A Training Method for Image Compression Networks to Improve Perceptual Quality of Reconstructions

Jooyoung Lee*, Donghyun Kim, Younhee Kim
Hyoungjin Kwon, Jongho Kim & Taejin Lee
Broadcasting and Media Research Laboratory
Electronics and Telecommunications Research Institute
Daejeon, Korea
leejy1003@etri.re.kr

Abstract

Recently, neural-network based lossy image compression methods have been actively studied and they have achieved remarkable performance. However, the classical evaluation metrics, such as PSNR and MS-SSIM, that the recent approaches have been using in their objective function yield sub-optimal coding efficiency in terms of human perception, although they are very dominant metrics in research and standardization fields. Taking into account that improving the perceptual quality is one of major goals in lossy image compression, we propose a new training method that allows the existing image compression networks to reconstruct perceptually enhanced images. By experiments, we show the effectiveness of our method, both quantitatively and qualitatively.

1. Introduction

Recently, neural-network based lossy image compression methods [23, 10, 4, 22, 5, 14, 16, 15] have been actively studied. They have been progressively improving the coding efficiency and the latest approach [15] has obtained a remarkable coding efficiency that outperforms Versatile Video Coding (VVC) Intra (VTM 7.1 [1]), which has been almost finalized for standardization by ISO/IEC MPEG, in terms of both PSNR and multi-scale structural similarity index (MS-SSIM) [25]. However, the classical metrics, such as PSNR and MS-SSIM, for which most of the recent neural-network based approaches [23, 10, 4, 22, 5, 14, 16, 15] have been developed, may yeild sub-optimal coding efficiency in terms of human perception. Although those metrics are dominantly used in research and standardization fields, they may be inappropriate to measure quality perceived by very complex human visual systems.

Meanwhile, in image restoration field, several studies [13, 24, 18, 9] have been conducted for improving perceptual quality of reconstructed images. Some approaches [13, 24, 18] have adopted GAN [8]-based generation models that make a distribution of reconstructed images as close to that of original images as possible, and the VGG [20]-based feature space loss [9] also has been exploited for some recent image restoration approaches [13, 24, 18, 9].

In image compression filed, based on the superiority of generative models in image restoration, some approaches [17, 19, 3] targeting better perceptual quality have been proposed. Rippel *et al*. [17] first introduced GAN [8]-based training for image compression. Santurkar*et al*. [19] also proposed the generative compression models based on the GAN network. They trained the encoder-decoder networks in a step-wise manner (the generator (decoder) first via the adversarial loss; the encoder next via the L2 norm and the perceptual loss). Agutsson *et al*. [3] also have introduced a generative compression method based on GAN, in which whole encoder and decoder networks are trained in an end-to-end manner, through a new R-D optimization scheme for which an adversarial loss is jointly exploited. These GAN-based approaches [17, 19, 3] have achieved visually pleasing reconstructions compared to the conventional codecs such as JPEG2000 [21] and BPG [6], especially at extremely low bit-rate range.

Taking into account that improving the perceptual quality can be viewed as one of major goals in lossy image compression, we propose a new training scheme allowing the perceptually improved reconstruction, in which the up-to-date perceptual losses are utilized. In contrast to the previous generative compression methods [17, 19, 3], we only fine-tune the existing image compression model originally trained using the classical reconstruction losses, such

*Corresponding author

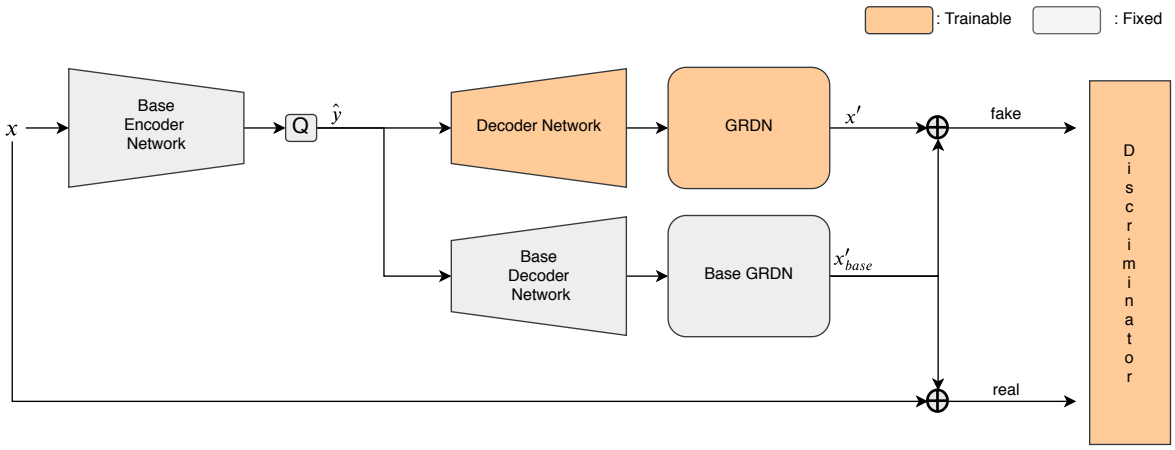

Figure 1: Training scheme of our perceptual quality oriented image Compression network.

as mean-squared-error (MSE) or MS-SSIM, from a more conservative perspective that the classical fidelity is as important as visual plausibility in image compression domain. For example, changing human faces cannot be acceptable in some applications, no matter how plausible the reconstructions are. For training our perceptual model, named *EIC-E2E-P*, we only fine-tune the reconstruction-related sub-networks of pre-trained base models, rather than optimizing the whole networks in an end-to-end manner.

2. Proposed method

As mentioned above, we pre-train a base image compression model, named *EIC-E2E-B*, and use the base model for parameter initialization of our model. As the base image compression model, we adopt a simplified version of JointIQ-Net [15] in which we only adopt the unified scheme of image compression and quality enhancement, and the model parameter refinement module (MPRM), for stable submission. After pre-training the base model using the MSE or MS-SSIM, we fine-tune only the reconstruction-related parts, including the decoder network and the quality enhancement network GRDN [11].

Fig. 1 shows our training scheme. Because we train only the reconstruction related sub-networks after pre-training, we omit some diagrams for compression related networks, such as the sub-networks for the hyperprior [5] and the context-adaptive model parameter estimating network, for clear illustration. As shown in Fig. 1, we only train the decoder network, the quality enhancement network GRDN [11], and the discriminator network. These trainable networks are highlighted with orange color. That is, we allow our decoder and quality enhancement networks to reconstruct more plausible images than those from the base model, although our model and the base model can share the same compressed files because the compression related elements are exactly same in both models.

To train our model, we exploit three different losses, the reconstruction loss $\mathcal{L}_{rec.}$, the perceptual loss $\mathcal{L}_{perc}$, and the adversarial loss $\mathcal{L}_{adv.}$, motivated by SRGAN [13] and ESRGAN [24]. Considering our decoder network and GRDN network to be a generator in adversarial training, the losses for the generator and the discriminator are formulated as follows:

$$\mathcal{L}_G = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{adv}\mathcal{L}_{adv_G} \quad (1)$$

$$\mathcal{L}_D = \mathcal{L}_{adv_D} \quad (2)$$

In Eq. 1 and 2, we basically adopt the typical reconstruction metric, 1 - MS-SSIM or MSE for the reconstruction loss $\mathcal{L}_{rec}$. We choose one out of the two metrics based on the type of an input image, which will be discussed later in this section. For the perceptual loss $\mathcal{L}_{perc}$, we use the MSE of the features extracted from the "conv5_4" layer of the VGG-19 [20] network. For the adversarial loss $\mathcal{L}_{adv}$, we do not use only the input $\boldsymbol{x}$ and the reconstruction $\boldsymbol{x'}$ as the real and fake data fed into the discriminator, but we use also the reconstruction $\boldsymbol{x'_{base}}$ from the base model in a pair-wise manner with the input $\boldsymbol{x}$ and reconstruction $\boldsymbol{x'}$ to compose the real and fake data in adversarial training. The reconstructions $\boldsymbol{x'_{base}}$ from the base model can be viewed as an additional context for the discriminator. $\mathcal{L}_{adv_G}$ and $\mathcal{L}_{adv_D}$ represents two symmetric adversarial losses for the generator and discriminator, respectively, which are calculated based on the relativistic average discriminator, as in ESRGAN [24]. λ_{rec}, λ_{perc}, and λ_{adv} represent the pre-defined weight parameters for $\mathcal{L}_{rec}$, $\mathcal{L}_{perc}$, and $\mathcal{L}_{adv}$, respectively.

Fig. 2 shows the sample images reconstructed from two versions of our base model, one optimized for MSE and the other optimized for MS-SSIM, respectively. As shown in

(a) (b)

Figure 2: Visual quality comparison between two versions of our base model (best viewed in digital format). One version is optimized for MSE, whereas the other is optimized for MS-SSIM (a) MSE optimized, bpp: 0.1307, PSNR: 33.7671, MS-SSIM: 0.9750 (b) MS-SSIM optimized, bpp: 0.1387, PSNR: 31.7666, MS-SSIM: 0.9845

Fig. 2, The base models trained for MSE have strength in maintaining structural information, whereas those trained for MS-SSIM preserve more textures. Consequently, we divide our models into two classes, a structure-oriented model and a texture-oriented model, according to the type of reconstruction loss $\mathcal{L}_{rec}$. The structure-oriented model adopts MSE as $\mathcal{L}_{rec}$, whereas the texture-oriented model uses MS-SSIM as $\mathcal{L}_{rec}$. The base model optimized with the same type of $\mathcal{L}_{rec}$ is used for each type of our model.

3. Implementation

We trained the base model first in the way described in [15], and then we initialized our model by duplicating the parameters from the pre-trained base model. For the structure-oriented model, we set λ_{rec}, λ_{perc}, and λ_{adv} to 40, 0.1, and 0.005, repectively, whereas we set them to 30, 0.1, and 0.005 for the texture-oriented models. To keep the fidelity and prevent undesired artifacts or structural distortions due to adversarial training, we increased the λ_{rec} values until those visual artifacts vanish. When we used the default setting of ESRGAN, our model suffered from the undesired artifacts. This may be due to the difference between image compression and super-resolution, in amount of given contexts.

In the training phase, we used 96×96 patches randomly extracted from CLIC [2] trainset, and we set the mini-batch size to 8. We trained each model using ADAM optimizer [12] for 500,000 iterations. We set the initial learning rate to 0.0001, and then gradient decaying was applied by decreasing the learning rate by half at every 10,000 steps during the final 50,000 steps.

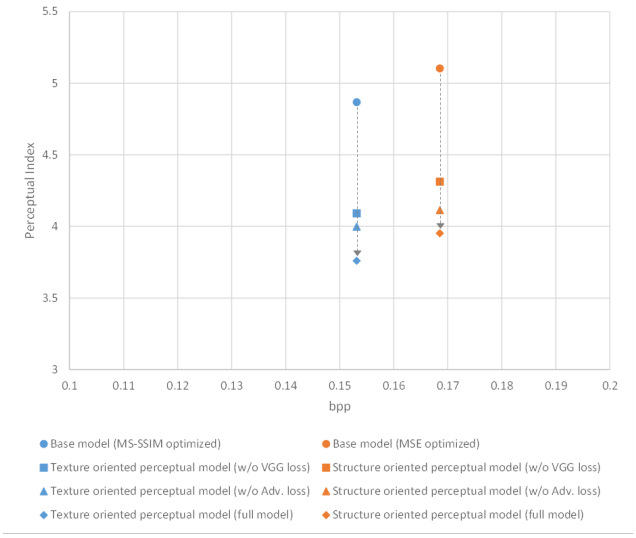

Figure 3: Perceptual index (PI) of the reconstructed images from the base models and their corresponding perceptual models.

4. Experiments

To verify the effectiveness of our method, we compared the average perceptual index (PI) values of reconstructed images from the base models with those from the perceptual models, over the CLIC [2] validation set images. The perceptual index is a metric for measuring perceptual quality of images, which was used in the PIRM-SR Challenge [7]. The higher perceptual quality is represented by the lower perceptual index. Fig. 3 shows the average perceptual index values of the reconstructed images from the base models and their corresponding perceptual models. The blue circle represents the base model optimized for MS-SSIM, and its corresponding texture-oriented perceptual model is denoted as the blue diamond. Likewise, the base model optimized

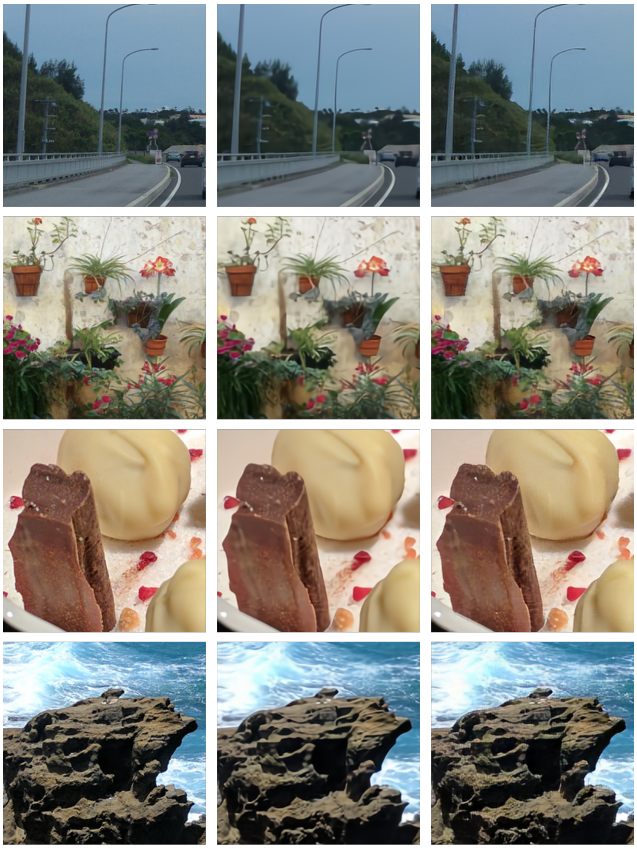

Figure 4: Visual quality comparison between the base models and their corresponding perceptual models (best viewed in digital format). (a) ground-truth images, (b) the reconstructions from the base models, (c) the reconstructions from the perceptual models.

for MSE and its corresponding structure-oriented perceptual model are denoted with orange color. As indicated with the dotted arrows, our perceptual models definitely improve the perceptual index values, compared with those from their corresponding base models. The base model optimized for MS-SSIM shows better results in terms of the perceptual index although they consume lower bit-rates compared with the base model optimized for MSE. Likewise, the texture-oriented perceptual model outperforms the structure-oriented perceptual model in terms the perceptual index. Correspondingly, we use the structure-oriented perceptual model only when there exists significant structural distortion in a reconstructed image. For further investigation on how each term in the objective function affects reconstructions in terms of the perceptual index, we evaluate two more models trained without the $\mathcal{L}_{perc}$ and $\mathcal{L}_{adv}$, respectively. As shown in Fig. 3, in which the two models are denoted as square and triangle markers, respectively, our full models show better results in terms of the perceptual index, and this represents that $\mathcal{L}_{perc}$ and $\mathcal{L}_{adv}$ are complementary for enhancing the perceptual index. Note that we trained three texture-oriented perceptual models with the different λ values to meet the bit-rate constraint of the challenge, but we omitted two out of the three models in Fig. 3, for simple illustration.

Fig. 4 shows the visual comparison results of our base models and the corresponding perceptual models. Compared to the reconstructions from the base model (Fig. 4 (b)), those from our perceptual models (Fig. 4 (c)) are more visually pleasing with clearer edges and richer textures. Note that two types of images in Fig. 4 (b) and Fig. 4 (c) are reconstructed from the same compressed binary files.

5. Conclusion

In this paper, we proposed a new training method for perceptual-quality oriented image compression. We utilized the existing entropy minimization based image compression approach, and fine-tuned the reconstruction-related sub-netowrks using the latest objective functions for better perceptual quality. We verified the effectiveness of our method by measuring the perceptual index, the metric indicating the perceptual quality of images. In addition, we provided visual comparison results, in which the reconstructions from our perctual models show better quality with clearer edges and richer textures.

Acknowledgments

This work was supported by Institute for Information and communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIP) 2017-0-00072, Development of Audio/Video Coding and Light Field Media Fundamental Technologies for Ultra Realistic Tera-media.

References

[1] Versatile video coding reference software version 7.1 (VTM-7.1), December 2019. 1

[2] Workshop and challenge on learned image compression, 2019. 3

[3] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. *arXiv preprint arXiv:1804.02958*, 2018. 1

[4] Johannes Ballé, Valero Laparra, and Eero P. Simoncelli. End-to-end optimized image compression. In *the 5th Int. Conf. on Learning Representations*, 2017. 1

[5] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *the 6th Int. Conf. on Learning Representations*, 2018. 1, 2

[6] Fabrice Bellard. Bpg image format, 2014. 1

[7] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *Workshop and Challenge on Perceptual Image Restoration and Manipulation in conjunction with ECCV 2018*, Sep 2018. 3

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. 1

[9] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016. 1

[10] Nick Johnston, Damien Vincent, David Minnen, Michele Covell, Saurabh Singh, Troy Chinen, Sung Jin Hwang, Joel Shor, and George Toderici. Improved lossy image compression with priming and spatially adaptive bit rates for recurrent networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 1

[11] Dong-Wook Kim, Jae Ryun Chung, and Seung-Won Jung. Grdn:grouped residual dense network for real image denoising and gan-based real-world noise modeling. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019. 2

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *the 3rd Int. Conf. on Learning Representations*, 2015. 3

[13] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2

[14] Jooyoung Lee, Seunghyun Cho, and Seung-Kwon Beack. Context-adaptive entropy model for end-to-end optimized image compression. In *the 7th Int. Conf. on Learning Representations*, May 2019. 1

[15] Jooyoung Lee, Seunghyun Cho, and Munchurl Kim. An end-to-end joint learning scheme of image compression and quality enhancement with improved entropy minimization. *arXiv preprint arXiv:1912.12817*, 2019. 1, 2, 3

[16] David Minnen, Johannes Ballé, and George Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *Advances in Neural Information Processing Systems*, May 2018. 1

[17] Oren Rippel and Lubomir Bourdev. Real-time adaptive image compression. In *International Conference on Machine Learning*, 2017. 1

[18] Mehdi S. M. Sajjadi, Bernhard Schölkopf, and Michael Hirsch. EnhanceNet: Single Image Super-Resolution through Automated Texture Synthesis. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 4501–4510. IEEE, 2017. 1

[19] Shibani Santurkar, David M. Budden, and Nir Shavit. Generative compression. In *The 33rd Picture Coding Symposium*, 2018. 1

[20] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations (ICLR)*, 2015. 1, 2

[21] David S. Taubman and Michael W. Marcellin. *JPEG 2000: Image Compression Fundamentals, Standards and Practice*. Kluwer Academic Publishers, Norwell, MA, USA, 2001. 1

[22] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy image compression with compressive autoencoders. In *the 5th Int. Conf. on Learning Representations*, 2017. 1

[23] George Toderici, Damien Vincent, Nick Johnston, Sung Jin Hwang, David Minnen, Joel Shor, and Michele Covell. Full resolution image compression with recurrent neural networks. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1

[24] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *The European Conference on Computer Vision Workshops (ECCVW)*, September 2018. 1, 2

[25] Zhou Wang, Eero P. Simoncelli, and Alan C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers*, 2003. 1