FDLC: Fast Decoupled Lossless Compression with State Space Models

Pei Wu[1,2], Xueming Fu[1,2], S. Kevin Zhou[1,2,3,4,*]
[1]*School of Biomedical Engineering, Division of Life Sciences and Medicine*,
University of Science and Technology of China (USTC), Hefei Anhui, China
[2]*Center for Medical Imaging, Robotics, Analytic Computing & Learning (MIRACLE)*,
Suzhou Institute for Advance Research, USTC, Suzhou Jiangsu, China
[3]*Jiangsu Provincial Key Laboratory of Multimodal Digital Twin Technology*, Suzhou Jiangsu, China
[4]*State Key Laboratory of Precision and Intelligent Chemistry, USTC*, Hefei Anhui, China
[*]Corresponding Author: skevinzhou@ustc.edu.cn

***Abstract*—Lossless image compression is essential for scientific and medical applications where preserving data accuracy is critical. Recent deep learning methods employ a decoupled framework, combining a lossy compressor with a residual compressor, to achieve high lossless compression rates while enabling efficient lossy reconstruction for practical applications. However, existing methods rely on computationally expensive Transformer components, hindering practical deployment. To address this limitation, we propose Fast Decoupled Lossless Compression (FDLC), the first architecture based on State Space Models that leverages their linear computational complexity. FDLC integrates Visual State Space blocks as the core component for effectively capturing global dependencies, generating compact latent representations, and accurately modeling the probability distribution of residuals. Experiments demonstrate that our model achieves a new state-of-the-art compression effectiveness on diverse datasets. Moreover, it achieves average speedups of $1.3\times$ for compression and $1.6\times$ for decompression compared to Transformer-based counterparts, demonstrating superior efficiency and effectiveness.**

***Index Terms*—Lossless Image Compression, State Space Models, Lossy Plus Residual Coding**

I. Introduction

Modern imaging technologies in precision medicine and life sciences—such as digital pathology, functional Magnetic Resonance Imaging, and Cryo-Electron Microscopy—generate massive-scale image data, with a single brain slice reaching several gigabytes. This explosive data growth necessitates efficient compression for storage and transmission. However, compression quality is equally critical: even the slightest loss of detail can lead to severe misinterpretations, directly affecting clinical diagnoses and scientific discoveries. Therefore, lossless image compression has become essential for preserving data fidelity in these critical applications.

Traditional lossless compression algorithms, including PNG [1], CALIC [2], JPEG-LS [3], and JPEG2000 [4], rely on predefined mathematical models to remove redundancy. For instance, JPEG2000 utilizes a reversible integer wavelet transform to decompose the image into frequency sub-bands, followed by entropy coding. While effective for general purposes, these methods struggle to capture the complex statistical characteristics inherent in specific imaging modalities. Their hand-crafted designs limit adaptability and prevent them from approaching the theoretical compression limits for domain-specific data.

Deep learning has fundamentally transformed lossless compression by learning complex probability distributions directly from data, demonstrating significant potential for improved compression rates. However, early deep learning methods [5]–[10] —predominantly based on predictive coding—achieved limited compression gains and failed to support lossy reconstruction, which is essential for many practical scenarios. For instance, in data center workflows, fast and bandwidth-efficient image previews generated from lossy reconstructions are highly desirable before full lossless transmission. A compression framework that provides such lossy previews while guaranteeing fully lossless reconstruction would therefore offer substantial practical value.

Recent work [11] addresses this challenge through *a decoupled lossless compression* framework comprising two components: *a lossy compressor* that supports efficient lossy reconstruction, and *a residual compressor* that encodes the difference between the original and lossy images to enable perfect lossless recovery. This architecture elegantly balances both requirements—achieving state-of-the-art lossless compression rates while enabling fast lossy reconstruction for practical applications. However, existing implementations rely heavily on computationally expensive Transformer [12] components, resulting in substantial computational overhead that hinders deployment in real-world, time-sensitive scenarios.

To address this limitation, we propose **F**ast **D**ecoupled **L**ossless **C**ompression (FDLC), a novel architecture based on State Space Models (SSMs) [13]–[15] that exploits their linear complexity. FDLC integrates Visual State Space (VSS) [16] blocks as its core component to efficiently capture global dependencies, generate compact latent representations, and accurately model residual probability distributions. Our VSS-based compressors efficiently remove redundancies and while delivering $1.3\times$ and $1.6\times$ average speedups in compression and decompression times, respectively, compared to Transformer-based counterpart. Furthermore, in the entropy model, we address redundancies from both spatial and chan-

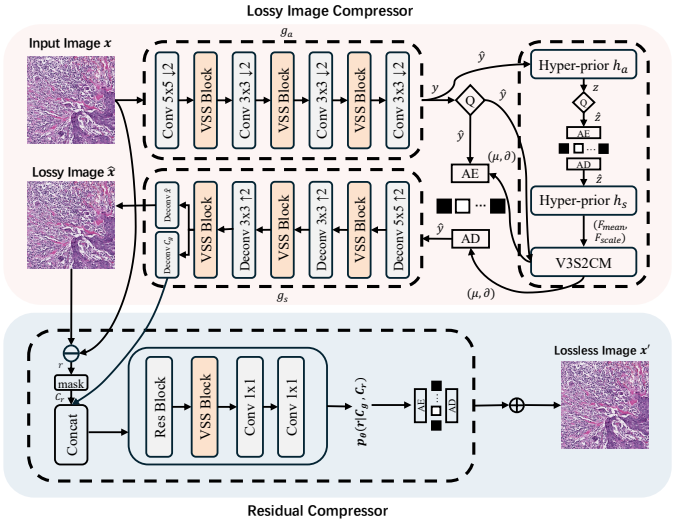


Fig. 1: Overview of proposed FDLC. AE, AD, and Q represent Arithmetic Encoding, Arithmetic Decoding, and Quantization, respectively. ↑2 or ↓2 means that the feature size is enlarged/reduced by two times compared to the previous layer.

nel perspectives by proposing a VSS-based Spatial-Channel Context Model.

In summary, our contributions are as follows: (1) To our best knowledge, we are the first to propose a SSM-based lossless compression with a decoupled architecture, which supports both lossless and lossy compression. (2) We introduce an effective VSS-based Spatial-Channel Context Model that exploits redundancy across both spatial and channel dimensions for precise entropy coding. (3) Extensive experiments demonstrate that our decoupled architecture achieves state-of-the-art performance across multiple datasets, consistently delivering higher compression efficiency and effectiveness compared with Transformer-based counterparts.

II. Related Work

A. Learned Lossless Image Compression

Lossless compression combines statistical modeling with entropy coding (e.g., arithmetic coding [17]). Deep generative models emerged to address the probability modeling challenge, initially led by autoregressive works like PixelRNN [18] and PixelCNN [19]. Although these methods offer high compression rates, their reliance on serial sampling causes prohibitive latency.

To improve efficiency, research expanded into parallelizable architectures. Flow-based models (e.g., IDF [20], iFlow [9]) utilize invertible transforms to map images to latent spaces for parallel inference. Simultaneously, VAE-based approaches introduced the ”bits-back” coding scheme (BB-ANS) [21], with methods like L3C [5] and HiLLoC [6] employing hierarchical structures to refine probabilistic modeling. However, these approaches often suffer from high implementation complexity or rigid architectural constraints, such as the bijection required in flows.

B. Learned Lossy Image Compression

Learned Lossy Image Compression has made remarkable strides, surpassing traditional codecs like VVC in objective metrics. Mainstream methods rely on an end-to-end Variational Autoencoder (VAE) framework, where early works primarily employed Convolutional Neural Networks (CNNs) combined with Generalized Divisive Normalization (GDN) [22] to eliminate spatial redundancy. To improve rate-distortion performance, research focus initially shifted toward advanced entropy modeling. The introduction of the hyperprior [23] enabled the capture of global spatial dependencies, a capability further refined by autoregressive context models [24] and discretized Gaussian Mixture Models (GMMs) [25] that exploit local correlations for precise entropy estimation. However, the serial nature of autoregressive decoding created a severe computational bottleneck, prompting the development of parallelizable alternatives, such as checkerboard [26] and channel-wise [24] context models, to enhance inference speed.

Simultaneously, the transform architecture evolved to address the limited receptive field of CNNs. Transformer-based models [27] utilizing self-attention mechanisms were introduced to effectively model long-range dependencies, yielding significant performance gains. Despite these advantages, the quadratic computational complexity of Transformers poses challenges for high-resolution inputs. Consequently, recent studies have explored hybrid CNN-Transformer architectures [28] to reconcile the trade-off between compression performance and computational efficiency, setting the stage for linear-complexity solutions like State Space Models.

C. Decoupled Paradigm for both Lossy and Lossless Compression

While traditional codecs offer versatile support for both lossy and lossless modes, deep learning approaches typically lack such flexibility. To bridge this gap, recent works like DLPR [11] have adopted a decoupled paradigm, decomposing the task into a coarse-to-fine hierarchy consisting of a lossy base compressor and a residual coder. However, these methods often suffer from high computational overhead due to the quadratic complexity of their underlying Transformer architectures. Identifying this quadratic complexity as the primary bottleneck, we propose a novel framework anchored in State Space Models (SSMs). Our method incorporates a VSS-based Spatial-Channel Context Model to capture long-range dependencies with linear complexity, effectively balancing high-fidelity reconstruction with practical inference speeds. Experiments demonstrate that our approach surpasses state-of-the-art methods in both bitrate savings and computational efficiency.

D. State Space Models

State Space Models (SSMs) [13], [29] have emerged as a powerful alternative for modeling long-range dependencies with linear computational complexity. The Structured State

Space model (S4) [14] laid the foundation for efficient sequence modeling, while Mamba [15] introduced an input-dependent selection mechanism to further enhance performance.

Given their efficiency, SSMs have recently been adapted for computer vision. Architectures such as MambaVision [30] and VMamba [31] incorporate bi-directional scanning and 2D-selective mechanisms to capture visual contexts effectively. Other variants, including LocalMamba [32], optimize scanning strategies for multi-scale feature learning. Despite their success in discriminative tasks, the potential of SSMs in lossless image compression remains largely unexplored. Our work bridges this gap by leveraging SSMs to achieve state-of-the-art compression performance with superior inference efficiency.

III. Methods

A. Overview

Fig. 1 illustrates our proposed framework of FDLC, which comprises two key components: a Lossy Compressor and a Residual Compressor. The lossy compressor is responsible for the image compression and lossy decompression of the image. The residual then encodes the residual between the original and the lossy image to enable lossless reconstruction.

B. Lossy Compressor

The Lossy Compressor (LC) comprises a main encoder g_a, a main decoder g_s, and an entropy model. During compression, the encoder g_a transforms an input image x into a compact latent representation y, effectively removing its spatial redundancy. The latents y are then quantized to $\hat{y}$. Concurrently, the unquantized representation y provides global context to the entropy model, which in turn generates a precise probability distribution, parameterized by μ and σ. Finally, an arithmetic coder compresses $\hat{y}$ into a bitstream using this probability distribution. During decompression, the arithmetic decoder reconstructs the quantized latent representation $\hat{y}$ from the bitstream using the transmitted probability parameters (μ, σ). The main decoder g_s then uses $\hat{y}$ to synthesize the final lossy image $\hat{x}$.

Main Encoder. Prior work has employed Transformers to capture long-range dependencies, introducing substantial computational overhead. Motivated by the parameter efficiency and long-range modeling capabilities of SSMs, we adopt VSS blocks as our foundational modules for main encoder g_a. The VSS block(Fig. 2) comprises layer normalization, linear projection, depthwise convolution, SiLU activation, and the core 2D Selective Scan (SS2D) mechanism. Formally, given an input tensor $T \in \mathbb{R}^{H\times W\times C}$, the computation within a VSS block is defined as:

$$\begin{aligned} T_{main} &= \text{LN}(\text{SS2D}(f(T))) \\ T_{gate} &= \text{Linear}(T) \\ T_{out} &= \text{Linear}(T_{main} \odot T_{gate}) + T \end{aligned} \tag{1}$$

where LN denotes Layer Normalization. DWConv is a depth-wise convolution, $\odot$ denotes element-wise multiplication, and f is defined as $f = SiLU \circ DWConv \circ Linear \circ LN$.

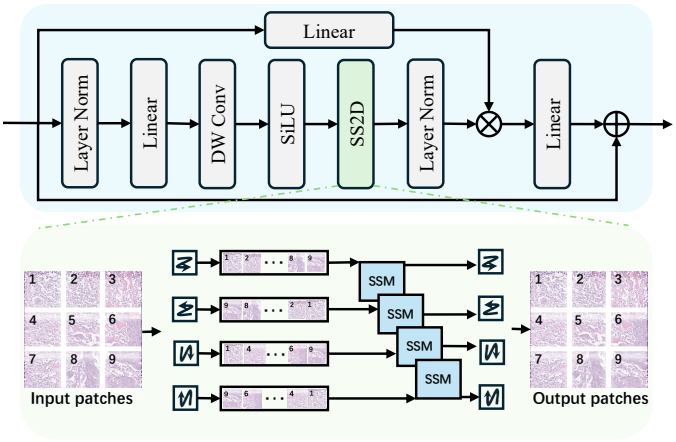


Fig. 2: Structure of Visual State Space (VSS) block with 2D Selective Scan (SS2D). Top: The overall architecture, consisting of a depth-wise convolution and the core SS2D module. $Bottom$: Detailed view of the SS2D operation, where input patches are traversed along four scanning paths, processed by SSMs, and aggregated to form the output features.

SS2D scans input patches along four distinct directional paths, processes each with a separate SSM, and merges them into a unified 2D output. This cross-directional scanning effectively captures global spatial dependencies, enabling efficient redundancy removal and compact latent representations for entropy coding. Specifically, the core recursive operation within each SSM path is governed by the following discrete-time state-space equations:

$$\begin{aligned} \bar{\mathbf{B}}_t &= \Delta_t \mathbf{B}, \\ \mathbf{s}_t &= \exp(\Delta_t \mathbf{A})\mathbf{s}_{t-1} + \bar{\mathbf{B}}_t \mathbf{i}_t, \\ \mathbf{o}_t &= \mathbf{W}_{out}\mathbf{s}_t + \mathbf{D}\mathbf{i}_t, \end{aligned} \tag{2}$$

where $\mathbf{i}_t$ and $\mathbf{o}_t$ denote the input and output features at step t, and $\mathbf{s}_t$ represents the hidden state. $\mathbf{A}$, $\mathbf{B}$, $\mathbf{W}_{out}$, and $\mathbf{D}$ are the system parameters: $\mathbf{A}$ and $\mathbf{B}$ govern the state evolution, $\mathbf{W}_{out}$ performs the output projection, and $\mathbf{D}$ provides a skip connection. The discrete parameter $\bar{\mathbf{B}}_t$ is derived from its continuous counterpart $\mathbf{B}$ via the dynamic timescale Δ_t.

Entropy Model. While the main encoder g_a produces compact latent representations, redundancy still exists across spatial and channel dimensions, particularly in biomedical images. To further exploit this structure for efficient entropy coding, following [23], we adopt a hyper-prior encoder-decoder to provide global context from g_a and propose VSS Spatial-Channel Context Model (V3S2CM), which integrates the global context with local spatial-channel context to model the conditional probability distribution of latent variables, which we assume to be a Gaussian: $p(\hat{y}|\hat{z}) \sim \mathcal{N}(\mu, \sigma^2)$.

$$z = h_a(y),\ \hat{z} = Q(z),\ (F_{mean}, F_{scale}) = h_s(\hat{z}) \tag{3}$$

$$(\mu, \sigma) = V3S2CM(\hat{y}, F_{mean}, F_{scale}) \tag{4}$$

where h_a and h_s represent the hyper-prior encoder and decoder, respectively. The encoder h_a extracts side information

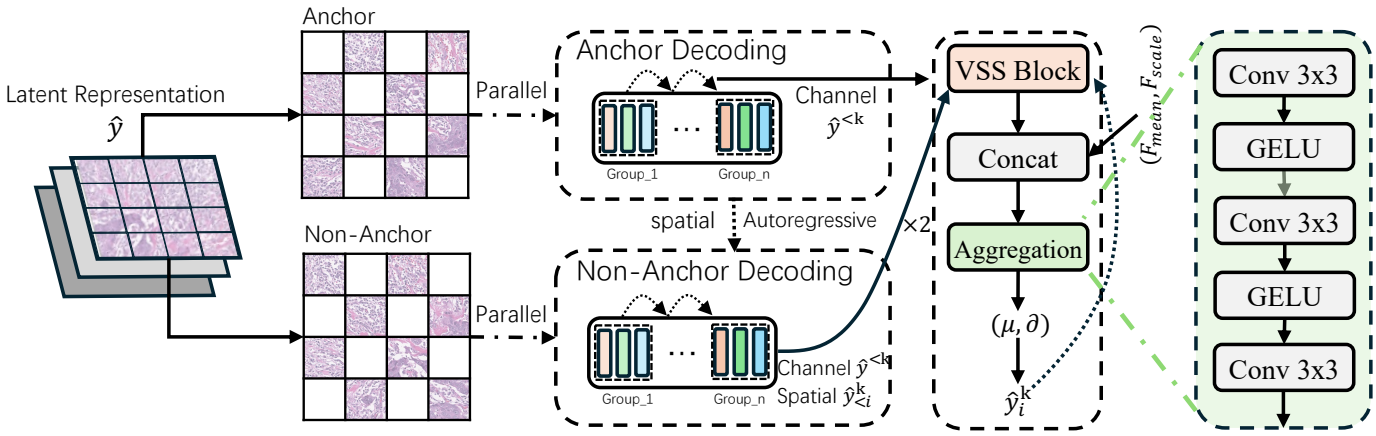


Fig. 3: Architecture of the proposed VSS Spatial-Channel Context Model (V3S2CM). The V3S2CM ($right$) utilizes a VSS Block to extract local context, which is then concatenated with global features and processed by the Aggregation module ($left$) to predict parameters.

z from y to capture underlying spatial dependencies. Subsequently, the quantized hyper-latent $\hat{z}$ is decoded by h_s to yield global context features (F_{mean}, F_{scale}), which serve as conditions for the following entropy estimation.

Spatially, V3S2CM adopts a checkerboard pattern to partition latent representation $\hat{y}$ into anchor and non-anchor groups, capturing spatial dependencies while enabling parallel decoding. In the channel dimension, we employ autoregressive channel grouping to progressively refine probability estimates using previously decoded channel groups. As detailed in Fig. 3, during anchor decoding, the context from previously decoded channel groups is processed by the VSS Block. This local feature is then concatenated with the global features (F_{mean}, F_{scale}) and passed to the Aggregation module to estimate the probability parameters for the anchor locations. Conversely, non-anchor decoding additionally incorporates spatial neighbors from decoded anchors into the input. By utilizing channel context to compensate for reduced spatial dependencies, this design breaks strict seriality to enable high-degree parallelism, achieving significantly faster inference with minimal performance compromise.

Main Decoder. Diverging from conventional designs that only reverse the encoding process, our main decoder g_s produces two outputs from the quantized latents $\hat{y}$. It simultaneously reconstructs the lossy image $\hat{x}$ and extracts a global context feature C_g for the residual compressor. Both outputs branch only at the final layer, as expressed by:

$$\hat{x}, C_g = g_s(\hat{y}) \tag{5}$$

This design allows C_g to capture high-level structural information while maintaining computational efficiency through parameter sharing.

C. Residual Compressor

To achieve lossless compression, the residual compressor encodes the residual $r = x - \hat{x}$ between the original and lossy reconstructed images. By losslessly compressing this residual, the framework enables perfect reconstruction of the original image as $x' = \hat{x} + r$. The residual compressor predicts the probability mass function (PMF) of r and encodes it using arithmetic coding. The prediction network also incorporates VSS blocks, consistent with the lossy compressor architecture.

To achieve accurate probability estimation, we provide the network with two types of side information: global context C_g from the synthesis transform and local context C_r from previously decoded residuals. We hypothesize that C_g encapsulates high-level structural information about r conditioned on $\hat{y}$. However, C_g alone is insufficient to exploit spatial redundancies in r. Therefore, C_r provides complementary local context from already decoded residual regions, enabling more precise probability modeling. To facilitate parallel training, we employ masked convolutions to simulate the autoregressive decoding process. The PMF of the residual can be expressed as $P_\theta(r \mid C_g, C_r)$, and the rate for the residual compressor (RC) is formulated as:

$$R(r) = \mathbb{E}\left[-\log_2 P_\theta(r \mid C_g, C_r)\right] \tag{6}$$

D. Loss Function

During training, we jointly optimize the lossy compressor and residual compressor to achieve globally optimal compression efficiency. Combining the rate terms from both components, the overall loss function is formulated as:

$$\mathcal{L} = R(\hat{y}) + R(\hat{z}) + R(r) + \lambda D(x, \hat{x}) \tag{7}$$

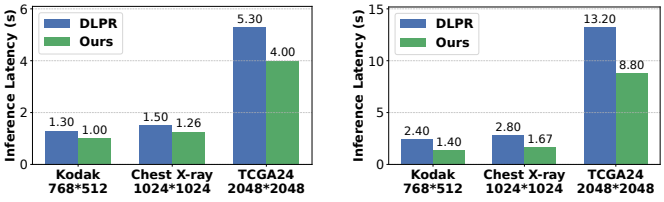
where $R(\hat{y})$, $R(\hat{z})$, and $R(r)$ represent the bitrates for the latent representation, hyper-prior, and residual, respectively, and $D(x, \hat{x})$ measures the distortion between the original and reconstructed images using MSE loss. Expanding the rate terms yields:

$$\begin{aligned}\mathcal{L} =& \mathbb{E}\left[-\log_2 P_{\hat{y}|\hat{z}}(\hat{y}|\hat{z})\right] + \mathbb{E}\left[-\log_2 P_{\hat{z}}(\hat{z})\right] + \\ & \mathbb{E}\left[-\log_2 P_\theta(r \mid C_g, C_r)\right] + \lambda D(x, \hat{x})\end{aligned} \tag{8}$$

The hyperparameter λ controls the rate-distortion trade-off.

TABLE I: Comparison of performance in BPSP(↓). Bold font denotes the best performance among all methods. ”-” indicates an unavailable result.

Method	Chest X-ray	TCGA24	Kodak	CLIC
Traditional Compression				
PNG [1]	3.07	3.95	4.35	3.93
CALIC [2]	2.98	3.20	3.18	2.87
JPEG-LS [3]	2.88	3.05	3.16	2.82
JPEG2000 [4]	2.90	2.95	3.19	2.93
FLIF [33]	2.87	2.65	2.90	2.72
JPEG-XL [34]	2.93	2.55	2.87	2.63
Lossless-only Compression				
L3C [5]	2.89	-	3.26	2.94
HiLLoC [6]	3.02	-	-	-
LBB [7]	2.76	-	-	-
iVPF [8]	2.8	-	-	2.54
iFlow [9]	2.77	-	-	2.44
LVPNet [10]	2.65	-	-	-
Decoupled Compression				
DLPR [11]	1.29	1.75	2.86	2.38
Ours	**1.23**	**1.68**	**2.67**	**2.36**



(a) Compression Latency (b) Decompression Latency

Fig. 4: Computational Efficiency

IV. Experiments

A. Setup

Traning. Our model was trained on the DIV2K dataset, which contains 800 high-resolution (2K) images. We first cropped these images into 121,379 non-overlapping 128×128 patches. Data augmentation was performed on these patches through random horizontal and vertical flips, followed by a final random crop to a size of 64×64. We utilized the Adam optimizer with a batch size of 16 and trained the model for 600 epochs. Our framework is implemented in PyTorch, and all experiments were conducted on NVIDIA RTX 3090 GPUs.

Evaluation. We evaluated our proposed model's performance on four distinct datasets. These include a Chest X-ray dataset [35] consisting of 1,000 images at 1024×1024 resolution; the TCGA24 dataset, with 24 randomly selected images from the broader TCGA dataset [36], subsequently patched into 2K resolution; the CLIC Professional Validation dataset [37] , containing 41 images with resolutions up to 2K; the Kodak dataset [38], with 24 images at 768×512 resolution. Compression performance was measured in Bits Per SubPixel (BPSP).

TABLE II: Ablation study on the SSM block. The performance is compared against variants using Transformer blocks.

Main Blocks	Comp. latency(s)	Decomp. latency(s)	Bitrate(BPSP)
Kodak			
Transformer	1.25	2.35	2.83
Ours	**1.00**	**1.40**	**2.67**
Chest X-ray			
Transformer	1.48	2.67	1.32
Ours	**1.26**	**1.67**	**1.23**

B. Performance Comparison

Compression Effectiveness. As a decoupled compression method, we compare our method with the most direct baseline, DLPR, which also provides lossless compression while supporting lossy image preview. As shown in Table I, our method achieves lower bitrates, which indicates higher compression performance across all datasets. For example, Our method achieves 4% and 7% BPSP reductions on the Chest X-ray and Kodak datasets, respectively. To comprehensively demonstrate the superiority of our method, we further compare with early deep learning-based approaches that only support lossless compression, as well as with traditional compression methods. Our method still achieves substantial improvements. For instance, on the Chest X-ray dataset, our method achieves a 2.15× improvement over the lossless-only LVPNet method (2.65 → 1.23), and an even more remarkable 2.5× improvement over the widely traditional PNG compression (3.07 → 1.23), highlighting the effectiveness of our method.

Compression Efficiency. We further evaluate the efficiency gains brought by the linear computational complexity of the SSM-based architecture across datasets of varying resolutions. Specifically, the Kodak, Chest X-ray, and TCGA24 datasets cover a progression from low to high resolutions, ranging from 768×512, 1024×1024, to 2048×2048. Our method consistently demonstrates substantially higher efficiency in both compression and decompression times. For instance, on the Chest X-ray dataset, our method achieves 1.1× and 1.67× speedups in compression and decompression latency, respectively, compared to DLPR (Fig. 4). These results highlight the superiority of our SSM-based backbone, which not only enhances compression performance but also significantly boosts computational efficiency.

C. Residual Visualization

To verify that our decoupled architecture satisfies the requirements of practical lossy compression scenarios, we visualize the lossy reconstructions and corresponding residual maps in Fig. 5. We present two compression examples from the TCGA24 biomedical dataset and the CLIC natural image dataset. While our counterpart DLPR generally meets practical quality requirements, our approach still excels in fine detail preservation. In the biomedical example from TCGA24, images compressed by both our method and DLPR appear visually indistinguishable from the original. However, quantitative

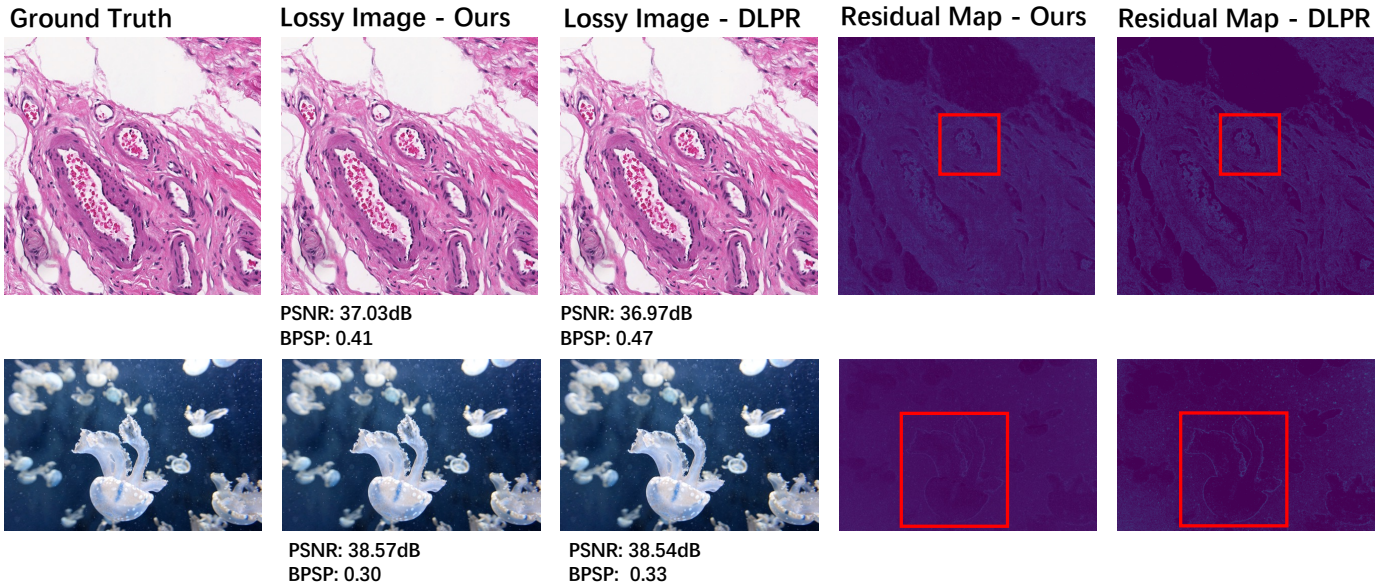


Fig. 5: Visual comparison on TCGA24 (*top*) and CLIC (*bottom*) datasets. Ours achieves higher PSNR and lower BPSP than DLPR. Red boxes highlight our superior residual sparsity and reduced structural artifacts.

metrics reveal that our method reduces BPSP from 0.47 to 0.41 relative to DLPR, while improving PSNR from 36.97 dB to 37.03 dB. We highlight these differences using residual maps calculated as the average absolute difference across channels. In the red-boxed regions, our method exhibits significantly lower reconstruction errors. This visual evidence, combined with improved PSNR values, confirms our method's superior performance in lossy compression.

D. Ablation Study

Comprehensive ablation studies were performed on the Kodak and Chest X-ray datasets to validate the effectiveness of our core module. As detailed in Table II, we substituted the VSS blocks with Transformer blocks while keeping all other configurations fixed to ensure a fair comparison. The results indicate that our SSM block demonstrates superior compression performance, achieving the lowest BPSP across both datasets. Notably, our approach strikes an optimal balance between efficiency and performance, as it significantly outperforms the Transformer variant in compression ratio while maintaining competitive inference speeds and capturing long-range dependencies more efficiently than standard attention mechanisms.

V. Conclusion

In this work, we introduce fast decoupled lossless compression, called FDLC, the first State Space Model-based architecture for lossless image compression. We leverage highly efficient Visual State Space blocks to form the core, which model global context at linear computational cost—a significant advantage over prior Transformer-based designs. This is complemented by our VSS-based Spatial-Channel Context Model, which further minimizes redundancy for precise entropy coding. Experiments confirm that FDLC sets a new state-of-the-art in compression effectiveness across multiple datasets. However, we recognize that there is still room for improvement when deploying FDLC to specialized biomedical domains, as an amortization gap inevitably exists. While addressing this domain shift remains a primary focus of our future work, by uniting top-tier performance with high efficiency and the support for lossy previews, FDLC ultimately offers a practical and powerful solution for modern large-scale imaging.

References

[1] T. Boutell, "Png (portable network graphics) specification version 1.0," Tech. Rep., 1997.

[2] X. Wu and N. Memon, "Calic-a context based adaptive lossless image codec," in *1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings*, vol. 4. IEEE, 1996, pp. 1890–1893.

[3] S. SECTOR and O. ITU, "Information technology–lossless and near-lossless compression of continuous-tone still images–baseline," 1998.

[4] 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, 2002.

[5] F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L. V. Gool, "Practical full resolution learned lossless image compression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 10 629–10 638.

[6] J. Townsend, T. Bird, J. Kunze, and D. Barber, "Hilloc: Lossless image compression with hierarchical latent variable models," *arXiv preprint arXiv:1912.09953*, 2019.

[7] J. Ho, E. Lohn, and P. Abbeel, "Compression with flows via local bits-back coding," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[8] S. Zhang, C. Zhang, N. Kang, and Z. Li, "ivpf: Numerical invertible volume preserving flow for efficient lossless compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 620–629.

[9] S. Zhang, N. Kang, T. Ryder, and Z. Li, "iflow: Numerically invertible flows for efficient lossless compression via a uniform coder," *Advances in Neural Information Processing Systems*, vol. 34, pp. 5822–5833, 2021.

[10] C. Song, C. Hui, Q. Lin, W. Zhang, S. Li, H. Zhu, S. Zhang, Z. Li, S. Liu, F. Jiang *et al.*, "Lvpnet: A latent-variable-based prediction-driven end-to-end framework for lossless compression of medical images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2025, pp. 288–298.

[11] Y. Bai, X. Liu, K. Wang, X. Ji, X. Wu, and W. Gao, "Deep lossy plus residual coding for lossless and near-lossless image compression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 5, pp. 3577–3594, 2024.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[13] A. Gu, I. Johnson, K. Goel, K. Saab, T. Dao, A. Rudra, and C. Ré, "Combining recurrent, convolutional, and continuous-time models with linear state space layers," *Advances in neural information processing systems*, vol. 34, pp. 572–585, 2021.

[14] A. Gu, K. Goel, and C. Ré, "Efficiently modeling long sequences with structured state spaces," *arXiv preprint arXiv:2111.00396*, 2021.

[15] A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," in *First conference on language modeling*, 2024.

[16] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, and X. Wang, "Vision mamba: Efficient visual representation learning with bidirectional state space model," *arXiv preprint arXiv:2401.09417*, 2024.

[17] I. H. Witten, R. M. Neal, and J. G. Cleary, "Arithmetic coding for data compression," *Communications of the ACM*, vol. 30, no. 6, pp. 520–540, 1987.

[18] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *International conference on machine learning*. PMLR, 2016, pp. 1747–1756.

[19] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves *et al.*, "Conditional image generation with pixelcnn decoders," *Advances in neural information processing systems*, vol. 29, 2016.

[20] E. Hoogeboom, J. Peters, R. Van Den Berg, and M. Welling, "Integer discrete flows and lossless compression," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[21] J. Townsend, T. Bird, and D. Barber, "Practical lossless compression with latent variables using bits back coding," *arXiv preprint arXiv:1901.04866*, 2019.

[22] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," *arXiv preprint arXiv:1611.01704*, 2016.

[23] J. Ballé, D. Minnen, S. Singh, S. J. Hwang, and N. Johnston, "Variational image compression with a scale hyperprior," *arXiv preprint arXiv:1802.01436*, 2018.

[24] D. Minnen and S. Singh, "Channel-wise autoregressive entropy models for learned image compression," in *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2020, pp. 3339–3343.

[25] Z. Cheng, H. Sun, M. Takeuchi, and J. 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*, 2020, pp. 7939–7948.

[26] D. He, Y. Zheng, B. Sun, Y. Wang, and H. Qin, "Checkerboard context model for efficient learned image compression," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 771–14 780.

[27] Y. Zhu, Y. Yang, and T. Cohen, "Transformer-based transform coding," in *International conference on learning representations*, 2022.

[28] J. Liu, H. Sun, and J. Katto, "Learned image compression with mixed transformer-cnn architectures," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 14 388–14 397.

[29] H. Mehta, A. Gupta, A. Cutkosky, and B. Neyshabur, "Long range language modeling via gated state spaces," *arXiv preprint arXiv:2206.13947*, 2022.

[30] A. Hatamizadeh and J. Kautz, "Mambavision: A hybrid mamba-transformer vision backbone," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 25 261–25 270.

[31] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, and Y. Liu, "Vmamba: Visual state space model," *Advances in neural information processing systems*, vol. 37, pp. 103 031–103 063, 2024.

[32] T. Huang, X. Pei, S. You, F. Wang, C. Qian, and C. Xu, "Localmamba: Visual state space model with windowed selective scan," in *European Conference on Computer Vision*. Springer, 2024, pp. 12–22.

[33] J. Sneyers and P. Wuille, "Flif: Free lossless image format based on maniac compression," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 66–70.

[34] J. Alakuijala *et al.*, "Jpeg xl next-generation image compression architecture and coding tools," in *Applications of digital image processing XLII*, vol. 11137. SPIE, 2019, pp. 112–124.

[35] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2097–2106.

[36] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, "Review the cancer genome atlas (tcga): an immeasurable source of knowledge," *Contemporary Oncology/Współczesna Onkologia*, vol. 2015, no. 1, pp. 68–77, 2015.

[37] L. Theis and G. Toderici, "Clic, workshop and challenge on learned image compression," in *CVPR*, 2020.

[38] E. Kodak, "Kodak lossless true color image suite (photocd pcd0992)," *URL http://r0k. us/graphics/kodak*, vol. 6, no. 2, p. 5, 1993.