SIGVIC: SPATIAL IMPORTANCE GUIDED VARIABLE-RATE IMAGE COMPRESSION

Jiaming Liang[1] *Meiqin Liu*[1] *Chao Yao*[2] *Chunyu Lin*[1] *Yao Zhao*[1]

[1] Institute of Information Science, Beijing Jiaotong University, Beijing 100044, China
[2] School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China

ABSTRACT

Variable-rate mechanism has improved the flexibility and efficiency of learning-based image compression that trains multiple models for different rate-distortion tradeoffs. One of the most common approaches for variable-rate is to channel-wisely or spatial-uniformly scale the internal features. However, the diversity of spatial importance is instructive for bit allocation of image compression. In this paper, we introduce a Spatial Importance Guided Variable-rate Image Compression (SigVIC), in which a spatial gating unit (SGU) is designed for adaptively learning a spatial importance mask. Then, a spatial scaling network (SSN) takes the spatial importance mask to guide the feature scaling and bit allocation for variable-rate. Moreover, to improve the quality of decoded image, Top-K shallow features are selected to refine the decoded features through a shallow feature fusion module (SFFM). Experiments show that our method outperforms other learning-based methods (whether variable-rate or not) and traditional codecs, with storage saving and high flexibility.

Index Terms— variable-rate, image compression, spatial importance, scale factor, shallow feature

1. INTRODUCTION

With the explosive demand for storing and sharing images, image compression gradually becomes a vital research field. It can reduce the required storage space and transmission cost of images within acceptable distortion. Recently, many learning-based methods [1, 2, 3, 4, 5] have outperformed the traditional codecs (e.g. JPEG[6], JPEG2000[7], BPG[8]). However, these methods rely on training multiple models with fixed rate-distortion (RD) tradeoffs, which leads to a proportionate increase in storage occupation and inflexibility in actual situation. Hence the variable-rate mechanism, which has been available in traditional codecs, is required for reducing the storage space and improving the flexibility.

Some methods have explored learning-based variable-rate image compression, which typically scale the features in encoder to obtain coarser or finer quantized features and inverse the scaling in decoder. Yang et al.[9] design a modulated autoencoder, which scales the internal features for adapting to different RD tradeoffs. Choi et al.[10] propose a similar conditional autoencoder with quantization bin-sizes for continuous bit-rates. However, improper selection of bin-size can lead to degradation of RD performance [11]. Yin et al.[12] only scale the bottleneck features, but this scaling strategy is unstable and difficult to adapt to a wide range of RD tradeoffs when training in an end-to-end manner [9]. It is noted that, all these methods only involve the channel-wise relationship, without utilizing the spatial importance of images.

When the bit-rate of encoding an image is limited, the spatial importance of the image has guiding significance for bit allocation and rate control. The works [11, 13] have attempted to utilize the spatial importance, but additional quality map need to be generated for each source image in dataset as input. Moreover, Song et al.[11] manually pre-define the uniform quality maps for image compression, the spatially different importance is not fully utilized.

In this paper, we propose a Spatial Importance Guided Variable-rate Image Compression method named SigVIC, which can adapt to arbitrary bit-rates without additional inputs. Specifically, a spatial gating unit (SGU) is designed to adaptively generate a spatial importance mask of image features. Then, a spatial scaling network (SSN) is proposed to employ the spatial importance mask to guide the generation of spatial scale factors for variable-rate mechanism. Besides, to improve the quality of reconstructed image, we propose a Top-K shallow feature fusion strategy to refine the decoded features through a shallow feature fusion module (SFFM).

We evaluate our method on Kodak and CLIC datasets in terms of MSE and MS-SSIM. Experiments illustrate that our method achieves better performance than other variable-rate methods, traditional codecs and some well-known learning-based compression methods. Specifically, by calculating the BD-Rate of MSE results on Kodak dataset, our method saves 17.84%, 10.85% and 2.78% bit-rate compared with BPG[8], Song et al.[11] and Cheng et al.[5], respectively.

2. PROPOSED METHOD

The overview of our method is depicted in Fig.1, where the encoder and decoder networks consist of several stages. In each stage t, a spatial gating unit (SGU) and a spatial scaling network (SSN) are designed to introduce spatial importance

Corresponding author: Meiqin Liu.

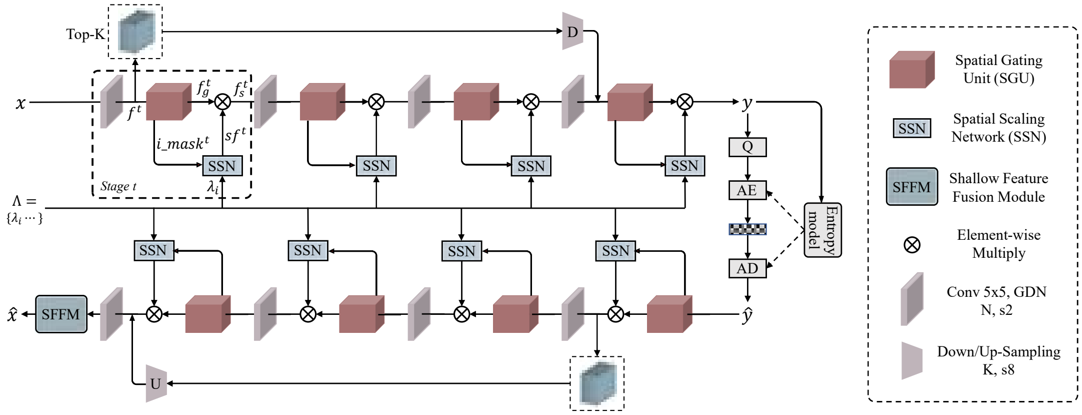

Fig. 1. The overview of our proposed spatial importance guided variable-rate image compression method, SigVIC.

mask to guide RD optimization for variable-rate. Besides, a shallow feature fusion module (SFFM) is employed at the end of decoder to refine the decoded features.

2.1. Spatial Gating Unit

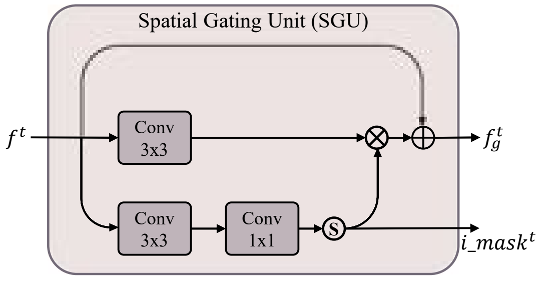

Fig. 2. The detials of spatial gating unit (SGU).

For the input features f^t at stage t, a spatial gating unit (SGU) is designed for adaptively generating a spatial importance mask i_mask^t, as shown in Fig.2. To generate i_mask^t, a 3×3 convolution is adopted to initially extract the spatial features from f^t. Then, a 1×1 convolution with a sigmoid operation is adopted to produce the weights of i_mask^t. Through the activation of sigmoid, different weights are produced in i_mask^t for different pixels of f^t, and most of them tend to be 0 or 1.

$$i\_mask^t = \text{Sigmoid}(C^1(C^3(f^t))) \tag{1}$$

By multiplying with equivalent spatial features extracted in another branch using 3×3 convolution, i_mask^t can play the role of gating unimportant features of f^t. After adding with the identity features f^t, the gated features f_g^t is obtained as:

$$f_g^t = i\_mask^t \cdot C^3(f^t) + f^t \tag{2}$$

The i_mask^t is also be passed into our spatial scaling network (SSN) to guide the generation of spatial scale factor.

2.2. Spatial Scaling Network

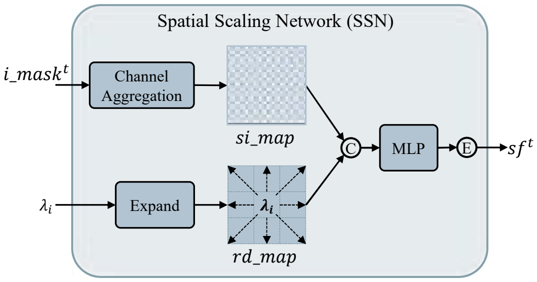

Fig. 3. The structure of spatial scaling network (SSN).

The proposed spatial scaling network (SSN) is depicted in Fig.3. To introduce the spatial importance mask for guiding feature scaling and RD optimization, SSN takes the adaptively learned i_mask^t and RD tradeoff λ_i as the joint inputs. Firstly, the spatial resolution of both inputs should be converted to the same as the gated features f_g^t. We aggregate the channel-wise information of $i\_mask^t \in \mathbb{R}^{H \times W \times N}$ and produce a spatial importance map $si\_map \in \mathbb{R}^{H \times W \times 1}$. The constant λ_i cannot be directly converted by learning, because H and W are uncertain with different size of input image. Our solution is to tile and expand λ_i into a $rd\_map \in \mathbb{R}^{H \times W \times 1}$. After that, we channel-wisely concatenate and input them into a MLP network to generate the spatial scale factor sf^t, which has been related to both spatial importance and bit-rate. The MLP consists of two full-connected layers with 64 hidden units and a ReLU in the middle. An exponential function is applied behind for positive outputs, which is beneficial for training process [9].

$$sf^t = \exp(MLP(Concat(si\_map, rd\_map))) \tag{3}$$

By spatial-wisely multiplying with sf^t, different positions of f_g^t are scaled to different fineness and the scaled feature f_s^t is obtained. In this way, the bit allocation for different

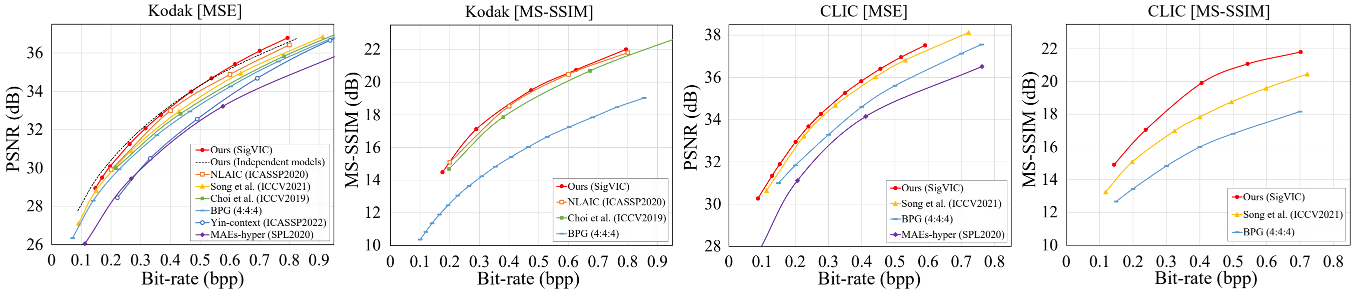

Fig. 5. The RD curves, using MSE and MS-SSIM as the distortion term of loss function, on Kodak and CLIC datasets.

bit-rates is related to spatial importance after quantization.

2.3. Top-K Shallow Feature Fusion Module

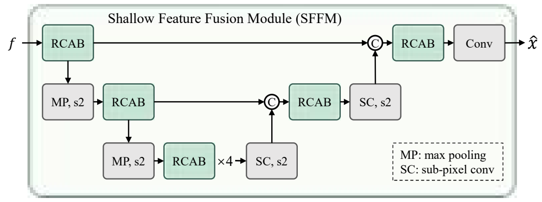

Fig. 4. Structure of shallow feature fusion module (SFFM).

To enrich the details of decoded image $\hat{x}$, Top-K shallow features that rich in edge and texture information are selected and incorporated into the encoded feature y. These features are restored in decoder to refine the decoded features through a shallow feature fusion module (SFFM). Because most of the information in features is only contributed by a few channels[14], we select the most informative K feature maps for balancing the performance and computation.

SFFM leverages a U-shaped network as shown in Fig.4. Multiple residual channel attention blocks (RCAB)[15] are utilized to provide the learning of channel-wise information. In addition, SFFM can be trained end-to-end in the whole network with only a few parameters.

2.4. Loss function

Our network receives variable λ_i for adapting to arbitrary bit-rates, while the guidance of spatial importance has been adaptively completed. The loss function is formulated as:

$$L = R(\hat{y}, \hat{z}, \lambda_i) + \lambda_i D(x, \hat{x}, \lambda_i) \;\; (\lambda_i \in \Lambda) \tag{4}$$

where Λ represents a range between two selected boundary values, containing all the possible values of λ_i.

3. EXPERIMENTS

A subset of ImageNet[16] with 13600 images is used for training our models, which are randomly cropped to patches with resolution of 256×256. The batch size is set to 8, and the models are trained for $2M$ iterations with Adam optimizer [17]. The learning rate is initially set to 1×10^{-4}, and decreased to 1×10^{-5} for the last $0.1M$ iterations. We train our models using MSE and MS-SSIM loss, where the selected boundaries of Λ are respectively (0.0016, 0.045) and (5, 120) for covering wide range of bit-rates. The number of selected shallow features K is empirically set to 32, and the number of filters N of network is set to 192.

Table 1. BD-results against BPG on Kodak and CLIC

Methods	Kodak		CLIC	
	BD-PSNR	BD-Rate	BD-PSNR	BD-Rate
MAEs-hyper[9]	-0.96 dB	21.35%	-0.59 dB	14.88%
Yin-context[12]	-0.87 dB	17.30%	--	--
Choi et al.[10]	0.22 dB	-4.96%	--	--
Lee et al.[4]	0.22 dB	-5.05%	0.56 dB	-12.25%
Song et al.[11]	0.37 dB	-8.06%	1.05 dB	-21.73%
NLAIC[18]	0.71 dB	-14.50%	--	--
Cheng et al.[5]	0.78 dB	-16.43%	1.21 dB	-25.82%
Ours (SigVIC)	**0.92 dB**	**-17.84%**	**1.23 dB**	**-26.16%**

3.1. Rate-distortion (RD) performance

We compare the RD curves of our method with other variable-rate compression methods including learning-based methods [9, 10, 11, 12, 18] and traditional codec BPG[8]. As shown in Fig.5, our method outperforms all the other methods in terms of both PSNR and MS-SSIM (values are converted to decibels by $-10log_{10}(1-\text{MS-SSIM})$ for clarity) on Kodak[19] and CLIC[20] datasets.

We further compare the BD-results(i.e.BD-PSNR and BD-Rate) with the above methods and some well-known learning-based compression methods[4, 5] that need to train multiple models for different bit-rates. Higher BD-PSNR and smaller BD-Rate indicate better RD performance, which correspond to more PSNR gains and bit-rate savings. As shown in Table.1, all the results are calculated against BPG. Obvoiusly, our method achieves the maximum PSNR gain of 0.92dB and saves the maximum bit-rate of 17.84% on Kodak dataset, which outperforms all the other methods. On the CLIC dataset with higher resolution, our method gets higher BD-PSNR of 1.23dB and saves more bit-rate of 26.16%.

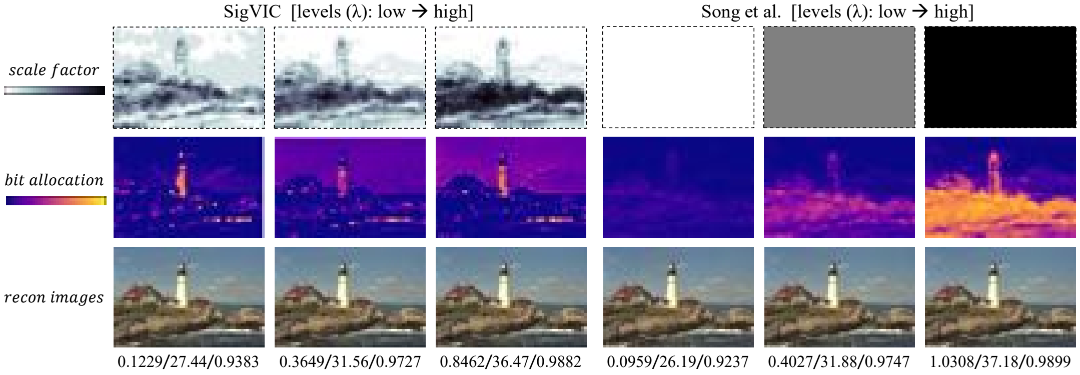

Fig. 7. The visualizations (including scale factor and bit allocation) of our SigVIC and Song et al.

Besides, our results are slightly better than Cheng et al.[5], which achieves comparable performance with VVC[21].

In order to test the robustness of our variable-rate mechanism, we also train 6 independent models without SSN. As shown in the *Kodak[MSE]* results in Fig.5, our SigVIC achieves comparable performance with corresponding independent models. This indicates that our variable-rate mechanism will not affect the performance of compression networks, while improving the efficiency and flexibility.

3.2. Subjective Performance

Fig.6 shows some reconstructed images to evaluate the subjective performance of our method. All of them are at the similar bit-rate of about 0.22 bpp. In the enlarged regions, it can be seen that our method restores more details, such as the words on helmet and pants, the texture of motorcycle tyres and mudguards, and so on.

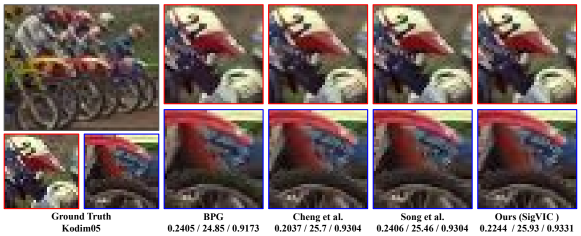

Fig. 6. Visual quality of reconstructed images.

3.3. Visualizations

To evaluate the effectiveness of our spatial importance guided feature scaling and bit allocation, we visualize the scale factors with corresponding bit allocation maps under several levels of RD tradeoff λ. As shown in Fig.7, our scale factors can adaptively learn the spatial importance of image and scale features of different regions with different weights, while the quality maps of Song et al.[11] scale different regions using equal weights. Therefore, when the bit-rate is limited, our method can preferentially allocate more bits to informative regions under the guidance of spatial importance.

3.4. Ablation Study

We study the effectiveness of SGU and SFFM in our method, the results are shown in Table.2. Scheme A adopts a uniform scale factor for variable-rate mechanism, where SSN is used but SGU is not. Compared with Scheme A, Scheme B saves 3.56% bit-rate, which benefits from the guidance of spatial importance on bit allocation and RD optimization by SGU. In addition, more bit-rate of 6.92% is saved by Scheme C. This demonstrates that the Top-K shallow feature fusion strategy with SFFM can further improve the quality of decompressed image and RD performance.

Table 2. Results of ablation study

Schemes	SSN	SGU	SFFM	BD-Rate	Params.
A	✓			0%	15.3M
B	✓	✓		-3.56%	23.21M
C	✓	✓	✓	-6.92%	24.85M

4. CONCLUSION

In this paper, we propose a Spatial Importance Guided Variable-rate Image Compression method, called SigVIC. Specifically, a spatial gating unit (SGU) and a spatial scaling network (SSN) are designed for using spatial importance to guide the feature scaling and bit allocation for variable-rate. Besides, a shallow feature fusion module (SFFM) is adopted to refine the decoded features with Top-K shallow features. Experimental results show that our method can yield a state-of-the-art performance, with significantly storage saving and flexibility improvement of compression methods.

5. ACKNOWLEDGEMENT

This work was supported in part by the National Natural Science Foundation of China under Grant 61972028, Grant 61902022, and Grant 62120106009.

6. REFERENCES

[1] Johannes Ballé, Valero Laparra, and Eero P Simoncelli, “End-to-end optimized image compression,” *arXiv preprint arXiv:1611.01704*, 2016.

[2] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston, “Variational image compression with a scale hyperprior,” *arXiv preprint arXiv:1802.01436*, 2018.

[3] David Minnen, Johannes Ballé, and George Toderici, “Joint autoregressive and hierarchical priors for learned image compression,” *arXiv preprint arXiv:1809.02736*, 2018.

[4] Jooyoung Lee, Seunghyun Cho, and Seung-Kwon Beack, “Context-adaptive entropy model for end-to-end optimized image compression,” *arXiv preprint arXiv:1809.10452*, 2018.

[5] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto, “Learned image compression with discretized gaussian mixture likelihoods and attention modules,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 7939–7948.

[6] Gregory K Wallace, “The JPEG still picture compression standard,” *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.

[7] Athanassios Skodras, Charilaos Christopoulos, and Touradj Ebrahimi, “The JPEG 2000 still image compression standard,” *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 36–58, 2001.

[8] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand, “Overview of the high efficiency video coding (HEVC) standard,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[9] Fei Yang, Luis Herranz, Joost Van De Weijer, José A Iglesias Guitián, Antonio M López, and Mikhail G Mozerov, “Variable rate deep image compression with modulated autoencoder,” *IEEE Signal Processing Letters*, vol. 27, pp. 331–335, 2020.

[10] Yoojin Choi, Mostafa El-Khamy, and Jungwon Lee, “Variable rate deep image compression with a conditional autoencoder,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019, pp. 3146–3154.

[11] Myungseo Song, Jinyoung Choi, and Bohyung Han, “Variable-rate deep image compression through spatially-adaptive feature transform,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 2380–2389.

[12] Shanzhi Yin, Chao Li, Youneng Bao, Yongsheng Liang, Fanyang Meng, and Wei Liu, “Universal efficient variable-rate neural image compression,” in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2022, pp. 2025–2029.

[13] Rushil Gupta, Suryateja BV, Nikhil Kapoor, Rajat Jaiswal, Sharmila Reddy Nangi, and Kuldeep Kulkarni, “User-guided variable rate learned image compression,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1753–1758.

[14] Dailan He, Ziming Yang, Weikun Peng, Rui Ma, Hongwei Qin, and Yan Wang, “ELIC: Efficient learned image compression with unevenly grouped space-channel contextual adaptive coding,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5718–5727.

[15] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu, “Image super-resolution using very deep residual channel attention networks,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 286–301.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[17] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[18] Tong Chen and Zhan Ma, “Variable bitrate image compression with quality scaling factors,” in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2020, pp. 2163–2167.

[19] “Kodak PhotoCD dataset. [online],” 2013, Available: `http://r0k.us/graphics/kodak/`.

[20] “CLIC dataset. [online],” 2019, Available: `http://www.compression.cc/`.

[21] Jens-Rainer Ohm and Gary J Sullivan, “Versatile video coding–towards the next generation of video compression,” in *Proceedings of Picture Coding Symposium (PCS)*, 2018, vol. 2018.