# FEAD: Fine-Grained Epipolar Attention Diffusion for Large-Disparity Light Field Spatial Super-Resolution

**Wenbin Wang**[1] **Youfang Lin**[1,2] **Chen Gao**[1] **Shuo Zhang**[1,2]*

[1]School of Computer Science & Technology, Beijing Jiaotong University, Beijing, China
[2]Beijing Key Laboratory of Traffic Data Mining and Embodied Intelligence, China
{wangwenbin,yflin,gaochen,zhangshuo}@bjtu.edu.cn

## Abstract

Light field spatial super-resolution hinges on exploiting cross-view correspondences to recover high-frequency details while preserving geometric structure. However, existing methods often rely on implicit feature interaction or coarse disparity maps, which suffer noticeable performance degradation in regions with large disparities. In this paper, we propose Fine-Grained Epipolar Attention Diffusion (FEAD), which adopts epipolar attention as an explicit mechanism to characterize cross-view geometric correspondence along epipolar lines. To enable more accurate feature alignment and long-range spatial–angular interaction, we introduce a diffusion-based epipolar attention refinement strategy that progressively improves the initial attention. With the refined attention as guidance, cross-view features are explicitly warped and fused to enforce geometric consistency and aggregate complementary details across views for fine-grained reconstruction. Extensive experiments demonstrate that FEAD achieves state-of-the-art performance, with particularly strong gains in large-disparity scenarios.

## 1 Introduction

Light Field (LF) imaging captures both spatial and angular radiance information of light rays, enabling a wide range of applications such as post-capture refocusing [24], depth estimation [8], novel view synthesis [18], and 3D reconstruction [32]. However, due to the inherent trade-off between spatial and angular sampling in LF cameras [37], the spatial resolution of sub-aperture images (SAIs) is often limited, which significantly restricts the applicability of LF in high-precision visual tasks. As a result, LF spatial super-resolution (LFSSR), which aims to reconstruct high-resolution (HR) LF from their low-resolution (LR) counterparts, has become an important research problem [1, 6].

Unlike single-image, which relies on local spatial context for SR, LFSSR requires integrating redundant information across multiple SAIs to preserve the geometric structure. Early methods typically formulated LFSSR as an optimization problem, where explicit depth or disparity priors are imposed to guide reconstruction via global energy minimization or probabilistic inference. Recent deep LFSSR methods largely replace explicit geometric optimization with learned spatial-angular feature interaction, using convolutional neural network (CNNs) [16, 26, 27], Transformer-based architectures [2, 12], or Mamba-based models [5, 17]. Despite their promising performance, most existing networks still perform cross-view interaction implicitly, without explicitly verifying whether the aggregated features correspond to the same HR scene points. When the disparity becomes large, even a small correspondence error along an EPI may shift features across object boundaries or repetitive textures, causing the aggregation module to mix misaligned details rather than complementary information. Therefore, the key bottleneck is not only how to model long-range

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

| Warping method | MSE×100 ↓ | PSNR↑ | SSIM↑ |
|---|---|---|---|
| GT-depth warping | 0.067556 | 35.6040 | 0.936147 |
| GT-driven attention | 0.010717 | 41.8535 | 0.985996 |

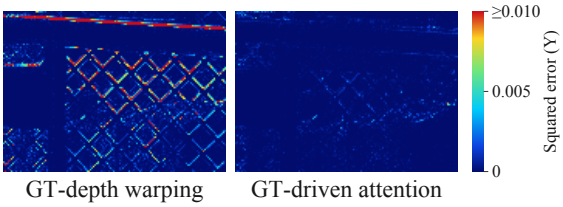



Figure 1: Comparison of GT-depth and GT-driven attention warping. Left: Metrics over 576 source–center pairs. Right: Pixelwise squared Y-channel errors for the same crop from *Boxes* (Cam049 → Cam040), computed against the center-view GT and shown on a shared color scale.

spatial-angular dependencies, but whether the model can obtain sufficiently accurate pixel-level correspondences to guide cross-view alignment.

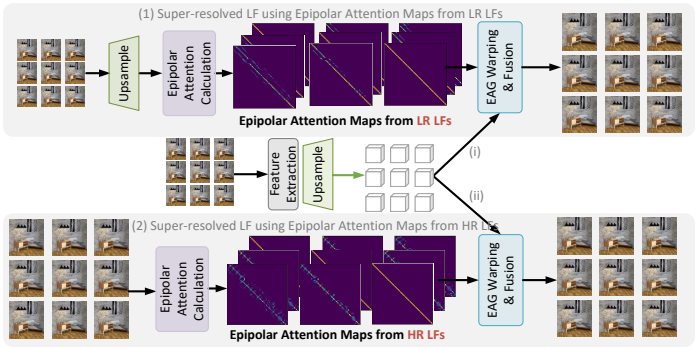
Depth-based warping uses depth and camera geometry to map each target pixel to a single sampling location in the source view. Even with GT depth, the corresponding scene point may be occluded in the source view or appear different across views. The parallax attention mechanism (PAM) [22] instead uses feature similarity to weight multiple candidates along the source epipolar line, allowing the warp to combine samples from several plausible matches. We compare the two warps on 24 depth-annotated $5 \times 5$ scenes from the 4D Light Field Benchmark [7], independently warping each non-central view to the center to obtain 576 source–center pairs. GT-depth warping uses the provided center-view GT depth and is evaluated on valid pixels after excluding out-of-bounds samples. GT-driven attention warping uses attention computed from all 25 HR GT views and is evaluated over the full image. The error maps in Fig. 1 show lower squared Y-channel errors along the crate rim and diagonal lattice bars with GT-driven attention.

We then explore how the quality of these attention-based correspondences affects LFSSR. As illustrated in Fig. 2, the attention map shows a correspondence distribution along the horizontal epipolar line: for each target pixel in the reference SAI, each element indicates the matching similarity to candidate pixels at different spatial offsets in another SAI. We plug this attention into two representative LFSSR backbones, EPIT and DistgSSR, and use it to guide the same HR-space warping and fusion process. Based on this framework, two variants are designed for guiding feature interaction: (i) $\mathbf{A}_{\mathrm{h}}\&\mathbf{A}_{\mathrm{v}}$: epipolar attention estimated from upsampled LR inputs, and (ii) $\mathbf{A}_{\mathrm{h}}^{GT}\&\mathbf{A}_{\mathrm{v}}^{GT}$: epipolar attention computed from HR ground truth (GT). Importantly, the feature extractor, warped features, and fusion pipeline are kept unchanged in both variants; only the attention guidance differs. With GT-driven epipolar attention, EPIT and DistgSSR improve by 9.01 dB and 9.44 dB over their original backbones, respectively, and by 8.89 dB and 9.28 dB over their LR-attention-guided counterparts (Fig. 2). This gap highlights the importance of accurate correspondence guidance for cross-view interaction and motivates refining epipolar attention before feature aggregation. The comparison isolates the effect of attention guidance, rather than establishing the intrinsic superiority of HR-space over LR-space attention.

Motivated by this finding, we propose **F**ine-grained **E**pipolar **A**ttention **D**iffusion (**FEAD**), an LFSSR framework that explicitly estimates, refines, and exploits epipolar attention for correspondence-aware cross-view interaction. FEAD first upsamples LR features and estimates epipolar attention in the HR space, so that correspondence candidates are defined at the same resolution as subsequent warping and fusion. Based on this attention, the Epipolar Attention Guided (EAG) Warping and Fusion module explicitly aligns and adaptively aggregates cross-view features across SAIs. The epipolar attention estimated from upsampled LR features is inherently coarse and noisy. Therefore, FEAD introduces a conditional residual diffusion strategy to progressively correct the initial attention toward a sharper and more geometrically reliable correspondence representation.

The main contributions of this work are threefold:

- We reveal and quantitatively analyze the reconstruction performance gap induced by LR-predicted and GT-driven epipolar attention, demonstrating that the precision of cross-view correspondence is a critical factor for LFSSR performance;




| Model | PSNR | SSIM |
|---|---|---|
| EPIT | 39.39 | 0.9869 |
| EPIT $+A_{\mathrm{h}}\&A_{\mathrm{v}}$ (1) | 39.51 (+0.12) | 0.9870 (+0.0001) |
| EPIT $+A_{\mathrm{h}}^{GT}\&A_{\mathrm{v}}^{GT}$ (2) | 48.40 (+9.01) | 0.9987 (+0.0118) |
| DistgSSR | 38.94 | 0.9867 |
| DistgSSR $+A_{\mathrm{h}}\&A_{\mathrm{v}}$ (1) | 39.10 (+0.16) | 0.9869 (+0.0002) |
| DistgSSR $+A_{\mathrm{h}}^{GT}\&A_{\mathrm{v}}^{GT}$ (2) | 48.38 (+9.44) | 0.9986 (+0.0119) |

Figure 2: Illustration and quantitative verification of the evaluation framework. Left: epipolar attention maps computed from upsampled LR inputs or HR GT are used to guide cross-view feature warping and aggregation directly in HR space. Right: quantitative results on $\times 2$ LFSSR demonstrate the reconstruction benefit of GT-driven epipolar attention. The gains in parentheses are measured relative to the original backbones.

- We propose FEAD, an epipolar attention–driven LFSSR framework that performs explicit cross-view interaction and incorporates a diffusion-based strategy to progressively refine epipolar attention, enabling accurate and stable geometric modeling.
- Experiments on multiple LF benchmarks verify the effectiveness of FEAD, particularly on medium- and large-disparity scenes, where it outperforms recent state-of-the-art methods such as EPIT-SSR and LF-DET by up to +0.52 dB and +0.73 dB PSNR at $\times 2$, respectively.

## 2 Related Work

Deep learning has significantly advanced LFSSR by designing increasingly effective spatial–angular representation learning mechanisms. LFCNN [31] is among the earliest CNN-based methods to adopt 4D convolutions for modeling inter-view correspondence in LFs. To capture spatial–angular dependencies efficiently, Yeung et al. [30] introduced Spatial-Angular Separable (SAS) convolution to approximate 4D LF convolution. LFNet [28] designs a bidirectional recurrent network to capture view-to-view dependencies along both horizontal and vertical EPIs. Multi-branch LF networks [35, 34] group LFs into multiple angular directions and model multi-directional spatial–angular correlations via parallel sub-networks. To better exploit information from all views, the All-to-One framework [9] reconstructs each target SAI by aggregating complementary information from the entire LF.

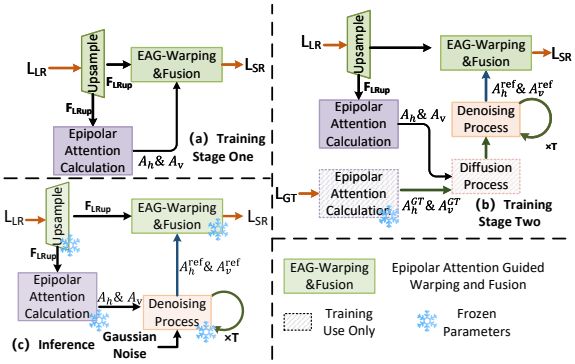



Figure 3: Training and inference strategy of the proposed FEAD framework.

More recently, Transformer-based LFSSR networks [23] explore sequence modeling to capture long-range dependencies across views. LFT [11] adopts a spatial–angular decoupled Transformer architecture to more effectively model LF correlations, while other Transformer-based methods [2] separately encode spatial and angular information to enhance global interaction. Furthermore, EPIT [12] performs attention modeling directly on EPIs, enabling effective exploitation of redundancy across SAIs. EPIT-SSR [13] further strengthens this architecture by enlarging the feature channel dimension. Inspired by the success of state space models in vision, Mamba-based methods have also been preliminarily applied to LFSSR to model long-range dependencies more efficiently [5, 17]. In contrast to these methods that focus on feature or dependency modeling, SEIIF [3] revisits the upsampling process and adopts an upsampling-aware implicit modeling strategy. Despite these advances, most

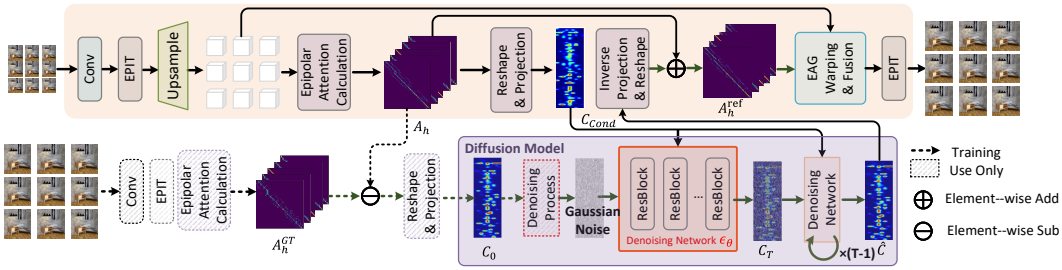



Figure 4: Overall architecture of our FEAD.

existing LFSSR methods mainly improve feature representation or spatial–angular dependency modeling, while cross-SAI feature interaction is still largely performed in an implicit manner. Without explicitly estimating reliable pixel-level correspondences or using them to guide cross-view feature alignment, these methods may aggregate misaligned information when the disparity between views becomes large. This issue is relatively mild in small-disparity scenes, where adjacent SAIs are nearly aligned, but becomes critical in large-disparity scenarios that require accurate correspondence-aware alignment. Therefore, explicitly modeling fine-grained inter-view correspondence and using it to guide cross-view warping and aggregation is important for fully exploiting complementary angular information in LFSSR.

## 3 Method

LF is commonly represented as a four-dimensional function $\mathbf{L}(u, v, h, w)$, where $(u, v)$ denotes the angular resolution and $(h, w)$ represents the spatial resolution of the LF. Our network takes an LR LF $\mathbf{L}_{\mathrm{LR}} \in \mathbb{R}^{U\times V\times h\times w}$ as input and aims to reconstruct a HR LF $\mathbf{L}_{\mathrm{SR}} \in \mathbb{R}^{U\times V\times H\times W}$, where $H = \alpha h$, $W = \alpha w$, and $\alpha$ denotes the spatial upsampling factor. As illustrated in Fig. 2, our analysis reveals a clear performance gap between LFSSR models guided by epipolar attention estimated from upsampled LR features and those guided by epipolar attention generated from GT features. This observation indicates that the geometric accuracy of correspondence modeling is crucial for cross-view alignment and directly impacts reconstruction performance.

Motivated by this observation, we design a two-stage FEAD framework, as illustrated in Fig. 3. In the first stage, LR features are first upsampled to the HR domain, where epipolar attention is estimated to explicitly model cross-view correspondences and guide epipolar attention–based alignment and aggregation (Sec. 3.1). This stage provides a stable HR-space correspondence initialization and enables preliminary geometry-aware feature fusion. In the second stage, the initially estimated epipolar attention is further refined by a conditional residual diffusion module and then exploited to facilitate more robust spatial–angular feature interaction and final reconstruction (Sec. 3.2).

### 3.1 Stage One: HR Epipolar Attention-guided Routing

As illustrated in Fig. 4, given the LR input $\mathbf{L}_{\mathrm{LR}}$, Stage One first extracts preliminary spatial-angular features at the LR scale and then promotes them to the HR space for subsequent processing. Specifically, three lightweight $3 \times 3$ convolutional layers followed by the AltFilter module in EPIT [12] are employed to model coarse spatial–angular correlations, producing LR features $\mathbf{F}_{\mathrm{LR}}$. These features are then upsampled to the HR space via PixelShuffle, yielding $\mathbf{F}_{\mathrm{LRup}} \in \mathbb{R}^{U\times V\times H\times W\times C}$.

Then the *Epipolar Attention Calculation* module estimates parallax-aware epipolar attention to explicitly model cross-view geometric correspondences. Subsequently, in the *EAG Warping and Fusion* module, we first align cross-view features under the guidance of the estimated epipolar attention and then use the *View Adaptive Fusion (VAF)* to fuse the warped features in a view-aware manner.

**Epipolar Attention Calculation.** To explicitly establish cross-view geometric correspondences in the HR space, we adopt a PAM-style [22] modeling strategy to estimate disparity-constrained cross-view attention on epipolar plane images (EPIs). Given HR features $\mathbf{F}_{\mathrm{LRup}}$, we first construct horizontal and vertical EPI feature groups and perform cross-view correlation modeling to obtain

parallax-aware attention representations. Taking the horizontal direction as an example, we rearranged $\mathbf{F}_{\mathrm{LRup}}$ into $\mathbf{E}_h \in \mathbb{R}^{B' \times C \times U \times H}$ by fixing the angular index $v$ and spatial coordinate $x$, where $B' = B \cdot V \cdot W$ denotes the number of horizontal EPI tokens.

Following the PAM-style correlation formulation [22], we calculate the similarity of the horizontal EPI tokens and obtain the horizontal HR parallax attention $\mathbf{A}_h \in \mathbb{R}^{B' \times (UH) \times (UH)}$, which explicitly encodes pixel-level cross-view correspondences in the HR space. More concretely, each horizontal epipolar attention map visualized in Fig. 2 corresponds to a $(UH) \times (UH)$ affinity matrix defined on a horizontal EPI. Each row represents a query pixel located at a specific angular index and spatial position in the reference view, while each column enumerates candidate pixels sampled along the same epipolar line across different views and spatial offsets. The highlighted responses indicate the most likely matching cross-view correspondences for a given query pixel. Due to the epipolar geometry of LF, valid correspondences are constrained to lie along a 1D epipolar line with approximately linear disparity-induced shifts across views. As a result, accurate correspondence estimation leads to attention responses concentrated around a near-diagonal structure, where the diagonal implicitly encodes consistent disparity slopes across angular positions. Pixels far away from this diagonal violate the epipolar constraint and therefore receive negligible attention weights, yielding the sparse, diagonal-dominant patterns observed in the visualization. The vertical direction follows a symmetric pipeline, yielding vertical parallax attention $\mathbf{A}_v \in \mathbb{R}^{B_v \times (VW) \times (VW)}$, where $B_v$ denotes the number of vertical EPI tokens. More visualizations of epipolar attention maps across different scenes and views are provided in the supplementary material.

**Epipolar Attention Guided Warping and Fusion.** Guided by the estimated HR epipolar attention, we perform explicit cross-view feature alignment directly in the HR space. Taking the horizontal direction as an example, the upsampled HR feature $\mathbf{F}_{\mathrm{LRup}}$ is rearranged into horizontal EPI sequences and projected into a value embedding $\mathbf{X}_h$. The horizontal attention $\mathbf{A}_h$ is then used to aggregate source-view features along the epipolar line:

$$\widetilde{\mathbf{X}}_h = \mathrm{MatMul}(\mathbf{A}_h, \mathbf{X}_h), \tag{1}$$

where $\widetilde{\mathbf{X}}_h$ denotes the parallax-aligned features gathered from different source SAIs. The vertical branch is constructed symmetrically with $\mathbf{A}_v$.

After attention-guided alignment, the warped candidates from different source views are fused in a view-adaptive manner to account for view-dependent reliability caused by occlusions and appearance variations. The detailed implementation of the view-adaptive fusion module is provided in Appendix A.

**Training Objective.** We train our network in stage one by minimizing a hybrid loss:

$$\mathcal{L}_{\mathrm{s1}} = \mathcal{L}_{\mathrm{sr}} + \lambda_{\mathrm{warp}}\mathcal{L}_{\mathrm{warp}} + \lambda_{\mathrm{KL}}\mathcal{L}_{\mathrm{KL}} \tag{2}$$

where $\mathcal{L}_{\mathrm{sr}}$ is the $\ell_1$ reconstruction loss of the $\mathbf{L}_{\mathrm{SR}}$ and GT $\mathbf{L}_{\mathrm{HR}}$. $\mathcal{L}_{\mathrm{KL}}$ is an attention-level Kullback–Leibler (KL) divergence loss that encourages the predicted epipolar attention $\mathbf{A}_{\mathrm{LRup}}$ to match the GT-driven epipolar attention $\mathbf{A}^{GT}$, which is treated as the target distribution:

$$\mathcal{L}_{\mathrm{KL}} = \mathrm{KL}\big(\mathbf{A}_{\mathrm{h}}^{GT} \parallel \mathbf{A}_{\mathrm{h}}\big) + \mathrm{KL}\big(\mathbf{A}_{\mathrm{v}}^{GT} \parallel \mathbf{A}_{\mathrm{v}}\big) \tag{3}$$

$\mathcal{L}_{\mathrm{warp}}$ is the warp-based reconstruction loss [22, 4] that supervises the geometric alignment induced by the epipolar attention:

$$\mathcal{L}_{\mathrm{warp}} = \big\|\mathrm{Warp}(\mathbf{L}_{\mathrm{HR}}, \mathbf{A}_{\mathrm{h}} \& \mathbf{A}_{\mathrm{v}}) - \mathbf{L}_{\mathrm{HR}}\big\|_1 \tag{4}$$

### 3.2 Stage Two: Epipolar Attention Refinement using Diffusion Module

The epipolar attention $\mathbf{A}_h$ and $\mathbf{A}_v$ estimated in Stage One are computed from $\mathbf{F}_{\mathrm{LRup}}$ and directly used to guide cross-view feature alignment. Such attention inevitably suffers from limited geometric precision due to the coarse nature of $\mathbf{F}_{\mathrm{LRup}}$, especially when applied to HR cross-view alignment. To address this limitation, as illustrated in Fig. 4, we introduce a diffusion-based optimization strategy in Stage Two to progressively refine the initial epipolar attention.

Rather than applying diffusion in the image space, we formulate attention refinement as a conditional residual diffusion problem [15]. Specifically, the residual between the GT-derived HR attention and

the initial attention is used as the learning target:

$$\mathbf{z}_h = \mathbf{A}_h^{GT} - \mathbf{A}_h, \quad \mathbf{z}_v = \mathbf{A}_v^{GT} - \mathbf{A}_v. \tag{5}$$

It should be noted that the diffusion model does not directly operate on the raw high-dimensional residuals $\mathbf{z}_h$ and $\mathbf{z}_v$. Instead, the residuals are first transformed into an internal compact correspondence space through *Reshape & Projection*, where conditional denoising is performed under the guidance of the initial epipolar attention. The denoised outputs are then mapped back to the attention-residual space through *Inverse Projection & Reshape*, yielding residual corrections $\hat{\mathbf{z}}_h$ and $\hat{\mathbf{z}}_v$. The refined attention is obtained by:

$$\mathbf{A}_h^{ref} = \mathbf{A}_h + \hat{\mathbf{z}}_h, \quad \mathbf{A}_v^{ref} = \mathbf{A}_v + \hat{\mathbf{z}}_v. \tag{6}$$

The refined attention maps replace the initial ones in Eq. (1) to guide cross-view feature alignment and aggregation. The detailed formulation of the internal projection and denoising process is provided in Appendix B. We further analyze the effect of diffusion-based attention refinement in Sec. 4.2, where the initially estimated, diffusion-refined, and GT epipolar attention maps are visualized and compared.

**Training and Inference.** It is worth noting that $\mathbf{A}_h^{GT}$ and $\mathbf{A}_v^{GT}$ are used only during training to construct the residual target and attention-level supervision. During inference, the refinement module takes only the initial attention estimated from the LR input as its condition and predicts the residual correction without accessing any HR ground-truth information. Stage Two adopts the same training objective as Stage One, except that $\mathcal{L}_{\mathrm{KL}}$ and $\mathcal{L}_{\mathrm{warp}}$ are applied to the refined epipolar attention $\mathbf{A}_{\mathrm{h}}^{\mathrm{ref}}$ and $\mathbf{A}_{\mathrm{v}}^{\mathrm{ref}}$ produced by the diffusion model. In this way, the refined attention is explicitly guided toward $\mathbf{A}_{\mathrm{h}}^{\mathrm{GT}}$ and $\mathbf{A}_{\mathrm{v}}^{\mathrm{GT}}$ while being jointly optimized with the SR network. As shown in Fig. 3, during inference, LR features are first extracted and upsampled to the HR space. The initial epipolar attention $\mathbf{A}_{\mathrm{h}}$ and $\mathbf{A}_{\mathrm{v}}$ serve as the condition of the diffusion model, which generates a refined residual $\hat{\mathbf{z}}_0$ via iterative denoising. After attention decoding, the refined epipolar attention is applied by EAG Warping & Fusion to align HR features across views, which are further processed by subsequent AltFilter-based interaction modules to produce the final HR LF.

# 4 Experiments

We conduct experiments on five commonly used public LF datasets, including two synthetic datasets (HCInew [7] and HCIold [29]) and three real-world datasets (EPFL [19], INRIA [10], and STFgantry [20]). Among them, scenes in the STFgantry dataset exhibit relatively large disparity variations due to the gantry-based acquisition setup, with disparity ranges spanning from $-4$ to $4$.

## 4.1 Implementation Details

Following the [11, 12, 34, 32], all experiments are conducted on LF extracted from the central $5 \times 5$ SAIs. During training, each SAI is first cropped into HR patches of size $32 \times 32$, and the corresponding LR inputs are then obtained via bicubic downsampling for $\times 2$ and $\times 4$ SR tasks. We follow the same training settings of EPIT [12] and adopt a unified training configuration for all experiments. The network parameters are initialized using Xavier initialization and optimized with the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$). The initial learning rate is set to $2 \times 10^{-4}$ and decayed by a factor of $0.5$ every $15$ epochs. Random horizontal flipping, vertical flipping, and $90^\circ$ rotation are applied for data augmentation. We adopt a two-stage training strategy. In Stage One, the network is trained end-to-end without the diffusion-based attention refinement module to obtain a stable initialization. In Stage Two, we load the pretrained weights from Stage One and jointly optimize together with the diffusion-based epipolar attention refinement module. In the denoising network, we set the number of ResBlocks to $5$ and the diffusion step $T$ to 4. The loss weights are set to $\lambda_{\mathrm{warp}} = 0.1$ and $\lambda_{\mathrm{KL}} = 0.1$. All models are implemented using PyTorch and trained for $81$ epochs on a single NVIDIA RTX A4000 GPU.

## 4.2 Comparison with the State-of-the-Arts

We compare our model with 13 representative LFSSR methods, including resLF [34], SAS [30], LF-ATO [9], LF_InterNet [25], LF-DFNet [26], MEG-Net [35], LF-IIINet [16], DPT [23], LFT [11],

Table 1: Quantitative comparison (PSNR/SSIM) with state-of-the-art LF-SSR methods on five LF datasets under $\times 2$ and $\times 4$ SR settings. Best and second-best results are highlighted in red and blue, respectively.

| Methods | $\times 2$ | | | | | $\times 4$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | EPFL $[-1, 1]$ | INRIA $[-1, 1]$ | HCInew $[-2, 2]$ | HCIold $[-2, 2]$ | STFgantry $[-4, 4]$ | EPFL $[-1, 1]$ | INRIA $[-1, 1]$ | HCInew $[-2, 2]$ | HCIold $[-2, 2]$ | STFgantry $[-4, 4]$ |
| Bicubic | 29.74/0.9376 | 31.33/0.9577 | 31.89/0.9356 | 37.69/0.9785 | 31.06/0.9498 | 25.14/0.8324 | 26.82/0.8867 | 27.61/0.8517 | 32.42/0.9344 | 25.93/0.8452 |
| VDSR | 32.50/0.9598 | 34.43/0.9741 | 34.37/0.9561 | 40.61/0.9867 | 35.54/0.9789 | 27.25/0.8777 | 29.19/0.9204 | 29.31/0.8823 | 34.81/0.9515 | 28.51/0.9009 |
| EDSR | 33.09/0.9629 | 34.97/0.9764 | 34.83/0.9592 | 41.01/0.9874 | 36.29/0.9818 | 27.84/0.8854 | 29.66/0.9257 | 29.60/0.8869 | 35.18/0.9536 | 28.70/0.9072 |
| RCAN | 33.16/0.9634 | 35.01/0.9769 | 34.98/0.9603 | 41.05/0.9875 | 36.33/0.9831 | 27.88/0.8863 | 29.76/0.9276 | 29.63/0.8886 | 35.20/0.9548 | 28.90/0.9131 |
| resLF | 33.62/0.9706 | 35.40/0.9804 | 36.69/0.9739 | 43.42/0.9932 | 38.35/0.9904 | 28.26/0.9035 | 30.34/0.9412 | 30.72/0.9107 | 36.71/0.9682 | 30.19/0.9372 |
| SAS | 33.67/0.9744 | 35.28/0.9832 | 36.80/0.9749 | 43.81/0.9938 | 37.94/0.9898 | 28.60/0.9118 | 30.59/0.9467 | 30.93/0.9145 | 36.91/0.9696 | 30.57/0.9426 |
| LF-ATO | 34.27/0.9757 | 36.17/0.9842 | 37.24/0.9767 | 44.21/0.9942 | 39.64/0.9929 | 28.51/0.9115 | 30.71/0.9484 | 30.88/0.9135 | 37.00/0.9699 | 30.61/0.9430 |
| LF_InterNet | 34.11/0.9760 | 35.83/0.9843 | 37.17/0.9763 | 44.57/0.9946 | 38.44/0.9909 | 28.81/0.9162 | 30.78/0.9491 | 30.96/0.9161 | 37.15/0.9716 | 30.37/0.9409 |
| LF-DFNet | 34.51/0.9755 | 36.42/0.9840 | 37.42/0.9773 | 44.20/0.9941 | 39.43/0.9926 | 28.77/0.9165 | 30.83/0.9503 | 31.23/0.9196 | 37.32/0.9718 | 31.15/0.9494 |
| MEG-Net | 34.31/0.9773 | 36.10/0.9849 | 37.42/0.9777 | 44.10/0.9942 | 38.77/0.9915 | 28.75/0.9160 | 30.67/0.9490 | 31.10/0.9177 | 37.29/0.9716 | 30.77/0.9453 |
| LF-IIINet | 34.73/0.9773 | 36.57/0.9853 | 37.77/0.9790 | 44.85/0.9948 | 39.89/0.9936 | 29.04/0.9188 | 31.03/0.9515 | 31.33/0.9208 | 37.62/0.9734 | 31.26/0.9502 |
| DPT | 34.49/0.9758 | 36.41/0.9843 | 37.35/0.9771 | 44.30/0.9943 | 39.43/0.9926 | 28.94/0.9170 | 30.96/0.9503 | 31.20/0.9188 | 37.41/0.9721 | 31.15/0.9488 |
| LFT | 34.80/0.9781 | 36.59/0.9855 | 37.84/0.9791 | 44.52/0.9945 | 40.51/0.9941 | 29.26/0.9210 | 31.21/0.9524 | 31.46/0.9218 | 37.63/0.9735 | 31.86/0.9548 |
| DistgSSR | 34.81/0.9787 | 36.59/0.9859 | 37.96/0.9796 | 44.94/0.9949 | 40.40/0.9942 | 28.99/0.9195 | 30.99/0.9519 | 31.38/0.9217 | 37.56/0.9732 | 31.65/0.9535 |
| HLFSR-SSR | 35.31/0.9800 | 37.06/0.9867 | 38.32/0.9807 | 44.98/0.9950 | 40.85/0.9947 | 29.20/0.9222 | 31.24/0.9534 | 31.57/0.9238 | 37.83/0.9742 | 31.64/0.9537 |
| LF-DET | 35.26/0.9797 | 36.95/0.9864 | 38.32/0.9807 | 44.99/0.9950 | 41.76/0.9955 | 29.47/0.9230 | 31.39/0.9534 | 31.56/0.9235 | 37.84/0.9744 | 32.14/0.9573 |
| EPIT-SSR | 34.92/0.9799 | 37.14/0.9896 | 38.42/0.9814 | 45.20/0.9951 | 42.31/0.9958 | 29.51/0.9197 | 31.49/0.9540 | 31.60/0.9241 | 37.89/0.9744 | 32.29/0.9585 |
| **Ours** | 34.92/0.9775 | 36.67/0.9850 | 38.70/0.9817 | 45.72/0.9954 | 42.49/0.9961 | 29.40/0.9210 | 31.36/0.9530 | 31.64/0.9240 | 37.92/0.9748 | 32.47/0.9600 |

DistgSSR [27], HLFSR-SSR [21], LF-DET [2], and EPIT-SSR [13]. We also include Bicubic and representative single-image SR baselines, i.e., VDSR, EDSR, and RCAN [14, 36], by applying them to each SAI independently. We adopt peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) as evaluation metrics, computed on the Y channel of all SAIs.

**Quantitative Results.** Quantitative results are reported in Table 1.

The intervals $[-1, 1]$, $[-2, 2]$ and $[-4, 4]$ represent the disparity ranges of the HR images, which are used to measure the pixel displacements of scene points between different SAIs caused by the multi-view acquisition process. A larger disparity range implies a greater cross-view displacement and a more complex correspondence relationship modeling.

Compared with existing approaches, the performance gains of our method are particularly pronounced on medium- and large-disparity datasets. On HCInew, HCIold, and STFgantry, FEAD achieves the best PSNR under both $\times 2$ and $\times 4$ settings. Compared with the recent EPIT-SSR, FEAD improves PSNR by 0.28/0.52/0.18 dB at $\times 2$ on HCInew/HCIold/STFgantry, respectively. Compared with LF-DET, the corresponding gains are 0.38/0.73/0.73 dB at $\times 2$, respectively.

This advantage stems from our explicit modeling of fine-grained, parallax-aware geometric correspondences, which enforces stronger cross-view geometric consistency and enables more accurate feature alignment and aggregation for LFSSR. On small-disparity datasets such as EPFL and INRIA, the geometric variations across SAIs are relatively limited, and different views are nearly aligned. Under such conditions, cross-view correspondence is easier to establish, and convolution-based methods with large model capacity (e.g., HLFSR) can effectively aggregate inter-view information through strong local feature modeling, leading to competitive performance.

**Qualitative Results.** Fig. 5 presents the qualitative comparisons of methods for $\times 2/\times 4$ SR. As observed from the zoomed-in regions, most existing LFSSR methods still suffer from blurry texture and detail reconstruction. In contrast, our method explicitly performs cross-view interaction among SAIs, enabling more effective aggregation of multi-view information and resulting in more faithful details with fewer artifacts (e.g. in the $\times 2$ results, the reconstructed bracket structure appears clearer and more complete, while in the $\times 4$ results, the texture details in the bottom-right region are better preserved with significantly reduced artifacts). Compared with most existing LFSSR approaches, our method achieves superior visual quality with high angular consistency, as evidenced by the more continuous and well-aligned epipolar lines observed in the EPI visualizations. Additional pixel-wise error maps and per-view PSNR distributions are provided in Appendices C.1 and C.2, respectively.

**Computational Efficiency.** Table 2 summarizes the cost and accuracy comparison under the $\times 2$ LF-SSR setting on medium- and large-disparity scenes. Avg. $[-2, 2]$ is computed by averaging the

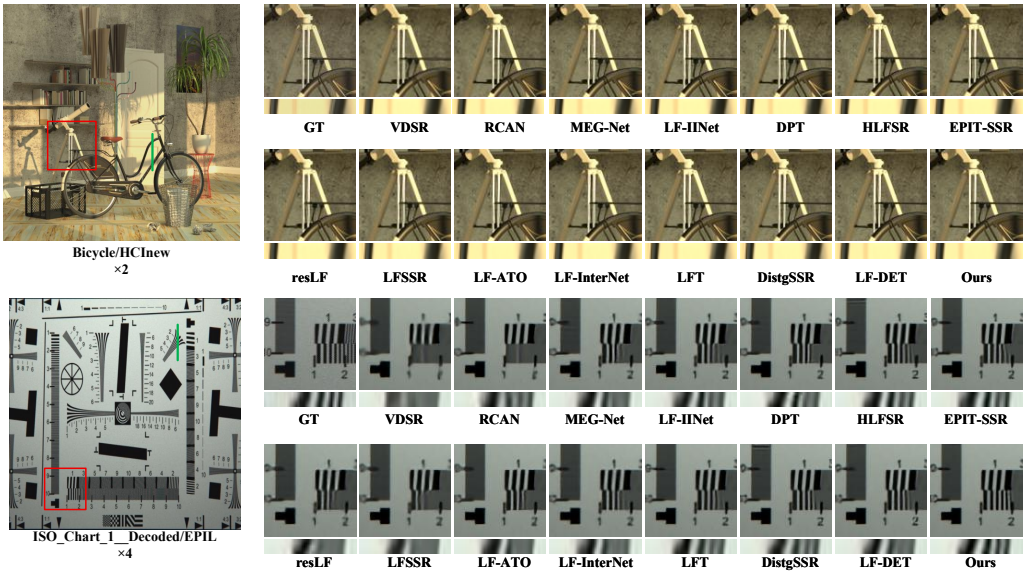



Figure 5: Qualitative comparison of different LF-SSR methods for $\times 2$ and $\times 4$ SR.

Table 2: Cost and accuracy comparison for $\times 2$ LF-SSR on medium and large-disparity settings.

| Methods | Para. (M) | FLOPs (G) | Train (h) | Infer (s) | $[-2, 2]$ | $[-4, 4]$ |
|---|---|---|---|---|---|---|
| HLFSR-SSR | 13.72 | 156.15 | 2.5 | 6.83 | 41.65 | 40.85 |
| LF-DET | 1.59 | 48.50 | 3.0 | 5.72 | 41.66 | 41.76 |
| EPIT-SSR | 5.68 | 216.19 | 3.5 | 34.31 | 41.81 | 42.31 |
| Ours ($T = 1$) | 2.78 | 209.19 | 0.8 (S1) + 1.4 (S2) | 5.26 | 41.26 | 40.51 |
| Ours ($T = 2$) | 2.78 | 354.81 | 0.8 (S1) + 2.1 (S2) | 18.72 | 42.10 | 42.40 |
| **Ours** ($T = 4$) | **2.78** | **646.06** | **0.8 (S1) + 3.1 (S2)** | **44.19** | **42.21** | **42.49** |
| Ours ($T = 8$) | 2.78 | 1228.54 | 0.8 (S1) + 5.7 (S2) | 108.43 | 42.23 | 42.74 |

results on HCInew and HCIold, while Avg. $[-4, 4]$ corresponds to the performance on STFgantry. These settings better reflect the scenarios where accurate cross-view correspondence modeling is critical. Under the default setting of $T = 4$, our FEAD achieves 42.21 dB on Avg. $[-2, 2]$ and 42.49 dB on Avg. $[-4, 4]$, outperforming EPIT-SSR by +0.40 dB and +0.18 dB, respectively. Compared with LF-DET, FEAD further brings PSNR gains of +0.55 dB and +0.73 dB on the two settings. Meanwhile, FEAD uses fewer parameters than EPIT-SSR, i.e., 2.78M vs. 5.68M, while maintaining comparable total training time, i.e., 3.9h vs. 3.5h. Although the explicit EAG Warping & Fusion and diffusion-based attention refinement introduce additional inference cost, they provide clear reconstruction benefits on challenging scenes with larger cross-view disparities. We further analyze the effect of diffusion steps. The model size remains fixed at 2.78M for all $T$, while FLOPs and inference time increase as $T$ grows. From $T = 2$ to $T = 4$, the PSNR improves from 42.10 dB to 42.21 dB on Avg. $[-2, 2]$ and from 42.40 dB to 42.49 dB on Avg. $[-4, 4]$. Although $T = 8$ further improves the large-disparity result, it substantially increases the computational cost, with FLOPs rising to 1228.54G and inference time increasing to 108.43s. Therefore, we adopt $T = 4$ as the default setting to balance reconstruction accuracy and computational efficiency.

**Angular Consistency.** Beyond pixel-wise reconstruction accuracy, preserving angular consistency is essential for LF enhancement, since the reconstructed LF is often used by downstream geometry-related applications such as depth or disparity estimation. Therefore, we further evaluate the geometric consistency of different SR methods by applying the Spinning Parallelogram Operator (SPO) LF disparity estimation method [33] to the reconstructed $\times 2$ LF results. As shown in Fig. 6, the disparity maps estimated from our reconstructed LFs are visually closer to GT than those produced by competing methods, especially around object boundaries and structured regions. Quantitatively,

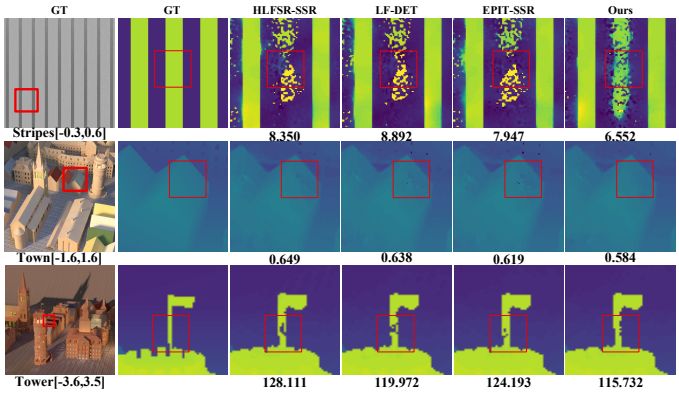



Figure 6: Quantitative and qualitative (MSE) comparisons of disparity estimation results achieved by SPO using different $2\times$ SR results. The MSE ($\downarrow$) is the mean square error multiplied by 100.

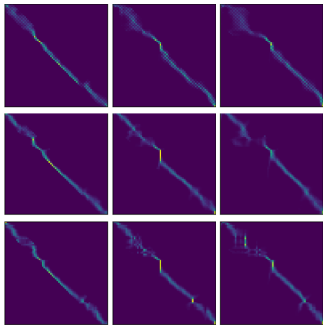


Figure 7: Visualization of epipolar attention maps. From top to bottom: the initially estimated, the diffusion-refined, and the GT-derived epipolar attention.

FEAD consistently achieves the lowest MSE$\times$100 across low-, medium-, and large-disparity scenes, obtaining 6.552, 0.584, and 115.732 on Stripes, Town, and Tower, respectively. Compared with the recent EPIT-SSR, FEAD reduces the disparity estimation error from 7.947 to 6.552 on Stripes, from 0.619 to 0.584 on Town, and from 124.193 to 115.732 on Tower. These results demonstrate that FEAD better preserves cross-view geometric consistency, indicating that the proposed HR epipolar attention refinement and attention-guided warping improve not only reconstruction fidelity but also the geometric structure of LFs.

**Attention Map Refinement.** Figure 7 visualizes the initially estimated attention, the diffusion-refined attention, and the GT-derived HR epipolar attention computed from HR features. As can be observed, the initially estimated epipolar attention is relatively coarse and blurred, exhibiting unclear correspondence structures and noticeable deviations from the true epipolar geometry. Such imprecise attention distributions make it difficult to faithfully capture fine-grained cross-view correspondences, especially under large-disparity variations. After applying the proposed diffusion-based refinement, the epipolar attention becomes significantly sharper and more structured. The refined attention exhibits clearer and more continuous responses along the epipolar lines, with substantially reduced noise and ambiguity. More importantly, its spatial distribution aligns much better with the GT epipolar attention, indicating that the refinement process effectively corrects geometric inaccuracies and recovers reliable correspondence patterns. These results show that the proposed refinement effectively narrows the gap between initial and true epipolar correspondences, providing more accurate geometric guidance for cross-view alignment and aggregation. Additional attention-level MSE analysis and the comparison with LR-attention conditioning are provided in Appendix C.3.

### 4.3 Ablation Studies

Table 3 presents a comprehensive ablation study for $\times 2$ SR. The reported Avg. in $[-2, 2]$ is computed by averaging the results on HCIold and HCInew, while Avg. in $[-4, 4]$ corresponds to the performance on the STFgantry.

**Effectiveness of EAG Warping and Fusion.** By comparing model (1) and (2), introducing EAG Warping and Fusion already brings a noticeable performance improvement, +0.295 dB on Avg$[-2, 2]$. Model (6) and (7) investigate different fusion strategies under the same diffusion refinement setting. While simple Mean or Sum fusion yields reasonable results, compared with (11), the proposed VAF consistently performs better, with an average improvement of +0.12 dB.

**Effectiveness of Attention Refinement.** The results from model (2) to (5) further demonstrate that explicitly refining epipolar attention is beneficial. Among different refinement strategies, Transformer and diffusion-based methods consistently outperform SAS and 3D alternatives. In particular, the proposed diffusion-based refinement model (11) achieves the best overall performance among all

Table 3: Module ablation study on $\times 2$ LF-SSR.

| Num | Stage One | | Stage Two | Att Loss | | PSNR / SSIM | |
|---|---|---|---|---|---|---|---|
| | Warping | Fusion | Attention Refine | $\mathcal{L}_{\mathrm{warp}}$ | $\mathcal{L}_{\mathrm{KL}}$ | Avg. $[-2, 2]$ | Avg. $[-4, 4]$ |
| 1) | × | × | × | × | × | 41.680 / 0.9863 | 42.184 / 0.9937 |
| 2) | ✓ | VAF | × | ✓ | ✓ | 41.975 / 0.9874 | 42.209 / 0.9942 |
| 3) | ✓ | VAF | SAS | ✓ | ✓ | 41.996 / 0.9874 | 42.228 / 0.9953 |
| 4) | ✓ | VAF | 3D Conv | ✓ | ✓ | 41.986 / 0.9873 | 42.278 / 0.9948 |
| 5) | ✓ | VAF | Transformer | ✓ | ✓ | 42.063 / 0.9882 | 42.396 / 0.9959 |
| 6) | ✓ | Mean | Diffusion | ✓ | ✓ | 42.089 / 0.9876 | 42.372 / 0.9951 |
| 7) | ✓ | Sum | Diffusion | ✓ | ✓ | 42.096 / 0.9882 | 42.399 / 0.9957 |
| 8) | ✓ | VAF | Diffusion | × | × | 42.073 / 0.9883 | 42.326 / 0.9957 |
| 9) | ✓ | VAF | Diffusion | × | ✓ | 42.110 / 0.9885 | 42.370 / 0.9959 |
| 10) | ✓ | VAF | Diffusion | ✓ | × | 42.123 / 0.9885 | 42.457 / 0.9959 |
| 11) | ✓ | VAF | Diffusion | ✓ | ✓ | **42.214 / 0.9886** | **42.494 / 0.9961** |

variants, surpassing the Transformer-based counterpart model (5) by +0.151 dB and +0.098 dB on Avg$[-2, 2]$ and Avg$[-4, 4]$, respectively. This result highlights the advantage of progressive denoising in correcting ambiguous and noisy attention patterns.

**Effectiveness of Loss Design.** Model (8) to (11) show the impact of loss function design. Removing either $\mathcal{L}_{warp}$ or $\mathcal{L}_{KL}$ leads to noticeable performance degradation, as the generated epipolar attention lacks explicit supervision during training. Specifically, $\mathcal{L}_{KL}$ encourages the refined attention distribution to align with the reference attention derived from HR features, while the $\mathcal{L}_{warp}$ directly enforces the effectiveness of attention for the correspondence of GT. Omitting either loss weakens the refinement process, whereas the complete objective model (11) achieves the best performance. These results indicate that both geometric consistency and distribution-level regularization are essential for stable and effective diffusion-based attention refinement.

## 5 Conclusion

In this paper, we first quantitatively demonstrate that fine-grained epipolar correspondence representation is critical for LFSSR performance. Motivated by this observation, we design an EAG-Warping & Fusion module to explicitly align and aggregate features across SAIs, enabling effective modeling of cross-view geometric relationships. Furthermore, a diffusion-based refinement strategy is introduced to progressively transform the initially estimated attention into a more accurate and fine-grained epipolar correspondence. Experimental results show that the proposed design achieves significant performance improvements on large-disparity LF datasets.

## Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 62372032).

## References


[1] Zhen Cheng, Zhiwei Xiong, Chang Chen, Dong Liu, and Zheng-Jun Zha. Light field super-resolution with zero-shot learning. In *CVPR*, pages 10010–10019, 2021.

[2] 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.

[3] Ruixuan Cong, Yu Wang, Mingyuan Zhao, Da Yang, Rongshan Chen, and Hao Sheng. Rethinking the upsampling process in light field super-resolution with spatial-epipolar implicit image function. In *ICCV*, pages 7559–7569, 2025.

[4] Chen Gao, Shuo Zhang, and Youfang Lin. Epipolar consistent attention aggregation network for unsupervised light field disparity estimation. In *ICCV*, pages 6488–6497, 2025.

[5] Ruisheng Gao, Zeyu Xiao, and Zhiwei Xiong. Mamba-based light field super-resolution with efficient subspace scanning. In *ACCV*, pages 531–547, 2024.

[6] Mantang Guo, Junhui Hou, Jing Jin, Jie Chen, and Lap-Pui Chau. Deep spatial-angular regularization for light field imaging, denoising, and super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6094–6110, 2021.

[7] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *ACCV*, pages 19–34. Springer, 2016.

[8] Jing Jin and Junhui Hou. Occlusion-aware unsupervised learning of depth from 4-d light fields. *IEEE Transactions on Image Processing*, 31:2216–2228, 2022.

[9] Jing Jin, Junhui Hou, Jie Chen, and Sam Kwong. Light field spatial super-resolution via deep combinatorial geometry embedding and structural consistency regularization. In *CVPR*, pages 2260–2269, 2020.

[10] 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.

[11] Zhengyu Liang, Yingqian Wang, Longguang Wang, Jungang Yang, and Shilin Zhou. Light field image super-resolution with transformers. *IEEE Signal Processing Letters*, 29:563–567, 2022.

[12] Zhengyu Liang, Yingqian Wang, Longguang Wang, Jungang Yang, Shilin Zhou, and Yulan Guo. Learning non-local spatial-angular correlation for light field image super-resolution. In *ICCV*, pages 12376–12386, 2023.

[13] 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.

[14] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *CVPRW*, July 2017.

[15] Chengxu Liu, Lu Qi, Jinshan Pan, Xueming Qian, and Ming-Hsuan Yang. Frequency domain-based diffusion model for unpaired image dehazing. *arXiv preprint arXiv:2507.01275*, 2025.

[16] 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.

[17] Yao Lu, Shunzhou Wang, Ziqi Wang, Peiqi Xia, Tianfei Zhou, et al. Lfmamba: Light field image super-resolution with state space model. *arXiv preprint arXiv:2406.12463*, 2024.

[18] Yifan Mao, Zeyu Xiao, Ping An, Deyang Liu, and Caifeng Shan. Deep sparse-to-dense inbetweening for multi-view light fields. *IEEE Transactions on Image Processing*, 34:6302–6317, 2025. doi: 10.1109/TIP.2025.3612257.

[19] Martin Rerabek and Touradj Ebrahimi. New light field image dataset. In *QoMEX*, 2016.

[20] Abhilash Sunder Raj, Michael Lowney, Raj Shah, and Gordon Wetzstein. The stanford lytro light field archive. `http://lightfields.stanford.edu/LF2016.html`, 2016. Accessed: Aug. 22, 2020.

[21] 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.

[22] Longguang Wang, Yulan Guo, Yingqian Wang, Zhengfa Liang, Zaiping Lin, Jungang Yang, and Wei An. Parallax attention for unsupervised stereo correspondence learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2108–2125, 2020.

[23] Shunzhou Wang, Tianfei Zhou, Yao Lu, and Huijun Di. Detail-preserving transformer for light field image super-resolution. In *AAAI*, volume 36, pages 2522–2530, 2022.

[24] Yingqian Wang, Jungang Yang, Yulan Guo, Chao Xiao, and Wei An. Selective light field refocusing for camera arrays using bokeh rendering and superresolution. *IEEE Signal Processing Letters*, 26(1):204–208, 2018.

[25] Yingqian Wang, Longguang Wang, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. Spatial-angular interaction for light field image super-resolution. In *ECCV*, pages 290–308. Springer, 2020.

[26] Yingqian Wang, Jungang Yang, Longguang Wang, Xinyi Ying, Tianhao Wu, Wei An, and Yulan Guo. Light field image super-resolution using deformable convolution. *IEEE Transactions on Image Processing*, 30:1057–1071, 2020.

[27] 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, 2022.

[28] Yunlong Wang, Fei Liu, Kunbo Zhang, Guangqi Hou, Zhenan Sun, and Tieniu Tan. Lfnet: A novel bidirectional recurrent convolutional neural network for light-field image super-resolution. *IEEE Transactions on Image Processing*, 27(9):4274–4286, 2018.

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

[30] 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.

[31] 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.

[32] Jingyang Zhang, Yao Yao, and Long Quan. Learning signed distance field for multi-view surface reconstruction. In *ICCV*, pages 6525–6534, 2021.

[33] 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.

[34] Shuo Zhang, Youfang Lin, and Hao Sheng. Residual networks for light field image super-resolution. In *CVPR*, pages 11046–11055, 2019.

[35] Shuo Zhang, Song Chang, and Youfang Lin. End-to-end light field spatial super-resolution network using multiple epipolar geometry. *IEEE Transactions on Image Processing*, 30:5956–5968, 2021.

[36] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018.

[37] Hao Zhu, Mantang Guo, Hongdong Li, Qing Wang, and Antonio Robles-Kelly. Revisiting spatio-angular trade-off in light field cameras and extended applications in super-resolution. *IEEE Transactions on Visualization and Computer Graphics*, 27(6):3019–3033, 2019.

## Appendix Overview

This appendix provides additional technical details and experimental analyses to complement the main paper. Appendix A details the epipolar attention-guided aggregation and view-adaptive fusion modules. Appendix B presents the formulation of the residual diffusion refinement process. Appendix C provides additional qualitative and quantitative analyses, including pixel-wise error visualizations, view-wise reconstruction consistency, attention accuracy comparisons, loss-weight sensitivity, and limitations.

## A Epipolar Attention-Guided Aggregation

Figure 8 illustrates the detailed implementation of epipolar attention-guided warping and view-adaptive fusion. Taking the horizontal direction as an example, the upsampled HR feature $\mathbf{F}_{\mathrm{LRup}}$ is rearranged into horizontal EPI sequences and projected into a value embedding space: $\mathbf{X}_h \in \mathbb{R}^{B' \times U \times H \times C}$. The predicted horizontal epipolar attention $\mathbf{A}_h \in \mathbb{R}^{B' \times (UH) \times (UH)}$ is used in a source-view-preserving manner. Specifically, for implementation, $\mathbf{A}_h$ can be equivalently reshaped as $\mathbf{A}_h \in \mathbb{R}^{B' \times U \times H \times U \times H}$, where the first $(U, H)$ indexes the target EPI position and the second $(U, H)$ indexes the source EPI candidate. Thus, $\mathbf{A}_h^{(u,p,u',p')}$ denotes the correspondence weight from the target position $(u, p)$ to the source candidate $(u', p')$ on the same horizontal epipolar line.

For each target view $u$ and source view $u'$, the attention-guided warped candidate is obtained by aggregating source-view features along the spatial dimension of the EPI:

$$\tilde{\mathbf{X}}_h^{(u,u')}(p) = \sum_{p'=1}^{H} \mathbf{A}_h^{(u,p,u',p')} \mathbf{X}_h^{(u')}(p'), \tag{7}$$

where $p$ and $p'$ denote spatial positions on the horizontal EPI, and $\tilde{\mathbf{X}}_h^{(u,u')}$ denotes the parallax-aligned feature candidate gathered from source view $u'$ for target view $u$. In this way, EAG warping preserves the source-view index $u'$ and produces $U$ aligned candidates for each target view, rather than collapsing all source views into a single feature immediately.

Although the warped candidates are aligned by epipolar attention, different source views may contribute unequally due to occlusion, boundary effects, and view-dependent appearance changes. Therefore, we adopt a lightweight view-adaptive fusion module to predict source-view reliability weights:

$$\mathbf{W}_h = \mathrm{softmax}(\psi_h(\mathbf{F}_{\mathrm{LRup}})), \tag{8}$$

where $\psi_h(\cdot)$ denotes a lightweight convolution and linear projection. Given the aligned candidates $\tilde{\mathbf{X}}_h^{(u,u')}$, the feature for target view $u$ is obtained by weighted aggregation:

$$\mathbf{Y}_h^{(u)} = \sum_{u'=1}^{U} \mathbf{W}_h^{(u,u')} \cdot \tilde{\mathbf{X}}_h^{(u,u')}. \tag{9}$$

The vertical branch follows the same procedure to produce $\mathbf{Y}_v$.

## B Formulation of Residual Diffusion Refinement

This section provides the detailed formulation of the residual diffusion module used for epipolar attention refinement. We take the horizontal attention as an example, and the vertical attention follows the same procedure.

Given the initial HR epipolar attention $\mathbf{A}_h$ and the GT-derived attention $\mathbf{A}_h^{GT}$, the raw attention residual is defined as

$$\mathbf{z}_0 = \mathbf{A}_h^{GT} - \mathbf{A}_h. \tag{10}$$

Since $\mathbf{z}_0$ is a high-dimensional attention residual, we do not directly perform diffusion on it. Instead, we apply *Reshape & Projection* to transform it into a compact correspondence representation:

$$\mathbf{c}_0 = \mathcal{P}(\mathbf{z}_0), \tag{11}$$

where $\mathcal{P}(\cdot)$ denotes the *Reshape & Projection* operation. Specifically, the attention residual is first rearranged along the angular dimensions and then projected into a fixed-dimensional compact correspondence space. The diffusion process is conducted on $\mathbf{c}_0$ rather than on the raw residual $\mathbf{z}_0$.

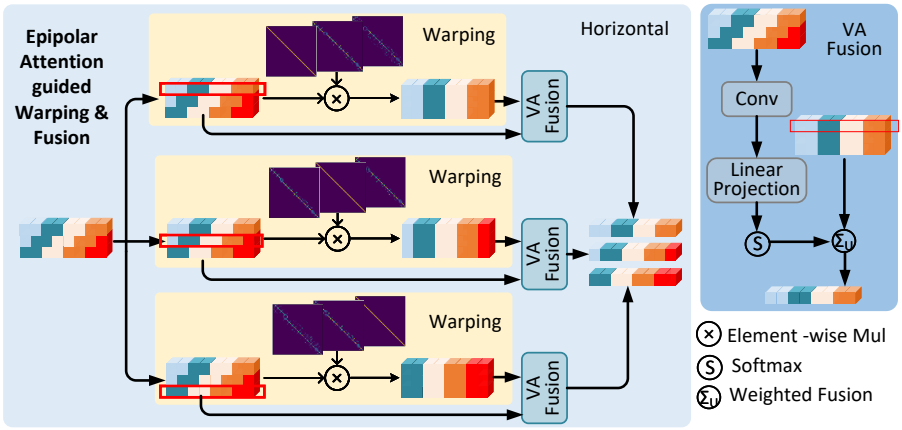



Figure 8: Detailed structure of epipolar attention-guided warping and view-adaptive fusion.

**Forward diffusion.** Following the standard diffusion formulation, Gaussian noise is gradually added to $\mathbf{c}_0$ over $T$ steps:

$$q(\mathbf{c}_T \mid \mathbf{c}_0) = \mathcal{N}\left(\mathbf{c}_T; \sqrt{\bar{\alpha}_T}\mathbf{c}_0, (1-\bar{\alpha}_T)\mathbf{I}\right), \tag{12}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$.

**Reverse denoising.** The reverse process progressively removes noise from $\mathbf{c}_T$ to recover the compact correspondence representation. The denoising process is conditioned on the initial epipolar attention $\mathbf{A}_h$, which is also transformed into the same representation space:

$$\mathbf{c}_{\text{cond}} = \mathcal{P}_{\text{cond}}(\mathbf{A}_h), \tag{13}$$

where $\mathcal{P}_{\text{cond}}(\cdot)$ denotes the condition projection operation.

At each reverse step, the transition is formulated as

$$p_\theta(\mathbf{c}_{t-1} \mid \mathbf{c}_t, \mathbf{c}_{\text{cond}}) = \mathcal{N}\left(\mathbf{c}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{c}_t, \mathbf{c}_{\text{cond}}, t), \boldsymbol{\Sigma}_t\right), \tag{14}$$

with

$$\boldsymbol{\mu}_\theta(\mathbf{c}_t, \mathbf{c}_{\text{cond}}, t) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{c}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta(\mathbf{c}_t, \mathbf{c}_{\text{cond}}, t)\right), \tag{15}$$

and

$$\boldsymbol{\Sigma}_t = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t\mathbf{I}. \tag{16}$$

Here, $\boldsymbol{\epsilon}_\theta(\cdot)$ denotes the denoising network, which predicts the noise in the compact correspondence space under the condition of the initial epipolar attention.

After $T$ denoising steps, we obtain the recovered compact correspondence representation $\hat{\mathbf{c}}$. It is then mapped back to the original attention-residual space through *Inverse Projection & Reshape*:

$$\hat{\mathbf{z}} = \mathcal{P}^{-1}(\hat{\mathbf{c}}), \tag{17}$$

where $\mathcal{P}^{-1}(\cdot)$ denotes the *Inverse Projection & Reshape* operation. Finally, the refined epipolar attention is obtained by adding the recovered residual correction to the initial attention:

$$\mathbf{A}_h^{ref} = \mathbf{A}_h + \hat{\mathbf{z}}. \tag{18}$$

The vertical epipolar attention $\mathbf{A}_v^{ref}$ is refined using the same procedure.

# C Additional Experimental Analysis

## C.1 Qualitative Comparison via Pixel-Wise Error Visualization

Figure 9 and Figure 10 show qualitative comparisons under the $\times 2$ and $\times 4$ SR settings, respectively, by visualizing the pixel-wise absolute reconstruction errors with respect to GT. It can be observed that

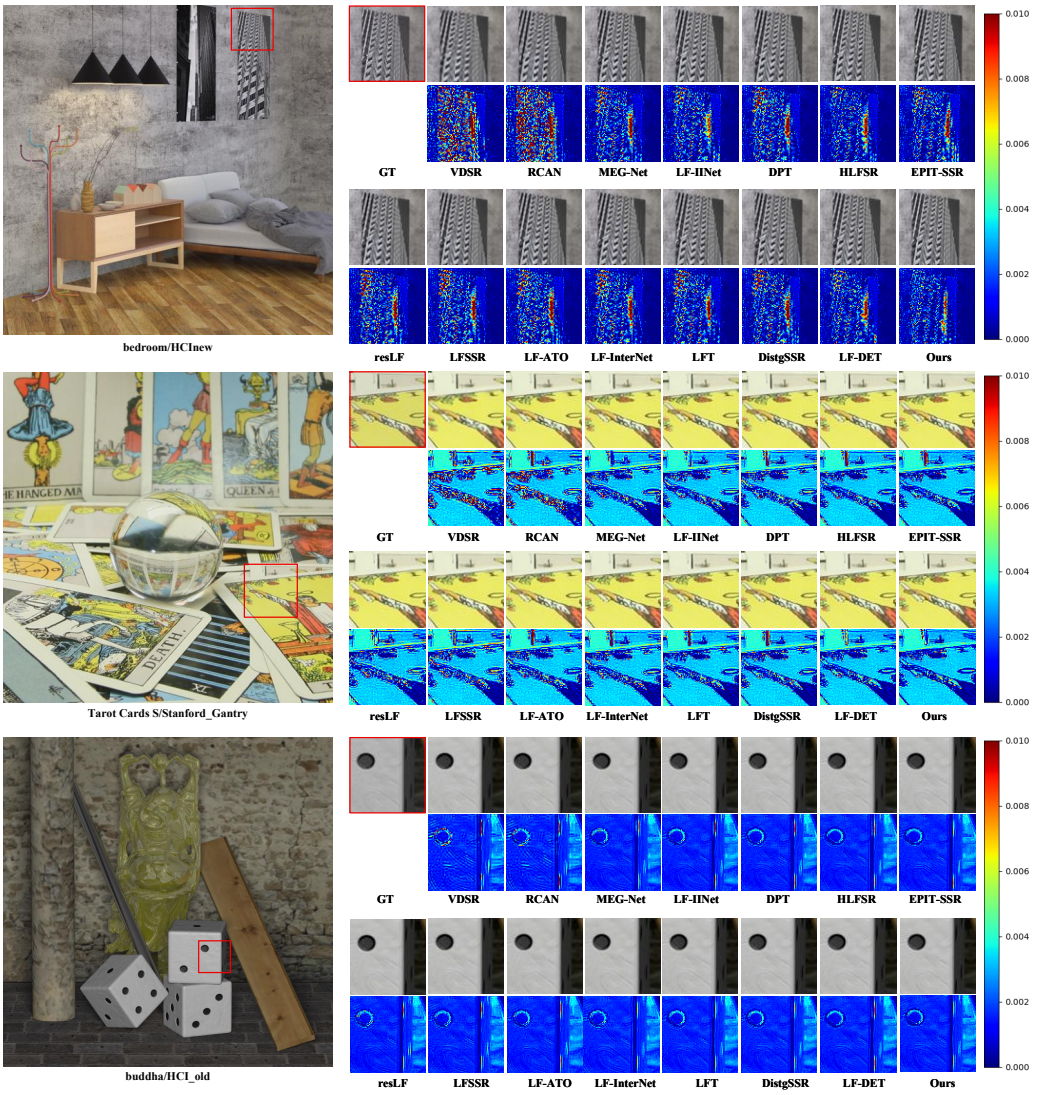



Figure 9: Pixel-wise absolute error visualization on the Y channel between reconstructed results and GT for different SR methods under the $\times 2$ setting. The error maps are visualized using a unified color scale within each scene, where warmer colors indicate larger reconstruction errors.

most existing methods suffer from relatively large prediction errors in regions with complex textures, which are visualized as pronounced high-error responses. In contrast, FEAD produces error maps with lower overall intensity and more spatially uniform distributions across all scenes. High-error responses around object boundaries and repetitive structures are effectively suppressed under both scaling factors, indicating more accurate reconstruction and improved geometric consistency.

### C.2 View-wise Reconstruction Consistency

To further evaluate the reconstruction performance across different angular perspectives, we conduct a per-view PSNR analysis on representative LF scenes under the $\times 2$ SR setting. For each reconstructed $5 \times 5$ LF, we compute the PSNR of all 25 SAIs and report both the average value and the standard deviation. As shown in Fig. 11, FEAD achieves consistently high PSNR values over almost all angular views, indicating that the reconstruction improvement is not limited to a few specific perspectives.

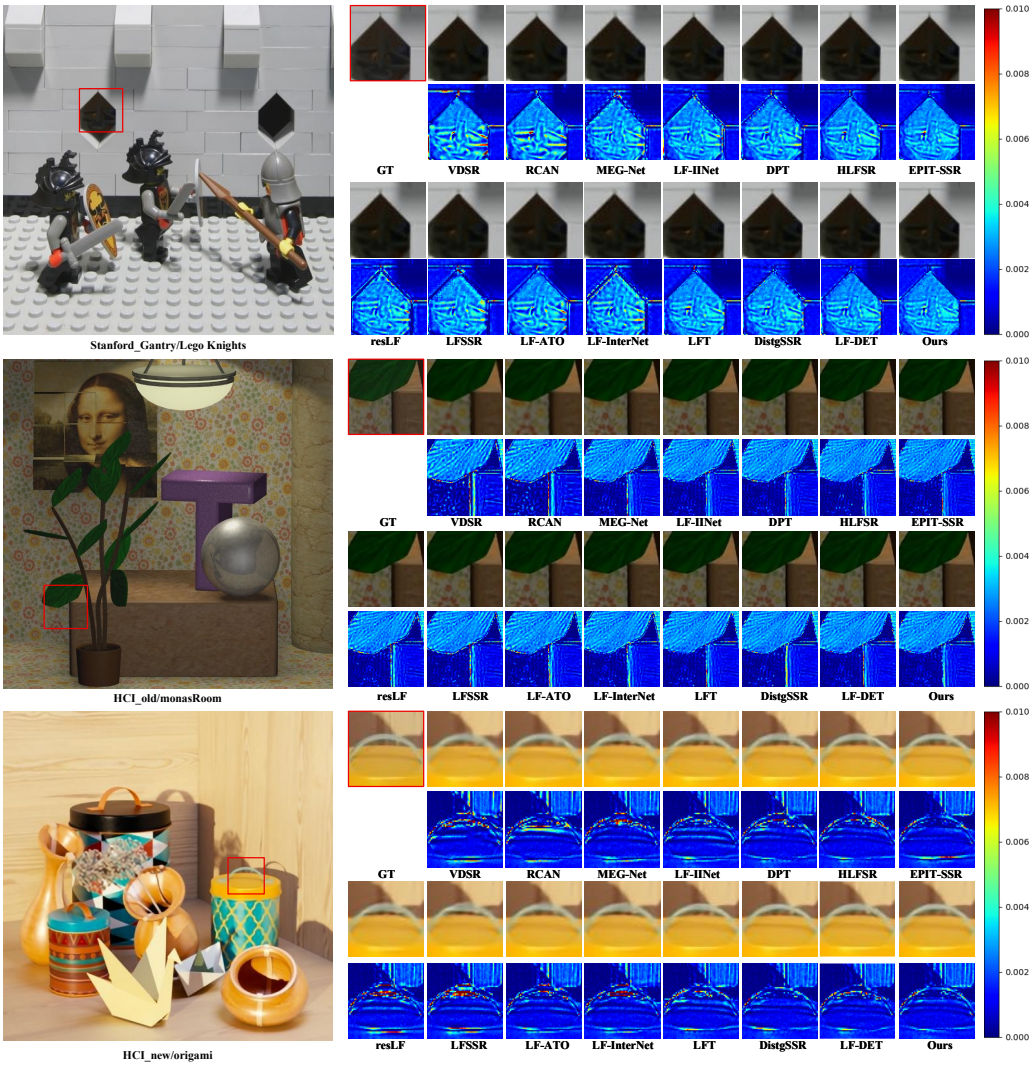



Figure 10: Pixel-wise absolute error visualization on the Y channel between reconstructed results and GT for different SR methods under the $\times 4$ setting. The error maps are visualized using a unified color scale within each scene, where warmer colors indicate larger reconstruction errors.

On the Bedroom scene, FEAD obtains the highest average PSNR of 36.95 dB, outperforming HLFSR, LF-DET, and EPIT-SSR by 0.20 dB, 0.31 dB, and 0.30 dB, respectively. More importantly, its per-view PSNR distribution remains uniformly high across the $5 \times 5$ angular grid, showing stable reconstruction quality from central to peripheral views. On the more challenging Tarot Cards S scene, FEAD further achieves an average PSNR of 39.13 dB, exceeding HLFSR, LF-DET, and EPIT-SSR by 1.21 dB, 0.35 dB, and 0.45 dB, respectively. Although EPIT-SSR shows a slightly lower standard deviation on some scenes, its per-view PSNR values are consistently lower than ours. These results demonstrate that FEAD improves the reconstruction quality for nearly all angular views, rather than benefiting only a small subset of SAIs, which confirms its effectiveness in producing view-wise reliable LF reconstruction.

### C.3 Attention Accuracy and Condition Analysis

To further evaluate whether the proposed refinement module improves the quality of epipolar correspondence, we measure the attention-level error between the predicted HR epipolar attention and the GT-derived HR epipolar attention. Specifically, we compute the mean squared error multiplied

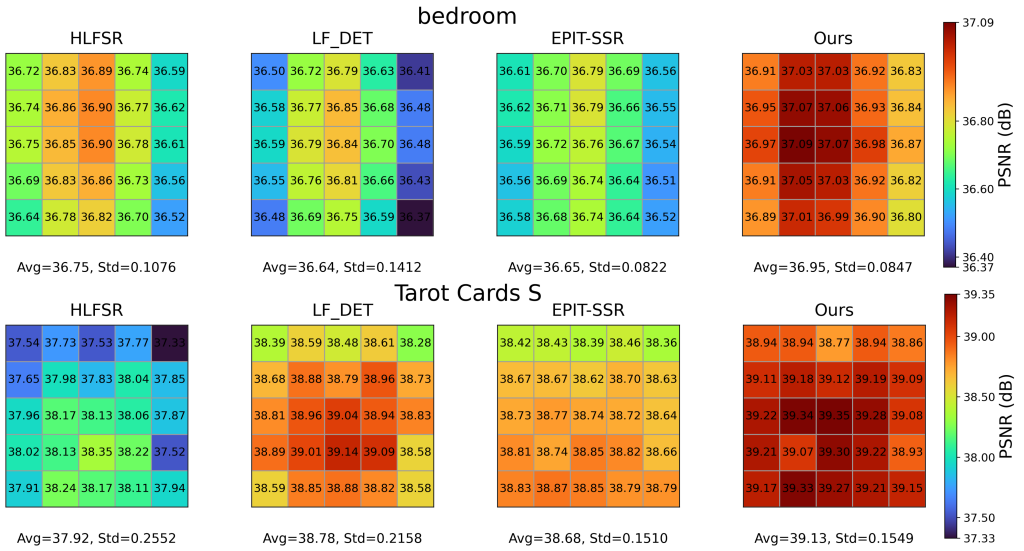



Figure 11: Per-view PSNR distributions under different angular perspectives for the $\times 2$ SR setting. The average value and standard deviation of each scene are also reported.

Table 4: Comparison of different attention refinement strategies on $\times 2$ LF-SSR. We report both reconstruction accuracy and attention accuracy. MSE$\times 100$ denotes the mean squared error between the predicted HR epipolar attention and the GT-derived HR epipolar attention.

| Method | Condition | Avg. $[-2, 2]$ | | Avg. $[-4, 4]$ | |
|---|---|---|---|---|---|
| | | PSNR / SSIM ↑ | MSE$\times 100$ ↓ | PSNR / SSIM ↑ | MSE$\times 100$ ↓ |
| No Refine | HR feature | 41.975 / 0.9874 | 1.2547 | 42.209 / 0.9942 | 1.3949 |
| Refined by 3D Conv | HR feature | 41.986 / 0.9873 | 0.9997 | 42.278 / 0.9948 | 1.1482 |
| Refined by Transformer | HR feature | 42.063 / 0.9882 | 0.7475 | 42.396 / 0.9959 | 0.8705 |
| Refined by Diffusion | HR feature | **42.214 / 0.9886** | **0.4196** | **42.494 / 0.9961** | **0.6286** |
| Refined by Diffusion | LR attention | 42.092 / 0.9884 | 0.7073 | 42.356 / 0.9957 | 0.9195 |

by $100$ between the refined attention and the GT-derived attention, and report the corresponding SR performance in Table 4. Lower attention MSE indicates more accurate correspondence estimation.

As shown in Table 4, all refinement strategies reduce the attention error compared with the unrefined baseline, confirming the necessity of explicitly improving the initially estimated epipolar attention. Among different refinement strategies, diffusion-based refinement achieves the lowest attention MSE and the best SR performance. Compared with the no-refinement baseline, it reduces the attention MSE$\times 100$ from $1.2547$ to $0.4196$ on Avg. $[-2, 2]$ and from $1.3949$ to $0.6286$ on Avg. $[-4, 4]$, while improving PSNR by $0.239$ dB and $0.285$ dB, respectively. This indicates that the proposed conditional residual diffusion module can effectively correct noisy and ambiguous correspondence patterns.

We also evaluate an alternative design that computes attention in the LR space and then upsamples it as the diffusion condition. This variant performs worse than using HR-space attention estimated from upsampled features. Specifically, using LR attention as the condition increases the attention MSE$\times 100$ from $0.4196$ to $0.7073$ on Avg. $[-2, 2]$ and from $0.6286$ to $0.9195$ on Avg. $[-4, 4]$, with corresponding PSNR drops from $42.214$ dB to $42.092$ dB and from $42.494$ dB to $42.356$ dB. These results support our design choice of first upsampling features and then estimating epipolar attention in the HR space, since directly interpolating LR attention can distort the pixel-wise correspondence distribution required for HR feature warping and fusion.

Table 5: Sensitivity analysis of attention-loss weights on $\times 2$ LF-SSR.

| $(\lambda_{\mathrm{warp}}, \lambda_{\mathrm{KL}})$ | (0.05, 0.05) | (0.05, 0.10) | (0.10, 0.05) | (0.10, 0.10) |
|---|---|---|---|---|
| Avg. $[-2, 2]$ | 42.12 / 0.9879 | 42.20 / 0.9880 | 42.16 / 0.9884 | **42.21 / 0.9886** |
| Avg. $[-4, 4]$ | 42.37 / 0.9957 | 42.45 / 0.9959 | 42.47 / 0.9960 | **42.49 / 0.9961** |

### C.4 Attention Loss Weights

We further analyze the sensitivity to the loss weights $(\lambda_{\mathrm{warp}}, \lambda_{\mathrm{KL}})$, as shown in Table 5. The results show that FEAD is relatively stable under different weight combinations. Assigning sufficient weights to both $\mathcal{L}_{\mathrm{warp}}$ and $\mathcal{L}_{\mathrm{KL}}$ consistently improves the reconstruction performance, indicating that both geometric consistency supervision and distribution-level regularization are beneficial for attention refinement. In particular, $(\lambda_{\mathrm{warp}}, \lambda_{\mathrm{KL}}) = (0.10, 0.10)$ achieves the highest PSNR/SSIM on both Avg. $[-2, 2]$ and Avg. $[-4, 4]$, and is therefore adopted as the default setting in FEAD.

## D Limitations

Although FEAD achieves clear advantages on medium- and large-disparity LF datasets, its performance gains are relatively smaller on low-disparity datasets such as EPFL and INRIA. This is because the pixel offsets between adjacent views are small in low-disparity scenes, and different SAIs are nearly aligned. Under such conditions, existing implicit spatial–angular feature interaction methods can already aggregate cross-view information effectively, making the additional benefit of explicit attention-guided warping less pronounced. In contrast, when the disparity range becomes larger, accurate cross-view correspondence modeling becomes more critical, and the proposed HR epipolar attention refinement and EAG warping bring more evident improvements.

## NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: **[TODO]**

   Justification: **[TODO]**

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: **[TODO]**

   Justification: **[TODO]**

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: **[TODO]**

    Justification: **[TODO]**

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: **[TODO]**

    Justification: **[TODO]**

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.