S2D-LFE: Sparse-to-Dense Light Field Event Generation

Yutong Liu Wenming Weng Yueyi Zhang Zhiwei Xiong*
University of Science and Technology of China
{ustclyt, wmweng}@mail.ustc.edu.cn {zhyuey, zwxiong}@ustc.edu.cn

Abstract

In this paper, we present S2D-LFE, an innovative approach for sparse-to-dense light field event generation. For the first time to our knowledge, S2D-LFE enables controllable novel view synthesis only from sparse-view light field event (LFE) data, and addresses three critical challenges for the LFE generation task: simplicity, controllability, and consistency. The simplicity aspect eliminates the dependency on frame-based modality, which often suffers from motion blur and low frame-rate limitations. The controllability aspect enables precise view synthesis under sparse LFE conditions with view-related constraints. The consistency aspect ensures both cross-view and temporal coherence in the generated results. To realize S2D-LFE, we develop a novel diffusion-based generation network with two key components. First, we design an LFE-customized variational auto-encoder that effectively compresses and reconstructs LFE by integrating cross-view information. Second, we design an LFE-aware injection adaptor to extract comprehensive geometric and texture priors. Furthermore, we construct a large-scale synthetic LFE dataset containing 162 one-minute sequences using simulator, and capture a real-world testset using our custom-built sparse LFE acquisition system, covering diverse indoor and outdoor scenes. Extensive experiments demonstrate that S2D-LFE successfully generates up to 9×9 *dense LFE from* 2×2 *sparse inputs and outperforms existing methods on both synthetic and real-world data. The datasets and code are available at* https://github.com/Yutong2022/S2D-LFE.

1. Introduction

Event cameras capture per-pixel brightness changes asynchronously with microsecond temporal resolution [1], while Light Field (LF) cameras record both the accumulated intensity and direction of light rays, thereby revealing scene geometric structure [48]. The integration of these technologies leads to Light Field Event (LFE) cameras, which capture multi-view event streams arranged in a fixed baseline array. This combination enables simultaneous recording of high-quality spatial depth information and dynamic motion data with exceptional temporal resolution. Such capabilities demonstrate significant potential for various applications, including 6-degree-of-freedom virtual reality [23] and real-time dynamic depth estimation [40]. The rich spatio-temporal information provided by LFE data enables effective capture of structural information in high-speed dynamic scenes.

In this paper, we aim to address a largely unexplored task of LFE generation. To date, the only relevant work is Ev-LFV [29], which combines an RGB LF camera with an event camera aligned to the LF center-view for LFE synthesis. However, this approach faces several critical challenges: (1) Simplicity. Ev-LFV's reliance on multi-view LF frame data as auxiliary information introduces complications. The RGB data is vulnerable to motion blur and exposure issues, potentially compromising LFE generation quality. Moreover, achieving precise spatial-temporal alignment between the LF and event cameras presents substantial technical challenges. (2) Controllability. The fixed view synthesis constraint of Ev-LFV prevents the generation of free viewpoints under sparse LFE conditions with view-related constraints. (3) Consistency. Due to the limitations of low frame-rate RGB data and insufficient structural modeling, Ev-LFV struggles to maintain both angular and temporal continuity in the generated LFE sequences.

To address these challenges, we propose S2D-LFE for **S**parse-to-**D**ense **L**ight **F**ield **E**vent generation, which is a novel approach for LFE generation. S2D-LFE demonstrates the capability to generate temporally consistent LFE from free viewpoints using only sparse LFE input, making it particularly suitable for real-world applications without the need for frame cameras. Our approach comprises two key components: an LFE-customized variational auto-encoder for efficient LFE compression and reconstruction, and a LFE-aware injection adapter integrated into a diffusion-based generation network that learns to generate the latent representation of LFE.

The first key component, LFE-VAE, is designed to effectively compress high-resolution LFE while preserving spatial and structural information. Specifically, we introduce

*Corresponding author

two significant improvements over conventional pretrained VAE architectures used in diffusion models [25, 26, 35, 36, 53]. First, we incorporate angular interaction blocks into the VAE decoder to fully exploit cross-view information within LFE. Second, we implement a view-enhancement mechanism that utilizes input sparse LFE views to improve the quality of reconstructed novel views. These enhancements effectively leverage the inherent cross-view information in LFE and mitigate compression artifacts, resulting in high-fidelity reconstruction. The second key component, LFE-adapter, is designed to utilize geometric and texture priors from the input sparse LFE to guide the diffusion process. Specifically, LFE-adapter generates pyramid condition signals that provide targeted conditioning at multiple resolution scales for the U-Net during diffusion. This multi-scale conditioning enables more precise injection of geometric and texture priors, thereby achieving superior cross-view consistency.

Moreover, we develop the first sparse LFE capture system to facilitate the acquisition of continuous light field event data. The system consists of four Davis346 cameras arranged in a 2×2 array configuration. Using this setup, we collect a diverse dataset comprising 25 sequences, including 15 indoor scenes and 10 autonomous driving scenarios. To support comprehensive model training, we additionally construct a large-scale synthetic dataset using Carla simulator [14], generating 162 one-minute dense LFE sequences.

The main contributions of this paper can be summarized as follows:

- We propose S2D-LFE, a novel approach for generating free-viewpoint and temporally consistent light field events from sparse-view LFE input.
- Our S2D-LFE incorporates two key components: an LFE-VAE for efficient LFE compression, and an LFE-adapter integrated into a diffusion-based generation network that reconstructs the latent representation of LFE.
- We develop the first sparse LFE capture system and collect a dataset of 25 real-world sparse LFE sequences. Additionally, we synthesize 162 dense LFE sequences using the Carla simulator to facilitate comprehensive training.
- Our approach outperforms existing methods on both synthetic and real-world datasets, while presenting enhanced robustness in generalizing to free-viewpoint synthesis.

2. Related Work

Multi-view Event Processing. While event cameras offer exceptional temporal resolution and low latency [1], monocular event data alone is insufficient for capturing comprehensive structural information of dynamic scenes. Various approaches have been proposed to address this limitation: 1) NeRF-based scene representation learning [4], 2) hybrid systems combining events with conventional cameras [9, 29], and 3) stereo event systems [10, 19, 40, 55]. NeRF-based methods, while promising, require per-scene training and multiple viewpoints for effective scene representation. For example, EvDNeRF [4] introduces an event-based dynamic NeRF pipeline for reconstructing event streams in rapidly deforming scenes, but its performance degrades significantly with limited viewpoints (< 8). Hybrid systems face challenges in reconciling the temporal resolution disparity between different modalities. For instance, Lu *et al.* [29] encounter practical limitations in real-world applications due to the substantial difference in temporal resolution between LF and event cameras. While stereo event systems can capture geometric information, the resulting geometry tends to be sparse. In contrast, the Light Field event modality offers distinct advantages by capturing dense structural information of dynamic scenes without requiring per-scene training, while enabling multiple angular resolution generation.

Light Field View Synthesis. The reconstruction of dense Light Fields from sparse views, commonly known as LF angular super-resolution, remains an active research area. Existing approaches can be broadly categorized into two groups: depth-dependent [5, 24, 43, 52] and depth-independent methods [12, 21, 23, 44]. Depth-dependent methods rely on precise disparity estimation but are vulnerable to occlusions, while depth-independent methods often struggle with generalization across varying disparity ranges. Recent advances in Neural Light Field representations [2, 3, 15, 28, 38, 42] have brought significant improvements to LF view synthesis by addressing limitations of conventional angular super-resolution methods. Compared with the widely adopted learnable 3D representation methods [7, 11, 34, 41], these approaches can more effectively exploit the spatial priors inherent in LF. However, these NeRF-based approaches are constrained by their requirement for numerous input views and scene-specific training. In contrast, our proposed S2D-LFE conquers these limitations by enabling free-viewpoint synthesis from sparse inputs without the need for per-scene training, while maintaining high recovery quality and temporal consistency.

Diffusion-based Image Generation. Diffusion models have shown remarkable progress in novel view synthesis [6, 8, 20, 39, 46, 50, 51]. For example, NeRDi [20] leverages language priors for multi-view synthesis, effectively connecting image semantics to appearance reconstruction. These advances demonstrate the robust capability of diffusion models in modeling multi-view distributions. Furthermore, diffusion models have achieved impressive results across diverse data modalities, including depth [26], point clouds [30], optical flow [26, 37], and human pose [25]. Motivated by this versatility and strong fitting capability across different modalities, we adopt diffusion models for dense LFE generation.

3. Real-world and Synthetic LFE Datasets

3.1. LFE Capture System and Real-world Dataset

System Setup. To facilitate the collection of real-world LFE data and validate our approach, we develop a custom LFE capture system. Given the lack of commercially available LFE cameras, our system design prioritizes precise spatial alignment and temporal synchronization while maintaining cost-effectiveness. The system, shown in Fig. 1, consists of a 2×2 array of Davis346 cameras mounted on a square aluminum alloy frame, with uniform horizontal and vertical baseline lengths of 6 cm. To ensure consistent imaging characteristics across views, all four cameras are equipped with identical lenses and configured with matching parameters, including focal length and exposure time. For precise temporal synchronization, we implement a master-slave configuration using external cables (depicted as black lines in Fig.1). The upper-left DAVIS346 serves as the master camera, transmitting trigger signal pulses to the three slave cameras through these cables. This system captures event data at a spatial resolution of 346 × 260 pixels, which can be readily upgraded to a higher resolution by updating event cameras. Detailed parameter configurations are provided in the supplementary materials.

Camera Calibration. Precise camera calibration is crucial for accurate Light Field Event capture, as it ensures geometric consistency across multiple views and enables reliable reconstruction of scene structure. This is particularly important for our system where accurate spatial relationships between cameras directly impact the quality of captured LFE data. We utilize the Kalibr toolbox [16] to estimate camera intrinsics and extrinsics by capturing fixed-rate frames of a checkerboard pattern. Since the DAVIS346 cameras output both frames and event streams through the same optical system, the calibration parameters obtained from frames are directly applicable to event streams. Following calibration, we perform systematic rectification of the captured sparse LFE. Using the upper left camera as reference, we first apply perspective transformation to project the upper right camera onto a virtual plane parallel to the reference camera's horizontal plane, establishing the baseline length between them. We then align the two lower cameras through perspective transformations to their corresponding positions in the sparse LF, ensuring uniform baseline lengths across the array. This comprehensive calibration and rectification process ensures accurate structural information capture of dynamic scenes.

Real-world LFE Dataset. Following system calibration and rectification, we collect a real LFE dataset comprising 25 dynamic scenes using our LFE capture system, which is termed "LFE-real". The dataset consists of two main categories: 15 indoor sequences captured by moving our camera array through various indoor environments (featuring subjects such as people, plants, and interior layouts), and 10 autonomous driving sequences recorded from displayed autonomous driving footage. Each sequence spans between 3 to 5 minutes, ensuring sufficient temporal coverage of the dynamic events. The details about scenes are in the supplementary material.

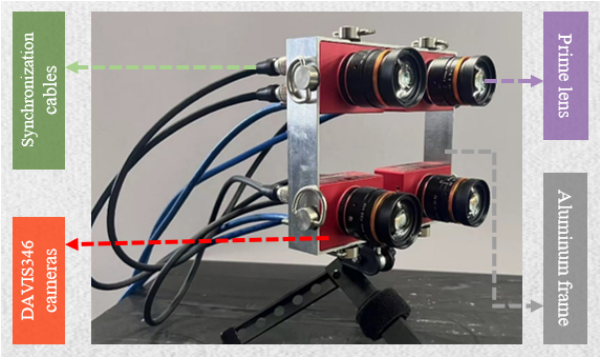

Figure 1. The proposed LFE capturing system.

3.2. Synthetic Dataset

Given that physical capture devices can only provide sparse view inputs without dense multi-view supervision, we leverage the Carla simulator [14] to generate a large synthetic LFE dataset for training and evaluation, which is termed "LFE-syn". We utilize 18 different maps provided by the Carla environment for data simulation. Each map is divided into 9 regions, with a designated starting point in each region. At these starting points, we deploy virtual vehicles equipped with 5×5 event camera arrays. The baseline lengths between adjacent cameras are randomly set between 3 cm and 9 cm. The event trigger thresholds are configured asymmetrically: 0.1 to 0.15 for positive events and -0.15 to -0.1 for negative events. Using autonomous navigation, each vehicle generates one-minute LFE sequences, resulting in a total of 162 synthetic sequences. Additionally, to facilitate the evaluation, our synthetic dataset share the same resolution with the real-world dataset. Detailed parameter configurations are provided in the supplementary materials.

4. S2D-LFE Generation Network

Our custom-built system captures sparse LFE with coarse geometric information of dynamic scenes. To enhance the structural representation capabilities, we propose S2D-LFE, a novel approach for synthesizing accurate dense LFEs from sparse inputs. We detail our approach in the following sections.

4.1. Overview

In our setting, each event stream view in the LFE is accumulated as an Event Count Map (ECM) [17] with shapes of $\mathbb{R}^{B\times TB\times 2\times H\times W}$, where B represents the batch size,

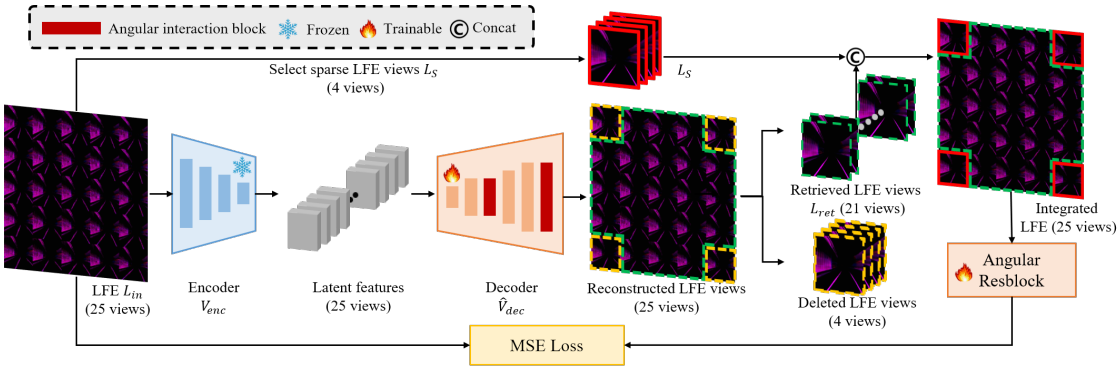

Figure 2. The second training stage of the LFE-VAE architecture. The input to the LFE-VAE supports multiple angular resolutions of LFE. Here, we take the case of an input with 25 views as an example. The LFE-VAE is composed of the encoder V_{enc}, the decoder $\hat{V}_{dec}$ inserted by angular interaction blocks, and the following angular resblocks. The training at this stage is aimed at updating the decoder $\hat{V}_{dec}$ and the angular resblocks.

TB represents the temporal bins, 2 denotes the polarities, and $H \times W$ denotes the spatial resolution. To facilitate the subsequent processing of the VAE, we transform the two-polarity channels of the event stream into three channels based on [49]. A LFE can be represented as $L \in \mathbb{R}^{B\times 3\times M\times N\times H\times W}$, where $M \times N$ denotes the angular resolution. The LFE synthesis task aims to generate a dense LFE $L_D \in \mathbb{R}^{B\times 3\times M\times N\times H\times W}$ from a sparse LFE $L_S \in \mathbb{R}^{B\times 3\times m\times n\times H\times W}$, where $m \times n$ represents a lower angular resolution. To effectively model the complex distribution of LFE data, we adopt a diffusion-based pipeline. Given the high-dimensional nature of LFE data, we first introduce LFE-VAE to encode LFE into a compact latent representation. The encoded sparse LFE, together with the encoded angular coordinates, serve as conditioning signals for the diffusion model to synthesize views at specified angular positions. To effectively incorporate these conditioning signals into the diffusion backbone, we design an LFE-adapter for optimal signal processing. Through iterative generation of all unknown views, we achieve the transformation from sparse to dense LFE. Notably, our diffusion network demonstrates the capability to synthesize novel views at arbitrary angles beyond the training set by accepting different angular coordinates as input.

4.2. LFE-VAE

To handle the high spatial resolution of LFE data efficiently, we employ Variational Autoencoder (VAE) architectures commonly used in diffusion models [25, 26, 35, 36, 53] for data compression. However, directly applying conventional VAEs to LFE data presents two significant challenges: 1) Traditional VAEs used in stable diffusion are designed for single-view processing, failing to leverage the inherent cross-view structural information in LFE data, and 2) The lossy nature of VAE compression inevitably results in information loss in the event streams.

To address these limitations, we propose LFE-VAE, building upon the pretrained VAE model from SDXL [35]. Our architecture introduces two key innovations: (1) angular interaction blocks in the decoder to incorporate cross-view complementary information, and (2) a view-enhancement mechanism utilizing four known views to minimize compression artifacts.

In specific, for each dense LFE input L_{in}, the encoder first generates a latent code Z_s. This latent representation is processed by our enhanced decoder, which embeds cross-view geometric information through angular interaction blocks. Each block consists of two cascaded layers of spatial and angular convolutions. The corresponding positions of the decoded target views are replaced with four known sparse LFE views, and then the LFE is refined through several cascaded resblocks defined on the LFE angular slices.

Our training process follows a two-stage approach. First, we fine-tune the stable diffusion VAE using single-view events to obtain updated weights for the encoder V_{enc} and decoder V_{dec}. Subsequently, while keeping V_{enc} frozen, we fine-tune the proposed decoder $\hat{V}_{dec}$ and angular resblocks. The second stage of the LFE-VAE architecture is illustrated in detail in Fig. 2. Specifically, we first select the sparse LFE views L_S, which represent the condition signal for the subsequent diffusion pipeline. The input LFE L_{in} is then encoded by V_{enc} and decoded by the updated $\hat{V}_{dec}$, resulting in an intermediate reconstructed LFE. Next, we replace the four corner views of the recovered LFE, with the four corresponding views in L_S selected in L_{in}. Finally, the compression degradation is restored through the utilization of angular resblocks. After training, the encoder V_{enc} is integrated into the diffusion training pipeline to compress input sparse LFE, enabling the subsequent diffusion network to handle higher-resolution LFE generation.

4.3. Diffusion Pipeline

Following mainstream diffusion architectures [13, 22, 25, 26, 35, 36, 53], our diffusion backbone adopts a UNet structure. While existing diffusion models primarily han-

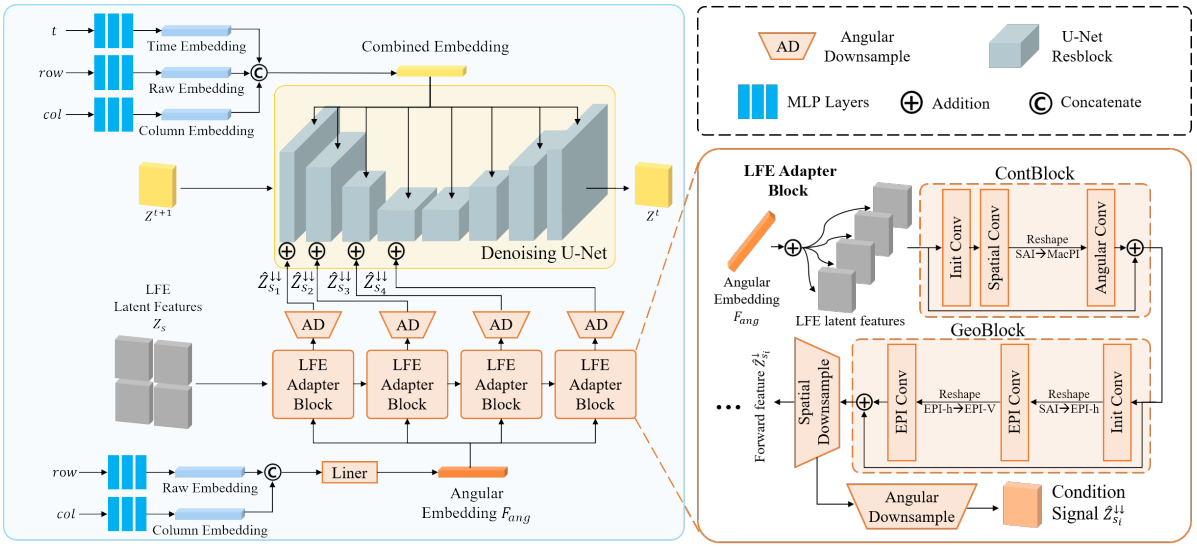

Figure 3. The diffusion pipeline of S2D-LFE at the t-th time step. The pipeline primarily consists of two parts: (1) The left part, highlighted with a blue background, represents the main pipeline diagram. The diffusion pipeline of S2D-LFE adopts a UNet backbone, where the latent features Z_s from four viewpoints are injected into the network as conditional signals via the LFE-adapter. In addition to the time embedding, our prompt incorporates positional encodings of both horizontal and vertical coordinates. (2) The right section, highlighted with an orange background, depicts the LFE-adapter block within the LFE-adapter. After cascading the condition information through the ContBlock and GeoBlock, the output features are embedded into the corresponding layers of the UNet. The LFE-adapter consists of four LFE-adapter blocks, each producing a different conditional signal $\hat{Z}_{s1}^{\downarrow\downarrow}$, $\hat{Z}_{s2}^{\downarrow\downarrow}$, $\hat{Z}_{s3}^{\downarrow\downarrow}$, $\hat{Z}_{s4}^{\downarrow\downarrow}$ corresponding to the respective layers of the UNet.

dle unstructured conditioning data such as images, videos, and text, the potential of structured conditioning data, which contains rich structural and texture priors, remains largely unexplored. To address this gap, we propose a specialized diffusion pipeline tailored for LFE data. At its core is the novel LFE-adapter, which serves two key functions: 1) extracting features that encode both structural and texture information from input sparse LFE and 2) injecting these extracted features into the diffusion backbone. With the structural and texture information effectively embedded, the diffusion pipeline generates specific LFE views based on input view prompts. Our approach employs continuous, linear angular coordinate encoding for view prompts, enabling the pipeline to generalize to sampling LFE views across varying densities and provide flexibility in LFE generation. The complete architecture of our diffusion pipeline at the t-th time step is illustrated in Fig. 3. Then, we will detail each of these innovations.

LFE Angular Prompt. Accurate positional priors of target views are essential for novel LF view generation, which we achieve through angular coordinate prompting. As illustrated in Fig. 3, we process 2D coordinates through a two-stage encoding scheme. First, we independently encode row and column coordinates. These encoded coordinates are then transformed through cascaded fully connected layers to obtain the raw and column embeddings, which are then concatenated and fed into a linear layer to get the angular embedding F_{ang}. It serves dual purposes in our pipeline. It is combined with time embedding as input to each UNet layer, while simultaneously providing angular positional priors to the LFE-adapter for extracting view-specific conditioning information for the target view generation.

LFE-adapter. Our pipeline leverages priors from a sparse LFE containing four known views. The design incorporates both texture and geometry priors from these input views to provide specific conditioning signals for novel view generation. Drawing inspiration from [33], we propose an LFE-adapter to effectively inject structural and textural information from the known views. The architecture of our LFE-adapter is depicted in the orange region of Fig. 3. The processing pipeline begins by using the V_{enc} of LFE-VAE to obtain compressed latent codes Z_{s} from the sparse LFE. We then integrate the angle embedding F_{ang} with the latent features of the four input views to obtain the updated latent feature Z_{s}^{*}. This feature is processed through a content extraction block (contblock) that extracts textures and details using spatial and angular convolutions. The spatial convolution embeds individual view information, while the angular convolution captures complementary cross-view information. A geometric extraction block (geoblock) further processes the features using cascaded Epipolar Plane Image (EPI) convolutions [5, 24, 43, 52] to embed geometric information, and thus we obtain the updated feature $\hat{Z}_{\text{s}}$. This feature undergoes spatial downsampling to produce $\hat{Z}_{\text{s}}^{\downarrow}$ for subsequent network updates, followed by angular downsampling to generate $\hat{Z}_{\text{s}}^{\downarrow\downarrow}$ as conditioning signals for corresponding UNet layers. The LFE-adapter ultimately generates a set of condition signals $\hat{Z}_{\text{s1}}^{\downarrow\downarrow}$, $\hat{Z}_{\text{s2}}^{\downarrow\downarrow}$, $\hat{Z}_{\text{s3}}^{\downarrow\downarrow}$, $\hat{Z}_{\text{s4}}^{\downarrow\downarrow}$. Each sig-

Table 1. Quantitative evaluation of event-based view synthesis on the in-training-scale setting. We tested various methods under synthetic and real settings and evaluated their performance across three metrics. The best results are highlighted in red, the second-best results are highlighted in blue, and the third-best results are highlighted in green. '↑': the higher the better performance, '↓': the opposite.

Category	Method	Synthetic			Real-world		
		PSNR ↑	SSIM ↑	LPIPS ↓	NIQE ↓	BRISQUE ↓	MUSIQ ↑
AngularSR	Jing *et al.* [23]	21.12	0.580	0.362	31.56	67.32	28.14
	DistgASR [44]	**23.32**	**0.676**	**0.278**	28.88	63.95	29.90
	SAV [12]	22.84	0.614	**0.282**	29.15	62.77	29.25
	Guo *et al.* [21]	**23.46**	**0.687**	0.290	**28.23**	**59.13**	**30.72**
Hybrid	ET-Net [47]+SAV [12]+Vid2E [18]	20.23	0.574	0.373	32.11	67.11	25.45
	ET-Net [47]+Guo *et al.*+Vid2E [18]	20.88	0.589	0.375	31.95	66.58	25.59
NeRF-based	R2L [42]	19.88	0.563	0.401	**28.44**	**60.25**	**30.45**
Generation	**Ours**	**24.06**	**0.701**	**0.239**	**27.65**	**53.45**	**31.14**

nal is added to the corresponding same-size input features of the UNet blocks.

LFE View Sampling. Through our LFE-adapter, the diffusion pipeline is enriched with comprehensive structural and textural information from the input LFE. The angular prompts effectively encode target view positions, ensuring angular continuity in the generated LFE views. The pipeline samples individual views sequentially, reconstructing the complete continuous LFE by iteratively processing coordinates for all target views. The framework accommodates LFEs of varying densities through relative coordinate mapping, where $(0, 0)$ represents the top-left view and $(1, 1)$ represents the bottom-right view. For example, generating a view at position (i, j) in an $N \times N$ LFE would use normalized coordinates $(i/(N-1), j/(N-1))$. This normalized coordinate system provides a unified approach for sampling views across different LFE densities while maintaining consistent spatial relationships.

5. Experiment

5.1. Experiment Settings

Implementation Details. Our networks are trained on a computing cluster equipped with eight NVIDIA A100 GPUs. Following [44], we employ data augmentation techniques including random horizontal flipping, vertical flipping, and 90° rotation to enhance model robustness. The optimization process utilizes the Adam optimizer with hyperparameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, processing batches of size 32. We adopt a learning rate scheduling strategy where the initial rate of 2×10^{-4} is halved every 400 epochs. Training concludes after 1500 epochs when convergence is achieved. To ensure fairness in the experiments, all metrics are calculated excluding the input views.

Datasets and Metrics. For experimental validation, we utilize two distinct datasets: a synthetic dataset generated using Carla for training and testing, and a real-world dataset captured by our system for additional testing. We use two sets of complementary metrics. For data with ground-truth references, we utilize PSNR measured in dB, SSIM [45], and LPIPS [54]. For data without ground-truth references, we employ NIQE [32], BRISQUE [31], and MUSIQ [27].

Comparison Methods. We establish comparisons with three categories of related approaches[1]. The first category comprises LF angular super-resolution methods, including Jing *et al.*[23], DistgASR[44], SAV [12], and Guo *et al.*[21]. The second category involves a three-stage approach: we first convert sparse LFE to sparse RGB LF using event-based video reconstruction methods like ET-Net[47], then perform angular super-resolution to get dense RGB LF, and finally convert the dense RGB LF back to dense LFE using Vid2E [18]. The third category explores neural radiance field-based reconstruction using R2L [42]. All the aforementioned baseline methods are re-trained on our dataset loading their original pre-trained parameters.

5.2. Comparison Results

In-training-scale Comparisons. We first conduct in-distribution evaluation on 5×5 LFE generation, which matches our training configuration. As shown in Table 1, our S2D-LFE outperforms all baselines across all metrics on both synthetic and real-world testsets. On the synthetic testset, S2D-LFE achieves a notable PSNR improvement of 0.60 dB over the second-best method Guo et al.[21], demonstrating superior reconstruction accuracy. The strong LPIPS performance further indicates enhanced perceptual quality, leveraging the advantages of generative modeling. This is visually evident in Fig.4, where S2D-LFE reconstructs building details with minimal artifacts and better ground truth alignment. In addition, although R2L [42] achieves decent visual results in Fig. 4, it fails to reconstruct in certain scenes, leading to lower metric performance on the synthetic dataset. Meanwhile, S2D-LFE maintains its superior performance across all metrics on the real-world testset, demonstrating effective generalization to real-world scenarios. This quantitative advantage is corroborated by

[1]Since EV-LFV [29] and our S2D-LFE utilize different input modalities, a direct comparison between the two methods is not feasible.

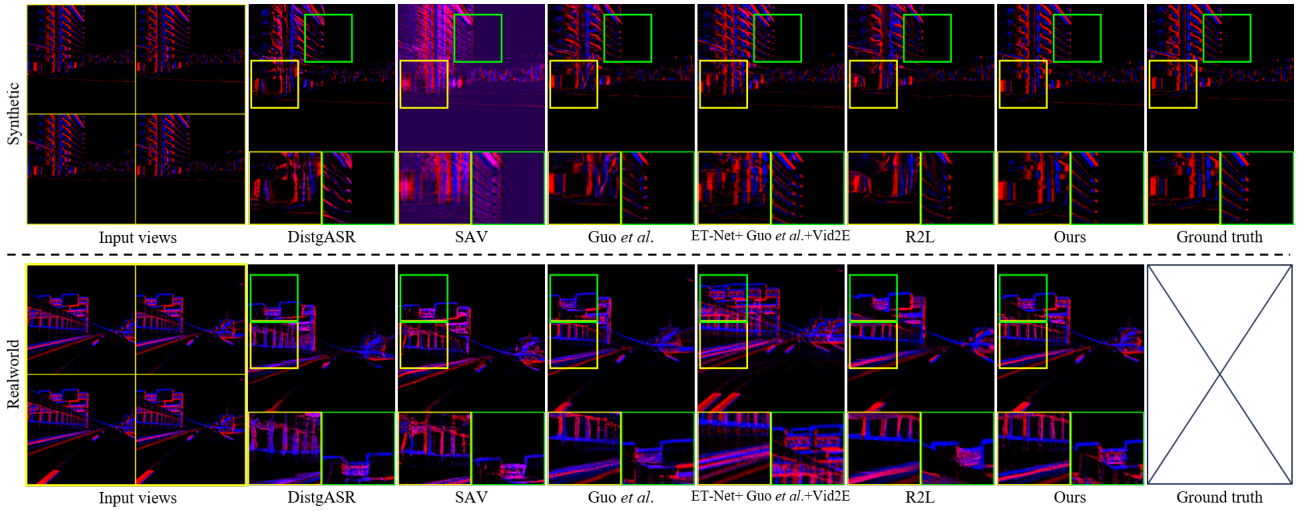

Figure 4. The figure presents a qualitative comparison on the central view of the generated LFE, using various methods for both synthetic (top row) and real-world (bottom row) testsets. The proposed S2D-LFE method (labeled as "Ours") is compared against other existing techniques, including SAV [12], Guo et al. [21], ET-Net [47] + Guo et al. + Vid2E [18], R2L [42], DistgASR [44], and the ground-truth. The highlighted regions (yellow and green boxes) magnify specific areas to emphasize the differences in event generation quality, with S2D-LFE exhibiting improved fidelity and preservation of fine details in both synthetic and real-world scenarios.

the qualitative results shown in Fig. 4, where S2D-LFE produces higher quality reconstructions with better preservation of details. Due to incorrect estimation of view position, SAV [12] and DistgASR [44] exhibit misaligned results compared to other methods. Furthermore, the ET-Net [47] + SAV + Vid2E [18] approach exhibits severe view aliasing in this scene.

Out-of-training-scale Comparisons. We extend our evaluation to test the model's generalization capability beyond its training distribution of 5×5 LFE, examining its performance on denser LFEs. Among the seven baselines we selected, only R2L [42] and Guo *et al.* [21] are capable of handling out-of-training-scale LFE generation. Therefore, we only compare our method with these two approaches in this section. As shown in Table 2, our S2D-LFE consistently outperforms baseline methods across all tested LFE densities. While all methods show performance degradation as LFE density increases, S2D-LFE demonstrates greater resilience with a slower degradation rate. The qualitative results in Fig.5 further validate S2D-LFE's superior performance in maintaining perceptual quality and detail preservation for out-of-distribution scenarios. Additional visual comparisons are provided in the supplementary material.

Angular Consistency. In this section, we analyze and compare the angular continuity of the LFE generated by our S2D-LFE. Given that the angular continuity of LFE data is strongly correlated with the gradient of the Epipolar Plane Image (EPI) [44, 48], we compute metrics (PSNR, SSIM) on the EPI gradient to evaluate the angular continuity of Guo et al. [21], R2L [42], and S2D-LFE when generating LFEs with different numbers of views (ranging from 5×5 to 9×9). It can be observed from Fig. 6 that as the number of views increases, all three methods exhibit a decreasing trend in both PSNR and SSIM metrics while the decline for our S2D-LFE is more gradual. It can be interpreted as S2D-LFE exhibits better angular consistency and robustness compared to the other methods. Moreover, we exhibit a qualitative comparison when generating 9×9 LFE in Fig. 7. It can be observed that S2D-LFE preserves fidelity details while producing EPI slices that more closely align with the ground truth.

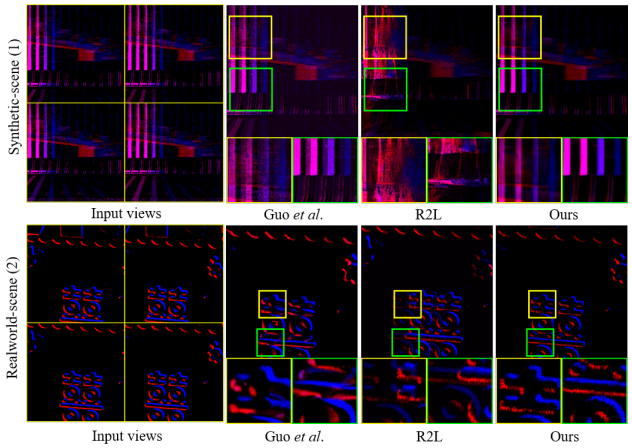

Figure 5. The figure presents the out-of-training-scale qualitative comparison using different methods including Guo et al. [21], R2L [42], and our proposed S2D-LFE (Ours). Besides, synthetic-scene (1) is the (1,1) view from the generated 8×8 LFE, and realworld-scene (2) is the (7,7) view from the generated 8×8 LFE,. The yellow and green boxes highlight specific regions to better illustrate the differences in event generation quality.

Table 2. Quantitative evaluation of event-based view synthesis on the out-of-training-scale setting. We tested various methods under three different settings (generating 36 views, 49 views, and 64 views) and evaluated their performance across three metrics. The best results are highlighted in bold. '↑': the higher the better performance, '↓': the opposite.

	Method	NIQE ↓			BRISQUE ↓			MUSIQ ↑		
		36-view	49-views	64-views	36-views	49-views	64-views	36-views	49-views	64-views
Syn	Guo *et al.* [21]	16.85	18.62	20.10	59.45	64.12	70.78	31.28	30.02	28.31
	R2L [42]	16.96	19.05	21.54	61.72	65.93	71.56	30.27	29.08	27.29
	Ours	**15.84**	**17.65**	**19.67**	**54.15**	**58.21**	**63.49**	**32.30**	**31.89**	**30.34**
Real	Guo *et al.* [21]	28.81	29.58	30.29	62.83	66.15	70.58	30.27	29.78	29.30
	R2L [42]	28.94	29.72	30.61	64.00	68.18	71.37	29.99	29.48	29.02
	Ours	**27.99**	**28.47**	**29.01**	**56.17**	**58.91**	**61.29**	**30.93**	**30.65**	**30.21**

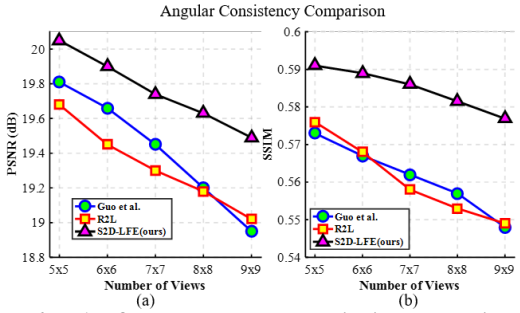

Figure 6. The figure presents a quantitative comparison of angular consistency across different methods, including Guo et al., R2L, and our proposed S2D-LFE. Subfigures (a) and (b) depict PSNR (dB) and SSIM metrics respectively, when generating different numbers of views ranging from 5×5 to 9×9. All methods exhibit a decline in both metrics as the number of views increases, but S2D-LFE exhibits a relatively slower decline.

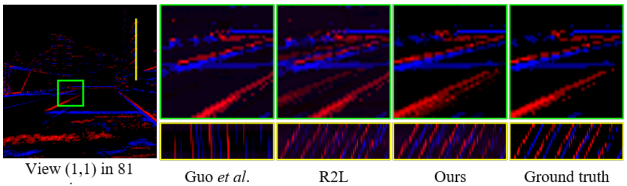

Figure 7. The figure presents a qualitative comparison when generating 9×9 LFE using Guo et al., R2L, and our proposed S2D-LFE. The green boxes highlight spatial slices of view (1,1), while the yellow boxes highlight EPI slices.

6. Ablation Study

LFE-VAE. We examine the effectiveness of our LFE-VAE by making comparisons with two variants: one using a standard VAE (S-VAE) and another using a fine-tuned VAE (FT-VAE) with single-view event data. As shown in Table 3, S-VAE variant shows performance degradation across all metrics, indicating reduced reconstruction quality and perceptual similarity. While FT-VAE variant shows modest improvements with PSNR of 37.12 dB and SSIM of 0.865, its LPIPS score of 0.151 still indicates suboptimal perceptual quality. In contrast, our LFE-VAE better captures the structural and perceptual characteristics of light field events.

LFE-adapter. To validate the effectiveness of our LFE-adapter design, we make comparisons with a variant that replaces the LFE-adapter with simple feature concatenation (concat). As shown in Table 4, the concatenation variant shows substantial performance degradation across all metrics, indicating significant reductions in reconstruction accuracy, structural fidelity, and perceptual quality. The comparison clearly validates that our LFE-adapter provides more effective feature integration compared to naive concatenation.

Table 3. Quantitative evaluation of the ablation on the LFE-VAE. The table compares the performance under a setting of generating 25 views in the synthetic dataset. The best results are highlighted in bold. '↑': the higher the better performance, '↓': the opposite.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
S-VAE	36.89	0.859	0.158
FT-VAE	37.12	0.865	0.151
LFE-VAE (Ours)	**38.54**	**0.881**	**0.112**

Table 4. Quantitative evaluation of the ablation on the LFE-adapter. The table compares the performance under a setting of generating 25 views in the synthetic dataset. The best results are highlighted in bold. '↑': the higher the better performance, '↓': the opposite.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Concat	22.54	0.631	0.275
LFE-adapter	**24.06**	**0.701**	**0.239**

7. Conclusion

In this paper, we present S2D-LFE, a novel approach for generating consistent, full-view LFE. Our key contributions include developing the first real-world LFE capture system and proposing effective technical components: an LFE-VAE for LFE compression and an LFE-Adapter for prior injection. S2D-LFE achieves simplicity by eliminating multi-view RGB dependencies, controllability through precise view synthesis, and consistency via cross-view and temporal coherence. These advances establish S2D-LFE as an effective solution for dense LFE generation.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grants 62472399, 62131003 and 62021001.

References

[1] A N Angelopoulos, J N Martel, A P Kohli, J Conradt, and G Wetzstein. Event based, near eye gaze tracking beyond 10,000 hz. *IEEE Transactions on Visualization and Computer Graphics*, page 1, 2020. 1, 2

[2] Benjamin Attal, Jia-Bin Huang, Michael Zollhöfer, Johannes Kopf, and Changil Kim. Learning neural light fields with ray-space embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19819–19829, 2022. 2

[3] Mojtaba Bemana, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel. X-fields: Implicit neural view-, light- and time-image interpolation. *ACM TOG*, 39(6):1–15, 2020. 2

[4] A Bhattacharya, R Madaan, F Cladera, et al. Evdnerf: Reconstructing event data with dynamic neural radiance fields. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5846–5855, 2024. 2

[5] Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. pages 497–504. 2001. 2, 5

[6] Eric R Chan, Koki Nagano, Matthew A Chan, Alexander W Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. Generative novel view synthesis with 3d-aware diffusion models. *arXiv preprint arXiv:2304.02602*, 2023. 2

[7] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14124–14133, 2021. 2

[8] H Chen, J Gu, A Chen, et al. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2416–2425, 2023. 2

[9] P Chen, W Guan, F Huang, et al. Ecmd: An event-centric multisensory driving dataset for slam. *IEEE Transactions on Intelligent Vehicles*, 2023. 2

[10] P Chen, W Guan, and P Lu. Esvio: Event-based stereo visual inertial odometry. *IEEE Robotics and Automation Letters*, 8 (6):3661–3668, 2023. 2

[11] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pages 370–386. Springer, 2024. 2

[12] Zhen Cheng, Yutong Liu, and Zhiwei Xiong. Spatial-angular versatile convolution for light field reconstruction. *IEEE Transactions on Computational Imaging*, 8:1131–1144, 2022. 2, 6, 7

[13] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 4

[14] A Dosovitskiy, G Ros, F Codevilla, et al. Carla: An open urban driving simulator. In *Conference on Robot Learning*, pages 1–16. PMLR, 2017. 2, 3

[15] Brandon Yushan Feng and Amitabh Varshney. Signet: Efficient neural representation for light fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14224–14233, 2021. 2

[16] P Furgale, J Rehder, and R Siegwart. Unified temporal and spatial calibration for multi-sensor systems. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1280–1286. IEEE, 2013. 3

[17] Daniel Gehrig, Antonio Loquercio, Konstantinos G Derpanis, and Davide Scaramuzza. End-to-end learning of representations for asynchronous event-based data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5633–5643, 2019. 3

[18] Daniel Gehrig, Mathias Gehrig, Javier Hidalgo-Carrió, and Davide Scaramuzza. Video to events: Recycling video datasets for event cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3586–3595, 2020. 6, 7

[19] M Gehrig, W Aarents, D Gehrig, and D Scaramuzza. Dsec: A stereo event camera dataset for driving scenarios. *IEEE Robotics and Automation Letters*, 6(3):4947–4954, 2021. 2

[20] Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *International Conference on Machine Learning*, pages 11808–11826. PMLR, 2023. 2

[21] Mantang Guo, Junhui Hou, Jing Jin, Hui Liu, Huanqiang Zeng, and Jiwen Lu. Content-aware warping for view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8):9486–9503, 2023. 2, 6, 7, 8

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 4

[23] Jing Jin, Junhui Hou, Hui Yuan, and Sam Kwong. Learning light field angular super-resolution via a geometry-aware network. In *Proceedings of the AAAI conference on artificial intelligence*, pages 11141–11148, 2020. 1, 2, 6

[24] Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi. Learning-based view synthesis for light field cameras. *ACM TOG*, 35(6):1–10, 2016. 2, 5

[25] J Karras, A Holynski, T C Wang, et al. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22623–22633, 2023. 2, 4

[26] B Ke, A Obukhov, S Huang, et al. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9492–9502, 2024. 2, 4

[27] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 6

[28] Zhong Li, Liangchen Song, Celong Liu, Junsong Yuan, and Yi Xu. Neulf: Efficient novel view synthesis with neural 4d light field. *arXiv preprint arXiv:2105.07112*, 2021. 2

[29] Zhicheng Lu, Xiaoming Chen, Vera Yuk Ying Chung, Weidong Cai, and Yiran Shen. Ev-lfv: Synthesizing light field event streams from an event camera and multiple rgb cameras. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 1, 2, 6

[30] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021. 2

[31] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21 (12):4695–4708, 2012. 6

[32] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 6

[33] C Mou, X Wang, L Xie, et al. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 5

[34] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5480–5490, 2022. 2

[35] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 4

[36] R Rombach, A Blattmann, D Lorenz, et al. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 2, 4

[37] S Saxena, C Herrmann, J Hur, et al. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[38] Vincent Sitzmann, Semon Rezchikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks: Neural scene representations with single-evaluation rendering. *NeurIPS*, 34:19313–19325, 2021. 2

[39] Hung-Yu Tseng, Qinbo Li, Changil Kim, Suhib Alsisan, Jia-Bin Huang, and Johannes Kopf. Consistent view synthesis with pose-guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16773–16783, 2023. 2

[40] S M N Uddin, S H Ahmed, and Y J Jung. Unsupervised deep event stereo for depth estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7489–7504, 2022. 1, 2

[41] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9065–9076, 2023. 2

[42] Huan Wang, Jian Ren, Zeng Huang, Kyle Olszewski, Menglei Chai, Yun Fu, and Sergey Tulyakov. R2l: Distilling neural radiance field to neural light field for efficient novel view synthesis. In *European Conference on Computer Vision*, pages 612–629. Springer, 2022. 2, 6, 7, 8

[43] Yunlong Wang, Fei Liu, Zilei Wang, Guangqi Hou, Zhenan Sun, and Tieniu Tan. End-to-end view synthesis for light field imaging with pseudo 4dcnn. In *ECCV*, pages 333–348, 2018. 2, 5

[44] Yingqian Wang, Longguang Wang, Gaochang Wu, Jungang Yang, Wei An, Jingyi Yu, and Yulan Guo. Disentangling light fields for super-resolution and disparity estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):425–443, 2022. 2, 6, 7

[45] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6

[46] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 2

[47] Wenming Weng, Yueyi Zhang, and Zhiwei Xiong. Event-based video reconstruction using transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2563–2572, 2021. 6, 7

[48] Gaochang Wu, Belen Masia, Adrian Jarabo, Yuchen Zhang, Liangyong Wang, Qionghai Dai, Tianyou Chai, and Yebin Liu. Light field image processing: An overview. *IEEE Journal of Selected Topics in Signal Processing*, 11(7):926–954, 2017. 1, 7

[49] Ziyi Wu, Xudong Liu, and Igor Gilitschenski. Eventclip: Adapting clip for event-based object recognition. *arXiv preprint arXiv:2306.06354*, 2023. 4

[50] J Wynn and D Turmukhambetov. Diffusionerf: Regularizing neural radiance fields with denoising diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4180–4189, 2023. 2

[51] J Xiang, J Yang, B Huang, et al. 3d-aware image generation using 2d diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2383–2393, 2023. 2

[52] Henry Wing Fung Yeung, Junhui Hou, Jie Chen, Yuk Ying Chung, and Xiaoming Chen. Fast light field reconstruction with deep coarse-to-fine modeling of spatial-angular clues. In *ECCV*, pages 137–152, 2018. 2, 5

[53] L Zhang, A Rao, and M Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 4

[54] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[55] Y Zhou, G Gallego, and S Shen. Event-based stereo visual odometry. *IEEE Transactions on Robotics*, 37(5):1433–1450, 2021. 2