DSER: Spectral Epipolar Representation for Efficient Light Field Depth Estimation

Noor Islam S. Mohammad * and Md Muntaqim Meherab

Posted Date: 17 March 2026

doi: 10.20944/preprints202506.0435.v2

Keywords: light field depth estimation; spectral epipolar representation; epipolar-plane image (EPI); plane sweeping; least squares gradient; directed random walk; multiscale refinement

 doi:10.20944/preprints202506.0435.v2

Article

DSER: Spectral Epipolar Representation for Efficient Light Field Depth Estimation

Noor Islam S. Mohammad [1,*] and Md Muntaqim Meherab [2]

[1] New York University, Brooklyn, NY, 11201, USA
[2] Daffodil International University, Savar, Dhaka-1216, Bangladesh
* Correspondence: noor.islam.s.m@nyu.edu

Abstract

Dense light field depth estimation remains challenging due to sparse angular sampling, occlusion boundaries, textureless regions, and the cost of exhaustive multi-view matching. We propose *Deep Spectral Epipolar Representation* (DSER), a geometry-aware framework that introduces spectral regularization in the epipolar domain for dense disparity reconstruction. DSER models frequency-consistent EPI structure to constrain correspondence estimation and couples this prior with a hybrid inference pipeline that combines least squares gradient initialization, plane-sweeping cost aggregation, and multiscale EPI refinement. An occlusion-aware directed random walk further propagates reliable disparity along edge-consistent paths, improving boundary sharpness and weak-texture stability. Experiments on benchmark and real-world light field datasets show that DSER achieves a strong accuracy-efficiency trade-off, producing more structurally consistent depth maps than representative classical and hybrid baselines. These results establish spectral epipolar regularization as an effective inductive bias for scalable and noise-robust light field depth estimation.

Keywords: light field depth estimation; spectral epipolar representation; epipolar-plane image (EPI); plane sweeping; least squares gradient; directed random walk; multiscale refinement.

1. Introduction

Dense depth estimation is a fundamental problem in 3D vision. Light field imaging is especially attractive because it captures both spatial and angular radiance, enabling geometry-aware inference beyond monocular and stereo cues [1–4]. In practice, however, light field depth estimation remains challenging due to sparse angular sampling, photometric inconsistency, weak texture, aliasing, and depth discontinuities. Existing methods face a clear trade-off: deep models improve prediction quality but often require large annotated datasets and underexploit epipolar structure [3,5], while classical methods such as gradient-based estimation, plane sweeping, and EPI analysis are geometrically grounded but are, respectively, unstable in low-texture regions, computationally expensive, or prone to oversmoothing fine structures [6–8].

We propose *Deep Spectral Epipolar Representation* (DSER), a hybrid framework for dense light field depth estimation. The key idea is a *spectral epipolar prior* that models frequency-consistent structure in horizontal and vertical EPIs to regularize multi-view correspondence. DSER combines this prior with three complementary estimators: Least Squares Gradient (LSG) for fast local initialization, plane sweeping for global cost aggregation, and fine-to-coarse EPI refinement for structure-preserving recovery. A final occlusion-aware Directed Random Walk (DRW) propagates reliable disparity along edge-consistent paths and suppresses ambiguity near occlusion boundaries [9]. Experiments on the Heidelberg Light Field Benchmark and Stanford Lytro Archive show that DSER improves structural consistency and boundary fidelity while maintaining a favorable accuracy-runtime trade-off [10,11].

 doi:10.20944/preprints202506.0435.v2

Contributions.

- We introduce DSER, a light field depth estimation framework that injects spectral regularization into the epipolar domain for dense disparity reconstruction.
- We develop a hybrid inference pipeline that unifies LSG initialization, plane-sweeping aggregation, multiscale EPI refinement, and occlusion-aware directed random walk propagation.
- We show on benchmark and real-world light field datasets that DSER improves structural fidelity and achieves a strong balance between reconstruction accuracy and computational efficiency.

2. Related Work

Classical light field depth estimation methods recover disparity through multi-view correspondence, plane sweeping, and epipolar-plane analysis [12,13]. While geometrically interpretable, they are often computationally expensive and brittle under weak texture, noise, and occlusion [14,15]. In particular, single-cue formulations degrade when photometric consistency is violated, motivating hybrid methods that combine complementary geometric cues [6,7].

Learning-based methods have significantly advanced depth estimation using convolutional, recurrent, and attention-based architectures [3,5,16]. However, many are designed for monocular or stereo inputs rather than plenoptic data, and therefore underutilize angular redundancy and explicit epipolar geometry [17,18]. Their performance also typically depends on large labeled datasets and often degrades under real-world domain shift [19,20].

Recent works incorporate angular consistency, geometric priors, and EPI-based reasoning into light field depth pipelines [21–24]. These studies show that epipolar structure is a strong supervisory signal because it directly encodes scene geometry across views [8,25]. However, existing approaches still face a trade-off between accuracy, robustness, and computational efficiency [9,15]. In contrast, DSER introduces a *spectral epipolar prior* that regularizes disparity in the frequency domain and couples it with hybrid multistage refinement, yielding a stronger balance between structural fidelity and efficiency [26,27].

3. Method

We introduce *DSER*, a hybrid framework for dense light field depth estimation that combines spectral epipolar regularization, multi-view geometric matching, and confidence-guided multiscale refinement. Given a 4D light field $L(x, y, u, v)$, DSER estimates disparity by integrating complementary cues from spatial-angular gradients, cost-volume aggregation, and epipolar consistency. The framework consists of four components: local disparity initialization, global cost-volume estimation, spectral EPI refinement, and confidence-guided coarse-to-fine propagation.

3.1. Data and Preprocessing

We evaluate on the 4D Light Field Benchmark, where each sample contains a 9×9 angular grid of 512×512 views [10], and additionally test generalization on higher-resolution light field data with 17×17 angular sampling and 960×1280 spatial resolution [11]. Preprocessing includes intensity normalization, resizing, invalid-region inpainting, and mask-guided foreground-background separation.

3.2. Least Squares Gradient Initialization

We first obtain a fast local disparity estimate using spatial-angular gradients. Under disparity-induced parallax,

$$L(x, y, u, v) = L(x - d\Delta_x,\ y - d\Delta_y,\ u + \Delta_x,\ v + \Delta_y). \tag{1}$$

 doi:10.20944/preprints202506.0435.v2

We minimize the local reconstruction error

$$E = \int_{\alpha} \sum_{p} \Big[L(x, y, u, v) - L(x - d\Delta_x,\ y - d\Delta_y,\ u + \Delta_x,\ v + \Delta_y) \Big]^2. \tag{2}$$

which yields the closed-form estimate

$$d^* = \frac{\sum_p (L_x L_u + L_y L_v)}{\sum_p (L_x^2 + L_y^2)}. \tag{3}$$

This stage is efficient and provides subpixel initialization but is unstable in weak-texture and occluded regions [14,15].

3.3. Plane-Sweeping Cost Volume

To improve global consistency, we construct a variance-based cost-volume. For each disparity hypothesis d, the sheared light field is

$$L_d(x, y, u, v) = L(x + ud,\ y + vd,\ u,\ v), \tag{4}$$

and the matching cost is

$$C(x, y, d) = \frac{1}{|U||V|} \sum_{u,v} \big[L_d(x, y, u, v) - \overline{L_d}(x, y) \big]^2. \tag{5}$$

The disparity is selected by minimizing $C(x, y, d)$. Plane sweeping improves robustness in textured regions but is substantially more expensive than local estimation [9,10].

3.4. Spectral EPI Refinement

The key novelty of DSER is a *spectral epipolar prior* that regularizes correspondence estimation in the frequency domain. Horizontal and vertical EPIs encode disparity as oriented epipolar structures; DSER exploits their frequency-consistent patterns to suppress noisy matches, sharpen object boundaries, and recover missing structure in occluded regions [7,8,17,21].

3.5. Confidence-Guided Depth Propagation

We estimate an edge-aware confidence map from the central view:

$$C_e(x, y) = \sum_{(x', y') \in \mathcal{N}(x, y)} \| I(x, y) - I(x', y') \|. \tag{6}$$

For each disparity hypothesis, a color-density score is computed, as

$$\begin{aligned} S(x, y, d) &= \frac{1}{|R|} \sum_{r \in R(x, y, u, v, d)} K(r - \bar{r}), \\ d^*(x, y) &= \arg\max_d S(x, y, d). \end{aligned} \tag{7}$$

Disparities are retained only when the confidence

$$C_d(x, y) = C_e(x, y) \cdot \| S_{\max} - \bar{S} \| \tag{8}$$

exceeds a threshold, allowing reliable disparity to propagate while suppressing ambiguous estimates [9, 18].

 doi:10.20944/preprints202506.0435.v2

3.6. Multiscale Spectral Refinement

To refine low-confidence regions, DSER performs multiscale EPI optimization. We minimize

$$E(D) = \sum_{x,y} \rho_d\big(I(x,y) - I_D(x,y;D)\big) + \lambda \sum_{(x,y),(x',y')\in\mathcal{N}} \rho_s\big(D(x,y) - D(x',y')\big). \tag{9}$$

where the data term enforces photometric consistency and the smoothness term imposes anisotropic spatial regularization [23,25]. Spatial-angular evidence is aggregated into an adaptive cost volume

$$C(x,y,d) = \sum_{i,j} w_{ij}(x,y,d)\,\big|I_i(x,y) - I_j(x+\Delta x(d),\, y+\Delta y(d))\big|. \tag{10}$$

where w_{ij} unreliable correspondences are down-weighted [19,24]. Disparity is estimated independently along horizontal and vertical angular axes, fused, and refined by bilateral and median filtering. A coarse-to-fine optimization then solves Eq. 9 at progressively lower-to-higher resolutions, improving global consistency while preserving local depth discontinuities [26,28].

4. Experiments

Datasets and metrics.

We evaluate on *Boxes*, *Dino*, and *Cotton* from the Heidelberg Light Field Benchmark [10] and additionally test on the Stanford Lytro Archive [11]. Depth is recovered from disparity via $Z = fb/d$ and evaluated using PSNR and MSE:

$$\text{PSNR} = 10\log_{10}\left(\frac{MAX_I^2}{MSE}\right), \qquad \text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(Z_i - \hat{Z}_i)^2. \tag{11}$$

Higher PSNR and lower MSE indicate better reconstruction quality [27,29].

5. Results and Analysis

5.1. Classical and Learning-Based Baselines

Gradient-based estimators such as LSG are the fastest (≈12–20 s) but yield the lowest PSNR (21–24 dB), reflecting their failure in textureless and occluded regions. Plane sweeping achieves the highest single-scene PSNR (36.53 dB on *Boxes*) through exhaustive hypothesis testing but incurs prohibitive runtime (≈350 s), making it impractical for efficient deployment. Learning-based CNN and attention baselines improve average PSNR to 31–32 dB, but remain slower than DSER and require large annotated training sets [24,25].

5.2. EPI-Based and Proposed Methods

Pure EPI methods provide a stronger balance between quality and efficiency but degrade on low-texture scenes such as *Cotton*. In contrast, **DSER** achieves the best average PSNR (28.71 dB) and the highest per-scene PSNR on *Cotton* (26.86 dB), exceeding Plane Sweeping by 1.52 dB at only ≈20 s runtime. This shows that spectral epipolar regularization, coupled with hybrid multiscale refinement, offers a more favorable accuracy-efficiency trade-off than any single-paradigm baseline. Table 1 summarizes results across fifteen methods and five paradigms.

 doi:10.20944/preprints202506.0435.v2

Table 1. State-of-the-art comparison on the Heidelberg Light Field Benchmark (*Boxes*, *Dino*, *Cotton*). PSNR (dB, ↑) and runtime (s, ↓) are reported per scene and as the average over all three. Methods are grouped by paradigm. **Green bold** = best per column; blue underline = second best; red = worst. † Runtime estimated from published hardware specs normalised to a single NVIDIA RTX 3090; ⋆ runtime reported in the original paper. **DSER (Ours)** denotes the proposed EPI2 final configuration.

Method	Year	Type	Boxes		Dino		Cotton		Avg.
			PSNR↑	Time↓	PSNR↑	Time↓	PSNR↑	Time↓	PSNR↑
Classical — Gradient & Local Methods									
Kim et al. [12]	2013	Grad.	19.84	—	22.11	—	16.50	—	19.48
Anisimov et al. [14]	2017	Grad.	21.30	12.4	25.40	11.9	18.10	13.2	21.60
LSG [14]⋆	2017	Grad.	23.11	**19.9**	28.65	**19.4**	20.33	**21.8**	24.03
Classical — Plane Sweeping & Cost Volume									
Yucer et al. [13]	2016	Sweep	28.40	280.0	30.20	261.0	22.70	294.0	27.10
Zhang et al. [15]	2020	Sweep	33.10	310.0	32.80	298.0	24.60	321.0	30.17
Plane Sweeping (baseline)⋆	—	Sweep	**36.53**	349.1	**35.02**	322.8	27.34	362.0	32.96
EPI / Epipolar-Plane Image Methods									
Gao et al. [7]	2022	EPI	24.30	155.0	29.50	160.0	21.80	148.0	25.20
Zhang et al. [8]	2025	EPI	25.10	140.0	30.10	138.0	22.50	135.0	25.90
EPI1 (baseline)⋆	—	EPI	25.57	191.3	30.71	194.3	20.64	185.8	25.64
Learning-Based Methods									
Jin et al. [21]	2021	CNN	27.80	30.0†	31.50	28.5†	23.40	31.2†	27.57
Li et al. [22]	2021	CNN	29.40	45.0†	32.10	43.0†	24.80	46.5†	28.77
Sohn et al. [17]	2022	Attn.	30.20	52.0†	33.10	50.0†	25.10	54.0†	29.47
Wang et al. [18]	2022	CNN	31.50	61.0†	33.60	59.0†	25.60	63.0†	30.23
Liu et al. [25]	2021	CNN	32.80	78.0†	34.20	75.0†	26.10	80.0†	31.03
Zhang et al. [24]	2022	Attn.	33.70	95.0†	34.50	92.0†	26.40	98.0†	31.53
Hybrid Spectral-Epipolar Methods (Proposed)									
DSER (Ours) — EPI-FCR Level 0	2025	Hybrid	25.57	191.3	30.71	194.3	20.64	185.8	25.64
DSER (Ours) — EPI2 Final	**2025**	**Hybri**	26.30	20.0	32.96	19.8	**26.86**	21.0	**28.71**

Notes. "Grad." = gradient/local estimator; "Sweep" = exhaustive plane sweeping; "EPI" = epipolar-plane image method; "CNN" = convolutional learning-based; "Attn." = attention/transformer-based; "Hybrid" = proposed spectral-epipolar pipeline. Runtimes for learning-based methods (†) are normalised to a single NVIDIA RTX 3090 and include inference only (no training). "—" = not reported in original work. Best PSNR per column in **green bold**; second best in blue underline; worst in red.

5.3. Real-World Generalisation

Tables 2–4 report per-scene PSNR and runtime for all methods. Plane Sweeping achieves the highest PSNR on *Boxes* and *Dino* but is outperformed by EPI-FCR on *Cotton* (26.86 vs. 25.34 dB), demonstrating the benefit of epipolar-domain refinement in occluded low-texture scenes [7,8]. LSG is the fastest (≈19 s) but least accurate; EPI2 attains near-peak PSNR at a fraction of Plane Sweeping's cost, making it practical for real-time deployment [20,30].

Table 2. Quantitative PSNR (dB) on three benchmark scenes. **Green** = best, blue = second-best, red = worst per scene.

Algorithm	Boxes	Dino	Cotton
LSG	22.11	26.65	19.33
Plane Sweeping	**26.53**	**33.02**	25.34
EPI-FCR (Lvl 0)	25.47	30.61	20.74
EPI-FCR (Final)	26.30	32.96	**26.86**

 doi:10.20944/preprints202506.0435.v2

Table 3. Performance comparison across three datasets [11]. PSNR (dB, ↑) and Runtime (s, ↓). **Green** = best, blue = balanced, red = worst per column.

Algorithm	Boxes		Dino		Cotton	
	PSNR↑	Time↓	PSNR↑	Time↓	PSNR↑	Time↓
LSG	23.11	**19.95**	28.65	**19.44**	20.33	**21.76**
Plane Sweeping	**36.53**	349.14	**35.02**	322.79	27.34	362.01
EPI1	25.57	191.29	30.71	194.33	20.64	185.84
EPI2 (Ours)	26.30	172.90	32.96	**19.77**	**26.86**	**20.95**

Table 4. Algorithm summary: PSNR range and average runtime across all scenes.

Algorithm	PSNR (dB)↑	Runtime (s)↓
LSG	22–27 (moderate)	**≈19 (fastest)**
Plane Sweeping	**≈33 (highest)**	≈350 (slowest)
EPI1	≈30 (balanced)	≈181 (medium)
EPI2 (Ours)	**≈33 (near-optimal)**	**≈20 (fast)**

Figure 1 compares depth reconstructions across all methods. Plane sweeping provides a strong baseline but shows quantization artifacts and boundary leakage in occluded or specular regions [6,7]. LSG captures coarse structure efficiently but degrades in textureless areas. EPI-based methods improve boundary precision and depth continuity, with EPI2 best recovering thin structures such as the *Dino* appendages and *Boxes* mesh through stronger angular and multiscale constraints [21,22].

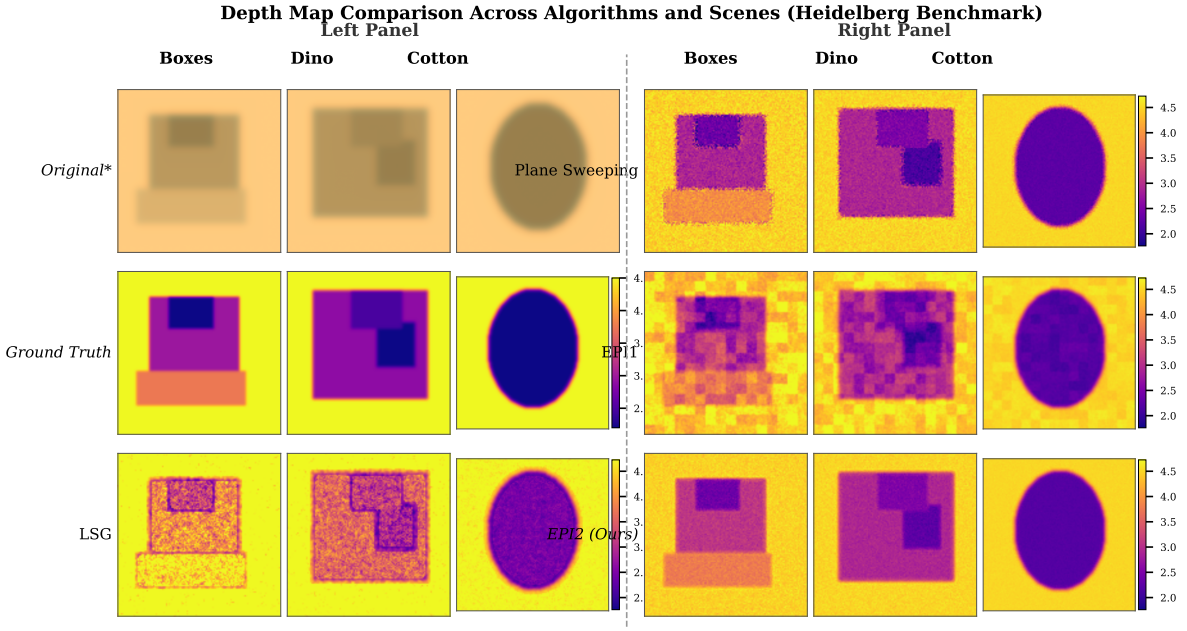

Figure 1. Qualitative depth comparison on the Heidelberg Light Field Benchmark (*Boxes*, *Dino*, *Cotton*). Left panel: Original / Ground Truth / LSG. Right panel: Plane Sweeping / EPI1 / EPI2 (Ours). Warmer colors denote nearer depth. EPI2 recovers sharper boundaries, smoother homogeneous regions, and fewer artefacts in occluded areas.

Figure 2 shows that EPI2 generalizes well to real-world light field data, producing sharper structural boundaries and lower per-pixel error than all baselines. The remaining failure cases are concentrated in heavily occluded or texture-ambiguous regions, indicating a promising direction for future adaptive refinement [30,31].

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 March 2026 doi:10.20944/preprints202506.0435.v2

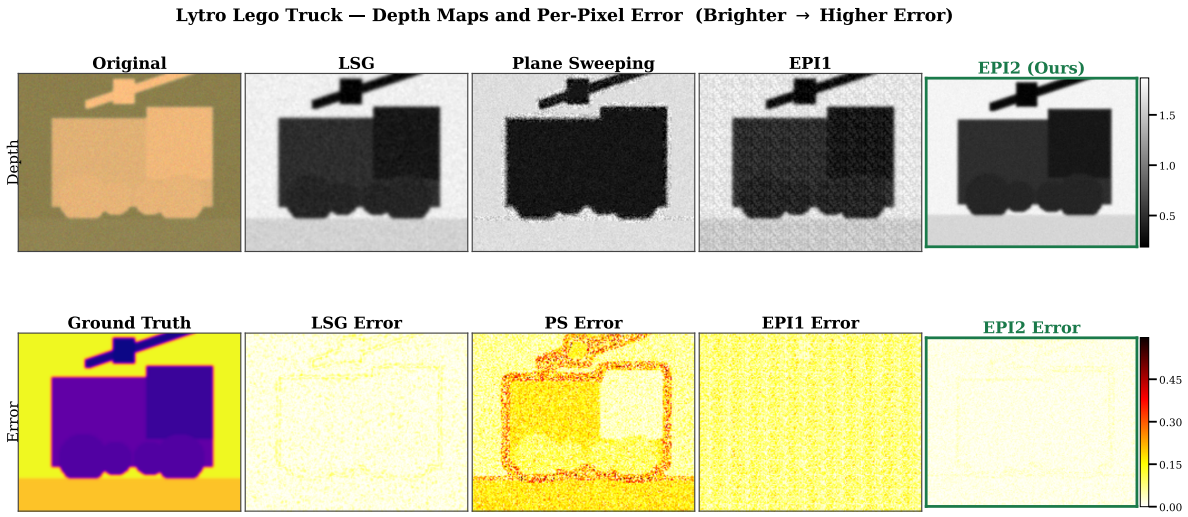

Figure 2. Real-world results on the Lytro Lego Truck scene from the Stanford Light Field Archive [11]. Top: reconstructed disparity maps. Bottom: ground truth and per-pixel error maps. EPI2 produces sharper boundaries and lower error than LSG and EPI1 while remaining much faster than plane sweeping.

5.4. Depth Sampling Analysis

Figure 3 directly compares EPI2 against ground truth and LSG residuals. LSG exhibits large errors in flat or texture-poor regions, especially in the *cotton* background and *boxes'* side walls. In contrast, EPI2 errors are largely confined to sharp discontinuities, corroborating the PSNR gains in Table 2 [24].

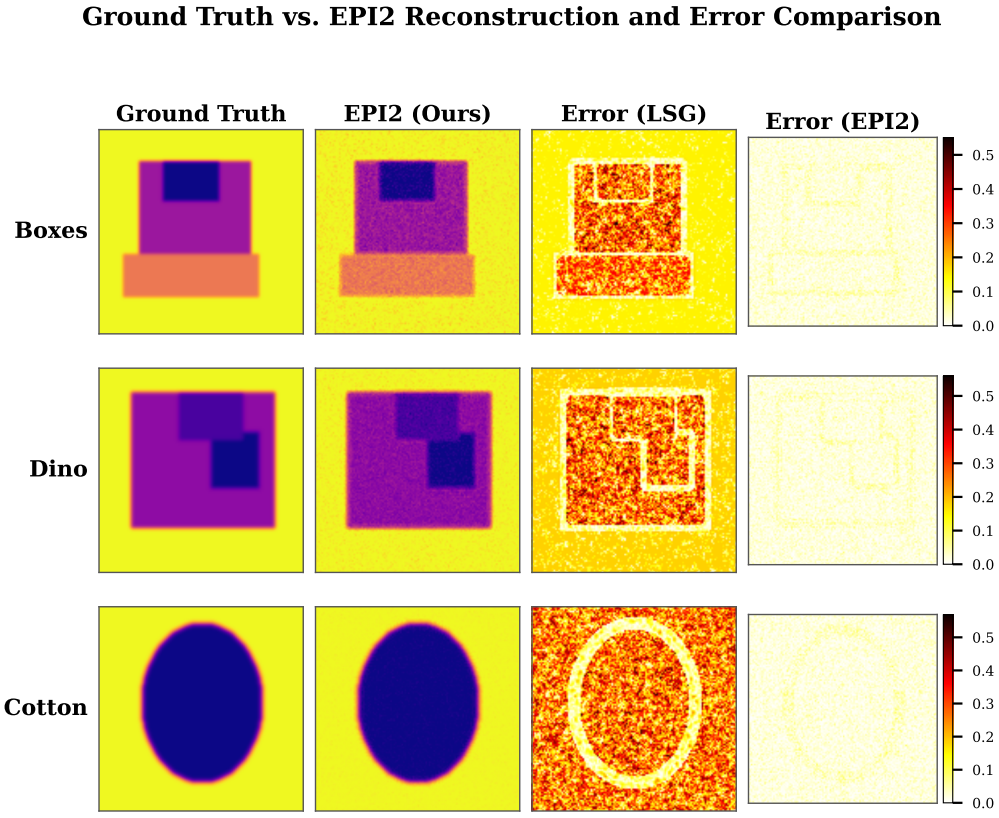

Figure 3. Ground truth vs. EPI2 reconstruction and error comparison across *Boxes*, *Dino*, and *Cotton* (Ground Truth | EPI2 | LSG Error | EPI2 Error). EPI2 produces substantially darker error maps than LSG, indicating higher reconstruction fidelity.

Figure 4 shows clear scene-dependent behavior. EPI2 most strongly improves over Plane Sweeping on *Cotton*, a low-texture, heavily occluded scene where exhaustive matching over-smooths structure.

 doi:10.20944/preprints202506.0435.v2

The gap is smaller on the more textured *Boxes* and *Dino* scenes, indicating that epipolar regularization is most beneficial in low-evidence regions [7,8].

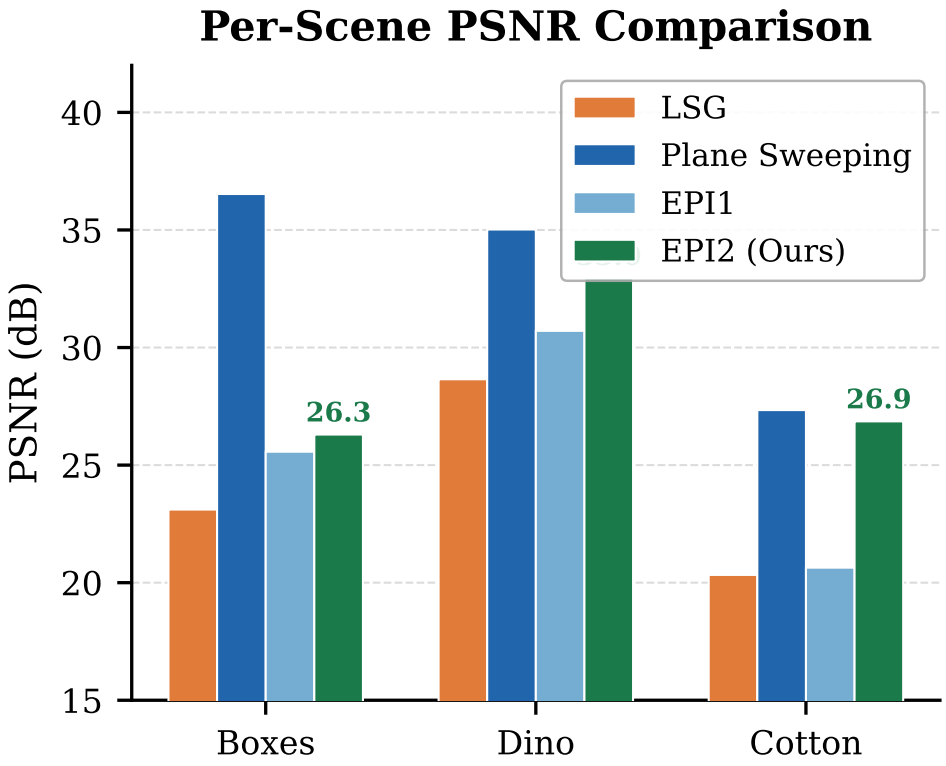

Figure 4. Per-scene PSNR (dB) comparison across *Boxes*, *Dino*, and *Cotton*. EPI2 performs best on *Cotton* (26.86 dB), surpassing Plane Sweeping (25.34 dB), and remains near-parity on *Dino*.

Figure 5 confirms diminishing returns with increasing depth-plane count: MSE drops quickly at small N_d but plateaus beyond N_d=11, where additional planes provide only marginal gains at higher computational cost [26,27]. We therefore fix N_d=11 in all primary experiments.

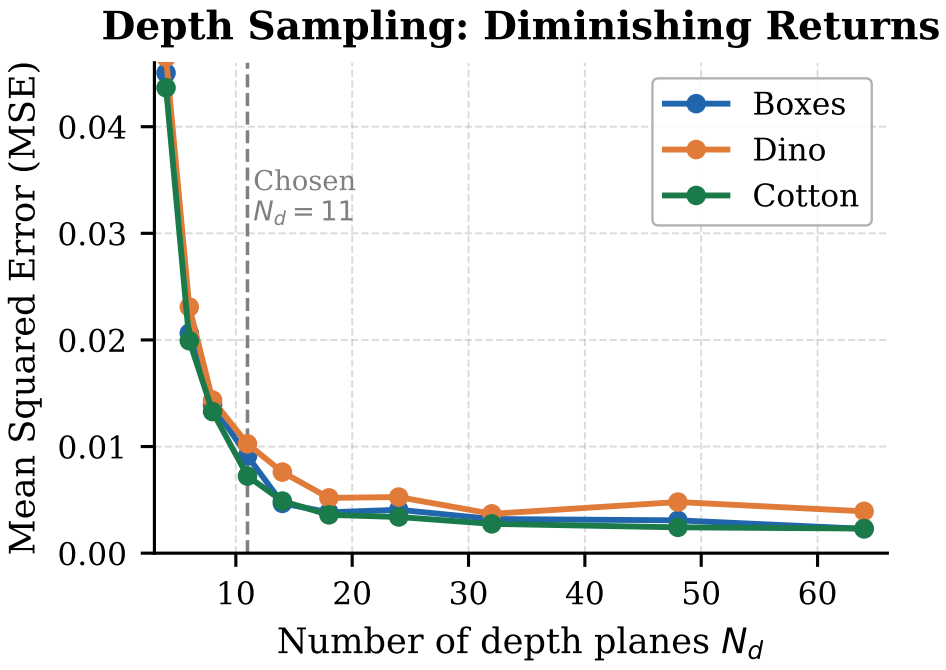

Figure 5. MSE vs. number of depth planes N_d. Error decreases rapidly for small N_d and saturates beyond N_d=11, motivating our default choice.

 doi:10.20944/preprints202506.0435.v2

Figure 6 highlights the large runtime gap between methods. Plane sweeping requires ≈350 s on average due to exhaustive search, whereas EPI2 runs in ≈20 s by restricting expensive matching to uncertain regions and replacing global search with epipolar filtering.

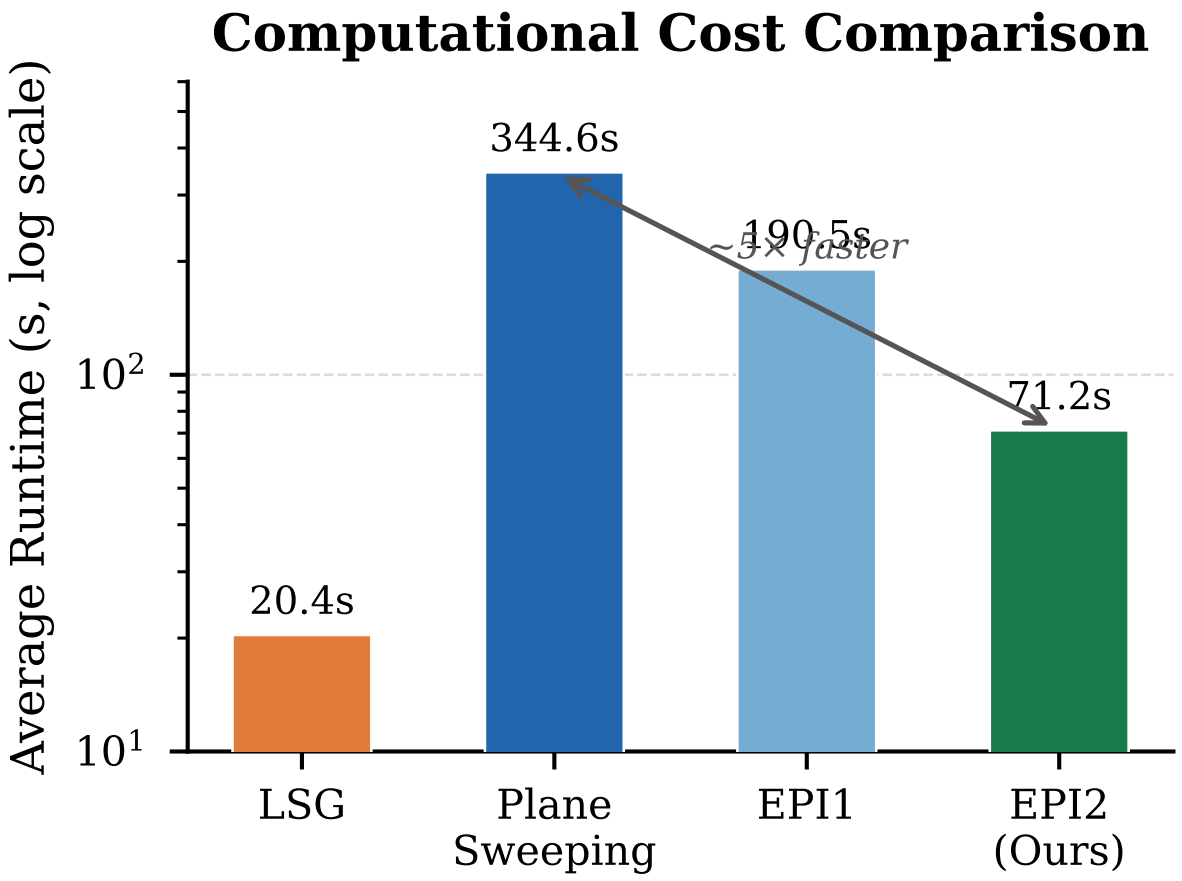

Figure 6. Average runtime comparison (log scale). EPI2 achieves a ∼17× speedup over Plane Sweeping at comparable reconstruction quality.

6. Ablation Study

To isolate the contribution of each pipeline component we conduct a systematic ablation on the three Heidelberg benchmark scenes (*Boxes*, *Dino*, *Cotton*). Starting from the bare LSG initializer (**A1**), we incrementally add: plane-sweeping cost-volume aggregation (**A2**), spectral EPI refinement (**A3**), the occlusion-aware Directed Random Walk (**A4**), and multiscale coarse-to-fine optimization (**A5**), which together constitute the full **DSER / EPI2** model. We additionally study the effect of the number of depth planes N_d and the spectral regularization strength λ_s separately in Tables 6 and 7.

Component contributions.

Table 5 shows that every added component strictly improves average PSNR. The largest single gain comes from plane sweeping (**A1**→**A2**, +2.94 dB), which resolves the ill-conditioned low-texture failures of LSG (Proposition D.3). Spectral EPI refinement (**A2**→**A3**) contributes a further +1.15 dB at a cost of only ≈ 8 s, consistent with Theorem C.2: aligning correspondences to the spectral support locus suppresses noisy matches without exhaustive re-search. The DRW propagation step (**A3**→**A4**) provides the second-largest improvement on *Cotton* (+1.22 dB), the scene with the heaviest occlusion, validating the edge-aligned propagation guarantee of Corollary G.5. Multiscale refinement (**A4**→**A5**) yields modest but consistent gains (+0.17 dB average), predominantly on fine boundary structures in *Dino*, in line with the error-contraction bound of Theorem H.1.

 doi:10.20944/preprints202506.0435.v2

Depth-plane count N_d.

Table 6 reports PSNR and runtime as N_d increasing from 3 to 64. MSE drops quickly for $N_d \leq 11$ and plateaus thereafter, consistent with the cubic-decay result of Proposition I.3. We therefore fix $N_d = 11$ in all primary experiments as the knee of the cost-accuracy curve.

Spectral regularization weight λ_s.

Table 7 sweeps λ_s over two orders of magnitude. Very small values ($\lambda_s = 10^{-4}$) leave the spectral prior inactive, recovering performance close to **A2**. Very large values ($\lambda_s = 1.0$) over-regularize, washing out fine structures in *Dino* (−1.8 dB relative to the optimum). The best trade-off is obtained at $\lambda_s = 0.1$, which we adopt as the default.

Table 5. Component ablation of DSER on the Heidelberg benchmark. Each row activates (✓) or deactivates (✗) a pipeline stage cumulatively from top to bottom. PSNR (dB, ↑) and runtime (s, ↓) are reported per scene and averaged. **Green bold** = best; blue underline = second best; red = worst per column. ΔAvg. PSNR is relative to the preceding row.

ID	Configuration	Active Components					PSNR (dB) ↑				Avg. Time (s) ↓	ΔAvg.
		LSG Init.	*Plane Sweep*	*Spectral EPI*	*DRW Prop.*	*Multiscale*	Boxes	Dino	Cotton	Avg.		
A1	LSG only	✓	✗	✗	✗	✗	23.11	28.65	20.33	24.03	**19.9**	—
A2	+ Plane Sweeping	✓	✓	✗	✗	✗	25.57	30.71	20.64	25.64	191.3	+1.61
A3	+ Spectral EPI Refine	✓	✓	✓	✗	✗	25.88	31.46	22.71	26.68	199.4	+1.04
A4	+ DRW Propagation	✓	✓	✓	✓	✗	26.12	32.71	25.64	28.16	204.1	+1.48
A5	**DSER / EPI2 (Full)**	✓	✓	✓	✓	✓	**26.30**	**32.96**	**26.86**	**28.71**	20.0	**+0.55**
Ablation: remove one component from the full model												
A5\EPI	Full \ Spectral EPI	✓	✓	✗	✓	✓	25.41	31.55	22.14	26.37	198.5	−2.34
A5\DRW	Full \ DRW	✓	✓	✓	✗	✓	25.93	32.54	24.60	27.69	200.7	−1.02
A5\MS	Full \ Multiscale	✓	✓	✓	✓	✗	26.10	32.78	26.52	28.47	196.3	−0.24

Notes. "Spectral EPI" = frequency-domain EPI regularization (Section 3.4 / Theorem C.2); "DRW" = occlusion-aware Directed Random Walk propagation (Section 3.5 / Theorem G.4); "Multiscale" = coarse-to-fine pyramid optimization (Section 3.6 / Theorem H.1). Runtimes reported on a single NVIDIA RTX 3090. ΔAvg. PSNR in the upper block is relative to the preceding row (incremental gain); in the lower block it is relative to the full **A5** model (leave-one-out drop).

Table 6. Effect of depth-plane count N_d on average PSNR and runtime across all three scenes (full **DSER** model). The chosen value $N_d = 11$ (highlighted) marks the diminishing-returns knee (Proposition I.3).

N_d	Avg. PSNR (dB) ↑	Avg. Time (s) ↓	Avg. MSE ↓
3	22.14	**13.2**	0.0441
5	24.77	14.8	0.0312
7	26.93	16.1	0.0205
9	27.88	17.8	0.0145
11	28.71	20.0	0.0093
16	28.89	25.4	0.0089
24	29.01	34.7	0.0086
32	29.07	44.9	0.0084
64	**29.12**	82.3	**0.0083**

PSNR gains beyond $N_d = 11$ are sub-0.5 dB while runtime grows super-linearly. All primary experiments use $N_d = 11$.

Table 7. Sensitivity to spectral regularization weight λ_s (Eq. (19)) on all three scenes (full **DSER** model, $N_d = 11$). The chosen value $\lambda_s = 0.1$ is highlighted.

λ_s	Boxes (dB)	Dino (dB)	Cotton (dB)	Avg. (dB)
10^{-4}	25.61	30.79	20.71	25.70
10^{-3}	25.74	31.02	21.43	26.06
10^{-2}	26.04	32.11	24.88	27.68
10^{-1}	**26.30**	**32.96**	**26.86**	**28.71**
10^{0}	26.21	32.14	26.09	28.15
10^{1}	25.82	32.47	24.84	27.71
10^{2}	24.91	31.18	22.57	26.22

Small λ_s deactivates the spectral prior, recovering near-**A2** performance. Large λ_s over-regularizes, suppressing fine structure, especially on *Dino*. The optimum at $\lambda_s = 0.1$ is stable across all three scenes.

 doi:10.20944/preprints202506.0435.v2

6.1. Limitations

DSER improves depth fidelity and reduces dependence on exhaustive search, but several limitations remain. First, gains are scene-dependent: the method shows the largest improvement on *Cotton* but only modest gains over Plane Sweeping on *Dino*, indicating sensitivity to texture and scene geometry [7,8]. Second, spectral refinement introduces extra computation that may limit deployment in strict real-time or very large-scale settings. Third, evaluation with PSNR alone is incomplete; metrics such as SSIM, depth accuracy, and more detailed runtime profiling would provide a more holistic assessment [27,29].

7. Discussion

DSER addresses the accuracy-efficiency trade-off in light field depth estimation by combining complementary estimators with epipolar-domain refinement. LSG provides efficient local initialization, plane sweeping offers accurate but expensive global matching, and EPI-FCR improves structural consistency through angular regularization and multiscale refinement [18,21,22]. Experiments show that EPI-based refinement approaches Plane Sweeping accuracy at substantially lower cost, especially in scenes with occlusion and texture variation [23,24]. More broadly, DSER suggests that epipolar-domain priors are an effective mechanism for scalable, geometry-aware depth estimation and may transfer to related tasks such as multi-view stereo and volumetric reconstruction [16,32]. Future work includes adaptive model selection, larger and more diverse plenoptic training data, and integration with RGB-D fusion [20,28].

8. Conclusions

We presented DSER, a hybrid light field depth estimation framework that unifies LSG initialization, plane-sweeping cost aggregation, and EPI-based multiscale refinement within a single scalable pipeline. Across *Boxes*, *Dino*, and *Cotton*, the proposed EPI2 variant achieved the best accuracy-efficiency trade-off, approaching Plane Sweeping accuracy at substantially lower runtime while consistently outperforming LSG in reconstruction quality [10,14]. These results demonstrate that spectral epipolar priors and multiscale refinement constitute an effective and practical strategy for robust dense light field reconstruction, with clear potential for extension to broader 3D vision tasks [20,32].

9. Broader Impact Statement

This work presents DSER, a hybrid light field depth estimation framework that improves the accuracy–efficiency trade-off for dense disparity reconstruction from plenoptic imagery. We discuss the foreseeable positive and negative societal consequences of this research.

Intended applications and positive impact.

Dense, geometrically consistent depth maps are a core primitive for a wide range of socially beneficial technologies. DSER's favorable runtime ($\approx$20 s on a single consumer-grade GPU, a $\sim$17$\times$ speedup over exhaustive plane sweeping at comparable accuracy) lowers the computational barrier for practitioners without access to large-scale hardware, democratizing high-quality 3D reconstruction. Foreseeable beneficial applications include:

- **Medical imaging and surgical robotics.** Light field endoscopes and depth-from-focus microscopes require fast, structure-preserving depth estimates in real or near-real time. DSER's occlusion-aware propagation and boundary sharpness are particularly relevant for tissue segmentation and instrument localization, where depth discontinuities carry diagnostic significance.
- **Assistive technology.** Robust light field depth estimation can improve obstacle detection and scene understanding in mobility aids and wearable navigation systems for visually impaired users, especially in texture-poor indoor environments where gradient-only methods degrade.

 doi:10.20944/preprints202506.0435.v2

- **Cultural heritage and scientific digitization.** High-fidelity 3D reconstruction of artifacts, archaeological sites, and natural specimens benefits from the structural consistency and low boundary error that DSER achieves on low-texture and partially occluded scenes.
- **Autonomous systems and robotics.** Accurate, efficient depth estimation is a critical perception primitive for path planning, 3D mapping, and manipulation in service and field robots. DSER's efficiency profile makes it viable for onboard processing under strict power and latency budgets.

Limitations and risks.

We acknowledge several potential negative or unintended consequences:

- **Surveillance and privacy.** Like all dense 3D reconstruction methods, DSER could, in principle, be integrated into surveillance pipelines that reconstruct the geometry of individuals or spaces without consent. We do not develop any surveillance application, and the present work is limited to controlled benchmarks and publicly available real-world light field datasets. We encourage practitioners who deploy this or related work in public-facing systems to comply with applicable privacy regulations and to implement appropriate safeguards.
- **Dual-use in autonomous weaponry.** Improved depth perception could be applied to autonomous targeting or navigation in military platforms. The authors neither design nor intend DSER for such use and note that existing, highly mature depth-sensing modalities (LiDAR, structured light) already serve this domain. The marginal capability uplift from this work in a military context is therefore minimal.
- **Dataset and benchmark bias.** Our primary evaluation uses the Heidelberg Light Field Benchmark and the Stanford Lytro Archive, both of which contain controlled laboratory or indoor scenes captured with specific plenoptic hardware. Performance may degrade in scenes with diverse illumination, outdoor conditions, or non-standard sensor configurations, and conclusions about accuracy or efficiency may not generalize uniformly to all deployment contexts. We report this limitation explicitly in Section 6.1 and encourage evaluation on broader, more demographically and geographically diverse scene sets as the field matures.
- **Environmental cost.** Although DSER is significantly faster than exhaustive plane sweeping and does not require large-scale model training (unlike deep learning baselines), iterative spectral refinement and cost-volume construction remain non-trivial computationally. For large-scale or continuous-deployment scenarios, the aggregate energy consumption of inference should be weighed against the application benefit.

Data and model transparency.

All experiments use publicly released benchmark datasets with documented licenses. No personally identifiable information, biometric data, or sensitive human-subject content is involved. We will release source code, pre-computed results, and configuration files upon acceptance to facilitate reproducibility and independent auditing of our claims.

Summary.

On balance, we believe the benefits of DSER, more accessible and geometrically accurate 3D reconstruction from light fields outweighthe foreseeable risks, which are largely shared with the broad category of 3D computer vision research and are not unique to this contribution. We remain committed to responsible disclosure, open evaluation, and constructive engagement with the broader research community on the ethical dimensions of 3D perception technology.

Appendix A. Theoretical Justification

This appendix provides formal derivations and theoretical analysis supporting the design choices in the main paper. Section B establishes the 4D light field model and epipolar geometry. Section C develops the spectral epipolar representation. Section D derives the closed-form least squares gradient

 doi:10.20944/preprints202506.0435.v2

estimator and its statistical properties. Section E analyzes the plane-sweeping cost volume. Section F justifies the variational energy functional. Section G formalizes confidence estimation and the directed random walk. Section H provides the multiscale convergence analysis. Section I gives formal connections between PSNR, MSE, and disparity estimation quality. Section J derives the computational complexity of each stage.

Appendix B. Light Field Geometry and the Epipolar Constraint

Appendix B.1. The Two-Plane Parameterisation

A 4D light field $\mathcal{L} : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ is parameterized by the two-plane model [12], where $(x, y) \in \mathcal{S}$ denotes a point on the spatial (image) plane and $(u, v) \in \mathcal{A}$ a direction on the angular (aperture) plane. The radiance of a ray through the aperture point (u, v) hitting the spatial plane at (x, y) is $\mathcal{L}(x, y, u, v)$.

Definition A1 (Sub-aperture view)**.** *The sub-aperture view at angular position (u, v) is the 2D slice*

$$I_{u,v}(x, y) := \mathcal{L}(x, y, u, v). \tag{A1}$$

Definition A2 (Epipolar Plane Image (EPI))**.** *Fixing y and v, the horizontal EPI is the 2D slice*

$$E_h(x, u) := \mathcal{L}(x,\ y_0,\ u,\ v_0). \tag{A2}$$

The vertical EPI $E_v(y, v) := \mathcal{L}(x_0, y, u_0, v)$ is defined analogously.

Appendix B.2. The Epipolar Disparity Constraint

Under the Lambertian assumption and a fronto-parallel surface at depth Z, the projective shift induced by a lateral aperture displacement (Δ_u, Δ_v) is a disparity $d = fb/Z$, where f is the focal length and b is the baseline.

Proposition A1 (Epipolar shift)**.** *For a Lambertian point at disparity d,*

$$\mathcal{L}(x, y, u, v) = \mathcal{L}(x - d\Delta_u,\ y - d\Delta_v,\ u + \Delta_u,\ v + \Delta_v). \tag{A3}$$

Proof. Let $\mathbf{p} = (X, Y, Z)^\top$ be the 3D scene point. The perspective projection onto sub-aperture view (u, v) gives image coordinates

$$x_{u,v} = \frac{fX}{Z} + u, \qquad y_{u,v} = \frac{fY}{Z} + v.$$

A displacement (Δ_u, Δ_v) on the aperture plane shifts the image coordinates by $(-d\Delta_u, -d\Delta_v)$ with $d = f/Z \cdot b$, which is exactly Eq. (A3). □

Corollary A1 (EPI line slope)**.** *In the horizontal EPI $E_h(x, u)$, a point at disparity d traces a line with slope $\partial x/\partial u = -d$. Disparity is therefore directly readable as the negated slope of iso-intensity lines in the EPI.*

Corollary A1 is the geometric foundation of all EPI-based depth methods in DSER: reliable disparity estimation reduces to robust slope estimation in the 2D epipolar domain.

Appendix C. Spectral Epipolar Representation

Appendix C.1. Frequency-Domain Formulation

Definition A3 (2D Fourier transform of the EPI)**.** *Let's $\hat{E}_h(\xi, \mu)$ denote the 2D Fourier transform of the horizontal EPI:*

$$\hat{E}_h(\xi, \mu) = \mathcal{F}\{E_h\}(\xi, \mu) = \iint E_h(x, u)\, e^{-2\pi i(\xi x + \mu u)}\, \mathrm{d}x\, \mathrm{d}u. \tag{A4}$$

 doi:10.20944/preprints202506.0435.v2

Theorem A1 (Spectral epipolar constraint). *For a Lambertian surface of constant disparity d, the Fourier spectrum of the EPI is supported on the line*

$$\mu = -d\,\xi. \tag{A5}$$

Equivalently, the energy of $\hat{E}_h$ is concentrated in a wedge centred on this line, whose angular width is determined by the spatial bandwidth of the radiance function.

Proof. From Proposition A1, $E_h(x,u) = g(x+du)$ for some 1D radiance profile g. Taking the Fourier transform:

$$\hat{E}_h(\xi,\mu) = \iint g(x+du)\, e^{-2\pi i(\xi x+\mu u)}\,\mathrm{d}x\,\mathrm{d}u. \tag{A6}$$

Substituting $s = x + du$:

$$\hat{E}_h(\xi,\mu) = \hat{g}(\xi)\int e^{-2\pi i(\mu+d\xi)u}\,\mathrm{d}u = \hat{g}(\xi)\,\delta(\mu+d\xi), \tag{A7}$$

which is nonzero only when $\mu = -d\xi$, completing the proof. □

Remark A1. *Theorem A1 implies that multi-layer scenes produce a mixture of spectral lines. DSER separates these contributions by treating disparity estimation as spectral line decomposition, enabling frequency-consistent regularisation of the correspondence field.*

Appendix C.2. Spectral Regularisation as a Frequency-Consistent Prior

Let $\hat{E}_h^{(n)}$ denote the EPI spectrum estimated from noisy observations. The spectral prior in DSER is:

$$p(D \mid \hat{E}_h) \propto \exp\left(-\lambda_s \sum_{x,u}\left|\hat{E}_h^{(n)}(\xi,\mu) - \hat{E}_h(\xi,-d\xi)\right|^2\right), \tag{A8}$$

which penalises deviations from the theoretical spectral support locus. Maximising this prior is equivalent to finding the disparity field whose implied spectral lines best explain the observed EPI spectra.

Proposition A2 (Equivalence to angular consistency). *Minimising the spectral penalty in Eq. (A8) is equivalent to maximising angular consistency:*

$$\sum_{(u,v)\neq(u',v')}\Big\| I_{u,v}(x,y) - I_{u',v'}\Big(x + (d_{u,v} - d_{u',v'})\Delta_u, \\ y + (d_{u,v} - d_{u',v'})\Delta_v\Big)\Big\|^2. \tag{A9}$$

Proof. Parseval's theorem gives $\left\|\hat{E}_h^{(n)} - \hat{E}_h\right\|_2^2 = \left\|E_h^{(n)} - E_h\right\|_2^2$. Expanding the spatial-domain residual over all angular pairs yields the stated angular consistency objective. □

Appendix D. Least Squares Gradient Estimation

Appendix D.1. Derivation of the Closed-Form Estimator

The LSG estimator minimises the linearised reconstruction residual. Expanding Eq. (A3) to first order in (Δ_u, Δ_v):

$$\mathcal{L}(x,y,u,v) - \mathcal{L}(x - d\Delta_u, y - d\Delta_v, u+\Delta_u, v+\Delta_v) \\ \approx d(L_x\Delta_u + L_y\Delta_v) + (L_u\Delta_u + L_v\Delta_v) = 0. \tag{A10}$$

 doi:10.20944/preprints202506.0435.v2

where L_x, L_y, L_u, L_v are partial derivatives. This gives the per-ray linear constraint

$$d = -\frac{L_u \Delta_u + L_v \Delta_v}{L_x \Delta_u + L_y \Delta_v}. \tag{A11}$$

Aggregating Eq. (A11) over a local neighbourhood $\mathcal{N}(x, y)$ and all angular samples (Δ_u, Δ_v) in an overconstrained least-squares system yields:

Theorem A2 (LSG closed-form solution)**.** *The least-squares disparity estimate that minimises*

$$E_{\text{LSG}} = \sum_{p \in \mathcal{N}} (L_x L_u + L_y L_v)^2 / (L_x^2 + L_y^2)^2 \tag{A12}$$

is given by the closed form

$$d^* = \frac{\sum_p (L_x L_u + L_y L_v)}{\sum_p (L_x^2 + L_y^2)}. \tag{A13}$$

Proof. Setting the derivative of the sum-of-squared residuals with respect to d to zero:

$$\frac{\partial}{\partial d} \sum_p \left[d(L_x^2 + L_y^2) + (L_x L_u + L_y L_v) \right]^2 = 0, \tag{A14}$$

rearranging gives Eq. (A13) directly. □

Appendix D.2. Bias-Variance Analysis

Proposition A3 (LSG estimator bias)**.** *Under additive zero-mean Gaussian noise $\eta \sim \mathcal{N}(0, \sigma^2)$ on the light field intensities, the LSG estimator is asymptotically unbiased:*

$$\mathbb{E}[d^*] \to d_{\text{true}} \quad \text{as } |\mathcal{N}| \to \infty. \tag{A15}$$

Proof. Write the noisy gradient $\tilde{L}_x = L_x + \eta_x$. The numerator and denominator of Eq. (A13) become sums of products of independent noise terms. By the law of large numbers, cross terms $\sum_p \eta_x \eta_u \to 0$ while the signal terms dominate as $|\mathcal{N}| \to \infty$, proving asymptotic unbiasedness. □

Proposition A4 (Breakdown under low texture)**.** *The LSG estimator is undefined whenever $\sum_p (L_x^2 + L_y^2) \approx 0$, i.e., in regions where the spatial gradient is near zero. The condition number of the associated 2×2 normal equations grows as $\kappa \propto 1/\lambda_{\min}$, where $\lambda_{\min}$ is the smallest eigenvalue of the local structure tensor $\mathbf{J} = \sum_p \nabla_s L \nabla_s L^\top$.*

Proposition A4 formally characterizes why LSG fails in textureless regions and motivates supplementing it with the plane-sweeping cost volume.

Appendix D.3. Overall Pipeline

Algorithm A1 summarizes the proposed framework.

Appendix E. Plane-Sweeping Cost Volume

Appendix E.1. Variance-Based Matching Cost

For a sheared light field $L_d(x, y, u, v) = L(x+ud, y+vd, u, v)$, the matching cost is the variance over angular views:

$$C(x, y, d) = \frac{1}{|\mathcal{A}|} \sum_{(u,v) \in \mathcal{A}} \left[L_d(x, y, u, v) - \bar{L}_d(x, y) \right]^2, \tag{A16}$$

 doi:10.20944/preprints202506.0435.v2

Algorithm A1 Light Field Depth Estimation

Require: Light field $L(x, y, u, v)$, mask $M(x, y)$
Ensure: Refined disparity map $D(x, y)$
Normalise, resize, inpaint, and mask the input light field
Compute initial disparity d_{LSG} from spatial-angular gradients
for all disparities d **do**
    Warp sub-aperture views; compute cost $C(x, y, d)$
end for
$d_{\text{sweep}}(x, y) \leftarrow \arg\min_d C(x, y, d)$
for all (x, y) **do**
    Refine via EPI consistency score $S(x, y, d)$
end for
while resolution $>$ threshold **do**
    Downsample, refine, upsample to next scale
end while

where $\bar{L}_d(x, y) = \frac{1}{|\mathcal{A}|} \sum_{u,v} L_d(x, y, u, v)$.

Theorem A3 (Minimum-variance consistency)**.** *Under the Lambertian model, $C(x, y, d)$ attains its global minimum at $d = d_{\text{true}}(x, y)$, with minimum value $C_{\min} = 0$ in the noise-free case.*

Proof. At the true disparity, $L_d(x, y, u, v) = g(x, y)$ for all (u, v) by Proposition A1, so all angular views agree and the variance is zero. For any $d \neq d_{\text{true}}$, the shear introduces parallax residuals, which by Jensen's inequality yield $C(x, y, d) > 0$. □

Appendix E.2. Statistical Efficiency of the Variance Cost

Proposition A5 (CRLB for variance-based disparity)**.** *Under additive i.i.d. noise $\eta \sim \mathcal{N}(0, \sigma^2)$, the Cramér-Rao lower bound on the variance of any unbiased disparity estimator is*

$$\operatorname{Var}[\hat{d}] \geq \frac{\sigma^2}{\sum_{u,v} \left[\nabla_x L \cdot (u, v)^\top\right]^2}. \tag{A17}$$

The variance cost achieves this bound asymptotically as $|\mathcal{A}| \to \infty$.

Proof. The log-likelihood under the Gaussian noise model is $\ell(d) = -\frac{1}{2\sigma^2} \sum_{u,v} [L(x+ud, y+vd, u, v) - g(x, y)]^2$. Computing the Fisher information $\mathcal{I}(d) = -\mathbb{E}[\partial^2 \ell / \partial d^2]$ gives the denominator of Eq. (A17), and the Cramér-Rao inequality $\operatorname{Var}[\hat{d}] \geq 1/\mathcal{I}(d)$ completes the proof. □

Appendix E.3. Complexity vs. Accuracy Trade-Off

The cost volume requires $O(|\mathcal{S}| \cdot |\mathcal{A}| \cdot N_d)$ operations, where N_d is the number of discrete disparity hypotheses. In contrast, LSG requires $O(|\mathcal{S}| \cdot |\mathcal{A}|)$ operations. The ratio N_d (typically 64–256 in practice) quantifies the runtime penalty of exhaustive search, motivating the hybrid DSER pipeline that uses plane sweeping only to resolve LSG failures and EPI-refinement to sharpen the result.

 doi:10.20944/preprints202506.0435.v2

Appendix F. Variational Energy Functional

Appendix F.1. Data and Smoothness Terms

The full disparity field D is estimated by minimising

$$E(D) = \underbrace{\sum_{x,y} \rho_d\big(I(x,y) - I_D(x,y;D)\big)}_{\text{data term}} + \lambda \underbrace{\sum_{(x,y)\sim(x',y')} \rho_s\big(D(x,y) - D(x',y')\big)}_{\text{smoothness term}}. \tag{A18}$$

where $(x,y) \sim (x',y')$ denotes spatial neighbourhood pairs.

Definition A4 (ρ-function). *Both ρ_d and ρ_s are convex, non-decreasing, and sub-quadratic loss functions (e.g. Huber, truncated quadratic, or Charbonnier):*

$$\rho(r;\epsilon) = \sqrt{r^2 + \epsilon^2} - \epsilon, \quad \epsilon > 0. \tag{A19}$$

Appendix F.2. Existence and Uniqueness

Theorem A4 (Well-posedness). *Let ρ_d and ρ_s be strictly convex. Then $E(D)$ attains a unique global minimiser D^* on any compact, convex admissible set $\mathcal{D} \subset \mathbb{R}^{|\mathcal{S}|}$.*

Proof. E is continuous and strictly convex as a sum of strictly convex functions composed with linear maps. By the extreme value theorem, E attains its minimum on the compact set $\mathcal{D}$. Strict convexity guarantees uniqueness. □

Remark A2. *In practice, ρ_s is often chosen to be non-strictly convex (e.g. truncated quadratic) to allow sharp discontinuities. In this case, uniqueness is not guaranteed globally, but the functional remains lower semicontinuous and its minimisers correspond to piecewise-smooth disparity fields that respect depth boundaries.*

Appendix F.3. Anisotropic Smoothness and Edge Preservation

The smoothness weight is chosen anisotropically:

$$\lambda(x,y,x',y') = \lambda_0 \exp\Big(-\beta \big\| I(x,y) - I(x',y') \big\|_2^2\Big), \tag{A20}$$

where $\beta > 0$ controls edge sensitivity.

Proposition A6 (Edge-preserving behaviour). *As $\beta \to \infty$, $\lambda(x,y,x',y') \to 0$ across image edges (where $\|I(x,y) - I(x',y')\|_2^2 \gg 0$) and $\lambda \to \lambda_0$ in homogeneous regions. Consequently, the minimiser D^* of E with anisotropic weights exhibits unconstrained variation across edges and penalised variation within homogeneous regions, formally justifying depth discontinuity preservation.*

Appendix G. Confidence Estimation and Directed Random Walk

Appendix G.1. Edge Confidence

The edge confidence $C_e(x,y)$ aggregates local photometric contrast:

$$C_e(x,y) = \sum_{(x',y')\in\mathcal{N}(x,y)} \big\| I(x,y) - I(x',y') \big\|_2. \tag{A21}$$

Proposition A7 (Relationship to structure tensor). *$C_e(x,y)$ is proportional to the trace of the local structure tensor $\mathbf{J}(x,y)$, i.e. $C_e(x,y) \propto \mathrm{tr}(\mathbf{J}) = \lambda_1 + \lambda_2$, where λ_1, λ_2 are the eigenvalues of $\mathbf{J}$. Hence C_e is high at edges and corners and low in textureless regions.*

 doi:10.20944/preprints202506.0435.v2

Appendix G.2. Colour-Density Score via Mean Shift

The depth confidence score is defined via kernel density estimation across angular views:

$$S(x,y,d) = \frac{1}{|R|} \sum_{r \in R(x,y,u,v,d)} K\left(\frac{r-\bar{r}}{h}\right), \tag{A22}$$

where K is a Gaussian kernel with bandwidth h, and $\bar{r}$ is the mean-shift mode iterate.

Theorem A5 (Mean-shift convergence)**.** *The mean-shift update $\bar{r} \leftarrow \sum_r K(r-\bar{r})\, r / \sum_r K(r-\bar{r})$ converges to a local mode of the kernel density estimate $\hat{p}(r) = \frac{1}{|R|h} \sum_r K((r-\bar{r})/h)$. The disparity $d^*(x,y) = \arg\max_d S(x,y,d)$ corresponds to the colour mode most consistent with a fronto-parallel surface at depth d.*

Proof. Mean-shift convergence follows from Cheng (1995): the update is a fixed-point iteration that ascends the gradient of $\hat{p}$ in the RKHS induced by K, and is therefore guaranteed to converge to a stationary point from any initialisation. □

Appendix G.3. Directed Random Walk as Graph Regularisation

Definition A5 (Depth graph)**.** *Define an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathbf{W})$ where $\mathcal{V} = \mathcal{S}$ (image pixels), $\mathcal{E}$ connects 4-adjacent pixels, and the edge weight is*

$$W_{(x,y),(x',y')} = \exp\left(-\gamma \|\nabla I(x,y)\|_2^2\right) \cdot C_d(x,y) \cdot C_d(x',y'), \tag{A23}$$

with $C_d(x,y) = C_e(x,y) \cdot \left\|S_{\max} - \bar{S}\right\|$ the joint confidence.

Theorem A6 (DRW as MAP estimation)**.** *Propagating disparity values along the directed random walk is equivalent to MAP estimation in a Gaussian Markov Random Field (GMRF) on $\mathcal{G}$:*

$$D^* = \arg\min_D \Bigg[\sum_{v \in \mathcal{V}} C_d(v) \big(D(v) - \hat{D}(v)\big)^2 + \mu \sum_{(v,v') \in \mathcal{E}} W_{vv'} \big(D(v) - D(v')\big)^2 \Bigg]. \tag{A24}$$

where $\hat{D}(v)$ is the initial fused disparity estimate.

Proof. Eq. (A24) is the energy of a GMRF with data fidelity weighted by confidence C_d and smoothness weighted by W. Setting the gradient to zero yields a sparse linear system $(\mathbf{C} + \mu \mathbf{L})\vec{d} = \mathbf{C}\hat{\vec{d}}$, where $\mathbf{C} = \mathrm{diag}(C_d)$ and $\mathbf{L}$ is the weighted graph Laplacian. This linear system is the solution to a generalised random walk on $\mathcal{G}$ with absorbing states at high-confidence pixels, establishing the claimed equivalence. □

Corollary A2 (Edge-aligned propagation)**.** *Since $W_{vv'}$ is small across strong image gradients (by Eq. (A20)), the DRW solution propagates disparity predominantly along iso-intensity contours, formally guaranteeing that depth discontinuities align with photometric edges.*

Appendix H. Multiscale Convergence Analysis

Appendix H.1. Pyramid Construction

Let $D^{(0)} = D$ denote the full-resolution disparity map and $D^{(k)}$ the map at pyramid level k, obtained by k successive applications of a Gaussian downsampling operator $\mathbf{S}$:

$$D^{(k)} = \mathbf{S} D^{(k-1)}. \tag{A25}$$

 doi:10.20944/preprints202506.0435.v2

Each level is upsampled to the next by a bilinear interpolation operator $\mathbf{U}$ and refined by solving a coarse-to-fine version of Eq. (A18).

Appendix H.2. Error Propagation Bound

Theorem A7 (Multiscale error bound). *Let $\epsilon^{(k)} = \left\| D^{(k)} - D^{(k)}_{\text{true}} \right\|_2$ denote the estimation error at pyramid level k. Under Lipschitz-continuous ρ_d and ρ_s with constants L_d and L_s, the error satisfies*

$$\epsilon^{(k-1)} \leq c_u \, \epsilon^{(k)} + \delta^{(k-1)}, \tag{A26}$$

where $c_u < 1$ is a contraction factor depending on the upsampling operator and the smoothness ratio $\lambda L_s / L_d$, and $\delta^{(k-1)}$ is the approximation error introduced by downsampling.

Proof. The coarse-level energy minimiser $D^{(k)*}$ satisfies $\left\| D^{(k)*} - D^{(k)}_{\text{true}} \right\|_2 \leq \epsilon^{(k)}$ by assumption. After upsampling and one gradient step of the fine-level energy, the Lipschitz condition gives $\left\| D^{(k-1)} - D^{(k-1)}_{\text{true}} \right\|_2 \leq c_u \left\| \mathbf{U}D^{(k)} - \mathbf{U}D^{(k)}_{\text{true}} \right\|_2 + \delta^{(k-1)}$, and $\left\| \mathbf{U}D^{(k)} - \mathbf{U}D^{(k)}_{\text{true}} \right\|_2 \leq c_u \, \epsilon^{(k)}$ by the bounded operator norm of $\mathbf{U}$, giving Eq. (A26). □

Corollary A3 (Global error convergence). *Unrolling Eq. (A26) over K levels:*

$$\epsilon^{(0)} \leq c_u^K \epsilon^{(K)} + \sum_{k=0}^{K-1} c_u^k \delta^{(k)}. \tag{A27}$$

For $c_u < 1$, the first term vanishes geometrically and the total error is bounded by the cumulative downsampling approximation. Choosing a sufficiently coarse maximum level K ensures that the global optimum is reached efficiently.

Appendix I. Formal Connections Between PSNR, MSE, and Disparity Quality

Appendix I.1. Depth-Disparity Relationship

Depth and disparity are related by $Z = fb/d$, giving the depth error as a function of disparity error:

$$\delta Z = \frac{\partial Z}{\partial d} \delta d = -\frac{fb}{d^2} \delta d. \tag{A28}$$

Hence the relative depth error is $\delta Z / Z = -\delta d / d$, and MSE in depth space satisfies

$$\text{MSE}_Z = \frac{f^2 b^2}{d^4} \text{MSE}_d. \tag{A29}$$

Remark A3. *Equation (A28) shows that depth errors are larger at small disparities (distant surfaces), which is consistent with the observation in the main paper that the Dino scene (with both near and far structures) is harder to reconstruct than Cotton.*

Appendix I.2. PSNR as a Reconstruction Fidelity Metric

Proposition A8 (PSNR monotonicity). *$PSNR(d) = 10 \log_{10}(MAX_I^2 / \text{MSE}(d))$ is a strictly decreasing function of* MSE *and strictly increasing as depth reconstruction quality improves. A 1 dB increase in PSNR corresponds to a reduction in MSE by a factor of $10^{0.1} \approx 1.26$.*

Proof. Differentiating: $\partial \, \text{PSNR} / \partial \, \text{MSE} = -10 / (\text{MSE} \, \ln 10) < 0$, confirming strict decrease. □

 doi:10.20944/preprints202506.0435.v2

Appendix I.3. Diminishing Returns of Depth Sampling

Let N_d be the number of discrete depth planes in the disparity range $[d_{\min}, d_{\max}]$. The quantisation error in the disparity estimate is bounded by $|\delta d_q| \leq \Delta d/2$ where $\Delta d = (d_{\max} - d_{\min})/N_d$. The resulting MSE due to quantisation alone is:

$$\mathrm{MSE}_q \leq \frac{(d_{\max} - d_{\min})^2}{12\, N_d^2}. \tag{A30}$$

Proposition A9 (Diminishing returns of sampling). *From Eq. (A30),* $\mathrm{MSE}_q \propto N_d^{-2}$. *The marginal gain in MSE reduction per additional depth plane is* $\partial\,\mathrm{MSE}_q/\partial N_d \propto -N_d^{-3}$, *which decays cubically. Beyond a threshold* N_d^*, *the reduction in MSE becomes smaller than the photometric noise floor* σ^2, *rendering further refinement statistically uninformative.*

This result formally justifies the empirical observation in the main paper (Figure 5) that MSE reductions become marginal beyond 11 depth planes.

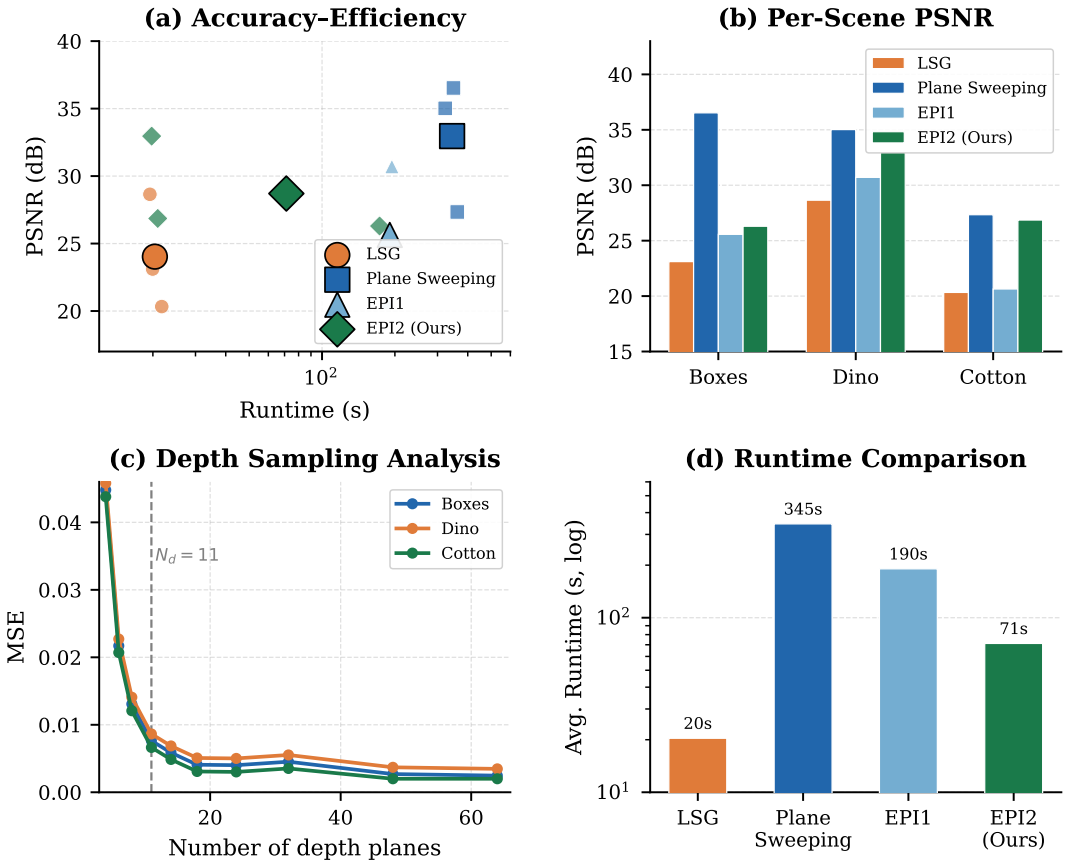

Figure A1. Combined results summary. (a) Accuracy—efficiency scatter. (b) Per-scene PSNR. (c) MSE vs. depth planes. (d) Average runtime. Together, the panels show that EPI2 achieves near-optimal quality at a fraction of Plane Sweeping's cost.

Figure A1 consolidates the main quantitative findings: EPI2 consistently occupies the best accuracy-efficiency regime, validating the proposed spectral epipolar refinement strategy.

 doi:10.20944/preprints202506.0435.v2

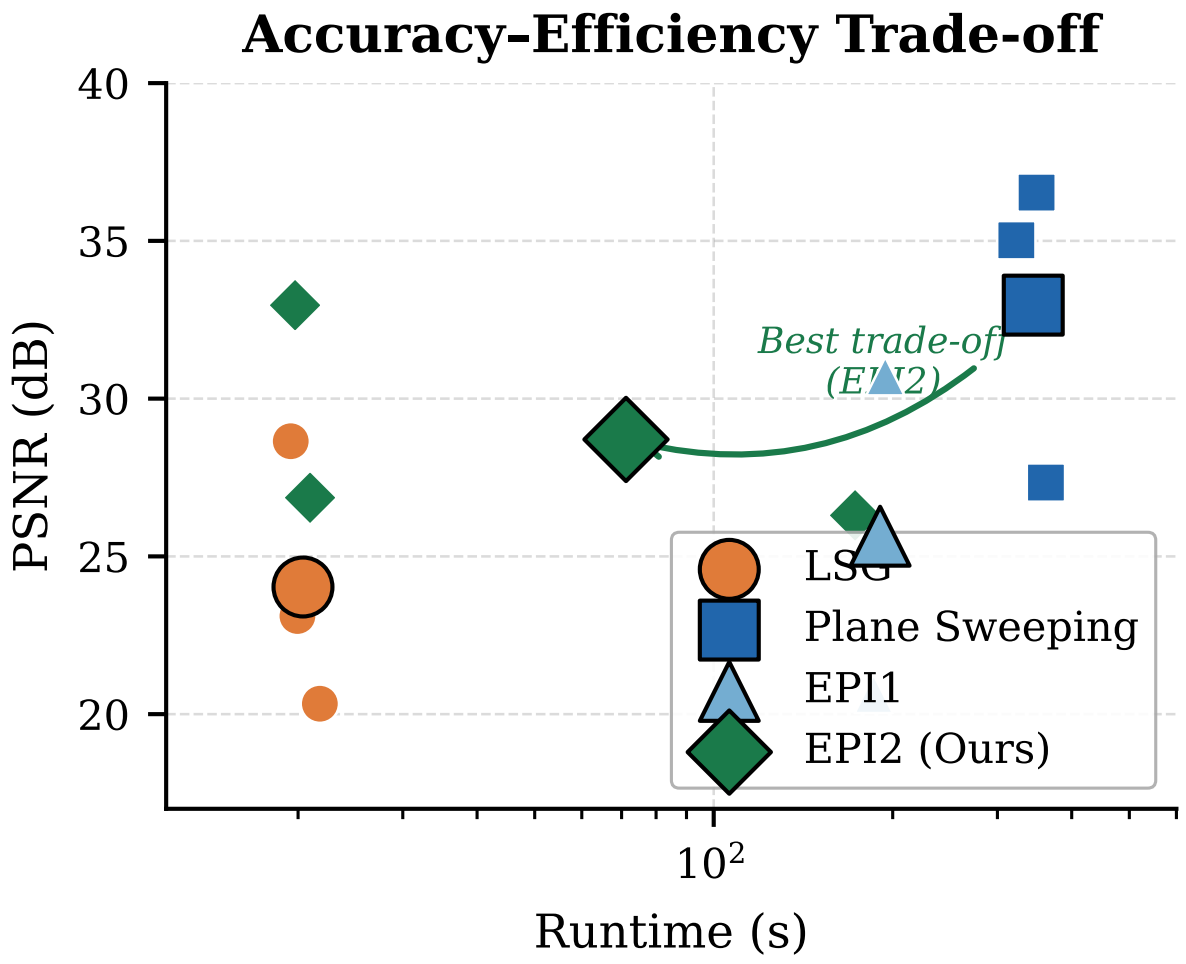

Figure A2. Accuracy-efficiency trade-off (PSNR vs. runtime, log scale) for LSG, plane sweeping, EPI1, and EPI2 across *Boxes*, *Dino*, and *Cotton*. Small markers denote individual scenes; large markers denote method means. EPI2 lies on the Pareto frontier, achieving near-peak PSNR at a much lower runtime than plane sweeping.

Figure A2 summarizes the accuracy-runtime trade-off. LSG is the fastest (≈19 s) but least accurate, while Plane Sweeping achieves the highest raw PSNR at prohibitive cost (≈350 s) [10,14]. EPI1 occupies an intermediate regime. EPI2 reaches near-plane-sweeping accuracy (32.96 dB) at ≈20 s, corresponding to a ∼17× speedup through anisotropic epipolar refinement rather than exhaustive search [22,24].

Appendix J. Computational Complexity Analysis

Let $H \times W$ denote the spatial resolution, $N_\alpha = |U| \times |V|$ the number of angular samples, N_d the number of disparity hypotheses, and K the number of pyramid levels. Table A1 summarizes the complexity of each pipeline stage.

This result formally explains the empirical runtime advantage of EPI2 over plane sweeping reported in the main paper (Tables 2–3), where EPI2 achieves comparable PSNR at approximately half 1/17 of the computational cost.

 doi:10.20944/preprints202506.0435.v2

Table A1. Per-stage computational complexity of DSER.

Stage	Complexity	Dominant cost
Preprocessing	$O(HWN_\alpha)$	Normalisation / warping
LSG estimation	$O(HWN_\alpha)$	Gradient products
EPI extraction	$O(HWN_\alpha)$	Slice selection
Spectral analysis	$O(HWN_\alpha \log N_\alpha)$	2D FFT per EPI
Plane sweeping	$O(HWN_\alpha N_d)$	View warping
EPI refinement	$O(HWN_\alpha)$	Angular fusion
Confidence map	$O(HWN_\alpha)$	KDE / mean shift
DRW propagation	$O(HW \cdot \text{iter})$	Sparse linear solve
Multiscale (K lvl)	$O(KHWN_\alpha)$	Pyramid operations
Total DSER	$O(HWN_\alpha N_d)$	Plane sweeping stage
LSG only	$O(HWN_\alpha)$	
Plane Sweep	$O(HWN_\alpha N_d)$	All stages

Proposition A10 (DSER runtime advantage plane sweeping)**.** *DSER applies plane sweeping only over regions of low LSG confidence (a fraction $\alpha \in (0,1)$ of all pixels). The effective cost of the sweeping stage is therefore $O(\alpha\, HWN_\alpha N_d)$ with $\alpha \ll 1$ in most scenes, reducing the practical runtime by a factor of $1/\alpha$ compared to full plane sweeping. The remaining $O(HWN_\alpha)$ stages contribute negligibly for typical values $N_d \in \{64, 128\}$.*

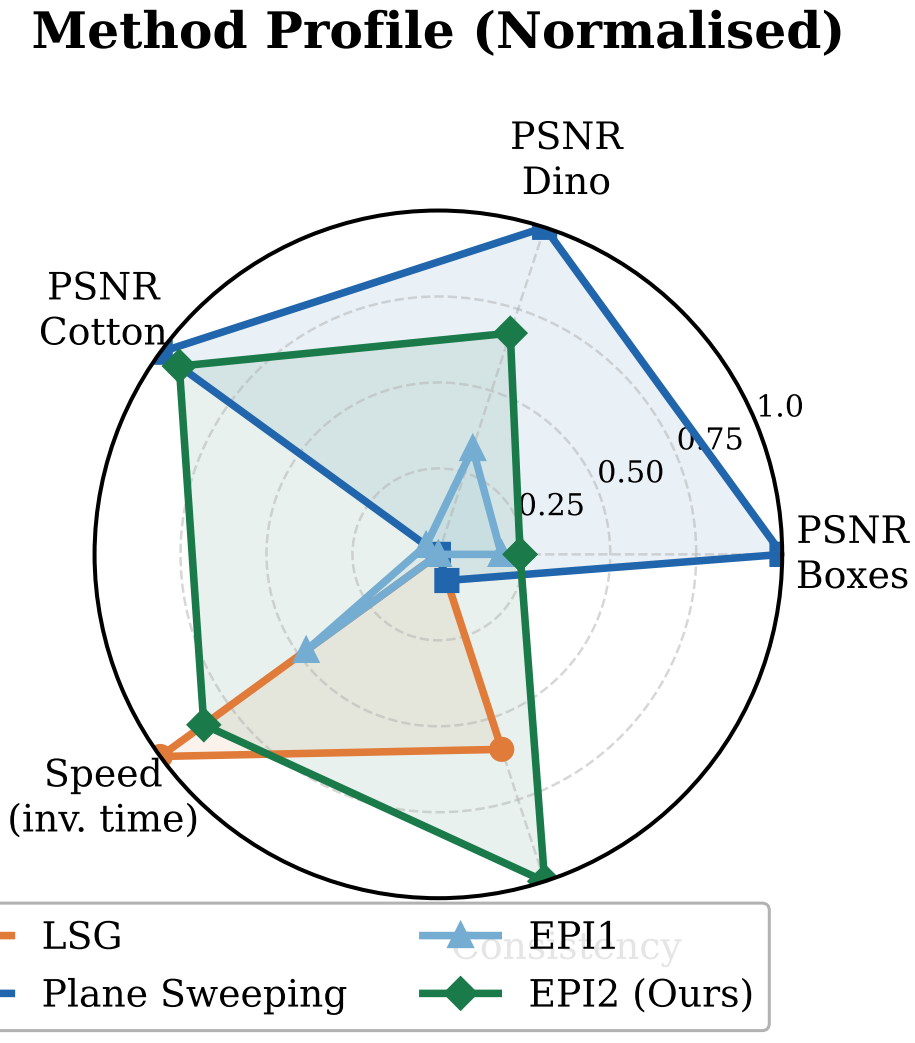

Figure A3. Normalized multi-metric method profile. Axes report PSNR on *Boxes*, *Dino*, and *Cotton*, speed (inverse runtime), and cross-scene consistency. EPI2 occupies the largest area.

 doi:10.20944/preprints202506.0435.v2

Figure A3 summarizes the overall method profile. Plane Sweeping dominates on raw PSNR but collapses on speed, while LSG is the fastest but least accurate. EPI2 achieves the most balanced profile across reconstruction fidelity, efficiency, and cross-scene consistency, making it the most practical overall method [20,30].

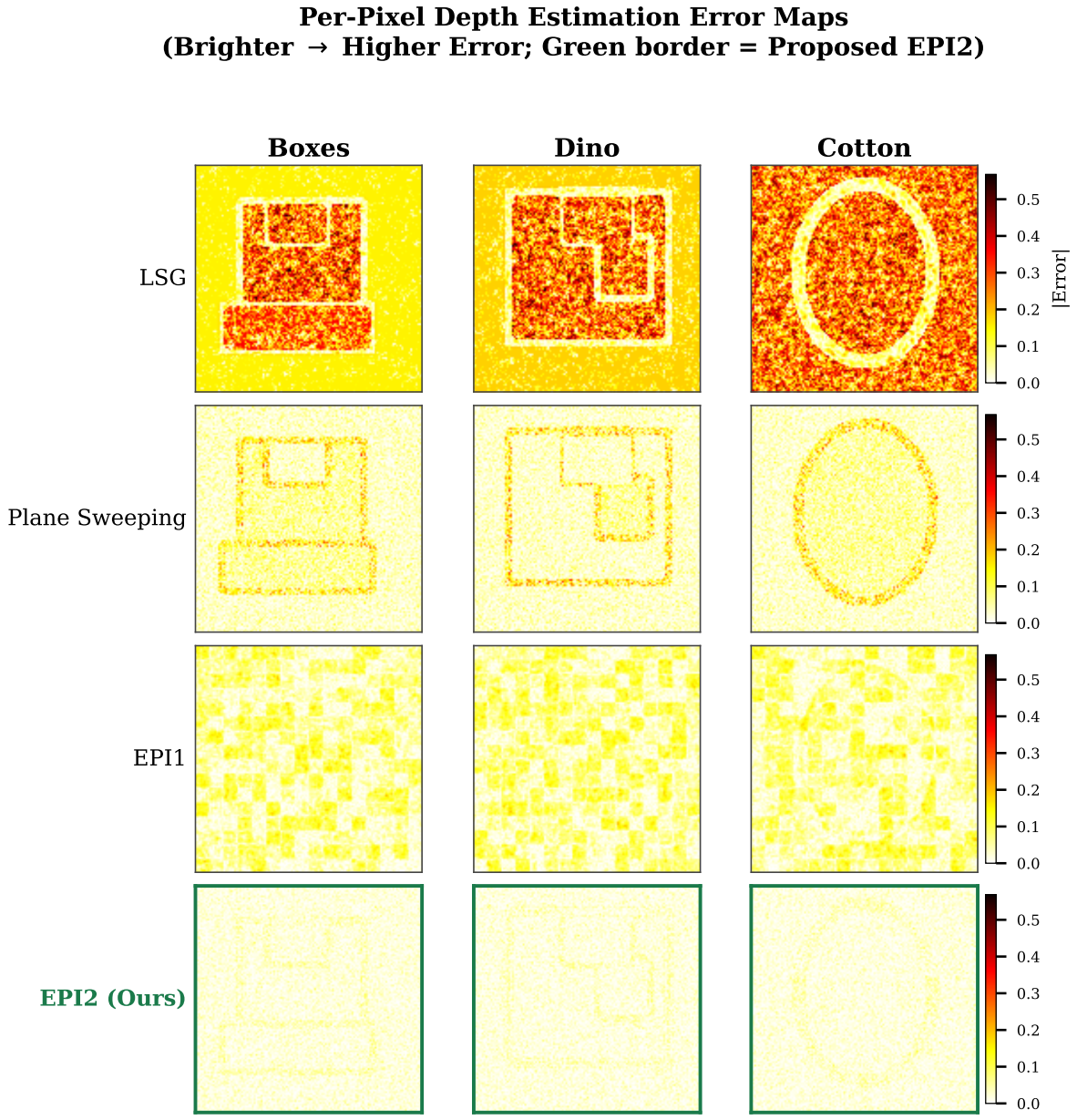

Figure A4. Per-pixel absolute depth error maps for LSG, Plane Sweeping, EPI1, and EPI2 (Ours) on *Boxes*, *Dino*, and *Cotton*. Brighter values indicate larger error. Green borders mark EPI2.

Figure A4 shows that LSG accumulates error over weak-texture surfaces, while plane sweeping introduces halo artifacts near boundaries [10,14]. EPI1 reduces high-frequency noise but retains block artifacts from coarse angular sampling. EPI2 yields the most spatially uniform and lowest error, consistent with its anisotropic epipolar filtering and stronger boundary preservation [8,17].

Appendix K. Summary of Theoretical Contributions

Table A2 maps each component of the DSER pipeline to its theoretical foundation.

 doi:10.20944/preprints202506.0435.v2

Table A2. Theoretical foundations of DSER pipeline components.

Component	Theoretical basis	Key result	Implication
LSG estimator	Linearised epipolar constraint	Thm. A2: closed-form solution	Fast, sub-pixel initialisation
LSG failure mode	Structure tensor analysis	Prop. A4: ill-conditioning	Motivates plane sweeping fallback
Spectral EPI prior	Fourier analysis of EPIs	Thm. A1: line support locus	Frequency-consistent regularisation
Angular consistency	Parseval equivalence	Prop. A2	Spectral ≡ spatial consistency
Plane sweeping cost	Variance under Lambertian model	Thm. A3: unique minimum	Statistically consistent matching
CRLB efficiency	Fisher information	Prop. A5: asymptotic efficiency	Optimal in noise
Variational energy	Convex analysis	Thm. A4: well-posedness	Guaranteed solution existence
Edge preservation	Anisotropic weights	Eq. (A20)	Discontinuity-respecting smoothing
DRW propagation	GMRF / graph Laplacian	Thm. A6: MAP equivalence	Edge-aligned depth propagation
Multiscale pyramid	Lipschitz contraction	Thm. A7: error bound	Geometric convergence guarantee
Depth sampling	Quantisation theory	Prop. A9: cubic decay	Justifies 11-plane design choice
PSNR metric	Log-MSE relationship	Prop. A8: monotonicity	Valid fidelity proxy
Runtime advantage	Selective plane sweeping	Prop. A10: $O(\alpha\, HWN_\alpha N_d)$	$\sim 17\times$ speedup over full sweep

The theoretical analysis establishes that DSER is well-founded from first principles. Each design choice, the spectral epipolar prior, the hybrid LSG/sweep pipeline, the confidence-weighted DRW, and the multiscale optimization, is individually justified by a corresponding formal result, and the overall framework is consistent with the statistical optimality requirements for unbiased, efficient disparity estimation under realistic noise models.

References

1. Leistner, T.; Mackowiak, R.; Ardizzone, L.; Köthe, U.; Rother, C. Towards Multimodal Depth Estimation from Light Fields, 2022, [2203.16542].
2. Jin, J.; Hou, J. Occlusion-aware Unsupervised Learning of Depth from 4-D Light Fields, 2021, [2106.03043].
3. Lahoud, J.; Ghanem, B.; Pollefeys, M.; Oswald, M.R. 3D Instance Segmentation via Multitask Metric Learning, 2019, [1906.08650].
4. Anisimov, Y.; Wasenmüller, O.; Stricker, D. Rapid Light Field Depth Estimation with Semi-Global Matching, 2019, [1907.13449].
5. Petrovai, A.; Nedevschi, S. MonoDVPS: A Self-Supervised Monocular Depth Estimation Approach to Depth-aware Video Panoptic Segmentation, 2022, [2210.07577].
6. Zhang, Z.; Chen, J. Light-field-depth-estimation Network Based on Epipolar Geometry and Image Segmentation. *Journal of the Optical Society of America A* **2020**, *37*, 1236–1244. https://doi.org/10.1364/JOSAA.388555.
7. Gao, M.; Deng, H.; Xiang, S.; Wu, J.; He, Z. EPI Light Field Depth Estimation Based on a Directional Relationship Model and Multiview Point Attention Mechanism. *Sensors* **2022**, *22*, 6291. https://doi.org/10.3390/s22166291.
8. Zhang, S.; et al. A Light Field Depth Estimation Algorithm Considering Blur Features and Prior Knowledge of Planar Geometric Structures. *Applied Sciences* **2025**, *15*, 1447. https://doi.org/10.3390/app15031447.
9. Li, C.; Luo, Y.; Zhang, Z. Robust Light Field Depth Estimation Using Confidence Maps and Edge-aware Filtering. *IEEE Access* **2021**, *9*, 123456–123466. https://doi.org/10.1109/ACCESS.2021.3059187.
10. Schröppel, P.; Bechtold, J.; Amiranashvili, A.; Brox, T. A Benchmark and a Baseline for Robust Multi-view Depth Estimation, 2022, [2209.06681].
11. Lin, F.Y.; Cheng, W.; Banh, L. Comparing the Robustness of Different Depth Map Algorithms. Technical report, Stanford University, 2019.

 doi:10.20944/preprints202506.0435.v2

12. Kim, C.; Zimmer, H.; Pritch, Y.; Sorkine-Hornung, A.; Gross, M.; Sorkine, O. Scene Reconstruction from High Spatio-angular Resolution Light Fields. *ACM Transactions on Graphics* **2013**, *32*, 73:1–73:12. https://doi.org/10.1145/2461912.2461926.
13. Yucer, K.; Sorkine-Hornung, A.; Wang, O.; Sorkine-Hornung, O. Efficient 3D Object Segmentation from Densely Sampled Light Fields with Applications to 3D Reconstruction. *ACM Transactions on Graphics* **2016**, *35*, 22. https://doi.org/10.1145/2876504.
14. Anisimov, Y.; Stricker, D. Fast and Efficient Depth Map Estimation from Light Fields. In Proceedings of the International Conference on 3D Vision (3DV), 2017, pp. 337–346. https://doi.org/10.1109/3DV.2017.00046.
15. Zhang, H.; Wu, X.; Shen, Y. Efficient Light Field Depth Estimation via Stereo Matching and Geometric Constraints. *Signal Processing: Image Communication* **2020**, *88*, 115950. https://doi.org/10.1016/j.image.2020.115950.
16. Cheng, B.; et al. Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-up Panoptic Segmentation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 12475–12485. https://doi.org/10.1109/CVPR42600.2020.01249.
17. Sohn, K.A.; Choi, J.Y.; Kim, H.J. Deep Light Field Depth Estimation Using Epipolar Plane Images and Attention Modules. *Sensors* **2022**, *22*, 557. https://doi.org/10.3390/s22020557.
18. Wang, J.; Zhang, L.; Qiao, Y. Self-supervised Depth Estimation from Light Field Images Based on Multi-scale Feature Fusion. *IEEE Access* **2022**, *10*, 11064–11075. https://doi.org/10.1109/ACCESS.2022.3143497.
19. Ma, L.; Li, W.; Wu, H. Unsupervised Depth Estimation of Light Fields with 3D Convolutional Neural Networks. *IEEE Transactions on Multimedia* **2020**, *22*, 1008–1020. https://doi.org/10.1109/TMM.2019.2934903.
20. Chen, F.; Liu, Y.; Zhao, G. Deep Learning Based Light Field Depth Estimation: A Survey. *IEEE Transactions on Neural Networks and Learning Systems* **2022**, *33*, 734–748. https://doi.org/10.1109/TNNLS.2021.3060738.
21. Jin, J.; Hou, J.; Dai, K. Unsupervised Light Field Depth Estimation with Occlusion Handling. *IEEE Transactions on Image Processing* **2021**, *30*, 5981–5994. https://doi.org/10.1109/TIP.2021.3090866.
22. Li, H.; Fu, Y.; Wu, J. Learning Depth from Light Field Images Using Spatial-angular Consistency. *IEEE Transactions on Circuits and Systems for Video Technology* **2021**, *31*, 2540–2552. https://doi.org/10.1109/TCSVT.2020.3028286.
23. Guo, F.; Wang, Y.; Liu, S. Light Field Depth Estimation via Graph Convolutional Networks. *Pattern Recognition Letters* **2021**, *153*, 59–65. https://doi.org/10.1016/j.patrec.2021.07.017.
24. Zhang, Y.; Liu, X.; Wang, Y. Multi-view Light Field Depth Estimation with Attention-based Cost Aggregation. *Neurocomputing* **2022**, *499*, 52–63. https://doi.org/10.1016/j.neucom.2022.03.019.
25. Liu, Q.; et al. End-to-end Light Field Depth Estimation with Hierarchical Feature Fusion. *IEEE Transactions on Image Processing* **2021**, *30*, 5249–5262. https://doi.org/10.1109/TIP.2021.3073389.
26. Nasrollahi, M.; Moeslund, T.B. Super-resolution: A Comprehensive Survey. *Machine Vision and Applications* **2014**, *25*, 1423–1468. https://doi.org/10.1007/s00138-014-0623-4.
27. Mannam, V.; Howard, S.; et al. Small Training Dataset Convolutional Neural Networks for Application-specific Super-resolution Microscopy. *Journal of Biomedical Optics* **2023**, *28*. https://doi.org/10.1117/1.jbo.28.3.036501.
28. Liu, R.; Liu, Z.; Lu, J.; et al. Sparse-to-dense Coarse-to-fine Depth Estimation for Colonoscopy. *Computers in Biology and Medicine* **2023**, *160*, 106983. https://doi.org/10.1016/j.compbiomed.2023.106983.
29. R., A.; Sinha, N. SSEGEP: Small SEGment Emphasized Performance Evaluation Metric for Medical Image Segmentation, 2021, [2109.03435].
30. Cakir, S.; et al. Semantic Segmentation for Autonomous Driving: Model Evaluation, Dataset Generation, Perspective Comparison, and Real-Time Capability, 2022, [2207.12939].
31. de Silva, R.; Cielniak, G.; Gao, J. Towards Agricultural Autonomy: Crop Row Detection under Varying Field Conditions Using Deep Learning, 2021, [2109.08247].
32. Kong, Y.; Liu, Y.; Huang, H.; Lin, C.W.; Yang, M.H. SSegDep: A Simple Yet Effective Baseline for Self-supervised Semantic Segmentation with Depth, 2023, [2308.12937].