GAN-Based Image Compression Using Mutual Information for Optimizing Subjective Image Similarity

Shinobu KUDO1,2, Shota ORIHASHI1, Ryuichi TANIDA1, Members, Seishi TAKAMURA1, and Hideaki KIMATA1, Senior Members

Summary
Recently, image compression systems based on convolutional neural networks that use flexible nonlinear analysis and synthesis transformations have been developed to improve the restoration accuracy of decoded images. Although these methods that use objective metric such as peak signal-to-noise ratio and multi-scale structural similarity for optimization achieve high objective results, such metric may not reflect human visual characteristics and thus degrade subjective image quality. A method using a framework called a generative adversarial network (GAN) has been reported as one of the methods aiming to improve the subjective image quality. It optimizes the distribution of restored images to be close to that of natural images; thus it suppresses visual artifacts such as blurring, ringing, and blocking. However, since methods of this type are optimized to focus on whether the restored image is subjectively natural or not, components that are not correlated with the original image are mixed into the restored image during the decoding process. Thus, even though the appearance looks natural, subjective similarity may be degraded. In this paper, we investigated why the conventional GAN-based compression techniques degrade subjective similarity, then tackled this problem by rethinking how to handle image generation in the GAN framework between image sources with different probability distributions. The paper describes a method to maximize mutual information between the coding features and the restored images. Experimental results show that the proposed mutual information amount is clearly correlated with subjective similarity and the method makes it possible to develop image compression systems with high subjective similarity.

Key words: image compression, CNN, GAN, mutual information

1. Introduction

Recently, image compression techniques have become increasingly important due to the development and popularity of high-definition multimedia content. Conventional image coding systems, e.g., Better Portable Graphics (BPG) [1], WebP [2], and JPEG2000 [3], based on handcraft-designed predictions and transformations have poor performance against complex textures and unexpected signals. Furthermore, since they are designed to faithfully reproduce signals based on peak signal-to-noise ratio (PSNR), one of the objective evaluation indexes, they suffer from visual artifacts such as blurring, ringing, and blocking at low bit rate. Thus, subjective image quality is substantially degraded.

To improve the restoration accuracy, machine learning-based image compression systems using convolutional neural networks (CNNs) have been actively studied during the last few years [4]–[11]. Many of these methods perform end-to-end optimization using flexible nonlinear analysis and synthesis by using CNN-based autoencoders, and it has been reported that they show higher performance than conventional image compression systems. However, a problem still remains in that PSNR based optimization leads to degradation of subjective image quality at low bit rates such as 0.1 bits per pixel (bpp). To address this problem, a method using a framework called a generative adversarial network (GAN) [12] has been reported as one of the methods aiming to improve the subjective image quality. It optimizes the distribution of restored images to be close to that of natural images. In [15], the authors achieved higher performance than that for a High Efficiency Video Coding (HEVC)-based image compression system BPG [1] by adopting pyramidal decomposition and multi-scale structural similarity (MS-SSIM) based optimization. In [16], the authors proposed an optimization scheme similar to that proposed by [15], as well as a system in which the decoding area of the autoencoder is specified by the user.

However, since these methods are optimized to focus on whether the restored image is subjectively natural or not, components that are not correlated with the original image are mixed into the restored image during the decoding process. Thus, even though the appearance looks natural, subjective similarity may be degraded. For example, the decoded image may be subjectively seen as a different object from the original image or the impression may be changed. This is not preferable because the original role of the compression system is lost.

To address this problem, we first investigated why this phenomenon occurs. Then, we revealed that the conventional GAN-based methods could not retain the coding features during the decoding process. This means that the mutual information amount between the coding feature and the restored image decreases. From this investigation, this paper proposes image compression method that optimizes coding features obtained from the encoder and explicitly correlates them to the restored images by introducing mutual information maximization as regularization. In experimental section, we evaluate the relationship between the proposed mutual information amount and subjective similarity, indicating that the proposed mutual information amount is clearly correlated with subjective similarity and the method makes it possible to provide image compression system with higher subjective similarity.
Our contributions are twofold: 1) We show that the conventional GAN-based compression techniques cannot retain the coding features during the decoding process, which degrades subjective similarity between original images and restored images. 2) We propose a GAN-based compression system that can retain coding features by maximizing mutual information between coding features and restored images. Experimental results show that the proposed mutual information amount is correlated with subjective similarity and the proposed compression system obtains higher subjective similarity compared with the conventional GAN-based method.

This paper is an extension of previous study [13]. The differences from previous study are as follows: 1) We investigate the correlation between the coding features and its re-encoded features with optimization by the conventional GAN-based method. The results show that there is no correlation between the original image and its restored image in the conventional GAN-based method, and our method can solve this problem. 2) We describe the technical differences between our method and InfoGAN [14]. This makes it clear that the purpose of this paper is different from that of InfoGAN and that the equations have also been modified from InfoGAN. 3) The proposed objective function is described in detail, assuming that the posterior distribution is a Gaussian distribution. 4) Some simulations have been added. One of them is a comparison with non-GAN-based compression method. The results show that the non-GAN-based compression method has lower naturalness and similarity than the GAN-based compression methods. We also calculate PSNR and MS-SSIM metrics for all results and investigate how these metrics are correlated with subjective similarity, showing that these metrics cannot reflect subjective image quality, but the proposed mutual information amount can. The result also indicates that even the non-GAN-based compression method does not maximize the mutual information amount, which suggests that regardless of whether the adversarial error is used or not, components that are not correlated with the original image are mixed during the decoding process. Simulations with other datasets (Imagenet for train/test, ADE20k for train and Kodak for test), show that our method is also effective for natural images.

The rest of this paper is organized as follows. Related work and its problems are explained in Sect. 2. Sect. 3 presents the proposed image compression method. Sect. 4 describes simulations we conducted and the results obtained. Finally, we conclude the paper with a brief summary in Sect. 5.

2. Related Work

Machine learning-based image compression approaches have been studied during the last few years [4]–[11]. Many of these methods use CNN-based autoencoders in which the encoder transforms the input image into a bitstream and the decoder reconstructs the restored image. To improve the coding performance, some techniques were proposed. In [4] and [5], the authors introduced recurrent neural networks (RNNs) for feature extractions. In [6] and [7], the authors studied quantization process to approximate the entropy of the quantized representation and in [8] and [9], the authors directly trained the quantization. In [10], the authors proposed pixel-wise loss weighted by SSIM and spatially adaptive bit allocation algorithm. In [11], the authors developed the context models to directly model the entropy of the latent representation. Although these methods attained high objective results, PSNR or (MS-)SSIM metric for loss function does not reflect human visual characteristics and is not suitable for improving subjective image quality.

2.1 GAN-based Image Compression

The GAN-based image compression method [15], [16] introduces a GAN framework into CNN-based image compression systems. This method can improve the subjective image quality by optimizing the distribution of the restored images to be close to that of natural images. This model consists of the encoder E, the decoder G, and the discriminator D. The encoder maps the input images x to the coding features w( = E(x)), performs quantization Q, then outputs the bitstream z( = Q(w)). The decoder reconstructs the restored images ˆx ( = G(z)) from the bitstream. The discriminator discriminates whether the input is the input image or the restored image. The models are trained by optimizing the following objective function L that can be formulated as a min-max rate-distortion optimization problem as shown in Eq. (1)

\[
\min_{E,G} \max_{D} L = \mathbb{E}[f(D(x))] + \mathbb{E}[g(D(G(z)))] + \lambda_d \mathbb{E}[d(x, G(z))] + \lambda_r r(z),
\]

(1)

where d(x, ˆx) represents reconstruction error between the input image x and its reconstruction ˆx, ˆz, ˆr represents its weighting factor, r(z) denotes the coding bit loss, and \(\lambda_d\) denotes its weighting factor. f and g are scalar functions. One of the choices is f(y) = log(y), g(y) = log(1 − y) called Vanilla GAN [12], which corresponds to the minimization of the KL divergence of x and G(z). Since there is a non-differentiable quantizer Q during the backpropagation process, applying techniques such as approximation with uniform noise [7] or a straight-through estimator [17] enables end-to-end optimization.

The first and second terms in Eq. (1) represent the adversarial error. These factors enable subjectively natural output to be achieved by considering whether the restored image is natural or not.

2.2 Problem with Related Work

In the related work, the adversarial error optimizes the distribution of restored images to be close to that of original images, which improves the subjective image quality. However, there is no guarantee that the similarity between the
original image and its restored image is retained. In other words, components that are not correlated with the original image are mixed into the restored image during the decoding process. This is illustrated by Fig. 1 that is a plot of the coding features \( w \) and its re-encoded features \( w' \) (re-encode the restored image of \( w \)) with optimization by Eq. (1). If the similarity between the original image and its restored image is retained, the plot should show correlation \( w'_i = w_i \), but it is not. This means that the mutual information amount \( I(w; G(z)) \) between the coding feature \( w \) and the restored image \( G(z) \) decreases. As a result, the restored image may look different from the original image, or the impression may be changed even though the appearance is natural. In particular, human eyes are thought to be sensitive to changes in images related to the recognition of human faces and characters and so on. Therefore, for human face input images, even slight changes in facial expressions, eye shapes or hair styles may make a person look different from another person or give a different impression. This is not preferable because the original role of the compression system is lost.

### 3. Proposed Method

#### 3.1 Motivation

In order to solve the problems that occur with conventional methods, it is necessary to make sure that the components uncorrelated with the original image are not mixed into the restored image during the decoding process. Thus, optimization is required such that the coding feature and the restored image are completely correlated. In other words, this is equivalent to maximizing the mutual information amount \( I(w; G(w)) \) between the coding feature \( w \) and the restored image \( G(w) \). The mutual information amount can also be expressed as the difference between the true information source of the coding feature and the one of the restored image.

Although the technique for maximizing mutual information is similar to InfoGAN [14], its purpose is different from that of our work. InfoGAN aims at maximizing mutual information of latent feature distribution in terms of realizing disentangled representation of image features, but our work is intended to maximize mutual information of coding feature distribution in order to realize a coding system that provides high subjective image quality. Therefore, our method modifies the following points to apply the mutual information maximization to the compression task:

1. InfoGAN conditions only some latent features, but our work does all latent (coding) features.
2. InfoGAN does not model auxiliary distribution that makes problem solvable by using a technique known as variational information maximization, but our work explicitly models it as the encoder to solve the compression task and evaluates how closely the distributions match.

In this paper, we show how we focused on the mutual information amount between the coding features and the restored images and how we proposed to introduce its maximization into the objective function as regularization. In the proposed method, since the decoded image is explicitly correlated with the coding features, random decode processing is suppressed, the change of the appearance impression can be reduced, while the restored image is subjectively natural. Therefore, it is expected to make it possible to develop compression systems that are suitable for compression tasks.

#### 3.2 Image Compression Using Mutual Information Maximizing Regularization

The objective function \( L \) of the proposed method is shown in Eq. (2):

\[
\min_{E,G} \max_D \left[ \mathbb{E}[f(D(x))] + \mathbb{E}[g(D(G(z)))] + \lambda_r \mathbb{E}[d(x, G(z))] + \lambda_f r(z) - \lambda I(w; G(w)) \right]
\]

where \( I(w; G(w)) \) represents the mutual information amount between the coding features and the restored images and \( \lambda \) represents its weighting factor.

Since it is difficult to analytically solve \( I(w; G(w)) \) by using the posterior distribution \( P(w|G(w)) \), we introduced an auxiliary distribution \( F \) and expressed it by a variational lower bound \( L(F, G) \) using a technique known as variational information maximization [14] as shown in Eq. (3).

\[
I(w; G(w)) = H(w) - H(w|G(w)) = H(w) + \mathbb{E}_{\tilde{x} \sim G(w)} \mathbb{E}_{w \sim P(w)} \log P(w|\tilde{x}) \\
= H(w) + \mathbb{E}_{\tilde{x} \sim G(w)} \mathbb{E}_{w \sim P(w|\tilde{x})} \log P(w|\tilde{x}) + D_{KL}(P(\tilde{x}|F)|P(\tilde{x}|G)) \\
\geq H(w) + \mathbb{E}_{\tilde{x} \sim G(w)} \mathbb{E}_{w \sim P(w|\tilde{x})} \log F(w|\tilde{x}) \\
= L_f(G, F),
\]

![Fig. 1](image-url)
The posterior distribution $P(w'|\hat{x})$, which is $F(w'|\hat{x})$, must be modeled with the encoder $E$ and we assumed this was a Gaussian distribution $N(\mu, 1)$. We also assumed $P(w) \sim N(0, 1)$, thus $H(w) = \log(\sqrt{2\pi}e)$. $L_f(G, F)$ is described as shown in Eq. (4)

$$
L_f(G, F) = H(w) + \mathbb{E}_{\hat{z} \sim G(w)}[\mathbb{E}_{w' \sim P(w|\hat{z})}[\log F(w'|\hat{z})]]
$$

$$
= \mathbb{E}_{\hat{z} \sim G(w)}[\mathbb{E}_{w' \sim P(w|\hat{z})}[\log \frac{1}{\sqrt{2\pi}} e^{-\frac{(w-w')^2}{2}}]]
+ \log(\sqrt{2\pi}e)
$$

$$
= \mathbb{E}_{\hat{z} \sim G(w)}[\mathbb{E}_{w' \sim P(w|\hat{z})}[-\frac{(w-w')^2}{2} - \log(2\pi)]]
+ \log(\sqrt{2\pi}e)
$$

$$
= -\mathbb{E}_{\hat{z} \sim G(w)}[\mathbb{E}_{w' \sim P(w|\hat{z})}[(w-w')^2 + 0.5].
$$

(4)

This also shows that when the lower bound attains its maximum $L_f(G, F) = 0.5$, the lower bound becomes tight and the maximal mutual information is achieved.

The final proposed objective function is shown in Eq. (5)

$$
\min_{E} \max_{G} \min_{b} L = \mathbb{E}[f(D(x))] + \mathbb{E}[g(D(G(z)))]
+ \lambda_d \mathbb{E}[d(x, G(z))] + \lambda_r r(z)
+ \lambda_f \mathbb{E}_{\hat{z} \sim G}[(w-w')^2].
$$

(5)

4. Experiments

We conducted simulations to verify the efficiency of the proposed method.

4.1 Network Architecture

The network architecture was based on ResNet as shown in Fig. 2. The encoder output size was $N \times N \times C/64$ where the image size is $N \times N$. LeakyReLU ($\alpha = 0.2$) was applied to the output of each layer except for the last layer of each network and the tanh function was applied to the output of the encoder and the decoder. The quantizer $Q$ quantizes each element to 1 bit [-1, 1] and entropy coding was not performed for the data after quantization. Thus, quantized data was directly used as the bitstream. Vanilla GAN [12] was used for $f$ and $g$.

4.2 Experiment Conditions

We compared the method with BPG (version 0.9.7) [11] and the method [7] (optimized by PSNR) and [16]. Since we compared only the effect of the objective function in the experiments, we used conventional methods that had the same network architecture configuration as ours. In more detailed, for the method [7], the network architecture was changed to one in Fig. 2 and quantizer was also changed to binarizer [-1, 1]. That is, this method corresponds to using the reconstruction term in Eq. (1) for loss function (hereafter referred to as “PSNR Train”). For the method [16], the network architecture was changed to one in Fig. 2 and quantized bits to 1. The loss function was the same as original (hereafter referred to as “GAN Train”). For the proposed method, we used the mean square error for $d(x|G(z))$. The other parameters were $C = 8$, $\lambda_d = 10.0$, $\lambda_r = 0$, and $\lambda_f = 10.0$. BPG changes the quantization parameter and chooses the one closest to the coding bits obtained with the others. The other parameters are defaults (cfm4=420, color_space=ycbcr, bit_depth=8, encoder=x265, level=8). We used the human face dataset (CelebA [18]) to perform the simulations under the severe condition where the human eye seems to be especially sensitive to changes. 200,000 training data elements and 25 test data elements were extracted from the dataset and resized to 128 \times 128 pixels. All methods except for BPG used the same optimization settings. Adam [19] was used for the optimization, with the batch size set to 16, the learning rate to 0.0001, and the number of iterations to 200,000. We used $R_1$ regularization [20] and the weighted average [21]. For the non-differentiable quantizer $Q$, we used a straight-through estimator [17].

4.3 Evaluation Method

We conducted subjective evaluation experiments using 10 image experts as subjects. The evaluation method was the Absolute Category Rating (ACR) method and the Degradation Category Rating (DCR) method in which the degree of image quality and the degree of degradation with the reference (original) image for each evaluation image were evaluated by five stage evaluation scores [22]. As an evaluation score for the ACR method, we defined “naturalness” that represents the naturalness of the image when the evaluation image is viewed independently, and for the DCR method, “similarity” that represents the similarity when compared with the reference image.

4.4 Results and Discussions

Figure 3 shows the transition of the lower bound $L_f(G, E)$ for the mutual information amount in “GAN Train” and the proposed method over training iterations. The results confirmed that the lower bound increases more with the proposed method and comes close to the upper limit ($= 0.5$). This means the mutual information can be maximized.

Figure 4 shows the plot of the coding features $w$ and its re-encoded ones $w'$ in our method. This also indicated that the correlation between $w$ and $w'$ is retained.

The subjective image quality evaluation results are shown in Fig. 5. The evaluation scores in the figure show the average scores of all subjects for each of the 25 evaluation images. These results indicate that BPG has a noticeably low degree of naturalness and similarity (1.2 points), indi-
Fig. 2  Network architecture used in the experiments

Fig. 3  Lower bound $L_i$ in “GAN Train” and the proposed method over training iterations

cating there were insufficient coding bits. “PSNR Train” has higher degree of naturalness and similarity than BPG (respectively 3.1 points and 2.7 points), but is not as good as the other methods. This indicates that the method using PSNR metric is not suitable for subjective image quality improvement. Our method and “GAN Train” had almost the same naturalness (respectively 3.6 points and 3.4 points), but the similarity for the former was 0.6 points higher than that for the latter (respectively 3.4 points and 2.8 points). These results made it clear that our method outperforms “GAN Train” reported in providing good restoration image quality so that images close to the original images are obtained.

Figure 6 shows restored images obtained for each of the methods in question in comparison with the ground truth. They show that the BPG images are significantly degraded; this is due to lost high frequency components. It is

Fig. 4  100 randomly sampled coding features $w$ and re-encoded ones $w'$ in the proposed method

Fig. 5  Subjective evaluation results
also found that “PSNR Train” lost high frequency components. In contrast, both “GAN Train” and our method preserved texture and achieved natural restoration. However, our method produced clearly better results for both the eye lines in (a) and (b) and the mouth shape in (c) and (d). We consider that this accounts for the differences in the subjective image quality evaluation results obtained for the two methods.

We evaluate how well each metric correlates with subjective image quality. Figure 7 plots the relationship between subjective image similarity and each metric (respectively (a) PSNR, (b) MS-SSIM and (c) mutual information) for the image results obtained by each method. The averages of the correlation coefficients for each image are (a) 0.108,
Fig. 7  The relationship between subjective image similarity and each metric (respectively (a) PSNR, (b) MS-SSIM and (c) mutual information (Lower bound $L_I$)) for the image results obtained by each method. Each color represents the result for the same image.

(b) $0.162$ and (c) $0.801$, respectively. These results indicate that PSNR and MS-SSIM metric cannot reflect subjective image similarity, on the contrary, the proposed mutual information amount is clearly correlated with it. The reason PSNR and MS-SSIM metric degrades subjective image similarity is considered that these metrics has no spatial phase because it directly minimizes the difference of the signals in spatial domain. On the other hand, since the proposed one minimizes the difference of the coding features obtained by CNN, it is assumed that spatial phase is taken into account in the optimization, which is the reason our method worked well. It is worth noting that even “PSNR Train”, which is optimized only by PSNR, does not maximize the mutual information amount. This suggests that, regardless of whether the adversarial error is used or not, components that are not correlated with the original image are mixed during the decoding process. Thus, the proposed mutual information amount is effective for non-GAN-based methods as well.
4.5 Results for Other Datasets

We also conducted simulations for Cityscapes dataset [23], Imagenet dataset [24] and ADE20k dataset [25] (Kodak [26] was used for test phase), respectively. We made some changes to experiment conditions; the dataset was resized to 256 × 256 pixels, reshape and linear layer on encoder and decoder were removed, Residual Block with 8ch was inserted on encoder, and instance norm was inserted in front of each leaky ReLU layer. We set the number of iterations to 500,000 for Cityscapes dataset and 2,000,000 for Imagenet and ADE20k dataset. The other parameters were same as described in 4.2.

Figure 8, 9 and 10 show restored images obtained for each of the methods. They also show that our method produced higher similarity compared with both “PSNR Train” and “GAN Train”. On the other hand, there were parts that could not be improved even by our method. It was a hotel signboard (green square in Fig. 8 (b)), which was replaced of each leaky ReLU layer. We set the number of iterations to 500,000 for Cityscapes dataset and 2,000,000 for Imagenet and ADE20k dataset. The other parameters were same as described in 4.2.

Figure 9  Restored images produced by different compression systems (Imagenet). L to R: ground truth, BPG [1], “PSNR Train”, “GAN Train”, our method.
by a road sign. This seems that the hotel signboard was not included in training dataset. Therefore, we can see that what cannot be expressed itself cannot be restored even with our method.

5. Conclusion

In this paper, we show that the conventional GAN-based compression techniques cannot retain the coding features during the decoding process, which degrades subjective similarity between original images and restored images. To solve this problem, we propose a GAN-based compression system that can retain coding features by maximizing mutual information between coding features and restored images. Subjective evaluation experiment results show that the proposed compression system obtains higher subjective similarity compared with the conventional GAN-based method.
References

[1] F. Bellard, “BPG Image format,” https://bellard.org/bpg/.
[2] “WebP Image format,” https://developers.google.com/speed/webp/.
[3] D.S. Taubman and M.W. Marcellin, “JPEG 2000: Image Compression Fundamentals, Standards and Practice,” Kluwer Academic Publishers, Norwell, MA, USA, 2001.
[4] G. Toderici, S.M. O’Malley, S.J. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell, and R. Sukthanker, “Variable rate image compression with recurrent neural networks,” arXiv preprint arXiv:1511.06085, March 2016.
[5] G. Toderici, D. Vincent, N. Johnston, S.J. Hwang, D. Minnen, J. Shor, and M. Covell, “Full Resolution Image Compression with Recurrent Neural Networks,” arXiv preprint arXiv:1608.05148, July 2017.
[6] L. Theis, W. Shi, A. Cunningham, and F. Huszar, “Lossy image compression with compressive autoencoders,” arXiv preprint arXiv:1703.00395, March 2017.
[7] J. Ballé, V. Laparra, and E.P. Simoncelli, “End-to-end optimized image compression,” arXiv preprint arXiv:1611.01704, March 2017.
[8] J. Cai and L. Zhang, “Deep Image Compression with Iterative Non-Uniform Quantization,” 2018 25th IEEE Int. Conf. Image Process. (ICIP), Sept. 2018.
[9] J. Cai, Z. Cao, and L. Zhang, “Learning a Single Tucker Decomposition Network for Lossy Image Compression with Multiple Bits-Per-Pixel Rates,” arXiv preprint arXiv:1807.03470, July 2018.
[10] N. Johnston, D. Vincent, D. Minnen, M. Covell, S. Singh, T. Chinen, S.J. Hwang, J. Shor, and G. Toderici, “Improved Lossy Image Compression with Priming and Spatially Adaptive Bit Rates for Recurrent Networks,” arXiv preprint arXiv:1703.10114, March 2017.
[11] F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L.V. Gool, “Conditional Probability Models for Deep Image Compression,” arXiv preprint arXiv:1801.04260, June 2019.
[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” Advances in Neural Information Processing Systems, pp.2672–2680, Dec. 2014.
[13] S. Kudo, S. Orihashi, R. Tanida, and A. Shimizu, “GAN-based Image Compression Using Mutual Information Maximizing Regularization,” IEEE Conf. Picture Coding Symposium (PCS), 2019.
[14] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, “InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets,” arXiv preprint arXiv:1606.03657, June 2016.
[15] O. Rippel and L. Bourdev, “Real-time adaptive image compression,” arXiv preprint arXiv:1705.05823, May 2017.
[16] E. Agustsson, M. Tschannen, F. Mentzer, R. Timofte, and L.V. Gool, “Generative Adversarial Networks for Extreme Learned Image Compression,” arXiv preprint arXiv:1804.02958, April 2018.
[17] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, “Binarized neural networks: Training deep neural networks with weights and activations constrained to +1 or -1,” arXiv preprint arXiv:1602.02830, 2016.
[18] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep learning face attributes in the wild,” Proc. Int. Conf. Comput. Vis. (ICCV), 2015.
[19] D. Kingma and J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6984, 2014.
[20] L. Mescheder, A. Geiger, and S. Nowozin, “Which training methods for GANs do actually converge?,” arXiv preprint arXiv:1801.04406, July 2018.
[21] Y. Yazici, C.-S. Foo, S. Winkler, K.-H. Yap, G. Pililiouras, and V. Chandrasekhar, “The unusual effectiveness of averaging in GAN training,” arXiv preprint arXiv:1806.04498, Feb. 2019.
[22] Recommendation ITU-T P. 910, “quality assessment methods for multimedia applications,” April 2008.
[23] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The Cityscapes dataset for semantic urban scene understanding,” IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2016.
[24] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” Int. J. Comput. Vis. (IJCV), vol.115, no.3, pp.211–252, 2015.
[25] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso and A. Torralba, “Scene parsing through ade20k dataset,” IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), 2017.
[26] “Kodak PhotoCD dataset,” http://r0k.us/graphics/kodak/.
Seishi Takamura received the B.E., M.E. and Ph.D. degrees from Department of Electronic Engineering, Faculty of Engineering, the University of Tokyo in 1991, 1993 and 1996 respectively. In 1996 he joined Nippon Telegraph and Telephone (NTT) Corporation, where he has been engaged in research on efficient video coding and ultra-high quality video coding. He has fulfilled various duties in the research and academic community in current and prior roles including Associate Editor of IEEE Transactions on Circuits and Systems for Video Technology (2006-2014), Editor-in-Chief of the Institute of Image Information and Television Engineers (ITE, 2015-2017), Executive Committee Member of IEEE Tokyo Section (2011-2018), Japan Council (2013-2016), and Region 10 (2019-). He has also served as Chair of ISO/IEC JTC 1/SC 29 Japan National Body, Japan Head of Delegation of ISO/IEC JTC 1/SC 29, and as an International Steering Committee Member of the Picture Coding Symposium. From 2005 to 2006, he was a Visiting Scientist at Stanford University, California, USA. He is currently a Senior Distinguished Engineer at NTT Media Intelligence Laboratories. Dr. Takamura is a Fellow of IEEE, a senior member of IEICE and IPSJ and a member of MENSA, APSIPA, SID and ITE.

Hideaki Kimata received the B.E. and M.E. degrees in applied physics, and the Ph.D. degree in electrical engineering respectively from Nagoya University, Nagoya, Japan, in 1993, 1995, and 2006. He joined Nippon Telegraph and Telephone Corporation (NTT) in 1995, and has been involved in the research and development of video processing of coding, realistic communication, computer vision, and machine learning. He is currently a Senior Research Engineer and also the Supervisor at NTT Media Intelligence Laboratories. He is the Chair of Technical Committee on Image Engineering of the Institute of Electronics, Information and Communication Engineers of Japan.