Attentional Multi-scale Back Projection Network for Low Bit-rate Image Compression

Ge Gao, Rong Pan, Pei You, Shunyuan Han, Yuanyuan Zhang, Hojae Lee
Samsung R&D Institute China Xi'an
{ge1.gao,rong.pan,pei.you,shuny.han,yuan2.zhang,hojae72.lee}@samsung.com

Abstract

In this paper, we provide the description of our approach for participating Workshop and Challenge on Learned Image Compression (CLIC) 2021: SRCX_DLIC. Our method is an end-to-end network based on variational autoencoder, for which we develop a novel back projection method with attentional and multi-scale feature fusion. Our back projection method recalibrates the current estimation by establishing feedback connections between high-level and low-level attributes in a discriminative and efficient manner. Further, our network recovers the fine spatial details by decomposing the input image and separately processing the distinct frequency components, whose derived latents are recombined using a novel dual attention module, so that details inside regions of interest could be explicitly manipulated. We also describe the training strategies for variable bit-rates and perceptual quality enhancement.

1. Introduction

Lately, the demand for image compression has increased dramatically to cope with the enormous amount of high-resolution images. Based on deep neural networks (DNNs), neural image compression has reinvigorated this domain with its superb capacity to learn in a data- and metric-driven manner, as opposed to their conventional counterparts [13].

Neural image compression typically employs autoencoders to model image down-sampling and up-sampling as a unified task and optimize the rate-distortion trade-off jointly. Such methods map the input image into a latent intermediate via an encoder and inversely transform the quantized latent back to generate the reconstructed image on the decoder side. Many researches concentrate on optimizing the network architecture, e.g., GDN [4], residual blocks [22, 18], RNNs [23, 16, 24], to facilitate both decorrelation and recovery of the image signal. Meanwhile, some other works focus on further reducing the entropy of the latent representations to attain fewer encoding bits. Earlier works [6, 22] in this respect incorporated elementwise entropy models to encode each element independently. Later advancements introduced hierarchical hyperprior networks [5] and autoregressive components [14, 20] into the VAE framework to explicitly estimate the entropy of the latent representation by utilizing prior information. Currently, the rate-distortion performance of the state-of-the-art methods have surpassed that of reigning compression codecs, such as BPG [7] and VVC [21], on both PSNR and MS-SSIM.

Nonetheless, existing schemes are limited in capturing the mappings between the source image and its compact form, leading to over-smoothed reconstructions at low compression rates. One major issue is that, while the autoencoder excels at extracting contextualized, non-linear information for effective decorrelation, it stumbles in preserving spatial image details that are crucial to faithful reconstruction. In addition, the fact that the input image is usually processed in its RGB format, in which these easily-lost high-frequency details are mingled with large-scale variations, makes it even harder for the network to preserve or infer fine-grained details for optimal reconstruction.

In this paper, to enable mutual facilitation between low- and high-level image properties, we replace the standard feedforward up-sampling layers with a novel Attentional Multi-scale Back Projection (AMBP) module. Our AMBP module aggregates intermediate features from higher to lower layers of the network, allowing it to attain semantically rich features, on the one hand, and extrapolate fine spatial details, on the other. Retaining the desired properties of both gives the network a greater flexibility to decide which information should be preserved for better rate-distortion trade-off. Our simplified design discards the iterative up- and down-sample procedure [11, 17] by including both in-scale and cross-scale feature fusion within one back projection step. To extract richer visual representations, we further leverage a soft attention mechanism that consolidates the input feature maps in a weighted average fashion.

Moreover, we propose to decompose the original image by extracting and processing its frequncy components separately so that the network could yield further efficiencies in representation by exploiting different pieces of information

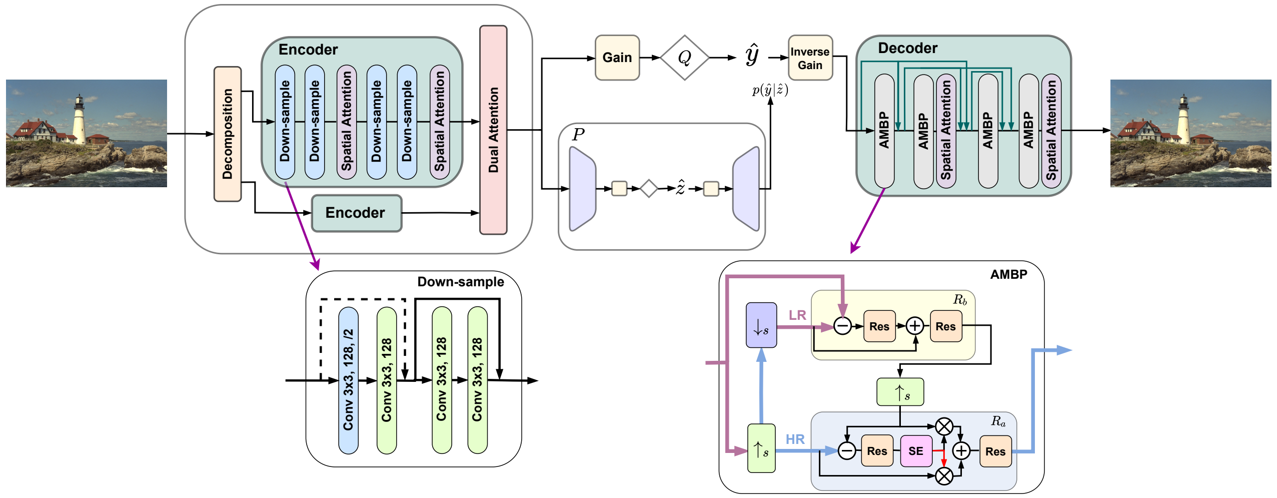

Figure 1. Network architecture of the proposed method. Q denotes the quantization. $\hat{y}$ and $\hat{z}$ denote the quantized latent and the quantized side information, respectively. P denotes the context model for estimating the entropy of $\hat{y}$. Gain and Inverse Gain denotes the gain unit for continuously variable bit-rates control. $\uparrow_s$ and $\downarrow_s$ denotes down-sampling and up-sampling by a scale factor s, respectively, via residual blocks. The black and red arrows in the attentional fusion block R_a denotes reweighting by W and $(1 - W)$, respectively. $\otimes$ denotes element-wise multiplication.

that contains distinctive frequency characteristics. Besides, to enhance the perceptual quality, we finetune the model by adding lpips [25] to the loss function, which enforces the network to focus on enhancing perceptual quality.

2. Proposed Method

2.1. Network Architecture

The network architecture is shown in Fig.1. The encoder side of our design consists of a decomposition module, a dual-branch encoder and a dual attention module. Instead of processing the input image in its RGB form, we extract its low- and high-frequency components and compress them separately using the dual-branch encoder, in which each branch contains four down-sampling modules/blocks and two spatial attention modules/blocks in between [9]. The down-sampled latents of the frequency components are then rescaled and combined into the complete latent representation y via the dual attention module. The hyperprior model and the context model follows the same design as [5]. The single-branch decoder consisting of four AMBP module that up-sample the quantized latent $\hat{y}$ into the reconstruction image. We further adopt dense connections where the current AMBP module process the concatenation of outputs of all previous modules [11]. Gain and Inverse-Gain refers to the gain units used to accomplish variable bit-rates in a single model [10].

2.2. Attentional Mutli-scale Back Projection

Back projection was first put forward in DBPN [11] for image super-resolution. The back projection technique iteratively utilizes the feedback residual to refine HR images, based on the assumption that the projected, down-sampled version of a SR image should be as close as possible to the original LR image. We extend the similar idea to image compression task and construct our building blocks entitled as the AMBP modules.

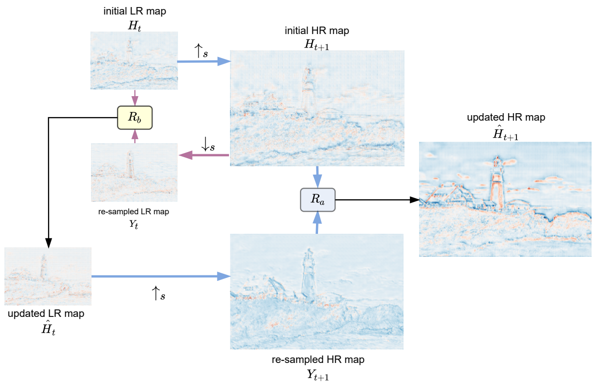

Figure 2. Illustration of the back projection procedure using feature maps sampled from the decoder when reconstructing *kodim21.png*. The updated HR map $\hat{H}_{t+1}$ contains better defined details than the initial HR map H_{t+1}.

Convolution layers of the autoencoder trade fine spatial details for copious semantic information through the repeated use of down-sampling operations, making it less reliable for faithful image reconstruction. To address this issue, AMBP aggregates multi-scale features across stages in a trainable way. That is, the current stage features are consolidated by the complementary information (spatially accurate) from later computations. The refined feature maps in turn produce features of higher quality in the next stage, thereby achieving progressive improvement to the intermediate features that propagate throughout the computation.

As shown in Fig. 2, our AMBP module refines a HR map H_{t+1}, up-sampled from H_t, by applying reverse mapping to recover its original resolution. Despite having the same resolution, the re-sampled feature map Y_t encloses details from H_{t+1} that are then integrated into H_t using a fusion module R_b, and the updated LR map is up-sampled again to yield a re-sampled HR map Y_{t+1}. To facilitate in-scale feature fusion, we leverage another fusion module R_a that aggregates H_{t+1} and Y_{t+1} to update the later as $\hat{H}_{t+1}$ that contains more detailed features. The described process can be written as:

$$\begin{aligned} Y_t &= \downarrow_s (H_{t+1}) = \downarrow_s (\uparrow_s (H_t)) \\ \hat{H}_t &= R_b(H_t, Y_t) \\ Y_{t+1} &= \uparrow_s (\hat{H}_t) \\ \hat{H}_{t+1} &= R_a(H_{t+1}, Y_{t+1}). \end{aligned} \tag{1}$$

The feature fusion is based on residual calculation. As shown in Fig. 1, the residual fusion module R_b aggregates H_t and Y_t according to their residual e_t. Inituitively, the residual $e_t = H_t - Y_t$ represents distinctive information available in one source while missing in the other. The key modification we made is to incorporate another attentional fusion module R_a for the (H_{t+1}, Y_{t+1}) pair. We leverage channel attention [12] to capture the channel-wise depedencies of the residual $e_{t+1} = H_{t+1} - Y_{t+1}$ at a global scale, and adopt a soft fusion scheme that reweights the respective inputs by W and $(1 - W)$, where W is the normalized attention map. The reisudal fusion modules is formulated as:

$$R_b(H_t, Y_t) = \mathcal{R}(Y_t + \mathcal{R}(e_t))$$

$$W = \mathcal{S}(\mathcal{R}(e_{t+1})) \tag{2}$$

$$R_a(H_{t+1}, Y_{t+1}) = \mathcal{R}(W \otimes H_{t+1} + (1 - W) \otimes Y_{t+1}),$$

where $\mathcal{R}$ denotes residual blocks, W denotes the attention map, $\mathcal{S}$ denotes channel attention, and $\otimes$ denotes element-wise multiplication.

The benefits of the additional feature fusion operations are threefold. First, it further optimizes the re-sampled feature map Y_{t+1} and faciliates both in-scale and cross-scale feature fusion without iterations. Second, the proposed soft content selection scheme efficiently enables more adaptive feature fusion by implictly reguarlizing the values of the attention map. Third, fusion based on residual allows the network to focus only on distinctive information, making the gradient update better guided, and incorporating another residual-based fusion operation could further accelerate the training procedure.

2.3. Decomposition

"High-frequency content will get lost during sample rate conversion", as pointed out by the Nyquist-Shannon Sampling Thorem. Decomposition could help alleviate this issue by enabling the network to explicitly manipulate the high-frequency contents.

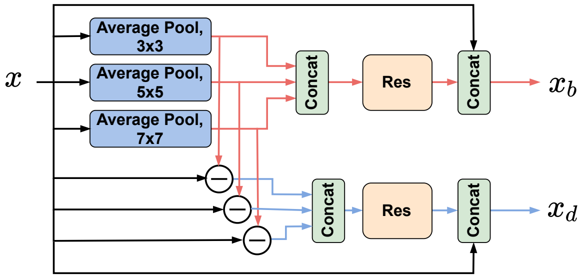

Figure 3. The frequency decomposition module, where the red and blue arrows denote the low- and high-frequency components, respectively

As illustrated in Fig.3, the low-frequency components across scales are obtained using average pooling with various kernel sizes. The high-frequency components are attained by subtracting the corresponding low-frequency component from the input image x. To produce the base layer x_b, we pass the concatenated low-frequency components to a residual block. The detail layer x_d containing high-frequency information is attained in a similar way. As the original image x also contains rich information, it is concatenated with x_b and x_d and then processed separately by the dual-branch encoder, which progressively down-samples the respective components into their latent representations y_{lf} and y_{hf}.

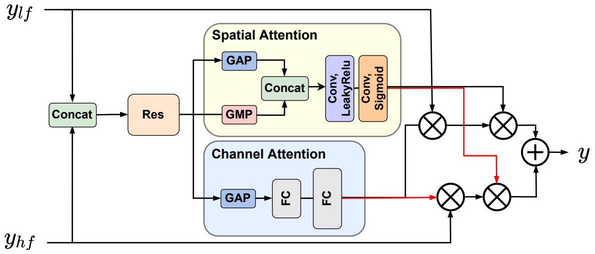

Figure 4. Dual attention module. FC denotes fully connected networks. The black and red arrows indicate multiplying the feature map by the corresponding attention weight W and (1-W), respectively. GAP and GWP denote global average pooling and global max pooling, respectively.

Afterwards, the latents y_{lf} and y_{hf} are aggregated using a dual attention module (Fig.4). The latents of the respective components are concatenated along the channel dimension to produce feature map F, which is then transformed by a residual block and passed to the channel and the spatial attention module. To reduce computation, the spatial attention module independently applies global average pooling and global max pooling to F along the channel dimension and concatenates the results to form feature map $F_s \in \mathbb{R}^{H\times W\times 2}$, from which is the spatial attention map $W_s \in \mathbb{R}^{H\times W\times 1}$ extracted. The channel attention feature map $W_c \in \mathbb{R}^{1\times 1\times C}$ is generated using SE blocks [12]. We adopt the soft selection trick to improve representations as well. The low-frequency latent y_{lf} is rescaled by W_c and then W_s while the high-frequency latent y_{hf} is rescaled by $(1 - W_c)$ and then $(1 - W_s)$. The re-weighted latents are then summed to yield the final latent representation y.

2.4. Variable Bit Compression

Considering the bit-rates constraint, we adopt the gain units [10] to achieve continuously variable rate with a single model. As shown in Fig.1, the gain unit and the reverse gain unit are added after the encoder and before the decoder, respectively. The pairs are inserted into the hyperprior model in a similar manner.

2.5. Finetune Strategies

Enhancing the perceptual quality of our method is imperative, as the challenge performs evaluation based on human perception. Thus, we finetune the decoder D in our baseline model by incorporating the up-to-date perceptual loss - LPIPS [25] to improve the perceptual quality. Besides, we notice that the reconstruction quality could be considerably enhanced by decoding the original rather than quantized latent without tuning, so we added a rounding loss $d_r = MSE(y, \tilde{y})$ to the loss function, where $\tilde{y}$ is the updated latent map by the residual fusion module. Inituitively, variations of a pixel's neighboring pixels in the high-resolution map could help the network infer the pixel's value before quantization. The loss function for finetuning becomes:

$$\mathcal{L}_f = d(x, \hat{x}) + \beta \cdot d_r + \gamma \cdot d_{lpips}, \tag{3}$$

where β and γ controls the weight of the rounding loss term and perceptual loss term.

3. Implementation Details and Results

We trained the proposed networks using cropped images of size 256x256 from DIV2K [2], Flickr2K [15], and CLIC training dataset [1] without augmentation. Formulating E, D and $P(y)$ in our network allows them to be trained jointly by minimizing the rate-distortion trade-off:

$$\mathcal{L} = \sum_{i=0}^{n} \lambda_i \cdot d(x, \hat{x}) + R, \tag{4}$$

where d represents distotion, R represents the required number of bits, and λ_i represents the index of the gain vectors in the gain metrix [10]. We used the Adam algorithm to jointly optimize the networks for *1.2M* iterations with a mini-batch size of 4. The initial learning rate was set to 1×10^{-4} and decreased to 5×10^{-5} at *800k* iterations. The networks were optimized with respect to two quality metrics, i.e., mean square error (MSE) and multi-scale structural similarity index (MS-SSIM). The distortion d is defined as d = MSE($\boldsymbol{x}$,$\hat{\boldsymbol{x}}$) and d = 1 - MS-SSIM($\boldsymbol{x}$,$\hat{\boldsymbol{x}}$), respectively. When optimized by MSE, the value of λ belongs to the set $\{0.0004, 0.0025, 0.009\}$, $\{0.0009, 0.0085, 0.02\}$, $\{0.005, 0.009, 0.045\}$ for the three different bit-rates, respectively. As to networks optimized for MS-SSIM, the value of λ belongs to the set $\{1, 4, 9\}$, $\{2, 10, 32\}$, $\{10, 35, 120\}$ for the three different bit-rates, respectively. After that, we finetuned the decoder for *500k* iterations at the learning rate 5×10^{-5}, the coefficient β and γ was set to 1 and 100 for MSE optimized model.

The results on CLIC validation dataset are summarized in Table 1, we could attain the bit-rates close to the constraint value through the gain unit. It can be seen that, while MSE optimized models attain higher PSNR than MS-SSIM optimized models, they decode images of lower MS-SSIM in comparison. Further, finetuned models achieve higher perceptual scores (in this case, the lower the lpips the better), which suggests the effectiveness of our strategies in enhancing the perceptual quality of reconstructed images.

Table 1. Results on CLIC validation dataset.

Method	BPP	PSNR	MS-SSIM	lpips
MSE Optimized	0.0732 0.1481 0.2964	27.836 30.825 34.014	0.9279 0.9605 0.9783	0.1676 0.0946 0.1059
Fintuned MSE Optimized	0.0732 0.1481 0.2964	26.866 29.124 30.355	0.9513 0.9694 0.9799	0.1602 0.0927 0.1013
MS-SSIM Optimized	0.0743 0.1475 0.2974	26.375 28.621 29.819	0.9555 0.9718 0.9813	- - -

4. Conclusion

In this paper, we propose a neural image compression scheme using back projection techniques and frequency decomposition. We reformulate the iterative projection operations into a multi-scale feature fusion module and incorporate channel attention with soft content selection. We also enable the network to focus on respective frequency components of the input image via decomposition, where their derived latents are adaptively rescaled and integrated using an efficient dual attention module. Further, we adopt a finetune strategy that helps enhance the perceptual quality and reduce the latent reisudal.

References

[1] Workshop and challenge on learned image compression. https://compression.cc. 4

[2] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017. 4

[3] Eirikur Agustsson, Michael Tschannen, Fabian Mentzer, Radu Timofte, and Luc Van Gool. Generative adversarial networks for extreme learned image compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 221–231, 2019.

[4] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. Density modeling of images using a generalized normalization transformation. In *International Conference on Learning Representations*, 2016. 1

[5] Johannes Ballé, David Minnen, Saurabh Singh, Sung Jin Hwang, and Nick Johnston. Variational image compression with a scale hyperprior. In *International Conference on Learning Representations*, 2018. 1, 2

[6] Johannes Ballé, Valero Laparra, and Eero Simoncelli. End-to-end optimized image compression. In *International Conference on Learning Representations*, 2017. 1

[7] Fabrice Bellard. Bpg image format. https://bellard.org/bpg. 2014. 1

[8] Yochai Blau, Roey Mechrez, Radu Timofte, Tomer Michaeli, and Lihi Zelnik-Manor. The 2018 pirm challenge on perceptual image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, September 2018.

[9] 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*, pages 7939–7948, 2020. 2

[10] Ze Cui, Jing Wang, Bo Bai, Tiansheng Guo, and Yihui Feng. G-vae: A continuously variable rate deep image compression framework, 2020. 2, 4

[11] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1664–1673, 2018. 1, 2

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 3, 4

[13] Yueyu Hu, Wenhan Yang, Zhan Ma, and Jiaying Liu. Learning end-to-end lossy image compression: A benchmark. *arXiv preprint arXiv:2002.03711*, 2020. 1

[14] Jooyoung Lee, Seunghyun Cho, and Seung-Kwon Beack. Context-adaptive entropy model for end-to-end optimized image compression. In *International Conference on Learning Representations*, 2018. 1

[15] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1132–1140, 2017. 4

[16] Chaoyi Lin, Jiabao Yao, Fangdong Chen, and Li Wang. A spatial rnn codec for end-to-end image compression. In *International Conference on Computer Vision and Pattern Recognition*, 2020. 1

[17] Zhi-Song Liu, Li-Wen Wang, Chu-Tak Li, Wan-Chi Siu, and Yui-Lam Chan. Image super-resolution via attention based back projection networks. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3517–3525. IEEE, 2019. 1

[18] Fabian Mentzer, Eirikur Agustsson, Michael Tschannen, Radu Timofte, and Luc Van Gool. Conditional probability models for deep image compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4394–4402, 2018. 1

[19] Fabian Mentzer, George Toderici, Michael Tschannen, and Eirikur Agustsson. High-fidelity generative image compression. In *International Conference on Neural Information Processing Systems*, 2020.

[20] David Minnen, Johannes Ballé, and George Toderici. Joint autoregressive and hierarchical priors for learned image compression. In *Advances in Neural Information Processing Systems*, pages 10794–10803, 2018. 1

[21] Gary Sullivan and Jens-Rainer Ohm. Versatile video coding. *JVET-T2002*, 2020. 1

[22] Lucas Theis, Wenzhe Shi, Andrew Cunningham, and Ferenc Huszár. Lossy image compression with compressive autoencoders. In *International Conference on Learning Representations*, 2017. 1

[23] George Toderici, Sean M. O'Malley, Sung Jin Hwang, Damien Vincent, David Minnen, Shumeet Baluja, Michele Covell, and Rahul Sukthankar. Variable rate image compression with recurrent neural networks. In *International Conference on Learning Representations*, 2016. 1

[24] 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. 1

[25] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018. 2, 4