

Physical Priors Augmented Event-Based 3D Reconstruction

Jiaxu Wang¹ [◇] Junhao He¹ [◇], Ziyi Zhang¹ and Renjing Xu¹ [†]

Abstract—3D neural implicit representations play a significant component in many robotic applications. However, reconstructing neural radiance fields (NeRF) from realistic event data remains a challenge due to the sparsities and the lack of information when only event streams are available. In this paper, we utilize motion, geometry, and density priors behind event data to impose strong physical constraints to augment NeRF training. The proposed novel pipeline can directly benefit from those priors to reconstruct 3D scenes without additional inputs. Moreover, we present a novel density-guided patch-based sampling strategy for robust and efficient learning, which not only accelerates training procedures but also conduces to expressions of local geometries. More importantly, we establish the first large dataset for event-based 3D reconstruction, which contains 101 objects with various materials and geometries, along with the groundtruth of images and depth maps for all camera viewpoints, which significantly facilitates other research in the related fields. The code and dataset will be publicly available at <https://github.com/Mercerai/PAEv3d>.

I. INTRODUCTION

3D representations serve as the foundation for many robotic applications such as navigation, manipulation, and 3D understanding. However, images captured by standard cameras are hardly used to reconstruct entire 3D scenes on account of the lack of information under suboptimal illumination environments, especially under extreme lighting conditions including over- and under-exposures. On the other side, event cameras, as neuromorphic sensors, have been demonstrated to perform well in such environments due to their high dynamic ranges which is because each pixel in event cameras individually detects the changes of brightness and only outputs a sequence of asynchronous events composed of the polarity rather than the absolute intensities.

Unfortunately, event-based 3D representation and reconstruction tasks remain challenging because event cameras only record relative brightness changes. Several approaches combine other devices like depth sensors or standard cameras with event cameras to reconstruct 3D scenes [56], [13], [43]. However, these methods sacrifice the advantages of event sensors, such as high temporal resolution. Other approaches tackle the problems by stereo visual odometry (VO) [53], [9], [13], [31], [56] or SLAM [11], [30], [52]. These methods only can reconstruct sparse 3D models such as point clouds. The sparsity limits their usage in many scenarios. Besides, another branch represents objects as rough templates initially, then updates their deformations to align with events [26], [34]. Nevertheless, they rely on the initialization of templates and are only constrained by specific object categories.

NeRF [25] gains great success in computer vision communities because it can densely represent entire 3D scenes and merely learn from images. Very recently, [10] and [35] propose a novel paradigm that reconstructs NeRF from event streams. Even if the paradigm

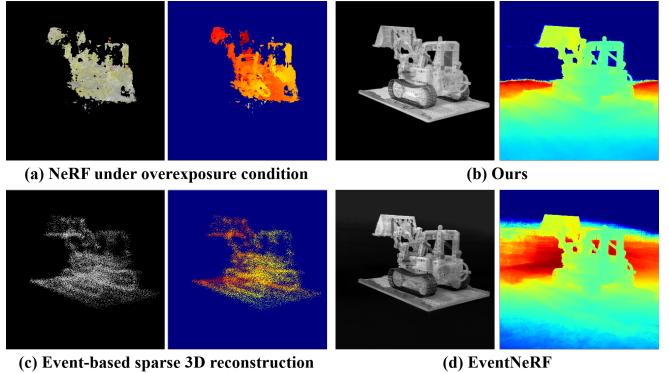


Fig. 1. Reconstruction results of original NeRF, semi-dense point cloud from event-based approach, EventNeRF, and Ours under extreme overexposure condition. The left figure is the rendering image and the right side is the depth map.

can perform well in certain situations, it is hard to faithfully reconstruct objects in real event data with complex geometries, textures, or realistic noises. Figure 1 indicates examples of reconstruction with different approaches in extreme illumination scenarios. The results are produced by overexposing the original NeRF training set and generating events with the event simulator. It can be seen the event-based approaches are less affected by the intense illumination whereas image-based NeRF is significantly destroyed. Additionally, the conventional event-based reconstruction (c) introduces severe sparsities while NeRF-based can continuously represent the 3D object. Furthermore, there is still a gap in high-quality event-based 3D reconstruction datasets. Current event-based real datasets either contain a simple handful of objects or lack corresponding image labels, which hinders the development of related tasks.

In this paper, we tackle the problem of NeRF reconstruction from raw event data in more realistic scenes. We analyze that event data actually contain rich priors including density, motion, and geometry because events are triggered by relative movement and edges. Our aim is to embed these priors into the NeRF pipeline to physically enhance its training and improve the reconstruction quality in the aspect of textures and geometries. Moreover, we fill the gap in the lack of high-quality event-based 3D reconstruction datasets. We experimentally prove that our method outperforms the recent benchmark by a considerable margin for both synthesis and realistic datasets. Our primary technical contributions are summarized in the following:

- We analyze underlying motion, density, and geometry priors behind events, which we incorporate into the NeRF pipeline through the warp field, deterministic event generation model, and disparity-flow relation.
- We propose the probabilistic patch sampling strategy based on the spatial event density to address the local minimum optimization caused by the event sparsity, which also benefits local feature representations.
- We first propose a large and real event-based 3D reconstruction dataset with accurate groundtruth of image frames, foreground

[†]Corresponding authors; [◇]Co-first authors

¹Jiaxu Wang is PhD student with MICS Thrust, HKUST(GZ), Email: jwang457@connect.hkust-gz.edu.cn

¹Junhao He and ¹Ziyi Zhang are Research Assistants with MICS Thrust, HKUST(GZ), Email: junhaohebright@outlook.com, ziyizhang@hkust-gz.edu.cn

¹Renjing Xu is the professor with MICS Thrust, HKUST(GZ), Guangzhou, China, Email: renjingxu@ust.hk

masks, and depth maps. The dataset contains more than 100 different objects with a wide range of materials and geometries, and would be publicly available for the community.

II. RELATED WORK

A. Event-based 3D Reconstruction

Approaches for reconstructing 3D scenes via event cameras have grown progressively, which can be roughly divided into different categories. The first branch usually uses a mix of input data modalities for additional information. Devo [56] reconstructs 3D models with the assistance of depth sensors. [15], [5], and [46] combine intensity frames, point clouds, or IMUs with event cameras respectively. Obviously, these methods require additional sensors and cannot fully utilize all the advantages of event cameras. Second, the task can also be addressed with SLAM or VO techniques. [12], [33], [3] use a pair of synchronized event cameras to obtain the sparse 3D points. [53] solves this problem by using spatial-temporal consistency principles. Moreover, [11], [30], [8] can produce semi-dense 3D reconstructions by utilizing the knowledge prior to camera motion to integrate events over a large time interval. However, the reconstruction results of these methods are not dense enough and only contain boundaries and edges resulting in events. The third branch of works tends to initialize objects with 3D scans or templates, then track the deformation of the initialization to align with input events [48], [28]. However, they are constrained to specific categories such as human bodies and hands.

Inspired by the recently popular 3D dense representation NeRF, [10] and [35] modify the traditional NeRF pipeline and make it trainable with only raw event data. These works preliminarily bridge event cameras and the neural implicit representation, achieving dense scene reconstruction. However, the event-based NeRF paradigm only supports recovering objects with simple geometries and regular textures because event data only contain relative illumination changes and much less information compared to the standard image, which causes ambiguity in the solution subspace. To reduce ambiguity, we propose the physically augmented event-based NeRF via underlying priors behind raw event data. Experiments illustrate that our proposed prior-augmented NeRF paradigm outperforms the benchmark considerably, especially in realistic scenarios.

B. 3D Scene Representations

Previous works have explored many different representations for modeling 3D scenes in various vision, graphics, and robotic applications. Traditional methods based on explicit representations such as point cloud [29], [1], [18], mesh [45], [40], [20], and voxel [22], [36] have inherent limitations of fixed topological structures and poor quality of novel view synthesis. To address these, 3D implicit scene representations [39], [51], [23], [25], [21], [27], [19] have been presented. For instance, DIST [19] and DVR [27] propose differentiable rendering formulations for implicit representations. But they require explicit extraction of surface information.

Recently, the Neural Radiance Field [25] has gained many successes. NeRF encodes a continuous volume representation of shape and color in the weights of an MLP, it supports efficient learning at arbitrary resolutions and enables rendering novel views with high-fidelity detail. The superiorities of NeRF inspire subsequent works in a wide variety of robotic applications, such as robotic policies [50], [14], [6], [2], [16], [37], SLAM [44], [55], [4], large scene reconstruction [49], [47], [38], robotic localization [54], [24], [17], and robotic safety [41]. SPARTN [50] augments robot trajectories via NeRF to improve the robustness of the grasping strategy. For street view or aerial images, Block-NeRF [38], Switch-NeRF [49],

and BungeeNeRF [47] enable city level reconstruction. NICER-SLAM [55] uses monocular geometric cues and optical flow as supervisions to optimize hierarchical neural implicit representation for building the SLAM system.

III. METHOD

First, we provide an overview diagram of our full method in Figure 2 which contains all main components.

A. Preliminary Background

This section introduces and analyzes the preliminary knowledge about volumetric rendering and event-based radiance fields. Conventional radiance fields store scene information in the parameters of MLPs and one can explore the scene by repeatedly accessing the neural network that is defined by $c, \sigma = f(x, y, z, d)$. The color c is not only dependent on locations but also the direction from which we observe. The opacity can be interpreted as the occupancy rate of this location. After these definitions, volumetric rendering can be applied to composite the results of sampled points on a ray into a pixel, as in Equation 1.

$$C(r) = \sum_{t=t_0}^{t_f} W(\sigma(t), t)c(t, d) \quad (1)$$

where r refers to a sampled ray, c is the color for each point on the ray. W is the blending weight calculated from the opacity by Equation 2.

$$W(\sigma, t) = \sum_{t=t_0}^T [(1 - \exp(\sigma(t)\delta))T(t)] \quad (2)$$

in which $T(t) = \exp(-\sum_{s=t_0}^t \sigma(s)\delta)$, δ represents the distance between adjacent points on the same ray. Besides, we can approximate the depth value at a certain pixel by the following equation:

$$D(r) = \sum_{t=t_0}^T W(\sigma(t), t)z(t) \quad (3)$$

However, such NeRF model is hardly trained with pure event streams directly [32].

A simple self-supervised paradigm for NeRF reconstruction from a single event stream is first proposed by [10] and [35]. It considers accumulated event frames as the intensity changes between two images. The event-based paradigm computes losses between a pair of images as in the following:

$$\Delta \hat{L}(u) = \log(\hat{I}_t + b) - \log(\hat{I}_{t-1} + b) \quad (4)$$

in which I_t denotes the rendering image at time t , b is an infinitesimal number to prevent $\log(0)$. The difference is defined in the log domain due to the principle of event cameras. The event frame is accumulated as follows:

$$\Delta L(u) = \sum_{t \in \Delta t} p_k C \delta(u - u_k), u \in (X, Y) \quad (5)$$

In this equation, p_k denotes event polarities, C represents the event number, and δ is the impulse function. u stands for pixel coordinates. It sums all events within the time interval on the frame. The training loss function is defined as:

$$L_{event} = \|\Delta \hat{L} - \Delta L\|_2^2 \quad (6)$$

However, this paradigm cannot recover correct 3D scenes with complex geometries, irregular textures, or realistic noises. We incorporate strong priors into the training pipeline to improve the reconstruction quality of realistic data. This will be introduced in the next section.

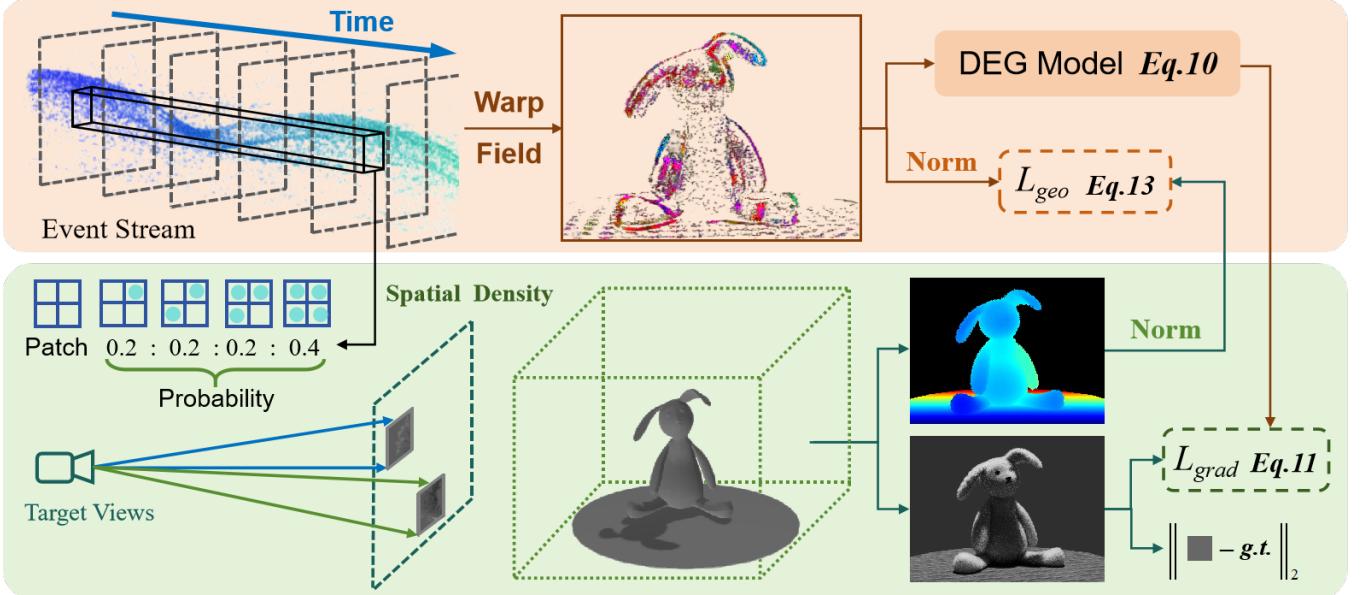


Fig. 2. The whole pipeline of the proposed approach. There are two main branches, i.e. the prior extraction and NeRF rendering branches. The priors are incorporated into the NeRF pipeline at sampling and loss parts.

B. Extraction of motion priors

Events are triggered by apparent motions parallel to the intensity gradient because event cameras respond to the apparent motion of edges. Since the event stream only reports pixel-wise brightness changes, it carries rich information about motion priors that contain relative motions, underlying geometric structures, and appearance gradients for events to be generated. These implicit priors are beneficial for 3D reconstruction from pure event streams and improve the details of geometries.

The problems of extraction of these priors can be converted to establish data association among events. Data association is to determine which events were triggered by the same edge. All events can be warped back along their trajectories into a reference view with a timestamp t_{ref} to obtain a sharper edge image. Therefore, the data association implicitly contains priors like optical flows, motion patterns, and depths. Next, we introduce how to extract those priors and how to use them to enhance the quality of reconstruction.

Since we only focus on the reconstruction of static scenes, motion information where the effect of camera motion can be described by a homography θ . Assume we are given a set of events $\varepsilon = \{e_i^N\}$, a general function for warping events is defined as $x'_k = W(x_k, t_k; \theta)$, which warps event $e_k = (x_k, t_k, p_k)$ to $e_k = (x'_k, t_{ref}, p_k)$ according to the motion parameter θ . Then following Equation 5, an image patch of warped events is built. The resulting sum is denoted by $H(x)$, which measures how well the events agree with the candidate trajectories. After that, we compute the variance of H by the following equation:

$$V_H = \frac{1}{N_e} \sum_{i,j} (h_{ij} - u_H)^2 \quad (7)$$

where N_e is the total number of pixels of H , u_H is the mean of H . It is clear that correcter motion parameters result in sharper event accumulated maps, i.e. the maximum variance. Thus we optimize $\theta = argmax_\theta V_H$ using gradient ascending to solve for the warp field. One by-product is that a compensated edge map can also be obtained from this optimization.

C. Learning from motion priors

We incorporate the prior warp field, which essentially encodes the per-event optical flow, event density, and a deterministic generative

event model into the NeRF training. First, the process of event generation within a given time interval Δt can be described as:

$$|\Delta L(u)| = |L(u, t) - L(u, t - \Delta t)| \geq C \quad (8)$$

in which C is a preset threshold. If we substitute the brightness consistency assumption (Equation 9) into its Taylor's approximation, we can get 10

$$\frac{\delta L}{\delta t}(\mathbf{u}, t) + \nabla L(\mathbf{u}, t) \cdot \mathbf{v}(\mathbf{u}) = 0 \quad (9)$$

$$\Delta L(\mathbf{u}) \approx \frac{\delta L}{\delta t}(\mathbf{u}, t) \Delta t = -\nabla L(\mathbf{u}) \cdot \mathbf{v}(\mathbf{u}) \Delta t \quad (10)$$

The left side of the equation can be regarded as event accumulation frames (defined by Equation 5) in a Δt . On the right side, $\mathbf{v}(\mathbf{u})$ refers to the velocity on the trajectories, which can be obtained by accessing the warp field. Additionally, ∇L is the intensity gradient in the log domain. This term can be considered as the self-supervised target and obtained from the NeRF model. To leverage this supervision, we modified two details in the NeRF pipeline. First, conventional NeRF randomly selects some rays in a batch for training. Nevertheless, isolated sampling pixels are not allowed to compute gradients. Instead, we use patch-based random sampling to randomly select several patches and compute their gradients. Then we substitute gradients to Equation 10 to build the l1 gradient loss:

$$L_{grad} = \frac{1}{N} \sum_p \|\nabla L_p / \Delta t - \Delta L_p(u) \cdot v_p\|_1 \quad (11)$$

in which p refers to each independent patch.

For each independent patch, it is a square of size n . We use the event spatial density to guide the sampling processes. Patches are sampled with weights of the number of pixels containing events, which we call event pixels, (the number of pixels ranges from 0 to n^2). However, we observed from experiments that if we evenly sample patches with different event numbers, this could lead to the exclusion of certain event patches from the sampling process. Hence we propose to sample patches with different event numbers in an uneven manner. In detail, we define a probability for each patch being sampled according to its number of pixels containing events. The probability of patches with the most event pixels

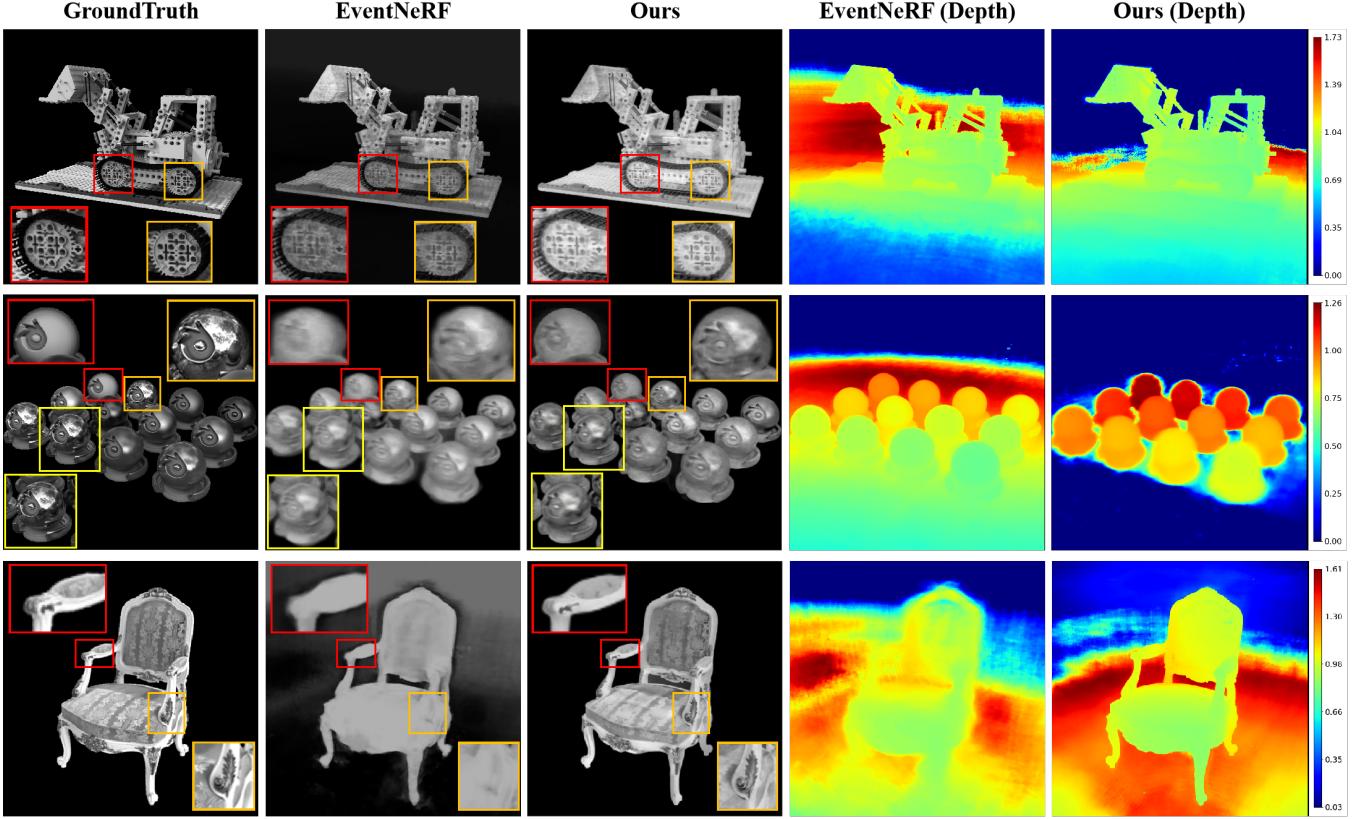


Fig. 3. Qualitative Comparisons between ours and benchmark on the synthetic dataset.

reaches the maximum, whereas patches without events have the least confidence. In Equation 12, n_e denotes the number of event pixels, f is a monotone increasing function and here we simply choose linear function as f . It is noted that other functions may reach the same effect.

$$P = \frac{1}{\sum_0^{n^2} f(n_e)} f(n_e) \quad (12)$$

This density prior not only ensures the sampling of a sufficient number of event pixels but also captures local features present within sparse event patches, thereby significantly enhancing the quality of the final texture reconstruction.

Additionally, the warp field also implies geometric priors of the scene because the scene is static and all events are caused by moving cameras. A common sense is that objects closer to cameras lead to larger displacement on the camera plane. The optical flows are roughly proportional to the disparities when only the camera is translationally moving. If we divide the camera trajectories into infinitesimal segments, each can be approximated as a translation. We use such features to regularize the geometries learned in NeRF. For each batch, we have the equation below

$$L_{geo} = \frac{1}{\hat{N}} \sum_{u \in \epsilon}^{\hat{N}} \left\| \frac{F_u}{F_{max}} - \frac{1}{D_u D_{min}} \right\|_1 \quad (13)$$

In this equation, $u \in \epsilon$ represents the (x, y) of sampled rays origins that contain at least one event. F_u denotes the warping flow that can be attained by accessing the warp field ($F_u = W(u, t; \theta)$). F_{max} is the maximum value over all F_u in this batch. $D_u = D'_u + \epsilon$, where D'_u is the output of NeRF models and can be computed

via Equation 3, ϵ is a very small number to prevent the divisor from reaching zero. We normalize each F_u and D_u over the batch to reduce the influence of noises and the infinitesimal assumption. This loss directly regularizes the learning of geometries and efficiently improves the reconstruction of details.

D. Implementation details

We use the original NeRF framework containing the coarse and fine models in accordance with the benchmark. Then we compose the above loss functions to guide the model training. The total losses are described as:

$$L_{total} = \alpha L_{event} + \beta L_{geo} + \gamma L_{grad} \quad (14)$$

where α , β and γ denote the weights to balance losses. In all experiments, we uniformly choose $\alpha = 1$, $\beta = 0.01$, $\gamma = 0.01$. For the density-based patch sampling, we fix n in Equation 12 as 2. Adam optimizer with 1e-5 learning rate and $\beta_1 = 0.9$, $\beta_2 = 0.999$ is used to train the model. We train our model on two NVIDIA 3090 GPUs with 150,000 steps, whereas train the benchmark with 500,000 on the same devices because our model converges more quickly due to the assistance of priors.

The depth maps rendered by NeRF usually display anomalies, and substantial inaccuracies may occur. Since the event depth is influenced by event noise, if we simply use the noisy event depth prior and the depth map rendered by NeRF for training, it can introduce significant errors, affecting the results of texture and geometric reconstruction. Therefore, we use Ambient Occlusion (AO) [42] to measure the confidence of the depth map rendered by NeRF.

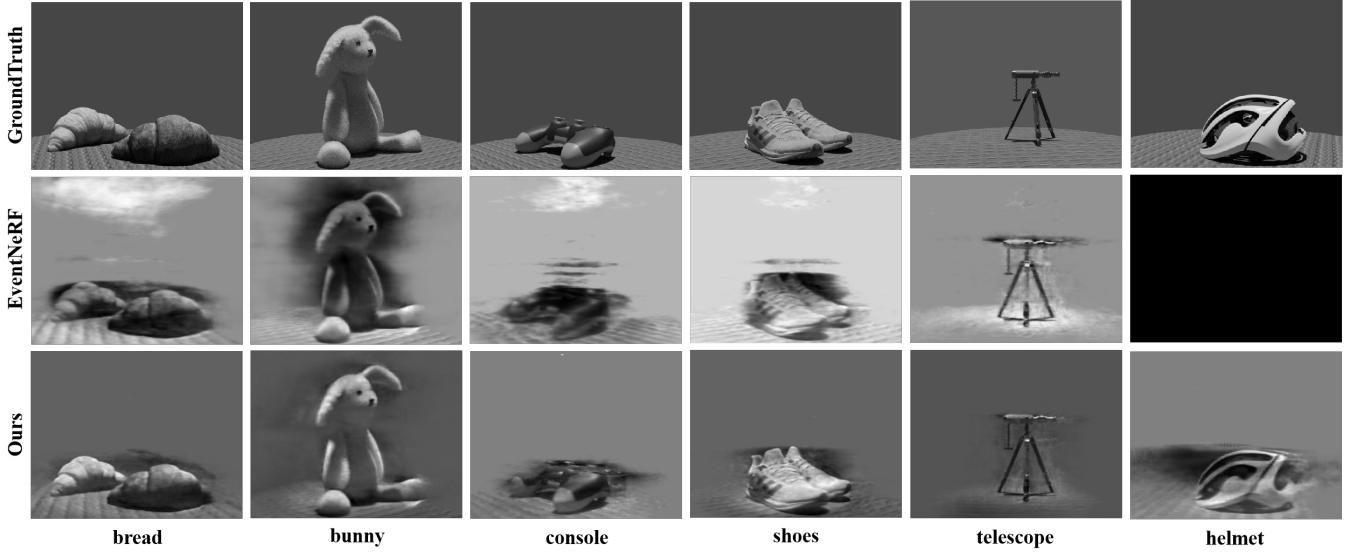


Fig. 4. Qualitative Comparisons between ours and benchmark on the realistic dataset.

$$AO = \sum_{i=1}^N T_i \alpha_i, \quad \alpha_i = 1 - \exp(-\sigma_i \delta_i) \quad (15)$$

Different from [42], we start with an AO initial value of 0 and gradually increase it to a predetermined maximum value as the number of training steps increases. This is because the NeRF trained with sparse event frames has a relatively poor fit when the number of training steps is not high. Therefore, we employ a warm-up filtering mechanism to gradually filter out geometric loss.

IV. EXPERIMENTS

Dataset. We test our methods and the benchmark (EventNeRF) on both synthetic and realistic datasets. Notably, EvNeRF and EventNeRF both worked at the same time and they follow a very similar paradigm, thus we only test the EventNeRF by its official source code. For the synthetic dataset, we simply transfer the original NeRF dataset to event streams by the video2event algorithm [7]. However, there is a large gap between synthetic and real event data in the aspect of event distributions, noises, and resolutions. Therefore, we mainly concentrate on realistic data.

Currently, there is no high-quality 3D reconstruction dataset captured by real event cameras, which obstacles the development and evaluations of algorithms, therefore we present a large dataset for event-based 3D reconstruction, which includes 101 different objects and scenarios. This dataset is recorded by a realistic DVXplore event camera and contains various materials and textures, such as plush, leather, alloy, wood, flour, metal, plastics, etc. Moreover, we provide the groundtruth of intensity frames, depth maps, and foreground masks associated with all camera poses. Although EventNeRF offers a few real event data, they only contain 10 objects and do not include the groundtruth of frames and depth maps, which cannot be used for quantitative analysis. As the space is limited, we selectively established six different scenarios including "bread", "bunny", "console", "helmet", "shoes", and "telescope". The whole dataset will be publicly available for the convenience of other researchers.

Metrics. As event data reflect the relative differences in intensity rather than absolute brightness, we normalize the rendering results to align with the groundtruth before computing metrics. Then we compute PSNR, SSIM, and LPIPS for comparisons.

TABLE I
METRICS OF OURS AND THE BENCHMARK ON SYNTHETIC DATA.

	EventNeRF			Ours		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
lego	21.91	0.921	0.073	24.51	0.961	0.067
materials	18.64	0.915	0.135	23.47	0.945	0.092
chair	20.74	0.917	0.121	23.61	0.972	0.112

TABLE II
METRICS OF OURS AND THE BENCHMARK ON REALISTIC DATA.

		bread bunny console shoes telescope helmet					
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
EventNeRF	lego	18.59	0.840	0.968	0.783	0.870	0.944
	materials	22.60	0.918	0.085	0.304	0.224	0.280
	chair	17.35	0.917	0.121	23.27	22.51	24.74
Ours	lego	27.15	0.962	0.923	0.950	0.911	0.935
	materials	20.03	0.950	0.059	0.116	0.059	0.142
	chair	19.63	0.944	0.067	0.017	0.063	N/A

A. Evaluation on Synthetic Event Data

We first evaluate the proposed method on the NeRF synthetic data. We only report the comparison results on "lego", "material", and "chair" on account of the limited space. Our main focus is on realistic event data. It is observed that our method maintains correct structures, especially at geometry discontinuities, such as the wheel of the "lego" and the depression in the "material". The counterpart method causes severe background noises at the depth map, whereas our method attains clearer depths. Moreover, the proposed approach delivers better contrast in the renderings, while images produced by the benchmark are blurry. The quantitative and qualitative comparisons are shown in Table I and Figure 3 respectively.

B. Evaluation on Realistic Event Data

The proposed dataset includes more than 100 objects and we randomly select seven items with different materials for establishment. Event NeRF only receives the 3D coordinates and view direction as input, and considers the event stream as learning targets. However,

TABLE III

ABLATION STUDIES FOR ALL COMPONENTS OF LOSSES.

	PSNR↑	SSIM↑	LPIPS↓
<i>Ours_{wo/geo}</i>	20.38	0.865	0.201
<i>Ours_{wo/grad}</i>	19.49	0.820	0.199
Ours	21.31	0.910	0.167

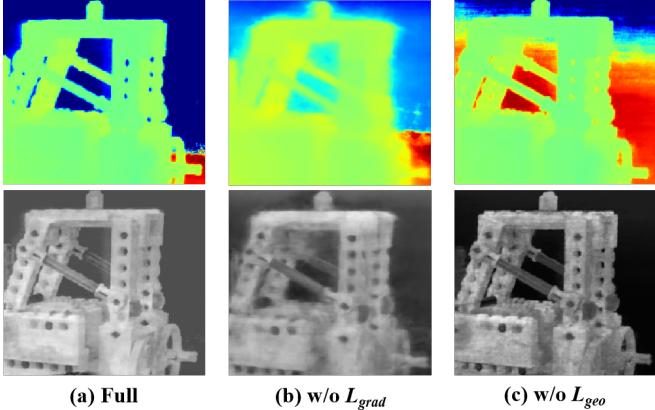


Fig. 5. Qualitative Results of ablations on synthetic data.

as stated in Figure 2, we additionally input event streams into an optimization-based prior extraction branch to build priors that guide sampling and training procedures. This intuitively will decrease the training speed. But we precompute the warp field for each camera pose timestamp. Hence, we only compute them once during training. In "helmet", EventNeRF cannot converge properly, thereby learning nothing from event streams. Our approach faithfully reconstructs the main structures of objects even if there are some fog noises around them. In "bread", "telescope", and "console", the benchmark infers wrong geometries and leads to large variances in depth predictions. In contrast, ours maintains relatively clean shapes and sharper boundaries. In "bunny" and "shoes", the benchmark incorrectly predicted the material to be translucent, while our method gives higher confidence to the geometry. All the above results are shown in Figure 4. We additionally compute the metrics for the above six scenes and the results are listed in Table II.

C. Ablation studies

We ablate the work's main contributions, including two prior-based loss functions and the density-based patch sampling strategy on synthetic and realistic event data. The quantitative results of losses on the "Lego" scene are shown in Table III. We also indicate the qualitative comparisons in Figure 5. Clearly, the gradient loss enables the model to correctly learn detailed local structures, including holes or poles. While the geometry loss optimizes global transmittance fields with less noise. Figure 6 gives ablation examples on two realistic event data, namely "wooden chair" and "telescope". It can be seen that the full model maintains the best complicated geometries and local structures such as poles and hinged joints.

To evaluate the sampling strategy, we did three counterparts containing our density-based patch sampling, random patch sampling, and anchor-based patch sampling. Notably, our gradient loss must require patches, therefore single-pixel sampling strategies such as that in EventNeRF cannot be applied. For a fair comparison, we instead use anchor-based sampling to replace, which means that we first sample several independent pixels by the EventNeRF's sampling strategy. Then we extend these single pixels as patches.

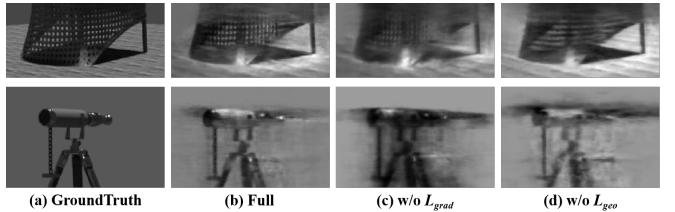


Fig. 6. Qualitative Results of ablations on realistic data.

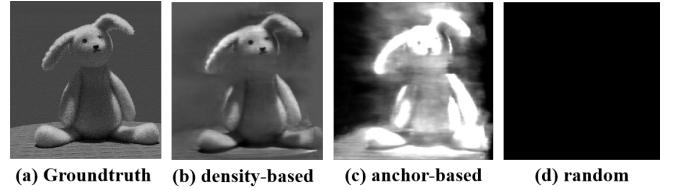


Fig. 7. Ablation studies of different sampling strategies.

Qualitative comparisons are listed in Figure 7. The random patch sampling results in complete failure because of the spatial sparsity of event data. The anchor-based method can resolve the influence of sparsity, however, it only considers the distributions at the centers of patches but ignores the sparsity over whole patches. Our density-based patch sampling outperforms the counterparts and recovers fine-grained local structures.

D. Efficiency

The proposed method learns from event data to represent 3D scenes in a more efficient way, especially when temporal event distribution is sparse. We evaluate our model and EventNeRF with different maximum window sizes to compare the learning efficiencies on different levels of sparsities. We observe that EventNeRF is not robust to the window size selection. For certain scenarios, it requires a fine hyperparameter search to ensure convergence, while our method can perform relatively well regardless of which window size we choose. Moreover, our approach converges faster than its counterpart. EventNeRF consumes about 500,000 iterations for full convergence, which approximately needs 20 hours, whereas ours only costs 200,000 steps and the time required is about 10 hours.

V. CONCLUSION

In this work, we present a novel paradigm to reconstruct NeRF from event data with strong physical augmentations by motion, geometry, and density priors. In detail, we bridge motion and geometry priors with NeRF training via the event warping field and the deterministic event generation model. Additionally, we propose the density-guided patch sampling strategy to enable the model to train more efficiently. Furthermore, we propose the first large dataset for event-based 3D representation with high-quality image labels, which contains 101 objects with various materials and geometries. This dataset contributes significantly to advancing research in the related field. We evaluate our approach on both synthetic and real datasets and it is clear that our approach is much superior to the counterpart, especially on realistic and noisy event data. Moreover, the proposed method converges more than 2 times faster. However, this method also has some limitations. For example, the performance depends on the results of the prior extraction branch. In the future, we plan to unify the two stages into a whole optimization formular to optimize the two branches simultaneously. Besides, with the help of large datasets, it is possible to explore generalizable event-based 3D representation without the need for per-scene optimization.

REFERENCES

- [1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018.
- [2] Arunkumar Byravan, Jan Humplík, Leonard Hasenclever, Arthur Brussee, Francesco Nori, Tuomas Haarnoja, Ben Moran, Steven Bohez, Fereshteh Sadeghi, Bojan Vujatovic, et al. Nerf2real: Sim2real transfer of vision-guided bipedal motion skills using neural radiance fields. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9362–9369. IEEE, 2023.
- [3] Luis A Camuñas-Mesa, Teresa Serrano-Gotarredona, Sio H Ieng, Ryad B Benosman, and Bernabé Linares-Barranco. On the use of orientation filters for 3d reconstruction in event-driven stereo vision. *Frontiers in neuroscience*, 8:48, 2014.
- [4] Chi-Ming Chung, Yang-Che Tseng, Ya-Ching Hsu, Xiang-Qian Shi, Yun-Hung Hua, Jia-Fong Yeh, Wen-Chin Chen, Yi-Ting Chen, and Winston H Hsu. Orbeez-slam: A real-time monocular visual slam with orb features and nerf-realized mapping. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9400–9406. IEEE, 2023.
- [5] Mingyue Cui, Yuzhang Zhu, Yechang Liu, Yunchao Liu, Gang Chen, and Kai Huang. Dense depth-map estimation based on fusion of event camera and sparse lidar. *IEEE Transactions on Instrumentation and Measurement*, 71:1–11, 2022.
- [6] Qiyu Dai, Yan Zhu, Yiran Geng, Ciyu Ruan, Jiazhao Zhang, and He Wang. Graspnerf: multiview-based 6-dof grasp detection for transparent and specular objects using generalizable nerf. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1757–1763. IEEE, 2023.
- [7] Tobi Delbrück, Yuhuang Hu, and Zhe He. V2e: From video frames to realistic dvs event camera streams. *arXiv e-prints*, pages arXiv–2006, 2020.
- [8] Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3867–3876, 2018.
- [9] Antea Hadviger, Igor Cvišić, Ivan Marković, Sacha Vražić, and Ivan Petrović. Feature-based event stereo visual odometry. In *2021 European Conference on Mobile Robots (ECMR)*, pages 1–6. IEEE, 2021.
- [10] Inwoo Hwang, Junho Kim, and Young Min Kim. Ev-nerf: Event based neural radiance field. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 837–847, 2023.
- [11] Hamne Kim, Stefan Leutenegger, and Andrew J Davison. Real-time 3d reconstruction and 6-dof tracking with an event camera. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VI 14*, pages 349–364. Springer, 2016.
- [12] Jurgen Kogler, Martin Humenberger, and Christoph Sulzbachner. Event-based stereo matching approaches for frameless address event stereo data. In *Advances in Visual Computing: 7th International Symposium, ISVC 2011, Las Vegas, NV, USA, September 26–28, 2011. Proceedings, Part I 7*, pages 674–685. Springer, 2011.
- [13] Beat Kueng, Elias Mueggler, Guillermo Gallego, and Davide Scaramuzza. Low-latency visual odometry using event-based feature tracks. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 16–23. IEEE, 2016.
- [14] Soomin Lee, Le Chen, Jiahao Wang, Alexander Liniger, Suryansh Kumar, and Fisher Yu. Uncertainty guided policy for active robotic 3d reconstruction using neural radiance fields. *IEEE Robotics and Automation Letters*, 7(4):12070–12077, 2022.
- [15] Marc Levoy and Pat Hanrahan. Light field rendering. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 441–452, 2023.
- [16] Yunzhi Lin, Thomas Müller, Jonathan Tremblay, Bowen Wen, Stephen Tyree, Alex Evans, Patricio A Vela, and Stan Birchfield. Parallel inversion of neural radiance fields for robust pose estimation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9377–9384. IEEE, 2023.
- [17] Jianlin Liu, Qiang Nie, Yong Liu, and Chengjie Wang. Nerf-loc: Visual localization with conditional neural radiance field. *arXiv preprint arXiv:2304.07979*, 2023.
- [18] Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Hyeongwoo Kim, Florian Bernard, Marc Habermann, Wenping Wang, and Christian Theobalt. Neural rendering and reenactment of human actor videos. *ACM Transactions on Graphics (TOG)*, 38(5):1–14, 2019.
- [19] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2019–2028, 2020.
- [20] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. A general differentiable mesh renderer for image-based 3D reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):50–62, 2020.
- [21] Shichen Liu, Shunsuke Saito, Weikai Chen, and Hao Li. Learning to infer implicit surfaces without 3d supervision. *Advances in Neural Information Processing Systems*, 32, 2019.
- [22] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: learning dynamic renderable volumes from images. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.
- [23] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019.
- [24] Dominic Maggio, Marcus Abate, Jingnan Shi, Courtney Mario, and Luca Carlone. Loc-nerf: Monte carlo localization using neural radiance fields. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4018–4025. IEEE, 2023.
- [25] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.
- [26] Jalees Nehvi, Vladislav Golyanik, Franziska Mueller, Hans-Peter Seidel, Mohamed Elgharib, and Christian Theobalt. Differentiable event stream simulator for non-rigid 3d tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1302–1311, 2021.
- [27] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020.
- [28] Liyuan Pan, Cedric Scheerlinck, Xin Yu, Richard Hartley, Miaomiao Liu, and Yuchao Dai. Bringing a blurry frame alive at high frame-rate with an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6820–6829, 2019.
- [29] Charles R. Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017.
- [30] Henri Rebecq, Guillermo Gallego, Elias Mueggler, and Davide Scaramuzza. Emvs: Event-based multi-view stereo—3d reconstruction with an event camera in real-time. *International Journal of Computer Vision*, 126(12):1394–1414, 2018.
- [31] Henri Rebecq, Timo Horstschäfer, Guillermo Gallego, and Davide Scaramuzza. EVO: A geometric approach to event-based 6-dof parallel tracking and mapping in real time. *IEEE Robotics and Automation Letters*, 2(2):593–600, 2016.
- [32] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza. High speed and high dynamic range video with an event camera. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):1964–1980, 2019.
- [33] Paul Rogister, Ryad Benosman, Sio-Hoi Ieng, Patrick Lichtsteiner, and Tobi Delbrück. Asynchronous event-based binocular stereo matching. *IEEE Transactions on Neural Networks and Learning Systems*, 23(2):347–353, 2011.
- [34] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022.
- [35] Viktor Rudnev, Mohamed Elgharib, Christian Theobalt, and Vladislav Golyanik. Eventnerf: Neural radiance fields from a single color event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4992–5002, 2023.
- [36] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent

- 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019.
- [37] Niko Sünderhauf, Jad Abou-Chakra, and Dimity Miller. Density-aware nerf ensembles: Quantifying predictive uncertainty in neural radiance fields. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9370–9376. IEEE, 2023.
- [38] Matthew Tancik, Vincent Casser, Xincheng Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022.
- [39] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019.
- [40] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics (TOG)*, 38(4):1–12, 2019.
- [41] Mukun Tong, Charles Dawson, and Chuchu Fan. Enforcing safety for vision-based controllers via control barrier functions and neural radiance fields. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10511–10517. IEEE, 2023.
- [42] Fabio Tosi, Alessio Tonioni, Daniele De Gregorio, and Matteo Poggi. Nerf-supervised deep stereo. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 855–866, June 2023.
- [43] Antoni Rosinol Vidal, Henri Rebecq, Timo Horstschaefer, and Davide Scaramuzza. Ultimate slam? combining events, images, and imu for robust visual slam in hdr and high-speed scenarios. *IEEE Robotics and Automation Letters*, 3(2):994–1001, 2018.
- [44] Hengyi Wang, Jingwen Wang, and Lourdes Agapito. Co-slam: Joint coordinate and sparse parametric encodings for neural real-time slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13293–13302, 2023.
- [45] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 52–67, 2018.
- [46] Ziyun Wang, Kenneth Chaney, and Kostas Daniilidis. Evac3d: From event-based apparent contours to 3d models via continuous visual hulls. In *European conference on computer vision*, pages 284–299. Springer, 2022.
- [47] Yuanbo Xiangli, Lining Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahu Lin. Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In *European conference on computer vision*, pages 106–122. Springer, 2022.
- [48] Lan Xu, Weipeng Xu, Vladislav Golyanik, Marc Habermann, Lu Fang, and Christian Theobalt. Eventcap: Monocular 3d capture of high-speed human motions using an event camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4968–4978, 2020.
- [49] MI Zhenxing and Dan Xu. Switch-nerf: Learning scene decomposition with mixture of experts for large-scale neural radiance fields. In *The Eleventh International Conference on Learning Representations*, 2022.
- [50] Allan Zhou, Moo Jin Kim, Lirui Wang, Pete Florence, and Chelsea Finn. Nerf in the palm of your hand: Corrective augmentation for robotics via novel-view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17907–17917, 2023.
- [51] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multi-plane images. *arXiv preprint arXiv:1805.09817*, 2018.
- [52] Yi Zhou, Guillermo Gallego, Henri Rebecq, Laurent Kneip, Hongdong Li, and Davide Scaramuzza. Semi-dense 3d reconstruction with a stereo event camera. In *Proceedings of the European conference on computer vision (ECCV)*, pages 235–251, 2018.
- [53] Yi Zhou, Guillermo Gallego, and Shaojie Shen. Event-based stereo visual odometry. *IEEE Transactions on Robotics*, 37(5):1433–1450, 2021.
- [54] Zhenxin Zhu, Yuantao Chen, Zirui Wu, Chao Hou, Yongliang Shi, Chuxuan Li, Pengfei Li, Hao Zhao, and Guyue Zhou. Latitude: Robotic global localization with truncated dynamic low-pass filter in city-scale nerf. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8326–8332. IEEE, 2023.
- [55] Zihan Zhu, Songyou Peng, Viktor Larsson, Zhaopeng Cui, Martin R Oswald, Andreas Geiger, and Marc Pollefeys. Nicer-slam: Neural implicit scene encoding for rgbd slam. *arXiv preprint arXiv:2302.03594*, 2023.
- [56] Yi-Fan Zuo, Jiaqi Yang, Jiaben Chen, Xia Wang, Yifu Wang, and Laurent Kneip. Devo: Depth-event camera visual odometry in challenging conditions. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2179–2185. IEEE, 2022.