Image Compression with Encoder-Decoder Matched Semantic Segmentation

Trinh Man Hoang[1], Jinjia Zhou[1,2], and Yibo Fan[3]

[1]Graduate School of Science and Engineering, Hosei University, Tokyo, Japan
[2]JST, PRESTO, Tokyo, Japan
[3]Fudan University, Shanghai, China

Abstract

In recent years, the layered image compression is demonstrated to be a promising direction, which encodes a compact representation of the input image and apply an up-sampling network to reconstruct the image. To further improve the quality of the reconstructed image, some works transmit the semantic segment together with the compressed image data. Consequently, the compression ratio is also decreased because extra bits are required for transmitting the semantic segment. To solve this problem, we propose a new layered image compression framework with encoder-decoder matched semantic segmentation (EDMS). And then, followed by the semantic segmentation, a special convolution neural network is used to enhance the inaccurate semantic segment. As a result, the accurate semantic segment can be obtained in the decoder without requiring extra bits. The experimental results show that the proposed EDMS framework can get up to 35.31% BD-rate reduction over the HEVC-based (BPG) codec, 5% bitrate and 24% encoding time saving compare to the state-of-the-art semantic-based image codec.

1. Introduction

The typical lossy image encoding standards such as JPEG, JPEG2000 or HEVC-based BPG codec[2] are mostly processed based on block-wise transformation and quantization. In the case of limited transmission bandwidth, the large quantization parameter is usually assigned to achieve low bit-rate coding. However, it also leads to extreme blurring and block-type artifacts.

Many works of the deep learning-based image compression, which could outperform the traditional approach, were replied on the use of additional information. This additional information is varied from semantic information [4][9], attention map[8] to a low-dimension version[5] of the image. Recently, thanks to the rapid evolution of the semantic segmentation technique, several techniques have achieved high performance whose results can be used in other tasks[16].

The state-of-the-art semantic-based image compression framework – DSSLIC[4] sent the down-sampled image and the semantic segment to the decoder. DSSILC then up-sampled the compact image and used the GAN-based[7] image synthesis technique[14] to reconstruct the image using the sending semantic segment and the up-sampled version. With the help of the semantic segment, DSSLIC could reconstruct better image quality than all the existed traditional image compression methods. However, DSSLIC requires the encoder to send extra bits of the semantic segment extracted from the original image to the decoder under a lossless compression.

To address this issue, we propose a new layered image compression framework with encoder-decoder matched semantic segmentation (EDMS). A semantic segmentation network is applied to the up-sampled image in both encoder and decoder. But the semantic segment extracted from the up-sampled image is not as accurate as of that from the original image. To obtain this quality gap, a convolution neural network (CNN) with a special structure is further applied to non-linear map the extracted segment to its original distribution. Experimental results show that our approach can get better performance than the state-of-the-art segmentation-based image compression.

2. Proposed image compression framework: EDMS

We propose a new layered image compression framework with encoder-decoder matched semantic segmentation (EDMS). Based on EDMS, the decoder can use the semantic segment to enhance the quality of the reconstructed image without any extra bits.

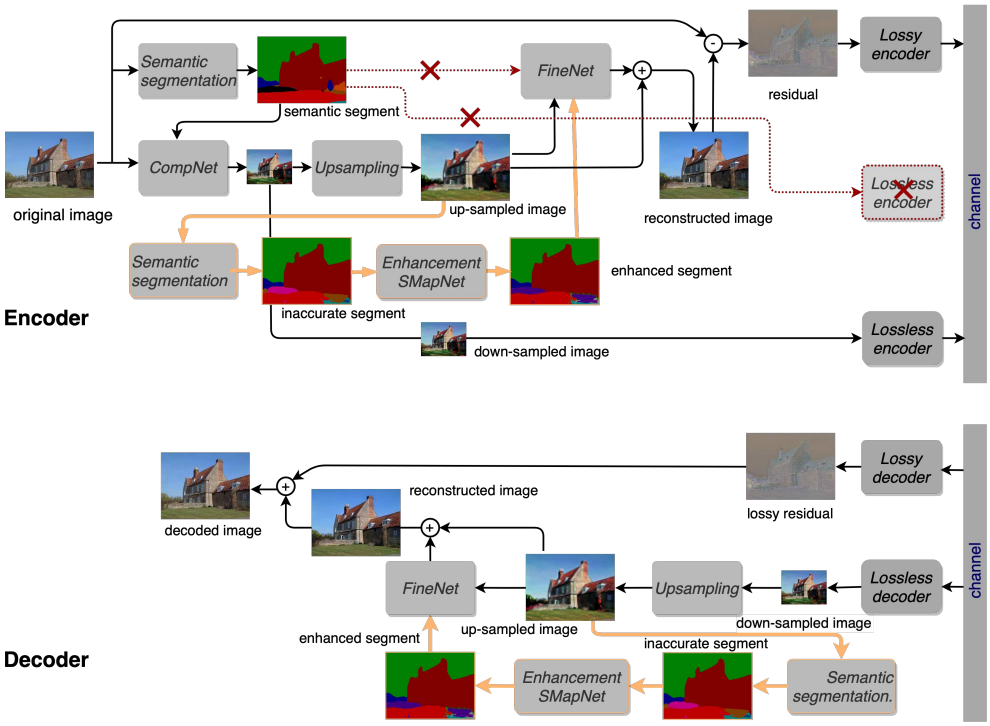

Figure 1. Our proposed framework - EDMS with extra branch for segmentation enhancement.

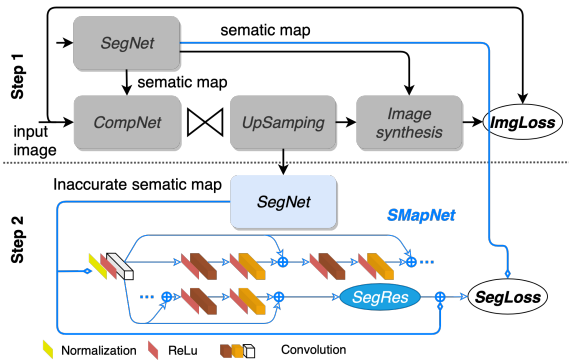

Figure 2. The specific training procedure.

2.1. Encoder-decoder matched approach

The layered image compression systems usually encode and send the compact version of the original images to the decoder. In the decoder side, a super-resolution neural network is applied to reconstruct the compacted image. To improve the quality of the reconstructed image, the semantic segment is also sent as a piece of evidence for the reconstruction task. Because of that, a noticeable number of bits are used to store this evidence.

There will be several problems if we discard the semantic segmentation information and then perform it only at the decoder-end. Firstly, the quality of the reconstructed image will not be good since the received residual was conducted based on the original semantic segment. Recently, Akbari et al[4] introduced a CompNet which used the semantic segment as side information to perform the downsampling. Since this compact version is conducted based on the semantic segment and usually lossless sent to the decoder, it is suitable to perform the semantic segmentation on this version of image. Besides, a GAN-based FineNet[14] was also demonstrated that it can generate a synthesis image from the up-sampled version and the semantic segment. Therefore, we chose the up-sampled version in the encoder to perform the semantic segmentation (see Figure 1), then replace it as the input for the image synthesis network instead of the original semantic segment. The new synthesis image is utilized to calculate the residual. Because this process can be repeated at the decoder, there is a correlation between the residual and the synthesis image without sending the semantic segment.

Secondly, the decoded image usually contains noise artifacts from the lossy compression, performing the segmentation directly on the decoded image will lead to inaccurate boundary decisions. Therefore, we need to map the deformation semantic to its true distribution. By wondering about the special type of mapping information, the semantic segment is much simpler than a general image. It means that it will be easy to fall into the overfitting situation or gradient exploding when training. Hence, we applied the Recursive

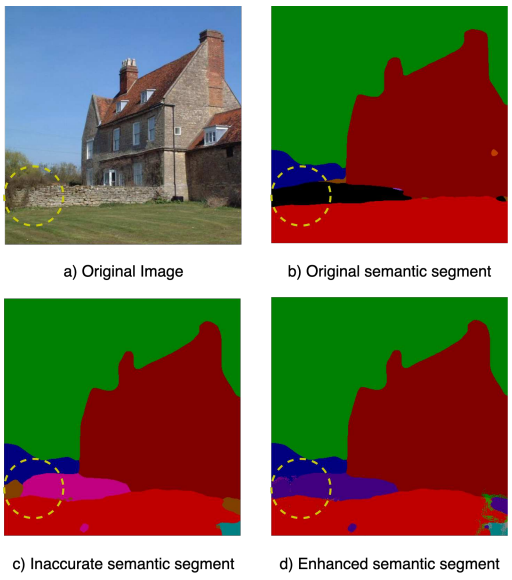

Figure 3. Our SMapNet performance in the semantic enhancement task. The inaccurate semantic segment is extracted from the up-sampled image.

Residual architecture [12] as a mapping operator for this type of semantic information - SMapNet. This architecture is demonstrated as a strong design again the overfitting issue and more stable for training than normal recurrent CNN.

2.2. Overall framework

Figure 1 shows our overall framework, on the encoder side, we extract the segment from the up-sampled version and use the SMapNet for semantic segmentation enhancement and input the SMapNet's output into the position of the semantic segment in the FineNet. The final residual will be calculated based on the output of FineNet forward with SMapNet segment as its input (see Figure 1). This residual then will be encoded by BPG[2] (state-of-the-art traditional lossy image codec), lossless FLIF codec[11] is applied for the compact version of the image and there is no extra bit used for transferring the semantic segment.

On the decoded side, we received only the down-sampled image and lossy residual from the channel. The semantic segment used to reconstruct the decoded image is conducted from the up-sampled image and enhanced by our SMapNet. Next, FineNet uses this enhanced segment and the up-sampled image as its input to perform the reconstruction. Since we also performed this process on the encoder side, there always is a correlation between the received residual and this reconstructed image. The reconstructed image is then sum up with the residual to output the final decoded image.

There are three main networks in our framework: CompNet, FineNet (leverage from [4]) and SMapNet (proposed in this work). Figure 2 shows the architecture of our SMapNet with our specific training procedure. Please refer to Section 1 of our supplementary document for more details of the training procedure and the network architecture.

2.3. Experiment settings

3. Experimental results

Our experiments were conducted on an NVIDIA Tesla V100 GPU while an Intel Core i7-8700K CPU was used to perform non-GPU tasks. We trained our models on the ADE20K dataset[1]. ADE20K test set and Kodak[3] dataset are used as testing sets in our experiments. The results are recorded by average value over all test images. We use PSNR, MS-SSIM[15] and Bjontegaard-delta (BD)-rate[6] metrics to evaluate our results.

3.1. Evaluating the overall performance

Performance gain and processing speed. As shown in Table 1, we first compare the performance of our codec with

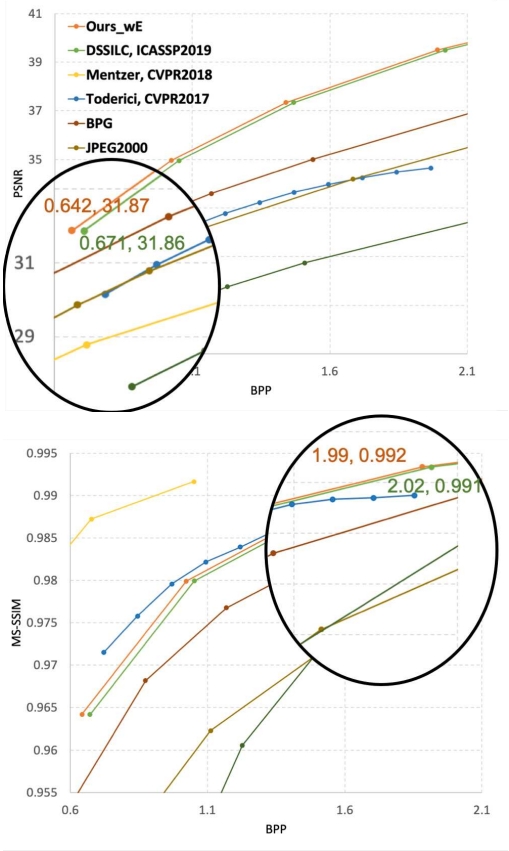

Figure 4. The comparison of compression techniques using PSNR(above) and MS-SSIM(below) on the Kodak test set with different training dataset

Dataset	DSSLIC[4]			Ours: w/o Sematic Enhancement			Ours: w/ Sematic Enhancement		
	bpp	PSNR (dB)/ MS-SSIM	Enc./Dec. Time (s)	bpp	PSNR (dB)/ MS-SSIM	Enc./Dec. Time (s)	bpp	PSNR (dB)/ MS-SSIM	Enc./Dec. Time (s)
ADE20K	0.762	33.57 / 0.977	0.838/0.315	0.75	33.33 / 0.974	0.579 / 0.247	**0.726**	33.57 / 0.977	**0.709** / 0.377
Kodak	0.671	31.86 / 0.967	1.08/0.342	0.657	31.77 / 0.959	0.703 / 0.264	**0.642**	31.87 / 0.964	**0.745** / 0.305

Table 1. Comparison of the compression quality and processing time (under the same BPG quantized parameter – QP = 32).

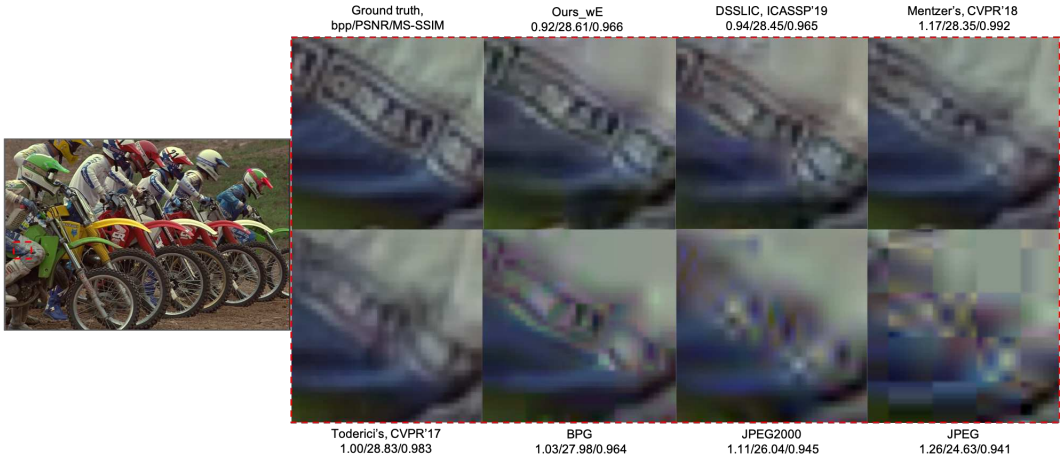

Figure 5. Qualitive comparison between different compression codec by bpp/PSNR/MS-SSIM. Note that our proposed method gets the best decoded quality with the smallest bitrate.

DSSLIC codec[4], which is state-of-the-art in semantic-based image compression. For a fair comparison, our residual is compressed by the BPG codec with the same QPs as the DSSLIC codec. Note that, in all figures and following discussion, Ours_woE and Ours_wE represent our approaches with and without applying SMapNet respectively. As shown in Table 1, our method – Ours_wE gets the same PSNR and MS-SSIM with DSSLIC with a lower 5% bitrate (bits per pixel) while reducing 24% encoding time and almost unchanged the decoding time.

Semantic segment enhancement. To demonstrate the effect of enhancing the semantic segmentation, we record the results without using it – Ours: w/o Semantic Enhancement (see Table 1). Since the up-sampled image is lack of details, the inaccurate semantic segment does not perform very well. Figure 3 shows the enhancing effect of our SMapNet, we can see clearly a wrong split wall in the raw segment has been connected again by SMapNet.

Generalization capability. We further test our ADE20K-trained model on the Kodak dataset. We compare our method with some learning-based image codecs like DSSLIC, Mentzer *et al.*'s[10] and Toderici *et al.*'s[13] and several traditional methods like BPG[2], JPEG2000 and JPEG. The RD-curve are shown in Figure 4. From Figure 4, we can observe that our method still achieves the upper-bound in the consideration of PSNR. In particular, our method gains 35.31% BD-rate reduction over BPG. This result demonstrates that our method generalizes well when the training and testing images are from different distributions. Note that, for Mentzer *et al.*'s and Toderici *et al.*'s works, since the provided models were designed and trained by MS-SSIM loss, it is easy to understand their poor performance on PSNR and the opposite results on MS-SSIM.

Subjective evaluation. A visual example from the Kodak dataset is shown in Figure 5. Our method could get the best image quality with the smallest bpp. When looking into the cropped part, we could clearly see that JPEG and JPEG2000 got a lot of block artifacts and noise. While BPG, Mentzer's and Toderici's models smoothed over some parts. And DSSILC needed more bits to reconstruct an image with clearly block artifact. Please refer to Section 2 of our supplementary document for more visual results.

4. Conclusion

This paper presents a novel layered image compression framework for leveraging the semantic segment without transferring any extra bit. With our idea of encoder-decoder matched semantic segmentation (EDMS), semantic segment enhancement and specific training procedure, our model could keep the quality of decoded images while saving the bits for transferring the semantic segment. Experimental results showed that the proposed approach could

outperform all traditional codecs and gain up to 5% bitrate and 24% encoding time reduction compare to DSSILC[4], the state-of-the-art semantic-based image codec. Since there still have a lot of information can be synchronously extracted from both encoder and decoder, our approach has the potential to be applied to other future work.

Acknowledgement

This work is supported by JST, PRESTO Grant Number JPMJPR1757 Japan.

References

[1] Ade20k dataset. https://groups.csail.mit.edu/vision/datasets/ADE20K/, 11 2019. 3

[2] Bpg image format. https://bellard.org/bpg/, 11 2019. 1, 3, 4

[3] True color kodak images. http://r0k.us/graphics/kodak/, 3 2020. 3

[4] Mohammad Akbari, Jie Liang, and Jingning Han. Dsslic: Deep semantic segmentation-based layered image compression. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2042–2046. IEEE, 2019. 1, 2, 3, 4, 5

[5] Pinar Akyazi and Touradj Ebrahimi. Learning-based image compression using convolutional autoencoder and wavelet decomposition. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, number CONF, 2019. 1

[6] Gisle Bjontegaard. Calculation of average psnr differences between rd-curves. *VCEG-M33*, 2001. 3

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 1

[8] Mu Li, Wangmeng Zuo, Shuhang Gu, Debin Zhao, and David Zhang. Learning convolutional networks for content-weighted image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3214–3223, 2018. 1

[9] Sihui Luo, Yezhou Yang, Yanling Yin, Chengchao Shen, Ya Zhao, and Mingli Song. Deepsic: Deep semantic image compression. In *International Conference on Neural Information Processing*, pages 96–106. Springer, 2018. 1

[10] Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. Conditional probability models for deep image compression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4394–4402, 2018. 4

[11] Jon Sneyers and Pieter Wuille. Flif: Free lossless image format based on maniac compression. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 66–70. IEEE, 2016. 3

[12] Ying Tai, Jian Yang, and Xiaoming Liu. Image super-resolution via deep recursive residual network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3147–3155, 2017. 3

[13] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5306–5314, 2017. 4

[14] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 1, 2

[15] 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*, volume 2, pages 1398–1402. Ieee, 2003. 3

[16] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 1