

SynLF: Zero-Shot Metric Depth from Light Field Cameras via Physics-Grounded Synthesis

Zhexuan Cao, Yuduo Guo, Peisheng Ding, Zhan Shi, and Hui Qiao*

Department of Automation, Tsinghua University, Beijing, China

Abstract. Accurate monocular metric depth estimation remains inherently ill-posed due to scale ambiguity. Light field cameras mitigate this ambiguity by capturing micro-baseline disparities within a single lens to recover metric depth, but current learning-based solutions are limited by scarce training data and often overfit Lambertian appearance patterns, restricting real-world robustness. To address this, we propose **SynLF**, which alleviates data scarcity via physics-grounded light field (PG-LF) synthesis and performs robust real-world metric depth estimation through VisDepth, a prior-initialized iterative estimator with explicit multi-view occlusion reasoning. During training, we synthesize light field supervision on-the-fly from RGB-D via physics-grounded light field synthesis, including geometry-consistent view synthesis, depth-dependent defocus, and non-Lambertian perturbations. At inference, the model is directly evaluated on real light field captures without fine-tuning. Extensive experiments demonstrate that SynLF generalizes zero-shot to our custom real-world light field dataset, which features highly challenging conditions such as transparency, reflections, specular highlights, and textureless regions. It achieves 30 mm absolute and 1.4% relative errors across 0.5–2.5 m, demonstrating that physics-grounded LF synthesis from large-scale RGB-D provides a scalable and storage-efficient alternative to conventional light field acquisition for accurate single-lens metric depth estimation.

Keywords: Light Field Depth Estimation · Single-lens Metric Depth · Physics-Grounded Synthesis · Zero-Shot Generalization · Occlusion Reasoning

1 Introduction

Accurate metric depth perception is essential for 3D applications such as autonomous driving and robotics [8, 9]. While active sensors (*e.g.*, LiDAR) are bulky and costly [22], monocular estimators offer a compact alternative by leveraging large-scale datasets and powerful foundation models [16, 32]. However, due to the inherent scale ambiguity of conventional cameras, recovering reliable absolute metric scale from a single image remains fundamentally ill-posed. In contrast, light field (LF) cameras provide a principled single-lens solution:

* Corresponding author

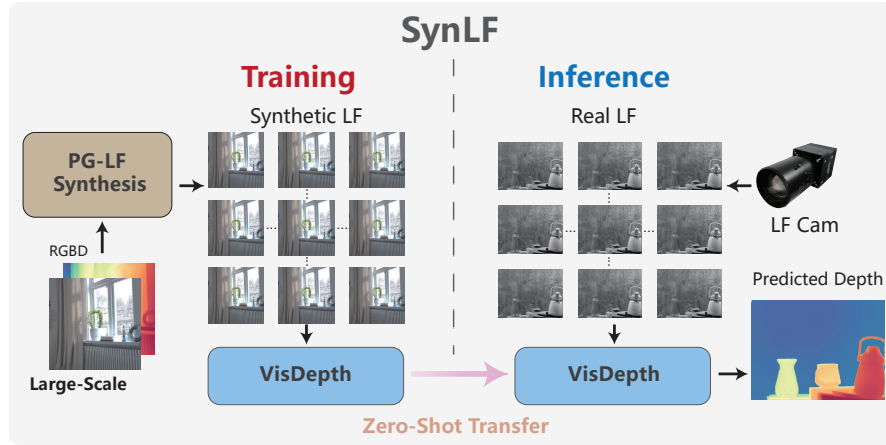


Fig. 1: Overview of SynLF. During training, SynLF synthesizes supervisory light fields on-the-fly from large-scale RGB-D data via PG-LF synthesis and trains VisDepth on synthetic light fields. At inference, the same model is directly applied to real light field captures for zero-shot metric depth estimation.

by capturing angular information in a single shot, they directly encode metric depth cues through micro-baseline disparities [1, 3].

Despite this theoretical advantage, current learning-based LF depth estimation is fundamentally bottlenecked by the scarcity of training data. Rendering multi-view LF supervision is computationally expensive, and available supervised LF datasets remain far smaller than modern monocular RGB-D corpora, ranging from tens of benchmark scenes to roughly thousand-sample rendered/captured datasets [15, 24]. This data scarcity often causes models to overfit idealized Lambertian appearances and generalize poorly to real-world complexities such as occlusion, non-Lambertian surfaces, and textureless regions [6, 30]. Furthermore, existing methods typically aggregate multi-view evidence through implicit weighting [28], lacking explicit visibility constraints. Such underconstrained geometric reasoning inevitably leads to brittle performance in challenging real-world scenes.

To address these limitations, we introduce SynLF, a framework that combines physics-grounded light field synthesis (**PG-LF synthesis**) with a visibility-aware depth estimation network (**VisDepth**), as summarized in Fig. 1. During training, SynLF synthesizes light field supervision on-the-fly from RGB-D inputs via PG-LF synthesis, which selectively models the imaging effects that directly corrupt cross-view disparity correspondence—the core cue for light field depth estimation—including occlusion-aware view synthesis, depth-dependent defocus, and non-Lambertian perturbations. Built upon this supervision, VisDepth performs prior-initialized, visibility-aware iterative disparity estimation with explicit multi-view occlusion reasoning. By coupling the synthesis pipeline and the iterative estimator via the same depth-aware projection operator, our

framework inherently enforces strict geometric reasoning rather than overfitting to appearance shortcuts, establishing a scalable and robust solution for light field metric depth estimation under real-world conditions.

We evaluate SynLF on real-world data captured by a custom-built light field camera paired with a high-precision structured-light sensor, covering challenging scenes with transparency, specular reflections, and textureless regions. SynLF achieves 30 mm mean absolute error and 1.4% mean relative error across a 0.5–2.5 m range, demonstrating strong zero-shot generalization and validating PG-LF synthesis as a scalable alternative to conventional light field data acquisition.

Our contributions are three-fold:

- We propose **PG-LF synthesis**, a physics-grounded light field synthesis pipeline that generates on-the-fly supervisory light fields from abundant RGB-D, eliminating the need for pre-rendered light field datasets.
- We introduce **VisDepth**, a prior-initialized, visibility-aware iterative estimator that explicitly reasons about multi-view occlusions for robust light field metric depth estimation.
- The resulting framework, **SynLF**, demonstrates strong zero-shot generalization on challenging real-world light field captures and achieves accurate metric depth estimation without fine-tuning.

2 Related Work

2.1 Monocular Metric Depth Estimation

Monocular metric depth estimation seeks to recover scene geometry from a single RGB image, yet the problem is fundamentally ill-posed due to scale ambiguity: multiple geometrically plausible 3D scenes can produce identical image observations with different absolute scales. Early approaches mitigated this issue by adopting scale-invariant objectives that emphasize relative depth relationships rather than metric accuracy [10].

Subsequent methods explored various strategies to reduce scale ambiguity, including discretized depth classification [2, 11], large-scale synthetic supervision [21], and scale/shift-invariant objectives or affine-invariant geometry representations to improve cross-domain robustness [20, 29, 32]. More recent works explicitly aim at metric depth prediction by modeling camera geometry, estimating camera parameters, or training for absolute metric scale over diverse data [4, 16, 19]. Nevertheless, even state-of-the-art monocular estimators can exhibit significant scale errors in simple scenes, underscoring the inherent limitations of single-view geometry.

2.2 Light Field Depth Estimation

Light field cameras capture multiple sub-aperture views within a single exposure, effectively forming a narrow-baseline camera array within a single-lens imaging

system [1, 13]. Compared with conventional monocular imaging, this multi-view structure provides additional geometric constraints that help resolve scale ambiguity.

Early approaches leveraged optimization-based formulations or epipolar plane image (EPI) analysis to estimate depth from geometric and photometric constraints [14, 26, 31, 33]. More recent learning-based methods construct multi-disparity cost volumes and employ neural networks for disparity regression and refinement [5, 28, 30]. These supervised approaches typically rely on fully rendered light field datasets with ground-truth disparity annotations, such as HCI [15] and the synthetic subset of UrbanLF [24].

However, due to the limited scale and diversity of available light field datasets, learning-based models often struggle to generalize beyond controlled synthetic settings. In addition, occlusion reasoning is commonly handled implicitly through view aggregation mechanisms without explicit visibility constraints, limiting geometric consistency across views [28]. Together, these data and modeling limitations substantially restrict the robustness and scalability of current light field depth estimation methods.

2.3 Light Field Data Synthesis

Common light field depth benchmarks and synthetic subsets are primarily generated using graphics rendering engines [15, 24], which construct full 3D scenes and render images from micro-baseline camera arrays. Although this strategy produces geometrically accurate supervision, it requires rendering many sub-aperture views per scene (typically 5×5 to 9×9), making large-scale generation much heavier than producing a single RGB-D pair. As a result, available datasets remain limited in scale and diversity.

To improve efficiency, recent work has explored synthesizing light fields directly from a single image using neural networks [12, 23, 27]. However, these methods learn to hallucinate plausible views via reconstruction losses without explicit geometric consistency guarantees, and typically require real light field captures for training. As a result, the synthesized views may appear visually reasonable but lack the multi-view geometric fidelity required for reliable depth supervision.

3 Method

3.1 Overview

As illustrated in Fig. 2, SynLF consists of two components: PG-LF synthesis, our physics-grounded light field synthesis pipeline, and VisDepth, our visibility-aware depth network. During training, PG-LF synthesis generates supervisory light fields on-the-fly from large-scale RGB-D data, providing scalable supervision without requiring pre-rendered light field datasets. Rather than pursuing generic photorealistic simulation, PG-LF synthesis selectively models the factors

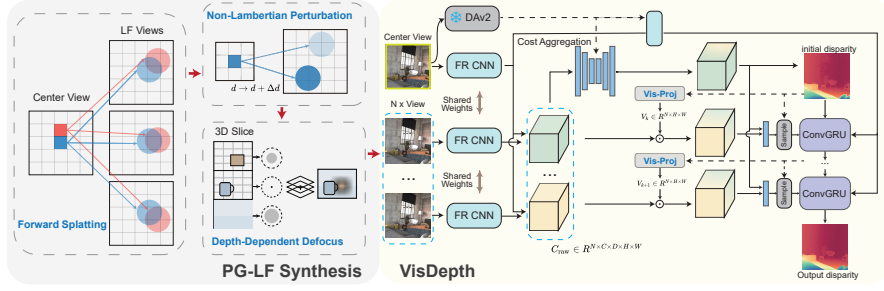


Fig. 2: Overview of our proposed **SynLF** framework. **Left:** PG-LF synthesis generates supervisory light fields on-the-fly from RGB-D through forward splatting, depth-dependent defocus, and non-Lambertian perturbation. **Right:** VisDepth first predicts an initial disparity using monocular priors and multi-view cost aggregation, and then iteratively updates disparity through explicit visibility-guided reasoning.

that directly disrupt disparity correspondence in real captures: given a center-view RGB-D image, it produces geometrically consistent sub-aperture views with occlusion-aware splatting, while simulating depth-dependent defocus and non-Lambertian perturbations.

Built upon this supervision, VisDepth learns to estimate disparity from input light fields through prior-initialized, visibility-aware iterative estimation. The monocular prior provides a stable initialization, while the iterative stage actively reuses the depth-aware projection from PG-LF synthesis to compute per-view visibility. By explicitly filtering incompatible observations through this shared geometric operator, VisDepth progressively improves disparity estimation under occlusions, weak textures, and view-dependent appearance changes.

At inference time, PG-LF synthesis is no longer needed. The trained VisDepth model is directly applied to real light field captures to estimate center-view disparity. Crucially, because the network is trained solely to extract geometric correspondences from micro-baseline disparities rather than memorize dataset-specific scales, its output can be converted to metric depth via the camera-specific disparity-to-depth mapping calibrated offline from the real light field system. In this way, SynLF transfers supervision from large-scale RGB-D data to real light field depth estimation and directly generalizes to real captures without fine-tuning.

3.2 Physics-Grounded Light Field Synthesis

The design of PG-LF synthesis is guided by a task-driven principle: we selectively model the imaging effects that directly corrupt the core cue exploited by light field depth estimation—cross-view disparity correspondence—rather than pursuing generic photorealistic simulation. Light field depth estimation fundamentally relies on establishing consistent correspondences across sub-aperture views. Three factors systematically disrupt this process in real captures: (1) *occlusion*,

where foreground objects block background content across views; (2) *depth-dependent defocus*, where out-of-focus blur attenuates the very textures used for correspondence matching; and (3) *non-Lambertian appearance*, where specular or transparent surfaces cause the apparent depth to diverge from the physical surface depth. PG-LF synthesis addresses each of these through geometry-consistent forward splatting, disparity-conditioned defocus, and targeted non-Lambertian perturbation, respectively.

Geometry-Consistent View Synthesis To overcome the scarcity of large-scale light field supervision, we synthesize sub-aperture views directly from RGB-D images instead of relying on fully rendered light field datasets. Given a central RGB image and its depth map Z_c , we first convert depth into light-field disparity

$$d = \frac{\alpha}{Z_c} - \beta, \quad (1)$$

where α controls the disparity scale, and β sets the zero-disparity reference corresponding to the focal plane in a setup-dependent manner.

For a target sub-aperture view indexed by angular coordinates (u_i, v_i) , each pixel (x_c, y_c) in the central view is projected according to

$$x_i = x_c + u_i d, \quad y_i = y_c + v_i d. \quad (2)$$

This forward projection explicitly establishes the geometric relationship between the center view and neighboring views. Unlike image-based neural synthesis, this formulation ties the generated sub-aperture views directly to the underlying scene depth.

A key challenge is visibility: multiple pixels from different depths may be projected to the same target location. To resolve this, we adopt a depth-aware forward splatting strategy that explicitly models occlusion. When collisions occur, we retain the closest sample in a z-buffer-like manner,

$$I_i(\mathbf{p}) = I_c(\mathbf{q}^*), \quad \mathbf{q}^* = \arg \min_{\mathbf{q} \in \mathcal{S}_i(\mathbf{p})} Z_c(\mathbf{q}), \quad (3)$$

where $\mathcal{S}_i(\mathbf{p})$ denotes the set of central-view pixels projected to location $\mathbf{p}$ in target view i . The visible samples are then bilinearly splatted to neighboring grid locations to reduce discretization artifacts and preserve smooth angular transitions.

Compared with backward warping, which can produce incorrect visibility ordering, depth-aware forward splatting resolves occlusions explicitly. To prevent shortcut learning from disocclusion artifacts, we apply stochastic gap handling using randomly selected filling strategies. As validated in Fig. 3 (Left), the synthesized EPIs exhibit highly consistent disparity-induced line structures and occlusion patterns despite visual domain differences from real captures.

Modeling Non-Ideal Imaging Effects

Depth-Dependent Defocus. Standard synthetic RGB-D datasets are typically rendered under an all-in-focus pinhole model, whereas real light field cameras exhibit pronounced depth-dependent defocus blur. We specifically model defocus because its CoC magnitude is directly coupled to scene disparity (Eq. 1), meaning it systematically attenuates the very correspondence cues that drive depth estimation. Other optical effects such as motion blur or cat-eye bokeh, while present in practice, are not primarily disparity-coupled and thus less critical for bridging the sim-to-real gap. We model defocus using a disparity-conditioned CoC formulation.

Under the thin-lens model, the CoC magnitude is determined by the mismatch between the scene depth z and the focal-plane depth z_f . Since light-field disparity is affine to inverse depth ($d = \alpha/z - \beta$), aligning the zero-disparity plane with the focal plane yields a direct mapping:

$$c(z) \propto \left| \frac{1}{z} - \frac{1}{z_f} \right| \implies c(z) \propto |d|, \quad (4)$$

indicating that the defocus magnitude is directly determined by disparity under fixed imaging settings.

In practice, we implement defocus via disparity slicing. We discretize the disparity map into bins, assign each bin a corresponding Gaussian kernel, and blur the image layer by layer. The blurred layers are then composited in back-to-front order to preserve correct occlusion boundaries and prevent foreground textures from bleeding into the background.

Non-Lambertian Perturbation. Real scenes often violate the Lambertian assumption due to specular reflections and transparent media. Physically, on such surfaces, the apparent depth perceived through multi-view matching diverges from the true surface geometry: reflections appear behind the surface, while transmitted light shifts toward background depth. Lambertian warping misses this discrepancy by tying radiance strictly to the surface. To mimic such effects without corrupting the underlying geometry, we perturb the disparity used for sub-aperture view synthesis, so that the projected source samples are intentionally displaced from their geometrically correct locations. This produces view-inconsistent reprojections that simulate the failure of strict photometric matching under non-Lambertian conditions.

To avoid conflating this perturbation with true geometric discontinuities, it is applied only to regions that are locally depth-flat yet visually textured, while excluding depth boundaries and weak-texture areas. Importantly, only the synthesized sub-aperture views are perturbed, whereas the original center view remains unchanged. As a result, trivial appearance matching across views is discouraged, and the model is forced to rely more on stable geometric cues from the synthesized light field together with reliable center-view cues, improving generalization to non-Lambertian scenes.

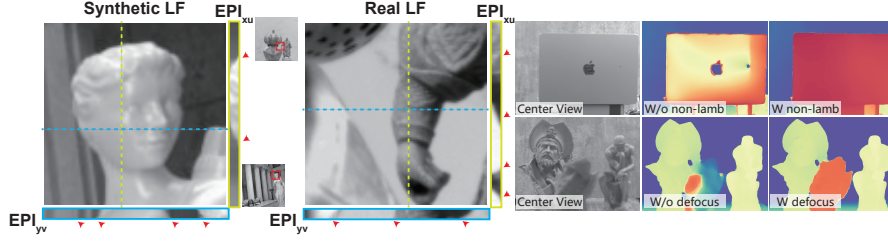


Fig. 3: Left: Center-view patches and their EPIs from a synthesized light field (RGB-D $\rightarrow$ PG-LF) and a real capture. Despite domain differences, the EPIs exhibit similar disparity-induced line structures and occlusion patterns. **Right:** Effects of the two non-ideal imaging components in PG-LF synthesis, evaluated in a strict zero-shot setting on real captures (no fine-tuning). *W/o non-lamb* denotes training without non-Lambertian perturbation, while *W/o defocus* denotes training without depth-dependent defocus. Non-Lambertian perturbation improves robustness to specular reflections, and depth-dependent defocus improves estimation on heavily blurred foreground objects.

3.3 Visibility-Aware LF Disparity Estimation

Prior-Guided Disparity Initialization Explicit visibility reasoning requires a reasonably accurate initial disparity estimate. In practice, directly starting iterative estimation from raw multi-view matching is highly unstable, especially in textureless and non-Lambertian regions where light field correspondence is ambiguous. We therefore introduce a lightweight prior-guided disparity initialization stage whose role is not to solve the full task, but to provide a reliable starting point for subsequent visibility-aware iterative estimation.

Specifically, we first extract full-resolution multi-view features using a full-resolution CNN encoder and construct a raw light field correlation volume

$$C_{\text{raw}} = \text{Corr}(f_c, \{f_v\}), \quad (5)$$

where f_c denotes the center-view feature and $\{f_v\}$ the source-view features. Since this initial matching stage does not explicitly model occlusions, we inject semantic priors from the center view, processed by a frozen monocular foundation model (Depth Anything V2-Small [32]), to obtain dense prior features F_{mono} . Because the foundation model provides structural context but inherently lacks absolute scale, we fuse F_{mono} with the raw correlation volume via a lightweight 3D cost aggregation network:

$$C_{\text{reg}} = \Phi(C_{\text{raw}}, F_{\text{mono}}), \quad (6)$$

where $\Phi(\cdot)$ denotes the regularization operator. The resulting regularized volume is converted into a probability distribution over disparity levels, and an initial disparity map d_0 is obtained by soft-argmin regression:

$$d_0 = \sum_d d \cdot \text{softmax}(C_{\text{reg}}(d)). \quad (7)$$

Explicit Visibility-Guided Iterative Estimation Starting from the initial estimate d_0 , we iteratively update disparity by explicitly modeling per-view visibility under the current geometry. Unlike conventional estimation schemes that aggregate source-view evidence through implicit or learned weighting alone, we explicitly determine which views are geometrically visible at each iteration and use only those valid observations for cost aggregation.

At iteration k , given the current disparity estimate d_k , we project the center-view geometry to each source view using the same depth-aware forward projection as in PG-LF synthesis. This produces per-view visibility maps $V_k^{(v)} \in [0, 1]$, which indicate whether a source-view observation is geometrically valid under the current disparity hypothesis. These visibility maps are then used to aggregate the raw per-view correlation costs before occlusion-inconsistent evidence is mixed during volume regularization:

$$C_{\text{vis}}^k = \frac{\sum_v C_{\text{raw}}^{(v)} \odot V_k^{(v)}}{\sum_v V_k^{(v)}}. \quad (8)$$

By explicitly suppressing occluded or inconsistent source views, this masked aggregation isolates valid photometric evidence, producing a visibility-aware geometric cue C_{vis}^k .

To preserve the rich structural context extracted during the initial matching stage, we fuse this dynamically masked cost C_{vis}^k with the foundational geometric feature volume C_{reg} (derived from the prior-guided initialization stage via Φ). Specifically, they are concatenated and processed by a 3D convolutional fusion block $\mathcal{F}_{\text{fuse}}$:

$$G_k = \mathcal{F}_{\text{fuse}}([C_{\text{reg}}, C_{\text{vis}}^k]), \quad (9)$$

where $[\cdot, \cdot]$ denotes channel-wise concatenation, and G_k serves as the fused dense geometric volume for the current iteration.

At each step, we sample a local 1D neighborhood of the fused volume G_k centered at the current disparity d_k , retrieving localized geometric matching features along the signed disparity dimension—reflecting the bidirectional nature of light field disparity around the focal plane. Because the micro-baseline disparity range of light field cameras is compact (typically tens of levels, compared to hundreds in stereo matching), the dense cost volume can be directly indexed without the pooled correlation pyramids required by large-range optical flow or stereo methods [18]. These features, concatenated with monocular-enriched context, are processed by a multi-scale ConvGRU that updates coarse-to-fine hidden states and predicts a scalar disparity residual Δd_k , yielding $d_{k+1} = d_k + \Delta d_k$. While the individual gated recurrence follows standard ConvGRU practice, the overall update block is substantially adapted for light field disparity: visibility is recomputed from the current disparity at each iteration and used to re-aggregate the raw cost volume (Eq. 8–9), the resulting visibility-aware geometric volume is queried along the local signed disparity dimension, and multi-scale hidden states interact across scales. This iterative coupling between disparity updates and visibility-aware matching allows later updates to leverage increasingly con-

sistent occlusion-aware features, helping preserve depth discontinuities under partial occlusion and view-dependent appearance changes.

3.4 Training Objective

We supervise both the prior-guided initialization and the iterative estimation using a combination of absolute disparity loss and scale-and-shift invariant gradient regularization:

$$\mathcal{L}_{\text{init}} = \|d_0 - d\|_1, \quad (10)$$

$$\mathcal{L}_{\text{gru}} = \sum_{k=1}^K \gamma^{K-k} \|d_k - d\|_1, \quad (11)$$

$$\mathcal{L}_{\text{grad}} = \sum_{l=1}^L \left(\left\| \nabla_x^l (\hat{d}_K - \hat{d}) \right\|_1 + \left\| \nabla_y^l (\hat{d}_K - \hat{d}) \right\|_1 \right), \quad (12)$$

$$\mathcal{L} = \mathcal{L}_{\text{init}} + \mathcal{L}_{\text{gru}} + \lambda \mathcal{L}_{\text{grad}}, \quad (13)$$

where d_0 is the initialized disparity, $\{d_k\}_{k=1}^K$ are the iteratively updated disparities, and d is the ground-truth disparity. $\hat{d}_K$ and $\hat{d}$ denote the final prediction and ground truth after applying the scale-and-shift invariant alignment in [20]. The gradient term enforces multi-scale structural consistency, and γ^{K-k} assigns larger weights to later iterations.

4 Experiments

4.1 Experimental Setup

We implement SynLF in PyTorch and train it on the Hypersim dataset, which contains 77,400 photorealistic RGB-D images of diverse indoor scenes with accurate per-pixel depth annotations. The disparity is computed from depth using Eq. (1), with $\alpha = 16$ and $\beta = -9$. Samples containing invalid depth values are removed, leaving 74,013 images for training.

During training, light field supervision is synthesized on-the-fly from RGB-D inputs using the proposed physics-grounded synthesis pipeline. Unless otherwise specified, the input image is randomly cropped to 768×768 . We train SynLF with the AdamW optimizer using a total batch size of 16. Leveraging the rich texture and geometric diversity of large-scale RGB-D data, convergence is fast and 8K optimization steps suffice (approximately two passes over the Hypersim training set; see the supplementary material for the training-loss trajectory). The learning rate is initialized at 5×10^{-5} , linearly warmed up to 5×10^{-4} , and then decayed polynomially to 1×10^{-5} . The disparity range is set to $d_{\min} = -9$ and $d_{\max} = 20$. During both training and inference, the ConvGRU module performs 24 iterative updates.

For quantitative evaluation, we report four standard depth metrics: mean absolute error (MAE, in mm), root mean squared error (RMSE, in mm), absolute relative error (AbsRel, in %), and threshold accuracy δ_1 .

Table 1: Zero-shot generalization on our real-world indoor light field dataset. We compare SynLF with prior light field methods and controlled variants of our model.

Method	MAE (mm)↓	RMSE (mm)↓	AbsRel (%) ↓	δ_1 (%) ↑
OAVC [14]	146.99	291.9	7.02	88.75
OACC-Net [30]	113.66	200.27	5.99	93.33
OPAL [17]	141.77	202.35	7.39	93.47
PlaneNet [7]	91.31	164.91	4.48	96.03
ESMNet [6]	110.86	208.99	5.36	92.20
Ours (HCI-trained)	52.95	148.40	3.01	95.01
Ours w/o Prior	35.75	158.50	1.63	98.50
Ours w/o Vis	39.03	160.34	1.90	98.04
SynLF	30.05	144.34	1.38	98.77

4.2 Real-World Light Field Dataset

To evaluate zero-shot generalization under realistic imaging conditions, we capture a real-world light field dataset using a custom acquisition system consisting of an industrial camera with a 9×9 micro-lens array, paired with a high-precision structured-light camera for metric depth. All scenes are static during capture. Using calibrated light-field intrinsics and the extrinsic transformation between the two sensors, we reproject the structured-light depth into the light-field center view to obtain per-pixel metric depth annotations.

For evaluation, we collected 173 annotated indoor captures spanning diffuse objects and challenging non-Lambertian materials (*e.g.*, specular metals, plastics, glass, and reflective displays), as well as 30 outdoor light field captures without annotations for qualitative evaluation. Since structured-light sensing may fail under severe transparency or strong reflections, we restrict quantitative evaluation to valid annotated pixels using the sensor validity mask; all methods are evaluated on the same valid regions for fair comparison. For metric evaluation, predicted center-view disparity is converted to depth using a camera-specific disparity-to-depth mapping calibrated offline. The geometric calibration yields sub-pixel mean reprojection error; details are provided in the supplementary material.

Ethics and privacy. Our indoor dataset contains only static objects and does not involve human participants. For outdoor qualitative examples, bystanders may incidentally appear in some captures; we do not use any personal identity information and we anonymize the figures by blurring/cropping identifiable regions (*e.g.*, faces, bodies, license plates). Any data released will undergo privacy review and anonymization.

4.3 Zero-Shot Generalization on Real-World Data

Comparison with Prior Light Field Methods We evaluate SynLF in a strict zero-shot setting on our real-world indoor light field dataset and com-

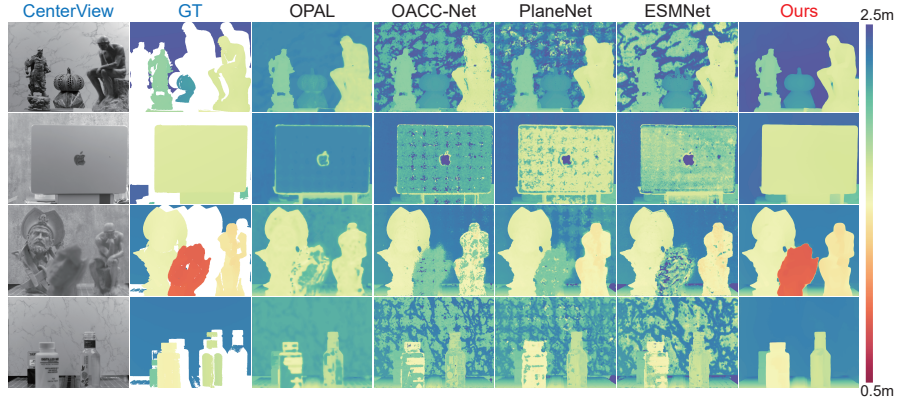


Fig. 4: Qualitative zero-shot comparisons on our real-world indoor light field dataset. In challenging scenes (*e.g.*, weak-texture, reflections, strong defocus, and semi-transparency), prior methods often produce noisy or inconsistent depth, whereas SynLF yields cleaner predictions with sharper structural boundaries.

pare it with representative prior light field depth estimation methods, including optimization-based approaches [14] and learning-based approaches [6, 7, 17, 30]. For all baselines, we use the officially released pretrained models when available and directly test on our captures without fine-tuning or test-time adaptation.

As shown in Tab. 1, SynLF achieves the best performance across all metrics, reducing MAE by 61.26 mm over PlaneNet [7] (91.31→30.05 mm) and by 80.81 mm over ESMNet [6] (110.86→30.05 mm). Even compared with the identical VisDepth architecture trained on the rendered HCI dataset [15], SynLF reduces MAE from 52.95 mm to 30.05 mm and AbsRel from 3.01% to 1.38%, suggesting that real-world generalization is largely limited by supervision mismatch rather than model capacity. The controlled variants further validate the necessity of our two model designs: removing the monocular *Prior* substantially degrades stability in ambiguous regions, while removing *Vis* (explicit visibility reasoning) reduces robustness around occlusions and depth discontinuities, confirming that a strong initialization and iterative visibility reasoning are both critical for accurate metric estimation on real captures.

Fig. 4 shows qualitative comparisons on four challenging regimes: weak-texture backgrounds, reflective textureless surfaces, strong foreground defocus, and semi-transparent objects. Prior methods often appear plausible on well-textured Lambertian regions, but degrade rapidly once correspondence becomes unreliable, leading to speckle noise, boundary leakage, and fragmented object structure. This indicates that the core difficulty is not the absence of multi-view cues, but the inability to identify trustworthy observations and propagate them coherently into ambiguous areas.

SynLF addresses this from both the data and model perspectives: PG-LF synthesis exposes the model to simulated non-Lambertian effects and defocus during training, while VisDepth explicitly filters per-view evidence via visibility

Table 2: Reciprocal control: Data vs. Architecture on our real-world dataset.

Architecture	Training Data	MAE↓
PlaneNet	HCI	91.31
PlaneNet	PG-LF	53.11
VisDepth	HCI	52.95
VisDepth	PG-LF	30.05

Table 3: Material-stratified depth error on foreground-separable real-world scenes, grouped by dominant surface type.

Material	#Scenes	Valid%	MAE↓
Diffuse	78	59.8	8.83
Specular	31	54.8	21.76
Transparent	34	37.2	39.81

Table 4: Cross-dataset evaluation on public light field benchmarks.

Method	Training Data	HCI 4D		Inria Dense	
		MSE×100	BP _{0.07}	MSE×100	BP _{0.07}
PlaneNet	HCI	0.69	4.19	1.97	5.24
OACC-Net	HCI	0.92	3.78	2.74	2.82
OAVC	—	2.23	8.75	2.17	4.74
SynLF (ours)	PG-LF	2.08	9.36	2.35	6.45

reasoning. Together, they produce cleaner surfaces, sharper discontinuities, and more coherent object-level geometry in challenging real scenes.

Disentangling Data and Architecture Contributions To isolate the respective contributions of PG-LF synthesis and the VisDepth architecture, we conduct a reciprocal control experiment by training both PlaneNet [7] and VisDepth on either HCI [15] or PG-LF supervision and evaluating on the same real-world test set. As shown in Tab. 2, replacing HCI with PG-LF improves both architectures by 42–43% in MAE: 91.31→53.11 for PlaneNet and 52.95→30.05 for VisDepth. This confirms that PG-LF synthesis provides a substantial data-level benefit independent of model design. Meanwhile, VisDepth consistently outperforms PlaneNet under identical supervision, contributing an orthogonal architectural gain of 23.06 mm.

We further examine whether the structured-light validity mask introduces evaluation bias by stratifying all 143 cleanly separable scenes by dominant surface material. As reported in Tab. 3, the GT valid-pixel ratio decreases and SynLF MAE increases monotonically with material difficulty: diffuse (59.8 %, 8.83 mm), specular (54.8 %, 21.76 mm), and transparent (37.2 %, 39.81 mm). This confirms that our metrics faithfully capture material-dependent challenges even under partial GT coverage.

Cross-Dataset Generalization on Public Benchmarks To calibrate our results against the broader literature, we evaluate SynLF in a zero-shot setting on two public synthetic light field benchmarks—HCI 4D Light Field [15] and Inria Dense Light Field [25]—and compare with both learning-based (PlaneNet,

Table 5: Effect of depth-dependent defocus and non-Lambertian perturbation.

Defocus	NL Pert.	MAE↓	AbsRel↓
×	×	50.01	3.41
×	✓	43.25	2.40
✓	×	46.28	2.51
✓	✓	30.05	1.38

**Fig. 5:** Qualitative zero-shot results on outdoor real-world light field scenes.

OACC-Net) and optimization-based (OAVC) methods. As shown in Tab. 4, HCI-trained methods achieve the lowest MSE on their training domain but degrade substantially on Inria (*e.g.*, PlaneNet 0.69→1.97, OACC-Net 0.92→2.74), whereas OAVC and SynLF maintain consistent MSE across both benchmarks. SynLF slightly outperforms OAVC on HCI (2.08 vs. 2.23) while remaining competitive on Inria, supporting the cross-dataset generalization benefit of PG-LF synthesis. Per-scene breakdowns are provided in the supplementary material.

Qualitative Generalization to Outdoor Scenes Beyond indoor scenes with depth annotations, we further evaluate SynLF on outdoor real-world light field images captured by our custom-built system. Although these scenes do not have ground-truth depth, the qualitative disparity results in Fig. 5 show that SynLF remains stable under broader scene variations and complex real-world imaging conditions, producing coherent disparity estimates with clear structural boundaries and strong robustness to natural outdoor appearance changes.

4.4 Ablation Study

For the ablation study, we report MAE (mm) and AbsRel (%) for brevity. As shown in Tab. 5, both non-Lambertian perturbation and depth-dependent defocus independently improve zero-shot transfer: the former discourages shortcut learning from strict photometric matching and improves robustness to reflective regions (Fig. 3), while the latter bridges the mismatch between all-in-focus RGB-D and the pronounced blur of real light field cameras, particularly benefiting heavily defocused foregrounds (Fig. 3). Notably, the largest gain comes from combining both components, suggesting that robust real-world transfer requires exposure to both appearance inconsistency and blur-induced ambiguity during training.

We further ablate two design choices in the supplementary material. Increasing GRU iterations from 6 to 24 steadily reduces MAE from 36.39 to 30.05 mm, confirming that iterative visibility-aware refinement progressively sharpens geometry in occluded or ambiguous regions. In contrast, scaling the frozen prior from DAv2-Small to DAv2-Large does not improve results, indicating that SynLF’s

accuracy stems from the visibility-aware iterative estimation rather than the strength of the monocular prior alone.

5 Discussion

Our results suggest that the main bottleneck of real-world light field depth estimation lies not only in model design, but in the mismatch between idealized supervision and real imaging conditions. SynLF shows that large-scale RGB-D data can be transformed into effective light field supervision when the synthesis process preserves multi-view geometry and introduces physically motivated non-ideal effects. More broadly, PG-LF synthesis shifts the paradigm from collecting large light field datasets to designing task-driven synthesis that preserves the geometric structure required by the downstream task while enriching the training distribution with realistic imaging variations.

At the same time, our synthesis model remains an approximation of real light field formation. Although it captures key factors such as visibility, defocus, and non-Lambertian perturbation, it does not explicitly model the full complexity of real optical effects, or severe material-dependent appearance changes. In addition, our quantitative evaluation focuses on static indoor scenes, and ground-truth depth in highly reflective or transparent regions remains inherently challenging even with active sensors.

These limitations point to several directions for future work, including richer optical modeling for light field synthesis, extension to more diverse outdoor or dynamic scenes, and broader applications of physics-grounded synthetic supervision to other light field perception tasks beyond depth estimation.

6 Conclusion

We presented SynLF, a framework for robust single-lens metric depth estimation from light field cameras that combines physics-grounded light field synthesis with visibility-aware iterative disparity estimation. By generating geometry-consistent supervisory light fields on the fly from large-scale RGB-D data, SynLF alleviates the scarcity of conventional light field supervision, while the proposed VisDepth explicitly reasons about multi-view visibility to improve robustness under occlusion, weak texture, reflections, and defocus. Extensive experiments on challenging real-world captures demonstrate strong zero-shot generalization and accurate metric depth estimation without fine-tuning. These results suggest that physics-grounded light field synthesis from RGB-D provides a scalable and effective path toward real-world light field depth estimation.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant No. 2023YFB3209700, and in part by the National Natural Science Foundation of China under Grant No. 62322110.

References

1. Adelson, E.H., Wang, J.Y.: Single lens stereo with a plenoptic camera. *IEEE transactions on pattern analysis and machine intelligence* **14**(2), 99–106 (1992)
2. Bhat, S.F., Alhashim, I., Wonka, P.: AdaBins: Depth estimation using adaptive bins. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 4009–4018 (2021)
3. Bishop, T.E., Favaro, P.: The light field camera: Extended depth of field, aliasing, and superresolution. *IEEE transactions on pattern analysis and machine intelligence* **34**(5), 972–986 (2011)
4. Bochkovskiy, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S., Koltun, V.: Depth Pro: Sharp monocular metric depth in less than a second. In: *International Conference on Learning Representations* (2025), <https://openreview.net/forum?id=aueXfY0Clv>, accessed 29 June 2026
5. Chen, J., Zhang, S., Lin, Y.: Attention-based multi-level fusion network for light field depth estimation. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 35, pp. 1009–1017 (2021)
6. Chen, R., Sheng, H., Cong, R., Yang, D., Cui, Z., Wang, S., Ke, W.: Stereo matching on epipolar plane image for light field depth estimation via oriented structure. *Engineering Applications of Artificial Intelligence* **151**, 110608 (2025)
7. Chen, R., Sheng, H., Yang, D., Cui, Z., Cong, R.: Surface-continuous scene representation for light field depth estimation via planarity prior. *IEEE Transactions on Circuits and Systems for Video Technology* **35**(5), 5051–5066 (2024)
8. Chen, T., Tippur, M., Wu, S., Kumar, V., Adelson, E., Agrawal, P.: Visual dexterity: In-hand dexterous manipulation from depth. In: *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems* (2023)
9. Dong, X., Garratt, M.A., Anavatti, S.G., Abbass, H.A.: Towards real-time monocular depth estimation for robotics: A survey. *IEEE Transactions on Intelligent Transportation Systems* **23**(10), 16940–16961 (2022)
10. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems* **27** (2014)
11. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. pp. 2002–2011 (2018)
12. Gao, R., Liu, Y., Xiao, Z., Xiong, Z.: Diffusion-based light field synthesis. In: *European Conference on Computer Vision*. pp. 1–19. Springer (2024)
13. Hahne, C., Aggoun, A., Velisavljevic, V., Fiebig, S., Pesch, M.: Baseline and triangulation geometry in a standard plenoptic camera. *International Journal of Computer Vision* **126**(1), 21–35 (2018)
14. Han, K., Xiang, W., Wang, E., Huang, T.: A novel occlusion-aware vote cost for light field depth estimation. *IEEE transactions on pattern analysis and machine intelligence* **44**(11), 8022–8035 (2021)
15. Honauer, K., Johannsen, O., Kondermann, D., Goldluecke, B.: A dataset and evaluation methodology for depth estimation on 4d light fields. In: *Asian Conference on Computer Vision*. pp. 19–34. Springer (2016)
16. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: Metric3D v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024)

17. Li, P., Zhao, J., Wu, J., Deng, C., Han, Y., Wang, H., Yu, T.: OPAL: Occlusion pattern aware loss for unsupervised light field disparity estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **46**(2), 681–694 (2023)
18. Lipson, L., Teed, Z., Deng, J.: RAFT-Stereo: Multilevel recurrent field transforms for stereo matching. In: 2021 International Conference on 3D Vision (3DV). pp. 218–227. IEEE (2021)
19. Piccinelli, L., Sakaridis, C., Yang, Y.H., Segu, M., Li, S., Abbeloos, W., Van Gool, L.: UniDepthV2: Universal monocular metric depth estimation made simpler. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **48**(3), 2354–2367 (2026). <https://doi.org/10.1109/TPAMI.2025.3628473>
20. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence* **44**(3), 1623–1637 (2020)
21. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10912–10922 (2021)
22. Roriz, R., Cabral, J., Gomes, T.: Automotive LiDAR technology: A survey. *IEEE Transactions on Intelligent Transportation Systems* **23**(7), 6282–6297 (2021)
23. Ruan, L., Chen, B., Lam, M.L.: Light field synthesis from a single image using improved Wasserstein generative adversarial network. In: 39th Eurographics Conference (EG 2018). pp. 19–20. Eurographics Association (2018)
24. Sheng, H., Cong, R., Yang, D., Chen, R., Wang, S., Cui, Z.: UrbanLF: A comprehensive light field dataset for semantic segmentation of urban scenes. *IEEE Transactions on Circuits and Systems for Video Technology* **32**(11), 7880–7893 (2022)
25. Shi, J., Jiang, X., Guillemot, C.: A framework for learning depth from a flexible subset of dense and sparse light field views. *IEEE Transactions on Image Processing* **28**(12), 5867–5880 (2019)
26. Shin, C., Jeon, H.G., Yoon, Y., Kweon, I.S., Kim, S.J.: EPINET: A fully-convolutional neural network using epipolar geometry for depth from light field images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4748–4757 (2018)
27. Srinivasan, P.P., Wang, T., Sreelal, A., Ramamoorthi, R., Ng, R.: Learning to synthesize a 4D RGB-D light field from a single image. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2243–2251 (2017)
28. Tsai, Y.J., Liu, Y.L., Ouhyoung, M., Chuang, Y.Y.: Attention-based view selection networks for light-field disparity estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 12095–12103 (2020)
29. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: MoGe: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5261–5271 (2025)
30. Wang, Y., Wang, L., Liang, Z., Yang, J., An, W., Guo, Y.: Occlusion-aware cost constructor for light field depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19809–19818 (2022)
31. Williem, W., Park, I.K.: Robust light field depth estimation for noisy scene with occlusion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4396–4404 (2016)

- 32. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth Anything V2. *Advances in Neural Information Processing Systems* **37**, 21875–21911 (2024)
- 33. Zhang, S., Sheng, H., Li, C., Zhang, J., Xiong, Z.: Robust depth estimation for light field via spinning parallelogram operator. *Computer Vision and Image Understanding* **145**, 148–159 (2016)