Semantic Map Based Image Compression via Conditional Generative Adversarial Network

Zhensong Wei, Zeyi Liao, Huihui Bai(✉), and Yao Zhao

Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
{zhswei,zyliao,hhbai,yzhao}@bjtu.edu.cn

Abstract. Recently, deep learning methods have been applied for image compression and achieved promising results. For lossy image compression at low bit rate, the traditional compression algorithms usually introduce undesired compression artifacts, such as blocking and blurry effects. In this paper, we propose a novel semantic map based image compression framework (SMIC), restoring visually pleasing images at significantly low bit rate. At the encoder, a semantic segmentation network (SS-Net) is designed to generate a semantic map, which is encoded as the first part of the bit stream. Furthermore, a sampled image of the input image is compressed as the second part of bit stream. Then, at the decoder, in order to reconstruct high perceptual quality images, we design an image reconstruction network (Rec-Net) conditioned on the sampled image and corresponding semantic map. Experimental results demonstrate that the proposed framework can reconstruct more perceptually pleasing images at low bit rate.

Keywords: Image compression · Semantic map · Generative adversarial network

1 Introduction

Over the past few years, there has been an active interest in making a prediction at every pixel in whole-image, named pixel-wise semantic segmentation. In a semantic segmentation map, each pixel is labeled with the class of its enclosing object or region. Semantic segmentation has a wide array of applications ranging from scene understanding, autonomous driving to inferring support-relationships among objects in images. Recently, some of the approaches based on deep learning (DL) particularly are designed for semantic segmentation, obtaining the promising results by learning the mapping from low resolution features to categories of the input image [1–3]. Recent advancements in generative models also show promise for the task of semantic segmentation [4–6]. In addition, it is further shown that generative models can synthesize a high-quality image using only a semantic map as input [7].

Y. Zhao et al. (Eds.): ICIG 2019, LNCS 11903, pp. 13–22, 2019.
https://doi.org/10.1007/978-3-030-34113-8_2

Image compression has been a fundamental and significant research topic in the field of image processing for several decades, which refers to the task of representing images using as little storage as possible. For the task of image compression, there are two main categories, named lossless compression and lossy compression. In lossless image compression, that is, an original image should be completely recovered with limited compression rate, while in lossy image compression, a greater reduction in storage can be achieved by allowing some reconstruction distortion. The traditional image compression algorithms, such as JPEG and JPEG2000, rely on handcrafted codec blocks. They usually consist of three parts: transform, quantization and entropy code. At the very low bit rate, the compressed image may incur serious blocking and blurring artifacts with quantization operation, leading to poor perceptual quality. These compression artifacts not only affect the expression of information in the image but also impact on the accuracy of high-level computer vision tasks.

Recently, DL-based approaches have the potential to improve the performance of image compression. Several methods have been proposed using different networks, achieving promising image compression results [8–10]. In [8], the authors proposed a framework for end-to-end optimization of an image compression model based on nonlinear transform. The work of [9] used learned context models for improved coding performance on their trained models when using adaptive arithmetic coding. In [10], the researchers proposed an end-to-end trainable model for image compression based on variational autoencoder, and the model incorporated a hyperprior to capture spatial dependencies in the latent representation. Furthermore, the main idea of GAN has enabled a significant process in photo-realistic image generation, which is particularly relevant to the real world and has visually pleasing results. In [11], they trained the synthesis transform as a generative model for generative compression, and demonstrated the potential of generative compression for orders-of-magnitude improvement in image compression.

In this paper, we propose a novel semantic map based image compression framework (SMIC), focusing on the low bit rate, as shown in Fig. 1. The compression framework consists of two parts: encoder and decoder modules. The semantic map contains the category information and location information of the original image, which is important for understanding the content of the image. In addition, a semantic map can be compressed to very low bit rate, requiring little storage space. In the encoder module, firstly, we propose a semantic segmentation network (SS-Net) for extracting semantic maps from the given input images. The extracted semantic maps are encoded as the first part of the bit stream. Then, the input images are down-sampled to obtain low-resolution images, which are losslessly encoded into the second part of the bit stream. Two parts of the bit stream are transmitted to the decoder through the channel. For the decoder module, the two parts of the bit stream are respectively decoded to the semantic maps and the low-resolution images by the corresponding decoder. The decoded low-resolution images are up-sampled to obtain the original resolution, which together with the decoded semantic maps for reconstructing the original image.

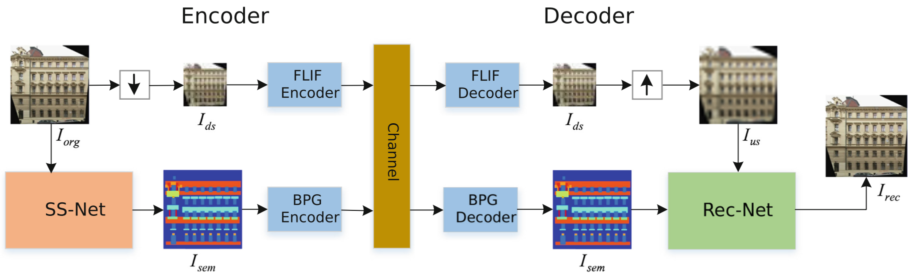


Fig. 1. The overall framework of our proposed SMIC.

Finally, we propose an image reconstruction network (Rec-Net) to obtain high-quality results by the decoded semantic maps and up-sampled images. We validate the proposed approach and compare our performance against the traditional compression algorithms including JPEG and JPEG2000. Experimental results show that our proposed image compression framework can yield visually more appealing results at low bit rate.

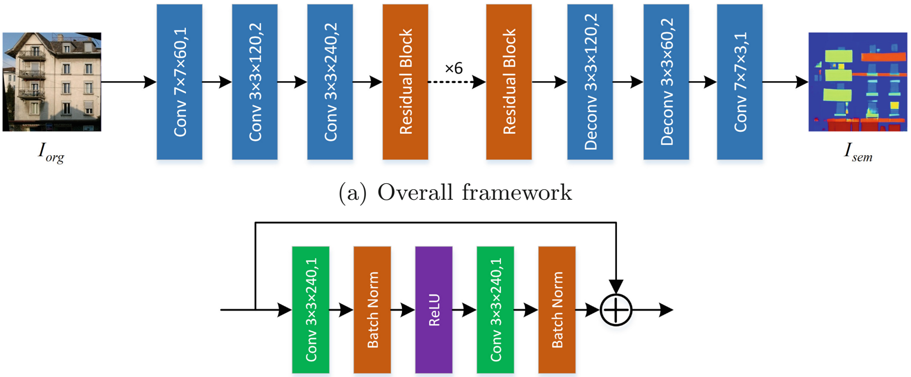
The remainder of this paper is organized as follows. Section 2 introduces the proposed SMIC in detail. The experimental results are demonstrated in Sect. 3. The conclusion of this paper is presented in Sect. 4.

2 Proposed Method

2.1 Encoder Framework

The overall image compression framework is shown in Fig. 1, which includes two parts: encoder and decoder. In order to extract the semantic maps, we propose a semantic segmentation network (SS-Net) based on conditional generative adversarial network, as shown in Fig. 2. The input image is first down-sampled to obtain a low-resolution image I_{ds}, which is losslessly encoded using the FLIF codec [12], which is state-of-the-art in lossless image codec. Then, the semantic map of the input image is extracted by our proposed SS-Net, which is encoded by a lossy BPG codec [13]. The BPG codec is based on the H.265/HEVC standard technology, which is a state-of-the-art lossy image codec. Two parts of the bit stream are transmitted to the decoder through the channel.

Our SS-Net model is based on the architecture of conditional GAN [14] and consists of two networks: generator and discriminator, which are alternately trained to compete with each other. The task of the generator of SS-Net is to extract features from the input image to generate a corresponding semantic map. The task of the discriminator is to determine whether the input image is from real or fake semantic map. By training the generator and discriminator alternately, we can improve the performance of the generator, generating an indistinguishable semantic map. For the SS-Net, the architecture of the generator is illustrated in Fig. 2(a), which consists of three parts: the encoder, the residual



(a) Overall framework

(b) Residual block

Fig. 2. The generator of our proposed SS-Net.

blocks and the decoder. In the encoder part, there is a 7×7 convolution layer which outputs a 60-channeled feature map. And then two 3×3 convolution layers are performed to extract high-dimensional features. In the residual blocks part, we use 6 residual blocks, which are designed to learn the mapping from the encoded features to corresponding semantic information. The residual block is shown in Fig. 2(b). Each residual block make small changes to the input feature map to make it better, and the last residual block can generate good enough feature maps. Finally, the decoder part consists of two 3×3 convolution layers and a 7×7 convolution layer. The 3×3 convolution layers are performed to up-sample the feature maps to ensure that the output size is the same with the input. Then, the feature maps pass through a 7×7 convolution layer, and finally output a semantic map.

The architecture of discriminator D_1 is illustrated in Table 1. For the discriminator D_1, two pairs of image are required as input. The input image is concatenated with the ground truth semantic map as the input of 'real' discriminator. Meanwhile, the input image is concatenated with the generated semantic map as the input of 'fake' discriminator. The concatenated results are fed through 5 convolution layers, producing a feature map that each pixel represents a classification result of the image patch. Finally, the discriminator tries to determine if each image patch is 'real' or 'fake'. Such a discriminator can run faster because it focuses on the image patches but not the entire image.

2.2 Decoder Framework

Here, we introduce the decoder module of our image compression framework, the decoder framework is shown in the right part of Fig. 1, which includes a deep learning based network Rec-Net. At the decoder, the semantic map and low-resolution image are decoded by the corresponding codec respectively. The low-resolution image is first up-sampled, together with the semantic map as

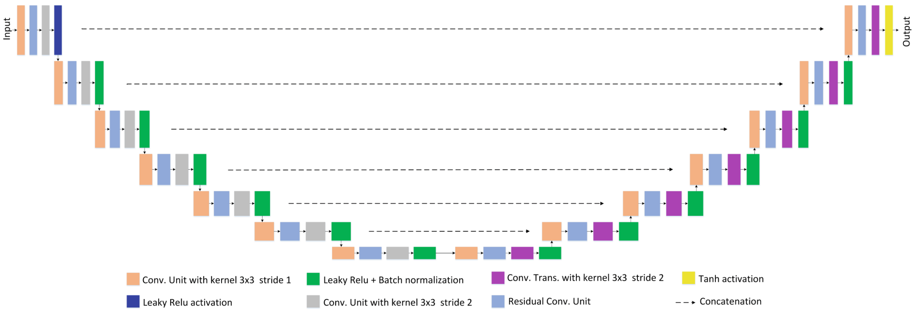

Input
Output
Conv. Unit with kernel 3x3 stride 1
Leaky Relu + Batch normalization
Conv. Trans. with kernel 3x3 stride 2
Tanh activation
Leaky Relu activation
Conv. Unit with kernel 3x3 stride 2
Residual Conv. Unit
Concatenation

Fig. 3. The generator of our proposed Rec-Net.

the input of Rec-Net. Although GAN-based network can synthesize an appealing image using only a semantic map, which is quite different from the original image in details. In order to reduce the difference between the synthesized image and the original image, we propose an image reconstruction network (Rec-Net) conditioned on the up-sampled image and corresponding semantic map. By adding the up-sampled image, the Rec-Net can be easy to generate a high-quality image, which is indistinguishable from the original image. By training to learn the difference between the reconstructed image and original image, our model can reconstruct a high perceptual quality image.

Our Rec-Net is also based on conditional GAN architecture and consists of a generator and a discriminator. The generator of Rec-Net is based on the architecture of basic U-Net [15], as shown in Fig. 3. We select the architecture of U-Net as our generator due to its simplicity and effectiveness for many image tasks. Basically, U-Net is fully convolution network, which includes a series of down-sampling layers followed by a series of up-sampling layers. The feature maps are cropped and copied from down-sampling layers to up-sampling layers. To keep the spatial size of the output same with the input, we modify the padding scheme in our Rec-Net. We also remove the cropping and copying unit from the basic U-Net model and use concatenation operation. We add a residual block in each layer of the generator of Rec-Net to learn the semantic information and low-frequency information, yielding an improved architecture that results in better performance. The residual block is shown in Fig. 2(b). As shown in Fig. 3, the network consists of two main parts: the encoding and decoding units. The convolution layers with kernel size 3×3, stride 1 are designed to extract more feature information in each unit. By the adversarial training, the residual block can learn the feature mapping relations from the input to the original image. The convolution operations are performed followed by Relu activation and Batch Normalization (BN) in both parts of the network, except that the first and the last one. We use the skip connections to concatenate feature maps from the encoding unit to the decoding unit. The skip connection has a benefit that gradients can flow from the higher layers to the lower layers, which can improve the performance of the generator and make the training process easier.

Table 1. The discriminator architecture.

Layer	D_1	D_2
Conv 1	$4 \times 4 \times 64$, s = 2, relu	$4 \times 4 \times 64$, s = 2, relu
Conv 2	$4 \times 4 \times 128$, s = 2, relu	$4 \times 4 \times 128$, s = 2, relu
Conv 3	$4 \times 4 \times 256$, s = 2, relu	$4 \times 4 \times 256$, s = 2, relu
Conv 4	$4 \times 4 \times 256$, s = 1, relu	$4 \times 4 \times 512$, s = 2, relu
Conv 5	$4 \times 4 \times 1$, s = 1, sigmoid	$4 \times 4 \times 512$, s = 2, relu
Conv 6	–	$4 \times 4 \times 512$, s = 1, relu
Conv 7	–	$4 \times 4 \times 1$, s = 1, sigmoid

For the discriminator D_2, we use an architecture similar to the discriminator D_1, adding two convolution layers, as shown in Table 1. Two pairs of image include the input and the original image, the input and the reconstructed image. The concatenated results are fed through 7 convolution layers, producing a feature map that each pixel represents a classification result of the image patch. Finally, the discriminator tries to determine if each small image patch is real or fake, allowing the generator to reconstruct an image with better details.

2.3 Loss Function

The loss function for our generator consists of the L_1 loss, the adversarial loss and the perpetual loss. For the task of image reconstruction, the generator can reconstruct the image closer to the original image in pixel-wise. The L_1 loss function can be formulated as:

$$L_1 = \lambda \frac{1}{N} \sum_{i=1}^{N} \| I_{GT} - I_{out} \|_1 \tag{1}$$

where I_{GT} represents the ground truth image, I_{out} is the output image by our generator and N is the total number of image elements.

For the adversarial loss, we use the regular loss form in [16]. The adversarial loss can encourage the generator to generate a high-quality image with more photo-realistic details. The conditional GAN trains the generator G and the discriminator D by alternatively minimizing L_{adv}^{G} and maximizing L_{adv}^{D}, which are defined as follows:

$$L_{adv}^{G} = E[\ log(1 - D(G(I_{in}), I_{in}))] \tag{2}$$

$$L_{adv}^{D} = E[logD(I_{GT}, I_{in})] + E[\ log(1 - D(G(I_{in}), I_{in}))] \tag{3}$$

where I_{GT} and I_{in} denote the ground truth image and the input image of the generator, respectively. We minimize $-log(D(G(I_{in}), I_{in}))$ instead of $log(1 - (D(G(I_{in}), I_{in})))$ for the generator, which can have a better gradient behavior.

For the semantic segmentation network SS-Net, the final loss for the generator can be represented as:

$$L_{SS-Net} = L_{adv}^{G} + L_1 \tag{4}$$

In order to improve the perceptual quality of the reconstructed image, our also use a perpetual feature-matching loss based on the VGG networks [17], named VGG loss. The VGG loss is based on the ReLU activation layers of the pre-trained 19 layers VGG network, which can be defined as:

$$L_{VGG/i,j} = \lambda \frac{1}{W_{i,j}H_{i,j}} \sum_{x=1}^{W_{i,j}} \sum_{y=1}^{H_{i,j}} \| F^{(i,j)}(I_{GT})_{x,y} - F^{(i,j)}(I_{rec})_{x,y} \|_2 \tag{5}$$

where I_{GT} and I_{rec} represents the ground truth image and the reconstructed image. $F^{(i,j)}$ denotes the feature map obtained by the j-th convolution before the i-th max-pooling layer in the VGG network. $W_{i,j}$ and $H_{i,j}$ represent the dimensions of the feature maps in the VGG network.

For the image reconstruction network Rec-Net, the final generator loss can be formulated as:

$$L_{Rec-Net} = L_{adv}^{G} + L_1 + L_{VGG} \tag{6}$$

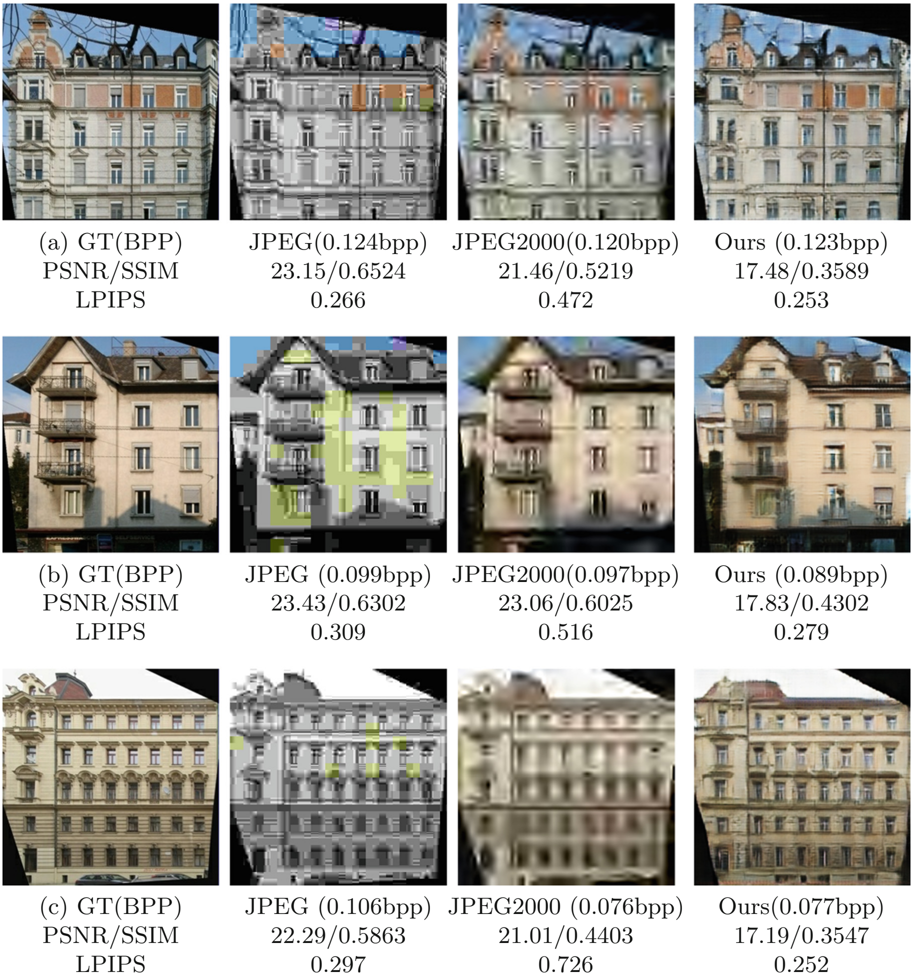
3 Experimental Results

3.1 Implementation Details

Our model is trained in a supervised fashion on pairs of images and semantic maps. In this paper, we use the CMP Facades dataset [18], which consists of just 400 images for training. We use the validation set for testing, which consists of 100 images. We sample the original images to 256 × 256 resolution and scale the range of the images to $[-1, 1\ ]$ for our experiments. We encode the down-sampling images using BPG codec with different sampling factors. We use the VGG loss $L_{VGG/5,4}$, which is defined on feature maps of higher level features from deeper VGG network layers, yielding better texture details. We consider the weight $\lambda = 100$ for L_1 and L_{VGG}. For the architecture of Rec-Net and two discriminators, all Relus are leaky with slope 0.2. In our experiment, we use the Adam [19] optimizer with a mini-batch size 1 and a momentum parameter 0.9 for training. The learning rate is fixed at 0.0002. We train the SS-Net and Rec-Net model for 200 epochs.

3.2 Perceptual Results

In this work, the ultimate goal of our work is not to achieve the best objective evaluation results, but instead to generate a restoration image with high perceptual quality. The traditional metrics used to evaluate the reconstructed image are PSNR and SSIM, both of which have been found to correlate poorly with human assessment of visual quality. At the extreme low bit rate, it becomes impossible to preserve the full image content. Because the PSNR and SSIM favor exact preservation of local structure (high-entropy), they are meaningless to evaluate the reconstructed images. We use a recently developed image quality assessment metric employing deep feature for measuring the perceptual quality, termed LPIPS [20], which tries to measure the perceptual similarity between two images. In Fig. 4, the perceptual results of our experiments are shown, compared with JPEG, JPEG2000 at low bit rate. Our results achieves better perceptual similarity scores than JPEG and JPEG2000.



(a) GT(BPP)
PSNR/SSIM
LPIPS

JPEG(0.124bpp)
23.15/0.6524
0.266

JPEG2000(0.120bpp)
21.46/0.5219
0.472

Ours (0.123bpp)
17.48/0.3589
0.253

(b) GT(BPP)
PSNR/SSIM
LPIPS

JPEG (0.099bpp)
23.43/0.6302
0.309

JPEG2000(0.097bpp)
23.06/0.6025
0.516

Ours (0.089bpp)
17.83/0.4302
0.279

(c) GT(BPP)
PSNR/SSIM
LPIPS

JPEG (0.106bpp)
22.29/0.5863
0.297

JPEG2000 (0.076bpp)
21.01/0.4403
0.726

Ours(0.077bpp)
17.19/0.3547
0.252

Fig. 4. Subjective comparison on several images compressed by JPEG, JPEG2000 and our method. Corresponding BPP (bits/pixel/channel), PSNR(dB), SSIM and LPIPS score (lower score is better) are shown in bottom.

As shown in Fig. 4(a), it can be found that there are some blocking artifacts and color distortion in JPEG images compressed at low bitrate. And there are some blurring artifacts in JPEG2000 images, which can not exhibit a good subjective quality. However, our method can produce very good details in reconstructed images and keep the edges sharper, which make the whole image perceptually pleasing.

As the bit rate decreases, we can see that the JPEG image has more serious blocking artifacts and color distortion, and it also has serious blurring artifacts for image reconstructed by JPEG2000. Due to the limitation of bitrate, the traditional methods can recover some the low-frequency information, but the recovery of high-frequency information is very difficult, which leads to the serious degradation of recovered image quality. As shown in Fig. 4(b), it can be observed that the other methods recover results with noticeable color distortion and artifacts such as blocking and blurring artifacts at low bitrate. Compared to other methods, our method effectively suppresses such artifacts and distortion through the semantic information and the robust perceptual loss function, generating an image with high perceptual quality.

When the bitrate is about 0.07bpp, our approach can still restore an image with high perceptual quality than the comparison methods. As shown in Fig. 4(c), we can see that the performance of JPEG and JPEG2000 has serious distortion, whereas our method can recover high-quality images with much cleaner and sharper details. In contrast, our method does a good performance in the perceptual results, reconstructing much more visually pleasant high-quality images.

4 Conclusion

In this paper, we propose a novel semantic map based image compression framework (SMIC) for image compression at low bit rate. Firstly, we propose a semantic segmentation network (SS-Net) to extract the semantic map from the input image. The semantic map and the down-sampled image of the input image are encoded into the bit stream respectively. Then we propose an image reconstruction network (Rec-Net) conditioned on the decoded semantic map and the up-sampled image of the input image, yielding more perceptually pleasing image at low bit rate. Contrast to the traditional compression codecs, our method can achieve good performance in perceptual quality. According to experimental results, our proposed method can reconstruct many perceptual details and generate sharp edges comparing with traditional methods.

Acknowledgment. This work was supported by Fundamental Research Funds for the Central Universities (2019JBZ102).

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE TPAMI **39**(12), 2481–2495 (2017)

2. Zhou, Q., Zheng, B., Zhu, W., Latecki, L.J.: Multi-scale context for scene labeling via flexible segmentation graph. Pattern Recogn. **59**, 312–324 (2016)
3. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE TPAMI **40**(4), 834–848 (2016)
4. Zhou, Q., et al.: Multi-scale deep context convolutional neural networks for semantic segmentation. World Wide Web-Internet Web Inf. Syst. **22**(2), 555–570 (2019)
5. Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: IEEE CVPR, pp. 5967–5976 (2017)
6. Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired Image-to-image translation using cycle-consistent adversarial networks. In: IEEE ICCV, pp. 2380–7504 (2017)
7. Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional GANs. In: IEEE CVPR (2018)
8. Ballé, J., Laparra, V., Simoncelli, E.P.: End-to-end optimized image compression. In: ICLR (2016)
9. Rippel, O., Bourdev, L.: Real-time adaptive image compression. In: IEEE ICML, vol. 70, pp. 2922–2930 (2017)
10. Ballé, J., Minnen, D., Singh, S., Hwang, S.J., Johnston, N.: Variational image compression with a scale hyperprior. In: ICLR (2018)
11. Santurkar, S., Budden, D., Shavit, N.: Generative compression. arXiv preprint arXiv:1703.01467 (2017)
12. Sneyers, J., Wuille, P.: FLIF: free lossless image format based on maniac compression. In: IEEE ICIP, pp. 66–70 (2016)
13. Bellard, F.: BPG image format (2017). http://bellard.org/bpg
14. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.178 (2014)
15. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
16. Goodfellow, I., et al.: Generative adversarial nets. In: NIPS, pp. 2672–2680 (2014)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
18. Tyleček, R., Šára, R.: Spatial pattern templates for recognition of objects with regular structure. In: Weickert, J., Hein, M., Schiele, B. (eds.) GCPR 2013. LNCS, vol. 8142, pp. 364–374. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40602-7_39
19. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
20. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)