COMPRESSION ARTIFACT REMOVAL WITH STACKED MULTI-CONTEXT CHANNEL-WISE ATTENTION NETWORK

Binglin Li[⋆] *Jie Liang*[⋆] *Yang Wang*[†]

[⋆]School of Engineering Science, Simon Fraser University, Burnaby, BC, Canada
[†]Department of Computer Science, University of Manitoba, Winnipeg, MB, Canada
{binglinl, jiel}@sfu.ca[⋆] ywang@cs.umanitoba.ca[†]

ABSTRACT

Image compression plays an important role in saving disk storage and transmission bandwidth. Among traditional compression standards, JPEG is one of the commonly used standards in lossy image compression. However, the decompressed JPEG images usually have inevitable artifacts due to the quantization step, especially at low bitrate. Many recent works leverage deep learning networks to remove the JPEG artifacts and have achieved notable progress. In this paper, we propose a stacked multi-context channel-wise attention model. The channel-wise attention adaptively integrates features along the channel dimension given a set of feature maps. We apply multiple context-based channel attentions to enable the network to capture features from different resolutions. The entire architecture is trained progressively from the image space of low quality factor to that of high quality factor. Experiments show that we can achieve the state-of-the-art performance with lower complexity.

Index Terms— Compression artifact removal, image restoration, hourglass network, attention

1. INTRODUCTION

We consider the problem of artifact removal in lossy image compression. Compared to lossless compression standards such as PNG [13], lossy compression methods (e.g. JPEG [23], JPEG2000 [12] and WebP [8]) can produce smaller compressed files at the expense of a small amount of information loss. JPEG is the most commonly used standard in lossy image compression nowadays. The main components in JPEG include DCT, quantization, and entropy coding. Among them, almost all the information loss is caused by the quantization, which introduces various artifacts (blocking, ringing, blurring) that degrade the reconstructed images at the decoder. Compression artifact removal is a post-filtering process that aims to restore the degraded image as close to the artifact-free image as possible. Recent works show that deep learning is a promising technique for artifact removal. Methods based on deep learning can significantly improve perceptual and metric similarities between the reconstructed and the original images.

Since JPEG operates at block level with block size of 8×8 pixels, it may cause one object to be divided into several blocks during the compression and lead to artifacts in the reconstructed image. As different parts of an object are highly correlated and share similar textures, contextual information may help capturing patterns in this region and reducing the artifacts. In this paper, we propose to use attentions to capture contextual information in an image for artifact removal. The attention mechanism has been used in other computer vision and image processing tasks. For example, Chu et al. [5] apply spatial attentions for human pose estimation. Spatial attention allocates different weights in different spatial positions in an image. In image restoration, the evaluation is based on the whole image and we average the computed values of all pixels, which means learning individual importance to different pixels may not work. Besides, spatial attention will produce a large number of zeros. When the attention map is applied to features of deep learning network, most of the features will be mapped to zeros.

Based on these observations, we propose a deep learning-based stacked multi-context channel-wise attention model and apply it to JPEG compression artifact removal. In our model, feature maps are integrated by a learnt rescaling attention vector along the channel dimension. In [26], the channel attention is incorporated into each residual block with the same architecture. Different from [26], we augment the channel attention with multiple contexts at different scales. This allows our model to effectively exploit multi-scale contextual information. During training, we also take advantage of decompressed images with different quality factors to progressively supervise the network. Experiments show that our method can achieve the state-of-the-art performance with lower complexity.

2. RELATED WORK

Several researches on compression artifact removal have been proposed in recent years. In [6], a 4-layer convolutional net-

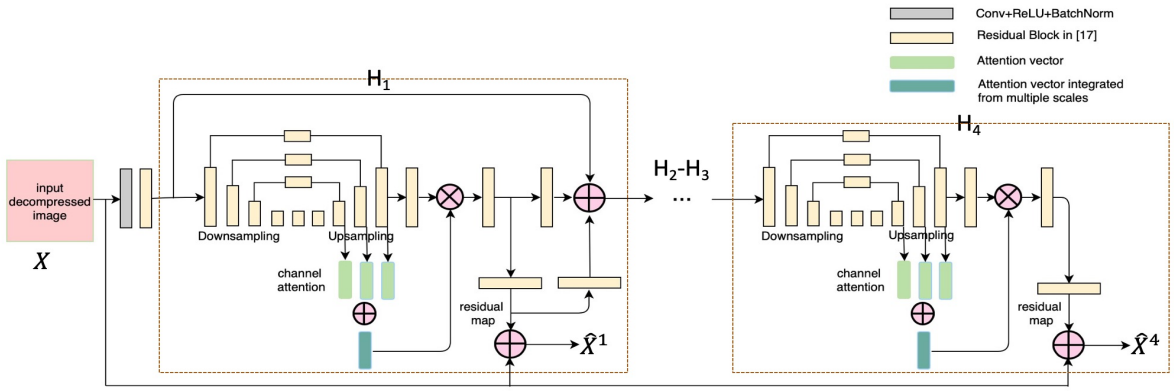


Fig. 1. Stacked multi-context channel-wise attention model. There are four stacks of hourglass networks H_1, H_2, H_3, H_4 in our system. H_2 and H_3 have the same architecture as H_1. $\widehat{X}^1$, $\widehat{X}^2$, $\widehat{X}^3$ and $\widehat{X}^4$ are reconstructed outputs for the four sub-networks respectively.

work is proposed, where the easy-to-hard transfer learning is used to initialize the parameters from a shallow network and transfer features learnt from high compression factors to that of low quality factors, which facilitates faster convergence than random initialization. The model in [20] includes 8 layers. It predicts residual map between the input and ground truth image, and uses skip connections to help to propagate information. It also combines the direct mapping loss with the Sobel edge loss to focus on high-frequency recovery for better perceptual reconstructions. In [3], a 12-layer deep network with hierarchical skip connections and a multi-scale loss is proposed. The architecture has multiple downsampling and upsampling, and predicts the reconstructed outputs at different scales. It demonstrates that deeper network has better capability to restore images and is also effective in low-level vision tasks.

Some works [10, 7] follow the spirit of Generative Adversarial Network (GAN). Basically GAN contains a generator and a discriminator, where the generator produces a candidate image output to fool the discriminator so that it is hard for the discriminator to distinguish whether the image is from the generator or it is a real image. These methods show that they can generate more realistic reconstructions, but may get relatively lower PSNR performance. They also apply extra perceptual loss where a pre-trained VGG network is used to share similar high-layer features between the predicted and the original images. The dual domain learning is adopted in [10, 9, 25] where features from both pixel domain and DCT domain are integrated to enhance the final reconstruction. However, it is not clear that whether the improvement is due to the proposed DCT-domain reconstruction or the increase in the number of parameters from the branch. In [21], a very deep MemNet consisting of many memory blocks is developed. Gate units are applied to control how much previous memory blocks and the current state are reserved. The densely connected structure helps restoring mid/high-frequency signals. In [7, 15], it is shown that image restoration can benefit subsequent high level computer vision tasks such as detection and segmentation.

Most recent works [14, 26, 11] focus on image super-resolution and achieve superior performance. However, their models have many more parameters (10M+) and require a larger training dataset to support. On the other hand, they are not specifically designed for compression artifact removal. As the noises in compression and super-resolution are quite different, the techniques in super-resolution do not necessarily achieve satisfactory results when applied to compression artifact removal task.

3. PROPOSED MODEL

In this section, we propose a deep learning-based multi-context channel-wise attention model to reduce JPEG compression artifacts. Our proposed model is based on the stacked hourglass network [17] which is originally developed for human pose estimation. Fig. 1 gives an overview of our proposed model. We use 4 stacks of hourglass networks $\{H_1, H_2, H_3, H_4\}$ to allow for iterative reconstructions. Each yellow box in Fig. 1 represents a single residual module same as in [17]. For the last few layers in each stack of the hourglass network, we collect the outputs of residual blocks from different scales, and apply the channel-wise attention for each scale (as shown in green boxes in Fig. 1).

Before the first hourglass network, we use a convolution layer and a residual block to obtain high frequency components. Each stack of hourglass network produces a 2D residual map between the input decompressed image and the target image with a channel dimension of 1. The residual map is added to the input decompressed image X to generate the reconstructed image $\widehat{X}^i$ at current stack. A 1×1 convolution

layer remaps the residual map to match the number of feature channels, and then add the output feature and input of this stack as an input for the next hourglass network. The last stack of hourglass network outputs $\widehat{X}^4$ without further steps. Given the input decompressed JPEG image X, the final reconstructed image $\widehat{X}^4$ is obtained by

$$\widehat{X}^4 = X + H_4(H_3(H_2(H_1(X, \theta_1), \theta_2), \theta_3), \theta_4). \quad (1)$$

$\Theta = \{\theta_1, \theta_2, \theta_3, \theta_4\}$ are parameters of the four sub-networks. The model parameters are trained end-to-end.

3.1. Channel-wise Attention Network

Channel attention network has shown great success in image super-resolution [26]. It adaptively integrates features by considering the interdependencies among different channels in a feature map.

Suppose the size of the feature map F is $C \times H \times W$, where C is the channel dimension, and $H \times W$ is the spatial size. First, we use a global average pooling to get a C-dimension channel vector $\mathbf{z}$ where the c-th element z_c is calculated as:

$$z_c = H_{GP}(F) = \frac{1}{H \times W} \Sigma_{i=1}^{H} \Sigma_{j=1}^{W} F_c(i, j). \quad (2)$$

$F_c(i, j)$ is the feature value at spatial location (i, j) of the c-th channel in the feature map F. This channel statistics can be viewed as a collection of the local descriptors for each feature map. To get the attention scores, we use a linear layer to map the pooled features $\mathbf{z}$ followed by a softmax operation. A ReLU layer is added for nonlinear interactions.

$$\begin{aligned} \mathbf{g} &= ReLU(\mathbf{W}\mathbf{z}) \\ a_c &= \frac{exp(g_c)}{\Sigma_{j=1}^{C} exp(g_j)} \end{aligned} \quad (3)$$

where $\mathbf{W} \in \mathbb{R}^{C \times C}$ is the weight of the linear layer and $\Sigma_{c=1}^{C} a_c = 1$. Attentions in each stack of hourglass network share the same parameter $\mathbf{W}$ to get scores and $\mathbf{W}$ is learnt during training.

3.2. Multi-context Channel-wise Attention Network

As shown in Fig. 1, the upsampling process generates features with different sizes r, *i.e.* $r = 8, 16, 32$. Previous work [2] has shown that deconvolution can produce checkerboard artifacts in the output image. PixelShuffle [19] is a popular alternative applied in super-resolution tasks [14, 26, 11], but it will lead to more parameters (from channel size c increased to $c \times s \times s$, where s is upsampling factor). Here we simply use bilinear layer for upsampling. The downsampling operation in the hourglass network not only reduces computation complexity, but also enlarges receptive field. The consecutive downsamplings and upsamplings enable our model to capture contextual information at different scales. We use $\mathbf{a}^r$ to represent the attention vector produced from the feature size r ($r = 8, 16, 32$). All attention vectors are summed up and then applied to the feature f to generate the refined feature h^{atten} by

$$h^{atten} = f \star (\Sigma_{r=8,16,32} \mathbf{a}^r) \quad (4)$$

where f is the output feature for an hourglass stack and $\star$ denotes the channel-wise multiplication operation.

Method	Classic 5 PSRN	Classic 5 SSIM	LIVE1 PSNR	LIVE1 SSIM
JPEG	27.82	0.7595	27.77	0.7730
ARCNN [6]	29.03	0.7929	28.96	0.8076
TNRD [4]	29.28	0.7992	29.15	0.8111
DnCNN [24]	29.40	0.8026	29.19	0.8123
CAS-CNN [3]	$\sim$	$\sim$	29.44$*$	0.8333$*$
MemNet [21]	29.69	0.8107	**29.45**	0.8193
hourglass	29.61	0.8100	29.37	0.8182
hourglass(PS)	29.63	0.8109	29.38	0.8186
Ours(PS+atten)	**29.70**	**0.8121**	**29.45**	**0.8201**
		0.8297$*$		0.8342$*$

Table 1. Average PSNR/SSIM on datasets Classic5 and LIVE1 with quality factor 10. $*$ indicates using different parameter setting for SSIM.

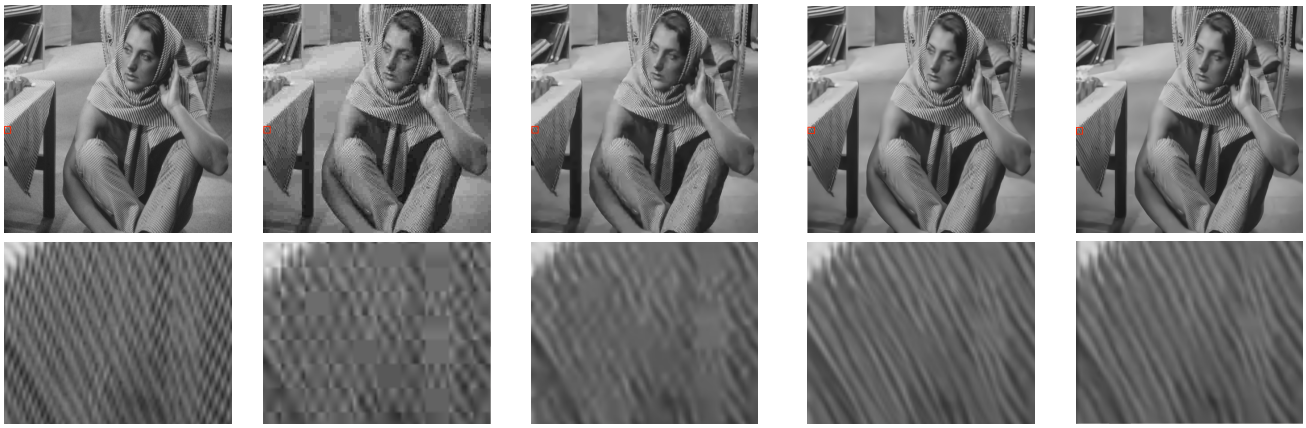
3.3. Progressive Supervision

Since we stack multiple hourglass networks and train them end-to-end by feeding the output of one stack as input of the next stack, the entire network can be quite deep. We can employ a loss at the end of each sub-network to avoid gradient vanishing problem. We use a sequence of JPEG decompressed images with higher quality factors for supervision after each stack, which allows the network to gradually learn the mapping from the space of low quality factor to the original image space (we call PS). The sequential decompressed images can guide the network to build a progressive path to predict the residual map at the end. Even without transferring features from high-quality decompressed images, we can still obtain good performance for low-quality decompressed images.

We use Mean Square Error (MSE) as the loss function for training. For an input decompressed image X with compression quality factor 10, $\Omega = \{X^1, X^2, X^3\}$ represent images with quality factors from 20 to 40 respectively. Y is the artifact-free image, therefore the final loss function is

$$\ell = \frac{1}{N} \Sigma_{j=1}^{N} (\Sigma_{i=1}^{1,2,3} ||\widehat{X}_j^i - X_j^i||^2 + ||\widehat{X}_j^4 - Y_j||^2) \quad (5)$$

where N is the number of samples, $\widehat{X}^i$ is the corresponding intermediate output and $\widehat{X}^4$ is the final output given in Eq. 1.



Original JPEG (25.78/0.762) ARCNN (26.92/0.797) MemNet (28.13/0.835) Ours (28.26/0.837)

Fig. 2. Quantitative results on image "barbara" from Classic 5 dataset with quality factor 10, shown as (PSNR/SSIM).

4. EXPERIMENT

Datasets: We train on images from the Berkeley Segmentation Dataset (BSD) [16]. We use the 400 images including the default split of training and test sets for learning our model. To compare with previous papers, we report testing results on the two popular benchmarks: Classic 5 (5 images) and LIVE1 (29 images) [18]. The LIVE1 dataset contains images with diverse properties.

Training Details: We modify the hourglass network from [17] in several ways. In order to reduce computation complexity, we fix all the layers with channel dimension as 64 instead of 128. Each hourglass contains 3 pairs of downsampling and upsampling. Our multi-context channel-wise attention network has 2.1M parameters in total, while the original hourglass network for human pose estimation [1] has around 40M parameters. Note that we calculate the released model for [21], it has around 2.91M parameters.

Before training, we first convert images from RGB color space to YCrCb, and then only use the luminance component during training and testing. We first train on 32×32 patches with a stride of 16 and then finetune on larger patch size 64×64 with a stride of 48. Results are reported on the entire test images. Following [21], data augmentation is applied for the training images. We do not use quality augmentation as in [21]. We apply the standard JPEG compression encoder in MATLAB to get the images with quality factors involved in the experiment.

Our model is built based on [1]. We use RMSprop [22] optimizer to train with a batch size of 256 for input size 32×32 and a batch size of 60 for size 64×64. We set the momentum parameter to 0.9 and weight decay of 10^{-4}. The initial learning rate is set to 0.0001 and then decreased by 10 every 10 epochs. PSNR and structural similarity (SSIM) are applied to evaluate the performance, where SSIM uses the same parameter setting as in [21].

Results and Discussions: Table 1 shows results (in terms of both PSRN and SSIM) for JPEG artifact removal on dataset Classic5 and LIVE1 for quality factor 10. We compare with several baseline approaches. Our method (*PS+atten*) with the proposed attention module achieves either comparable or better results compared to previous works. Note that CAS-CNN [3] uses 396k training images, which is far more than BSD dataset. They also use a different window for SSIM evaluation. We show our results on both parameter settings for SSIM.

The *hourglass* and *hourglass(PS)* models are the simplified hourglass network trained without and with progressive supervision as mentioned in Sec. 3.3 respectively. These two models do not use the multi-context channel-wise attention module. The performance of the baseline *hourglass* degrades 0.09dB in PSNR and 0.002 in SSIM compared with our proposed model. This shows the effectiveness of our proposed multi-context channel-wise attention model.

We show quantitative results on image "barbara" from Classic 5 dataset with quality factor 10 in Fig. 2. Compared to the original JPEG decompressed image, the reconstructed result from our approach has much clearer textures as well as less blocking artifacts. Even though the MemNet method in [21] has satisfactory restoration of the table cloth, ours can further eliminate the blurring artifact. Besides, our complexity is lower than that in [21].

5. CONCLUSION

In this paper, we propose a stacked multi-context channel-wise attention model which is trained in a progressive manner for JPEG artifact removal. The multi-context channel-wise attention can adaptively integrate information from different scales. Experiment results show that the proposed attention mechanism achieves the state-of-the-art performance with lower complexity.

6. REFERENCES

[1] A pytorch toolkit for 2d human pose estimation. *https://github.com/bearpaw/pytorch-pose*, 2017.

[2] A. Aitken, C. Ledig, L. Theis, J. Caballero, Z. Wang, and W. Shi. Checkerboard artifact free sub-pixel convolution: A note on sub-pixel convolution, resize convolution and convolution resize. *arXiv preprint arXiv:1707.02937*, 2017.

[3] L. Cavigelli, P. Hager, and L. Benini. Cas-cnn: A deep convolutional neural network for image compression artifact suppression. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pages 752–759. IEEE, 2017.

[4] Y. Chen and T. Pock. Trainable nonlinear reaction diffusion: A flexible framework for fast and effective image restoration. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1256–1272, 2017.

[5] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. *Conference on Computer Vision and Pattern Recognition*, 2017.

[6] C. Dong, Y. Deng, C. Change Loy, and X. Tang. Compression artifacts reduction by a deep convolutional network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 576–584, 2015.

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

[8] Google. Webp: Compression techniques. Accessed: 2018-11-26.

[9] J. Guo and H. Chao. Building dual-domain representations for compression artifacts reduction. In *European Conference on Computer Vision*, pages 628–644. Springer, 2016.

[10] J. Guo and H. Chao. One-to-many network for visually pleasing compression artifacts reduction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4867–4876, 2017.

[11] M. Haris, G. Shakhnarovich, and N. Ukita. Deep back-projection networks for super-resolution. In *Conference on Computer Vision and Pattern Recognition*, 2018.

[12] Information technology – jpeg 2000 image coding system: Core coding system. Standard, International Organization for Standardization, Dec. 2000.

[13] Information technology – computer graphics and image processing – portable network graphics (png): Functional specification. Standard, International Organization for Standardization, Mar. 2004.

[14] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. In *The IEEE conference on computer vision and pattern recognition (CVPR) workshops*, volume 1, page 4, 2017.

[15] D. Liu, B. Wen, J. Jiao, X. Liu, Z. Wang, and T. S. Huang. Connecting image denoising and high-level vision tasks via deep learning. *arXiv preprint arXiv:1809.01826*, 2018.

[16] 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*, volume 2, pages 416–423. IEEE, 2001.

[17] A. Newell, K. Yang, and J. Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016.

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

[19] W. Shi, J. Caballero, F. Huszár, J. Totz, A. P. Aitken, R. Bishop, D. Rueckert, and Z. Wang. Real-time single image and video super-resolution using an efficient subpixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1874–1883, 2016.

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

[21] Y. Tai, J. Yang, X. Liu, and C. Xu. Memnet: A persistent memory network for image restoration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4539–4547, 2017.

[22] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[23] G. K. Wallace. The jpeg still picture compression standard. *IEEE transactions on consumer electronics*, 38(1):xviii–xxxiv, 1992.

[24] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing*, 26(7):3142–3155, 2017.

[25] X. Zhang, W. Yang, Y. Hu, and J. Liu. Dmcnn: Dual-domain multi-scale convolutional neural network for compression artifacts removal. *arXiv preprint arXiv:1806.03275*, 2018.

[26] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu. Image super-resolution using very deep residual channel attention networks. *arXiv preprint arXiv:1807.02758*, 2018.