RSTSIC: Reparameterized Swin Transformer for Stereo Image Compression

Yuxuan Zhao HARVEY2077@WHUT.EDU.CN
Jiaxin Li LJXXX@WHUT.EDU.CN
Cheng Tan CHENG_TAN@WHUT.EDU.CN
Jianwen Xiang JWXIANG@WHUT.EDU.CN
Junwei Zhou* JUNWEIZHOU@MSN.COM
School of Computer Science and Artificial Intelligence, Wuhan University of Technology, 122 Luoshi Road, Wuhan, Hubei, P.R. China

Editors: Hung-yi Lee and Tongliang Liu

Abstract

Stereo image compression (SIC) aims to enhance compression performance and efficiency by exploiting cross-view redundancy in overlapping fields between stereo images. However, current SIC methods faces practical limitations in adequately exploiting inter-view correlations and contextual information due to occlusions, disparity variations, and computational overhead. To effectively extract contextual information and efficiently model cross-view dependencies in stereo images, we propose a novel distributed stereo image compression framework, Reparameterized Swin Transformer for Stereo Image Compression (RSTSIC) integrating Reparameterized Swin Block (RSB) and Cross Feature Enhancement Modules (CFEMs) in the joint decoder. CFEMs progressively aggregate cross-view dependencies and enhance cross feature interaction efficiency. RSB integrates window-based self-attention with convolutional operations to effectively leverage non-local contextual information, while maintaining inference efficiency through structural reparameterization. RSTSIC outperforms traditional codecs and deep stereo compression methods on both Cityscapes and InStereo2K datasets, with at least 58.57% reduction in model parameters and 36.43% decrease in FLOPs compared to state-of-the-art compression models. Ablation studies confirm the necessity of CFEMs and RSB for efficient compression and perceptual fidelity. Our code is available at https://github.com/SnowBlind0/RSTSIC.

Keywords: Stereo Image Compression, Swin Transformer, Structural Reparameterization, Feature Enhancement

1. Introduction

Image compression has emerged as a critical research frontier in modern visual signal processing, driven by the exponential growth of digital media and stringent bandwidth requirements across diverse applications including live streaming platforms Tang et al. (2018); Jeong et al. (2024), medical imaging systems Ungureanu et al. (2024); Liu et al. (2017), and mobile devices Schuster et al. (2014); Kim et al. (2015). This long-standing challenge remains an active area of investigation, with ongoing innovations spanning both traditional hand-engineered codecs Wallace (1991); Skodras et al. (2001); Sullivan et al. (2012); Trow (2020) and emerging data-driven approaches Ballé et al. (2016); Deng et al. (2021); Mital et al. (2022); Zhang et al. (2023); Wödlinger et al. (2022); Liu et al. (2024). Conventional

image codecs typically implement a multi-stage processing pipeline where input images are first partitioned into independently processed spatial blocks. These blocks subsequently undergo linear spectral decorrelation through transforms like DCT or wavelets, spatial redundancy reduction via intra-block prediction mechanisms, residual error encoding to capture prediction inaccuracies, quantization for discrete symbol conversion, and entropy-optimized bitstream compression. The decoding pipeline reconstructs the image through corresponding inverse operations, while irreversible information loss occurs during quantization.

Learned compression methods offer distinct advantages over traditional approaches, primarily through their symmetric encoding-decoding computational profiles, contrasting with conventional methods that typically suffer from disproportionate encoding delays Huang and Wu (2024). Learned image compression field has been predominantly dominated by convolutional architectures since the introduction of the variational autoencoder (VAE) framework by Ballé et al. Ballé et al. (2016), adopting an end-to-end trainable network. These neural approaches Deng et al. (2021); Mital et al. (2022); Zhang et al. (2023); Liu et al. (2023, 2024) process full-resolution images holistically without spatial partitioning, learning nonlinear transforms through deep neural networks to obtain compact latent representations. The latent features undergo quantization followed by arithmetic coding using learned probability distributions, with the entire pipeline optimized through rate-distortion tradeoffs. Distributed image compression Zhang et al. (2023); Mital et al. (2022) leverages the inherent correlation between stereo images to remove redundancy information and enhance compression performance. It is typically achieved by employing a joint encoding framework where one image is compressed independently as side information, while the other image is compressed conditioned on the extracted features or disparity information. By explicitly modeling and exploiting the inter-view dependencies, distributed compression approaches can significantly reduce second image bitrate compared to independently compressing each view, leading to superior overall rate-distortion performance for stereo images.

Stereo image compression is critically important in modern applications such as autonomous driving Fan et al. (2020, 2023) and virtual reality streaming platforms Abu Alhaija et al. (2018); Hou et al. (2024), where dual-camera systems demand high compression rates and low latency to enable continuous recording, real-time streaming, and efficient management of binocular visual data while maintaining strict processing throughput requirements. The primary objective of stereo image compression is to enhance compression performance and efficiency by effectively exploiting correlated information from overlapping fields of view, while preserving content integrity and perceptual quality. However, practical implementations face fundamental limitations Deng et al. (2023); Wödlinger et al. (2022) arising from scene geometry constraints to deployment computational complexity, particularly occluded regions and divergent fields of view between stereo cameras, introducing irreducible disparity variations and spatial information disparities that degrade compression efficiency. These inherent challenges in real-world stereo systems prevent current methods from fully realizing the theoretical bitrate upper bound while maintaining acceptable reconstruction fidelity across both views. To overcome these challenges, existing learned stereo image compression approaches Liu et al. (2019); Deng et al. (2021); Lei et al. (2022) primarily focused on disparity modeling to establish dense warp fields or rigid homography transformations between stereo images. While these methods enable effective stereo compression for similarity exploitation, they suffer from several major limitations: high computational complexity,

inadequate non-local contextual information mining, and inefficient modeling of cross-view feature interactions.

To address these challenges, we propose a novel distributed stereo image compression framework, Reparameterized Swin Transformer for Stereo Image Compression (RSTSIC), synergistically integrating vision transformer capabilities with cross-view feature dependency modeling. Specifically, the proposed framework employs Cross Feature Enhancement Modules in the joint decoder to progressively and efficiently aggregate inter-view feature dependency through cross feature interaction. Reparameterized Swin Block captures long-range spatial dependencies and effectively exploits non-local contextual information while maintaining deployment efficiency through reparameterization technique during the decoding stage. The main contributions of this paper can be summarized as follows:

- To adequately extract non-local contextual information between stereo images, we propose a novel distributed stereo image compression framework, RSTSIC, integrating vision transformer capabilities with cross-view feature dependency modeling, while employing reparameterization technique to optimize inference efficiency.
- To efficiently leverage and progressively aggregate inter-view feature dependency in stereo images, we propose lightweight Cross Feature Enhancement Module integrated into the joint decoder.
- Extensive experimental results show that the proposed framework achieves competitive compression performance on both Cityscapes and InStereo2K datasets compared to state-of-the-art compression methods.

2. Related Work

The stereo image compression field has evolved significantly by leveraging advancements in stereo matching and deep learning-based compression frameworks. Early works in stereo matching laid the groundwork for understanding inter-view correlations, which is critical for effective and efficient stereo compression. Concurrently, modern deep compression approaches have redefined compression paradigms by integrating geometric reasoning, context-aware entropy modeling and attention mechanisms.

2.1. Stereo Matching

Seminal works in stereo matching established methodologies for disparity estimation, which later informed disparity-aware compression strategies. Žbontar and LeCun (2016) pioneered the field by utilizing convolutional neural network to compare image patches, formulating stereo matching as a learned similarity metric problem. Their work demonstrated the potential of deep learning for capturing complex correspondences and addressing challenging stereo compression task. Luo et al. (2016) improved computational efficiency through a shallow architecture, balancing accuracy and speed for real-time compression systems. Pyramid Stereo Matching Network (PSMNet) proposed by Chang and Chen (2018) introduced spatial pyramid pooling to aggregate multi-scale contextual information, significantly enhancing disparity estimation in occluded regions. This hierarchical feature extraction inspired later stereo compression methods to exploit multi-scale redundancies. End-to-end learning

further bridged disparity estimation and compression. Kendall et al. (2017) proposed an end-to-end deep stereo regression network, integrating geometry and context through cost volume construction and 3D convolutions. Their architecture emphasized the importance of joint geometric and semantic reasoning in stereo regression tasks. Zhong et al. (2017) introduced self-supervised learning for stereo matching, enabling iterative refinement of disparity maps without ground truth labels.

2.2. Deep Stereo Image Compression

Recent researches explicitly target stereo image compression by combining disparity-aware mechanisms with advanced entropy models and attention modules. Deng et al. (2021) proposed a deep homography-based framework, aligning stereo pairs via learned homographic transformations to reduce inter-view redundancy. Wödlinger et al. (2022) introduced SASIC, which employs latent shifts and stereo attention module to dynamically allocate bitrates between views. By focusing attention on disparity-sensitive regions, SASIC achieved state-of-the-art rate-distortion performance. Deng et al. (2023) proposed deep mask stereo image compression (MASIC) integrating semantic masks to prioritize compression of disparity-sensitive regions. By jointly learning mask-guided bit allocation and entropy modeling, MASIC achieves superior reconstruction quality in occluded areas while minimizing bit consumption. Liu et al. (2024) proposed BiSIC, a bidirectional stereo compression framework with cross-dimensional entropy model, jointly leveraging spatial, channel-wise, and inter-view dependencies. DSIC, an end-to-end stereo compression network proposed by Liu et al. (2019), jointly optimizes disparity estimation and entropy coding, and leverages learned warping and context-aware probability models to balance compression quality and efficiency. Varodayan et al. (2007) proposed a joint bitplane distributed source coding with unsupervised disparity learning, enabling efficient distributed compression of grayscale stereo images. It achieves significant rate savings compared to ignoring disparity and performs nearly as well as an impractical system with perfect disparity knowledge at the decoder. Zhang et al. (2023) proposed a distributed coding framework inspired by distributed source coding theory, enabling independent encoding and joint decoding without geometric assumptions. The progression from stereo matching to deep compression frameworks underscores a trend toward integrating geometric priors, data-driven entropy optimization and attention mechanisms.

3. Proposed Method

The section first presents an architectural overview of the proposed Reparameterized Swin Transformer for Stereo Image Compression framework. Then we dive into the design of Cross Feature Enhancement Module, which enables explicit feature enhancement between stereo images through cross feature interactions. Finally, we detail the architecture of Reparameterized Swin Block, capturing long-range dependencies and contextual information while maintaining deployment-friendly complexity through structural reparameterization.

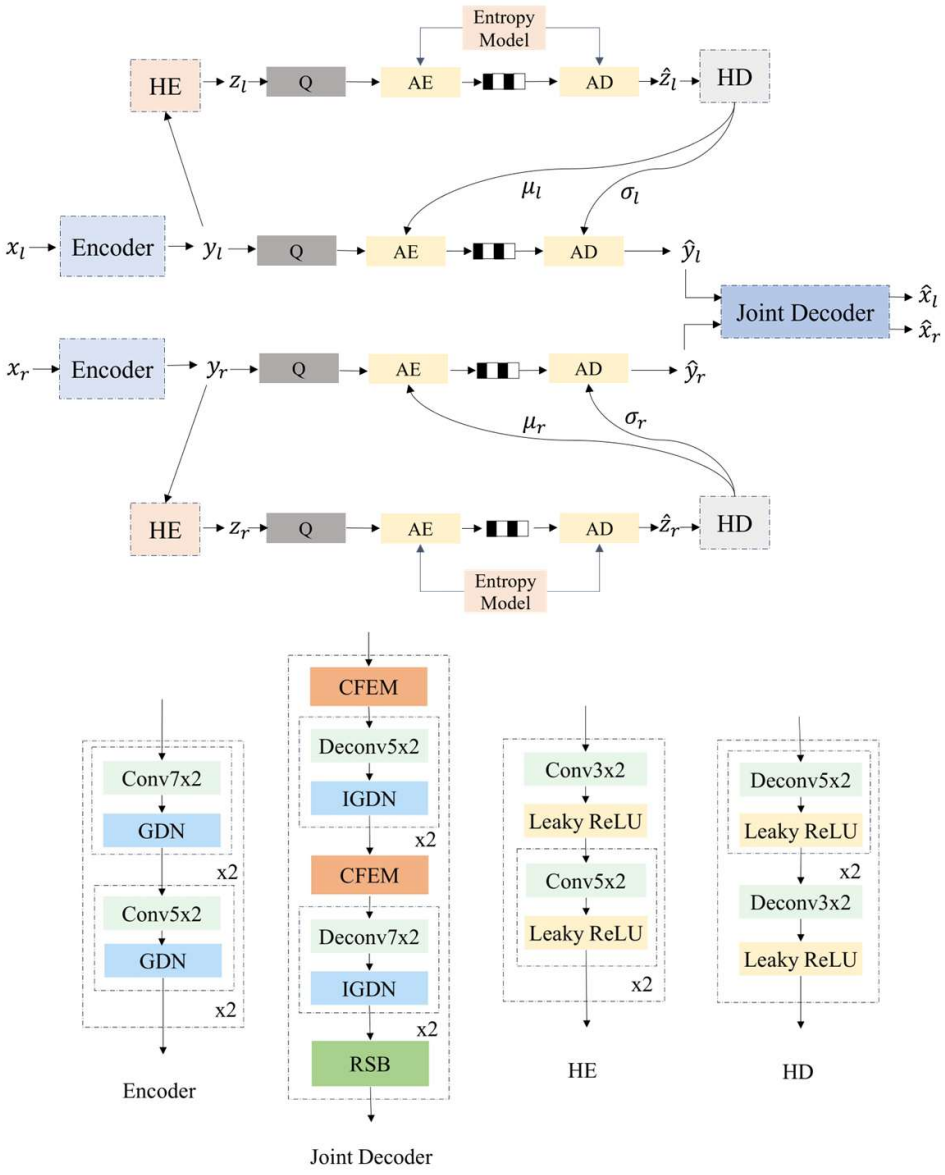

Figure 1: Architectural overview of Reparameterized Swin Transformer for Stereo Image Compression (RSTSIC) framework. The detailed architectures of Cross Feature Enhancement Module (CFEM) and Reparameterized Swin Block (RSB) are shown in Figure 2 and Figure 3. x_l denotes the left-view image, and the corresponding reconstructed image is denoted as $\hat{x}_l$. y_l represents the corresponding latent representation, while $\hat{y}_l$ represents the quantized latent representation. z_l denotes the hyper latent representation, while $\hat{z}_l$ represents the quantized hyper latent. μ_l and σ_l stand for the estimated mean and standard deviation of the distribution of y_l.Q stands for quantization, while AE and AD denote arithmetic encoder and decoder. HE and HD represent hyper encoder and decoder. Analogously for x_r, $\hat{x}_r$, y_r, $\hat{y}_r$, z_r, $\hat{z}_r$, μ_r and σ_r.

3.1. The Overall Architecture of RSTSIC

As illustrated in Figure 1, RSTSIC adopts a distributed compression architecture with independent encoder and joint decoder, integrating hyperprior network, Cross Feature Enhancement Modules (CFEMs) and Reparameterized Swin Block (RSB). The encoder employs cascaded convolutional layers to progressively downsample stereo images x_l and x_r while expanding channel dimensions, each followed by Generalized Divisive Normalization (GDN) layer to model local feature correlations and introduce nonlinearity. This hierarchical transformation produces compact latent representations y_l and y_r, which are quantized to discrete tensors $\hat{y}_l$ and $\hat{y}_r$ for entropy coding and subsequent bitstream generation. The joint decoder reconstructs compressed outputs $\hat{x}_l$ and $\hat{x}_r$ through transposed convolutions and inverse GDN layers, reversing the encoding operations. Notably, the joint decoder integrates RSB, leveraging window-based self-attention mechanism to capture long-range spatial dependencies and exploit non-local contextual information. CFEMs boost stereo feature interactions by progressively aggregating cross-view feature representations, improving reconstruction fidelity and efficiency. CFEMs enhance cross-view feature alignment, providing RSB with better-conditioned inputs for intra-view reconstruction. The synergy of CFEM and RSB leads to more effective fusion of cross-view features and improved compression performance. The hyper encoder processes y_l and y_r through multiple convolutional layers followed by Leaky ReLU activations. This transformation effectively extracts spatially enriched representations, yielding hyperprior representations z_l and z_r. The quantized hyperprior representations $\hat{z}_l$ and $\hat{z}_r$ are obtained through quantization, discretizing the continuous latent space for entropy coding. $\hat{z}_l$ and $\hat{z}_r$ are then fed into the hyper decoder to reconstruct spatially coherent distribution parameters of latent representations. Together, these components optimize the trade-off between compression rate and distortion, leveraging both spatial and statistical correlations and capturing both local and non-local contextual information between stereo images.

3.2. Cross Feature Enhancement

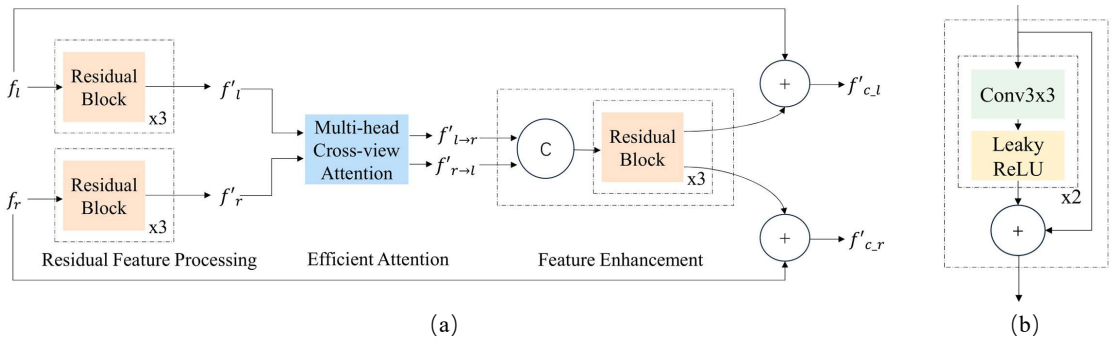

Figure 2: The architecture of Cross Feature Enhancement Module and Residual Block. f_l and f_r represent stereo features. f'_l and f'_r denote the residual processed stereo features. f'_{c_l} and f_{c_r} stand for the compact enhanced stereo features.

As illustrated in Figure 2, Cross Feature Enhancement Module operates through three stages: residual feature processing, efficient attention, and feature enhancement. CFEM progressively processes stereo features f_l and f_r, establishing cross-view dependencies and ultimately generating compact enhanced features f_{c_l} and f_{c_r} for stereo image reconstruction. The residual feature processing stage employs three cascaded residual blocks to transform stereo features through nonlinear operations. As detailed in Figure 2, each residual block implements 3×3 convolutional layers with Leaky ReLU activations, formulated as:

$$\begin{cases} f'_l = RB(f_l) + f_l \\ f'_r = RB(f_r) + f_r \end{cases} \tag{1}$$

where $RB(\cdot)$ denotes stacked convolution-activation operations of residual block. This architecture preserves the fidelity of stereo features through identity skip connections while enabling progressive feature abstraction. To address computational complexity in standard multi-head attention mechanism, we utilize a parameter-efficient multi-head cross-view attention, facilitating bidirectional feature interactions through optimized query-key-value projections. For the left-view feature (with analogous processing for the right-view feature), three parallel 1×1 convolutions project the feature into query Q_l, key K_l and value Q_l tensors. These projections are spatially flattened and divided into H attention heads along the channel axis. Within each head i, spatial and channel-wise softmax normalization is applied to the key and query tensors. The attention context is computed as $C_{l_i} = \tilde{K}_{l_i} V_{l_i}^T$, followed by value aggregation. The outputs from all heads are concatenated, reshaped, and reprojected via 1×1 convolution, preserving spatial structure for subsequent feature enhancement. The multi-head cross-view attention ensures efficient cross-view feature interaction while maintaining computational tractability through dimensional-controlled projections and parallelized multi-head processing. The feature enhancement stage concatenates cross-attentive stereo features $f'_{l->r}$ and $f'_{r->l}$ into one fused feature representation. Subsequently, the fused feature representation is enhanced through three residual blocks, progressively merging complementary information through channel-wise gating mechanisms and finally producing compact yet information-rich stereo features f_{c_l} and f_{c_r} through skip connections. The compact stereo feature generation of feature enhancement is formally expressed as:

$$\begin{cases} f'_{c\_l} = RBs(Concat(f'_{l\to r}, f'_{r\to l})) + f_l \\ f'_{c\_r} = RBs(Concat(f'_{l\to r}, f'_{r\to l})) + f_r \end{cases} \tag{2}$$

where $RBs(\cdot)$ consists of three residual blocks and $Concat(\cdot)$ denotes channel-wise concatenation. As shown in Figure 1, CFEMs are integrated in the joint decoder to facilitate progressive and efficient cross feature enhancement for stereo image compression.

3.3. Reparameterized Swin Block

Reparameterized Swin Block, illustrated in Figure 3, synergistically integrates contextual modeling with convolutional feature extraction to enhance complementary representations for stereo compression. During training, RSB maintains the core Swin Transformer architecture while incorporating parallel convolutional branch. The Transformer branch employs window-based multi-head self-attention (W-MSA) with relative positional encoding

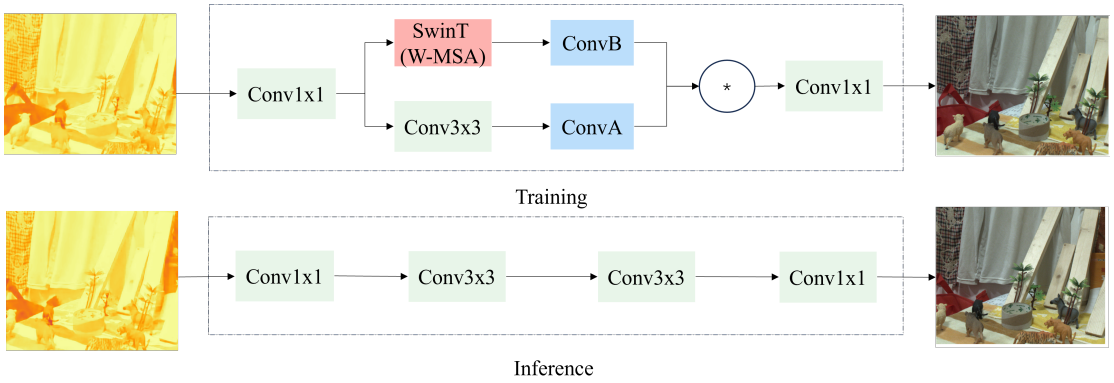

Figure 3: The detailed architecture of Reparameterized Swin Block. The upper branch depicts the training phase architecture, while the lower branch corresponds to the reparameterized inference configuration.

to capture long-range dependencies and hierarchical context, while the complementary convolutional branch extracts local features and texture details. The combined architecture effectively handles both global and local visual cues during training. For inference, structural reparameterization is applied to transform the complex dual-path architecture into a streamlined sequence of convolutional layers, reducing computational complexity and accelerating inference speed while preserving the feature representations learned during training. Specifically, input feature f is initially projected through 1×1 convolutional layer to generate intermediate embeddings, which are subsequently processed via dual-path feature extraction. In the Transformer branch, projected feature is partitioned into non-overlapping $M \times M$ windows, where W-MSA computes feature dependencies. The attention-enhanced features then refined by residual convolutional layer $ConvB$. Simultaneously, the complementary convolutional path processes projected embeddings through 3×3 convolutional and $ConvA$. Both $Conv_A$ and $Conv_B$ comprise multiple residual units and convolutional layers. The outputs from both branches are fused via element-wise multiplication to form an enhanced latent representation. The fused output of RSB during training is computed as follows:

$$f_{out} = Conv_{1\times1}(Conv_A(f_{att}) * \sigma(Conv_B(Conv_{3\times3}(f)))) \tag{3}$$

where $Conv_{k\times k}(\cdot)$ denotes $k \times k$ convolution, $*$ represents element-wise multiplication and $\sigma(\cdot)$ is the sigmoid function. For inference, the dual-path block undergoes structural reparameterization, transformed into an equivalent sequence of convolutional layers formally expressed as:

$$f_{out} = Conv_{1\times1}(Conv_{3\times3}(Conv_{3\times3}(Conv_{1\times1}(f)))) \tag{4}$$

The structural reparameterization converts the training-time dual-branch structure into inference-time convolutional blocks through kernel fusion and branch merging. During inference, we merge the dual paths by absorbing $Conv_A$ and $Conv_B$ into convolutional kernels and eliminating multiplicative interactions. The attention-enhanced features f_{att} are equivalently represented by reparameterized convolution kernel, which is achieved by converting W-MSA's linear projections into 1×1 convolutions, and fusing the kernels of $Conv_A$, $Conv_B$

and 3×3 convolution via kernel summation and zero-padding to align spatial dimensions. The element-wise multiplication is replaced by cascaded convolutions, as the combined kernels implicitly encode the multiplicative dependencies. A detailed mathematical derivation of the structural reparameterization process is provided in the Supplementary Material.

Crucially, the window partitioning induces structured sparsity in attention computation, enabling efficient non-local context extraction without quadratic complexity penalties while convolutional inductive bias stabilizes convergence. During inference, structural reparameterization transforms the dual-path architecture into a streamlined convolutional sequence, reducing computational complexity while preserving learned representations. RSB achieves an optimal balance between Transformer-based context modeling expressiveness and convolutional operational efficiency, effectively capturing long-range dependencies while preserving critical texture details.

4. Experiments and Results

4.1. Experimental Setup

Datasets. To evaluate compression performance of the proposed framework against state-of-the-art methods, we select two public stereo datasets, InStereo2K dataset and Cityscapes dataset. InStereo2K dataset Bao et al. (2020) is a large-scale real-world dataset for stereo matching in indoor scenes, containing 2050 stereo image pairs with high-accuracy disparity maps. Cityscapes dataset Cordts et al. (2016) consists of 5000 stereo image pairs capturing long-range urban scenes, divided into 2975 training pairs, 500 validation pairs, and 1525 testing pairs. These datasets are also chosen due to their distinct characteristics and the challenges they pose for stereo image compression, making them suitable for comprehensive evaluations of the proposed framework and state-of-the-art compression methods.

Metrics. The reconstruction quality is quantitatively assessed using two metrics, Peak Signal-to-Noise Ratio (PSNR) and Multi-Scale Structural Similarity Index Measure (MS-SSIM) Wang et al. (2003). The rate is measured by bit-per-pixel (bpp). PSNR measures pixel-wise intensity differences between the reconstructed and original image. Structure Similarity Index Measure evaluates luminance consistency, contrast preservation, and structural correspondence. MS-SSIM extends this analysis across multiple spatial scales to better align with hierarchical human visual perception.

Benchmarks. To evaluate the compression performance of RSTSIC, we conduct extensive evaluations against conventional codecs BPG, HEVC Sullivan et al. (2012), MV-HEVC Hannuksela et al. (2015), VVC Bross et al. (2021), learning-based stereo compression models DSIC Liu et al. (2019), HESIC Deng et al. (2021), SASIC Wödlinger et al. (2022), ECSIC Wödlinger et al. (2024) and distributed neural codec LDMIC Zhang et al. (2023). BPG processes images independently with full chroma resolution. For HEVC, stereo pairs are encoded as two-frame sequences without chroma subsampling. MV-HEVC operates in default two-view intra mode but remains constrained to 4:2:0 chroma subsampling, adversely affecting high-bitrate performance. VVC is evaluated using its reference software, VVC Test Model (VTM). DSIC and HESIC results are sourced from original publications, while SASIC, LDMIC and ECSIC implementations are reproduced on both datasets.

Implementation Details. The proposed model was optimized using the Adam optimizer, with an initial learning rate of 10^{-4} that was halved every 100 iterations. During

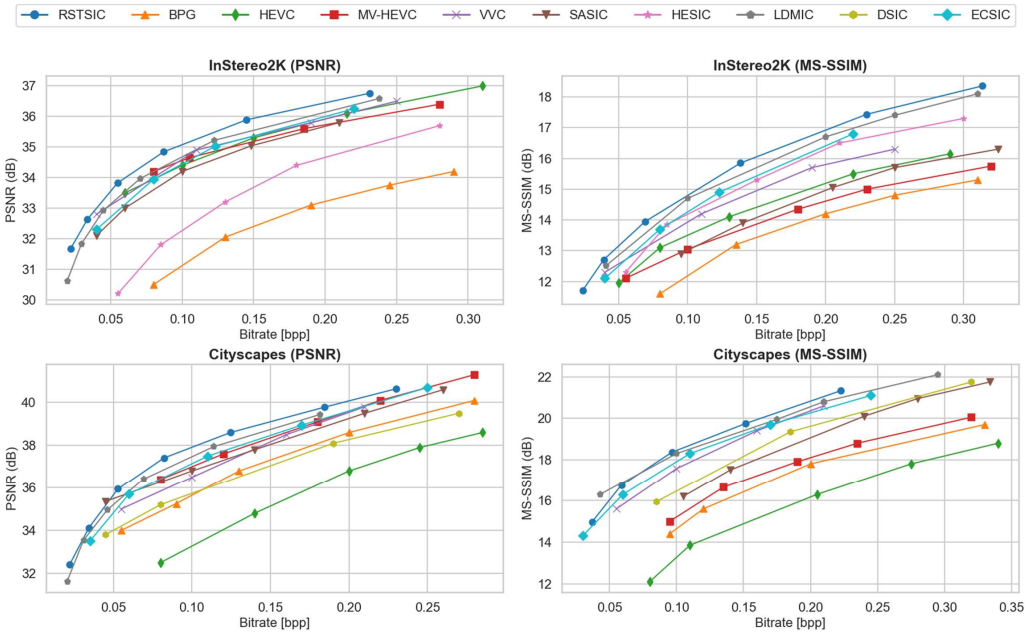

Figure 4: Rate-distortion curves of the proposed framework against state-of-the-art compression methods on InStereo2K and Cityscapes datasets.

training, each image was augmented by random flipping and cropping to a size of 256×256, which helps to diversify training data and improve the generalization ability of the model. The batch size for each training iteration was set to 16. We trained our model with varying λ values, specifically $\lambda = \{16, 32, 64, 128, 256, 512, 1024, 2048\}$. All experiments were conducted using NVIDIA Tesla T4 GPU with Python 3.9 and PyTorch 2.1.

4.2. Experimental Results

Compression Performance. To evaluate the compression performance of the proposed RSTSIC framework against state-of-the-art compression methods, we conduct extensive experiments on InStereo2K and Cityscapes datasets and present rate-distortion curves in Figure 4. On the InStereo2K dataset, RSTSIC achieves superior rate-distortion performance across diverse bitrates, outperforming both traditional codecs and learning-based compression methods in terms of PSNR and MS-SSIM. For the Cityscapes dataset, RSTSIC consistently surpasses these compression methods, demonstrating robust generalization to long-range urban scenes with varying depth distributions. These improvements stem from synergistic integration of cross-view feature enhancement and non-local contextual mining, precisely preserving structural details while effectively exploiting inter-view redundancies.

Qualitative Comparison. To evaluate the reconstruction quality of the proposed framework against state-of-the-art compression methods, we provide a qualitative comparison on InStero2K dataset, as presented in Figure 5. The top edge of the elephant in the original image is well retained in the RSTSIC reconstruction, whereas both LDMIC and SASIC outputs exhibit noticeable blurring artifacts in this region. Regarding shadow re-

construction, RSTSIC and SASIC achieve accurate restoration of the black cat's shadow, while LDMIC erroneously enlarges the shadow area.

(a) Ground Truth

(b) RSTSIC
bpp = 0.043
PSNR = 33.8

(c) LDMIC
bpp = 0.045
PSNR = 33.4

(d) SASIC
bpp = 0.052
PSNR = 32.8

Figure 5: Qualitative comparison of compressed image of the proposed framework against deep compression methods on InStereo2K dataset.

Computational Complexity. To assess the computational efficiency of the proposed RSTSIC framework against state-of-the-art compression methods, we evaluate key metrics including model parameters, FLOPs, encoding and decoding latency. These comparisons are performed with the input resolution fixed at 256 × 256 with three channel dimension. The specific comparison results are presented in Table 1. Moreover, we further provide a component-level complexity breakdown of the proposed framework in the Supplementary Material. The RSTSIC framework achieves significant parameter efficiency with 2.73M model parameters and 40.79G FLOPs, demonstrating at least 58.57% reduction in model parameters and 36.43% FLOPs reduction compared to state-of-the-art compression methods. To ensure fair comparison, we evaluate the average encoding and decoding latency of learning-based compression methods on the InStereo2K dataset. RSTSIC achieves the lowest latency, attributable to its structural reparameterization technique and distributed architecture. The computational efficiency makes RSTSIC practically deployable in hardware-constrained scenarios.

Table 1: Computational efficiency comparison of the proposed framework against state-of-the-art compression methods.

Model	Parameters (M)	FLOPs (G)	Encoding/Decoding Latency (ms)
RSTSIC	**2.73**	**40.79**	**121/96**
DSIC	17.13	64.17	522/537
LDMIC	18.23	80.95	206/869
SASIC	6.59	107.44	213/673
HESIC	34.87	73.66	127/323
ECSIC	23.98	85.70	139/127

4.3. Ablation Study

To rigorously evaluate the impact of individual architectural components in the proposed framework, we perform extensive ablation studies that measure the rate-distortion performance on the Cityscapes and InStereo2K datasets. These experiments employ controlled

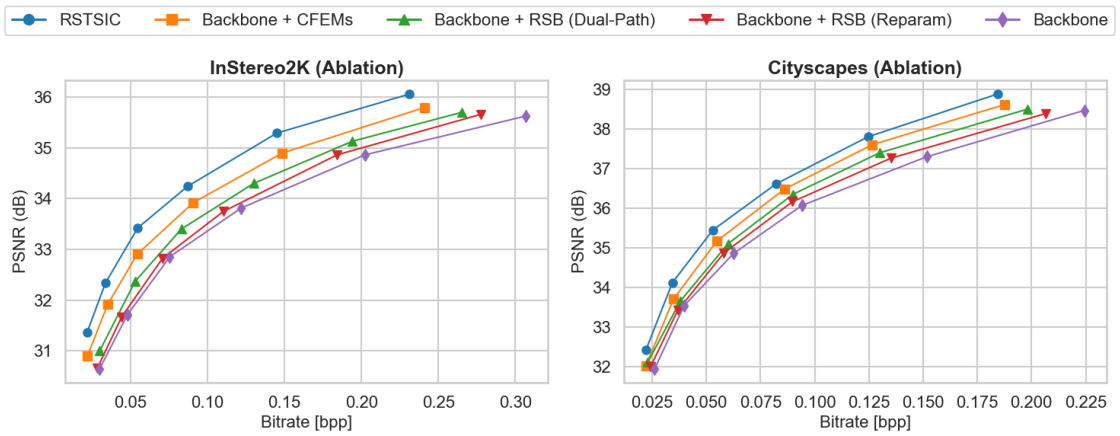

Figure 6: Ablation Study: Rate-distortion curves of the proposed framework with varying structural modifications on InStereo2K and Cityscapes datasets.

architectural variations of the proposed framework, with comparative curves specifically isolating the contributions of each component. The resulting rate-distortion curves are shown in Figure 6. To intuitively reflect the performance of architectural variants, we further provide Bjøntegaard Delta PSNR (BD-PSNR) Bjontegaard (2001) and BD-Rate comparison on both datasets in Table 2. BD-PSNR quantifies the average PSNR gain at equivalent bitrate, while BD-Rate measures the average bitrate reduction required to maintain identical reconstruction quality.

Backbone: To evaluate the individual contribution of key components to stereo compression performance, we derive a convolutional backbone architecture from the RSTSIC framework by eliminating both CFEM and RSB. Channel-wise concatenation is adopted in the backbone for feature fusion, serving as the baseline for ablating CFEM and RSB.

Backbone + CFEMs: For this case, we remove Reparameterized Swin Block from the RSTSIC architecture while retaining Cross Feature Enhancement Modules. As evidenced by the rate-distortion curves in Figure 6, the modified model exhibits degraded performance compared to the full model, suggesting RSB integrated in the joint decoder plays a critical role in achieving bitrate reduction without compromising reconstructed quality. Notably, despite the performance gap, the RSB-ablated variant still demonstrates consistent performance gains over the convolutional backbone architecture across both datasets.

Backbone + RSB (Reparam): In this part of ablation study, we selectively remove CFEMs while maintaining RSB with reparameterization technique. The resulting reparameterized architecture maintains performance superiority over the convolutional backbone across both datasets, yet demonstrates measurable performance degradation compared to the full RSTSIC implementation. This observation highlights the complementary roles of CFEMs and RSB in stereo compression. While the RSB alone ensures baseline performance improvements, the full performance potential is only realized through synergistic integration of both components.

Backbone + RSB (Dual-Path): To evaluate the impact of structural reparameterization and isolate the capability of RSB, we remove the reparameterization technique

from RSB and keep the dual-path training-time structure of RSB. While the modified architecture achieves superior performance compared to the reparameterized RSB structure, it incurs higher computational overhead due to the quadratic computational complexity relative to input resolution from self-attention mechanism.

RSTSIC: The RSTSIC framework, integrating both Reparameterized Swin Block and Cross Feature Enhancement Modules, demonstrates quantifiably superior rate-distortion performance across all evaluated architectural variants. This empirical superiority confirms that the synergistic combination of these components achieves optimal compression performance compared to partial implementations or convolutional backbone.

Table 2: BD-Rate and BD-PSNR comparison of architectural variants on InStereo2K and Cityscapes dataset.

Architectural Variants	InStereo2K		Cityscapes	
	BD-Rate	BD-PSNR	BD-Rate	BD-PSNR
Backbone	0	0	0	0
Backbone + RSB (Reparam)	-7.87	0.17 (+0.42%)	-7.61	0.23 (+0.56%)
Backbone + RSB (Dual-Path)	-15.76	0.35 (+1.04%)	-13.86	0.44 (+1.12%)
Backbone + CFEMs	-29.14	0.68 (+2.21%)	-21.84	0.72 (+1.83%)
RSTSIC	**-43.27**	**1.16 (+3.52%)**	**-29.74**	**1.22 (+2.83%)**

5. Conclusion

In this work, we propose a novel distributed stereo compression framework, Reparameterized Swin Transformer for Stereo Image Compression (RSTSIC), which integrates the global modeling capabilities of vision transformers with efficient inter-view feature correlation learning and structural reparameterization technique. CFEMs in the joint decoder efficiently model stereo feature dependencies through cross-view feature enhancement, coupled with RSB that employs window-based self-attention mechanism to capture long-range dependencies and exploit non-local contextual information. The reparameterization technique further ensures deployment-friendly complexity, reducing computational overhead without dramatically compromising compression performance. With extensive experiments on InStereo2K and Cityscapes datasets, RSTSIC demonstrates superior performance compared to both conventional codecs and deep compression methods. Ablation studies confirm the necessity of both CFEMs and RSB, highlighting their synergistic role in achieving optimal compression performance. Future research directions include extending RSTSIC to multi-view compression framework, integrating dynamic bitrate adaptation mechanisms, and exploring lightweight architectural variants for edge-device deployment.

Acknowledgments

This work was partially supported by the National Key Research and Development Program (2022YFB3104001), the National Natural Science Foundation of China (61672398), the Hubei Province Major Science and Technology Innovation Program (2024BAA011), and the Key Research and Development Program of Hubei Province (2022BAA050).

References

Hassan Abu Alhaija, Siva Karthik Mustikovela, Lars Mescheder, Andreas Geiger, and Carsten Rother. Augmented reality meets computer vision: Efficient data generation for urban driving scenes. *International Journal of Computer Vision*, pages 961–972, 2018.

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

Wei Bao, Wei Wang, Yuhua Xu, Yulan Guo, Siyu Hong, and Xiaohu Zhang. Instereo2k: A large real dataset for stereo matching in indoor scenes. *Science China Information Sciences*, pages 1–11, 2020.

Gisle Bjontegaard. Calculation of average PSNR differences between rd-curves. *ITU SG16 Doc. VCEG-M33*, 2001.

Benjamin Bross, Jianle Chen, Jens-Rainer Ohm, Gary J Sullivan, and Ye-Kui Wang. Developments in international video coding standardization after AVC, with an overview of versatile video coding (VVC). *Proceedings of the IEEE*, pages 1463–1493, 2021.

Jia-Ren Chang and Yong-Sheng Chen. Pyramid stereo matching network. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5410–5418, 2018.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.

Xin Deng, Wenzhe Yang, Ren Yang, Mai Xu, Enpeng Liu, Qianhan Feng, and Radu Timofte. Deep homography for efficient stereo image compression. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1492–1501, 2021.

Xin Deng, Yufan Deng, Ren Yang, Wenzhe Yang, Radu Timofte, and Mai Xu. MASIC: Deep mask stereo image compression. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 6026–6040, 2023.

Rui Fan, Li Wang, Mohammud Junaid Bocus, and Ioannis Pitas. Computer stereo vision for autonomous driving. *arXiv preprint arXiv:2012.03194*, 2020.

Rui Fan, Li Wang, Mohammud Junaid Bocus, and Ioannis Pitas. Computer stereo vision for autonomous driving: Theory and algorithms. *Recent Advances in Computer Vision Applications Using Parallel Processing*, pages 41–70, 2023.

Miska M Hannuksela, Ye Yan, Xuehui Huang, and Houqiang Li. Overview of the multiview high efficiency video coding (MV-HEVC) standard. *IEEE International Conference on Image Processing*, pages 2154–2158, 2015.

Qiqi Hou, Farzad Farhadzadeh, Amir Said, Guillaume Sautiere, and Hoang Le. Low-latency neural stereo streaming. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7974–7984, 2024.

Chen-Hsiu Huang and Ja-Ling Wu. Unveiling the future of human and machine coding: A survey of end-to-end learned image compression. *Entropy*, page 357, 2024.

Seunghwa Jeong, Bumki Kim, Seunghoon Cha, Kwanggyoon Seo, Hayoung Chang, Jungjin Lee, Younghui Kim, and Junyong Noh. Real-time cnn training and compression for neural-enhanced adaptive live streaming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. End-to-end learning of geometry and context for deep stereo regression. *Proceedings of the IEEE International Conference on Computer Vision*, pages 66–75, 2017.

Yong-Deok Kim, Eunhyeok Park, Sungjoo Yoo, Taelim Choi, Lu Yang, and Dongjun Shin. Compression of deep convolutional neural networks for fast and low power mobile applications. *arXiv preprint arXiv:1511.06530*, 2015.

Jianjun Lei, Xiangrui Liu, Bo Peng, Dengchao Jin, Wanqing Li, and Jingxiao Gu. Deep stereo image compression via bi-directional coding. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19669–19678, 2022.

Feng Liu, Miguel Hernandez-Cabronero, Victor Sanchez, Michael W Marcellin, and Ali Bilgin. The current role of image compression standards in medical imaging. *Information*, page 131, 2017.

Jerry Liu, Shenlong Wang, and Raquel Urtasun. DSIC: Deep stereo image compression. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3136–3145, 2019.

Jinming Liu, Heming Sun, and Jiro Katto. Learned image compression with mixed transformer-cnn architectures. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14388–14397, 2023.

Zhening Liu, Xinjie Zhang, Jiawei Shao, Zehong Lin, and Jun Zhang. Bidirectional stereo image compression with cross-dimensional entropy model. *European Conference on Computer Vision*, pages 480–496, 2024.

Wenjie Luo, Alexander G Schwing, and Raquel Urtasun. Efficient deep learning for stereo matching. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5695–5703, 2016.

Nitish Mital, Ezgi Özyılkan, Ali Garjani, and Deniz Gündüz. Neural distributed image compression using common information. *2022 Data Compression Conference*, pages 182–191, 2022.

Christopher Schuster, Bipeng Zhang, Rajan Vaish, Paulo Gomes, Jacob Thomas, and James Davis. Rti compression for mobile devices. *Proceedings of the 6th International Conference on Information Technology and Multimedia*, pages 368–373, 2014.

Athanassios Skodras, Charilaos Christopoulos, and Touradj Ebrahimi. The JPEG 2000 still image compression standard. *IEEE Signal Processing Magazine*, pages 36–58, 2001.

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*, pages 1649–1668, 2012.

Danhang Tang, Mingsong Dou, Peter Lincoln, Philip Davidson, Kaiwen Guo, Jonathan Taylor, Sean Fanello, Cem Keskin, Adarsh Kowdle, Sofien Bouaziz, et al. Real-time compression and streaming of 4d performances. *ACM Transactions on Graphics*, pages 1–11, 2018.

Ian Trow. AV1: Implementation, performance, and application. *SMPTE Motion Imaging Journal*, pages 51–56, 2020.

Vlad-Ilie Ungureanu, Paul Negirla, and Adrian Korodi. Image-compression techniques: Classical and "region-of-interest-based" approaches presented in recent papers. *Sensors*, page 791, 2024.

David Varodayan, Aditya Mavlankar, Markus Flierl, and Bernd Girod. Distributed grayscale stereo image coding with unsupervised learning of disparity. *2007 Data Compression Conference*, pages 143–152, 2007.

Gregory K Wallace. The JPEG still picture compression standard. *Communications of the ACM*, pages 30–44, 1991.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. *The Thrity-Seventh Asilomar Conference on Signals, Systems and Computers, 2003*, pages 1398–1402, 2003.

Matthias Wödlinger, Jan Kotera, Jan Xu, and Robert Sablatnig. SASIC: Stereo image compression with latent shifts and stereo attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 661–670, 2022.

Matthias Wödlinger, Jan Kotera, Manuel Keglevic, Jan Xu, and Robert Sablatnig. ECSIC: Epipolar cross attention for stereo image compression. *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3424–3433, 2024.

Jure Žbontar and Yann LeCun. Stereo matching by training a convolutional neural network to compare image patches. *Journal of Machine Learning Research*, pages 1–32, 2016.

Xinjie Zhang, Jiawei Shao, and Jun Zhang. LDMIC: Learning-based distributed multi-view image coding. *International Conference on Learning Representations*, 2023.

Yiran Zhong, Yuchao Dai, and Hongdong Li. Self-supervised learning for stereo matching with self-improving ability. *arXiv preprint arXiv:1709.00930*, 2017.