ENHANCED IMAGE DECODING VIA EDGE-PRESERVING GENERATIVE ADVERSARIAL NETWORKS

Qi Mao[1]*, Shiqi Wang*[2]*, Shanshe Wang*[1]*, Xinfeng Zhang*[3]*, Siwei Ma*[1]

[1] Institute of Digital Media, Peking University
[2] Department of Computer Science, City University of Hong Kong
[3] Ming Hsieh Department of Electrical Engineering, University of Southern California
{qimao, sswang, swma}@pku.edu.cn, shiqwang@cityu.edu.hk, xinfengz@usc.edu

ABSTRACT

Lossy image compression usually introduces undesired compression artifacts, such as blocking, ringing and blurry effects, especially in low bit rate coding scenarios. Although many algorithms have been proposed to reduce these compression artifacts, most of them are based on image local smoothness prior, which usually leads to over-smoothing around the areas with distinct structures, *e.g.,* edges and textures. In this paper, we propose a novel framework to enhance the perceptual quality of decoded images by well preserving the edge structures and predicting visually pleasing textures. Firstly, we propose an edge-preserving generative adversarial network (EP-GAN) to achieve edge restoration and texture generation simultaneously. Then, we elaborately design an edge fidelity regularization term to guide our network, which jointly utilizes the signal fidelity, feature fidelity and adversarial constraint to reconstruct high quality decoded images. Experimental results demonstrate that the proposed EP-GAN is able to efficiently enhance decoded images at low bit rate and reconstruct more perceptually pleasing images with abundant textures and sharp edges.

Index Terms— Compression artifact reduction; generative adversarial network (GAN); perceptual loss; edge prior; image restoration

1. INTRODUCTION

With the explosion of digital media services, there is a increasing demand to compress the images/videos to facilitate storage and transmission, which may degenerate their quality especially at low bit rate. The popular lossy compression standards (e.g., JPEG and HEVC) adopt the block-based compression architecture and quantize every block independently to reduce the amount of transform coefficients. This process determines the amount of bit rate and may incur obvious blurring, ringing and blocking artifacts with coarse quantization steps, leading to poor user experience. In addition, these compression artifacts also have negative impacts on the accuracy of high-level computer vision tasks, such as face recognition and object detection [1]. Consequently, compression artifact reduction is required at the decoder side to enhance the perceptual quality of the compressed images.

In order to reduce compression artifacts, there are numerous post-processing techniques proposed in the literatures. In [2–4], researchers focused on restoring high quality decoded images by removing blocking artifacts based on image prior distribution. These methods mainly considered the smoothness or the regularity of images, which may lead to over smooth the true edges or texture details. Recently, deep convolutional neural networks (DCNNs) have demonstrated superior performance in both high-level [5] and low-level [6] computer vision tasks compared with traditional handcrafted features based algorithms. Dong *et al.* [7] proposed a pioneer work using artifact reduction convolutional neural network (AR-CNN) to reduce the JPEG compression artifacts. However, the perceptual quality of the reconstructed images from AR-CNN is still not satisfactory, and many structural details, *e.g.*, edges and textures, have been blurred and even removed along with compression artifacts. This is because the AR-CNN only adopts pixel-wise L_2 distance as its loss function without considering image regular structures, and all the local areas are processed equally.

To solve the above problems, researchers further introduced new loss functions [8] by constraining the perceptual distortion between the high level image features extracted by pre-trained CNN. The latest generative adversarial network (GAN) [9], which was proposed to generate images in a min-max game way, has been confirmed to promisingly produce photo-realistic texture details in low-level tasks such as super-resolution [10]. The motivation is that when the generated images are difficult to be distinguished from the original ones, they should be "real" enough for human perception. Therefore, these methods not only utilize features to measure per-

This work was supported in part by National Natural Science Foundation of China (61632001, 61571017), the National Basic Research Program of China (973 Program, 2015CB351800), National Program for Support of Top-Notch Young Professionals, which are gratefully acknowledged.

978-1-5386-1737-3/18/$31.00

ceptual similarity, but also generate visually pleasing textures.

In this paper, we propose a multi-constraint based post-processing algorithm to enhance the perceptual quality of the compressed images, especially for the images compressed at low bit rate. To reconstruct a high quality decoded image, we first propose a novel edge-preserving generative adversarial network (EP-GAN). Then, we carefully design a multi-constraint loss function by incorporating the signal fidelity loss, feature fidelity loss and adversarial loss with our proposed edge fidelity loss. By minimizing the proposed multi-constraint loss function, the proposed EP-GAN can obtain a more perceptually pleasing reconstruction with abundant textures and sharp edges, comparing with AR-CNN method. The main contributions of this paper are as follows:

- We propose an ***edge preserving image prior*** applied in edge fidelity loss term, which enforces the generative adversarial network to better restore the edge structures besides maintaining semantic similarity and generating texture details.
- We take advantage of multiple image characteristics and propose a multi-constraint framework for ***perceptual reconstruction*** of compressed images.

The remainder of this paper is organized as follows. Section 2 introduces the proposed multi-constraint framework for enhanced image decoding in detail. Extensive experimental results and discussions are reported in Section 3. Finally we conclude this paper in Section 4.

2. PROPOSED FRAMEWORK FOR ENHANCED IMAGE DECODING

2.1. Problem Formulation

For a decoded image, I^d, the goal of enhanced image decoding is to further improve its visual quality by removing the compression artifacts and reconstruct more visually pleasing edge and texture structures corrupted during compression, obtaining a high quality restoration image I^r. In order to generate a reconstructed image favored by human perceptions using deep neural network, we specifically design a perceptual loss function by combining several statistical characteristics of the corresponding uncompressed original image I^o instead of using the traditional mean square error (MSE) based loss function. The proposed perceptual loss function is explained in Eq.(1),

$$\mathcal{L}_{Percept} = \mathcal{L}_{MSE} + \lambda_1 \mathcal{L}_{Feat} + \lambda_2 \mathcal{L}_{Edge} + \lambda_3 \mathcal{L}_{Adv}, \quad (1)$$

where the weights, $\{\lambda_i\}$, are the trade-off parameters to balance the multiple loss components. In this paper, we empirically set λ_1 to 2×10^{-6}, λ_2 to 1 and λ_3 to 10^{-3} according to experimental results to make these loss function in the same order of magnitude, which can well balance both the objective and subjective quality of the reconstructed images. Herein, a novel EP-GAN is proposed, as shown in Fig. 1, to well restore the destroyed edge structures during compression. By minimizing the above perceptual loss function, our EP-GAN can generate visually pleasing restoration images. In the following sections, we will detail the network architectures and introduce the individual loss functions used to guide such network to achieve visually pleasing restoration.

2.2. Edge-preserving Generative Adversarial Network

2.2.1. Generator network architecture

In lossy compressed images or videos, edge distortions are always sensitive to be observed by Human Visual System (HVS). However, traditional reconstruction methods treated all the pixels equally and failed to recover the sharp edges in an effective way. To deal with this issue, we propose an edge prior of the high quality original image I^o as a guidance and force the generative network to perform edge detection as a joint task in compressed image reconstruction.

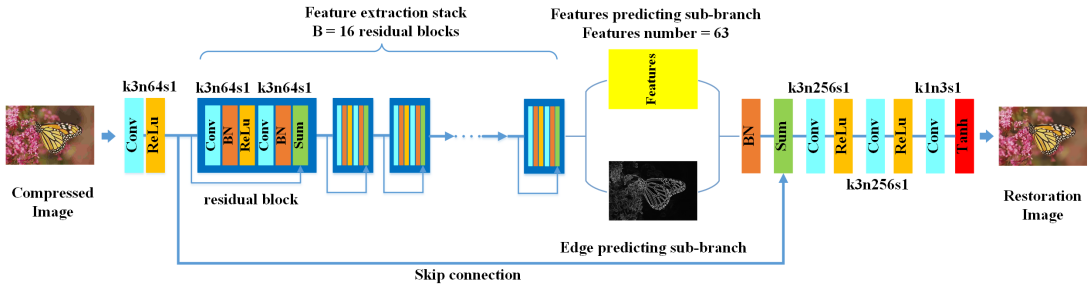
Feature extraction stack: We follow the network design of He *et al.*'s work in [11] and introduce skip connection, which has been confirmed to be efficient in training deep neural networks. We adopt the residual block proposed in [10] to build our neural network. Concretely, each residual block contains two convolutional layers with 3×3 kernel size and 64 feature maps, two Batch Normalization (BN) [12] layers and a ReLU [5] layer. As shown in Fig. 1, we utilize $B = 16$ residual blocks as a stack to fully extract the features from the compressed image for the following restoration.

Edge predicting sub-branch: We wish to incorporate our generative network with edge preserving mechanism to recover the distorted edges. Therefore, we propose to extract edge structure from the high quality image I^o as the prior label, denoted as I^E, which guides our deep generative network to perform multi-task learning. Specifically, the stacked residual blocks produce two sub-branches of features at the last layer blocks. One output is the features for image restoration and the other one is for edge prediction. Meanwhile, we introduce an edge fidelity loss as a regularization term in our loss function, which computes the L_2 distance between predicted and labeled edge maps.

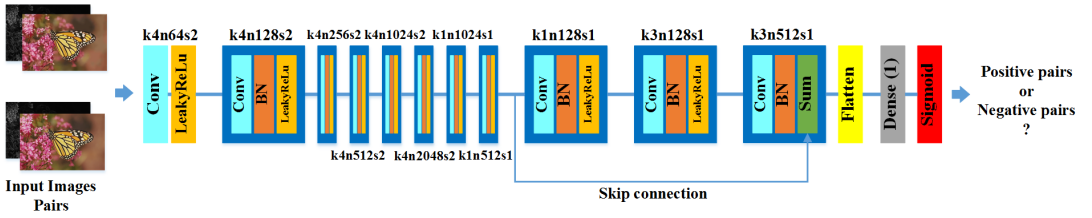
Then, we fuse the features and the edge map from the two sub-branchs through another two convolutional layers with 3×3 kernel size and 256 feature maps, followed by a ReLU layer. These combined features will be fed to the final layer to reconstruct the high quality restoration images, and the final layer is a convolutional layer with 1×1 kernel size and 3 feature maps followed by a *tanh* activation function.

2.2.2. Discriminator network architecture

In order to make full use of the edge prior we have introduced in section 2.2.1, we incorporate it as a condition into



(a) Generator network architecture



(b) Discriminator network architecture

Fig. 1: Proposed framework of edge-preserving generator network and discriminator network architectures. The k indicates the kernel size, n indicates the feature map numbers and s indicates the stride for each convolutional layer.

the discriminator network. Therefore, the positive examples $\{I^o|I^E\}$ and the negative examples $\{I^r|I^E\}$ are fed into the discriminative network. We take the architecture design proposed in [10] as a reference and use the LeakyReLU [13] activation ($\alpha = 0.2$) in our design, as shown in Fig. 1. The first six convolutional layers are all 4×4 filter kernels and the feature maps are in an increasing way by a factor of 2 from 64 to 2048 with stride 2. Except the first convolutional layer, other five layers are followed by BN layers. The seventh layer and the eighth layer utilize 1×1 filter kernels with stride 1, which correspond to 1024 and 512 feature maps. Then the two layers are followed by a special residual block with three convolutional layers, three BN layers and two LeakyReLU layers sequentially. The channels are 128, 128, and 512, with kernel sizes 1, 3 and 3. The outputs of the last convolutional unit are fed into a dense layer with *sigmoid* activation function to discriminate the positive pairs from the negative pairs.

2.3. Loss Functions for Perceptual Reconstruction

In this section, we introduce our loss function to train the proposed network. The loss function includes four components capturing distinct perceptual characteristics of the reconstructed image I^r aiming to obtain visually-pleasing images.

2.3.1. *Signal fidelity loss (MSE)*

To ensure the fidelity of the restored images, the pixel-wise MSE loss is utilized as follows,

$$\mathcal{L}_{MSE} = \frac{1}{WHC} \|I^o - I^r\|_2^2, \tag{2}$$

where W, H and C are the width, height and number of channels in the image.

2.3.2. *Feature fidelity loss*

Johnson *et al.* [8] first introduced the ***perceptual similarity measure*** by computing the distance in the pre-defined feature space instead of the image space. Similarly, we also define the feature space distance as the feature fidelity loss to encourage the network to reconstruct images preserving similar feature representation with the original image,

$$\mathcal{L}_{Feat} = \frac{1}{WHC} \|\phi(I^o) - \phi(I^r)\|_2^2, \tag{3}$$

where W, H and C are the width, height and number of channels in the feature maps, and $\phi(\cdot)$ denotes the feature space function, which is a pre-trained VGG-19 [14] network to map images into feature space. The fourth pooling layer is utilized to calculate the L_2 distances of the feature activations as our feature fidelity loss function.

2.3.3. *Edge fidelity loss*

As mentioned in section 2.2.1, edge distortions are easily observed by HVS. In order to reproduce sharp edges, we pro-

pose the edge fidelity loss by computing the edge map predicted by our network $\hat{I}^E$ and the label edge map I^E:

$$\mathcal{L}_{Edge} = \frac{1}{WH} \left\| \hat{I}^E - I^E \right\|_2^2. \quad (4)$$

where W, H are the width and height of the edge map. The labelled edge map I^E is extracted by a specific edge filter on the uncompressed image I^o. In our experiments, we choose the Sobel operator. $\hat{I}^E$ is predicted by edge predicting subbranch of the generator. The network is forced to perform edge-guided restoration by minimizing the edge fidelity loss.

2.3.4. Adversarial loss

We introduce the adversarial loss to further generate texture details which is favoured by Human. Our discriminative network D introduced in section 2.2.2, distinguishes whether the input image is original or reconstructed conditioned by the labelled edge map I^E and outputs the probability. We impose the negative log of this discrimination probability on the pair $\{I^r|I^E\}$ as the adversarial loss to the generative network:

$$\mathcal{L}_{Adv} = -\log\left(D(I^r|I^E)\right). \quad (5)$$

Simultaneously, the discriminative network minimizes

$$\mathcal{L}_D = -\log\left(D(I^o|I^E)\right) - \log\left(1 - D(I^r|I^E)\right). \quad (6)$$

It is worth mentioning that we use the labelled edge map I^E as the condition since the EP-GAN is expected to produce the realistic high-frequency in both texture and edge areas.

3. EXPERIMENTAL RESULTS

3.1. Implementation Details

We select 2060 pictures from Waterloo Exploration Database [15] in the training process. First, the MATLAB JPEG encoder is applied to these images using quality factor $QF = 10$ to generate low bit rate compressed images. Then, we randomly sampled patches in size of 224×224 from high quality images I^o and the corresponding compressed images I^d to make up pairs and trained our network on a NVIDIA Titan X GPU. We set the batch size as 4 and used Adam [16] with momentum term $\beta_1 = 0.9$ as our optimizer. In order to ensure the stability of the adversarial training, we first initialize and train the generator network by minimizing MSE only. Subsequently, we alternately train the discriminator network and generator network with the learning rate of 10^{-4} in the first half of the training epoches, and decrease the learning rate to 10^{-5} in the rest epoches.

3.2. Objective Evaluations

To verify the efficiency of the proposed method, we performed experiments on two commonly used datasets: LIVE1 [17] and the validation set of BSDS500 [18]. The traditional objective metrics, PSNR, PSNR-B [19] and SSIM [20], are used to evaluate the compression artifact reduction performance. We compare our EP-GAN method with the state-of-the-art approaches, SA-DCT [21], AR-CNN [7], L4 [22] and the latest compression artifact removal method using GAN proposed by Galteri *et al.* [23]. Because most of other state-of-the-art approaches (except GAN by Galteri *et al.*) only minimize the MSE loss, we replace the proposed perceptual loss by MSE to train our network as a "Baseline" for fair comparison. All experiments in this section are conducted on the image luminance component.

Table 1 shows the quantitative results of different compression artifact reduction methods. The baseline method outperforms SA-DCT, AR-CNN, L4 on all the datasets, which can be concluded that our network with MSE can recover the compressed image surpassing the state-of-the-art performance in objective metrics. Regarding GAN, although the performance of the method proposed by Galteri *et al.* is much lower than other approaches, our EP-GAN that uses edge prior to guide GAN training, achieves much better performance than that of Galteri *et al.*'s work.

Table 1: The average results of PSNR (dB), PSNR-B (dB), SSIM on the LIVE1 and BSDS500 images compressed by JPEG at QF = 10

Metrics	LIVE1			BSDS500		
	PSNR	PSNR-B	SSIM	PSNR	PSNR-B	SSIM
JPEG	27.77	25.33	0.791	27.61	24.98	0.777
SA-DCT	28.65	28.01	0.809	28.40	27.51	0.793
AR-CNN	28.98	28.70	0.822	28.82	28.49	0.805
L4	29.08	28.71	0.824	-	-	-
GAN	27.29	26.60	0.773	-	-	-
Baseline	**29.45**	**29.10**	**0.845**	**29.30**	**28.84**	**0.822**
EP-GAN	**28.80**	**28.04**	**0.823**	**28.45**	**27.59**	**0.798**

3.3. Perceptual Results

Since the ultimate goal of our work is not to achieve the best objective evaluation results, but a perceptual pleasing restoration image, we show the perceptual results in this section. The EP-GAN and baseline are compared with the existing available method AR-CNN. We select one image from the datasets of LIVE1 and BSDS500 respectively, and the subjective results are shown in Fig. 2.

We can see that there are obvious artifacts such as blocking and blurring artifacts in JPEG images compressed at low bit rate. The AR-CNN and the baseline methods can recover the low frequency information effectively, but many areas such as the grass in the restoration images are over-smoothed and blurred. However, our EP-GAN can produce very good details in the grassland and keep the fence edge sharp, thus making the whole image perceptually pleasing. We also ob-

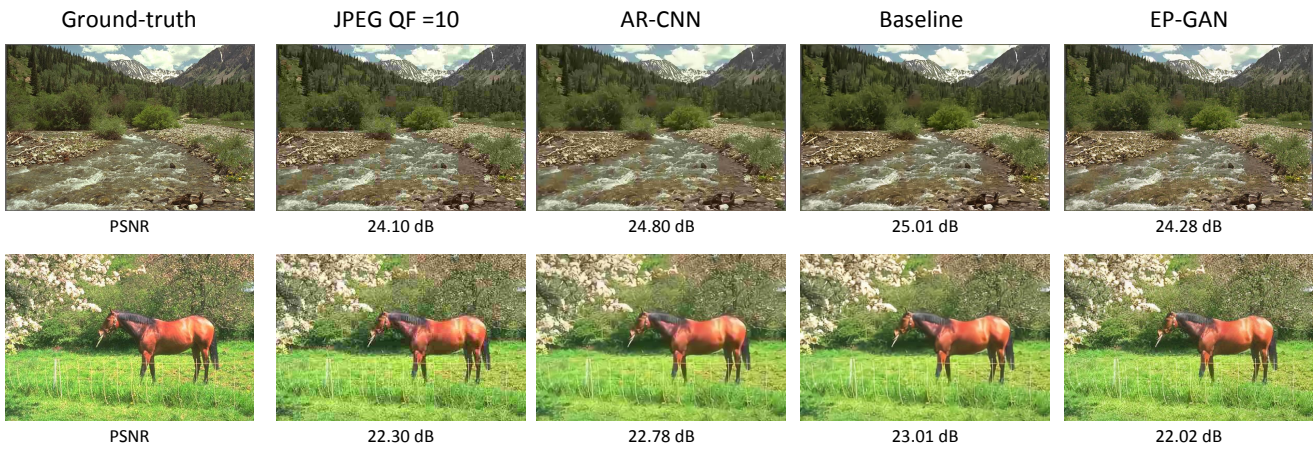


Fig. 2: Subjective comparison for LIVE1 and BSD100 images compressed by JPEG at QF=10, and restored by different compression artifact reduction methods.

serve that the EP-GAN does not introduce details randomly. For example, in the second row of Fig. 2, our approach largely enriches the texture of grass and flowers, while the body of horse remains smooth, which can efficiently prove that the EP-GAN reconstructs image in aware of semantics.

3.4. Investigations of Edge Preserving Scheme

Different from the Galteri *et al.*'s method, our EP-GAN introduces edge prior information to keep the edge fidelity and guide the generative network to concentrate on texture and edge areas simultaneously. In order to verify the effectiveness of edge preserving of EP-GAN, we abandon the edge predicting sub-branch in the generative network and remove the labelled edge map as the condition of the input in the discriminative network. Moreover, only the MSE loss, feature fidelity loss and the adversarial loss are utilized as the loss functions in network training. We name this variant of our EP-GAN as the V-GAN in the following.

We compare our EP-GAN with the V-GAN on the **Urban100** dataset which contains building pictures with sharp edges and compressed at QFs equalling to 10 and 20. Table 2 shows the results using the objective metrics. The EP-GAN performs better in all of the measurements, which confirms that EP-GAN can keep the edge fidelity better compared with the GAN without the designed edge loss function. Fig. 3 also obviously shows that the edges can be recovered much better using the proposed edge preserving prior information compared with V-GAN.

Table 2: The average results of PSNR (dB), PSNR-B (dB), SSIM, EPSNR on the Urban100 dataset with QF = 10, 20

Method	QF = 10			QF = 20		
	PSNR	PSNR-B	SSIM	PSNR	PSNR-B	SSIM
JPEG	26.95	24.58	0.832	29.29	26.85	0.893
V-GAN	28.33	27.66	0.859	29.69	29.37	0.905
EP-GAN	**28.58**	**27.84**	**0.868**	**30.75**	**29.54**	**0.912**

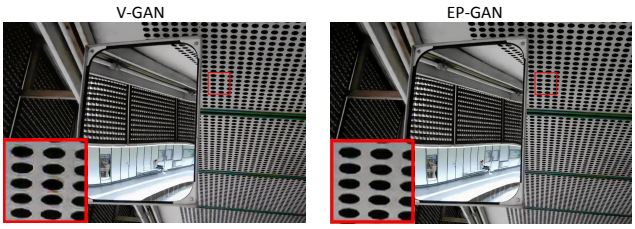


Fig. 3: Subjective comparison for Urban 100 dataset image compressed by JPEG at QF=10. It can be clearly seen that EP-GAN recovers the edge more efficiently.

4. CONCLUSION

In this paper, we proposed a novel framework to achieve the human-favored reconstructions for compressed images by well-preserving edge structures and predicting visually pleasing textures. Firstly, we proposed an EP-GAN by concentrating on edge restoration and texture generation simultaneously. Then we excavated distinctive visual characteristics of natural images by incorporating the signal fidelity loss, feature fidelity loss, adversarial loss and our proposed edge fidelity loss to train the network. According to extensive experimental results, the proposed EP-GAN outperforms the stat-of-the-art compression artifact reduction methods by obtaining more perceptually pleasing reconstruction results. Specifically, our perceptual reconstruction reproduces many details in texture area and generates shape edges comparing with AR-CNN method.

5. REFERENCES

[1] L. Zhao, X. Zhang, X. Zhang, S. Wang, S. Wang, S. Ma, and W. Gao, "Intelligent analysis oriented surveillance video coding," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2017, pp. 37–42.

[2] X. Zhang, R. Xiong, X. Fan, S. Ma, and W. Gao, "Compression artifact reduction by overlapped-block transform coefficient estimation with block similarity," *IEEE trans. on image processing*, vol. 22, no. 12, pp. 4613–4626, 2013.

[3] X. Zhang, R. Xiong, W. Lin, J. Ma, S.and Liu, and W. Gao, "Video compression artifact reduction via a spatio-temporal multi-hypothesis prediction," *IEEE Trans. on Image Processing*, vol. 24, no. 12, pp. 6048–6061, 2015.

[4] X. Zhang, X.and Lin, X. Xiong, R.and Liu, S. Ma, and W. Gao, "Low-rank decomposition-based restoration of compressed images via adaptive noise estimation," *IEEE Trans. on Image Processing*, vol. 25, no. 9, pp. 4158–4171, 2016.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[6] C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European Conference on Computer Vision*. Springer, 2014, pp. 184–199.

[7] C. Dong, Y. Deng, Chen C. C., and X. Tang, "Compression artifacts reduction by a deep convolutional network," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 576–584.

[8] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[10] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A.and Tejani, J. Totz, Z. Wang, et al., "Photo-realistic single image super-resolution using a generative adversarial network," *arXiv preprint arXiv:1609.04802*, 2016.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[12] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 2015, pp. 448–456.

[13] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, 2013, vol. 30.

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

[15] K. Ma, Z. Duanmu, Q. Wu, Z. Wang, H. Yong, and L. Li, H.and Zhang, "Waterloo Exploration Database: New challenges for image quality assessment models," *IEEE Transactions on Image Processing*, vol. 26, no. 2, pp. 1004–1016, Feb. 2017.

[16] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[17] H. R. Sheikh, Z. Wang, and A. Cormack, L.and Bovik, "Live image quality assessment database release 2 (2005)," 2016.

[18] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*. IEEE, 2001, vol. 2, pp. 416–423.

[19] C. Yim and A. Bovik, "Quality assessment of deblocked images," *IEEE Transactions on Image Processing*, vol. 20, no. 1, pp. 88–98, 2011.

[20] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[21] A. Foi, V. Katkovnik, and K. Egiazarian, "Pointwise shape-adaptive dct for high-quality denoising and deblocking of grayscale and color images," *IEEE Transactions on Image Processing*, vol. 16, no. 5, pp. 1395–1411, 2007.

[22] P. Svoboda, M. Hradis, D. Barina, and P. Zemcik, "Compression artifacts removal using convolutional neural networks," *arXiv preprint arXiv:1605.00366*, 2016.

[23] L. Galteri, L. Seidenari, M. Bertini, and A. Del Bimbo, "Deep generative adversarial compression artifact removal," *arXiv preprint arXiv:1704.02518*, 2017.