide range of quality factors while it can consistently deliver superior image artifacts reduction performance. Jiang *et al.* [Jiang *et al.*, 2021] propose a flexible blind CNN, namely FBCNN, which can predict the adjustable quality factor to control the trade-off between artifacts removal and details preservation. Witnessing the recent success of GAN in most image restoration tasks, some GAN-based JPEG artifacts reduction works [Galteri *et al.*, 2017; Galteri *et al.*, 2019] have been proposed, aiming to improve the subjective quality of compressed images. However, these methods show poor performance on recovering highly compressed image due to the serious loss of information. With the development of multimodal fusion learning technology, we can use the information of other modalities to assist in image deblocking.

Multimodal Machine Learning. The multimodal machine learning is a new and interesting topic, which simulates the cognitive process of humans by using information from multiple modalities. Some previous works [Xu *et al.*, 2018; Mittal *et al.*, 2020; Jiang *et al.*, 2023] have demonstrated the powerful advantages of multimodal machine learning in the field of computer vision. For example, AttnGAN [Xu *et al.*, 2018] employs attention mechanism on the descriptive text to produce images with fine-grained details. Therefore, benefiting from the semantic information and the prior information provided by the text description, JPEG artifacts reduction based on multimodal machine learning may have a greater possibility to obtain better deblocking results.

Contrastive Learning. Recently, contrastive learning has demonstrated its effectiveness in self-supervised representation learning [He *et al.*, 2020; Chen *et al.*, 2020]. The goal of the contrastive learning is to pull the target point toward the positive sample point and push the target point away from the negative sample point in a feature representation space. The contrastive learning enhances the contrast between positive and negative samples, which is beneficial for high-level vision tasks. For image pixel-level restoration, it is difficult to find a suitable feature space to construct positive and negative samples. In TGJAR, we propose to construct positive and negative samples in the image quality space, and design a novel contrastive loss for perceptual quality improvement.

3 The Proposed TGJAR

The process of compression algorithms can be expressed as

$$I^c = F_c(I, QF)\ , \tag{1}$$

where I^c is the compressed image, I is the original uncompressed image, F_c is the compression algorithm, and QF represents the quality factor determining the degree of compression. The goal of JPEG artifacts reduction is to reconstruct a deblocking image I^d from a compressed image I^c, aiming to keep I^d and I consistent in pixels.

When the image is highly compressed, the image information is severely lost, so it is difficult to recover a high-quality image through empirical modeling of the generator. In this scenario, we consider introducing other modal information (*e.g.* text information) to assist highly compressed image deblocking. Our main goal is to remove JPEG artifacts in a highly compressed image with the assistance of a corresponding text description. The corresponding parameter optimization can be defined as below,

$$\hat{\theta}_g = arg\ \min_{\theta} \frac{1}{N} \sum_{n=1}^{N} L(G(I^c, T; \theta_g), I)\ , \tag{2}$$

where G represents the generator, performing image artifacts reduction function, L represents the loss function, N represents the number of images in training dataset, and T represents the text description. Based on this consideration, we propose a text-guided JPEG artifacts reduction generative adversarial network (TGJAR).

3.1 Overview of Our TGJAR

The architecture design of the proposed TGJAR is shown in Fig. 2, which consists of five components: generator, discriminator, text encoder, image encoder and perceptual quality

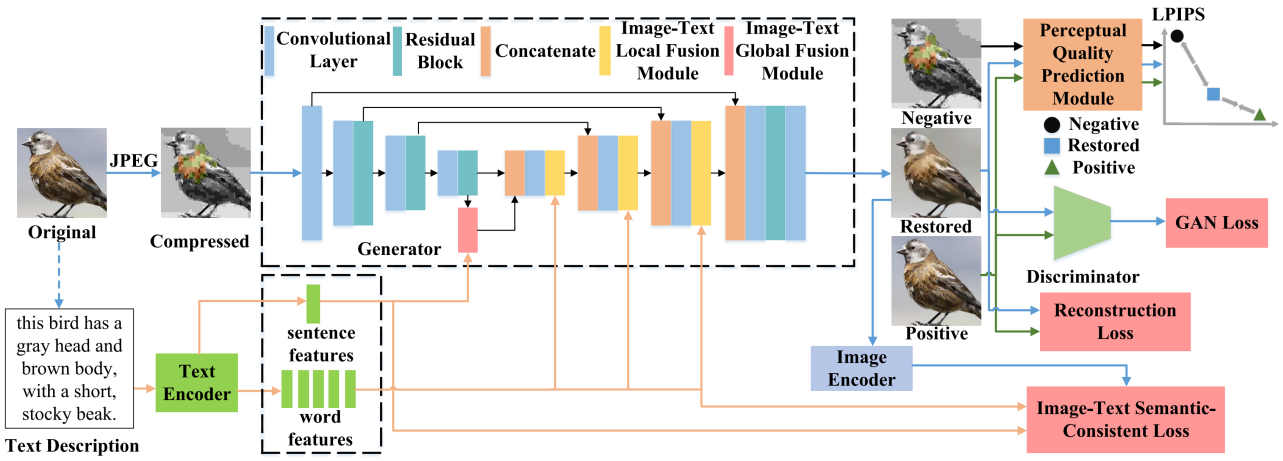

Figure 2: The architecture of TGJAR. In our network, both the compressed image and the corresponding text description are taken as input. Specially, the text information is used as auxiliary information to reconstruct high perceptual quality images.

prediction module. Firstly, the text features, including word and sentence features, are extracted from the text description by using the text encoder. Then the sentence and word features are input into the generator to assist in image deblocking. Specially, we design the image-text local fusion module (LFM) and image-text global fusion module (GFM) in the generator to better fuse the compressed image features and text features. With the aid of text information, the generator can recover deblocking results with high perceptual quality. Specifically, the architecture of the generator is based on the U-Net [Ronneberger *et al.*, 2015] structure, and equipped with some residual blocks [He *et al.*, 2016] in order to have a stronger deblocking ability. We also introduce the contrastive loss and the image-text semantic-consistent loss to further improve the perceptual quality of the deblocking results.

3.2 Image-Text Fusion Module

In the TGJAR, the text encoder is a bi-direction Long Short-Term Memory (LSTM) [Schuster and Paliwal, 1997], which extracts the semantic features from the text description. Then we can obtain the word and sentence features respectively. Note that the word features represent the local features of a text description, while the sentence features denote the global features. In this way, the semantic information of text description can be mapped into a feature space consistent with image semantic information. The function of the text encoder can be defined as

$$w, s = F_t(T) \,, \tag{3}$$

where w and s represent the word and sentence features respectively, F_t represents the text encoder. Considering that the global and local features of the text can provide global and local prior information for the image respectively, two kinds of image-text fusion module including GFM and LFM, are designed to fuse the text and image features from both global and local perspectives. The architectures of the two modules are shown in Fig. 3.

Image-Text Global Fusion Module (GFM). The architecture of GFM is shown in the Fig. 3 (a). We directly use the

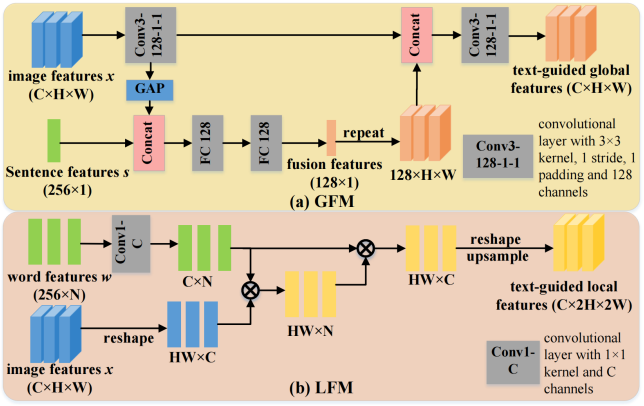

Figure 3: The architecture of the proposed image-text global fusion module (a) and image-text local fusion module (b).

features output by the third residual block in the generator to obtain the global image features. Given the $128 \times 32 \times 32$ feature map x, for global features, the feature map is further reduced to $128 \times 1 \times 1$ by employing a Conv3-128 and a global average pooling layer (GAP). The extracted image global features are concatenated with the sentence features s, and then two fully-connected (FC) layers are used to further extract the fused global features. The extracted $128 \times 1 \times 1$ fused global features are enlarged to the size of $128 \times 32 \times 32$ by repeating, and then concatenated with the input features. Finally, the concatenated features are processed by a Conv3-128 to get the text-guided global features.

Inspired by DPE [Chen *et al.*, 2018], GFM provides a way to fuse the global image features and the global text features, which utilizes the text-guided global features for image deblocking. With the aid of the text-guided global features, the image deblocking results achieve significant improvement on the overall image perceptual quality.

Image-Text Local Fusion Module (LFM). In TGJAR, we use three LFMs to adaptively fuse the local image features

and the word features, which is based on a multi-scale strategy. The main goal of the multi-scale design is to improve the effect of word features in image deblocking task. The architecture of the LFM is shown in Fig. 3 (b). In the LFM, one 1×1 convolutional layer is used to adjust the size of the word features w, so as to calculate the correlation between the image features x and the word features, and get the corresponding attention map. The input word features are weighted by the attention map, and then are reshaped and upsampled to obtain the text-guided local features.

LFM is designed to better fuse the image local features and the word features, which employs the attention mechanism to extract the corresponding attention regions of each word in compressed image. Then the attention regions in the compressed image can provide prior information for image deblocking. Thanks to the text-guided local features, the model removes the blocking artifacts and generates photorealistic details in the local region of the images, which makes the images look more real.

3.3 Contrastive Learning

Inspired by [He *et al.*, 2020; Chen *et al.*, 2020], we innovatively design a contrastive loss built upon contrastive learning to produce better results. The main difficulty lies in how to define the positive and negative samples and the corresponding feature constraint space. In TGJAR, we use compressed images as negative samples and uncompressed images as positive samples, and adopt the image perceptual quality space as the feature constraint space. The proposed contrastive loss constrains that the perceptual quality of the restored image is pulled to closer to that of the uncompressed image and pushed to far away from that of the compressed image. To the best of our knowledge, we are the first to propose a contrastive learning design constructed on the image quality assessment model. The details are as follows.

Here, we take the uncompressed image I_i as the positive sample and the compressed image I_i^c as the negative sample. The perceptual quality prediction module needs to be differentiable, so it can be any convolutional neural network-based model. We adopt the LPIPS [Zhang *et al.*, 2018] model in TGJAR, since LPIPS is highly consistent with human subjective evaluation. The image quality assessment (IQA) can be defined as $q_i = F_{lpips}(I_i^d, I_i)$, where q_i represents the quality of the restored image I_i^d with I_i as a reference, and F_{lpips} represents the function of LPIPS. Note that LPIPS is a reference-based IQA model. The predicted quality score is greater than or equal to zero, and lower score means that the perceptual quality of the predicted image is close to that of the reference image. Naturally, we can easily define the contrastive loss as

$$L_C = ||\frac{F_{lpips}(I_i^d, I_i) - F_{lpips}(I_i, I_i)}{F_{lpips}(I_i^d, I_i) - F_{lpips}(I_i^c, I_i)}||_1, \tag{4}$$

We can easily find that the $F_{lpips}(I_i, I_i)$ is equal to zero. Unfortunately, some related experiments prove that the above contrastive loss is not easy to converge, and even has a counterproductive effect. Thus, the L_C is further simplified as

$$L_C = \frac{F_{lpips}(I_i^d, I_i)}{F_{lpips}(I_i^d, I_i^c) + c}, \tag{5}$$

where c is a constant. This simplified contrastive loss L_C means that, taking I_i as a reference, the perceptual quality value of I_i^d should be small, and taking I_i^c as a reference, the perceptual quality value of I_i^d should be large. Its effectiveness has been verified in the experimental part.

3.4 Loss Functions

We use four different loss functions to optimize TGJAR, including contrastive loss L_C, reconstruction loss L_R, GAN loss L_G and image-text semantic-consistent loss L_{IT}. The overall loss function is defined as

$$L = \lambda_1 L_C + \lambda_2 L_R + \lambda_3 L_G + \lambda_4 L_{IT} \tag{6}$$

where the λ_1, λ_2, λ_3, and λ_4 are the weights to balance different losses. L_R and L_G are commonly used in GAN-based models, and defined as,

$$L_R = ||I_i^d - I_i||_1, \tag{7}$$

$$L_G = E_{I_i \sim p_{data}(I)}[logD(I_i)] + E_{I_i^d \sim I^c}[log(1 - D(I_i^d))] \tag{8}$$

where the $D(.)$ represents the mapping funtion of the discriminator.

The image-text semantic-consistent loss is introduced to make the deblocking results semantic consistent with the text description. AttnGAN [Xu *et al.*, 2018] proposes a deep attentional multimodal similarity model to calculate the similarity of the text features and the features of the generated image. Following AttnGAN, an image encoder and a text encoder are employed to map the image features and text features into a common semantic space. The image encoder is based on the AttnGAN, which consists of two parts including a Inception-v3 [Szegedy *et al.*, 2016] model and a mapping layer. Then the image-text semantic-consistent loss is defined as

$$L_{IT} = L_{word} + L_{sentence} \tag{9}$$

where $L_{sentence}$ is the negative log posterior probability between I_i^d and the corresponding sentence, and defined as $-(logP(s_i|I_i^d) + logP(I_i^d|s_i))$. Note that w_i and s_i are the word and sentence features of the corresponding text. Similarly, L_{word} is the negative log posterior probability between the local region of I_i^d and the corresponding words.

3.5 Training Implementation

In the proposed TGJAR, training is divided into two stages. In the first stage, following AttnGAN [Xu *et al.*, 2018], the image encoder and text encoder are pretrained, aiming to map the image features and text features into a common semantic space. In the second stage, we fix the weights of the image encoder, text encoder and the perceptual quality prediction module, and train the discriminator and generator. Note that the second stage is our main contribution. In the second training stage, Pytorch is used as the training toolbox, and the Adam optimization algorithm [Kingma and Ba, 2014] with a mini-batch of 4 is adopted for training. All the experiments are conducted on a NVIDIA GeForce RTX 1080 Ti. The learning rate is changed from 1×10^{-4} to 1×10^{-8} at the interval of twenty epochs. The hyper-parameter c of the L_C is set as 0.1, and the hyper-parameters λ_1, λ_2, λ_3, and λ_4 of the global loss function are empirically set as 0.01, 1, 0.001 and 0.0005, respectively.

Dataset	QF	JPEG	EDSR	RNAN	QGCN	FBCNN	TGJAR
CUB	1	0.505/202.5/11.45/19.43	0.349/92.8/8.17/10.36	0.356/93.4/7.89/9.74	0.341/90.8/8.54/11.00	0.332/74.2/8.56/11.06	**0.249/17.9/6.95/8.62**
	5	0.375/116.3/ 8.87 /14.39	0.248/59.6/7.19/ 9.34	0.253/54.3/6.99/8.81	0.239/53.9/7.56/ 9.91	0.237/47.5/7.58/ 9.84	**0.154/10.7/6.12/8.08**
	10	0.228/ 42.9 / 6.72 /10.32	0.156/34.3/6.45/ 8.71	0.161/36.9/6.08/8.14	0.154/33.0/6.72/ 9.09	0.155/31.2/6.65/ 8.90	**0.089/ 6.9 /5.60/7.69**
Oxford -102	1	0.493/222.3/10.68/18.64	0.309/74.1/7.91/9.70	0.315/75.7/7.62/9.14	0.294/72.1/7.97/10.04	0.290/66.6/7.94/9.98	**0.224/42.0/6.34/7.83**
	5	0.342/126.6/ 8.18 /13.29	0.201/60.2/6.90/8.49	0.206/59.3/6.76/8.14	0.190/58.5/6.96/ 8.77	0.190/53.8/6.97/8.70	**0.142/32.2/5.70/7.36**
	10	0.185/ 60.3 / 6.21 / 9.37	0.125/48.1/6.27/7.85	0.128/47.6/6.15/7.68	0.116/46.5/6.28/ 8.08	0.120/42.4/6.35/8.08	**0.080/25.4/5.53/7.11**

Table 1: Average LPIPS|FID|PI|NIQE values of various methods based on the color images from CUB and Oxford-102 datasets for QF = 1, 5 and 10. Lower is better. The best results are boldfaced.

Dataset	QF	JPEG	EDSR	RNAN	QGCN	FBCNN	TGJAR
CUB	1	21.8	24.4	24.3	24.6	24.7	24.4
	5	24.3	26.9	26.7	27.2	27.2	26.8
	10	27.4	29.8	29.6	30.0	29.9	29.7
Oxford -102	1	21.3	24.0	23.9	24.2	24.2	24.0
	5	23.7	26.6	26.4	26.9	26.9	26.5
	10	26.8	29.6	29.4	30.0	29.8	29.5

Table 2: Average PSNR values of various methods based on the color images from CUB and Oxford-102 datasets for QF = 1, 5 and 10. Higher is better.

Figure 4: Visual comparisons with SOTA methods on CUB and Oxford-102 datasets. Above each line of the images is the corresponding text description. Better zoom in.

4 Experimental Results

4.1 Datasets and Evaluation Methodology

We evaluate our TGJAR on the CUB [Wah *et al.*, 2011] and Oxford-102 [Nilsback and Zisserman, 2008] datasets, in which all images are annotated with corresponding text descriptions. CUB dataset contains 200 species of bird with a total of 11,788 images, of which 8,855 images are used for training and 2,933 images are used for testing. Oxford-102 Dataset consists of 102 flower categories, with a total of 8,189 images including 7,034 images for training and 1,155 images for testing. We preprocess the two datasets according to the methods in AttnGAN [Xu *et al.*, 2018], then crop and resize the images into patches of size 256×256.

Considering TGJAR aims at improving the perceptual quality of the highly compressed image, we adopt small QFs (i.e., 1, 5 and 10) of JPEG compression algorithm to process the training and testing datasets. Following [Blau *et al.*, 2018; Mentzer *et al.*, 2020], we evaluate the proposed TGJAR and the compared methods in four perceptual quality metrics, including LPIPS [Zhang *et al.*, 2018], FID [Heusel *et al.*, 2017], PI [Blau *et al.*, 2018] and NIQE [Mittal *et al.*, 2012], which are highly consistent with human perception of images. Besides, we use Peak Signal-to-Noise Ratio (PSNR) to measure the fidelity of the reconstructions.

4.2 Comparison with The SOTA Methods

In this part, TGJAR and the SOTA algorithms including EDSR [Lim *et al.*, 2017], RNAN [Zhang *et al.*, 2019], QGCN [Li *et al.*, 2020a] and FBCNN [Jiang *et al.*, 2021] are compared quantitatively and qualitatively. To conduct a fair comparison, EDSR and RNAN are retrained on these two datasets including CUB and Oxford-102. In particular, we remove the upsampling module and set the number of the residual block to 16 in EDSR, and use the RGB compressed images to train RNAN. For QGCN and FBCNN, they are finetuned on the training datasets since the pretrained models are available. Since the training codes for the existing GAN-based models are not available, TGJAR is not compared with them. Certainly, we compare the proposed TGJAR with our baseline model (i.e., a GAN-based model) in the ablation experiments.

Quantitative Comparisons. Tables 1 and 2 show the quantitative results on two datasets with JPEG QF 1, 5 and 10, respectively. In Table 1, the proposed TGJAR achieves the best performance on the four perceptual quality indexes (i.e. LPIPS, FID, PI and NIQE) at all JPEG QFs. Specially, we notice that our deblocking results of FID on two datasets at QF 1 are better than that of other methods on two datasets at QF 10. As shown in Table 2, it can be found that the deblocking results of these methods all achieve significant PSNR gain compared with the compressed images. Specifically, the proposed TGJAR achieves a competitive performance on PSNR compared to other four methods. EDSR, RNAN, QGCN and FBCNN are all MSE-based JPEG artifacts reduction meth-

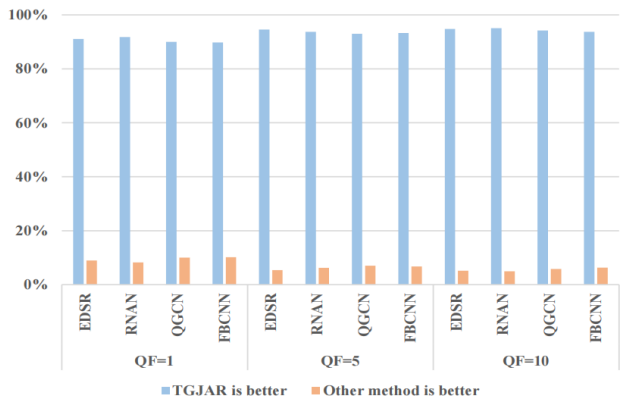

Figure 5: User study results. The reported value indicates the performance rate of the proposed TGJAR against the other methods at QF=1, 5 and 10, respectively.

ods, which usually produce overly smooth deblocking results. This results in good performance on PSNR, but poor performances on the other four perceptual quality indexes which are more consistent with human perception. Compared with them, our proposed TGJAR obtains competitive PSNR, and achieves significant improvement on LPIPS, FID, PI and NIQE. The main reason is that TGJAR uses text description as auxiliary information to improve the perceptual qualities of deblocking results, where text not only provides prior information of the compressed image, but also supplements semantic information losses. From Tables 1 and 2, we can find that TGJAR takes into account the faithfulness of the reconstruction to the uncompressed image, and greatly improves the perceptual quality of the deblocking results compared with existing algorithms.

Qualitative Comparisons. The proposed TGJAR not only outperforms the comparative methods in terms of overall quantitative evaluation, but also produces deblocking results containing high subjective quality. As shown in Fig. 4, four deblcocking methods all remove the JPEG artifacts to a large extent. However, the deblocking results of EDSR, RNAN, QGCN and FBCNN seem blurry, especially at QF=1. The main reason is that under a high compression rate, the image information is seriously lost resulting in reconstructing high-quality deblocking results difficultly. Unlike these four methods, TGJAR makes full use of text information, and generates appealing visual results containing clear edges and rich textures even in the case of QF=1.

4.3 User Study

We further conduct a user study with 13 subjects. Given a pair of images produced by two methods, the users are asked to judge which one owns higher perceptual quality and is more coherent with the given texts. We randomly select 50 images in each dataset at each QF value, and this user study requires 15,600 comparisons in total. The results are shown in Fig. 5. We can find that our results win more than 90%. This subjective comparisons are consistent with the quantitative comparisons in Table 1, which demonstrate that TGJAR outperforms other methods.

Attributes					Quality Index	
L_R	L_G	GFM $+L_{IT}$	LFM $+L_{IT}$	L_C	PSNR↑	LPIPS/FID/PI/NIQE↓
✓	✓				**24.4**	0.342/79.7/8.57/11.04
✓	✓	✓			**24.4**	0.314/28.4/8.43/10.93
✓	✓		✓		24.2	0.307/36.7/8.44/11.05
✓	✓	✓	✓		24.3	0.308/33.3/8.36/10.86
✓	✓	✓	✓	✓	**24.4**	**0.249/17.9/6.95/ 8.62**

Table 3: Performance comparisons between variations of our TGJAR on CUB at QF=1. The best results are boldfaced.

4.4 Ablation Study

We conduct the ablation study to verify the effectiveness of the proposed contrastive loss and two image-text fusion modules including GFM and LFM. Note that TGJAR only optimized by L_R and L_G is regarded as the baseline model. Considering that the L_{IT} can constrain deblocking results more consistent with the text description, we employ the L_{IT} when using the image-text fusion modules. This allows the TGJAR to make better use of text information.

The quantitative comparisons at QF 1 are shown in Table 3. It can be found that all variations of TGJAR obtain similar performance on PSNR index. This demonstrates that TGJAR still faithfully reconstructs the deblocking results to the uncompressed images after introducing text information. On the other four indicators, the final model achieves the best performance. Among them, the model with GFM or LFM shows similar performance. Compared with these two models, the model with both GFM and LFM shows intermediate performance on LPIPS and FID, and better performance on PI and NIQE, since LFM and GFM can complement each other. Finally, the contrastive loss improves the performance of the model on four indicators by simultaneously using positive and negative samples.

The qualitative comparisons of ablation study are shown in Fig. 6. By comparing the deblocking results, the final TGJAR produces the deblocking results with the highest subjective quality. In addition, we find that only applying the GAN design, the deblocking results are still a little blurry. By separately introducing LFM and GFM, the model can effectively remove the blocking artifacts, and greatly improve the subjective quality of the image. However, the deblocking results are still relatively blurry by only applying GFM, and some unnatural textures exist in the deblocking results only adopting LFM. Thus, using the two modules simultaneously can give full play to the advantages of both modules, and restrict the disadvantages of each module. Unfortunately, we find that there are still some unnatural textures exists in the deblocking results using both two modules. Recognizing this problem, we introduce contrast learning to obtain more real deblocking results with photo-realistic details.

4.5 Text-guided Explorative Image Deblocking

In order to explore more possibilities of our TGJAR, we also conduct experiments on COCO [Lin *et al.*, 2014] dataset with 80 types of objects at QF 1. To further verify the generalization performance of the algorithm, we also conduct the cross-dataset experiments on Kodak [Kodak, 1993]. Since

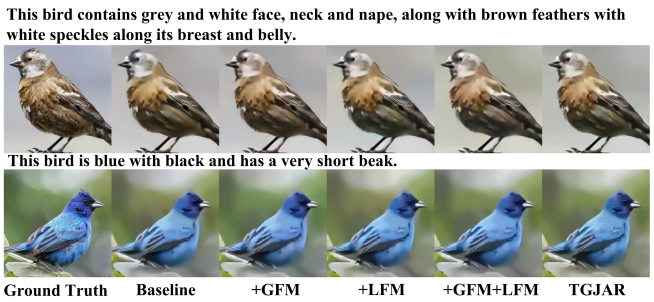

Figure 6: Ablation study on CUB dataset. Better zoom in.

Dataset	Method	LPIPS/FID/PI/NIQE
COCO	EDSR	0.458/165.3/7.63/9.19
	TGJAR	**0.269/ 70.6 /5.91/7.68**
Kodak	EDSR	0.561/255.3/6.17/7.05
	TGJAR	**0.348/120.5/4.43/5.51**

Table 4: Performance comparisons of our TGJAR and EDSR on COCO and Kodak datasets at QF = 1. 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.

Kodak does not have corresponding text descriptions, we consider using the image captions methods [Wang *et al.*, 2022a; Li *et al.*, 2020b; Zhang *et al.*, 2021a] to generate texts for our experiments. Here, we use OFA [Wang *et al.*, 2022a] to produce the corresponding texts as an example. The generated texts are of high quality and highly semantically consistent with the images. We compare our TGJAR with EDSR, since both use the residual block as the main feature extraction module. The quantitative and qualitative experimental results are shown in Table 4 and Fig. 7. We can find that TGJAR can produce much better results compared with EDSR, which confirms the great potential of our TGJAR at extreme low-birate image compression artifacts reduction.

4.6 Text-Guided Controllable Image Deblocking

An experiment is further conducted to explore the effect of text guidance, that is, using different texts to assist in image deblocking. The visual results are shown in Fig. 8. Here, we modify the most obvious information, i.e., the color information. It can be found that the modification of the text descriptions does have an impact on the results, and the results are consistent with the semantics of the text descriptions. The previous single-modal methods can only enhance the image based on experience, but the multimodal algorithm can be personalized to enhance the image.

4.7 Discussion

Text-guided JPEG artifacts reduction needs to introduce a new modality, that is, the corresponding text description of the image. In this regard, we believe that there are three ways to obtain the texts. Firstly, many images on social media (e.g. Twitter, Facebook) are tagged with text descriptions consisted of rich attributes, which can be used for multimodal deblocking. Secondly, users can provide reasonable descriptions to enhance the image according to their imagination. As a result, the image will have a personalized enhancement effect. Thirdly, image caption algorithms can be used to obtain the text description corresponding to the image before the image is compressed, since original images are accessible during image compression. Then we can transmit texts as labels with the compressed images to the decoder side, and use these texts to enhance the compressed images. Note that the text occupies very few bits and can be transmitted to the decoder side at marginal bandwidth cost. Even for extreme low bitrate, i.e. JPEG quality factor set to 1, texts use less than one twenty-fifth of the bits used by images on datasets [Wah *et al.*, 2011; Nilsback and Zisserman, 2008].

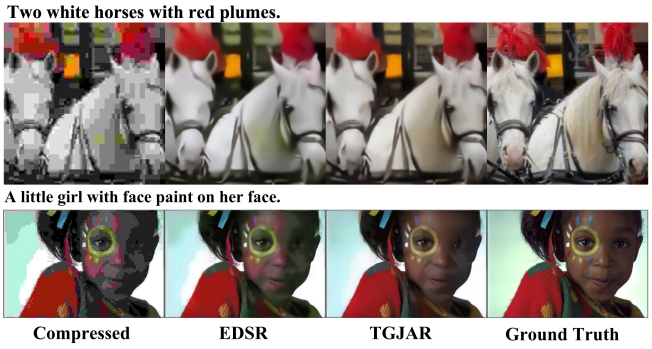

Figure 7: Visual comparisons on COCO (first row) and Kodak (second row) at QF 1. Above each line of the images is the corresponding text description. Better zoom in.

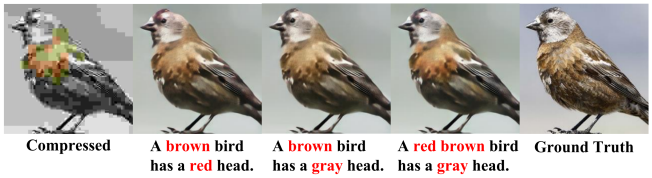

Figure 8: Our results generated by using different text descriptions at QF 1. Better zoom in.

5 Conclusion

In this paper, we propose a text-guided generative adversarial network for JPEG artifacts reduction (TGJAR). To better fuse image and text information, we design two fusion modules, including the image-text global fusion module and the image-text local fusion module. These two modules fuse the global and local features of the image and text information from the global and local perspectives, respectively. Besides, to further improve the subjective quality of the deblocking results, a well-designed contrastive loss is built upon contrastive learning to constrain that the restored image is pulled to closer to the uncompressed image and pushed to far away from the compressed image in perceptual quality space. Experimental results demonstrate that TGJAR outperforms the test SOTA methods. Especially, even when QF=1, the TGJAR can produce visually pleasing results.

References

[Blau *et al.*, 2018] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[Chen *et al.*, 2018] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6306–6314, 2018.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[Chen *et al.*, 2021] Zhengxin Chen, Xiaohai He, Chao Ren, Honggang Chen, and Tingrong Zhang. Enhanced separable convolution network for lightweight jpeg compression artifacts reduction. *IEEE Signal Processing Letters*, 28:1280–1284, 2021.

[Dong *et al.*, 2015] Chao Dong, Yubin Deng, Chen Change Loy, and Xiaoou Tang. Compression artifacts reduction by a deep convolutional network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 576–584, 2015.

[Fu *et al.*, 2019] Xueyang Fu, Zheng-Jun Zha, Feng Wu, Xinghao Ding, and John Paisley. Jpeg artifacts reduction via deep convolutional sparse coding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2501–2510, 2019.

[Fu *et al.*, 2021a] Xueyang Fu, Menglu Wang, Xiangyong Cao, Xinghao Ding, and Zheng-Jun Zha. A model-driven deep unfolding method for jpeg artifacts removal. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[Fu *et al.*, 2021b] Xueyang Fu, Xi Wang, Aiping Liu, Junwei Han, and Zheng-Jun Zha. Learning dual priors for jpeg compression artifacts removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4086–4095, 2021.

[Galteri *et al.*, 2017] Leonardo Galteri, Lorenzo Seidenari, Marco Bertini, and Alberto Del Bimbo. Deep generative adversarial compression artifact removal. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4826–4835, 2017.

[Galteri *et al.*, 2019] Leonardo Galteri, Lorenzo Seidenari, Marco Bertini, and Alberto Del Bimbo. Deep universal generative adversarial compression artifact removal. *IEEE Transactions on Multimedia*, 21(8):2131–2145, 2019.

[He *et al.*, 2016] 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.

[He *et al.*, 2020] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[Jiang *et al.*, 2021] Jiaxi Jiang, Kai Zhang, and Radu Timofte. Towards flexible blind jpeg artifacts removal. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4997–5006, 2021.

[Jiang *et al.*, 2022] Xuhao Jiang, Weimin Tan, Ri Cheng, Shili Zhou, and Bo Yan. Learning parallax transformer network for stereo image jpeg artifacts removal. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6072–6082, 2022.

[Jiang *et al.*, 2023] Xuhao Jiang, Weimin Tan, Tian Tan, Bo Yan, and Liquan Shen. Multi-modality deep network for extreme learned image compression. *arXiv preprint arXiv:2304.13583*, 2023.

[Jin *et al.*, 2020] Zhi Jin, Muhammad Zafar Iqbal, Wenbin Zou, Xia Li, and Eckehard Steinbach. Dual-stream multi-path recursive residual network for jpeg image compression artifacts reduction. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(2):467–479, 2020.

[Kim *et al.*, 2019] Yoonsik Kim, Jae Woong Soh, Jaewoo Park, Byeongyong Ahn, Hyun-Seung Lee, Young-Su Moon, and Nam Ik Cho. A pseudo-blind convolutional neural network for the reduction of compression artifacts. *IEEE Transactions on circuits and systems for video technology*, 30(4):1121–1135, 2019.

[Kim *et al.*, 2020] Yoonsik Kim, Jae Woong Soh, and Nam Ik Cho. Agarnet: adaptively gated jpeg compression artifacts removal network for a wide range quality factor. *IEEE Access*, 8:20160–20170, 2020.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Kodak, 1993] Eastman Kodak. Kodak lossless true color image suite (photocd pcd0992). *URL http://r0k.us/graphics/kodak*, 1993.

[Li *et al.*, 2020a] Jianwei Li, Yongtao Wang, Haihua Xie, and Kai-Kuang Ma. Learning a single model with a wide range of quality factors for jpeg image artifacts removal. *IEEE Transactions on Image Processing*, 29:8842–8854, 2020.

[Li *et al.*, 2020b] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[Liang *et al.*, 2021] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2021.

[Lim *et al.*, 2017] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

[Lin *et al.*, 2014] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[Mentzer *et al.*, 2020] Fabian Mentzer, George Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. *arXiv preprint arXiv:2006.09965*, 2020.

[Mittal *et al.*, 2012] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3):209–212, 2012.

[Mittal *et al.*, 2020] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emoticon: Context-aware multimodal emotion recognition using frege's principle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14234–14243, 2020.

[Nilsback and Zisserman, 2008] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[Schuster and Paliwal, 1997] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681, 1997.

[Szegedy *et al.*, 2016] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *Technical Report CNS-TR-2011-001, California Institute of Technology*, 2011.

[Wang *et al.*, 2021] Menglu Wang, Xueyang Fu, Zepei Sun, and Zheng-Jun Zha. Jpeg artifacts removal via compression quality ranker-guided networks. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 566–572, 2021.

[Wang *et al.*, 2022a] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. *CoRR*, abs/2202.03052, 2022.

[Wang *et al.*, 2022b] Xi Wang, Xueyang Fu, Yurui Zhu, and Zheng-Jun Zha. Jpeg artifacts removal via contrastive representation learning. In *European Conference on Computer Vision*, pages 615–631. Springer, 2022.

[Xu *et al.*, 2018] 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.

[Zhang *et al.*, 2017] Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE transactions on image processing*, 26(7):3142–3155, 2017.

[Zhang *et al.*, 2018] 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.

[Zhang *et al.*, 2019] Yulun Zhang, Kunpeng Li, Kai Li, Bineng Zhong, and Yun Fu. Residual non-local attention networks for image restoration. In *International Conference on Learning Representations*, 2019.

[Zhang *et al.*, 2021a] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2021.

[Zhang *et al.*, 2021b] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(7):2480–2495, 2021.

[Zheng *et al.*, 2019] Bolun Zheng, Yaowu Chen, Xiang Tian, Fan Zhou, and Xuesong Liu. Implicit dual-domain convolutional network for robust color image compression artifact reduction. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11):3982–3994, 2019.

[Zini *et al.*, 2020] Simone Zini, Simone Bianco, and Raimondo Schettini. Deep residual autoencoder for blind universal jpeg restoration. *IEEE Access*, 8:63283–63294, 2020.