DEEP UNFOLDING NETWORK VIA TRI-DOMAIN STRUCTURAL LEARNING FOR LIGHT FIELD IMAGE SUPER-RESOLUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Light field image super-resolution (LFSR) aims to reconstruct high-resolution LF images from low-resolution observations while preserving geometric structures. However, existing model-based and deep learning-based methods cannot simultaneously achieve rigorous mathematical modeling and powerful feature representation. To address this issue, we propose a deep unfolding network via tri-domain structural learning (LF-DUTS). Specifically, we formulate a 4D LF degradation model under the maximum a posteriori framework and unfold its optimization process into an interpretable multi-stage network. We further design a tri-domain structural refinement group to jointly exploit texture, multi-view, and disparity information from the spatial, angular, and epipolar plane image domains. In addition, we develop a cross-stage feature fusion module to enable feature interaction across unfolding stages. Experimental results on multiple benchmark datasets demonstrate that LF-DUTS consistently outperforms state-of-the-art methods, producing sharper textures, and better disparity consistency.

1 INTRODUCTION

Light field (LF) imaging captures 4D radiance information, encoding the relationship between scene geometry and ray propagation. This unique property enables a wide range of computational photography applications, including refocusing Alain & Smolic (2021), view synthesis Wu et al. (2021), depth estimation Wang et al. (2026), and 3D reconstruction Song et al. (2020). However, due to the physical limitations of microlens arrays, LF imaging inherently suffers from a trade-off between spatial and angular resolutions. Consequently, most LF cameras capture light fields with relatively low spatial resolution, limiting their applicability in high-level vision tasks. To address this issue, light field image super-resolution (LFSR) methods have attracted increasing attention. These methods aim to reconstruct high-resolution (HR) LF images from low-resolution (LR) observations while preserving the intrinsic 4D geometric structure.

Early LFSR studies mainly focused on model-based methods. These methods rely on hand-crafted priors and mathematical optimization formulations to explicitly characterize the geometric structure of LF images, solving the SR problem through iterative optimization algorithms (Fig. 1(a)). Representative approaches include Bayesian models for aliasing-free SR reconstruction Bishop et al. (2009), variational regularization for disparity-aware LFSR Wanner & Goldluecke (2013), graph regularization for enforcing EPI consistency in SR Rossi & Frossard (2018), and sparse coding in the 5D transform domain to exploit spatial-angular correlations for SR Alain & Smolic (2018). These methods are theoretically sound and highly interpretable.

Recently, deep learning-based LFSR methods have achieved significant progress. They rely on end-to-end networks to extract features and learn mappings from LR LF images to HR LF images (Fig. 1(b)). Early convolutional neural network (CNN)-based methods improve reconstruction performance through spatial-angular interactions Yeung et al. (2018); Wang et al. (2020), deformable convolutions Li et al. (2025b), and feature decoupling Wang et al. (2022b); Zhang et al. (2026). Subsequently, Transformer-based models, such as LF-DET Cong et al. (2023), M2MT-Net Hu et al. (2025), and EPIT-SSR Liang et al. (2026), further enhance reconstruction performance by capturing long-range dependencies through self-attention mechanisms. More recently, state space model

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(SSM)-based methods have emerged as an alternative to Transformers, employing multi-directional selective scanning Lu et al. (2024); Gao et al. (2024); Wei et al. (2025) to improve feature extraction while maintaining linear computational complexity. These methods demonstrate the strong representation capability of deep learning for LFSR.

Despite the respective advantages of both categories of methods, achieving rigorous mathematical modeling and effective feature representation remains a key challenge in LFSR. On the one hand, model-based methods combine physical degradation models, hand-crafted priors, and optimization algorithms to solve the LFSR problem, offering strong interpretability and theoretical grounding. However, the limited expressiveness of hand-crafted priors makes it difficult to capture the complex spatial-angular correlations of LF images, thereby limiting the ability of these methods to recover rich texture details. On the other hand, deep learning-based methods achieve superior reconstruction performance by learning nonlinear mappings from LR LF images to HR LF images. However, they usually lack explicit modeling of the underlying physical degradation process, resulting in insufficient geometric constraints among different views and making it difficult to preserve disparity consistency. Therefore, integrating physical models with deep neural networks provides a promising paradigm that combines rigorous degradation modeling with effective feature learning.

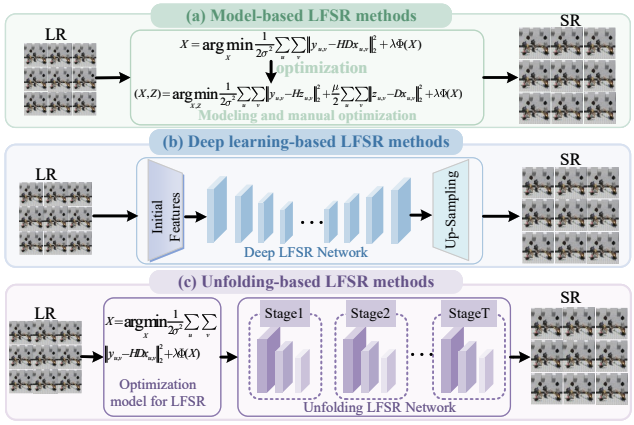


Figure 1: The comparison of representative LFSR frameworks. (a) Model-based LFSR. (b) Deep learning-based LFSR. (c) Unfolding-based LFSR (our proposed).

Based on the above considerations, we propose a deep unfolding network via tri-domain structural learning for light field image super-resolution (LF-DUTS). Specifically, we first formulate a 4D LF degradation model and derive the corresponding optimization framework based on maximum a posteriori (MAP) estimation. The optimization process is then unfolded into a multi-stage interpretable network, where each stage corresponds to one iteration of the optimization subproblems, thereby aligning the optimization procedure with the forward inference process of the network. To enhance the structural modeling capability of each unfolding stage, we embed a tri-domain structural refinement group into the prior update branch to jointly refine iterative features from the spatial, angular, and epipolar plane image (EPI) domains. Specifically, the spatial-angular domain modules employ Transformer blocks to model fine textures and multi-view structural information, while the EPI-domain module utilizes EPI deformable strip convolution to adaptively capture geometric structures with varying disparities. Furthermore, a cross-stage feature fusion module is introduced to enable information exchange across unfolding stages, leading to improved reconstruction performance.

2 Related Work

2.1 Light Field Image Super-Resolution

Early LFSR methods primarily relied on LF imaging models and handcrafted priors to exploit the geometric and statistical properties of LFs for SR reconstruction, including Bayesian inference Bishop et al. (2009), variational optimization Wanner & Goldluecke (2013), Gaussian mixture modeling Mitra & Veeraraghavan (2012), subspace learning Farrugia et al. (2017), BM5D-based extensions Alain & Smolic (2018), and graph regularization Rossi & Frossard (2018). With the rapid development of deep learning, CNN-based methods have become the dominant paradigm by progressively enhancing spatial-angular feature representation for LFSR through spatial-angular feature interaction Yoon et al. (2017), 4D convolution Meng et al. (2019), and spatial-angular decoupling strategies Wang et al. (2020; 2022b); Zhang et al. (2026). To further improve long-range dependency mod-

eling for LF reconstruction, Transformer-based methods have been introduced into LFSR, leveraging self-attention mechanisms to capture complementary information across spatial, angular, and EPI domains Cong et al. (2023); Hu et al. (2025); Liao et al. (2025); Liang et al. (2026). More recently, state space model (SSM)-based Mamba architectures have emerged as a promising alternative, achieving efficient spatial-angular dependency modeling for LFSR with linear computational complexity while maintaining competitive reconstruction performance Lu et al. (2024); Gao et al. (2024); Wei et al. (2025).

2.2 Deep Unfolding Network for Super-Resolution

Deep unfolding networks (DUNs) bridge model-based optimization and data-driven learning by unfolding iterative optimization algorithms into trainable networks, achieving promising results in various SR tasks. In single-image SR, DUNs integrate physical degradation modeling Zhang et al. (2020), deep priors Ning et al. (2021), blur kernel estimation Huang et al. (2020); Zheng et al. (2022), and structural constraints Ni et al. (2025) into optimization frameworks. In hyperspectral image SR, they further explore spatial-spectral prior modeling Liu et al. (2023); Tang et al. (2024), cross-stage propagation Hu et al. (2024), frequency-domain learning Fang et al. (2025), and physics-guided degradation modeling Li et al. (2025c). For multimodal image SR, unfolding-based methods investigate sparse optimization Marivani et al. (2020), second-order optimization Zhang et al. (2025), edge-preserving constraints Zhang et al. (2023), and degradation-aware reconstruction Zhao et al. (2023). However, existing DUN-based SR methods mainly focus on 2D image reconstruction, while their application to LF images remains underexplored due to the unique 4D spatial-angular structure.

3 Our proposed LF-DUTS

We propose LF-DUTS, as illustrated in Fig. 2(a). By unfolding the LF degradation model into a multi-stage learnable architecture, LF-DUTS effectively bridges physical modeling and deep learning, achieving an interpretable framework for LFSR.

3.1 Degradation Model

We represent a 4D LF as $\mathcal{L} \in \mathbb{R}^{U\times V\times H\times W}$ using the two-plane parameterization scheme, where (U, V) and (H, W) denote the angular and spatial resolutions, respectively. The degradation process of each sub-aperture image is formulated as

$$y_{u,v} = HDx_{u,v} + \mathcal{N}_{u,v}, \tag{1}$$

where H and D represent the blur and downsampling operators, respectively, and $\mathcal{N}_{u,v}$ denotes the additive noise term.

Based on the maximum a posteriori (MAP) framework, the data fidelity term, LF prior term, and geometric consistency term are jointly formulated as

$$X^* = \arg\min_X \frac{1}{2\sigma^2}\sum_u\sum_v \|y_{u,v} - HDx_{u,v}\|_2^2 + \Phi(X) + \mathcal{X}^T L\mathcal{X}, \tag{2}$$

where the three terms represent the data fidelity term, LF prior term, and graph Laplacian regularization term for cross-view geometric consistency, respectively.

3.2 Unfolding Optimization

To solve the above optimization problem (13), we adopt the half-quadratic splitting (HQS) algorithm by introducing an auxiliary variable Z. The original problem is transformed into the following equivalent form:

$$\begin{aligned}(X, Z) = \arg\min_{X,Z} &\frac{1}{2\sigma^2}\sum_u\sum_v \|y_{u,v} - Hz_{u,v}\|_2^2 \\ &+ \frac{\mu}{2}\sum_u\sum_v \|z_{u,v} - Dx_{u,v}\|_2^2 + \Phi(X) + \mathcal{X}^T L\mathcal{X},\end{aligned} \tag{3}$$

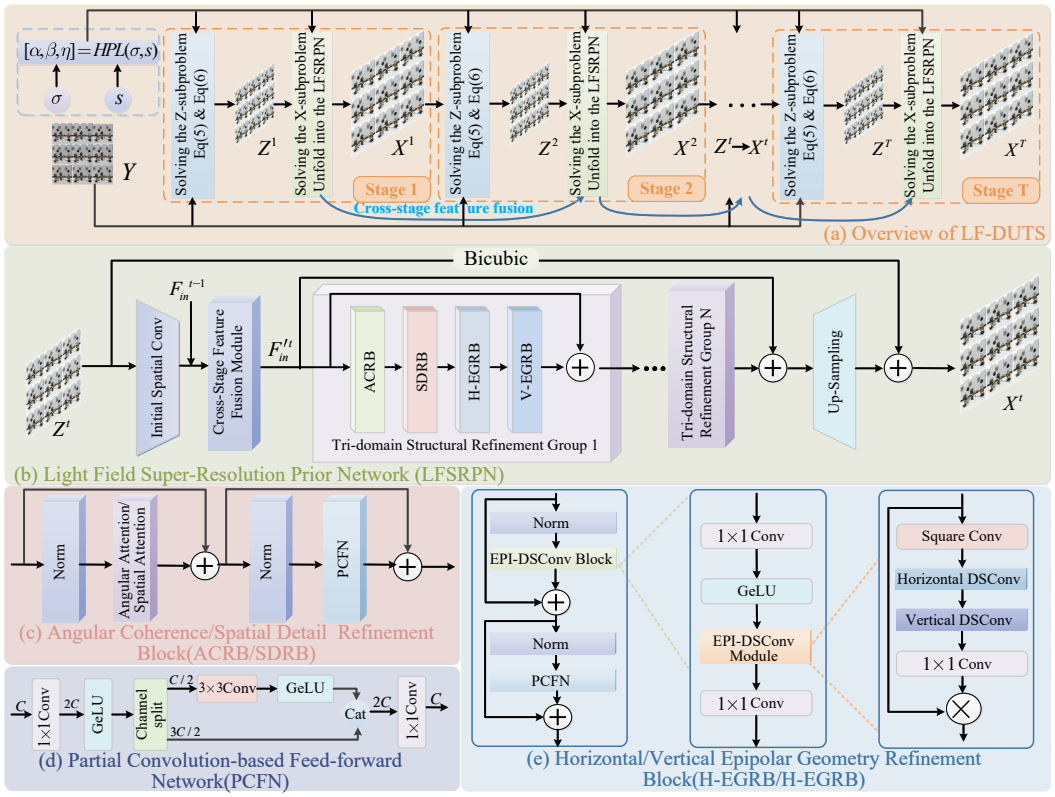


Figure 2: Overall architecture of the proposed LF-DUTS. (a) Overview of the deep unfolding framework. (b) Light Field Super-Resolution Prior Network (LFSRPN). (c) Angular Coherence/Spatial Detail Refinement Block (ACRB/SDRB). (d) Partial Convolution-based Feed-forward Network (PCFN). (e) Horizontal/Vertical Epipolar Geometry Refinement Block (H-EGRB/V-EGRB).

where μ is the penalty parameter controlling the consistency between Z and X. The above problem is then alternatively optimized by solving the following two subproblems:

$$\begin{cases} Z^t = \arg\min_Z \sum_u \sum_v \|y_{u,v} - H z_{u,v}^{t-1}\|_2^2 + \mu^t \sigma^2 \sum_u \sum_v \|z_{u,v}^{t-1} - D x_{u,v}^{t-1}\|_2^2, \\ X^t = \arg\min_X \sum_u \sum_v \frac{\mu^t}{2} \|z_{u,v}^t - D x_{u,v}^{t-1}\|_2^2 + \Phi(X^{t-1}) + (\mathcal{X}^{t-1})^T L \mathcal{X}^{t-1}, \end{cases} \tag{4}$$

where $t \in [1, T]$ denotes the iteration index.

Z-subproblem: The data fidelity subproblem is approximately solved by one-step gradient descent(the detailed derivation is provided in the Appendix A.1):

$$\begin{aligned} z_{u,v}^t &= z_{u,v}^{t-1} - \eta_t \left[H^T (H z_{u,v}^{t-1} - y_{u,v}) + \alpha_t (z_{u,v}^{t-1} - D x_{u,v}^{t-1}) \right] \\ &= \bar{H} z_{u,v}^{t-1} + \eta_t H^T y_{u,v} + \eta_t \alpha_t D x_{u,v}^{t-1} \end{aligned} \tag{5}$$

where $\alpha_t = \mu_t \sigma^2$ and $\bar{H} = (1 - \eta_t \alpha_t) I - \eta_t H^T H$. The updated SAIs are then aggregated according to their spatial-angular coordinates to reconstruct the LF:

$$Z^t = \left\langle z_{u,v}^t \right\rangle_{u=1,v=1}^{U,V} \tag{6}$$

X-subproblem: The X-subproblem can be rewritten as

$$X^t = \arg\min_X \frac{1}{2(\beta_t)^2} \sum_u \sum_v \|z_{u,v}^t - D x_{u,v}^{t-1}\|_2^2 + \Phi(X^{t-1}) + (\mathcal{X}^{t-1})^T L \mathcal{X}^{t-1} \tag{7}$$

where $\beta_t = \sqrt{1/\mu_t}$. To overcome the limited representation capability of handcrafted priors, we introduce a light field super-resolution prior network (LFSRPN) to model spatial, angular, and EPI priors for solving the X-subproblem. Following Zhang et al. (2020), we further employ a hyper-parameter network (HPN) to adaptively predict $\{\alpha_t, \beta_t, \eta_t\}_{t=1}^{T}$ under different degradation conditions, enabling end-to-end optimization of the unfolding network. The details of HPN are provided in the Appendix A.2.

3.3 Architecture of LFSRPN

As shown in Fig. 2(b), LFSRPN consists of three components: initial feature extraction, deep feature learning, and reconstruction. First, shallow features are extracted by an initial convolution layer and a cross-stage feature fusion module (CSFM). Then, these features are fed into a backbone composed of N cascaded tri-domain structural refinement groups (TSRGs) to learn deep representations. Each TSRG sequentially contains an angular consistency refinement block (ACRB), a spatial detail refinement block (SDRB), and the horizontal/vertical EPI geometry refinement block (H-EGRB/V-EGRB). Finally, shallow and deep features are fused and reconstructed through convolution and PixelShuffle with global residual learning. The details of CSFM, ACRB, SDRB, and EGRB are described below.

CSFM: Most deep unfolding networks lack direct feature interaction across stages, which hinders high-frequency information propagation Zhang et al. (2020); Ning et al. (2021). To address this issue, we design a CSFM, as shown in Fig. 3. Specifically, features from the previous stage are first processed by max-pooling to extract high-frequency responses: $F_{hf}^{t-1} = \mathrm{MaxPool}(F_{in}^{t-1})$. Then, the current-stage features are concatenated with the high-frequency features to generate adaptive gates for feature enhancement. The process is formulated as

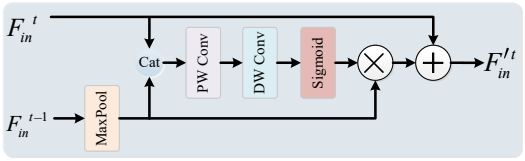


Figure 3: Implementation details of CSFM.

$$
\begin{aligned}
G &= \sigma\Big(\mathrm{DWConv}\Big(\mathrm{PWConv}\Big(\mathrm{Cat}(F_{in}^{t}, F_{hf}^{t-1})\Big)\Big)\Big), \\
F_{in}^{\prime t} &= F_{in}^{t} + G \odot F_{hf}^{t-1},
\end{aligned}
\tag{8}
$$

where $\odot$ denotes element-wise multiplication, $\mathrm{PWConv}(\cdot)$ and $\mathrm{DWConv}(\cdot)$ denote point-wise and depth-wise convolutions, respectively, $\mathrm{Cat}(\cdot,\cdot)$ denotes channel concatenation, and $\sigma(\cdot)$ denotes the Sigmoid activation.

ACRB and SDRB: To jointly reconstruct spatial details and preserve multi-view geometric consistency of light fields, we design an SDRB and an ACRB. As shown in Fig. 2(c), both blocks share the same residual architecture and perform domain-specific feature modeling via different attention mechanisms. Given the input feature F_{*in}^{t}, it is first normalized by layer normalization (LN) and fed into the corresponding domain attention module for feature enhancement, followed by a partial convolution-based feed-forward network (PCFN) for feature transformation. The process is formulated as

$$
\begin{aligned}
F_{*att}^{t} &= F_{*in}^{t} + \mathrm{Attention}_{*}\left(\mathrm{LN}(F_{*in}^{t})\right), \\
F_{*out}^{t} &= F_{*att}^{t} + \mathrm{PCFN}\left(\mathrm{LN}(F_{*att}^{t})\right),
\end{aligned}
\tag{9}
$$

where $* \in \{\mathrm{spa}, \mathrm{ang}\}$ denotes the spatial or angular domain, and $\mathrm{Attention}_{*}(\cdot)$ represents the corresponding attention operator. Specifically, ACRB employs angular attention to capture inter-view correlations and preserve LF geometric consistency, while SDRB adopts spatial attention to enhance high-frequency details such as edges and textures. As illustrated in Fig. 2(d), PCFN expands the channel dimension using a 1×1 convolution, performs local context modeling on partial channels with a 3×3 convolution, and restores feature dimensions through another 1×1 convolution, improving representation capability while maintaining computational efficiency.

EGRB: In the X-subproblem (17), the graph regularization term $(\mathcal{X}^{t-1})^T L \mathcal{X}^{t-1}$ is introduced to enforce the geometric consistency of LF EPIs. Accordingly, we unfold this constraint into the H-EGRB and the V-EGRB. Since both blocks share the same architecture with different EPI orientations, they are collectively referred to as EGRB. As shown in Fig. 2(e), EGRB adopts the same

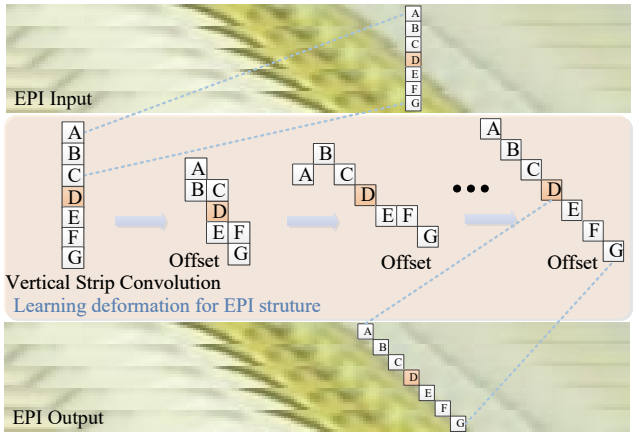


Figure 4: Implementation details of the EPI DSConv.

residual framework as ACRB and SDRB, while replacing the attention module in Eq. (9) with an EPI deformable strip convolution (EPI-DSConv) block for EPI geometry modeling. Specifically, the input feature is first projected by a 1×1 convolution, then fed into the EPI-DSConv module to extract direction-aware geometric features, and finally restored to the original channel dimension by another 1×1 convolution:

$$F^t_{epi_out} = \mathrm{Conv}\big(\mathrm{DSConv}\big(\mathrm{GeLU}\left(\mathrm{Conv}(F^t_{epi_in})\right)\big)\big)\,. \tag{10}$$

The EPI-DSConv module integrates local context modeling and direction-aware feature extraction. It first employs standard convolution to capture local information and then utilizes DSConv to adaptively model disparity-induced slanted structures in EPIs. As shown in Fig. 4, taking the vertical direction as an example, given a strip convolution sampling grid $\mathcal{R}_v = \{(0, -\lfloor K/2 \rfloor), \ldots, (0, \lfloor K/2 \rfloor)\}$ with kernel size $K \times 1$, the standard and deformable strip convolutions are formulated as

$$\begin{aligned} F_v(p_0) &= \sum_{p_n \in \mathcal{R}_v} \mathbf{W}_v(p_n) F(p_0 + p_n), \\ F_{v_def}(p_0) &= \sum_{p_n \in \mathcal{R}_v} \mathbf{W}_v(p_n) F(p_0 + p_n + \Delta p_n), \end{aligned} \tag{11}$$

where Δp_n denotes the dynamically predicted sampling offset. Finally, a 1×1 convolution is employed to generate a geometry attention map, which adaptively reweights the input features for EPI geometry refinement.

4 Experiments

4.1 Datasets and Implementation Details

Following previous LFSR works Wang et al. (2022b); Hu et al. (2025); Liang et al. (2026), we conduct experiments on five public LF datasets: EPFL Rerabek & Ebrahimi (2016), HCIold Wanner et al. (2013), HCInew Honauer et al. (2017), INRIA Le Pendu et al. (2018), and STFgantry Vaish & Adams (2008). The central 5×5 views are extracted from original 9×9 LF images and cropped into 32×32 and 64×64 patches. Bicubic downsampling with scale factors of 0.5 and 0.25 is applied to generate LR inputs for $2\times$ and $4\times$ SR, respectively. Random rotations and horizontal/vertical flips are used for augmentation. We optimize the network with the L_1 loss using Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$). The learning rate is initialized to 4×10^{-4} and halved every 15 epochs for 60 epochs. The batch size is 8, and all experiments are conducted on two NVIDIA RTX 4090 GPUs.

4.2 Comparison to State-of-the-art Methods

We compare the proposed method with several representative LFSR methods, including LF-InterNet Wang et al. (2020), DPT Wang et al. (2022a), LF-IINet Liu et al. (2021), DistgSSR Wang et al.

Table 1: Quantitative comparison of different LFSR methods in terms of PSNR/SSIM. The best and second-best results are highlighted in **bold** and underline, respectively.

Methods	2× SR					4× SR				
	EPFL	HCInew	HCIold	INRIA	STFgantry	EPFL	HCInew	HCIold	INRIA	STFgantry
Bicubic	29.74/.9376	31.89/.9356	37.69/.9785	31.33/.9577	31.06/.9498	25.26/.8324	27.72/.8517	32.58/.9344	26.95/.8867	26.09/.8452
LF-InterNet	34.11/.9760	37.17/.9763	44.57/.9946	35.83/.9843	38.44/.9909	28.81/.9162	30.96/.9161	37.15/.9716	30.78/.9491	30.37/.9409
DPT	34.49/.9758	37.36/.9771	44.30/.9943	36.41/.9843	39.43/.9926	28.94/.9170	31.20/.9188	37.41/.9721	30.96/.9503	31.15/.9488
LF-IINet	34.73/.9773	37.77/.9790	44.85/.9948	36.57/.9853	39.89/.9936	29.04/.9188	31.33/.9208	37.62/.9734	31.03/.9515	31.26/.9502
DistgSSR	34.81/.9787	37.96/.9796	44.94/.9949	36.59/.9859	40.40/.9942	28.99/.9195	31.38/.9217	37.56/.9732	30.99/.9519	31.65/.9535
HLFSR-SSR	35.31/.9800	38.32/.9807	44.98/.9950	37.06/.9867	40.85/.9947	29.20/.9222	31.57/.9238	37.78/.9742	31.24/.9534	31.64/.9537
LF-DET	35.26/.9797	38.31/.9807	44.99/.9950	36.95/.9864	41.76/.9955	29.47/.9230	31.56/.9235	37.84/.9744	31.39/.9573	32.14/.9573
MLFSR	35.22/.9801	38.14/.9803	44.90/.9950	36.92/.9865	40.98/.9949	29.28/.9218	31.56/.9235	37.83/.9745	31.24/.9531	32.03/.9567
TriFormer	35.93/.9826	38.41/.9811	45.15/.9951	37.58/.9880	41.51/.9953	29.77/.9246	31.79/.9258	38.07/.9752	31.72/.9546	32.50/.9593
LFMamba	35.76/.9824	38.37/.9801	44.99/.9950	37.06/.9876	40.95/.9948	29.84/.9256	31.70/.9249	37.91/.9748	31.81/.9551	31.85/.9553
OHT-B	35.66/.9829	38.50/.9814	45.04/.9950	37.10/.9877	41.40/.9953	29.45/.9254	31.79/.9263	37.90/.9745	31.36/.9550	32.24/.9590
M2MT-Net	35.64/.9815	38.43/.9810	**45.38/.9953**	37.22/.9870	40.99/.9949	29.85/.9284	31.76/.9264	37.98/.9749	31.77/.9563	32.21/.9584
HI-LLF	35.42/.9806	38.22/.9805	45.08/.9950	37.17/.9867	41.03/.9948	29.80/.9256	31.71/.9253	38.00/.9754	31.80/.9548	32.29/.9580
EPIT-SSR	34.83/.9775	38.23/.9810	45.08/.9949	36.67/.9853	**42.17/.9957**	29.34/.9197	31.51/.9231	37.68/.9737	31.37/.9526	32.18/.9571
LF-F^{3}Net	34.96/.9780	38.13/.9800	45.07/.9950	36.72/.9850	40.58/.9940	29.30/.9220	31.58/.9240	37.73/.9740	31.23/.9540	31.63/.9540
SimSSR	35.02/.9799	38.12/.9802	44.92/.9949	36.64/.9853	40.68/.9945	29.23/.9204	31.42/.9219	37.61/.9736	31.16/.9521	31.59/.9530
LF-DUTS	**36.49/.9839**	**38.60/.9817**	44.95/.9950	**38.07/.9882**	41.33/.9953	**30.40/.9316**	**32.02/.9291**	**38.21/.9763**	**32.30/.9581**	**32.59/.9605**

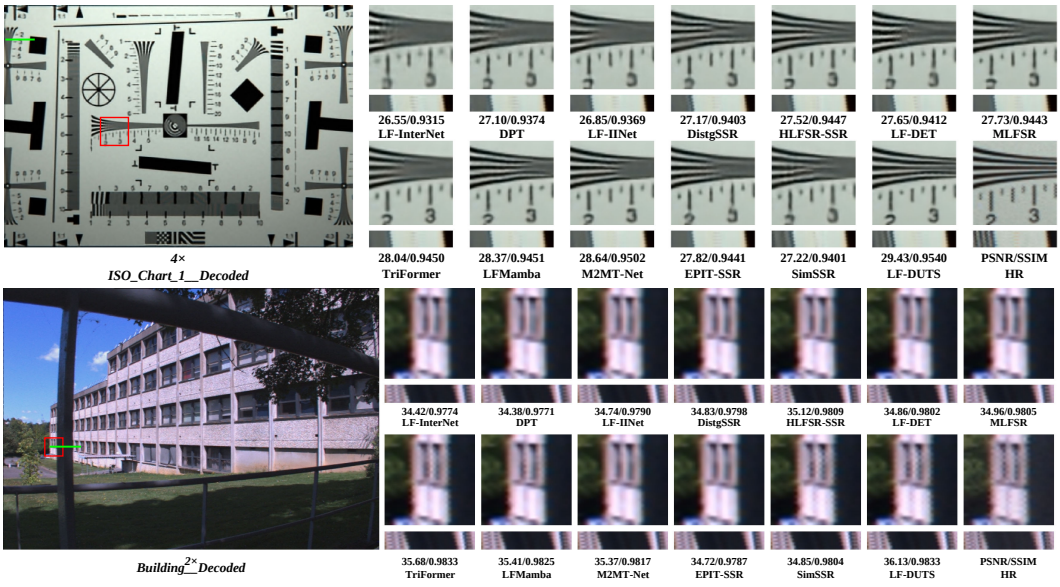


Figure 5: Visual results obtained by different methods for 2× and 4× LFSR.

(2022b), HLFSR-SSR Van Duong et al. (2023), LF-DET Cong et al. (2023), MLFSR Gao et al. (2024), TriFormer Wang et al. (2024), LFMamba Lu et al. (2024), OHT-B Xiao et al. (2025), M2MT-Net Hu et al. (2025), HI-LLF Li et al. (2025a), EPIT-SSR Liang et al. (2026), LF-P^{3}Net Yang et al. (2026), and SimSSR Zhang et al. (2026). Following the settings in Wang et al. (2022b), we adopt PSNR and SSIM calculated on the Y channel of the images as quantitative evaluation metrics. More experimental results and detailed analyses are shown in the Appendix A.3-A.8.

Quantitative Results. As presented in Table 1, our method achieves superior or competitive performance over state-of-the-art methods across all datasets under both 2× and 4× SR scales. Notably, under the 4× SR setting, our approach consistently achieves the best PSNR and SSIM results on all five benchmark datasets, outperforming the second-best method by up to 0.55 dB in PSNR on EPFL. These results demonstrate the effectiveness of our model in preserving 4D spatial-angular structures and reconstructing fine details.

Qualitative Results. Fig. 12 presents the visual comparison results for $2\times$ and $4\times$ SR. As can be seen, our method more effectively reconstructs complex geometric structures and fine details, while competing methods tend to produce blurred or structurally degraded results in challenging regions. By jointly exploiting the 4D physical degradation model and tri-domain structural learning, our method preserves structural information and restores high-frequency details, leading to superior reconstruction quality.

Disparity Consistency. To evaluate disparity consistency, we compare reconstructed EPI slices (Fig. 12) and perform depth estimation on the *herbs* scene using the SPO method Zhang et al. (2016) (Fig. 6). LF-DUTS reconstructs straighter EPIs with sharper edges and more continuous structures, while its estimated depth map is closer to the HR ground truth and achieves the lowest MSE among compared methods. These results demonstrate the effectiveness of LF-DUTS in preserving disparity consistency and recovering accurate geometric structures.

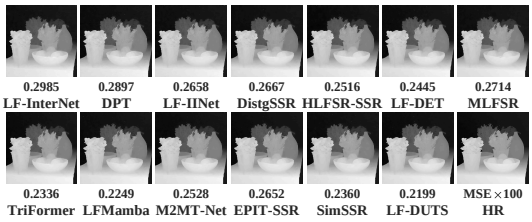


Figure 6: Depth estimation on $4\times$ LFSR images.

4.3 Ablation Study

In this section, we conduct ablation studies on $4\times$ SR to validate the effectiveness of LF-DUTS. We evaluate the contribution of each component through designed variants.

Effectiveness of the EGRB. To validate the effectiveness of EGRB, we construct several variants, including removing EGRB (w/o-EGRB), replacing deformable strip convolution with square convolution (SqConv), replacing it with standard strip convolution (w/o-Deformable), and adopting deformable strip convolutions with different kernel sizes. As shown in Table 2, removing EGRB decreases the PSNR to 32.83 dB, demonstrating its importance for LF reconstruction. Compared with SqConv and StripConv, deformable strip convolution better captures EPI geometric information by adaptively modeling disparity-induced structures. Different kernel sizes achieve different reconstruction performance, with the 9×1 configuration obtaining the best result (33.11 dB/0.9511), indicating that an appropriate strip receptive field is crucial for EPI geometry modeling. Furthermore, as shown in Figs. 7–9, LF-DUTS reconstructs sharper textures and structural details, produces straighter and more continuous EPIs, and extracts more discriminative EPI features, further demonstrating the effectiveness of EGRB.

Table 2: Comparison of different variants of EGRB.

Methods	EGRB	Deformable	PSNR	SSIM
w/o-EGRB	✗	✗	32.83	0.9500
SqConv	Square Conv	✗	32.89	0.9503
w/o-Deformable	9×1 Strip Conv	✗	32.92	0.9506
7×1 StripConv	7×1 Strip Conv	✓	32.96	0.9499
11×1 StripConv	11×1 Strip Conv	✓	32.94	0.9507
13×1 StripConv	13×1 Strip Conv	✓	32.97	0.9505
LF-DUTS (ours)	9×1 Strip Conv	✓	**33.11**	**0.9511**

Effectiveness of the CSFM. To validate the effectiveness of CSFM, we construct a network variant without CSFM (w/o-CSFM). As shown in Table 3, removing CSFM decreases the PSNR and SSIM to 32.95 dB and 0.9508, respectively, demonstrating its contribution to LF reconstruction. Furthermore, as shown in Fig. 9, CSFM produces more concentrated feature responses with clearer structural boundaries, enabling more discriminative high-frequency representations and improving reconstruction performance.

Parameter Analysis. To determine the optimal network architecture, we investigate the effects of two key hyperparameters, N and T, in Table 4. Specifically, N denotes the number of TSRGs, while T represents the number of unfolding iterations. First, with T fixed at 1, the reconstruction performance consistently improves as N increases from 6 to 8. When N is further increased to

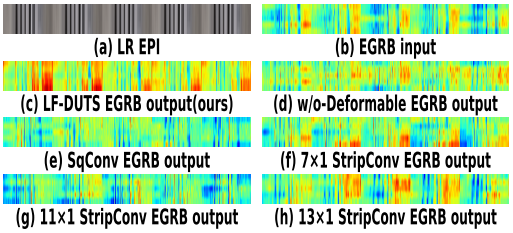


Figure 7: EPI feature visualization of EGRB.

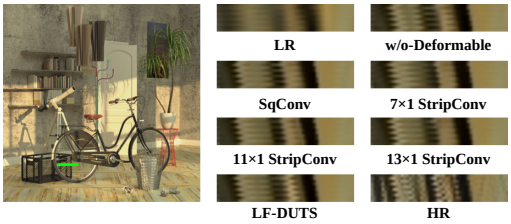


Figure 8: EPI visualization for the effectiveness of EGRB.

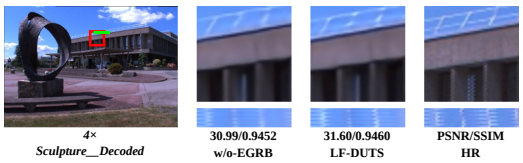


Figure 9: Visual comparison of w/o-EGRB and LF-DUTS.

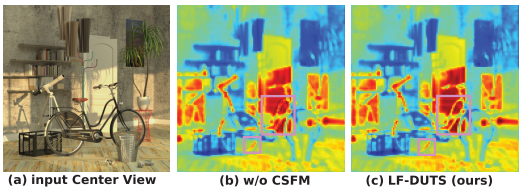


Figure 10: Feature visualization for the CSFM.

Table 3: Comparison of different variants of CSFM.

Methods	CSFM	PSNR	SSIM
w/o-CSFM	✗	32.95	0.9508
LF-DUTS (ours)	✓	**33.11**	**0.9511**

9, however, the performance gain becomes marginal despite the additional parameters and computational costs. Considering both reconstruction performance and computational efficiency, we set $N = 8$. Subsequently, with N fixed at 8, we evaluate the effect of T. Increasing T from 1 to 2 significantly improves the PSNR to 33.11 dB with only a modest increase in computational cost. Further increasing T to 3 yields only a slight improvement in SSIM but leads to a decrease in PSNR, indicating that additional unfolding iterations do not provide further performance gains. Therefore, the default configuration is set to $N = 8$ and $T = 2$.

Table 4: Ablation study of hyper-parameters N and T.

Parameter Values	#Param.	FLOPs	PSNR	SSIM
N=6 / T=1	3.63	53.77	32.76	0.9481
N=7 / T=1	4.21	62.25	32.80	0.9485
N=8 / T=1	4.80	70.72	32.85	0.9490
N=9 / T=1	5.39	79.20	32.83	0.9492
N=8 / T=2 (ours)	4.81	141.45	**33.11**	0.9511
N=8 / T=3	4.81	212.17	33.03	**0.9513**

5 Conclusion

In this paper, we propose LF-DUTS, an interpretable deep unfolding framework for light field image super-resolution. By unfolding a 4D degradation model into a learnable optimization process and incorporating tri-domain structural learning, the proposed framework effectively bridges model-driven optimization and deep feature learning. Extensive experiments demonstrate that LF-DUTS consistently achieves superior reconstruction performance, recovering sharper textures, more accurate EPI geometric structures, and better disparity consistency compared with existing methods. These results validate the effectiveness of LF-DUTS in achieving interpretable and high-quality LFSR.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing, grammatical correction, and improving the clarity and readability of parts of the manuscript. Generative AI tools were not used to generate experimental data, fabricate experimental results, or independently make scientific claims or conclusions.

All AI-assisted text was carefully reviewed, revised, and verified by the authors before being included in the manuscript. References, technical claims, experimental settings, and reported quantitative results were independently checked by the authors against the corresponding original sources and experimental records.

Generative AI was used only as an auxiliary tool and did not replace the authors' scientific reasoning, experimental design, implementation, or interpretation of the results. We take full responsibility for the final content of this work, including all text, claims, and other artifacts produced with the aid of generative AI.

REFERENCES

Martin Alain and Aljosa Smolic. Light field super-resolution via lfbm5d sparse coding. In *2018 25th IEEE international conference on image processing (ICIP)*, pp. 2501–2505. IEEE, 2018.

Martin Alain and Aljosa Smolic. A spatio-angular filter for high quality sparse light field refocusing. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pp. 1–6. IEEE, 2021.

Tom E Bishop, Sara Zanetti, and Paolo Favaro. Light field superresolution. In *2009 IEEE International Conference on Computational Photography (ICCP)*, pp. 1–9. IEEE, 2009.

Ruixuan Cong, Hao Sheng, Da Yang, Zhenglong Cui, and Rongshan Chen. Exploiting spatial and angular correlations with deep efficient transformers for light field image super-resolution. *IEEE Transactions on Multimedia*, 26:1421–1435, 2023.

Weisheng Dong, Peiyao Wang, Wotao Yin, Guangming Shi, Fangfang Wu, and Xiaotong Lu. Denoising prior driven deep neural network for image restoration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(10):2305–2318, 2018.

Yuan Fang, Yipeng Liu, Zhen Long, Chong-Yung Chi, and Ce Zhu. Content-adaptive unfolding wavelet transformer for hyperspectral image super-resolution. *IEEE Transactions on Image Processing*, 2025.

Reuben A Farrugia, Christian Galea, and Christine Guillemot. Super resolution of light field images using linear subspace projection of patch-volumes. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):1058–1071, 2017.

Ruisheng Gao, Zeyu Xiao, and Zhiwei Xiong. Mamba-based light field super-resolution with efficient subspace scanning. In *Proceedings of the Asian Conference on Computer Vision*, pp. 531–547, 2024.

Jinjin Gu and Chao Dong. Interpreting super-resolution networks with local attribution maps. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9199–9208, 2021.

Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part III 13*, pp. 19–34. Springer, 2017.

Qian Hu, Jiayi Ma, Yuan Gao, Junjun Jiang, and Yixuan Yuan. Maun: Memory-augmented deep unfolding network for hyperspectral image reconstruction. *IEEE/CAA Journal of Automatica Sinica*, 11(5):1139–1150, 2024.

Zeke Zexi Hu, Xiaoming Chen, Vera Yuk Ying Chung, and Yiran Shen. Beyond subspace isolation: Many-to-many transformer for light field image super-resolution. *IEEE Transactions on Multimedia*, 27:1334–1348, 2025.

Yan Huang, Shang Li, Liang Wang, Tieniu Tan, et al. Unfolding the alternating optimization for blind super resolution. *Advances in Neural Information Processing Systems*, 33:5632–5643, 2020.

Mikael Le Pendu, Xiaoran Jiang, and Christine Guillemot. Light field inpainting propagation via low rank matrix completion. *IEEE Transactions on Image Processing*, 27(4):1981–1993, 2018.

Meng Li, Bo Ma, and Shunzhou Wang. Hierarchical spatial–angular integration for lightweight light field image super-resolution. *Knowledge-Based Systems*, 315:113240, 2025a.

Yong-Chao Li, Guo-Yu Wang, Shao-Yu Liu, Sen Zhang, and Chang Liu. Cross-domain adaptive disparity-aware light field images spatial super-resolution method based on contrastive learning. *Expert Systems with Applications*, pp. 130762, 2025b.

Yuansheng Li, Yunhao Zou, Linwei Chen, and Ying Fu. Physical degradation model-guided interferometric hyperspectral reconstruction with unfolding transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13815–13825, 2025c.

Zhengyu Liang, Yingqian Wang, Longguang Wang, Jungang Yang, Yulan Guo, Li Liu, Shilin Zhou, et al. Diving into epipolar transformers for light field super-resolution and disparity estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Yigeng Liao, Le Song, Gaoyan Zhang, and Fengzhou Fang. Multi-level disparity-guided transformers for light field spatial super-resolution. *Pattern Recognition*, pp. 111803, 2025.

Denghong Liu, Jie Li, Qiangqiang Yuan, Li Zheng, Jiang He, Shuheng Zhao, and Yi Xiao. An efficient unfolding network with disentangled spatial-spectral representation for hyperspectral image super-resolution. *Information Fusion*, 94:92–111, 2023.

Gaosheng Liu, Huanjing Yue, Jiamin Wu, and Jingyu Yang. Intra-inter view interaction network for light field image super-resolution. *IEEE Transactions on Multimedia*, 25:256–266, 2021.

Yao Lu, Shunzhou Wang, Ziqi Wang, Peiqi Xia, Tianfei Zhou, et al. Lfmamba: Light field image super-resolution with state space model, 2024. URL https://arxiv.org/abs/2406.12463.

Iman Marivani, Evaggelia Tsiligianni, Bruno Cornelis, and Nikos Deligiannis. Multimodal deep unfolding for guided image super-resolution. *IEEE Transactions on Image Processing*, 29:8443–8456, 2020.

Nan Meng, Hayden K-H So, Xing Sun, and Edmund Y Lam. High-dimensional dense residual convolutional neural network for light field reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):873–886, 2019.

Kaushik Mitra and Ashok Veeraraghavan. Light field denoising, light field superresolution and stereo camera based refocussing using a gmm light field patch prior. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pp. 22–28. IEEE, 2012.

Zhangkai Ni, Yang Zhang, Wenhan Yang, Hanli Wang, Shiqi Wang, and Sam Kwong. Structural similarity-inspired unfolding for lightweight image super-resolution. *IEEE Transactions on Image Processing*, 34:3861–3872, 2025.

Qian Ning, Weisheng Dong, Guangming Shi, Leida Li, and Xin Li. Accurate and lightweight image super-resolution with model-guided deep unfolding network. *IEEE Journal of Selected Topics in Signal Processing*, 15(2):240–252, 2021.

Martin Rerabek and Touradj Ebrahimi. New light field image dataset. In *8th international conference on quality of multimedia experience (QoMEX)*, 2016.

Mattia Rossi and Pascal Frossard. Geometry-consistent light field super-resolution via graph-based regularization. *IEEE Transactions on Image Processing*, 27(9):4207–4218, 2018.

Zhengxi Song, Hao Zhu, Qi Wu, Xue Wang, Hongdong Li, and Qing Wang. Accurate 3d reconstruction from circular light field using cnn-lstm. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2020.

Yao Tang, Jie Li, Linwei Yue, Xinxin Liu, Yajie Li, Yi Xiao, and Qiangqiang Yuan. A cnn-transformer embedded unfolding network for hyperspectral image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

V. Vaish and A. Adams. The (new) stanford light field archive. In *Computer Graphics Laboratory, Stanford University*, volume 6, 2008.

Vinh Van Duong, Thuc Nguyen Huu, Jonghoon Yim, and Byeungwoo Jeon. Light field image super-resolution network via joint spatial-angular and epipolar information. *IEEE Transactions on Computational Imaging*, 9:350–366, 2023.

Shunzhou Wang, Tianfei Zhou, Yao Lu, and Huijun Di. Detail-preserving transformer for light field image super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2522–2530, 2022a.

Tun Wang, Hao Sheng, Ruixuan Cong, Da Yang, Zhenglong Cui, Guanqun Su, and Yi Zhang. Difference-guided full-view volume for light field depth estimation. *Expert Systems with Applications*, pp. 132256, 2026.

Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. Spatial-angular interaction for light field image super-resolution. In *European Conference on Computer Vision*, pp. 290–308. Springer, 2020.

Yingqian Wang, Longguang Wang, Gaochang Wu, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. Disentangling light fields for super-resolution and disparity estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):425–443, 2022b.

Ziqi Wang, Yao Lu, Shunzhou Wang, Wang Xia, Peiqi Xia, and Wenjing Wang. Trident transformer for light field image super-resolution. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2024.

Sven Wanner and Bastian Goldluecke. Variational light field analysis for disparity estimation and super-resolution. *IEEE transactions on pattern analysis and machine intelligence*, 36(3):606–619, 2013.

Sven Wanner, Stephan Meister, and Bastian Goldluecke. Datasets and benchmarks for densely sampled 4d light fields. In *VMV*, volume 13, pp. 225–226, 2013.

Zeqiang Wei, Kai Jin, Zeyi Hou, Kuan Song, and Xiuzhuang Zhou. l^2fmamba: Lightweight light field image super-resolution with state space model. *IEEE Transactions on Computational Imaging*, 2025.

Gaochang Wu, Yebin Liu, Lu Fang, and Tianyou Chai. Revisiting light field rendering with deep anti-aliasing neural network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5430–5444, 2021.

Zeyu Xiao, Zhuoyuan Li, and Wei Jia. Occlusion-embedded hybrid transformer for light field super-resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 8700–8708, 2025.

Yulei Yang, Zongju Peng, Huabo Zhang, Fen Chen, and Qianliang Zhang. Lf-f3net: Frequency-guided feature fusion network for light field image super-resolution. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 2026.

Henry Wing Fung Yeung, Junhui Hou, Xiaoming Chen, Jie Chen, Zhibo Chen, and Yuk Ying Chung. Light field spatial super-resolution using deep efficient spatial-angular separable convolution. *IEEE Transactions on Image Processing*, 28(5):2319–2330, 2018.

Youngjin Yoon, Hae-Gon Jeon, Donggeun Yoo, Joon-Young Lee, and In So Kweon. Light-field image super-resolution using convolutional neural network. *IEEE Signal Processing Letters*, 24 (6):848–852, 2017.

Chenxiao Zhang, Xin Deng, Jingyi Xu, Yongxuan Dou, and Mai Xu. Deep semi-smooth newton-driven unfolding network for multi-modal image super-resolution. *IEEE Transactions on Image Processing*, 34:7123–7137, 2025.

Fengyuan Zhang, Yingqian Wang, Xueying Wang, Zhengyu Liang, Longguang Wang, Lvli Tian, and Jungang Yang. Revisiting subspace disentangling for light field spatial super-resolution. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Jialong Zhang, Lijun Zhao, Jinjing Zhang, Ke Wang, and Anhong Wang. Explainable unfolding network for joint edge-preserving depth map super-resolution. In *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 888–893. IEEE, 2023.

Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3217–3226, 2020.

Shuo Zhang, Hao Sheng, Chao Li, Jun Zhang, and Zhang Xiong. Robust depth estimation for light field via spinning parallelogram operator. *Computer Vision and Image Understanding*, 145: 148–159, 2016.

Lijun Zhao, Ke Wang, Jinjing Zhang, Jie Zhao, and Anhong Wang. A joint model-driven unfolding network for degraded low-quality color-depth images enhancement. In *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 625–629. IEEE, 2023.

Hongyi Zheng, Hongwei Yong, and Lei Zhang. Unfolded deep kernel estimation for blind image super-resolution. In *European Conference on Computer Vision*, pp. 502–518. Springer, 2022.

A Appendix

A.1 Z-subproblem Optimization

A light field can be represented as an array of $U \times V$ sub-aperture images. Since the data fidelity term in the degradation model is independently defined for each angular view, the data fidelity subproblem can be decomposed into $U \times V$ view-wise subproblems. This decomposition reduces computational complexity and facilitates parallel implementation. Meanwhile, the angular correlations among different views are exploited and refined in the prior update stage through the proposed tri-domain structural refinement module. Specifically, for a given view (u, v), the corresponding proximal mapping is formulated as:

$$z_{u,v}^{t} = \arg\min_{z_{u,v}} \left\| y_{u,v} - H z_{u,v}^{t-1} \right\|_2^2 + \mu^t \sigma^2 \left\| z_{u,v}^{t-1} - D x_{u,v}^{t-1} \right\|_2^2. \tag{12}$$

Although subproblem (12) has a closed-form solution, the involved large-scale matrix inversion leads to considerable computational overhead. Therefore, following Dong et al. (2018), we adopt a single-step gradient descent scheme to obtain an approximate solution. Specifically, taking $z_{u,v}^{t-1}$ as the initialization, we update $z_{u,v}$ through one gradient descent step:

$$z_{u,v}^{t} = z_{u,v}^{t-1} - \eta \nabla f(z_{u,v}^{t-1}), \tag{13}$$

where η denotes the step size, and $f(\cdot)$ represents the objective function of the optimization problem, which is formulated as:

$$f(z_{u,v}^{t}) = \|y_{u,v} - H z_{u,v}^{t}\|_2^2 + \mu^{t+1} \sigma^2 \|z_{u,v}^{t} - D x_{u,v}^{t}\|_2^2. \tag{14}$$

By computing the gradient of $f(z_{u,v}^{t})$ and substituting it into Eq. (13), the final update of $z_{u,v}$ can be obtained as:

$$
\begin{aligned}
z_{u,v}^{t} &= z_{u,v}^{t-1} - \eta_t \left[H^T (H z_{u,v}^{t-1} - y_{u,v}) + \alpha_t (z_{u,v}^{t-1} - D x_{u,v}^{t-1}) \right] \\
&= \bar{H} z_{u,v}^{t-1} + \eta_t H^T y_{u,v} + \eta_t \alpha_t D x_{u,v}^{t-1},
\end{aligned} \tag{15}
$$

where $\alpha_t = \mu_t \sigma^2$, and $\bar{H} = (1 - \eta_t \alpha_t) I - \eta_t H^T H$. Finally, all updated SAIs are aggregated according to their inherent spatial-angular coordinates, naturally reconstructing the complete global light field:

$$
Z^t = \left\langle z_{u,v}^{t} \right\rangle_{u=1,v=1}^{U,V}. \tag{16}
$$

A.2 Hyper-Parameter Network

Given the high-dimensional complexity of light field data, the internal parameters in the unfolding framework, such as α, β, and η, are difficult to initialize manually. Therefore, we introduce a hyperparameter network (HPN) to adaptively control the balance between the data fidelity and prior terms. As shown in Fig. 11, HPN employs a dual-layer fully connected architecture to explicitly map degradation conditions, including noise level and scaling factor, into stage-specific parameter sequences:

$$
[\alpha, \beta, \eta] = \text{HPN}(\sigma, s), \tag{17}
$$

where $\alpha = [\alpha_1, \ldots, \alpha_t, \ldots, \alpha_T]$, $\beta = [\beta_1, \ldots, \beta_t, \ldots, \beta_T]$, and $\eta = [\eta_1, \ldots, \eta_t, \ldots, \eta_T]$.

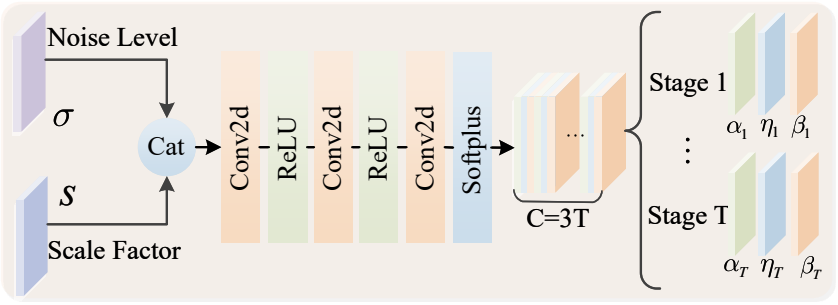


Figure 11: Implementation details of HPN.

A.3 Qualitative Results

To further demonstrate the visual reconstruction quality of LF-DUTS, we provide qualitative comparisons on the *origami* and *Palais_du_Luxembourg* scenes under $2\times$ and $4\times$ SR settings, as shown in Fig. 12. In the *Palais_du_Luxembourg* scene ($2\times$), LF-DUTS effectively recovers fine-grained details of distant shutters and preserves sharp boundaries in challenging regions, such as the junction between the foreground fence and wall, while suppressing reconstruction artifacts and blurring. In the *origami* scene ($4\times$), LF-DUTS maintains the continuity of geometric lines and intersecting edges of the folded paper model, avoiding edge discontinuities and texture degradation produced by competing methods. These results demonstrate the capability of LF-DUTS in preserving structural details under challenging reconstruction conditions.

Fig. 14 visualizes the SAI-wise PSNR values of two representative scenes, i.e., *ISO_Chart_1_Decoded* and *Sculpture_Decoded*. The results demonstrate that LF-DUTS achieves consistently higher PSNR values across different SAIs, indicating its superior reconstruction performance across angular views. Notably, in both scenes, the lowest PSNR values achieved by LF-DUTS remain higher than the highest PSNR values of competing methods for most SAIs, demonstrating its robustness in preserving angular consistency. These results further verify the effectiveness of LF-DUTS in exploiting 4D spatial-angular information for high-quality light field reconstruction.

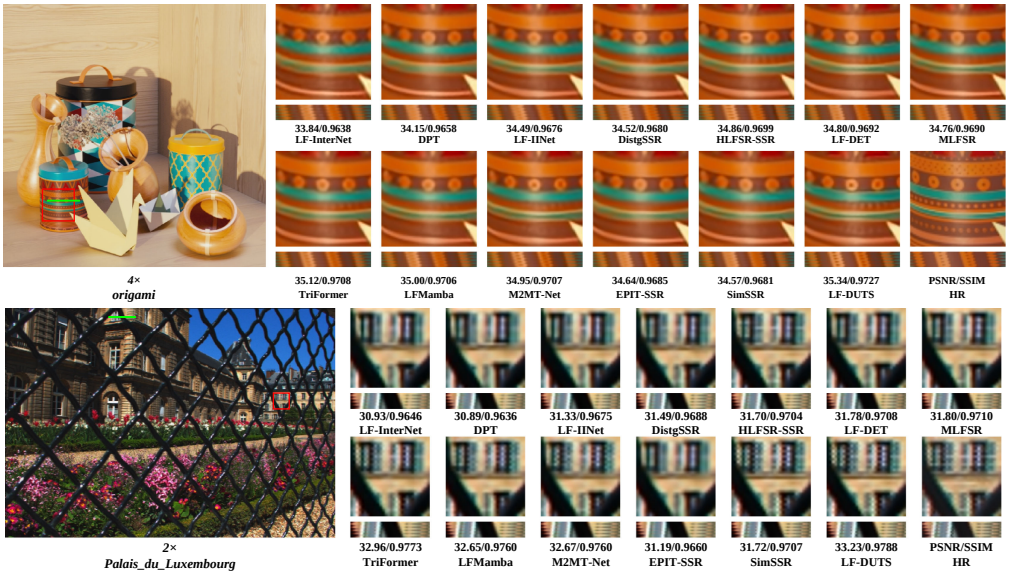


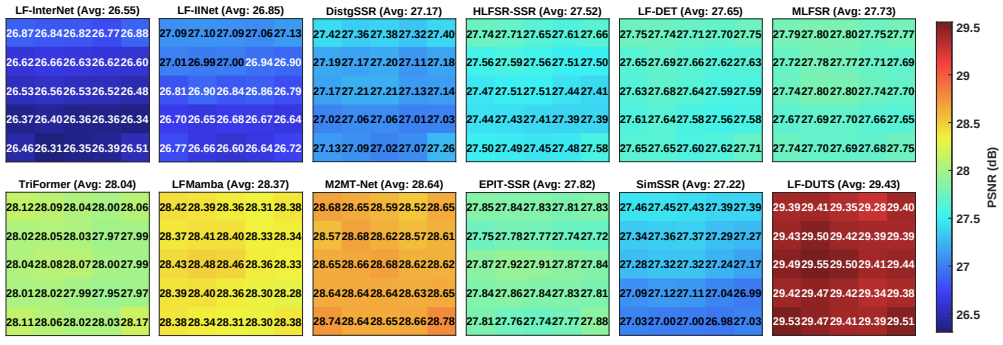
Figure 12: Visual results obtained by different methods for $2\times$ and $4\times$ LFSR.



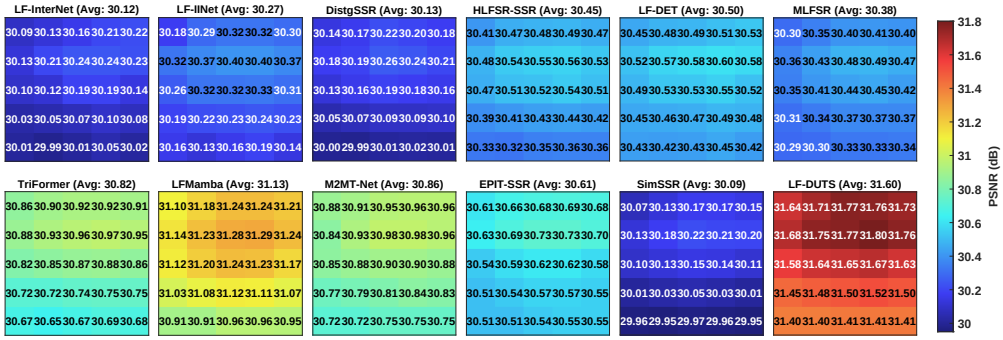
Figure 13: Visual results obtained by different methods on the $4\times$ SR real LF from the STFlytro dataset.

A.4 Results on real LF images

To evaluate the reconstruction performance of different SR methods on real-world scenes, we apply these methods to real LF images from the STFlytro dataset. Fig. 13 presents the visual reconstruction results of several representative scenes. Through qualitative comparison, LF-DUTS achieves superior visual reconstruction quality in both scenes. Specifically, in the *Stanford_Lytro_cars_34_esl* scene, LF-DUTS effectively recovers high-frequency character details. For the "CAMPUS" sign highlighted by the orange arrow, the proposed method reconstructs clear and sharp contours of the letters "P" and "U", while reducing character blurring, merging, and high-frequency artifacts commonly observed in other methods. In the more complex *Stanford_Lytro_people_16_esl* scene, LF-DUTS accurately restores key facial structures, such as eye and nose contours, while maintaining natural texture transitions and shading variations. Furthermore, the EPI slices shown at the bottom of both scenes demonstrate that LF-DUTS recovers more continuous and smoother disparity structures, highlighting its advantages in spatial-angular feature fusion and angular consistency preservation.



(a) ISO_Chart_1_Decoded



(b) Sculpture_Decoded

Figure 14: Visualization of SAI-wise PSNR to demonstrate the distribution of $4\times$ LFSR performance.

Overall, LF-DUTS maintains stable reconstruction performance across diverse real-world LF scenes and exhibits clear advantages in detail recovery, texture preservation, and visual naturalness, further validating its effectiveness and robustness for real-world LF image super-resolution.



Figure 15: Depth estimation on $4\times$ LFSR images.

A.5 DISPARITY CONSISTENCY

To evaluate the geometric consistency preservation of different methods, we perform depth estimation on the reconstructed results of the *buddha* and *monasRoom* scenes using the SPO method, and calculate the MSE $\times 100$ metric (lower values indicate smaller errors). As shown in Fig. 15, LF-DUTS achieves the best depth reconstruction performance on both scenes. In the *monasRoom* scene, LF-DUTS obtains the lowest MSE of 0.2173, indicating more accurate depth recovery than other methods. Similarly, in the *buddha* scene, LF-DUTS achieves an MSE of 0.0505, indicating better disparity consistency preservation. These results demonstrate that LF-DUTS effectively exploits the spatial-angular structures of light fields, enabling more accurate geometric modeling and depth reconstruction.

A.6 EFFICIENCY

Table 5 presents the comparison of different methods in terms of parameters, computational complexity, and reconstruction performance. It can be observed that LF-DUTS achieves the best reconstruction performance for both $2\times$ and $4\times$ SR, demonstrating its effectiveness in reconstructing high-quality LF images. Specifically, LF-DUTS obtains PSNR/SSIM values of 39.89 dB/0.9888 and 33.11 dB/0.9511 for $2\times$ and $4\times$ SR, respectively, with 4.78M and 4.81M parameters. Compared with the second-best method TriFormer under the $2\times$ SR setting, LF-DUTS improves the PSNR by 0.18 dB and SSIM by 0.0004. Under the $4\times$ SR setting, LF-DUTS surpasses TriFormer by 0.34 dB in PSNR and 0.0032 in SSIM. Although LF-DUTS requires slightly higher computational costs due to the incorporation of tri-domain structural learning, it achieves substantial performance improvements with a comparable model size. These results demonstrate that LF-DUTS achieves a favorable trade-off between reconstruction accuracy and model complexity.

Table 5: Comparison in terms of parameters, FLOPs, and average PSNR/SSIM values for $2\times$ and $4\times$ SR.

Method	#Param.(M)		FLOPs (G)		PSNR (dB)		SSIM	
	$2\times$	$4\times$	$2\times$	$4\times$	$2\times$	$4\times$	$2\times$	$4\times$
LF-InterNet	5.04	5.48	42.33	53.67	38.02	31.61	.9844	.9388
DPT	3.73	3.78	57.32	58.52	38.40	31.93	.9848	.9414
LF-IINet	4.84	4.89	56.08	57.35	38.76	32.06	.9862	.9429
DistgSSR	3.53	3.58	63.98	65.26	38.94	32.12	.9867	.9440
HLFSR-SSR	13.71	13.87	156.15	182.93	39.30	32.29	.9874	.9455
LF-DET	1.59	1.69	48.50	51.20	39.46	32.48	.9875	.9463
MLFSR	1.36	1.41	45.71	47.15	39.23	32.39	.9873	.9459
TriFormer	1.18	1.31	29.89	33.37	<u>39.71</u>	<u>32.77</u>	.9884	.9479
LFMamba	2.15	2.30	62.95	66.90	39.42	32.62	.9881	.9471
OHT-B	2.26	2.31	101.18	102.61	39.54	32.55	<u>.9885</u>	.9480
M2MT-Net	4.36	3.91	32.36	29.82	39.53	32.71	.9879	<u>.9489</u>
HI-LLF	0.80	0.85	25.20	25.80	39.44	32.72	.9875	.9478
EPIT-SSR	1.42	1.47	52.87	54.30	39.39	32.42	.9869	.9452
LF-F^{3}Net	6.43	6.49	88.04	89.79	39.09	32.29	.9870	.9450
SimSSR	3.47	3.52	121.20	112.50	39.08	32.21	.9872	.9442
LF-DUTS	4.78	4.81	139.77	141.45	**39.89**	**33.11**	**.9888**	**.9511**

A.7 LAM ANALYSIS

To further analyze the feature utilization mechanism of different models, we employ the Local Attribution Map (LAM) Gu & Dong (2021) for visualization analysis. LAM identifies input pixels that contribute significantly to the reconstruction of target regions, thereby revealing the feature dependencies learned by the model. Fig. 16 presents the LAM visualizations and corresponding DI values of different methods. It can be observed that LF-DUTS highlights more informative pixels both within individual views and across different views. Meanwhile, it achieves the highest DI values across all scenes and obtains the best reconstruction performance, with PSNR/SSIM values of 28.20 dB/0.9400 and 29.48 dB/0.9097 for the two evaluated scenes, respectively. Further analysis demonstrates that LF-DUTS effectively exploits repetitive textures and cross-view complementary

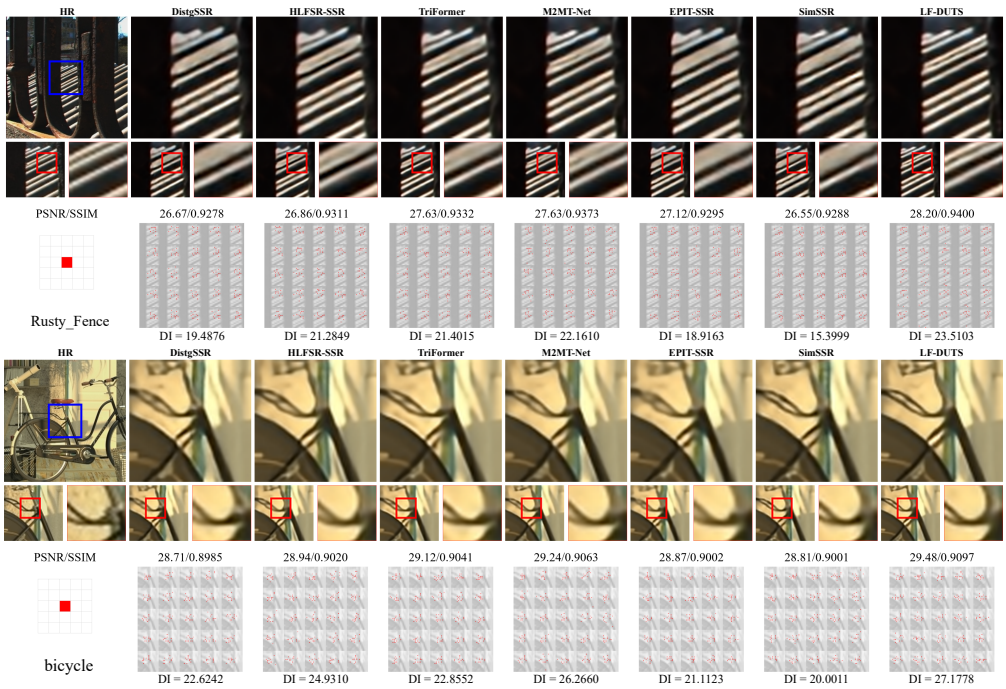


Figure 16: Visualization of LAM for analyzing feature utilization of different methods.

information in light fields, such as the fence structures in *Rusty_Fence* and continuous lines in *bicycle*, by establishing long-range pixel dependencies to facilitate high-frequency detail recovery. In contrast, other methods are limited by local receptive fields or insufficient long-range dependency modeling, making it difficult to fully exploit these informative features.

A.8 Effectiveness of the PCFN

To validate the effectiveness of PCFN in ACRB, SARB, and EGRB, we construct a variant w/o-PCFN, where PCFN is replaced with MLP. The average performance over five datasets is reported. As shown in Table 6, LF-DUTS with PCFN achieves a PSNR of 33.11 dB and an SSIM of 0.9511, achieving improvements of 0.17 dB and 0.0007, respectively, compared with the MLP-based variant. These results demonstrate that PCFN can better exploit local spatial information than the standard MLP, thereby enhancing feature representation capability and improving light field super-resolution performance.

Table 6: Comparison of different variants of PCFN.

Methods	FFN	PSNR	SSIM
w/o-PCFN	MLP	32.94	0.9504
LF-DUTS (ours)	PCFN	**33.11**	**0.9511**