

SELIC: Semantic-Enhanced Learned Image Compression via High-Level Textual Guidance

Haisheng Fu
Simon Fraser University
haisheng_fu@sfu.ca

Jie Liang
Simon Fraser University
jie_liang@sfu.ca

Zhenman Fang
Simon Fraser University
zhenman@sfu.ca

Jingning Han
Google
jingning@google.com

Abstract—Learned image compression (LIC) techniques have achieved remarkable progress; however, effectively integrating high-level semantic information remains challenging. In this work, we present a Semantic-Enhanced Learned Image Compression framework, termed SELIC, which leverages high-level textual guidance to improve rate-distortion performance. Specifically, SELIC employs a text encoder to extract rich semantic descriptions from the input image. These textual features are transformed into fixed-dimension tensors and seamlessly fused with the image-derived latent representation. By embedding the SELIC tensor directly into the compression pipeline, our approach enriches the bitstream without requiring additional inputs at the decoder, thereby maintaining fast and efficient decoding. Extensive experiments on benchmark datasets (e.g., Kodak) demonstrate that integrating semantic information substantially enhances compression quality. Our SELIC-guided method outperforms a baseline LIC model without semantic integration by approximately 0.1-0.15 dB across a wide range of bit rates in PSNR and achieves a 4.9% BD-rate improvement over VVC. Moreover, this improvement comes with minimal computational overhead, making the proposed SELIC framework a practical solution for advanced image compression applications.

Index Terms—semantic guidance, learned image compression, textual fusion

I. INTRODUCTION

In recent years, deep learning tools have been extensively applied to the field of image compression, achieving remarkable advancements that surpass traditional standards such as JPEG [1], JPEG 2000 [2], BPG (intra-coding of H.265/HEVC) [3], and H.266/VVC [4] in both objective and subjective metrics. Traditional image compression techniques typically consist of key components such as linear transform functions, quantization modules, and entropy coding. Similarly, end-to-end learned image compression frameworks follow a comparable pipeline but leverage learnable parameters to replace and enhance these modules.

Early learned image compression models primarily focused on enhancing coding performance by introducing various advanced neural network architectures, including residual blocks [5]–[7], self-attention mechanisms [8], [9], invertible structures [10], transformer-based blocks [11], [12], and wavelet-based blocks [13], [14]. These modules enable neural networks

to extract effective and efficient latent representations while reducing the redundancy in the input image.

Estimating a powerful and efficient entropy model is a critical topic in image compression. In [15], a hyperprior network is proposed to estimate the conditional probabilities of the latent representations. The hyper encoder extracts hyperpriors from the latents, which are encoded as side information and transmitted to a hyper decoder. The reconstructed hyperpriors enable the estimation of latent conditional probabilities, making the entropy model adaptive to both image content and spatial variations. This approach assumes a zero-mean Gaussian scale mixture (GSM) model for the latents, achieving superior performance compared to BPG (4:4:4). Subsequently, [16] extends this method by employing a non-zero-mean Gaussian mixture model (GMM). Furthermore, the autoregressive context model introduced in [17] is utilized to refine latent probability estimation by leveraging both hyperpriors and spatial context.

Dynamic bitrate allocation is also an essential technique for improving image compression performance. It involves allocating more bits to important regions while assigning fewer bits to flat areas. Most previous methods [7], [18] achieve bitrate allocation by introducing importance maps. However, these dynamic maps are learned by the network and do not represent the high-level semantic information conveyed by the image. A critical aspect of learned image compression is the integration of semantic information to enhance compression efficiency. While existing methods excel in capturing low-level image features, they frequently overlook the potential of high-level semantic guidance. Incorporating semantic information can provide a more nuanced understanding of image content, enabling more intelligent compression strategies that prioritize essential elements within an image.

In this paper, we introduce a novel image compression framework that leverages high-level semantic information extracted from images to guide the compression process. Our approach involves using a text encoder to translate image content into rich textual descriptions that encapsulate the image’s key semantics. This textual information is then transformed into a fixed tensor vector, which is integrated with the encoded image bitstream through various fusion techniques. By embedding semantic vectors directly into the compression pipeline, our method enhances the bitstream without necessitating additional inputs during decoding, thereby maintaining efficient

This work was supported by the Natural Sciences and Engineering Research Council of Canada (RGPIN-2020-04525, RGPIN-2019-04613, DGECR-2019-00120, ALLRP-552042-2020), Google Chrome University Research Program, MITACS Elevate Postdoc Program IT42556, CFI John R. Evans Leaders Fund.

decoding speeds.

The main contributions of this paper can be summarized as follows:

- We introduce a semantic-guided compression framework that leverages a pretrained text encoder to extract high-level semantic information from the input image. By embedding these semantic cues directly into the latent domain, our method preserves critical content and contextual nuances throughout the compression pipeline.
- We propose an image-text fusion module employing a channel-concatenation-based fusion strategy to effectively integrate the extracted semantic features with the image's latent representation. Compared to conventional element-wise addition or multiplication, this approach provides a more robust and flexible mechanism for aligning semantic and visual signals, thereby enhancing compression performance with minimal computational overhead.
- Our framework avoids the need for external information at the decoder, as the semantic guidance is fully embedded in the transmitted bitstream. This not only streamlines and accelerates the decoding process but also delivers superior rate-distortion efficiency. Experimental results show that our method outperforms recent learned image compression (LIC) approaches, achieving a better trade-off among coding performance, decoding time, and model complexity.

Extensive experiments on benchmark datasets (e.g., Kodak and Tecnick) demonstrate that incorporating semantic information leads to notable improvements in PSNR. Our findings confirm that semantic-aware integration within the compression pipeline produces more faithful reconstructions and enhances visual fidelity across a wide range of bit rates.

II. RELATED WORK

A. Learned Image Compression

Learned image compression methods have achieved superior rate-distortion performance in terms of PSNR and MS-SSIM metrics. These methods mainly improve coding performance through two ways. The first way is that different neural networks are provided to enhance encoder and decoder's learning abilities and better reduce the correlation of the latent representation. For example, different attention mechanisms [19]–[21] are introduced to extract compact and efficient latent representations. Also, the transformer-based modules [11], [21] are proposed to help the networks to extract global information of input image and better improve compression efficiency.

The second approach involves estimating a powerful autoregressive entropy model for latent representations. For instance, the GMM [19] and GLLMM [20] were proposed to accurately model the probability distribution of complex image regions, significantly reducing bit rates. Although these entropy models enhance compression performance, they introduce considerable decoding latency. To address this issue, parallelizable autoregressive entropy modules, such as

the checkerboard entropy model and the channel-wise autoregressive entropy model (ChARM), have been proposed. These methods aim to accelerate the decoding process while maintaining coding performance as much as possible.

Achieving a good trade-off between coding performance and decoding speed is a crucial challenge in image compression tasks. For instance, knowledge distillation methods [7], [22] have been proposed to reduce the complexity of decoding networks while preserving rate-distortion performance as much as possible.

B. Text-Conditioned Learned Image Compression

Early attempts at text-guided image compression, such as Text & Sketch (PICS) [23], combined caption information and sketch-based structures within diffusion models. While this approach achieved very low bitrates and preserved high-level semantics, it often produced images that differed significantly from the original content, as it prioritized semantic fidelity over pixel accuracy.

Later works like MISC [24] integrated both full-resolution images and text-driven diffusion refinements. Although MISC improved perceptual quality, it still used separate processing streams for text and image inputs. This separation, along with sparse textual encoding, made it difficult to maintain fine visual details (e.g., high PSNR) because semantic abstractions and low-level image signals were not fully aligned.

In contrast, our approach embeds textual semantics directly into the learned compression pipeline at the latent level. By fusing text and image representations before quantization, we remove the need for separate text-image pathways. This unified process retains key semantic cues throughout compression and reconstruction without relying on diffusion-based methods. As a result, we achieve strong semantic guidance while maintaining or even improving pixel-level fidelity, demonstrating a more direct and effective way to incorporate textual information into learned image compression.

III. THE PROPOSED IMAGE COMPRESSION FRAMEWORK

Our proposed framework integrates high-level semantic guidance into a learned image compression pipeline, as illustrated in Fig. 1. It comprises a main image encoder g_{a3} , a main image decoder g_s , a semantic image-to-text encoder (BLIP [26]) g_{a1} , a text-to-tensor encoder (BERT [27]) g_{a2} , a text-image latent fusion module, and hyperprior networks h_a and h_s . Unlike conventional methods that rely primarily on low-level visual signals, our method enriches the latent representation with contextual semantic information derived from a pretrained text model.

The process unfolds as follows. Given an input image x , the semantic branch first converts it into a textual description via g_{a1} . This textual output is then encoded by g_{a2} into a compact semantic feature vector of fixed dimension named y_2 . In parallel, the main image encoder g_{a3} processes the input image x through three residual blocks [25] and three downsampling steps to produce an image-derived latent y_3 .

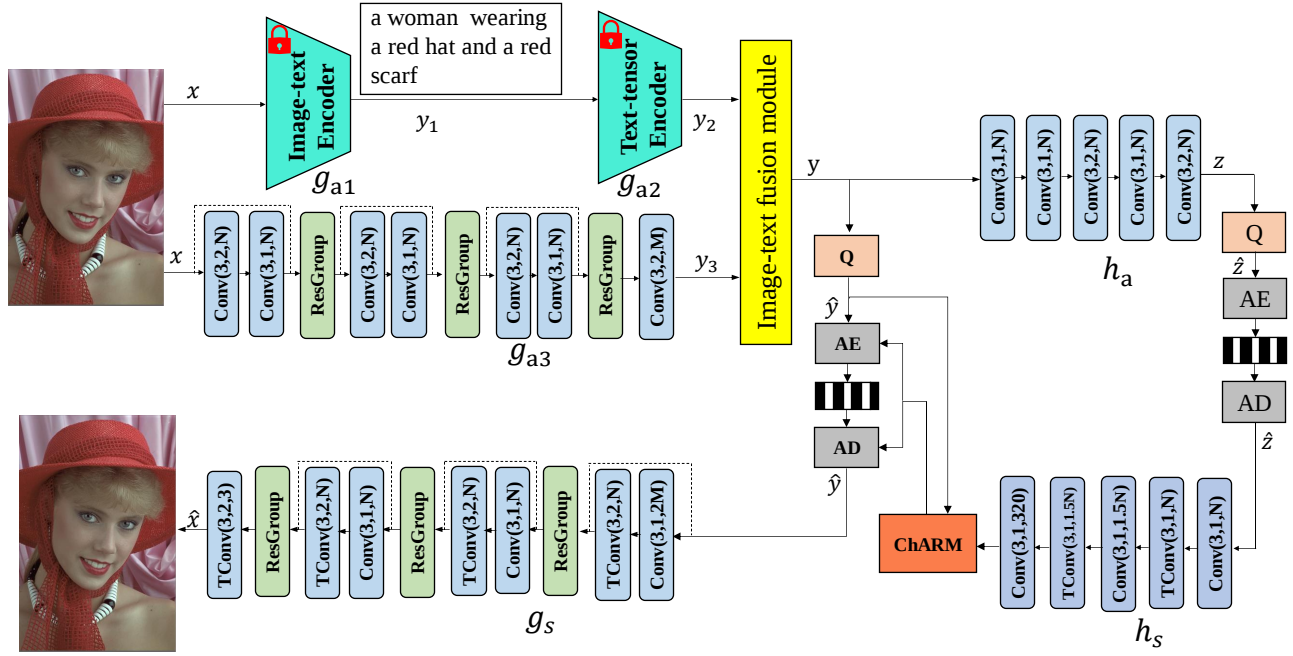


Fig. 1. The detailed architecture of the proposed framework. $\text{Conv}(3, s, n)$ denotes a convolutional layer with a 3×3 kernel size, stride s , and n filters. $\text{TConv}(3, s, n)$ represents a transposed convolution layer. Dashed shortcut connections represent changes in tensor size. AE and AD stand for Arithmetic Encoder and Arithmetic Decoder, respectively. The dotted lines represent the shortcut connection when size is change, as in [19], [25].

The tensors y_2 and y_3 are then passed to the image-text fusion module to obtain y , as described in Sec. III-A.

It is important to note that both g_{a1} (BLIP) and g_{a2} (BERT) are directly taken from publicly available pretrained models without updating their parameters during training. Since these models have been extensively trained on large-scale datasets, they provide robust and semantically rich textual features, ensuring that our semantic guidance is of high quality without adding significant training overhead.

As in [21], [28], we employ a channel-wise entropy coding (ChARM) approach for the latent representation y , as indicated in Fig. 1. The ChARM structure remains consistent with [29]. To enhance entropy coding performance, the hyperprior networks h_a and h_s encode and decode the hyperprior information z for y , as shown in Fig. 1.

A. Text-Image Latent Fusion Module

As shown in Fig. 2, the semantic representation y_2 , initially a single vector of dimension M , is broadcast across the spatial dimensions to form a $(H/16) \times (W/16) \times M$ tensor, aligning high-level semantic features with the image grid. In parallel, the main image encoder g_{a3} transforms x into a latent y_3 of identical spatial size and M channels. Ensuring both y_2 and y_3 share the same resolution and channel dimensions enables their seamless channel-wise concatenation.

This fusion produces a $(H/16) \times (W/16) \times 2M$ tensor that combines semantic latent y_2 and visual latent y_3 . A subsequent residual block refines the fused representation and restores its dimensionality to M channels, maintaining a compact, semantically enriched latent suitable for hyperprior analysis

and entropy coding. We evaluate the effectiveness of this fusion strategy through ablation studies.

IV. THE LOSS FUNCTION

Our loss function combines rate R and distortion D terms. The rate R is associated with the expected code length of the latents y and hyper-latents z , while the distortion D measures the reconstruction fidelity between x and $\hat{x}$. In this paper, we employ Mean Squared Error (MSE) as the distortion metric. A Lagrange multiplier λ balances the trade-off between rate and distortion:

$$\begin{aligned} L &= R + \lambda D, \\ D &= \mathbb{E}_{x \sim P_x} [d(x, \hat{x})], \end{aligned} \quad (1)$$

where P_x denotes the distribution of natural images.

V. EXPERIMENT

A. Training Details

Our models are trained on color PNG images from the CLIC dataset¹ and the LIU4K dataset [30]. Each model targets a specific rate setting by adjusting $\lambda \in \{0.0016, 0.0032, 0.0075, 0.015, 0.03, 0.045, 0.06\}$, optimizing primarily for PSNR. The number of filters N is fixed at 128. We train for 160 epochs using Adam with a batch size of 8. The learning rate is 1×10^{-4} for the first 130 epochs and reduced to 1×10^{-5} for the remaining 30 epochs. This training strategy ensures stable convergence and improved rate-distortion performance for all target bit rates.

¹<http://www.compression.cc/>

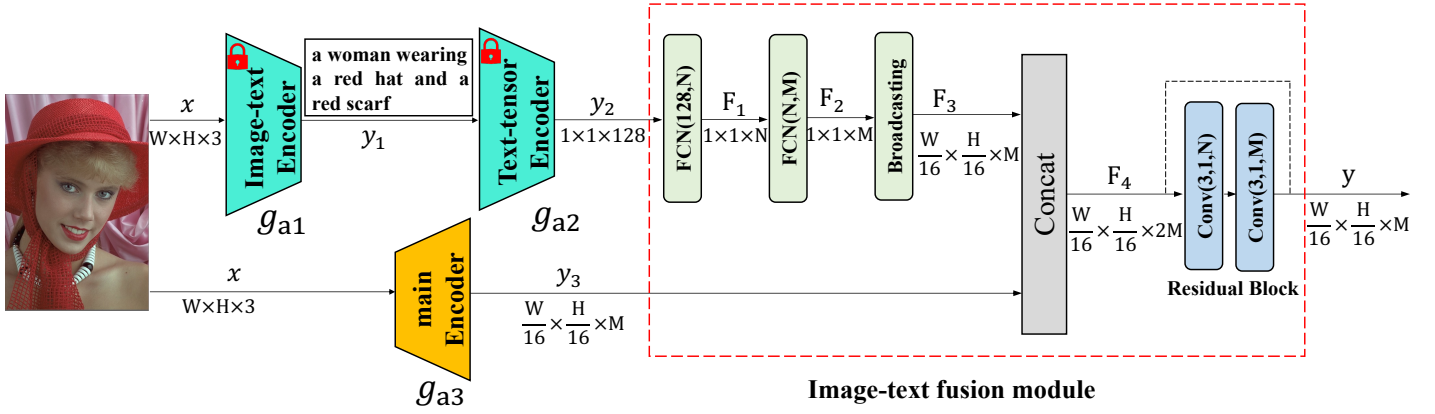


Fig. 2. Processing details of the text-image fusion module. Each $\text{FCN}(N, M)$ denotes a fully connected layer mapping from N input units to M output units.

TABLE I
COMPARISONS OF ENCODING/DECODING TIME, BD-RATE REDUCTION OVER VVC, AND MODEL PARAMETERS ON KODAK TEST SET.

Methods	Enc. Time	Dec Time	BD-Rate	#Params
VVC	402.3s	0.61s	0.0	-
Cheng2020 [19]	27.6s	28.8s	2.6 %	50.80 MB
Hu2021 [33]	32.7s	77.8s	11.1 %	84.60 MB
He2021 [34]	20.4s	5.2s	8.9 %	46.60 MB
Xie2021 [10]	4.097s	9.250s	-0.8 %	128.86 MB
Zhu2022 [11]	0.269s	0.183s	-3.9 %	32.34 MB
Zou2022 [12]	0.163s	0.184s	-2.2 %	99.86 MB
Qian2022 [31]	4.78s	85.82s	3.2 %	128.86 MB
Fu2023 [20]	420.6s	423.8s	-3.1 %	-
Ours	0.448s	0.167s	-4.9 %	52.22 MB

B. Comparisons

We compare our proposed method with recent learned compression methods and traditional codecs in terms of PSNR metric. The LIC methods include Fu2023 [20], Zhu2022 [11], Yi2022 [31], He2022 [32], Xie2021 [10], Cheng2020 [19]. The traditional methods are H.266/VVC Intra (4:4:4), and H.265/BPG Intra (4:4:4).

Fig. 3 shows the average rate-distortion curves on the 24 Kodak images and Tecnick 100 dataset. For Kodak, GLLMM achieves the top performance. Our method closely approaches GLLMM over a wide range of bit rates, outperforming other learning-based techniques and showing similar or better results than VVC (4:4:4) depending on the rate. For Tenick dataset, we remain competitive with GLLMM and surpass other learned methods.

C. Performance and Speed Trade-off

Table I compares encoding/decoding times, BD-rate reductions relative to VVC [35], and model complexities for various methods on the Kodak test set. All learned approaches, including ours, were evaluated on an NVIDIA Tesla V100 GPU with 16 GB memory, whereas VVC was tested on a 2.9GHz Intel Xeon Gold 6226R CPU. Parameter counts were obtained using the PyTorch Flops Profiler, except for [20],

for which an exact measure is unavailable. Nonetheless, [20] reports notably higher complexity than [19].

Some learned image compression methods [10], [19], [20], [31] rely on serial autoregressive models that limit parallelization and thus slow down inference. By contrast, more recent approaches [11], [12] employ parallelizable entropy models, substantially improving speed on GPUs.

Our proposed method surpasses [11] by about 1.0% in BD-rate reduction and achieves a total of -4.9% relative to VVC. While our parameter count (52.22 MB) is moderately larger than that of [11] (32.34 MB) and slightly higher than [19] (50.80 MB), it remains significantly smaller than other recent methods such as [12]. Although our encoding speed is slower than those of [12] and [11], our decoding time is on par or faster, and both are still orders of magnitude quicker than many traditional or highly complex learned models.

Notably, our encoding time increases because we invoke pretrained image-text and text-tensor encoders (e.g., BLIP and BERT) at runtime. These encoders are employed through API calls without updating their parameters, ensuring they do not inflate our model's parameter count. As a result, our overall parameter size remains stable, even though the initial encoding step is more time-consuming than the decoding phase.

In summary, our method balances compression performance, computational complexity, and decoding speed effectively. This trade-off makes our approach suitable for real-world scenarios, offering high-quality image compression guided by semantic information while maintaining manageable model size and practical runtime characteristics.

D. Ablation Study on Semantic Integration

We conduct an ablation study by removing the semantic encoders (g_{a1} , g_{a2}) and the text-image fusion module, treating the remaining structure as our baseline. As shown in Fig. 4, without the integrated semantic information, the performance shows a minor decrease at low bit rates. However, at higher bit rates, the absence of textual guidance leads to a more pronounced drop of approximately 0.1 dB-0.15 dB in PSNR.

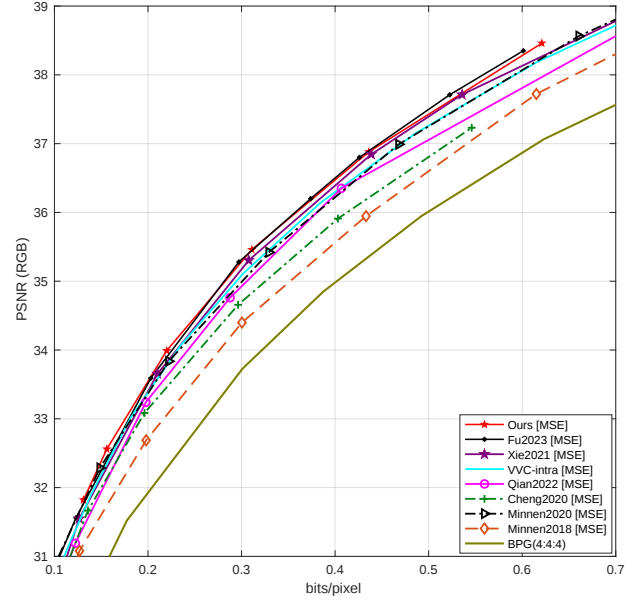
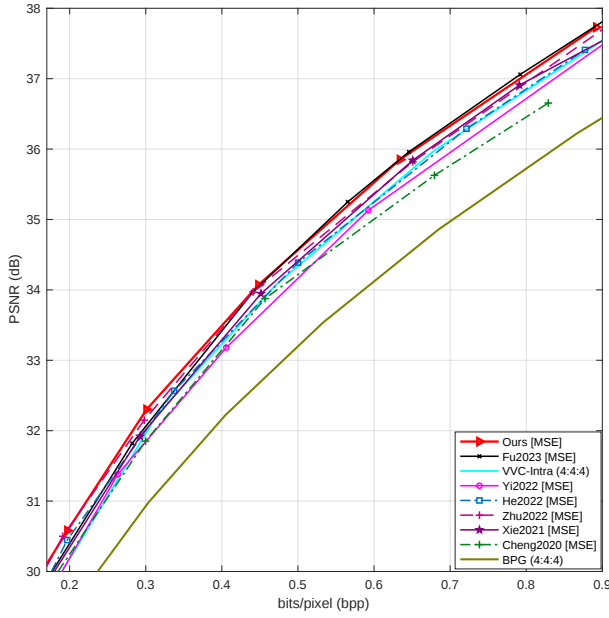


Fig. 3. Average PSNR and MS-SSIM performance on the 24 Kodak images and 100 Tecnick images.

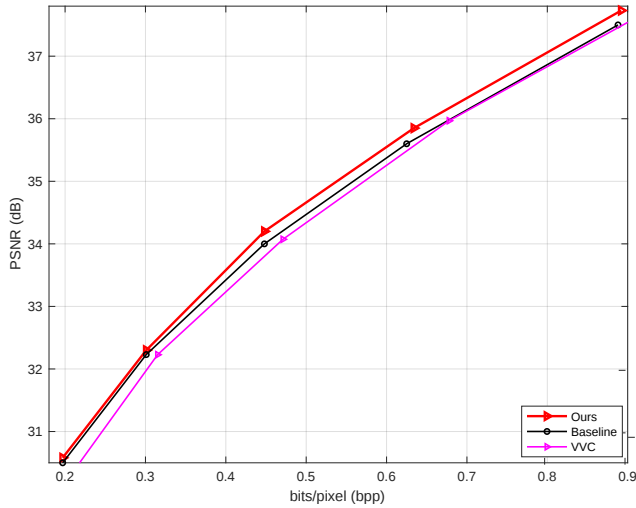


Fig. 4. The impact of removing semantic-related modules from the proposed framework.

The reason is that semantic cues become increasingly beneficial as more bits are allocated to represent subtle details. At higher bit rates, the model can exploit the textual semantics to guide its feature allocation more efficiently, improving the accuracy of subtle texture and structural reconstruction. When these semantic cues are removed, the model lacks a high-level contextual reference, making it harder to optimally distribute available bits. As a result, the rate-distortion trade-off deteriorates more noticeably in these higher-rate scenarios, underscoring the value of semantic guidance in enhancing overall compression performance.

E. Performance Comparison with Different Fusion Mechanisms

TABLE II
PERFORMANCE WITH DIFFERENT FUSION MECHNISMS

Methods	Bit rate	PSNR
Element-wise Multiplication	0.900	37.16 dB
Element-wise Addition	0.895	37.32 dB
(ours)Channel Concatenation	0.890	37.68 dB
Element-wise Multiplication	0.203	30.12 dB
Element-wise Addition	0.203	30.20 dB
(ours)Channel Concatenation	0.199	30.61 dB

Table II compares different fusion strategies for integrating image and text features at varying bit rates. Both element-wise addition and multiplication achieve suboptimal performance, as textual data is too sparse to be directly fused with image representations. This sparsity prevents effective modeling of joint dependencies, resulting in lower PSNR results.

In contrast, channel concatenation achieves consistently higher PSNR at similar or even lower bit rates. By simply merging the two feature sets along the channel dimension, it circumvents the limitations posed by sparse textual inputs. This approach preserves richer spatial information and facilitates more effective entropy modeling, ultimately enhancing overall compression quality.

VI. CONCLUSION

In this paper, we introduced a semantic-enhanced learned image compression framework named **SELIC** that directly integrates high-level textual information into the compression pipeline. By embedding textual information extracted from a dedicated image-to-text encoder, our method effectively lever-

ages semantic guidance to enhance the rate-distortion trade-off, achieving both improved fidelity and efficient decoding.

Experimental results confirm that this approach outperforms state-of-the-art LIC methods while maintaining practical computational costs. For future work, exploring more advanced image-text models could yield even more accurate semantic representations, and employing transformer-based modules for cross-modal fusion may further improve compression efficiency and quality. These directions hold promise for advancing semantic-driven learned image compression and broadening its applicability.

REFERENCES

- [1] G. K. Wallace, "The jpeg still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. 18–34, 1992.
- [2] A. Skodras, C. Christopoulos, and T. Ebrahimi, "The jpeg 2000 still image compression standard," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 36–58, 2001.
- [3] 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.
- [4] H Fraunhofer, "Vvc official test model vtm," 2019.
- [5] Honggang Chen, Xiaohai He, Hong Yang, Linbo Qing, and Qizhi Teng, "A feature-enriched deep convolutional neural network for jpeg image compression artifacts reduction and its applications," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 1, pp. 430–444, 2022.
- [6] Haisheng Fu, Feng Liang, Bo Lei, Nai Bian, Qian Zhang, Mohammad Akbari, Jie Liang, and Chengjie Tu, "Improved hybrid layered image compression using deep learning and traditional codecs," *Signal Processing: Image Communication*, vol. 82, pp. 115774, 2020.
- [7] Haisheng Fu, Feng Liang, Jie Liang, Binglin Li, Guohe Zhang, and Jingning Han, "Asymmetric learned image compression with multi-scale residual block, importance scaling, and post-quantization filtering," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 4309–4321, 2023.
- [8] Tong Chen, Haojie Liu, Zhan Ma, Qiu Shen, Xun Cao, and Yao Wang, "End-to-end learnt image compression via non-local attention optimization and improved context modeling," *IEEE Transactions on Image Processing*, vol. 30, pp. 3179–3191, 2021.
- [9] Mu Li, Kai Zhang, Jinxing Li, Wangmeng Zuo, Radu Timofte, and David Zhang, "Learning context-based nonlocal entropy modeling for image compression," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 3, pp. 1132–1145, 2023.
- [10] Yueqi Xie, Ka Leong Cheng, and Qifeng Chen, "Enhanced invertible encoding for learned image compression," in *Proceedings of the ACM International Conference on Multimedia*, 2021, pp. 162–170.
- [11] Yinhao Zhu, Yang Yang, and Taco Cohen, "Transformer-based transform coding," in *International Conference on Learning Representations*, 2022.
- [12] Renjie Zou, Chunfeng Song, and Zhaoxiang Zhang, "The devil is in the details: Window-based attention for image compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 17492–17501.
- [13] Haichuan Ma, Dong Liu, Ning Yan, Houqiang Li, and Feng Wu, "End-to-end optimized versatile image compression with wavelet-like transform," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 3, pp. 1247–1263, 2022.
- [14] Haisheng Fu, Jie Liang, Zhenman Fang, Jingning Han, Feng Liang, and Guohe Zhang, "Weconvene: Learned image compression with wavelet-domain convolution and entropy model," in *Computer Vision – ECCV 2024*, 2025, pp. 37–53.
- [15] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," in *International Conference on Learning Representations*, 2018, pp. 1–23.
- [16] David Minnen, Johannes Ballé, and George D Toderici, "Joint autoregressive and hierarchical priors for learned image compression," in *Advances in Neural Information Processing Systems*, 2018, pp. 10794–10803.
- [17] A. van den Oord, N. Kalchbrenner, O. Vinyals, L. Espeholt, A. Graves, and K. Kavukcuoglu, "Conditional image generation with pixelcnn decoders," in *Advances in Neural Information Processing Systems (NIPS)*, 2016, pp. 4797–4805.
- [18] 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 (CVPR)*, 2018, pp. 3214–3223.
- [19] 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.
- [20] Haisheng Fu, Feng Liang, Jianping Lin, Bing Li, Mohammad Akbari, Jie Liang, Guohe Zhang, Dong Liu, Chengjie Tu, and Jingning Han, "Learned image compression with gaussian-laplacian-logistic mixture model and concatenated residual modules," *IEEE Transactions on Image Processing*, vol. 32, pp. 2063–2076, 2023.
- [21] Jinming Liu, Heming Sun, and Jiro Katto, "Learned image compression with mixed transformer-cnn architectures," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 14388–14397.
- [22] Shaojie Li, Mingbao Lin, Yan Wang, Yongjian Wu, Yonghong Tian, Ling Shao, and Rongrong Ji, "Distilling a powerful student model via online knowledge distillation," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–10, 2022.
- [23] Eric Lei, Yiğit Berkay Uslu, Hamed Hassani, and Shirin Saeedi Bidokhti, "Text+ sketch: Image compression at ultra low rates," in *ICML 2023 Workshop on Neural Compression: From Information Theory to Applications*, 2023.
- [24] Chunyi Li, Guo Lu, Donghui Feng, Haoning Wu, Zicheng Zhang, Xiaohong Liu, Guangtao Zhai, Weisi Lin, and Wenjun Zhang, "Misc: Ultra-low bitrate image semantic compression driven by large multimodal model," 2024.
- [25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016, pp. 770–778.
- [26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022.
- [27] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," 2019.
- [28] David Minnen and Saurabh Singh, "Channel-wise autoregressive entropy models for learned image compression," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 3339–3343.
- [29] Haisheng Fu, Feng Liang, Jie Liang, Zhenman Fang, Guohe Zhang, and Jingning Han, "Learned image compression with dual-branch encoder and conditional information coding," in *2024 Data Compression Conference (DCC)*, 2024, pp. 173–182.
- [30] Jiaying Liu, Dong Liu, Wenhan Yang, Sifeng Xia, Xiaoshuai Zhang, and Yuanying Dai, "A comprehensive benchmark for single image compression artifact reduction," *IEEE Transactions on Image Processing*, vol. 29, pp. 7845–7860, 2020.
- [31] Yichen Qian, Ming Lin, Xiuyu Sun, Zhiyu Tan, and Rong Jin, "Entroformer: A transformer-based entropy model for learned image compression," in *International Conference on Learning Representations*, May 2022.
- [32] 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)*, June 2022, pp. 5718–5727.
- [33] Yueyu Hu, Wenhan Yang, Zhan Ma, and Jiaying Liu, "Learning end-to-end lossy image compression: A benchmark," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.
- [34] Dailan He, Yaoyan Zheng, Baocheng Sun, Yan Wang, and Hongwei Qin, "Checkerboard context model for efficient learned image compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021, pp. 14771–14780.
- [35] G. Bjontegaard, "Calculation of average PSNR differences between RD curves," 2001, VCEG-M33.