Learned Focused Plenoptic Image Compression with Local-Global Correlation Learning

Gaosheng Liu, Huanjing Yue, *Member, IEEE,* Bihan Wen, *Senior Member, IEEE,* and Jingyu Yang, *Senior Member, IEEE*

***Abstract*—The dense light field sampling of focused plenoptic images (FPIs) yields substantial amounts of redundant data, necessitating efficient compression in practical applications. However, the presence of discontinuous structures and long-distance properties in FPIs poses a challenge. In this paper, we propose a novel end-to-end approach for learned focused plenoptic image compression (LFPIC). Specifically, we introduce a local-global correlation learning strategy to build the nonlinear transforms. This strategy can effectively handle the discontinuous structures and leverage long-distance correlations in FPI for high compression efficiency. Additionally, we propose a spatial-wise context model tailored for LFPIC to help emphasize the most related symbols during coding and further enhance the rate-distortion performance. Experimental results demonstrate the effectiveness of our proposed method, achieving a 22.16% BD-rate reduction (measured in PSNR) on the public dataset compared to the recent state-of-the-art LFPIC method. This improvement holds significant promise for benefiting the applications of focused plenoptic cameras. *The code and pre-trained models are available at https://github.com/GaoshengLiu/LFPIC*.**

***Index Terms*—Focused plenoptic image compression, learned image compression, local-global correlation learning.**

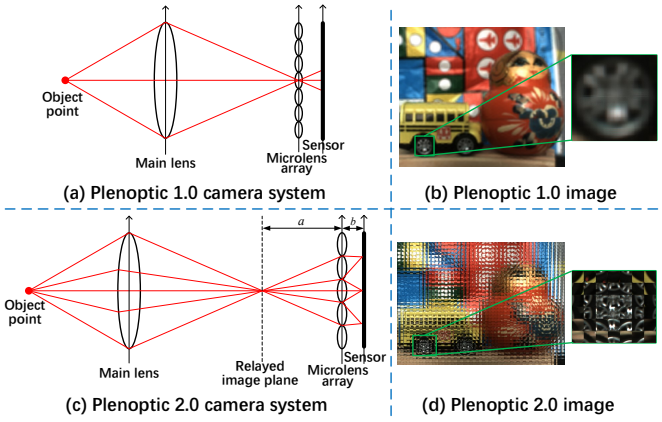


Fig. 1. An example of plenoptic 1.0 and 2.0 camera prototypes. (a) shows the imaging system of plenoptic 1.0 camera. (b) gives a recorded image by (a). (c) shows the imaging system of focused plenoptic (plenoptic 2.0) camera, where the microlens array is placed at a distance a of the image plane. The distance a, b, and the focal length f of microlens satisfy the lens equation, *i.e.*, $\frac{1}{a}+\frac{1}{b}=\frac{1}{f}$. (d) shows the same scene recorded by (c) and pre-processed by [8].

I. Introduction

In recent years, ***focused*** plenoptic cameras, also known as plenoptic 2.0 cameras, have emerged as promising devices due to their improved trade-off between spatial and angular sampling rates compared with ***unfocused*** ones, *i.e.* plenoptic 1.0 cameras. This capability opens up avenues for diverse research areas, including depth estimation [1], [2], view rendering [3], and 3D measurement [4], [5]. Fig. 1 illustrates the imaging pipeline of plenoptic 1.0 and 2.0 cameras with their recorded images. In plenoptic 1.0 cameras, *e.g.*, Lytro Illum camera, the microlens array is placed at the image plane of the main lens. Correspondingly, the microlens image (MLI) recorded by each microlens can be considered as a set of pixels with the same point but captured from different directions. While in focused plenoptic cameras, *e.g.*, Raytrix and Tsinghua Single-focused Plenoptic Camera (TSPC), the microlens array is placed at a distance of a behind the image plane of the main lens. As a result, each microlens re-images the scene at slightly different angles, and each microlens image in focused plenoptic image (FMLI) is a portion of the main-lens image from a different perspective. Despite the abundant geometric information provided by these plenoptic images, challenges arise due to the limited bandwidth and storage space. Consequently, efficiently compressing the plenoptic images becomes a crucial task. The JPEG Pleno [6], [7] has also initiated studies for the standardization of plenoptic image compression.

Image compression is a long-standing problem in engineering [9]–[15]. With the rapid development of learned image compression (LIC), the field of learned plenoptic 1.0 image compression has also witnessed significant progress [16]–[20]. Unfortunately, the application of these techniques to focused plenoptic images (FPIs) encounters obstacles due to the disparity in image formats. Notably, the FMLIs have larger size, *e.g.*, 69×69 *vs.* 14×14, and more complex textures compared with MLIs, as shown in Fig. 1 (b) and (d). Meanwhile, designing effective learning-based nonlinear transform architectures for learned focused plenoptic image compression (LFPIC) also faces challenges. Firstly, the boundaries between every two neighboring FMLIs present textural discontinuities. Secondly, since the spatial size of FMLI is large, the FPI exhibits long-

This work was supported in part by the National Natural Science Foundation of China under Grant 62231018 and Grant 62472308.

Gaosheng Liu, Huanjing Yue, and Jingyu Yang (corresponding author) are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mails: gaoshengliu@tju.edu.cn; huanjing.yue@tju.edu.cn; yjy@tju.edu.cn).

Bihan Wen is with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: bihan.wen@ntu.edu.sg).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the authors. The material includes more network analysis and details. This material is 340 KB in size.

distance spatial correlations among FMLIs. However, previous LFPIC method, GACN [8] applies sliding $K \times K$ convolutional kernels on the entire spatial domain of focused plenoptic images/features, which cannot effectively handle the textural discontinuities and model the correlations among FMLIs.

Our goal is to develop an effective nonlinear transform architecture for LFPIC, with a focus on reducing redundancies, compacting energy, and improving the rate-distortion (RD) performance. Toward this goal, we introduce a local-global correlation learning strategy to build nonlinear transforms. Specifically, we propose treating each FMLI as an independent unit for convolutional operations to model the ***local*** correlations. Additionally, we incorporate an efficient self-attention mechanism across the entire focused plenoptic feature to capture the ***global***, *i.e.*, long-distance correlations. By alternately stacking the local and global correlations learning in the analysis and synthesis transforms, we establish a robust framework for compressing FPIs.

In addition to the elegant design of nonlinear transforms, employing a context model, where previously decoded symbols are utilized for the entropy estimation of unknown symbols, has also been a typical strategy to improve the RD performance [21]–[23]. Drawing inspiration from the advantages of channel-wise autoregressive decoding [22] and spatial-wise checkerboard context modeling [23], we introduce channel-spatial context modeling in the proposed method. Given the discontinuities in latent focused plenoptic features, the original checkerboard context modeling, in which half (anchor) symbols are utilized to predict the distribution of ***adjacent*** other half (non-anchor) ones, might struggle to model spatial-wise context effectively. To this end, we further propose an FMLI-aware checkerboard attention (FCA) for spatial-wise context modeling. FCA can help our context model emphasize the most related symbol coding within each FMLI structure. The contributions of this work are summarized below:

- We propose a local-global correlation learning strategy to build nonlinear transform structures for LFPIC, which can efficiently leverage the discontinuous structures and long-distance correlations of FPIs.
- We introduce a spatial-wise context model tailored for LFPIC with an FMLI-aware checkerboard attention (FCA) to further enhance the RD performance.
- Experimental results demonstrate that our approach achieves state-of-the-art performance. Specifically, our method achieves an average BD-rate reduction of 22.16% compared with the state-of-the-art method, GACN [8].

The rest of this article is organized as follows. In Section II, we review the most related works on this topic. In Section III, we give the technical details of our proposed method. In Section IV, we present the experimental settings, results, comparisons, and ablations. Finally, we conclude the article in Section V.

II. Related Works

A. Plenoptic 1.0 Image Compression

The plenoptic 1.0 image can be reversibly re-organized from microlens pattern to multi-view pattern (sub-aperture images) by replacing pixels with the same angular coordinates. According to the input patterns, previous methods can be categorized into two groups. The first group tends to transform the input multiple views to a pseudo-video sequence by stacking views in different scan orders [24]–[26]. Subsequently, the inter-coding tools are applied to leverage correlations between views for bits saving. To further reduce the bitrate, several approaches were proposed to encode sparse key views and synthesize non-key views at the decoder side [16], [17], [27], [28]. For example, Hou et al. [27] proposed to encode selected key views and the residual between super-resolved non-key views and ground-truths. The second group treats the entire microlens-pattern plenoptic image as a single input, aiming at reducing the spatial redundancies among adjacent MLIs. To achieve this, previous methods have introduced various schemes, such as block matching [29], [30], self-similarity bi-directional prediction [31], disparity-assisted prediction [32], and transform coding [33], [34]. To improve the ability to handle non-homogeneous texture regions, Liu et al. [35] proposed combining a classification scheme to segment challenging areas. Huang et al. [36] introduced a disparity rectification model to refine the estimated disparity map for geometry-based compression. In addition to these conventional codecs, several end-to-end learned compression methods [18]–[20], [37], [38] have been introduced with promising compression efficiency. For example, Tong et al. [19] decoupled the spatial-angular information of plenoptic 1.0 images in the nonlinear transforms to improve the compression efficiency. Shi et al. [38] incorporated implicit neural representation for compression and introduced *descriptors* & *modulators* to control the rendering of different views.

B. Focused Plenoptic Image Compression

The FPI can also be rendered to sub-aperture images (SAIs) using rendering tools. However, as rendering is an irreversible process with inevitable information loss and rendering errors, performing compression on rendered SAIs is a sub-optimal choice. In the literature, conventional methods based on neighboring intra prediction [39], block matching [40], imaging-principle guided prediction [41], and FMLI-based two-step search [42] have been proposed to compress the focused plenoptic images/videos. Very recently, Liu et al. [43] introduced 5D Epanechnikov Mixture-of-Experts based on Epanechnikov Kernel [44] to model FMLI for compression. For LFPIC, Tong et al. [8] proposed GACN and a benchmark dataset captured using a TSPC. However, as mentioned in Section I, their method does not consider textural discontinuities in FPIs.

C. Learned Image Compression

Our work is also closely related to the LIC methods. In the area of LIC, the variational auto-encoder (VAE)-based approaches investigated in [12], [45], [46] have been influential in driving recent advancements. In [12], a hyperprior auto-encoder, serving as an entropy model, is introduced to predict the distribution of latent representations. Subsequently, various methods are developed with more powerful nonlinear

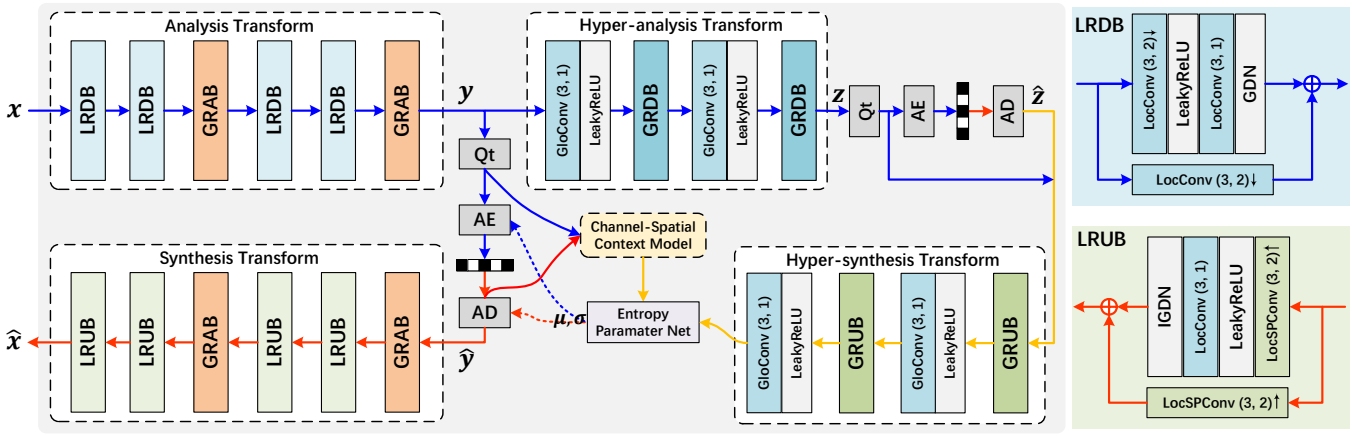


Fig. 2. Framework of our proposed method. Qt, AE, and AD denote the quantization, arithmetic encoder and decoder, respectively. In LRDB/LRUB, the LocConv (K_1, S_1) denotes local convolution with kernel size $K_1 \times K_1$ and stride S_1. The LocSPConv (K_2, S_2) denotes $K_2 \times K_2$ LocConv following a *PixelShuffle* layer with S_2-fold depth-to-space up-sampling. Similarly, the GloConv (K_3, S_3) denotes global convolution with $K_3 \times K_3$ kernel and stride S_3. In LRDB and LRUB, $\downarrow$ and $\uparrow$ indicate down-sampling and up-sampling operations, respectively. The lines colored in blue and red denote encoding and decoding information flow, respectively. The lines colored in yellow denote flows shared by both encoding and decoding.

transform architectures [47], [48] and entropy model [13], [21]–[23], [49]–[51]. It is worth noting that, the recent LIC methods have outperformed conventional codecs. However, these methods face significant challenges when handling FPIs due to the complex structures, as shown in Fig. 1.

III. Method

A. Problem Formulation and Overview

This work aims at compressing an FPI, $\boldsymbol{x} \in \mathbb{R}^{3\times\alpha H\times\beta W}$, to low bitrate symbols, which can also be decoded to a low-distortion counterpart, $\hat{\boldsymbol{x}} \in \mathbb{R}^{3\times\alpha H\times\beta W}$, where H, W are the spatial resolution of each FMLI, and α, β are the size of the FMLI array. As illustrated in Fig. 2, our method is designed based on the basic framework of hyperprior architecture [12]. Specifically, the main encoder, denoted as analysis transform g_a, maps an input FPI $\boldsymbol{x}$ to a latent representation $\boldsymbol{y}$. Then the quantization operation Qt is applied to get discrete representation:

$$\begin{aligned} \boldsymbol{y} &= g_a(\boldsymbol{x};\boldsymbol{\theta}), \\ \hat{\boldsymbol{y}} &= \mathrm{Qt}(\boldsymbol{y}-\boldsymbol{\mu})+\boldsymbol{\mu}, \end{aligned} \tag{1}$$

where $\boldsymbol{\theta}$ represents the trainable parameters of g_a. Following the VAE-based methods [12], [21], we model each element $\boldsymbol{y}(j)$ as a Gaussian distribution with standard deviation $\boldsymbol{\sigma}(j)$ and mean $\boldsymbol{\mu}(j)$. We adopt the approach from previous works [13], [21] to encode the rounded $\lceil \boldsymbol{y}-\boldsymbol{\mu} \rfloor$ using STE [52] to bit stream and restore $\hat{\boldsymbol{y}}$ as $\lceil \boldsymbol{y}-\boldsymbol{\mu} \rfloor + \boldsymbol{\mu}$ to facilitate entropy parameter estimation.

In g_a, we introduce local residual down-sampling blocks (LRDBs) and global residual attention blocks (GRABs) to capture the local and global correlations in FPIs. Conversely, the main decoder, *i.e.*, denoted as synthesis transform g_s, is composed of local residual up-sampling blocks (LRUBs) and GRABs. g_s maps the decoded $\hat{\boldsymbol{y}}$ from the lossless arithmetic encoder and decoder (AE, AD) to $\hat{\boldsymbol{x}}$:

$$\hat{\boldsymbol{x}} = g_s(\hat{\boldsymbol{y}};\boldsymbol{\phi}), \tag{2}$$

where $\boldsymbol{\phi}$ denotes the trainable parameters of g_s. The standard deviation $\boldsymbol{\sigma}$ and mean $\boldsymbol{\mu}$ are calculated from an entropy model, consisting of a hyperprior model to estimate the side information and a channel-spatial context model for contextual encoding/decoding.

The hyperprior model incorporates hyper-analysis transform h_a, to generate an additional hyper latent $\hat{\boldsymbol{z}}$, which is encoded to bit stream and transmitted with the bit stream of the $\hat{\boldsymbol{y}}$. The hyper-synthesis transform h_s reconstructs a hyperprior feature, which is involved in the entropy parameter estimation. In h_a and h_s, we introduce global residual down-sampling and up-sampling blocks (GRDB, GRUB) for nonlinear transforms. We adopt the non-parametric and fully factorized density model [12] to predict the distribution of the hyper latent $\hat{\boldsymbol{z}}$.

In our channel-spatial context model, we evenly divide $\hat{\boldsymbol{y}}$ into N segments, $\{\hat{\boldsymbol{y}}^1, \hat{\boldsymbol{y}}^2, \cdots, \hat{\boldsymbol{y}}^N\}$, along the channel dimension. For the coding of i-th segment $\hat{\boldsymbol{y}}^i$, we use a channel context net (CCN) to capture the cross-channel correlations from segments $\hat{\boldsymbol{y}}^{<i}$. Furthermore, we adopt the checkerboard pattern [23], where the latent tensor $\hat{\boldsymbol{y}}$ is divided into anchor component $\hat{\boldsymbol{y}}_{\text{anch}}$ and non-anchor component $\hat{\boldsymbol{y}}_{\text{nonan}}$. The spatial-wise correlations among anchor symbols are captured by a spatial context net (SCN) to help predict non-anchor symbols. In our context model, $\hat{\boldsymbol{y}}^i_{\text{anch}}$ is conditional on $\hat{\boldsymbol{y}}^{<i}$ and $\hat{\boldsymbol{z}}$, and $\hat{\boldsymbol{y}}^i_{\text{nonan}}$ is conditioned on $\hat{\boldsymbol{y}}^i_{\text{anch}}$, $\hat{\boldsymbol{y}}^{<i}$, and $\hat{\boldsymbol{z}}$. Therefore, the total bits $\mathcal{R}_{\hat{\boldsymbol{y}}}$ of the latent $\hat{\boldsymbol{y}}$ can be formulated as:

$$\mathcal{R}_{\hat{\boldsymbol{y}}} = \sum_{i=1}^{N} (\mathcal{R}_{\hat{\boldsymbol{y}}^i_{\text{anch}}} + \mathcal{R}_{\hat{\boldsymbol{y}}^i_{\text{nonan}}}). \tag{3}$$

The bits of anchor and non-anchor latent $\mathcal{R}_{\hat{\boldsymbol{y}}^i_{\text{anch}}}$ and $\mathcal{R}_{\hat{\boldsymbol{y}}^i_{\text{nonan}}}$ are given by:

$$\begin{aligned} \mathcal{R}_{\hat{\boldsymbol{y}}^i_{\text{anch}}} &= \mathbb{E}[-\log_2 p_{\hat{\boldsymbol{y}}^i_{\text{anch}}|\hat{\boldsymbol{z}},\hat{\boldsymbol{y}}^{<i}}(\hat{\boldsymbol{y}}^i_{\text{anch}}|\hat{\boldsymbol{z}},\hat{\boldsymbol{y}}^{<i})], \\ \mathcal{R}_{\hat{\boldsymbol{y}}^i_{\text{nonan}}} &= \mathbb{E}[-\log_2 p_{\hat{\boldsymbol{y}}^i_{\text{nonan}}|\hat{\boldsymbol{z}},\hat{\boldsymbol{y}}^i_{\text{anch}},\hat{\boldsymbol{y}}^{<i}}(\hat{\boldsymbol{y}}^i_{\text{nonan}}|\hat{\boldsymbol{z}},\hat{\boldsymbol{y}}^i_{\text{anch}},\hat{\boldsymbol{y}}^{<i})]. \end{aligned} \tag{4}$$

According to the RD optimization theory, a larger bitrate consumption leads to a lower distortion. Therefore, the loss function of our method is formulated as:

$$\ell = \mathcal{R}_{\hat{\boldsymbol{y}}} + \mathcal{R}_{\hat{\boldsymbol{z}}} + \lambda \mathcal{D}(\hat{\boldsymbol{x}}, \boldsymbol{x}), \tag{5}$$

where λ is a Lagrange multiplier for the trade-off of bitrate and distortion, $\mathcal{D}$. In the following subsections, we provide technical details of the key components in the main/hyper nonlinear transforms and the channel-spatial context model.

B. Main Analysis and Synthesis Transforms

As illustrated in Section I, the FMLIs in FPIs exhibit local and long-distance correlations. To leverage these correlations for improving RD performance, we introduce local-global correlation learning in the nonlinear transforms. More concretely, we propose LRDB, LRUB, and GRAB to construct the main analysis and synthesis transforms.

Local Residual Down- and Up-sampling Blocks. The structures of LRDB and LRUB are depicted on the right of Fig. 2. To capture the local correlations, we introduce local convolution (LocConv), whose kernel slides along the height and width dimensions within each FMLI. Taking the first LRDB in g_a as an example, we first unfold the input $\boldsymbol{x}$ to FMLI representations, $\boldsymbol{x}^k_{\text{FMLI}} \in \mathbb{R}^{3\times H\times W}, k \in \{1, \cdots \alpha \cdot \beta\}$. Then each FMLI is processed by a shared-weight LocConv with stride two for spatial-wise down-sampling and channel expanding. After a Leaky ReLU activation, another LocConv with stride one is deployed. Subsequently, generalized divisive normalization (GDN) [46] is used to remove statistical dependencies. Finally, a residual connection along with a LocConv (stride 2) is applied.

The structure of LRUB is symmetrical to LRDB. In LRUB, the up-sampling operation is realized by a local subpixel convolution (LocSPConv), which consists of a LocConv and a *PixelShuffle* layer. We adopt the inverse version of GDN, *i.e.*, IGDN [46] in LRUB.

Global Residual Attention Block. In the literature, leveraging self-attention for capturing long-distance correlations has been a common strategy [53], [54]. However, due to the substantial size of FPIs, directly applying vanilla self-attention to the entire focused plenoptic feature incurs high computational complexity, compromising coding efficiency. In this work, we adopt an efficient variant [55] of self-attention to capture global correlations. Specifically, we propose a global residual attention block (GRAB), as depicted in Fig. 3. Within each GRAB, we utilize three 1×1 convolutions followed by reshaping operations to map the input feature $\mathcal{F} \in \mathbb{R}^{C\times \alpha H_1 \times \beta W_1}$ to $\boldsymbol{Q} \in \mathbb{R}^{d_1 \times C}, \boldsymbol{K} \in \mathbb{R}^{d_1 \times C}$, and $\boldsymbol{V} \in \mathbb{R}^{d_1 \times C}$, where H_1 and W_1 are the spatial resolution of current FMLI structure and $d_1 = \alpha H_1 \times \beta W_1$. Following [55], the *Softmax* function is applied on the d_1 dimension of $\boldsymbol{K}$ and on the C dimension of $\boldsymbol{Q}$. Then the global attention (GA) can be formulated as:

$$\mathcal{F}_{\text{GA}} = \text{Softmax}(\boldsymbol{Q}) \otimes (\text{Softmax}(\boldsymbol{K})^\top \otimes \boldsymbol{V}), \tag{6}$$

where $\otimes$ denotes matrix multiplication and $\mathcal{F}_{\text{GA}}$ represents the output. By separately performing *Softmax* function on $\boldsymbol{Q}$ and $\boldsymbol{K}$ to approximate the *Softmax* on $\boldsymbol{Q} \otimes \boldsymbol{K}^T$ in vanilla self attention [53], the matrix multiplication operation between $\text{Softmax}(\boldsymbol{K})^T$ and $\boldsymbol{V}$ can be calculated first. Therefore, the complexity of GA is linear with the resolution, *i.e.*, $O(C^2 d_1)$. The $\mathcal{F}_{\text{GA}}$ is reshaped back to $C \times \alpha H_1 \times \beta W_1$ and processed by a 1×1 convolution and feed forward net (FFN) [53]. After a local residual connection, a final 1×1 convolution is deployed. Additionally, we incorporate global residual learning in GRAB.

C. Entropy Model

The entropy model aims at predicting the distributions of latent representations. It consists of a hyper auto-encoder to generate a hyperprior feature and a channel-spatial context model to incorporate channel- and spatial-wise correlations. An illustration of our entropy parameter prediction is shown in Fig. 5. Specifically, $\hat{\boldsymbol{y}}^i_{\text{anch}}$ is conditional on $\hat{\boldsymbol{y}}^{<i}$ and $\hat{\boldsymbol{z}}$, and $\hat{\boldsymbol{y}}^i_{\text{nonan}}$ is conditioned on $\hat{\boldsymbol{y}}^i_{\text{anch}}$, $\hat{\boldsymbol{y}}^{<i}$, and $\hat{\boldsymbol{z}}$.

1) Hyper-Analysis and Synthesis Transforms: The h_a maps $\boldsymbol{y}$ to hyper latent $\boldsymbol{z}$, which is then quantized (to $\hat{\boldsymbol{z}}$), compressed, and transmitted as side information. The h_s reconstructs a hyperprior feature from decoded $\hat{\boldsymbol{z}}$. Considering the spatial size of each FMLI structure in $\boldsymbol{y}$ is very small ($16\times$ down-sampled by g_a) and cannot be down-sampled more than two times, we treat the intermediate features passing through h_a and h_s as a single-image feature. Concretely, we introduce global residual down-sampling and up-sampling blocks (GRDBs, GRUBs) to build h_a and h_s. The GRDB/GRUB has the same structure as LGRB/LGUB except that LocConv is replaced with global convolution (GloConv), whose kernel slides along the entire spatial dimensions of intermediate features.

2) Channel-Spatial Context Model: In our approach, we apply channel- and spatial-wise context modeling to capture the dependencies among quantized elements. Specifically, each channel-wise segment $\hat{\boldsymbol{y}}^i$ is further divided into $\hat{\boldsymbol{y}}^i_{\text{anch}}$ and $\hat{\boldsymbol{y}}^i_{\text{nonan}}$.

The hyper decoder h_s reconstructs a hyperprior feature $\boldsymbol{\Phi}_{\text{hy}}$ from $\hat{\boldsymbol{z}}$. The CCN (network structure is provided in the ***supplementary material***), denoted as g^i_{ch}, generates a cross-channel context feature $\boldsymbol{\Phi}^i_{\text{ch}}$ from channel-wise concatenated $[\hat{\boldsymbol{y}}^1, \cdots, \hat{\boldsymbol{y}}^{i-1}]$. For coding $\hat{\boldsymbol{y}}^i_{\text{anch}}$, we predict its distribution $\tilde{\boldsymbol{\mu}}^i, \tilde{\boldsymbol{\sigma}}^i$ by feeding the concatenated $[\boldsymbol{\Phi}_{\text{hy}}, \boldsymbol{\Phi}^i_{\text{ch}}]$ to an entropy parameter net, denoted as g^i_{ep-a}. The whole process can be formulated as:

$$\begin{aligned} \boldsymbol{\Phi}_{\text{hy}} &= h_s(\hat{\boldsymbol{z}}), \\ \boldsymbol{\Phi}^i_{\text{ch}} &= g^i_{ch}([\hat{\boldsymbol{y}}^1 \cdots \hat{\boldsymbol{y}}^{i-1}]), \\ (\tilde{\boldsymbol{\mu}}^i, \tilde{\boldsymbol{\sigma}}^i) &= g^i_{ep-a}([\boldsymbol{\Phi}_{\text{hy}}, \boldsymbol{\Phi}^i_{\text{ch}}]). \end{aligned} \tag{7}$$

$\tilde{\boldsymbol{\mu}}^i, \tilde{\boldsymbol{\sigma}}^i$ are incorporated in the AE and AD for the coding of $\hat{\boldsymbol{y}}^i_{\text{anch}}$. For the coding of non-anchor symbols, we further adopt a spatial-wise context model. Specifically, the SCN, denoted as g^i_{sp}, is applied on $\hat{\boldsymbol{y}}^i_{\text{anch}}$ to generate a spatial context feature $\boldsymbol{\Phi}^i_{\text{sp}}$. Therefore, the estimation for the distribution of $\hat{\boldsymbol{y}}^i_{\text{nonan}}$ can be represented as:

$$\begin{aligned} \boldsymbol{\Phi}^i_{\text{sp}} &= g^i_{sp}(\hat{\boldsymbol{y}}^i_{\text{anch}}), \\ (\breve{\boldsymbol{\mu}}^i, \breve{\boldsymbol{\sigma}}^i) &= g^i_{ep-na}([\boldsymbol{\Phi}_{\text{hy}}, \boldsymbol{\Phi}^i_{\text{ch}}, \boldsymbol{\Phi}^i_{\text{sp}}]), \end{aligned} \tag{8}$$

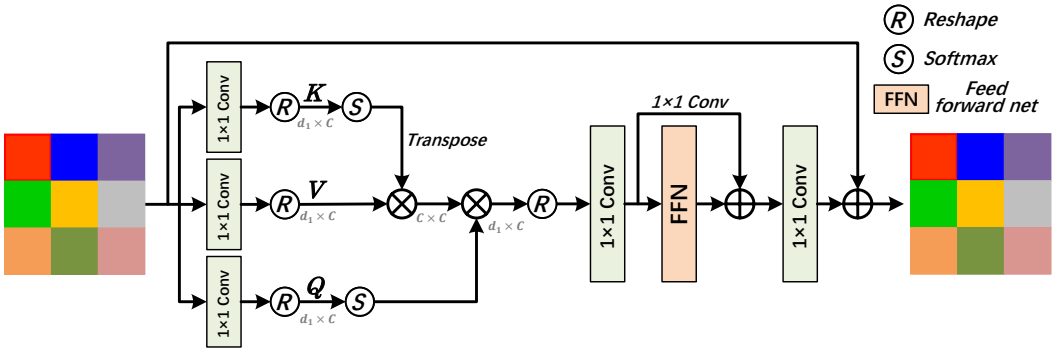


Fig. 3. Structure of global residual attention block (GRAB). In the input/output feature, patches with different colors indicate different FMLI structures. The channel dimension of feature maps is omitted for simplicity.

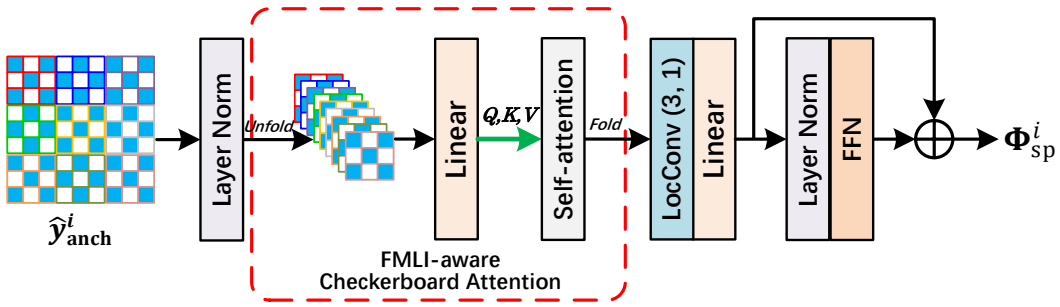


Fig. 4. Structure of spatial context net (SCN). In $\hat{\boldsymbol{y}}^i_{\text{anch}}$, the squares colored using light blue denote decoded anchor positions. The frames of each square in the same color indicate they belong to one FMLI structure. The channel dimension of $\hat{\boldsymbol{y}}^i_{\text{anch}}$ and feature reshaping operations are omitted for simplicity.

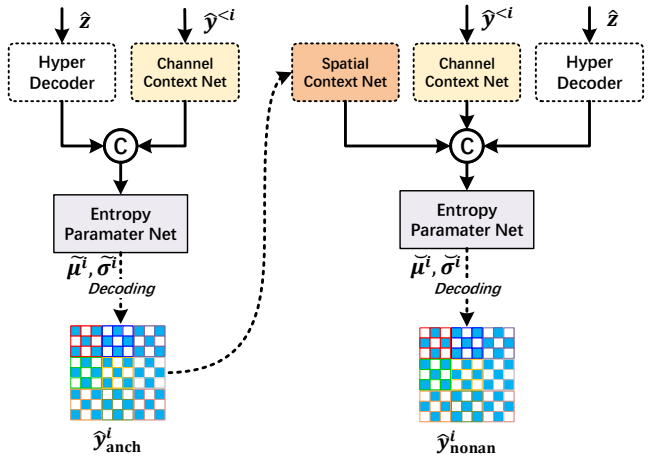


Fig. 5. Illustration of our entropy parameter prediction. The squares colored using light blue denote decoded anchor (non-anchor) positions in $\hat{\boldsymbol{y}}^i_{\text{anch}}$ ($\hat{\boldsymbol{y}}^i_{\text{nonan}}$). The frames of each square in the same color indicate that they belong to one FMLI structure.

where $(\breve{\boldsymbol{\mu}}^i, \breve{\boldsymbol{\sigma}}^i)$ are the estimated mean and standard deviation of $\hat{\boldsymbol{y}}^i_{\text{nonan}}$, and g^i_{ep-na} is the entropy parameter net for the distribution prediction.

Spatial Context Net. In our spatial-wise context model, we employ the checkerboard context strategy [23] that utilizes half symbols to predict the distribution of other adjacent symbols. However, considering the discontinuities between FMLI structures, it becomes crucial to emphasize the most relevant symbol during coding. Given the content-meaningful structure within each FMLI structure and the effective content-adaptive learning capability of self-attention [56], we introduce an FMLI-aware checkerboard attention (FCA) based on self-attention in SCN, as depicted in Fig. 4. In SCN, we unfold decoded anchor tensor of i-th segment, $\hat{\boldsymbol{y}}^i_{\text{anch}}$, after layer normalization into FMLI representations, $\hat{\boldsymbol{y}}^{i,k}_{\text{anch}} \in \mathbb{R}^{H_2 \times W_2 \times c}, k \in \{1, \cdots, \alpha \cdot \beta\}$, where H_2 and W_2 are the spatial size of FMLI structure in $\hat{\boldsymbol{y}}$, and $c = C/N$. Then $\hat{\boldsymbol{y}}^{i,k}_{\text{anch}}$ is linearly mapped to $\boldsymbol{Q}^{i,k}_{\text{anch}} \in \mathbb{R}^{d_2 \times c}$, $\boldsymbol{K}^{i,k}_{\text{anch}} \in \mathbb{R}^{d_2 \times c}$ and $\boldsymbol{V}^{i,k}_{\text{anch}} \in \mathbb{R}^{d_2 \times c}$, where $d_2 = H_2 \times W_2$. Then, FCA can be formulated as:

$$\mathcal{F}^{i,k}_{\text{FCA}} = \text{Softmax}\left(\frac{\boldsymbol{Q}^{i,k}_{\text{anch}} \otimes (\boldsymbol{K}^{i,k}_{\text{anch}})^\top}{\sqrt{c}}\right) \otimes \boldsymbol{V}^{i,k}_{\text{anch}}, \tag{9}$$

where $\mathcal{F}^{i,k}_{\text{FCA}}$ is the output, which is then reshaped to dimensions of $H_2 \times W_2 \times c$. In practice, we use the multi-head self-attention for the implementation. More analysis of FCA is provided in the ***supplementary material***. The FMLI representations $\mathcal{F}^{i,k}_{\text{FCA}}, k \in \{1, \cdots, \alpha \cdot \beta\}$ are folded back to plenoptic representations, and a LocConv is applied to aggregate the attention responses within each FMLI structure. Then the output feature is sequentially processed by a linear layer, a layer normalization, and an FFN with local residual learning to obtain spatial context feature $\boldsymbol{\Phi}^i_{\text{sp}}$.

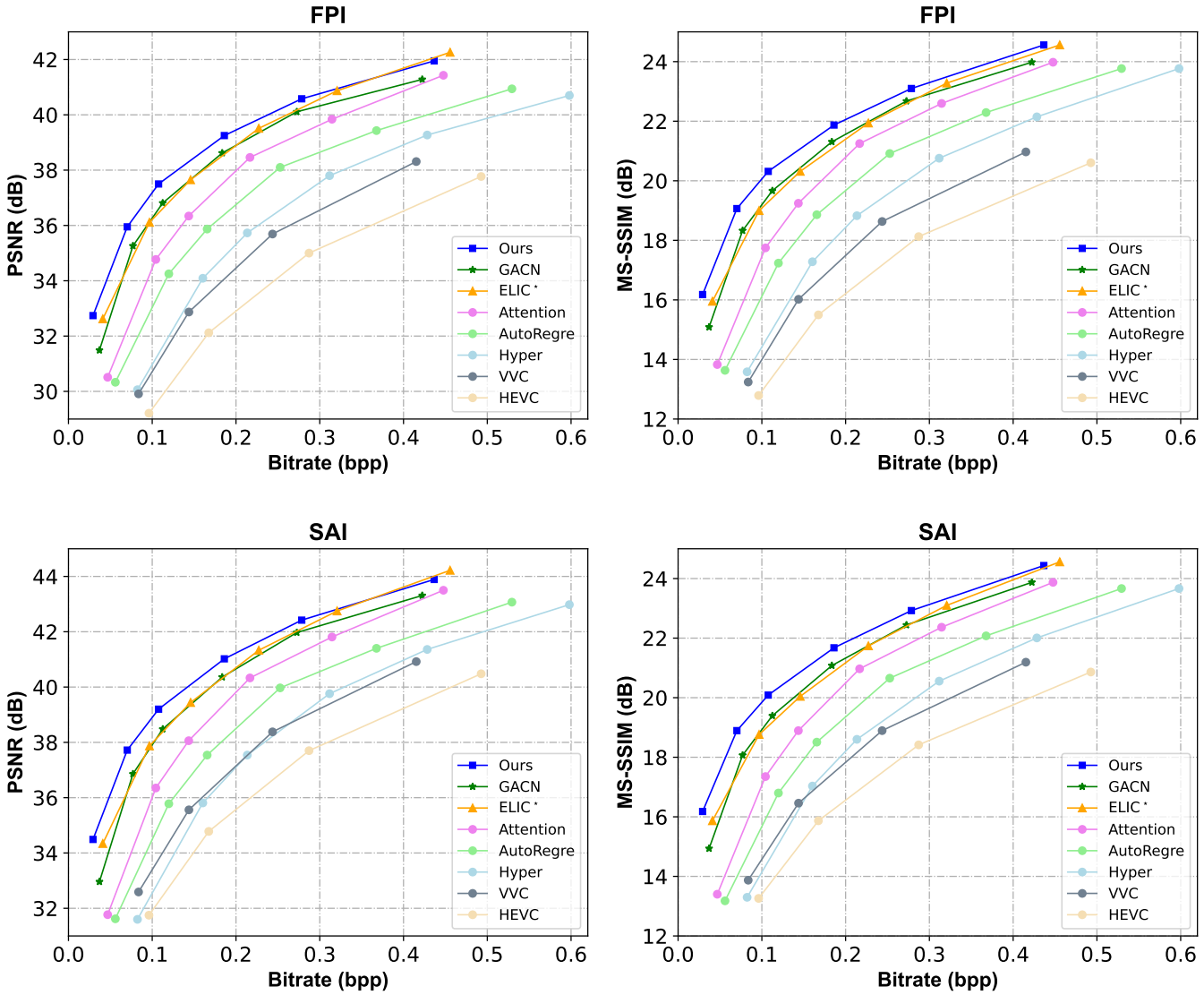


Fig. 6. PSNR-bpp and MS-SSIM-bpp curves of compared methods in both FPI and SAI domains on the testing set. The MS-SSIM is converted to $-10\log_{10}(1-\text{MS-SSIM})$ for a clearer comparison.

IV. Experiments

A. Experimental Setup

Model Settings. The number of channels (C) of both latent $\hat{\boldsymbol{y}}$ and $\hat{\boldsymbol{z}}$ are set to 192, and N is set to 6. Local residual prediction (LRP) [22] is applied to compensate for quantized errors. The model is optimized by the RD formula, as illustrated in Equation 5. Mean squared error (MSE) is employed to calculate the distortion, *i.e.*, $\mathcal{D}(\hat{\boldsymbol{x}}, \boldsymbol{x})$. The bits per pixel (bpp) is used to measure the rate term. Six models are trained with $\lambda \in \{0.1, 0.05, 0.025, 0.01, 0.005, 0.001\}$ to cover a variety of bitrate. Our method is implemented using CompressAI platform [57].

Training. To train our framework, we utilize the pre-processed FPIs from **FPI2K** [8] as the train set, which contains 1877 items captured in diverse scenarios. For training, the pre-processed FPIs (with a resolution of $3168 \times 2016 \times 3$) are cropped into patches with a size of $384 \times 384 \times 3$, which includes an 8×8 array of FMLI with a size of $48 \times 48 \times 3$. During training, a batch of 4 inputs is fed into the network. The initial learning rate is initially set to 1×10^{-4} and reduced by a factor of 0.5 for every 15 epochs. The total training epoch is set to 50. We use an NVIDIA GeForce RTX 3090 GPU and Adam optimizer [58] for training our models.

Testing. For the testing phase, 20 pre-processed FPIs with a resolution of $3168 \times 2016 \times 3$ from **FPI2K** are utilized as a testing set to evaluate the RD performance. Quality evaluation of the decoded FPIs is conducted using PSNR and MS-SSIM [59] metrics in RGB colorspace. We compare our approach with the recent LFPIC method, GACN [8], and five representative LIC approaches, including Factor [46], Hyper [12], AutoRegre [21], Attention [49], and ELIC*[1] [13]. These learned compression methods are trained using the same

[1]ELIC* is a re-implementation from https://github.com/VincentChandelier/ELiC-ReImplemetation since the official codes are not released.

TABLE I
AVERAGE BD-RATE (DISTORTION IS MEASURED BY PSNR IN FPI AND SAI DOMAINS) COMPARISON OF DIFFERENT METHODS ON THE TESTING SET. THE BEST RESULTS ARE HIGHLIGHTED IN **BLOD**. THE ANCHOR IS GACN. OUR METHOD ACHIEVES THE HIGHEST BD-RATE REDUCTION, *i.e.*, 22.16%.

	BD-rate (%)	
Method	FPI	SAI
HEVC	74.05	67.56
VVC	65.02	56.40
Factor [46]	65.81	62.96
Hyper [12]	57.45	55.15
AutoRegre [21]	43.21	42.26
Attention [49]	26.93	26.15
ELIC* [13]	−1.58	−3.44
GACN [8]	0.00	0.00
Ours	**−22.16**	**−22.70**

train set as ours. For conventional codecs, reference software, HM 16.22[2] and VTM 10.00[3] are employed for HEVC- and VVC-intra coding (in YUV 4:4:4 colorspace), respectively. Additionally, we render each decoder FPI by different methods to 5×5 SAIs with a resolution of $1149 \times 830 \times 3$ using open-source tool [8]. The distortion in FPI domain is measured by directly computing the PSNR and MS-SSIM scores of each decoded FPI. Following light field image super-resolution methods [60]–[62], the distortion in SAI domain is measured by averaging the PSNR and MS-SSIM scores of the rendered 5×5 SAIs. The bpp is calculated by dividing the number of bits by the number of pixels of the input FPI.

B. Rate-Distortion Performance

The Bjøntegaard-delta-rate (BD-rate) [63], which quantifies the average difference in bitrate between two methods at a given distortion, is utilized for quantitative RD performance comparison. The averaged results on decoded FPIs and rendered SAIs are reported in Table I. The GACN [8] is selected as anchor (BD-rate 0%) considering it is the first LFPIC method. From Table I, we have three interesting observations:

- The conventional codecs, HEVC and VVC, are inferior to learned compression methods. The primary challenge for these codecs lies in incorporating correlations of FPIs, arising from textural discontinuities and long-distance properties.
- The LIC method, ELIC* [13], achieves a better RD performance compared with GACN. This is because ELIC* incorporates complex channel-spatial contextural coding, while GACN applies only a masked convolution [21] for context modeling. The other LIC methods, such as Attention [49], Autoregre [21], and Hyper [12], incur significantly higher BD-rate cost due to the challenges in handling FPIs.
- Our method achieves the best BD-rate reduction, *i.e.*, 22.16%, demonstrating the superior RD performance of our method. The high performance of our method can be attributed to the effective local-global correlations modeling and channel-spatial contextual coding (detailed analyses are in Section IV-E).

The averaged RD curves of different methods are presented in Fig. 6. Notably, our method consistently outperforms GACN across a broad range of bitrate. The ELIC* achieves a comparable performance with our method at high bitrate.

Discussion. In Fig. 6, we can observe that ELIC* [13] also achieves good RD performance at high bitrate. This phenomenon can be attributed to that at high bitrate (resulting in more bits/information encoded per pixel), directly treating the FPI as a single entity without explicitly considering information discontinuity can still yield favorable RD performance. However, as the bitrate decreases, ELIC* experiences significant degradation in RD performance. In contrast, our proposed method demonstrates a more balanced RD performance across the entire bitrate range. Our method effectively addresses the challenges posed by information discontinuity, resulting in robust performance across different bitrate.

C. Visual Performance

The visual comparisons with different methods in the SAI domain and FPI domain are presented in Fig. 7 and 8, respectively. We present the central view of rendered 5×5 SAIs and the decoded FPI for comparison. The results of the learned methods are tested from models trained with the same λ value. We can observe that the traditional codecs, i.e., HEVC and VVC, require more bits to decode high-quality results. The LIC methods, such as Hyper and Attention, also suffer from obvious artifacts in decoded images. In contrast, our method recovers clearer structures and more visually pleasing details with a higher compression ratio.

D. Efficiency Analysis

To evaluate the coding efficiency of our method, we report the total encoding and decoding times (including entropy estimation) of different methods in Table II. For a fair comparison, all the methods are tested on the same desktop equipped with an Intel Xeon Platinum 8369B CPU (@2.90 GHz, 64GB RAM) and an NVIDIA GeForce RTX 3090 GPU. It can be observed that the context-free methods, *i.e.*, Factor [46] and Hyper [12] require shorter coding times. Among the other five methods deployed with context models, GACN [8] requires the most processing times because it employs high-complexity self-attention mechanisms in the nonlinear transforms. In contrast, our method takes reasonable time for coding and achieves the best compression performance, thanks to our efficient local-global learning strategy.

E. Ablation Investigation

In this subsection, we investigate the effectiveness of the proposed strategies by comparing our method with several

[2] https://vcgit.hhi.fraunhofer.de/jvet/HM
[3] https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM

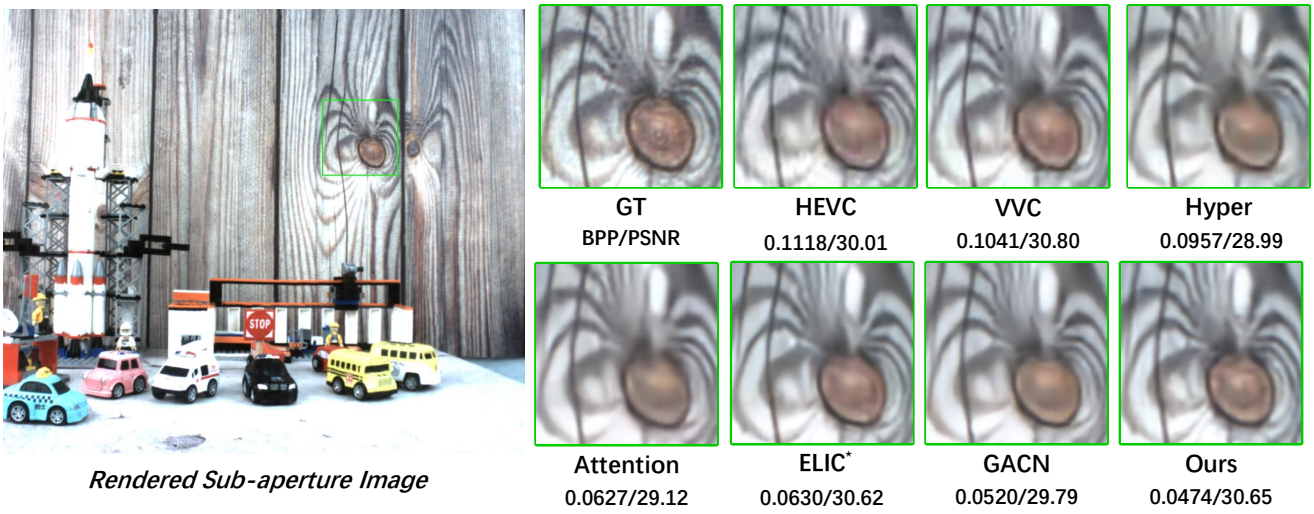


Fig. 7. Visual comparison of rendered SAI from decoded FPI. Compared to other methods, our method not only compresses image to the lowest bitrate, but also reconstructs the highest performance with fine-grained details.

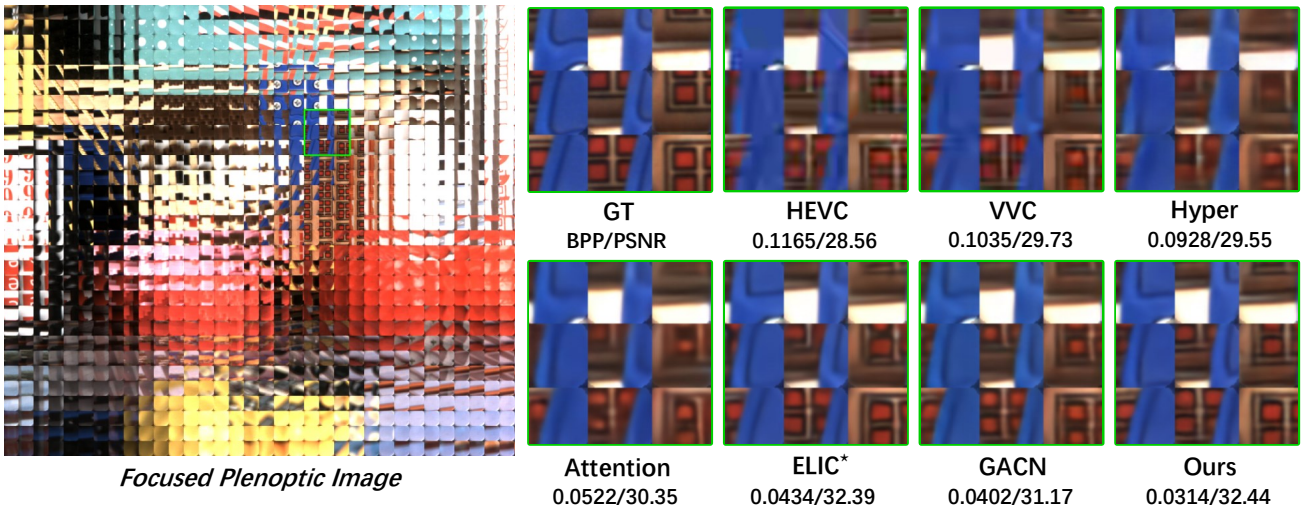


Fig. 8. Visual comparison of decoded FPI. Each enlarged patch contains 3×3 FMLIs. Compared to other methods, our method can reconstruct clearer and sharper structures with the lowest bitrate.

modified variants. We set $\lambda \in \{0.05, 0.025, 0.01, 0.005\}$ for RD performance comparison.

Local-Global Correlation Learning. In our method, we introduce a local-global correlation learning strategy to build the nonlinear transform architectures for LFPIC. Specifically, we treat each FMLI structure as an independent convolutional unit to incorporate local correlations, and we introduce GA to capture the long-distance correlations. To show their effectiveness, we introduce two variants. Firstly, we remove GRABs and replace LocConvs in g_a and g_s with GloConvs. This is similar to GACN [8], which applies $K \times K$ kernels on the entire focused plenoptic image/feature. In Table III, our method can achieve 12.07% RD-rate reduction compared with this variant, w/o_LocGlo. Secondly, we remove the GRABs in g_a and g_s, where only local correlations within each FMLI are considered. From Table III, our method can save 5.47% BD-rate when compared with this variant, w/o_GRAB.

Spatial-wise Context Model. In our method, we introduce a spatial-wise context model tailored specifically for LFPIC. To verify its effectiveness, we introduce two variants. Firstly, we investigate the influence of spatial-wise context modeling by directly removing this component. From Table III, this variant, w/o_SpaCtx is inferior to our method with a 6.99% more BD-rate cost. Furthermore, we modify one variant by incorporating the original masked checkerboard convolution (MCBConv)-based spatial context modeling [23], denoted as

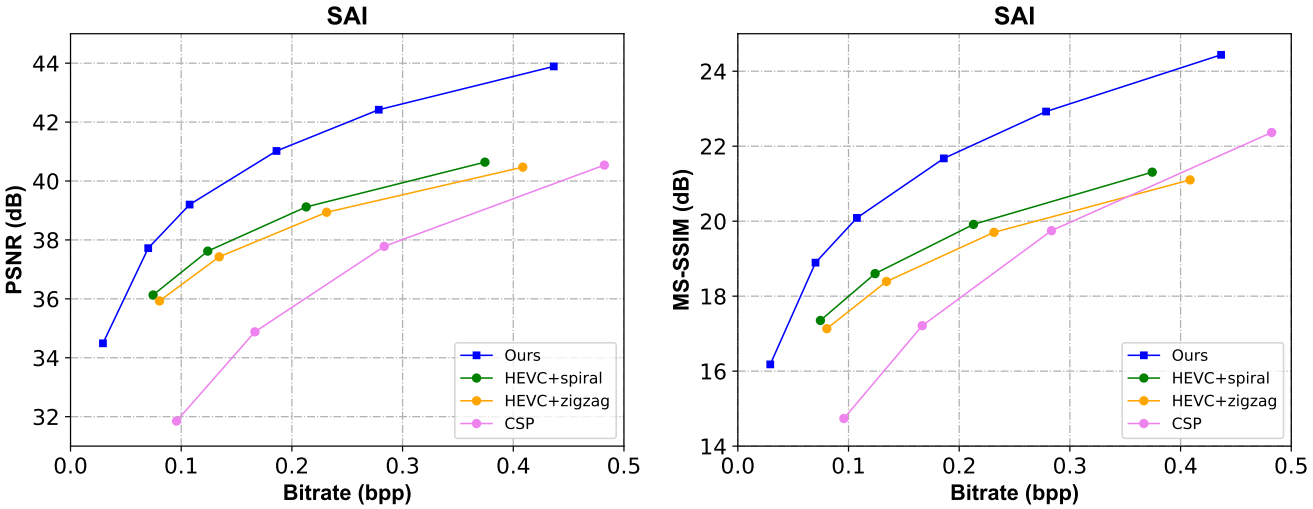


Fig. 9. PSNR-bpp and MS-SSIM-bpp curves (in SAI domain) of plenoptic 1.0 compression strategies and our method. The MS-SSIM is converted to $-10\log_{10}(1-\text{MS-SSIM})$ for a clearer comparison.

TABLE II
TOTAL ENCODING (ENC.) AND DECODING (DEC.) TIME OF DIFFERENT METHODS AVERAGED ON THE TESTING SET. NOTE THAT TO SAVE GPU MEMORY, THE INPUT FPIS ARE CROPPED INTO PATCHES FOR INFERENCE.

	Time (second)	
Method	Enc.	Dec.
Factor [46]	8.34	12.76
Hyper [12]	9.04	13.22
AutoRegre [21]	70.47	158.81
Attention [49]	58.02	146.56
ELIC* [13]	59.43	60.99
GACN [8]	167.11	261.23
w_MCBConv	35.78	49.32
Ours	35.23	47.87

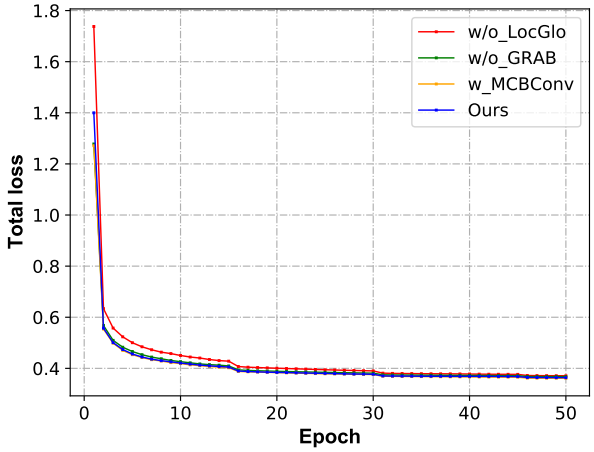


Fig. 10. Training convergence curves with different variants.

TABLE III
ABLATION STUDY RESULTS IN TERMS OF AVERAGE BD-RATE (DISTORTION IS MEASURED BY PSNR IN FPI DOMAIN) ON THE TESTING SET. THE BD-RATE RESULTS ARE REPORTED AS THE COMPARISON BETWEEN EACH VARIANT AND OUR METHOD.

		BD-rate (%)
Ours *v.s.*	w/o_LocGlo	−12.07
	w/o_GRAB	− 5.47
	w/o_SpaCtx	− 6.99
	w_MCBConv	− 0.49

w_MCBConv. As shown in Table III, our method achieves a 0.49% more BD-rate reduction compared with w_MCBConv without compromising the coding efficiency (time comparison can be found in Table II). This improvement is attributed to the fact that the original checkerboard context modeling does not sufficiently consider the most related symbols within each FMLI structure.

In Fig. 10, we provide the curves of total loss with respect to epochs, achieved by different variants during training at $\lambda = 0.01$. It can be observed that the training of all the variants is stable and converges.

TABLE IV
AVERAGE BD-RATE (DISTORTION IS MEASURED BY PSNR IN SAI DOMAIN) COMPARISON RESULTS ON THE TESTING SET. THE ANCHOR IS GACN. THE BEST RESULT IS HIGHLIGHTED IN **BOLD**.

	BD-rate (%)
HEVC + *spiral* scan [25]	32.49
HEVC + *zigzag* scan [24]	40.45
CSP [64]	66.76
GACN [8]	0.00
Ours	**−22.70**

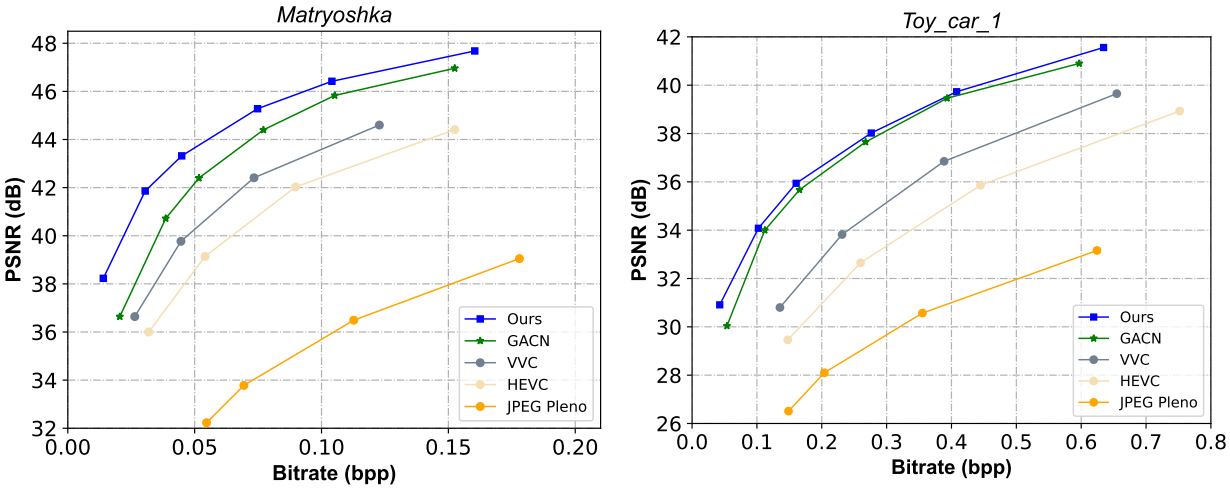


Fig. 11. PSNR-bpp curves of JPEG Pleno, HEVC, VVC, GACN, and our method on scene *Matryoshka* and *Toy_car_1*.

F. Comparison to Plenoptic 1.0 Compression Strategies

We also compare our method to plenoptic 1.0 compression strategies. The rendered SAIs from the testing set are transformed into video sequences using *zigzag* [24] and *spiral* [25] scan orders, and HEVC inter-coding is applied for compression. We also compare our method with an MLI-based method, CSP [64]. The BD-rate comparison results are listed in Table IV. It can be observed that both GACN and our method largely outperform these plenoptic 1.0 compression methods. The RD curves are presented in Fig. 9.

G. Comparison to JPEG Pleno Coding Standard

In this subsection, we compare our method to the JPEG Pleno coding standard [7], using the public reference software[4] for testing under the JPEG Pleno test conditions [65]. The RD curve comparisons on scene *Matryoshka* and *Toy_car_1* are shown in Fig. 11. It can be observed that both our method and GACN [8] significantly outperform the JPEG Pleno. The HEVC- and VVC-intra coding also achieve much better RD performance than JPEG Pleno. The main reason is that the JPEG Pleno transforms the input SAIs from 8-bit to 10-bit depth, resulting in input images with richer color information. However, the increased detailed information leads to larger bitrate and lower RD performance.

V. Conclusion

In this paper, we develop a novel end-to-end method for LFPIC. Motivated by the observations from the characteristics of FPIs, *e.g.*, texture discontinuous and long-distance properties, we introduce a local-global correlation learning strategy to build the nonlinear transforms. This strategy can effectively handle the discontinuous structures and leverage long-distance correlations in FPI for high compression efficiency. Additionally, we introduce channel and spatial contextual coding to further enhance the RD performance. The spatial-wise context model also helps our method emphasize the most related symbols during coding. The experimental results demonstrate that our method achieves state-of-the-art RD performance.

Limitation. Despite the high RD performance, the context model in our approach brings additional time and computational costs compared with the context-free methods [12], [46]. In addition, this work currently focuses on data captured by one specific focused plenoptic camera, TSPC [40]. Cameras with different shapes of the MLA result in different shapes and sizes of FMLI. And different lens setups, such as Galilean mode [66], also lead to different types of FPIs. When handling these FPIs, some LRDBs in the main auto-encoder might need to be replaced with GRDBs if the FMLI cannot be $16\times$ down-sampled. In future work, we plan to explore a more generalizable approach to model the correlations among FMLIs for more efficient LFPIC.

[4] https://gitlab.com/wg1/jpeg-pleno-refsw

References

[1] N. Zeller, F. Quint, and U. Stilla, "Depth estimation and camera calibration of a focused plenoptic camera for visual odometry," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 118, pp. 83–100, 2016.

[2] L. Palmieri, R. Koch, and R. op het Veld, "The plenoptic 2.0 toolbox: Benchmarking of depth estimation methods for mla-based focused plenoptic cameras," *IEEE International Conference on Image Processing (ICIP)*, pp. 649–653, 2018.

[3] J. N. Filipe, P. A. A. Assunção, L. M. N. Tavora, R. Fonseca-Pinto, L. A. Thomaz, and S. M. M. Faria, "Improved patch-based view rendering for focused plenoptic cameras with extended depth-of-field," *2020 28th European Signal Processing Conference*, pp. 680–684, 2021.

[4] J. Shi, H. Qi, Z. Yu, X. An, Y. Ren, and H. Tan, "Three-dimensional temperature reconstruction of diffusion flame from the light-field convolution imaging by the focused plenoptic camera," *Science China Technological Sciences*, vol. 65, no. 2, pp. 302–323, 2022.

[5] Z. Lu, Y. Liu, M. Jin, X. Luo, H. Yue, Z. Wang, S. Zuo, Y. Zeng, J. Fan, Y. Pang *et al.*, "Virtual-scanning light-field microscopy for robust snapshot high-resolution volumetric imaging," *Nature Methods*, vol. 20, no. 5, pp. 735–746, 2023.

[6] "Jpeg pleno," Accessed: Feb. 1, 2024. [Online]. Available: https://jpeg.org/jpegpleno/

[7] P. Schelkens, P. Astola, E. A. Da Silva, C. Pagliari, C. Perra, I. Tabus, and O. Watanabe, "Jpeg pleno light field coding technologies," in *Applications of Digital Image Processing XLII*, vol. 11137. SPIE, 2019, pp. 391–401.

[8] K. Tong, X. Jin, Y. Yang, C. Wang, J. Kang, and F. Jiang, "Learned focused plenoptic image compression with microimage preprocessing and global attention," *IEEE Transactions on Multimedia*, vol. 26, pp. 890–903, 2024.

[9] S. Zhu, M. Li, C. Chen, S. Liu, and B. Zeng, "Cross-space distortion directed color image compression," *IEEE Transactions on Multimedia*, vol. 20, no. 3, pp. 525–538, 2017.

[10] Y. Mei, L. Li, Z. Li, and F. Li, "Learning-based scalable image compression with latent-feature reuse and prediction," *IEEE Transactions on Multimedia*, vol. 24, pp. 4143–4157, 2021.

[11] G. Toderici, S. M. O'Malley, S. J. Hwang, D. Vincent, D. Minnen, S. Baluja, M. Covell, and R. Sukthankar, "Variable rate image compression with recurrent neural networks," in *International Conference on Learning Representations*, 2016, pp. 1–12.

[12] 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.

[13] D. He, Z. Yang, W. Peng, R. Ma, H. Qin, and Y. 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*, 2022, pp. 5718–5727.

[14] Z. Zhang, B. Chen, H. Lin, J. Lin, X. Wang, and T. Zhao, "Elfic: A learning-based flexible image codec with rate-distortion-complexity optimization," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 9252–9261.

[15] K. Wu, Y. Yang, Q. Liu, G. Jiang, and X.-P. Zhang, "Hierarchical independent coding scheme for varifocal multiview images based on angular-focal joint prediction," *IEEE Transactions on Multimedia*, vol. 26, pp. 2993–3006, 2024.

[16] N. Bakir, W. Hamidouche, O. Déforges, K. Samrouth, and M. Khalil, "Light field image compression based on convolutional neural networks and linear approximation," in *IEEE International conference on image processing (ICIP)*. IEEE, 2018, pp. 1128–1132.

[17] Z. Zhao, S. Wang, C. Jia, X. Zhang, S. Ma, and J. Yang, "Light field image compression based on deep learning," in *2018 IEEE International conference on multimedia and expo*. IEEE, 2018, pp. 1–6.

[18] T. Zhong, X. Jin, and K. Tong, "3d-cnn autoencoder for plenoptic image compression," in *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2020, pp. 209–212.

[19] K. Tong, X. Jin, C. Wang, and F. Jiang, "Sadn: learned light field image compression with spatial-angular decorrelation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 1870–1874.

[20] J. Shi and C. Guillemot, "Light field compression via compact neural scene representation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[21] D. Minnen, J. Ballé, and G. D. Toderici, "Joint autoregressive and hierarchical priors for learned image compression," *Advances in neural information processing systems*, vol. 31, pp. 1–10, 2018.

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

[23] 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.

[24] F. Dai, J. Zhang, Y. Ma, and Y. Zhang, "Lenselet image compression scheme based on subaperture images streaming," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2015, pp. 4733–4737.

[25] A. Vieira, H. Duarte, C. Perra, L. Tavora, and P. Assuncao, "Data formats for high efficiency coding of lytro-illum light fields," in *International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE, 2015, pp. 494–497.

[26] C. Perra and P. Assuncao, "High efficiency coding of light field images based on tiling and pseudo-temporal data arrangement," in *2016 IEEE International Conference on Multimedia & Expo Workshops*. IEEE, 2016, pp. 1–4.

[27] J. Hou, J. Chen, and L.-P. Chau, "Light field image compression based on bi-level view compensation with rate-distortion optimization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 2, pp. 517–530, 2018.

[28] D. Liu, Y. Huang, Y. Fang, Y. Zuo, and P. An, "Multi-stream dense view reconstruction network for light field image compression," *IEEE Transactions on Multimedia*, vol. 25, pp. 4400–4414, 2023.

[29] X. Jin, H. Han, and Q. Dai, "Plenoptic image coding using macropixel-based intra prediction," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3954–3968, 2018.

[30] R. J. Monteiro, P. J. Nunes, N. M. Rodrigues, and S. M. Faria, "Light field image coding using high-order intrablock prediction," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 7, pp. 1120–1131, 2017.

[31] C. Conti, P. Nunes, and L. D. Soares, "Light field image coding with jointly estimated self-similarity bi-prediction," *Signal Processing: Image Communication*, vol. 60, pp. 144–159, 2018.

[32] E. Dib, M. Le Pendu, X. Jiang, and C. Guillemot, "Local low rank approximation with a parametric disparity model for light field compression," *IEEE Transactions on Image Processing*, vol. 29, pp. 9641–9653, 2020.

[33] M. B. de Carvalho, M. P. Pereira, G. Alves, E. A. da Silva, C. L. Pagliari, F. Pereira, and V. Testoni, "A 4d dct-based lenslet light field codec," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2018, pp. 435–439.

[34] W. Ahmad, S. Vagharshakyan, M. Sjöström, A. Gotchev, R. Bregovic, and R. Olsson, "Shearlet transform-based light field compression under low bitrates," *IEEE Transactions on Image Processing*, vol. 29, pp. 4269–4280, 2020.

[35] D. Liu, P. An, R. Ma, W. Zhan, X. Huang, and A. A. Yahya, "Content-based light field image compression method with gaussian process regression," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 846–859, 2019.

[36] X. Huang, Y. Chen, P. An, and L. Shen, "Prediction-oriented disparity rectification model for geometry-based light field compression," *IEEE Transactions on Broadcasting*, vol. 69, no. 1, pp. 62–74, 2023.

[37] X. Jiang, J. Shi, and C. Guillemot, "An untrained neural network prior for light field compression," *IEEE Transactions on Image Processing*, vol. 31, pp. 6922–6936, 2022.

[38] J. Shi, Y. Xu, and C. Guillemot, "Learning kernel-modulated neural representation for efficient light field compression," *IEEE Transactions on Image Processing*, vol. 33, pp. 4060–4074, 2024.

[39] Y. Li, M. Sjöström, R. Olsson, and U. Jennehag, "Coding of focused plenoptic contents by displacement intra prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 7, pp. 1308–1319, 2015.

[40] L. Li, X. Jin, and T. Zhong, "Imaging-correlated intra prediction for plenoptic 2.0 video coding," in *IEEE International Conference on Multimedia and Expo*. IEEE, 2020, pp. 1–6.

[41] X. Jin, F. Jiang, L. Li, and T. Zhong, "Plenoptic 2.0 intra coding using imaging principle," *IEEE Transactions on Broadcasting*, vol. 68, no. 1, pp. 110–122, 2021.

[42] Y. Yang, X. Jin, K. Tong, C. Wang, and H. Huang, "Microimage-based two-step search for plenoptic 2.0 video coding," in *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023, pp. 2567–2572.

[43] B. Liu, Y. Zhao, X. Jiang, X. Ji, S. Wang, Y. Liu, and J. Wei, "5-d epanechnikov mixture-of-experts in light field image compression," *IEEE Transactions on Image Processing*, vol. 33, pp. 4029–4043, 2024.

[44] C.-Y. Chu, D. J. Henderson, and C. F. Parmeter, "On discrete epanechnikov kernel functions," *Computational statistics & data analysis*, vol. 116, pp. 79–105, 2017.

[45] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *International Conference on Learning Representations*, 2014, pp. 1–14.

[46] J. Ballé, V. Laparra, and E. P. Simoncelli, "End-to-end optimized image compression," in *International Conference on Learning Representations*, 2017, pp. 1–27.

[47] R. Zou, C. Song, and Z. 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*, 2022, pp. 17 492–17 501.

[48] 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.

[49] 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.

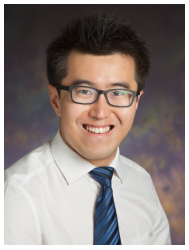
[50] Y. Qian, M. Lin, X. Sun, Z. Tan, and R. Jin, "Entroformer: A transformer-based entropy model for learned image compression," in *International Conference on Learning Representations*, 2022.

[51] W. Jiang, J. Yang, Y. Zhai, P. Ning, F. Gao, and R. Wang, "Mlic: Multi-reference entropy model for learned image compression," in *Proceedings*

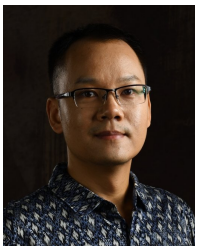
of the 31st ACM International Conference on Multimedia, 2023, pp. 7618–7627.
[52] L. Theis, W. Shi, A. Cunningham, and F. Huszár, "Lossy image compression with compressive autoencoders," in *International Conference on Learning Representations*, 2017, pp. 1–19.
[53] A. Vaswani, N. Shazeer, N. Parmar, 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, pp. 1–11, 2017.
[54] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.
[55] Z. Shen, M. Zhang, H. Zhao, S. Yi, and H. Li, "Efficient attention: Attention with linear complexities," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2021, pp. 3531–3539.
[56] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Restormer: Efficient transformer for high-resolution image restoration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5728–5739.
[57] J. Bégaint, F. Racapé, S. Feltman, and A. Pushparaja, "Compressai: a pytorch library and evaluation platform for end-to-end compression research," *arXiv preprint arXiv:2011.03029*, 2020.
[58] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[59] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. IEEE, 2003, pp. 1398–1402.
[60] G. Liu, H. Yue, J. Wu, and J. Yang, "Intra-inter view interaction network for light field image super-resolution," *IEEE Transactions on Multimedia*, vol. 25, pp. 256–266, 2023.
[61] G. Liu, H. Yue, K. Li, and J. Yang, "Disparity-guided light field image super-resolution via feature modulation and recalibration," *IEEE Transactions on Broadcasting*, vol. 69, no. 3, pp. 740–752, 2023.
[62] G. Liu, H. Yue, J. Wu, and J. Yang, "Efficient light field angular super-resolution with sub-aperture feature learning and macro-pixel upsampling," *IEEE Transactions on Multimedia*, vol. 25, pp. 6588–6600, 2023.
[63] G. Bjontegaard, "Calculation of average psnr differences between rd-curves," *ITU SG16 Doc. VCEG-M33*, 2001.
[64] H. Han, X. Jin, and Q. Dai, "Lenslet image compression using adaptive macropixel prediction," in *International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 4008–4012.
[65] F. Pereira, C. Pagliari, E. da Silva, I. Tabus, H. Amirpour, M. Bernardo, and A. Pinheiro, "Jpeg pleno light field coding common test conditions v3. 2," *Doc. ISO/IEC JTC*, vol. 1, 2019.
[66] A. Lumsdaine and T. Georgiev, "The focused plenoptic camera," in *2009 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 2009, pp. 1–8.

Gaosheng Liu received the B.E. degree from the School of Communication and Information Engineering, Shanghai University, Shanghai, China, in 2019. He is currently pursuing the Ph.D. degree in the School of Electrical and Information Engineering, Tianjin University, Tianjin, China. His research interests include light field imaging, super-resolution, and compression.

Huanjing Yue received the B.S. and Ph.D. degrees from Tianjin University, Tianjin, China, in 2010 and 2015, respectively. She was an Intern with Microsoft Research Asia from 2011 to 2012, and from 2013 to 2015. She visited the Video Processing Laboratory, University of California at San Diego, from 2016 to 2017. She is currently an Associate Professor with the School of Electrical and Information Engineering, Tianjin University. Her current research interests include image processing and computer vision. She received the Microsoft Research Asia Fellowship Honor in 2013 and was selected into the Elite Scholar Program of Tianjin University in 2017.



Bihan Wen received the B.Eng. degree in electrical and electronic engineering from Nanyang Technological University, Singapore, in 2012, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of Illinois at Urbana–Champaign, Champaign, IL, USA, in 2015 and 2018, respectively. He is currently a Nanyang Assistant Professor with the School of Electrical and Electronic Engineering, Nanyang Technological University. His research interests include machine learning, computational imaging, computer vision, image and video processing, and AI security. He was ranked the World's Top 2% Scientists in artificial intelligence by Stanford University, in 2021 and 2023, consecutively. He received the 2023 CASS VSPC Rising Star (Runner-Up). He received the 2022 Early Career Teaching Excellence Award, the 2021 Inspirational Mentor for Koh Boon Hwee Award from NTU, the 2016 Yee Fellowship from UIUC, and the 2012 Professional Engineers Board Gold Medal from Singapore. He was a recipient of the Best Paper Runner Up Award at IEEE ICME 2020, the Best Paper Award from IEEE ICIEA 2023, and the Best Paper Award from IEEE MIPR 2023. He is an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY; and is serving as the Guest Editor for IEEE SIGNAL PROCESSING MAGAZINE, IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING, and *Remote Sensing* (MDPI).



Jingyu Yang received the B.E. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2003, and Ph.D. (Hons.) degree from Tsinghua University, Beijing, in 2009. He has been a Faculty Member with Tianjin University, China, since 2009, where he is currently a Professor with the School of Electrical and Information Engineering. He was with Microsoft Research Asia (MSRA), Beijing, in 2011, within the MSRAs Young Scholar Supporting Program, and with the Signal Processing Laboratory, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, in 2012, and from 2014 to 2015. His research interests include computer vision, computational photography, and 3D vision. He has authored or co-authored over 170 high-quality research papers (including dozens of IEEE Transactions and top conference papers). As a co-author, he got the best 10% paper award in IEEE VCIP 2016 and the Platinum Best Paper award in IEEE ICME 2017. Dr. Yang was the Special Session Chair in the International Conference on Visual Communications and Image Processing 2016 and the Area Chair in the International Conference on Image Processing 2017. He was selected in the program for New Century Excellent Talents in University (NCET) from the Ministry of Education, China, in 2011, the Tianjin Municipal Innovation Talent Promotion Program in 2015, and a National High-Level Youth Talent Program, China, in 2020.