An End-to-End Mutual Enhancement Network Toward Image Compression and Semantic Segmentation

Junru Chen[1,2], Chao Yao[3], Meiqin Liu[1,2], and Yao Zhao[1,2(✉)]

[1] Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
[2] Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University, Beijing 100044, China
{19120289,mqliu,yzhao}@bjtu.edu.cn
[3] School of Computer and Communication Engineering, University of Science and Technology Beijing, 100083 Beijing, China
yaochao@ustb.edu.cn

Abstract. Image compression is to compress image data without compromising human vision feeling. However, the information loss through the image compression process may influence the following machine vision tasks, such as object detection and semantic segmentation. How to jointly consider the human vision and the machine vision to compress images for human and machine vision tasks is still an open problem. In this paper, we provide a multi-task framework for image compression and semantic segmentation. More specifically, an end-to-end mutual enhancement network is designed to efficiently compress the given image, and simultaneously segment the semantic information. Firstly, a uniform feature learning strategy is adopted to jointly learn the features for image compression and semantic segmentation in the encoder. Moreover, a multi-scale aggregation module in the encoder is employed to enhance the semantic features. Then, by transmitting the quantified features, both the decompressed image features and the learned semantic features can be reconstructed. Finally, we decode this information for the image compression task and the semantic segmentation task. On one hand, we can utilize the decompressed semantic features to implement semantic segmentation in the decoder. On the other hand, the quality of the decompressed image can be further improved depending on the obtained semantic segmentation map. Experimental results prove that our framework is effective to simultaneously support image compression and semantic segmentation, both in the subjective and objective evaluation.

Keywords: Learning-based compression · Video Coding for Machine · Semantic segmentation

This work was supported by National Natural Science Foundation of China (61972028, 61902022) and the Fundamental Research Funds for the Central Universities (2019JBM018, FRF-TP-19-015A1).

H. Ma et al. (Eds.): PRCV 2021, LNCS 13020, pp. 623–635, 2021.
https://doi.org/10.1007/978-3-030-88007-1_51

1 Introduction

Nowadays, a large number of image/video contents are produced and transmitted to the Internet every day. Reported from Cisco in 2018, Machine-to-Machine applications will occupy the greatest usage of Internet video traffic over the next following years. Moreover, machine learning algorithms tend to handle more contents directly instead of only by human perception. It is critical to establish the information that can be processed both by machine intelligence applications and human perception. Therefore, how to support the hybrid human-machine intelligence applications within the limited bandwidth is eager to be solved.

Recently, with the rapid development of deep learning, some learning-based compression methods [1–4] have been proposed. However, these methods are justly driven by the Rate-Distortion cost serving for the human perception, not compatible with the high-level machine vision tasks. Besides, when facing big data and high-level analysis, these methods are still questionable. Therefore, to interact the data compression with the machine intelligent analytics, a new video codec called VCM (Video Coding for Machine) [5] is organized which provides compression for machine vision as well as human-machine hybrid vision.

In this paper, we propose an end-to-end mutual enhancement network toward image compression and semantic segmentation, which not only makes the compression framework to be compatible with the semantic segmentation but also achieves the mutual enhancement to each other. The encoder consists of a base network and a multi-scale aggregation module. In particular, the multi-scale aggregation module is able to enhance the semantic features by suppressing the effect of the quantization. The decoder decompresses the latent representation obtained from the compression branch and the semantic branch, and obtains the decompressed image and the semantic segmentation map respectively. Then the enhancement module is utilized to enhance the quality of the decompressed image via the obtained semantic segmentation map. Our method is able to achieve mutual enhancement for both image compression and semantic segmentation tasks. Experimental results show that the proposed method can obtain improved decompressed image and semantic segmentation map.

In summary, the contributions of this paper are as follows:

(1) We propose a unified framework that integrates image compression with semantic segmentation to achieve mutual enhancement.
(2) We design a multi-scale aggregation module to suppress the impact of quantization in the encoder, which aims to enhance semantic features.
(3) We construct a post-enhancement module to improve the quality of the decompressed images by using the decompressed semantic segmentation map in the decoder.

2 Related Works

In this section, we briefly review some related works about learning-based image/video compression, especially several works in response to VCM.

2.1 Learning-Based Compression

Recently, lots of learning-based image/video compression methods are proposed [6]. In general, these methods can be classified into two categories based on the coding architecture. The first is to design the deep embedded modules in the traditional hybrid coding framework, and the second is the end-to-end deep compression framework.

Deep embedded modules aim to design an optimal network to replace the key parts in the traditional coding framework, such as in-loop filter [7], intra-prediction [8], inter-prediction [9], entropy coding [10], transform [11] and quantization [12]. For example, [7] proposed a post-processing learning-based method to enhance the decompressed image, instead of the in-loop filter. An intra-prediction convolutional neural network (IPCNN) was proposed in [8]. [9] utilized spatial adjacent pixels and temporal display order as additional inputs of the constructed CNN model to implement the dual prediction of video streaming. In addition, [12] proposed a fast quantization strategy of HEVC based on CNN.

The end-to-end compression architecture research starts from [1], which consists of nonlinear analysis transform, uniform quantizer and nonlinear synthesis transform. Then, many end-to-end compression methods are proposed to further improve compression performance. An end-to-end trainable image compression model based on variational autoencoder [2] was designed, where a hyperprior potential representation was incorporated for efficiently capturing spatial dependencies. A context-adaptive entropy model that can be used for the RD optimization in the end-to-end compression architecture [3]. Furthermore, [4] introduced the discrete Gaussian mixture likelihood to parameterize the distribution of latent code and reduced the number of coding bits required. Some latest works have achieved higher compression efficiency than that of the VVC (Versatile Video Coding) [13] or HEVC (High Efficiency Video Coding) [14].

2.2 Video Coding for Machines

Traditional video coding frameworks are optimized for HVS (Human Visual System). However, with the development of AI technology, a great amount of image/video is being analyzed by machines. Hence, the target of image/video coding is not only optimized for human vision but also machine vision. Toward collaborative compression and intelligent analytics, a new codec called VCM is proposed as the next-generation video codec, which attempts to bridge the gap between feature coding for machine vision and video coding for human vision.

In response to VCM, some researchers try to integrate machine vision tasks with image compression as a uniform framework. In [15], a hybrid resolution coding framework based on a reference-based DCNN was proposed to jointly solve the problem of the interference between the resolution loss and the compression artifacts. Similarly, an end-to-end restoration-reconstruction deep neural network (RR-DNCNN) [16] based on degradation sensing technology was proposed to answer the degradation problem caused by compression and sub-sampling due to various artifacts brought by compression to the super-resolution task. Besides, some interesting works which try to combine image compression with

high-level machine vision tasks have attracted various of attention. A framework called DSSLIC was proposed in [17], which combines the semantic map, coarse representation of the input image, and residuals of the input image in hierarchical coding, which can obtain a good compression reconstruction image and simultaneously facilitate other compression related computer vision tasks. A semantically structured image coding (SSIC) [18] framework was designed to generate a semantically structured bitstream (SSB), where each part of the bitstream represents a specific object, which can be directly used for various visual tasks. [19] proposed a encoder-decoder architecture that makes an image compression framework to support semantic segmentation. So far, the study on the relation between suitable compressed representations and the effectiveness of machine vision algorithms has been an active and fast-growing research area, how to standardize a bitstream format to enable both image compression and machine vision tasks will be worth noticing.

3 Proposed Method

The proposed method aims to achieve mutual enhancement for both the image compression task and the semantic segmentation task. Figure 1 shows the framework of our method which basically is an encoder-decoder structure. In the following, we will give detailed introductions.

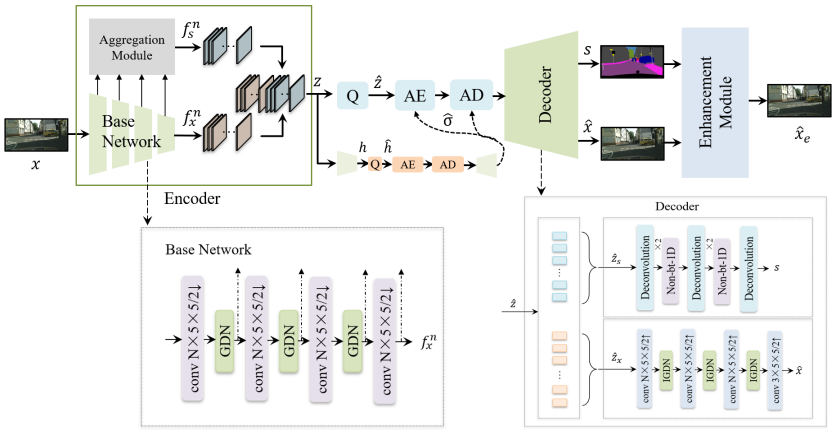


Fig. 1. The overall framework of our proposed method. "Q" denotes quantization. "AE" and "AD" mean the arithmetic encoder and decoder respectively.

3.1 Encoder

The encoder is consisted of two parts which correspond to compression branch and semantic segmentation branch respectively. One part is called base network. As shown in Fig. 1, several cascaded convolution layers are adopted to characterize the correlation between neighboring pixels, which is consistent with the hierarchical statistical properties of natural images. Here, to optimize the

features for image compression, the generalized divisive normalization (GDN) transform [1] is utilized to transfer the pixel-domain feature into a divisive normalization space.

An aggregation module is designed to learn and enhance the semantic features, which is shown in Fig. 2. It is worth noticing that all of the learned features should be quantified in our unified framework, even for the semantic features. Therefore, one key issue is to suppress the impact of the quantization. We try to explore some abundant features to enhance the semantic representations. More precisely, the hierarchical features from different layers of the base network are applied to learning the high-level semantic feature. For instance, f_x^i which is from the interlayer of the base network is added into the structure feature f_x^n by a hierarchical feature fusion block (HFFB). The operation can be represented as follows:

$$f_y^{j+1} = W_{j+1} \times f_y^j + f_x^i, i = n, n-1, ..., 1, j = 1, 2, ..., n, \tag{1}$$

where, f_x^i denotes the learned features from the i-layer of base network, f_y^i is the enhanced feature from the previous layer, while $f_y^1 = f_x^n$. W_i is the learnable parameter in the current layer.

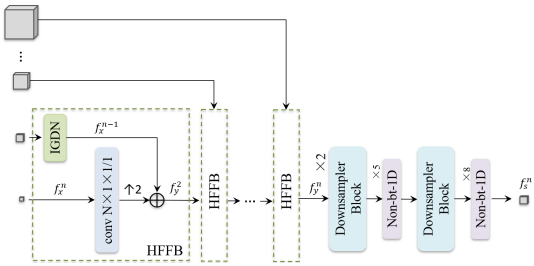


Fig. 2. The aggregation module of our proposed method, "IGDN" represents inverse GDN[20], "HFFB" represents hierarchical feature fusion block.

In the HFFB block, the feature f_x^i is first transformed to the pixel-domain by an IGDN layer associated with the GDN layer in the base network, and then the transferred feature is added to the previous fused feature f_y^i. To be noted, each HFFB block corresponds to the hierarchical features from different layers in the base network. This operation aims to suppress the additive noise by increasing the weight of the feature. To further improve the representation of semantic information, the special convolution layer non-bottleneck-1D (non-bt-1D) [21] is integrated into the HFFB blocks. Then, the features can be stretched and transformed into one-dimensional representation, which is more conducive to the subsequent pixel-level semantic classification and thus enhances the performance of the semantic segmentation task. Finally, the semantic feature f_s^n can be obtained. For the learned feature f_x^n and f_y^n, a quantization method depending on the additive noise and entropy encoding method [2] are applied to convert the learned feature into a piecewise bitstream. The bitstream is reverted to feature

by entropy decoding and sent to the decoder. It is worth mentioning that quantization operation in the traditional methods is to transform continuous data into discrete data to reduce the amount of data. So quantization operation is undesirable. However, learning-based methods depend on end-to-end optimization with gradient-based techniques. Many methods have made some contribution to solve this problem. Here, we follow [1] using additive noise. Specifically, we add uniform noise to approximate quantization operation in the training stage and we round it directly in the inference stage.

3.2 Decoder

As shown in Fig. 1, the received features are firstly divided into two parts in the decoder, including the semantic feature $\hat{z}_s$ and the compression feature $\hat{z}_x$. Correspondingly, the divided features $\hat{z}_s$ and $\hat{z}_x$ are fed into different decode branches respectively. To obtain the semantic image, several deconvolution layers and non-bottleneck-1D (non-bt-1D) layers are utilized as a semantic decoder to decompress $\hat{z}_s$. The non-bt-1D layers can gather more context from the received features, and deconvolution layers can up-sample the features to match the resolution of the input image. For the image decompression, we apply inverse operations on $\hat{z}_x$ to reconstruct image $\hat{x}$, which are corresponding to the base network in the encoder. Hence, the image decoder consists of several deconvolution layers and IGDN layers.

Considering all the factors in our framework, the loss function of the whole framework can be written as follows:

$$L = \lambda D + R + CE, \tag{2}$$

where λ is one hyper parameter, D represents the distortion between the input image and the reconstruction image, R denotes the bitrate which is approximated by using the entropy of the corresponding latent representations $\hat{z}$, CE represents the cross entropy of semantic segmentation map s and the ground truth. In general, $CE = \frac{1}{N}\sum_i - \sum_{c=1}^{M} s_{ic} log(p_{ic})$. M is the number of categories, s_{ic} has value 0 or 1. If the predicted category of the sample i is the same as the ground truth (equal to c), s_{ic} should be 1, otherwise 0. p_{ic} indicates the predicted probability that the sample i belongs to the category c.

3.3 Enhancement Module

Motivated by that semantic segmentation is able to recognize the category of each pixel, we take advantage of the semantic information to enhance the decompressed images. The semantic map where each pixel is labeled by category information can provide clearer spatial structure information for human to understand or intelligent analytics.

As shown in Fig. 3, we propose a post-enhancement module to improve the details of the decompressed image $\hat{x}$. The obtained semantic segmentation map s is fed into the post-enhancement module to learn the structure information.

First, the max pooling and the average pooling operations are separately conducted along the channel dimension, whose formulation is as follows:

$$s_s = [Max(s), Avg(s)], \tag{3}$$

here, $[\cdot, \cdot]$ represents the concatenation operation. Then, the weights of spatial structure features are obtained by a convolution layer and a sigmoid activation function. Finally, the weights are utilized to multiply by the semantic features which are learned on the semantic segmentation map s, and the output is the learned spatial structure features. The process can be represented as follows,

$$s_e = W_0 W_1 W_2 W_3 \sigma(s_s), \tag{4}$$

here, W_0, W_1, W_2, W_3 represent convolution operations and σ denotes sigmoid activation function. To embed the learned spatial structure information into the decompressed image $\hat{x}$. $\hat{x}$ is mapped to the feature space by a shadow convolutional layer. Then, some residual blocks are grouped as a frequency filter to learn high-frequency information $\hat{x}_r$. Finally, we concatenate s_e and $\hat{x}_r$ to embed the spatial structure information and obtain the final reconstruction image $\hat{x}_e$.

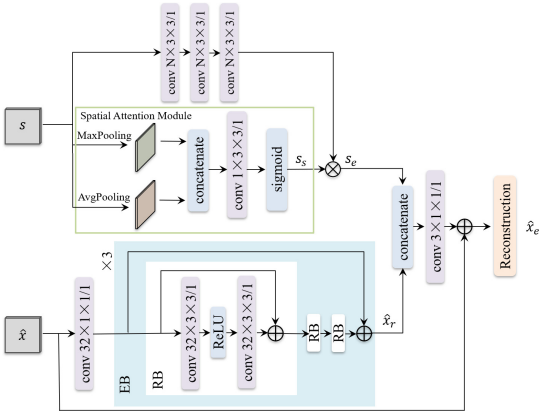


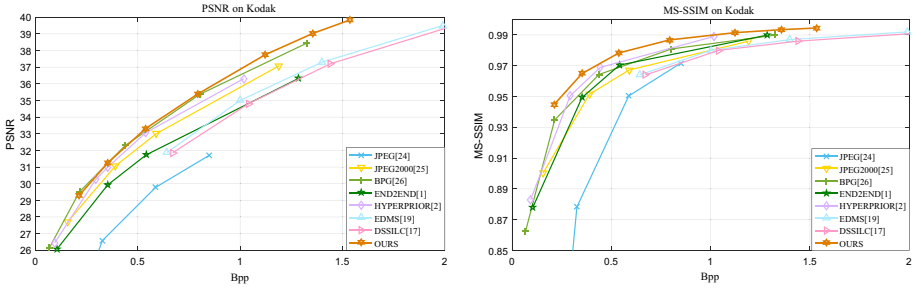
Fig. 3. The enhancement module in our proposed method. "RB" represents the residual block, "EB" refers to the enhancement block.

4 Experiments

In this section, we conduct a series of experiments to evaluate the performance of our proposed method. In our experiments, the widely used *Cityscapes* dataset is adopted. The *Cityscapes* dataset has 19 semantic labels, all 2,974 RGB images are resized to 512×1024. And the test dataset for the compression evaluation is constructed with 24 images from the *Kodak* image dataset [22]. For the semantic segmentation evaluation, we adopt the validation set and test set from *Cityscapes* dataset at the resolution of 1024×2048. The proposed framework is trained in

the end-to-end way and uses different λ values (256, 512, 1024, 2048, 4096, 6144, 8192) to control the quantization step. Adam optimizer [23] with the learning rate of 0.0001 is used, which is fixed in the first $2,000,000$ iterations but decreased to 0.00001 in the next $100,000$ iterations. All the experiments are conducted on the NVIDIA RTX 3090 with 24 GB memory.

To objectively evaluate the compression performance of our proposed method, we conduct comparable experiments with the following previous works [17,19]. Moreover, we use Multiscale Structural Similarity (MS-SSIM) and the Peak Signal to Noise Ratio (PSNR) between the original and the decompressed images as evaluation indicators. A larger MS-SSIM or PSNR means higher fidelity. Note that MS-SSIM is applied on RGB channels and averaged over the entire test set.



(a) (b)

Fig. 4. Rate-distortion curves of different image compression methods using the PSNR metric and MS-SSIM metric on *Kodak* [22].

4.1 Results on Image Compression

We compare several widely used image compression algorithms [1,2,24–26] and two hybrid compression methods [17,19] with our proposed method. The performances are shown in the Fig. 4(a) and 4(b). The curves report the PSNR and MS-SSIM at different bitrates respectively, where bpp means bits-per-pixel, referring to the averaged bitrate for each pixel.

As shown in Fig. 4(a), peak signal-to-noise ratio (PSNR) is adopted as the quality metric. It is obvious that the proposed method is better than the traditional methods JPEG [24], JPEG2000 [25] and the classical end-to-end learning-based method END2END [1], HYPERPRIOR [2]. Moverover, BPG [26] has achieved the state-of-the-art performance in traditional image compression methods, our method achieves comparable performance at low bitrates and achieves better performance apparently at high bitrates over BPG. Besides that, we also compare with the semantic-based image compression methods EDMS [19] and DSSILC [17], which are proposed recently and have excellent performance. As shown in Fig. 4(a), our method is apparently superior to both of them on PSNR quality metric.

As shown in Fig. 4(b), in order to clearly show the advantages of our method over other methods, we also carry out the experiments under multiscale structural similarity (MS-SSIM) quality metric. By comparing the curves in Fig. 4(b), it can be found that our method has the best performance of all comparable methods. Especially, the results of the proposed method have huge advantages over BPG in terms of the MS-SSIM quality metric while just can be comparable under the PSNR quality metric. Then, analyzing and comparing Fig. 4(a) and 4(b) together can be easily find out that the learning-based methods perform better than traditional methods under the MS-SSIM metric.

Moreover, the compression branch of our method has a similar structure with HYPERPRIOR [2]. When it is integrated into our method, the performance of our method is better than HYPERPRIOR. It shows that the semantic embedding method is reliable and can be used to improve the reconstruction effectively. Our enhancement module can improve decompressed image by using the semantic information extracting from semantic segmentation map.

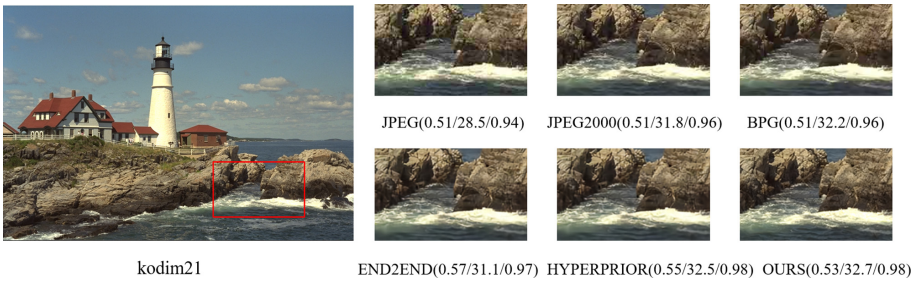


Fig. 5. Visualization of decompressed images "kodim21.png" from *Kodak* [22] and its ground truth. The numbers on the bottom of the images mean the value of (Bpp/PSNR/MS-SSIM).

To display the performance of our method intuitively, We exhibit the decompressed image of the proposed method and some competitive methods with similar bitrate in Fig. 5. A visual example of kodim21 from the *Kodak* dataset is provided. Our method obtains the best image quality at a similar bitrate. When looking carefully at the selected area, we can see that the wave of the sea in the images through JPEG and JPEG2000 methods are blurred. While the rocks in the selected area of the two methods have lots of noise and block artifacts. The more excellent traditional compression method at present BPG and the classical learning-based compression methods END2END, HYPERPRIOR are slightly better than JPEG and JPEG2000, but there are still some visible flaws. The image decompressed by our method is relatively clear in texture and color.

4.2 Results on Semantic Segmentation

In our method, the semantic segmentation branch can be compatible with many outstanding semantic segmentation networks. In this paper, ERFNet [21] is

Table 1. Results of our four semantic segmentation architectures on the *Cityscapes* Evaluation set. Per-class IoU(%) and mean classes IoU(%).

Methods	Roa	Sid	Bui	Wal	Fen	Pol	TLi	TSi	Veg	Ter	Sky	Ped	Rid	Car	Tru	Bus	Tra	Mot	Bic	Cla-IoU
Baseline	97.5	80.2	90.4	46.5	51.9	61.1	65.5	72.5	91.2	59.9	93.8	75.9	55.1	93.1	72.5	79.7	67.9	46.7	70.6	72.1
B+A	97.5	81.4	91.1	49.1	51.9	62.7	64.8	74.4	91.5	61.9	94.1	77.4	57.5	92.9	72.8	75.2	70.7	41.7	71.9	**72.8**
B+Q	97.2	80.6	90.1	51.4	52.6	59.4	63.2	72.1	91.1	61.2	92.2	75.5	55.3	92.2	66.5	75.3	63.5	44.4	70.3	71.2
B+Q+A	97.6	81.2	90.8	46.2	51.2	62.8	63.9	73.8	91.1	59.9	93.5	76.9	56.3	93.2	71.2	79.2	74.7	43.2	70.9	**72.5**

integrated into our semantic segmentation branch. Table 1 shows the segmentation results on 19 classes of the *Cityscapes* evaluation set under four conditions based on ERFNet, which are no quantization operation, only quantization operation, quantization operation plus aggregation module and only aggregation module. To conduct the segmentation experiments in four conditions, we correspondingly construct four models. We defined the original architecture of ERFNet in the semantic segmentation branch, without quantization operation and our aggregation module as baseline. Then over the architecture of baseline, only quantization is operated on the semantic segmentation branch (we called it B+Q) and only aggregation module is applied on the semantic segmentation branch (we called it B+A). The last model is our proposed model with quantization operation and aggregation module (we called it B+Q+A). As shown in Table 1, comparing baseline with B+Q, it is found that nearly 1% mean classes IoU (Cla-IoU) is declined because of quantization operation. Compared to the model B+Q, by using our aggregation module (B+Q+A), the accuracy is improved and better than baseline in the case of quantization operation. To verify the effectiveness of this aggregation module, we also compare this B+A model with the original unquantized baseline architecture. It turns out that the accuracy of our method is improved than before. Therefore, our multi-scale aggregation module is effective and the multi-scale feature information from the base network could suppress the impact of the quantization operation.

Table 2. Comparable results on *Cityscapes* Test sets.

Methods	Cla-IoU(%)	Cat-IoU(%)	Methods	Cla-IoU(%)	Cat-IoU(%)
RefineNet [27]	73.6	87.9	Dilation [28]	67.1	86.5
Adelaide-cntxt [29]	71.6	87.3	DPN [30]	66.8	86.0
LRR-4x [31]	69.7	88.2	B+A	**70.8**	**88.1**
Deeplabv2-CRF [32]	70.4	86.4	B+Q+A	**70.5**	**88.0**

Table 2 shows the semantic segmentation results of several comparable approaches. These results are obtained from the Cityscapes Test server. Baseline with the aggregation module (B+A) achieves a 70.8% mean Classes IoU (Cla-IoU) and an 88.1% mean Category IoU (Cat-IoU). The B+Q+A model achieves

a 70.5% Cla-IoU and an 88.0% Cat-IoU. Cla-IoU is improved compared to LRR-4x [31], Deeplabv2-CRF [32], Dilation10 [28] and DPN [30], and Cat-IoU is improved compared to RefineNet [27], Adelaide-cntxt [29], Deeplabv2-CRF [32], Dilation10 [28] and DPN [30]. It is proved that the aggregation module extracts hierarchical features from different layers in base network could not only reduce the impact of quantization operation but also improve the quality of semantic segmentation map. In general, benefiting from the aggregation module, the semantic segmentation branch in our proposed method is much more competitive.

5 Conclusion

To achieve mutual enhancement for image compression and semantic segmentation tasks, we propose a novel end-to-end mutual enhancement network. The whole framework of our method which is based on an encoder-decoder structure contains several creative designs. A multi-scale aggregation module in the encoder is designed to improve the accuracy of the semantic segmentation and an enhancement module after the decoder is designed to enhance the reconstruction of the compression. The experimental results show that our method is effective and achieves mutual enhancement for both image compression and semantic segmentation. In the future, we would expand this framework to support more machine intelligent tasks than semantic segmentation.

References

1. Ballé, J., Laparra, V., Simoncelli, E.P.: End-to-end optimized image compression. In: 5th International Conference on Learning Representations, ICLR 2017 (2017)
2. Ballé, J., Minnen, D., Singh, S., Hwang, S.J., Johnston, N.: Variational image compression with a scale hyperprior. In: 6th International Conference on Learning Representations, ICLR 2018 (2018)
3. Lee, J., Cho, S., Beack, S.K.: Context-adaptive entropy model for end-to-end optimized image compression. In: 6th International Conference on Learning Representations, ICLR 2018 (2018)
4. Cheng, Z., Sun, H., Takeuchi, M., Katto, J.: Learned image compression with discretized gaussian mixture likelihoods and attention modules. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7939–7948 (2020)
5. Duan, L., Liu, J., Yang, W., Huang, T., Gao, W.: Video coding for machines: a paradigm of collaborative compression and intelligent analytics. IEEE Trans. Image Process. **29**, 8680–8695 (2020)
6. Liu, D., Li, Y., Lin, J., Li, H., Wu, F.: Deep learning-based video coding: a review and a case study. ACM Comput. Surv. (CSUR) **53**(1), 1–35 (2020)
7. Lin, W., et al.: Partition-aware adaptive switching neural networks for post-processing in HEVC. IEEE Trans. Multimed. **22**(11), 2749–2763 (2019)
8. Cui, W., et al.: Convolutional neural networks based intra prediction for HEVC. In: 2017 Data Compression Conference (DCC), pp. 436–436. IEEE Computer Society (2017)

9. Mao, J., Yu, L.: Convolutional neural network based bi-prediction utilizing spatial and temporal information in video coding. IEEE Trans. Circ. Syst. Video Technol. **30**(7), 1856–1870 (2019)
10. Song, R., Liu, D., Li, H., Wu, F.: Neural network-based arithmetic coding of intra prediction modes in HEVC. In: Visual Communications and Image Processing (VCIP), pp. 1–4. IEEE (2017)
11. Liu, D., Ma, H., Xiong, Z., Wu, F.: CNN-based DCT-like transform for image compression. In: Schoeffmann, K., et al. (eds.) MMM 2018. LNCS, vol. 10705, pp. 61–72. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73600-6_6
12. Alam, M.M., Nguyen, T.D., Hagan, M.T., Chandler, D.M.: A perceptual quantization strategy for HEVC based on a convolutional neural network trained on natural images. In: Applications of Digital Image Processing, vol. 9599, p. 959918. International Society for Optics and Photonics (2015)
13. Bross, B., Chen, J., Ohm, J.R., Sullivan, G.J., Wang, Y.K.: Developments in international video coding standardization after AVC, with an overview of versatile video coding (VVC). In: Proceedings of the IEEE (2021)
14. Sullivan, G.J., Ohm, J.R., Han, W.J., Wiegand, T.: Overview of the high efficiency video coding (HEVC) standard. IEEE Trans. Circ. Syst. Video Technol. **22**(12), 1649–1668 (2012)
15. Hou, D., Zhao, Y., Ye, Y., Yang, J., Zhang, J., Wang, R.: Super-resolving compressed video in coding chain. arXiv preprint arXiv:2103.14247 (2021)
16. Ho, M.M., Zhou, J., He, G.: RR-DnCNN v2.0: enhanced restoration-reconstruction deep neural network for down-sampling-based video coding. IEEE Trans. Image Process. **30**, 1702–1715 (2021)
17. Akbari, M., Liang, J., Han, J.: DSSLIC: deep semantic segmentation-based layered image compression. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 2042–2046. IEEE (2019)
18. Sun, S., He, T., Chen, Z.: Semantic structured image coding framework for multiple intelligent applications. IEEE Trans. Circ. Syst. Video Technol. **31**(9), 3631–3642 (2020)
19. Hoang, T.M., Zhou, J., Fan, Y.: Image compression with encoder-decoder matched semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 160–161 (2020)
20. Ballé, J., Laparra, V., Simoncelli, E.P.: Density modeling of images using a generalized normalization transformation. In: 4th International Conference on Learning Representations, ICLR 2016 (2016)
21. Romera, E., Alvarez, J.M., Bergasa, L.M., Arroyo, R.: ERFNet: efficient residual factorized convnet for real-time semantic segmentation. IEEE Trans. Intell. Transp. Syst. **19**(1), 263–272 (2017)
22. Kodak, E.: Kodak lossless true color image suite (PhotoCD PCD0992), vol. 6. http://r0k.us/graphics/kodak (1993)
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
24. Wallace, G.K.: The JPEG still picture compression standard. IEEE Trans. Consum. Electron. **38**(1), 18–34 (1992)
25. Skodras, A., Christopoulos, C., Ebrahimi, T.: The JPEG 2000 still image compression standard. IEEE Signal Process. Mag. **18**(5), 36–58 (2001)
26. Bellard, F.: Better portable graphics. https://www.bellard.org/bpg (2014)
27. Lin, G., Milan, A., Shen, C., Reid, I.: RefineNet: multi-path refinement networks with identity mappings for high-resolution semantic segmentation. arXiv preprint arXiv:1611.06612

28. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)
29. Lin, G., Shen, C., Van Den Hengel, A., Reid, I.: Efficient piecewise training of deep structured models for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3194–3203 (2016)
30. Krešo, I., Čaušević, D., Krapac, J., Šegvić, S.: Convolutional scale invariance for semantic segmentation. In: Rosenhahn, B., Andres, B. (eds.) GCPR 2016. LNCS, vol. 9796, pp. 64–75. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-45886-1_6
31. Ghiasi, G., Fowlkes, C.C.: Laplacian reconstruction and refinement for semantic segmentation. arXiv preprint arXiv:1605.02264, vol. 4(4) (2016)
32. 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 Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2017)