Joint Depth and Reflectivity Estimation using Single-Photon LiDAR

Hashan K. Weerasooriya, *Student Member, IEEE,* Prateek Chennuri, *Student Member, IEEE,* Weijian Zhang, *Student Member, IEEE,* Istvan Gyongy, *Senior Member, IEEE,* and Stanley H. Chan, *Senior Member, IEEE*

***Abstract*—Single-Photon Light Detection and Ranging (SP-LiDAR) is emerging as a leading technology for long-range, high-precision 3D vision tasks. In SP-LiDAR, timestamps encode two complementary pieces of information: pulse travel time (depth) and the number of photons reflected by the object (reflectivity). Existing SP-LiDAR reconstruction methods typically recover depth and reflectivity separately, rely on one modality to estimate the other, or reconstruct only a single modality. Moreover, the conventional 3D histogram construction is effective mainly for slow-moving or stationary scenes. In dynamic scenes, however, it is more efficient and effective to directly process the timestamps. In this paper, we introduce an estimation method to simultaneously recover both depth and reflectivity in fast-moving scenes. We offer two contributions (1) A theoretical analysis demonstrating the mutual correlation between depth and reflectivity and the conditions under which joint estimation becomes beneficial. (2) A novel reconstruction method, "SP-LiDeR", which exploits the shared information to enhance signal recovery. On both synthetic and real SP-LiDAR data, our method outperforms existing approaches, achieving superior joint reconstruction quality.**
The code is available on the *project webpage*.

***Index Terms*—Single Photon-LiDAR, Feature Sharing, Joint Reconstruction, Depth, Reflectivity, Timestamp**

Hashan K. Weerasooriya, Prateek Chennuri, Weijian Zhang, and Stanley H. Chan are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47906, USA. E-mails: {hweeraso, pchennur, zhan5056, stanchan}@purdue.edu. Istvan Gyongy is with the School of Engineering, The University of Edinburgh, Edinburgh, U.K. E-mail: istvan.gyongy@ed.ac.uk.

The work is supported in part by the DARPA/SRC CogniSense JUMP2.0 Center, and in part by the National Science Foundation under the grants IIS-2133032 and ECSS-2030570.

I. Introduction

With the rapid growth of 3D applications from autonomous driving to virtual reality, Single-Photon Light Detection and Ranging (SP-LiDAR) has become one of the most important technologies for range-related products [1]–[7]. The sensing principle of modern SP-LiDAR involves sending a laser pulse train to an object and using a Single-Photon Avalanche Diode (SPAD) [8], [9] with a Time-Correlated Single-Photon Counting (TCSPC) technique [8] to record the time it takes for the pulse to travel [10], [11].

The measured timestamps of SP-LiDAR contain two pieces of information: (1) Depth—the time it takes for the pulse to travel, indicating how far the object is; and (2) Reflectivity—the number of photons reflected by the object, as shiny objects reflect photons while dull objects absorb them. Recovering both depth and reflectivity from timestamp measurements is of great interest. Typically, for better reconstruction results, conventional algorithms require multiple photon detections—around hundreds per pixel—which leads to slower acquisition speeds and limits applications to static or slow-moving scenes. However, recovering information in fast-changing environments remains an open challenge, where the choice of data acquisition strategy becomes critical to reconstruction quality.

In the literature, SP-LiDAR detections are typically processed using two approaches as summarized in Fig. 1.

1) 3D Histogram Cube: Conventionally, when a SPAD detector is employed, timestamp measurements are accumulated over multiple acquisition cycles until the desired photon

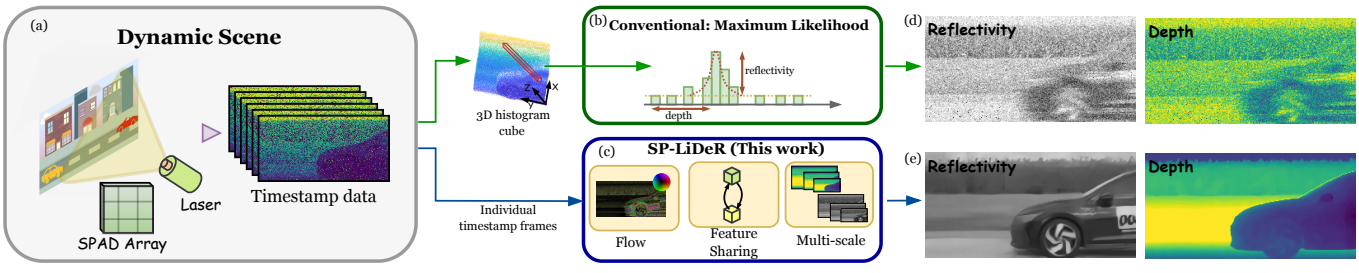


Fig. 1: **Different Processing Methods of Timestamp Data and the Corresponding Results:** (a) The SPAD sensor array captures noisy timestamp data from a dynamic scene at a high speed. (b) To mitigate noise in raw data, a conventional approach involves clustering multiple detections to form a 3D cube. Subsequently, the object's reflectivity and depth are determined by identifying the height and the location of the peak through algorithms such as maximum likelihood estimation. (c) We propose SP-LiDeR, a deep learning framework that leverages individual timestamp frames. (d) Conventional algorithms suffer from blurry results due to long integration times. (e) Our propsed method can yield better results in dynamic scene reconstruction.

level is achieved. This is because collecting more photons allows for better statistical estimation of depth and reflectivity by reducing the randomness due to shot noise, background photon detections, pulse width, and jitter time. For efficient representation of the data, a 3D histogram cube is constructed where each detection increments the corresponding time-of-flight bin count. Subsequent processing algorithms are applied to determine the location of the histogram peak which corresponds to depth, and the height of the peak which represents reflectivity. By incorporating 3D reconstruction algorithms [12]–[16], people have demonstrated 3D scene reconstruction even in cases where measurements are severely affected by scattering media or multiple surfaces [17]–[20].

However, constructing a 3D histogram cube means that we need to collect data for a longer period which has drawbacks when applied to fast-moving objects. As the object moves, accumulating photon detections over multiple acquisition cycles becomes challenging since each SPAD pixel captures timestamps of a different probability distribution. As a result, we will have multi-peak and broader-peak histograms.

2) Individual Timestamp Frame: In light of the above challenge, storing timestamp data as individual slices after each acquisition cycle is potentially more advantageous. This strategy, previously proposed by Altmann *et al*. in [21], enables timestamp frames to be captured over shorter integration periods, thereby minimizing effects caused by motion and making them well-suited for high-frame-rate reconstruction of fast-moving scenes.

Under this processing mechanism, each frame captures the first photon arrival at each pixel within a given time window. Under low-flux and high-speed conditions, the acquisition results in each pixel detecting at most one photon, yielding an average detection rate of less than one photon per pixel per frame [21]. Therefore, the amount of scene information captured per frame is limited, requiring more sophisticated algorithm designs for better reconstruction.

While both 3D histogram processing and individual timestamp frame processing can, in principle, support the simultaneous recovery of depth and reflectivity, most existing methods treat them separately. These approaches either reconstruct depth and reflectivity independently, rely on one modality to estimate the other, or focus exclusively on a single modality. However, since depth and reflectivity information is encoded in the timestamp detections, it seems feasible to *simultaneously* estimate both. This brings us to the two contributions of our paper.

1) *Does recovering depth help recover reflectivity, and vice versa?* To address this, we derive the Maximum-Likelihood Estimator (MLE) for the joint recovery problem and analyze the Cramer–Rao Lower Bound (CRLB) to theoretically establish conditions where information sharing is beneficial. This forms the foundation for joint estimation.
2) *Can we design an effective neural network for joint depth and reflectivity recovery in fast-moving scenes?* Building on our theoretical and experimental insights, We propose SP-LiDeR, a dual-channel joint estimation network with a feature-sharing mechanism. To address information loss from scarce photon detections, we align features from neighboring timestamp frames using optical flow. Our method extracts multiscale features for depth and reflectivity through two parallel branches, enhancing each modality and capturing both fine and coarse details for improved performance.

II. Prior Work

Depth Estimation. Depth reconstruction can generally be performed using Maximum-Likelihood Estimation (MLE) and its variants [22]–[26]. Under the photon-limited regime, where the photon acquisition rate is below 5% so that dead time becomes negligible [27]–[30], depth estimation is performed by match-filtering, aligning the photon timestamp registrations with the photon arrival flux function [31], [32]. In the absence of noise, the matched filter reduces to the sample mean of timestamps when the flux function follows a Gaussian shape. However, while this method performs well with high photon counts, its accuracy degrades under strong background illumination or low photon detections. This is because the matched filter is prone to being trapped in local optima due to its non-convexity. Therefore, pre-processing [12], [13], [22], [33] is required to reject outliers before estimation while leveraging spatial correlations. Advanced depth recovery methods using neural networks are also available [14], [34]–[37], often incorporating modalities such as intensity or monocular depth images [15], [37], [38]. All these methods use a 3D data cube as input.

Reflectivity Estimation. Reflectivity reconstruction is typically formulated via another MLE of photon counts [13], [22], [23], [39], as the refelectivity is proportional to the number of photons reflected by the object. Because of the photon-counting nature, the estimation is formulated by Poisson distribution and/or Binomial distribution for a sum of binary photon detections, similar to quanta image sensors [40]–[48]. Most previous work on reflectivity estimation is based on photon counts rather than the timestamps [12], [49]. To avoid the exhaustive grid search and the requirement of ground-truth depth in solving the joint MLE problem for reflectivity estimation, an alternative approach [12] proposes performing noise censoring first and then computing the photon count MLE. Although this method achieves better results, we observe that the information provided by timestamp detections is not fully utilized for reflectivity estimation.

Joint Estimation. Joint estimation of depth and reflectivity is theoretically beneficial but highly challenging due to the continuous, alternating iterative search between the two parameters. [50] adopts a Bayesian formulation, framing joint estimation as a single inference problem, while [18], [51] jointly estimates depth and reflectivity using Alternating Direction Method of Multipliers (ADMM) algorithm. To our knowledge, the only deep learning method that considers joint feature extraction from depth and reflectivity is [16], but its decoders are designed independently.

Dynamic Scenes. Depth and reflectivity estimation in dynamic scenes is even more challenging because the events can no longer be aggregated without compensating for motion,

which affects estimation accuracy. While current state-of-the-art methods can handle moving scenes, they rely on multiple timestamp detections per pixel. Consequently, rapid motion relative to the integration time leads to blurred results. Altmann *et al*. [21] proposed a depth map reconstruction method using individual photon detection frames, employing a Bayesian approach to model 3D profile dynamics. [52] proposes an alternative approach that differs from ours as it uses high-resolution photon-counting frames and multi-event low-resolution timestamp frames in an interleaved manner to achieve high-frame-rate reconstruction, whereas our method relies solely on timestamp data.

III. Depth and Reflectivity Information Sharing

Most Single Photon LiDAR (SP-LiDAR) reconstruction algorithms focus solely on depth reconstruction or first decode reflectivity before using it to infer depth. However, since a timestamp data encodes both depth and reflectivity, it would seem more beneficial to reconstruct them simultaneously. The only existing deep learning model for joint depth and reflectivity reconstruction employs two separate decoders to extract depth and reflectivity features independently [16]. However, this strict separation is not ideal, as depth and reflectivity features are inherently interconnected.

To that end, this section first theoretically examines how and when the dependency between depth and reflectivity manifests within the per-pixel Maximum-Likelihood Estimation (MLE) framework. Second, we illustrate the shared feature sets between depth and reflectivity using a toy problem. The insights gained from this analysis and the experiments will be incorporated into our proposed algorithm to enhance the feature-sharing mechanism. Proofs of theorems and corollaries are provided in the Appendix and Supplementary Materials.

A. Notation

The notations of this paper follow the literature, e.g., [12], [13], [32]. We assume that the SP-LiDAR operates in the low-flux regime so that the dead time is negligible. We also assume that there exists a single bounce per pixel with no depth ambiguity and that the object is quasi-static within the per-frame exposure time. Then, the photon arrival can be modeled as an inhomogeneous Poisson process with a mean rate [31], [53]

$$\lambda_{i,j}(\ell,t) = \eta\alpha_{i,j}(\ell)s\left(t - \frac{2z_{i,j}(\ell)}{c}\right) + b_\lambda(\ell), \qquad (1)$$

where c is the speed of light, $\tau_{i,j}(\ell) = 2z_{i,j}(\ell)/c$ is the time delay that carries the true depth $z_{i,j}$, and $s(\cdot)$ is a Gaussian-shaped pulse with energy S.

The reflectivity of the object is described by the parameter α. The uniformly distributed parameter $b_\lambda(\ell) = \eta\lambda_b(\ell) + \lambda_d(\ell) \sim \mathcal{U}(0, t_r)$ is the background noise which contains the ambient light $\lambda_b(\ell)$ and the dark current $\lambda_d(\ell)$, with η being the quantum efficiency. To specify the pixel index, we use (i, j). The index ℓ is the rank of the frame. Tab. I provides a summary of these notations.

TABLE I: Notation overview.

Symbols	Meaning	Symbols	Meaning
t_r	repetition period	$s(\cdot)$	pulse shape
N_r	# repetition / frame	σ_t	pulse width
z	ground truth depth	λ_b	background rate
$\tau = 2z/c$	true time delay	η	quant. eff.
α	true reflectivity	λ_d	dark current

Given $\lambda_{i,j}(\ell,t)$, the core quantity we are interested in is the distribution of the timestamps. Following prior work such as [12], [31], [32], [49], the distribution is defined according to the theorem below.

Theorem 1 (Joint density of M timestamps $\mathbf{t}_M$ [31], [49])**.** *Let* $\mathbf{t}_M = \{t_k\}_{k=1}^M$. *For* $M \geq 1$,

$$p\left[\mathbf{t}_M, M = m\right] = \frac{e^{-N_r\Lambda(\alpha)}}{m!}\prod_{k=1}^{m} N_r\lambda(t_k;\alpha,\tau). \qquad (2)$$

In this equation, $\Lambda(\alpha)$ specifies the per-cycle energy of the photon flux, obtained by integrating $\lambda(t)$ over the repetition period t_r, via $\Lambda(\alpha) = \int_0^{t_r}\lambda(t)\,dt = \eta\alpha S + B$, where $\eta\alpha S$ is the signal energy and $B = b_\lambda t_r$ is the noise energy. The total expected energy per frame is $N_r\Lambda(\alpha)$. We define Signal to Background Ratio (SBR) as $\text{SBR} = \eta\alpha S/B$.

The core signal estimation problem is formulated as a joint Constrained Maximum Likelihood (CML) estimation:

$$(\widehat{\tau},\widehat{\alpha}) = \underset{0<\tau<t_r,\alpha\geq 0}{\text{argmax}} \Big\{ -N_r\eta S\alpha + \sum_{k=1}^{m}\log\left(\eta\alpha s\left(t_k-\tau\right)+b_\lambda\right)\Big\}, \qquad (3)$$

where evidently depth and reflectivity rely on each other as long as $b_\lambda > 0$. In prior works such as [12], the joint estimation is solved via two separable problems when one eliminates skeptical outliers and then assumes *zero* background noise:

Corollary 1. *When* $b_\lambda = 0$, *Eq.* (3) *simplifies to*

$$\widehat{\alpha} = \underset{\alpha\geq 0}{argmax}\left\{m\log\alpha - N_r\eta\alpha S\right\} = \frac{m}{N_r\eta S}, \qquad (4)$$

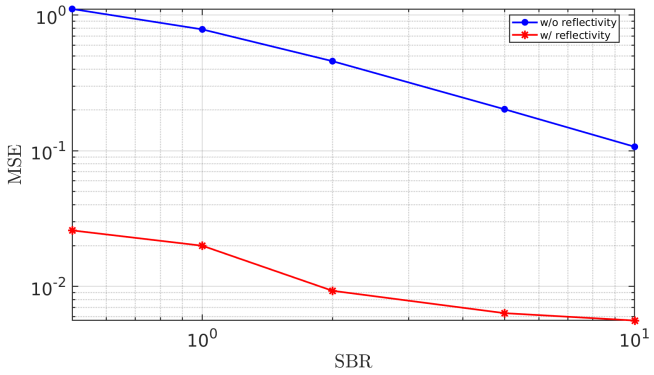
$$\widehat{\tau} = \underset{0<\tau<t_r}{argmax}\left\{\sum_{k=1}^{m}\log\left(s\left(t_k-\tau\right)\right)\right\} = \frac{1}{m}\sum_{k=1}^{m}t_k, \qquad (5)$$

meaning that they can be solved separately.

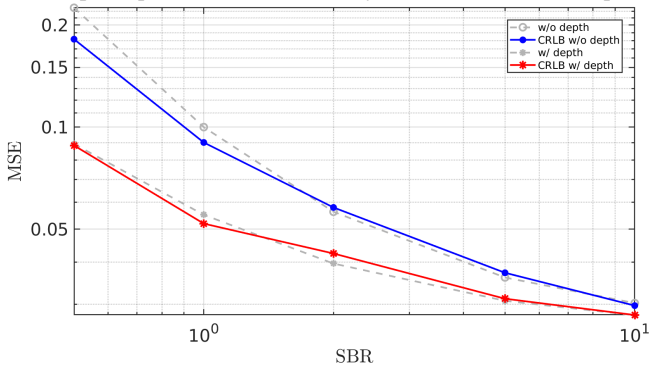
Corollary 1 states that when α and τ are separable (i.e., $b_\lambda = 0$), recovering one will *not* theoretically offer any help in recovering the other.

B. Information sharing under MLE

In this subsection, we discuss the theoretical intuition behind how depth and reflectivity can help each other when $b_\lambda \neq 0$. **Does reflectivity help depth?** This direction is straightforward because, without the knowledge of reflectivity, the least we can do is run a log-matched filter estimator, assuming no noise exists, as indicated in Eq. (5). This is a simple but less



(a) Depth estimation results with and without prior knowledge of reflectivity. When reflectivity is unknown, we use the mean value of timestamps (Eq. (5)). When reflectivity is known, we solve Eq. (3).



(b) Reflectivity estimation results with and without prior knowledge of depth. When depth is unknown, we use Eq. (6) to obtain reflectivity. When depth is known, we solve Eq. (8).

Fig. 2: **Performance Analysis of MLE.** The accuracy of various maximum likelihood estimations under per-pixel regime across varying signal-to-background ratios. The performance gap increases as noise becomes more dominant.

accurate approximation compared to solving Eq. (3) with prior knowledge of α and a search algorithm. Fig. 2a shows that the depth estimator with the knowledge of reflectivity consistently outperforms the one without, demonstrating reflectivity helps depth estimation.

Does depth help reflectivity? The reverse is non-trivial, as photon count is typically used to estimate reflectivity, and the additional information embedded in timestamps has not been explored for reflectivity estimation until [49]. The former estimation, Eq. (6), which is based on photon counts, does not require depth information, whereas the latter does. We compare these two estimators here.

Reflectivity Estimator w/o depth. Since the number of photon counts is a Poisson random variable that only conveys reflectivity information, we can derive a CML estimator of reflectivity that does not depend on depth

$$\widehat{\alpha}_{\mathrm{c}} = \max\Big\{\underbrace{\frac{1}{\eta S}\left(\frac{m}{N_r} - B\right)}_{\widehat{\alpha}_{\mathrm{c}}^*},\ 0\Big\}, \tag{6}$$

where $\widehat{\alpha}_{\mathrm{c}}^*$ is an **unconstrained** MLE for which we can derive the CRLB to analyze its performance.

Corollary 2. *The CRLB of the unconstrained MLE $\widehat{\alpha}_c^*$ is*

$$\mathrm{Var}\left[\widehat{\alpha}_c^*\right] \geq \frac{\eta S\alpha + B}{N_r\eta^2 S^2} = \frac{1 + 1/SBR}{N_r(\eta S/\alpha)}. \tag{7}$$

✠ *Remark: The variance of the estimator improves when we have more data, higher SBR, or larger system energy (ηS).*

Reflectivity Estimator w/ depth. The CML estimate depending on depth can be derived from Eq. (3) as

$$\widehat{\alpha}_t = \max\left\{\widehat{\alpha}_t^*, 0\right\},$$

where $\widehat{\alpha}_t^*$ is the largest root to the nonlinear equation below

$$\sum_{k=1}^{m} \frac{\eta s(t_k - \tau)}{\eta\widehat{\alpha}_t^* s(t_k - \tau) + B/tr} = N_r\eta S. \tag{8}$$

To solve Eq. (8), the knowledge of τ is required. More details on the optimization of this equation can be found in the Supplementary Materials. For this estimator, we can derive the CRLB of the **unconstrained** MLE $\widehat{\alpha}_t^*$.

Corollary 3. *The CRLB of the unconstrained MLE $\widehat{\alpha}_t^*$ is*

$$\mathrm{Var}\left[\widehat{\alpha}_t^*\right] \geq \left[N_r\eta^2 \int_0^{t_r} \frac{s^2(t-\tau)}{\eta\alpha s(t-\tau) + b_\lambda}\, dt\right]^{-1}. \tag{9}$$

Theorem 2 (CRLB comparison between the reflectivity estimators)**.** *The variance of $\widehat{\alpha}_t^*$ is uniformly lower than that of $\widehat{\alpha}_c^*$ that does not, i.e.*

$$\underbrace{\left[N_r\eta^2 \int_0^{t_r} \frac{s^2(t-\tau)}{\eta\alpha s(t-\tau) + b_\lambda}\, dt\right]^{-1}}_{\textit{w/ depth}} \leq \underbrace{\frac{\eta S\alpha + B}{N_r\eta^2 S^2}}_{\textit{w/o depth}}, \tag{10}$$

where the equality holds if and only if $b_\lambda = 0$, implying that they are equivalent when there is no noise.

The proof of this theorem can be found in the Appendix. Theorem 2 provides a theoretical justification for why depth can help reflectivity estimation. The numerical results in Fig. 2b confirm our theory. We remark that this results is also consistent with those from [49]. However, the context of our results is slightly different: In [49], a sequential estimation was developed whereas in our work, we aim for a joint estimation.

In Fig. 2b, we compare the MSE of $\widehat{\alpha}_{\mathrm{t}}$ and $\widehat{\alpha}_{\mathrm{c}}$. It can be seen that $\widehat{\alpha}_{\mathrm{t}}$ has a lower MSE[1]. This confirms that, under the per-pixel MLE regime, incorporating prior depth knowledge improves reflectivity reconstruction accuracy across all signal-to-background ratios, compared to reflectivity estimation based solely on photon counts arising from Poisson sampling.

In summary, we conclude the following:

When $\lambda_b > 0$, depth helps reflectivity, and reflectivity helps depth.

[1]The observed experimental MSE sometimes goes below the CRLB. This occurs because the positivity constraint imposed on α pulls negative estimations to zero, thereby reducing the error relative to the actual ground truth. Due to the randomness of Poisson realizations, the estimation performance can exceed the theoretical limit set by the CRLB under this positivity constraint.

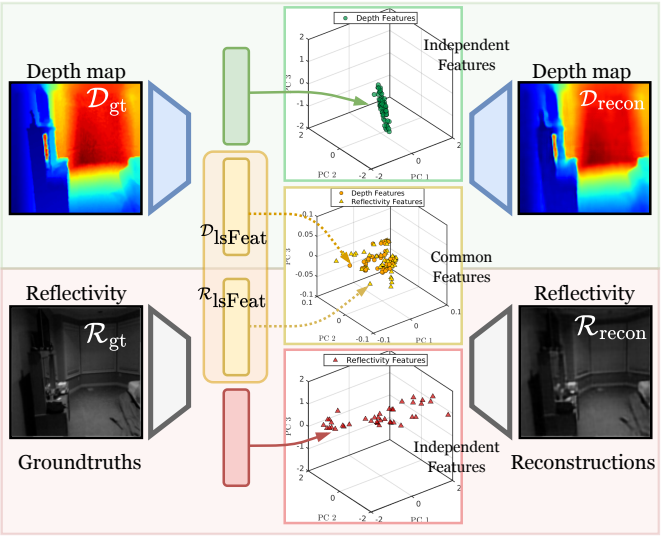


Fig. 3: **Information Sharing Pilot Study.** The setup involves two autoencoders attempting to reconstruct the inputs while isolating the common features in the latent space. The reconstruction results and the corresponding feature distribution verify the claims about shared features. "PC 1", "PC 2" and "PC 3" on the axes represent the 1th, 2nd and 3rd principal components.

C. Information Sharing in the Feature Space

Now that we have a theoretical justification for how depth and reflectivity are complementary, we ask the question: *Can this phenomenon be observed even in deep learning implementations?*

To answer this question, we design a toy experiment to assess the degree of feature overlap between depth and reflectivity. We employ two identical convolutional autoencoders: one to reconstruct reflectivity and the other to reconstruct depth. We simultaneously minimize the distance between specific latent features and the reconstruction losses by using the objective function

$$\begin{aligned}\mathcal{L}_{\text{all}} = \text{MSE}(\mathcal{D}_{\text{gt}}, \mathcal{D}_{\text{recon}}) + \text{MSE}(\mathcal{R}_{\text{gt}}, \mathcal{R}_{\text{recon}}) \\ + \sigma \cdot \text{MSE}(\mathcal{D}_{\text{lsFeat}}, \mathcal{R}_{\text{lsFeat}}).\end{aligned} \tag{11}$$

For the toy autoencoder network training, we utilized the NYU V2 RGB-D dataset [54] and assigned a value of 0.5 to σ.

Fig. 3 shows an interesting phenomenon of a randomly pick image. We compare the reconstruction results alongside the corresponding encoded features, visualized in a low-dimensional space. The excellent reconstruction results for both depth and reflectivity indicate that the encoder successfully captures and retains the crucial features needed for accurate reconstruction. The figure further demonstrates that some depth and reflectivity features are clustered in the latent space. This finding aligns with previous work [15], [55]–[58] and supports the hypothesis that there is a degree of feature sharing between depth and reflectivity. Additional experimental details can be found in the Supplementary Materials.

IV. SP-LiDeR Network

This section outlines the four main stages of the proposed SP-LiDeR (**S**ingle-**P**hoton **LiDAR** joint **De**pth and **R**eflectivity) architecture. The core module is Cross-Modal Information Fusion, supported by three additional modules. The overview of the architecture, highlighting the information-sharing module that enhances depth and reflectivity features, is shown in Fig. 4.

A. Main Module: Cross-Modal Information Fusion

Having established an understanding of the connection between depth and reflectivity feature sets, we now address the challenge of effectively implementing a feature-sharing mechanism in our proposed solution. We directly input noisy timestamp frames, from which noisy binary frames are generated via thresholding. This process does not introduce any new reflectivity information beyond what is already present in the timestamp data. However, the two-channel module estimation procedure for depth and reflectivity is designed to ensure that each modality retains its own distinct feature set for accurate reconstruction.

Due to noisy input data, the full feature set of each modality may not always be available, but some common features may be uniquely observed by one modality. A feature-sharing mechanism is therefore beneficial for completing missing features. The proposed two-channel network enables the sharing of common features between channels. We introduce a novel attention module, CCAM (Convolutional Cross-Attention Module), which extracts key temporal and spatial features from the other modality and enables feature sharing to improve overall reconstruction, as shown in Fig. 5. Since timestamp frames are the only unique input, we expect more information to flow from depth to reflectivity than vice versa.

The proposed CCAM module is inspired by Convolutional Block Attention Module (CBAM) [59] and the attention mechanism [60]. The main functionalities of the these modules are as follows:

1) CBAM: How to extract the features? To enable efficient cross-attention, we require a compact summarization of the warped feature maps $\mathcal{F}^w$, as the attention complexity $\mathcal{O}(NM)$ grows rapidly with input lengths N and M. To that end, we use CBAM attention modules to generate a channel attention map $\phi_c(\mathcal{F}^w) \in \mathbb{R}^{C\times 1}$ and a spatial attention map $\phi_s(\mathcal{F}^w) \in \mathbb{R}^{H\times W}$:

$$\phi_c(\mathcal{F}) = \sigma_s(\text{MLP}(\mathcal{F}_c^{\text{avg}} + \mathcal{F}_c^{\text{max}})) \tag{12}$$

$$\phi_s(\mathcal{F}) = \sigma_s(\text{Conv}([\mathcal{F}_s^{\text{avg}}; \mathcal{F}_s^{\text{max}}])) \tag{13}$$

where $\mathcal{F}_i^{\text{avg}}$ and $\mathcal{F}_i^{\text{max}}$ for $i \in \{c, s\}$ are the channel or spatial average and max-pooled maps of the warped features $\mathcal{F}^w$, and σ_s is the sigmoid function. MLP denotes a multilayer perceptron, while Conv indicates the convolution operation.

This process is repeated to produce P distinct spatial and channel attention maps, which are stacked into unified feature arrays $\Phi_c(\mathcal{F}^w) \in \mathbb{R}^{P\times C}$ and $\Phi_s(\mathcal{F}^w) \in \mathbb{R}^{P\times HW}$. The channel attention module output, $\Phi_c(\mathcal{F}^w)$, identifies salient features ('what') within the input warped frames, while the

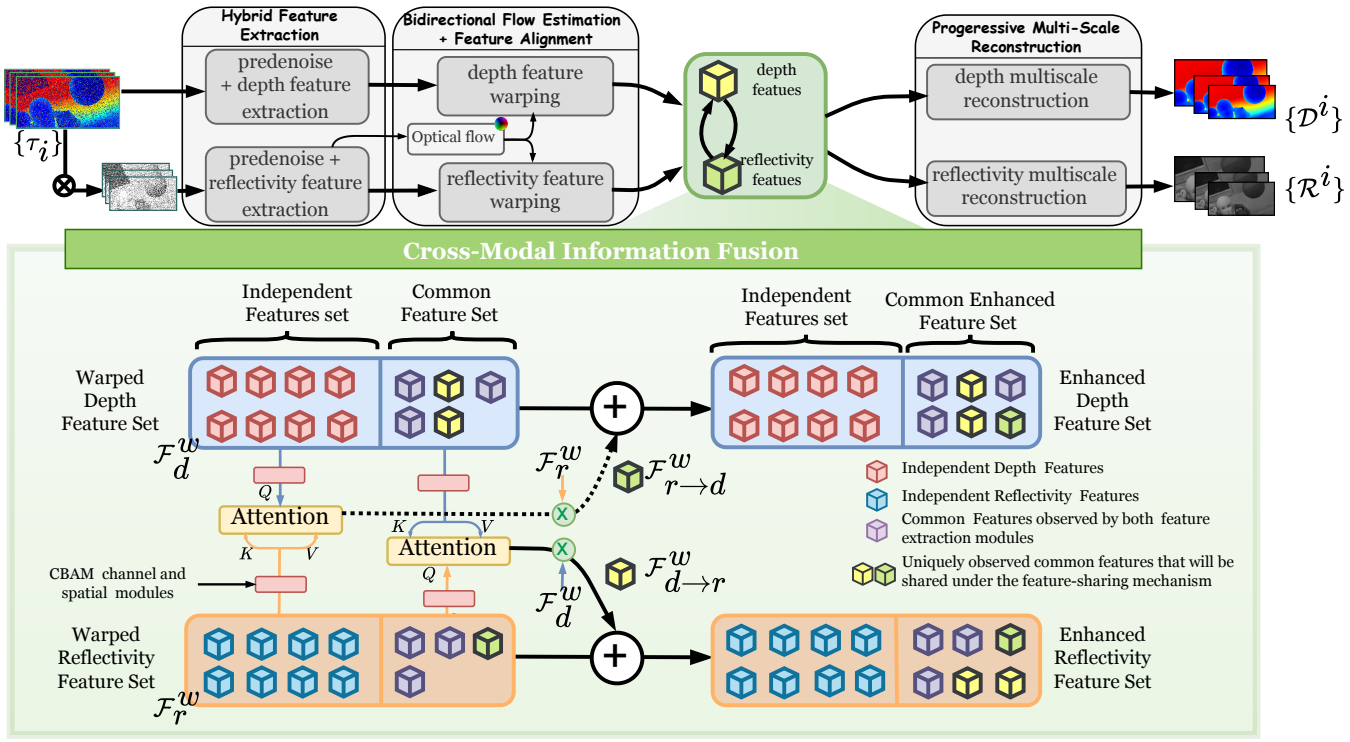


Fig. 4: **The Outline of the SP-LiDeR Network Architecture**. The proposed SP-LiDeR network consists of four main modules. Given several individual timestamp frames directly from a SPAD array, binary frames are generated via a thresholding process: a pixel is marked as 1 if a timestamp exists, and 0 otherwise. Next, depth and reflectivity features from several adjacent input frames are extracted using two feature extraction networks. These features are then warped together based on the optical flow between frames. Subsequently, a cross-model information fusion module is employed to enhance the feature sets of the respective modalities. This module identifies the uniquely observed features by the depth and reflectivity feature extraction mechanisms and fuses them cohesively. Warping and information sharing are performed at three resolution scales to capture both fine and coarse details, improving feature representation. Finally, the multi-scale reconstruction network simultaneously reconstructs depth and reflectivity.

spatial attention module output, $\Phi_s(\mathcal{F}^w)$, pinpoints their locations ('where'). Together, these modules yield a concise representation of both channel and spatial information that is useful in the feature-sharing mechanism. These channel and spatial attention maps are generated for both depth $\mathcal{F}_d^w$ and reflectivity $\mathcal{F}_r^w$ warped features independently, yielding $\Phi_s(\mathcal{F}_d^w)$, $\Phi_s(\mathcal{F}_r^w)$, $\Phi_c(\mathcal{F}_d^w)$, and $\Phi_c(\mathcal{F}_r^w)$.

2) Cross attention: How to share the features? Cross-attention is then implemented, enabling feature sharing. Cross-attention facilitates the model's ability to learn the interdependencies between two input sequences. We explain the feature-sharing mechanism for channel attention maps from reflectivity to depth as follows:

$$\begin{aligned}
&\text{Attention}(Q,K,V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,\\
&\text{head}_c^n = \text{Attention}(Q_c^d W_{n,c}^Q, K_c^r W_{n,c}^K, V_c^r W_{n,c}^V),\\
&\text{MHA}_c^d(Q_c^d, K_c^r, V_c^r) = \text{Concat}(\text{head}_c^1, \ldots, \text{head}_c^P)W_c^o, \quad (14)
\end{aligned}$$

where $Q_c^d = \Phi_c(\mathcal{F}_d^w)$, $K_c^r = V_c^r = \Phi_c(\mathcal{F}_r^w)$, and d_k denotes the size of Q or K. Cross-attention uses depth data (Q_c^d: queries) to find important parts of reflectivity data (K_c^r: keys). It then outputs a weighted version of the reflectivity data (V_c^r: values). We employ a single attention head, which proves sufficient for the feature-sharing mechanism. Spatial cross-attention is defined similarly. The depth-feature-infused reflectivity features for channel attention maps, denoted as $\mathcal{F}_{c,r\to d}^w$, are produced by the element-wise multiplication of the cross-attention maps with the warped reflectivity features. Similarly, the reflectivity-feature-infused depth features for channel attention maps, $\mathcal{F}_{c,d\to r}^w$, are generated through the same process. $\mathcal{F}_{s,d\to r}^w$ and $\mathcal{F}_{s,r\to d}^w$ are similarly defined for spatial attention maps. The generation of these infused feature maps is illustrated in Fig. 6.

Eventually, to obtain the overall infused feature maps from depth to reflectivity and reflectivity to depth, we use

$$\mathcal{F}_{d\to r}^w = \sigma(\mathcal{F}_{s,d\to r}^w + \mathcal{F}_{c,d\to r}^w) \quad (15)$$

$$\mathcal{F}_{r\to d}^w = \sigma(\mathcal{F}_{s,r\to d}^w + \mathcal{F}_{c,r\to d}^w) \quad (16)$$

as specified in [59]. This mechanism enables cross-modality fusion, enhancing the original feature map from the respective channel.

B. Supporting Modules

*1) **Hybrid Feature Extraction**:* SP-LiDeR simultaneously processes multiple photon timestamp slices, $\{\mathcal{T}_i\}_{i=0}^K$, where

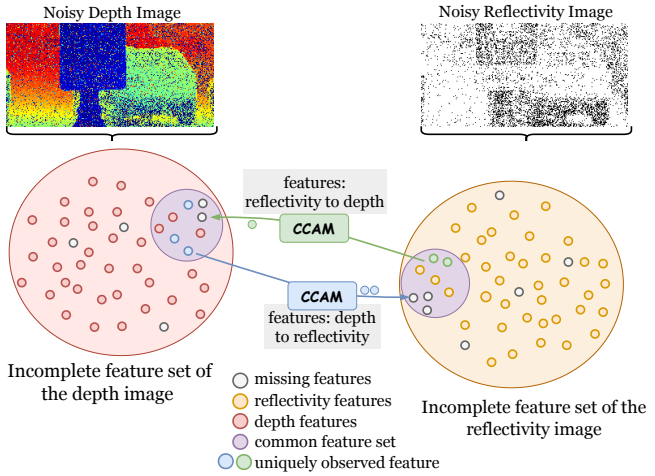


Fig. 5: **Functionality of the CCAM Module.** The CCAM module identifies common features uniquely observed by each channel's feature extraction mechanism and shares them across channels. Due to noisy input, the feature set of the input images is incomplete compared to the feature set of the ground truths.

K is the number of input timestamp frames, to enhance the reconstruction quality of the targeted reference frame.

Due to noise from dark current, background photons, pulse width, and shot noise, raw data often has a low Signal to Noise Ratio (SNR). To improve SNR, we denoise both timestamp frames and quantized binary frames obtained from the former. While denoising removes fine details such as textures and motion cues, it preserves the dominant scene structure. By combining high-SNR structural features with fine-grained details from noisy data, we achieve enhanced feature sets $\{\mathcal{F}_{p,i}^{\text{noisy}}\}_{i=1}^{K}$, $\{\mathcal{F}_{p,i}^{\text{denoised}}\}_{i=1}^{K}$, where $p \in \{d, r\}$ represents depth and reflectivity. Depth features are extracted from both denoised and noisy timestamps, whereas reflectivity estimation may require additional cues beyond direct noisy

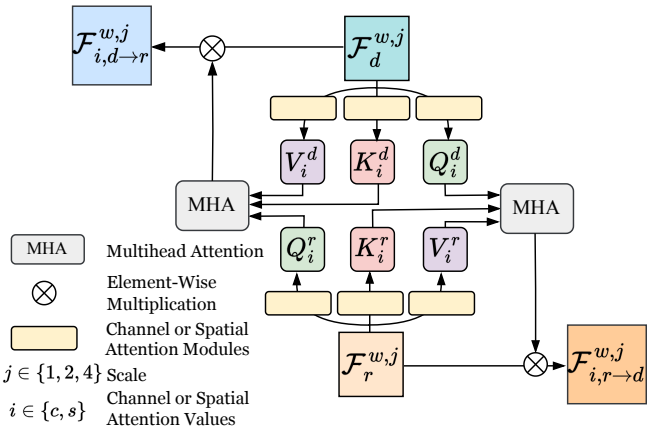


Fig. 6: **Convolutional Cross Attention Module** (CCAM) demonstrates how warped features are used to identify the most relevant features that can be shared across depth and reflectivity channels through an attention head.

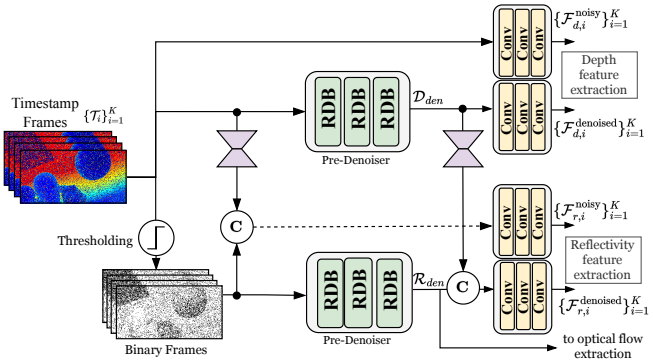


Fig. 7: **Hybrid Feature Extraction Module** denoises timestamp frames and corresponding binary frames separately. Depth and reflectivity features are extracted independently using both denoised and noisy features. Denoised binary frames are used to estimate the optical flow.

binary frames. Therefore, we use an autoencoder to share coarse feature sets with the reflectivity channel. The overview of the supporting module is depicted in Fig. 7.

*2) **Bidirectional Flow Estimation + Feature Alignment**:* Since the scene is dynamic, simply accumulating neighboring features without accounting for spatial motion across frames is suboptimal. To address this, we propose two modules for flow estimation and temporal data alignment. As illustrated in Fig. 8, The core idea of our model is to progressively accumulate features, extracted from the hybrid feature extraction module, from both directions of the reference frame N. This warping process is facilitated by an optical flow estimation mechanism, ultimately producing an improved set of features, $\mathcal{F}_N^w$, for frame reconstruction.

Inspired by [61], we designed the multi-scale Iterative and Bidirectional Flow Estimation (IBFE) module, which utilizes denoised reflectivity frames for robust flow estimation from noisy sensor data. Additionally, we introduced the Spatial-Temporal Alignment with Residual Refinement (STAR) module, which warps and integrates multi-scale depth and reflec-

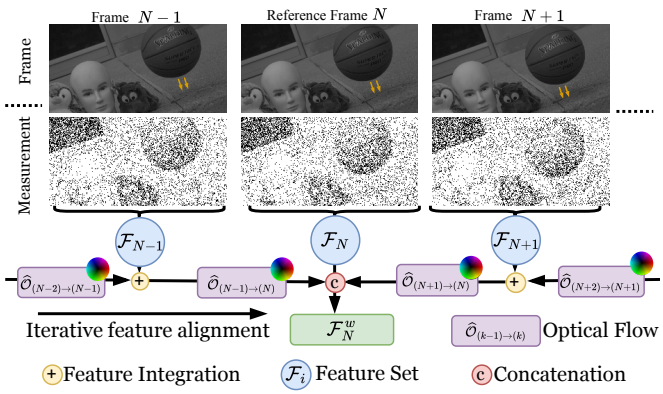


Fig. 8: **Bidirectional Feature Alignment Module** progressively propagates features from both directions to obtain the reference frame N warped features $\mathcal{F}_N^w$.

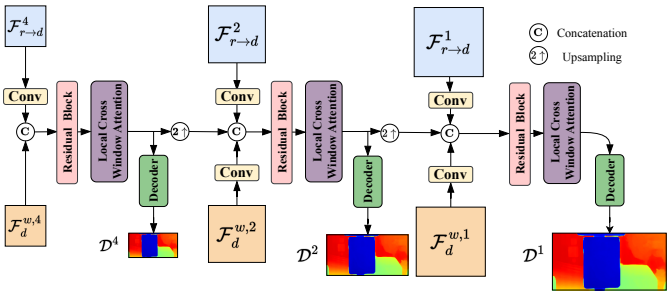


Fig. 9: **Progressive Multi-Scale Reconstruction Module** demonstrates how to reconstruct the multiscale depth maps $\{\mathcal{D}^j\}_j$ for $j = \{1, 2, 4\}$.

tivity features separately, resulting in feature sets $\mathcal{F}_p^{w,j}$, where $j \in \{1, 2, 4\}$ represents the different spatial scales.

*3) **Progressive Multi-Scale Reconstruction**:* Along with the multi-scale reflectivity $\{\mathcal{F}_r^{w,j}\}_j$ and depth $\{\mathcal{F}_d^{w,j}\}_j$ warped feature maps, we use the multi-scale infused features $\{\mathcal{F}_{d\to r}^{w,j}\}_j$ for reflectivity estimation and the multi-scale features $\{\mathcal{F}_{r\to d}^{w,j}\}_j$ for depth estimation.

Starting with the lowest scale $j = 1$, we combine information from the respective warped features with the infused feature maps using a customized residual network module. Next, the local cross-window attention module [62] is applied to generate the output for that scale. The infused feature map is then upscaled and refined at each subsequent scale using the attention and warped feature maps for that scale, repeating this process until the full-scale output is achieved as in Fig. 9.

Loss Function: We optimize the weights of channel branch simultaneously using the following loss function:

$$\begin{aligned}\mathcal{L}_{\mathcal{H}} = \lambda_1^{\mathcal{H}}\mathcal{L}(\mathcal{H}_{\text{gt}}^1, \mathcal{H}_{\text{den}}) + \lambda_2^{\mathcal{H}}\mathcal{L}(\mathcal{H}_{gt}^1, \mathcal{H}^1) + \\ \lambda_3^{\mathcal{H}}\mathcal{L}(\mathcal{H}_{\text{gt}}^2, \mathcal{H}^2) + \lambda_4^{\mathcal{H}}\mathcal{L}(\mathcal{H}_{\text{gt}}^4, \mathcal{H}^4) + \lambda_5^{\mathcal{H}}\mathcal{P}(\mathcal{H}_{\text{gt}}^1, \mathcal{H}^1),\end{aligned} \tag{17}$$

Where $\mathcal{H}^j \in \{\mathcal{D}^j, \mathcal{R}^j\}$,respectively, denotes the depth or reflectivity output, while j denotes the scale of the output. $\lambda_i^{\mathcal{H}}$ is a constant that controls the strength of the regularization for $i = \{1, 2, 3, 4, 5\}$. Moreover, $\mathcal{H}_{\text{den}}$ and $\mathcal{H}_{\text{gt}}^j$ refer to the corresponding denoised outputs from the initial denoisers and multiscale ground truth frames respectively. $\mathcal{P}$ represents the LPIPS loss, and $\mathcal{L}(\mathcal{A}, \mathcal{B})$ is defined as in Eq. (18).

$$\begin{aligned}\mathcal{L}(\mathcal{A}, \mathcal{B}) = \|\mathcal{A} - \mathcal{B}\|_1 \\ + \|\nabla_x\mathcal{A} - \nabla_x\mathcal{B}\|_1 + \|\nabla_y\mathcal{A} - \nabla_y\mathcal{B}\|_1.\end{aligned} \tag{18}$$

Here, ∇_x and ∇_y denotes the operations of horizontal and vertical gradients.

V. Experiments

In this section, we demonstrate the effectiveness of our proposed method through experiments on both simulated and real-world datasets. We further validate our approach by comparing it against SP-LiDAR and video reconstruction methods, and provide an ablation study to analyze the impact of key architectural components.

A. Timestamp Simulation Pipeline

We simulate the first-photon time-correlated imaging mechanism, as detailed in [63], for the training and testing of both existing and proposed deep learning methods. This data acquisition mechanism captures only the first photon during each acquisition cycle. As a result, some pixels may receive no photons, leading to an average of less than one photon detection per pixel per frame.

Existing SP-LiDAR architectures, such as those in [14]–[16], typically showcase reconstruction results for static scenes. In such scenarios, it is feasible to fix the signal photon levels (and thus the SBR) by adjusting the scene's reflectance, as signal photon levels are proportional to reflectance while the noise level remains constant. However, maintaining specific SBR values in dynamic scenes is impractical, as it would require altering the reflectance in every frame, This approach does not represent the actual reflectance variations across frames. Therefore, unlike previous methods, we showcases our results using a simulation mechanism that dynamically adjusts the SBR—ranging from 5 to 10—based on the reflectance of each frame.

Given that there is a photon detection in the (i, j), i.e., $M = 1$ in Eq. (2), the timestamp distribution of the pixel can be modeled as a mixture of distributions [13], [32]:

$$\begin{aligned}p[t_{i,j}|M = 1] = \frac{\eta\alpha_{i,j}S}{\eta\alpha_{i,j}S + B}\left(\frac{s\left(t_{i,j} - \frac{2z_{i,j}}{c}\right)}{S}\right) + \\ \frac{B}{\eta\alpha_{i,j}S + B}\left(\frac{1}{t_r}\right).\end{aligned} \tag{19}$$

We adopt the LiDAR setup specifications from [64] to calculate the parameters S, $\alpha_{i,j}$, η, B, and t_r.

Since publicly available depth datasets typically operate within a range of 30 Hz to 60 Hz, their temporal resolution is insufficient to replicate the rate of change in object positions captured by single-photon LiDAR sensors, which typically operate at several hundred kHz frame rates. To address this limitation, we use the I2-2000FPS high-speed RGB video dataset [42], which contains 280 videos. To generate depth maps $z_{i,j}$ from the RGB video dataset, we employ the pre-trained *Depth-Anything v2* network [65]. Finally, we use the Monte Carlo method to simulate 'first-photon behavior' to obtain timestamp frames. Further details of the simulation pipeline are provided in the Supplementary Materials.

Training of SP-LiDeR. We train the network by minimizing the loss function given in Eq. (17), with parameters $\lambda_1^{\mathcal{H}} = 0.2$, $\lambda_2^{\mathcal{H}} = 0.85$, $\lambda_3^{\mathcal{H}} = 0.1$, $\lambda_4^{\mathcal{H}} = 0.05$, $\lambda_5^{\mathcal{D}} = 0$, and $\lambda_5^{\mathcal{R}} = 0.05$. The Adam optimizer [66] is used with an initial learning rate of 10^{-4}. Every time the loss of the network plateaus, the learning rate is reduced by a factor of 0.5. Training is carried out on an NVIDIA A100 Tensor GPU. A subset of 249 out of the 280 depth videos is used for training, leaving 31 videos for evaluation purposes.

B. Synthetic Data Experiments

The comparative experimental results of the proposed approach and existing algorithms are presented in two parts. For

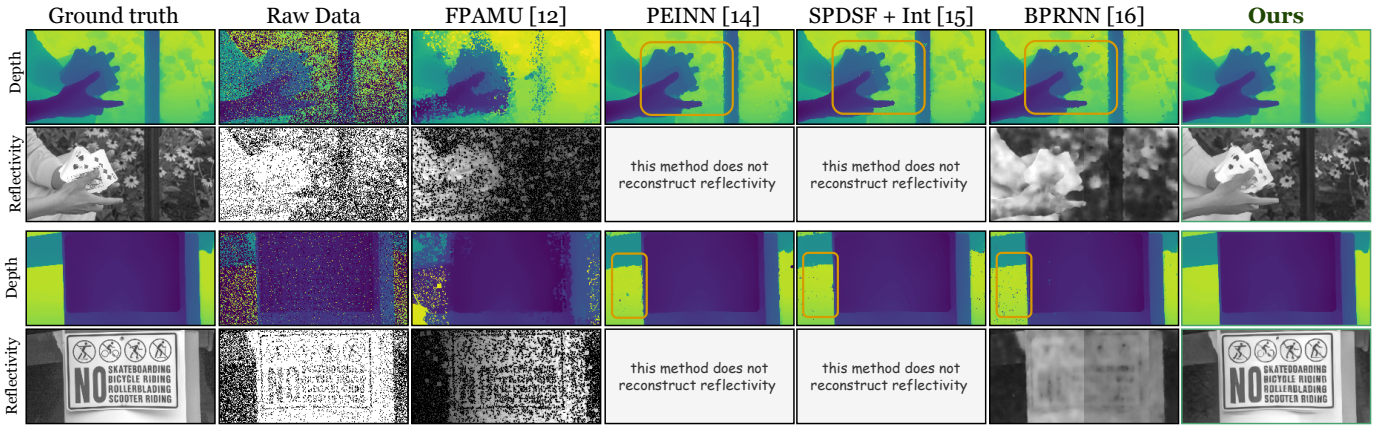


Fig. 10: **Comparison of Depth and Reflectivity Estimation on Simulated Data.** SP-LiDeR outperforms baseline methods in both reflectivity and depth estimation. The proposed method's robustness to low photon levels effectively suppresses the generation of spurious estimations, a common issue in other approaches (see artifacts in the highlighted regions). SP-LiDeR yields reflectivity reconstructions with substantially improved detail. *Best viewed in zoom.*

these experiments, all referenced deep learning methods are retrained from scratch using our timestamp simulation scheme.

1) Comparison with other SP-LiDAR Algorithms: We compare our estimation results quantitatively and qualitatively with other SP-LiDAR reconstruction algorithms. Specifically, we present benchmarking results for five existing algorithms: FPAMU [12], PEINN [14], SPDSF [15], and BPRNN [16]. Among these, FPAMU is a non-deep learning algorithm, and BPRNN is the only method that simultaneously estimates both depth and reflectivity. All methods require a 3D histogram cube (or a cluster of detections per pixel) as the input.

To evaluate existing baseline methods, we use a 3D histogram cube containing timestamp data collected during one acquisition period. The results are presented in Fig. 10 and Tab. II. We observe that existing deep learning algorithms produce depth maps with spurious estimations (see artifacts in the highlighted regions in Fig. 10), where the depth estimates at certain pixels are erroneous due to low photon levels—particularly when the level falls below 0.75 PPP per frame—resulting in less accurate depth reconstruction. However, our method is able to recover the depth maps without any spurious depth estimations. Notably, our SP-LiDeR surpasses the most recent SP-LiDAR deep learning method, BPRNN, in terms of both reconstruction quality and similarity metrics.

Since SP-LiDeR processes multiple frames, one might wonder how other baseline methods perform when provided with a 3D histogram cube constructed from timestamp measurements across several acquisition periods. While the increased average number of detections per pixel helps reduce spurious depth estimations, longer acquisition interval introduces blurry reconstruction results due to scene movement, as illustrated in Fig. 11.

2) Comparison with other Video Reconstruction Algorithms: We evaluate the reconstruction results of state-of-the-art video denoising/deblurring algorithms, as they can be utilized for reflectivity recovery from noisy timestamp data. For benchmarking, we select MemDeblur [69], Spk2ImgNet [70], RVRT [67], FloRNN [68], and QUIVER [42].

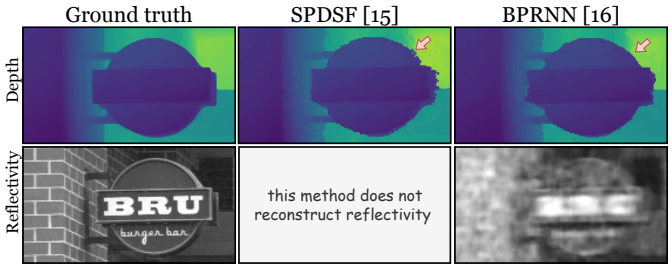


Fig. 11: **The Effect of Longer Acquisition Time for Existing Methods**. When multiple timestamp frames are compiled to construct a 3D cube, conventional deep learning methods yield blurry reconstruction results. *Best viewed in zoom.*

TABLE II: Quantitative evaluation of the proposed method against other SP-LiDAR and video reconstruction methods shows that our method achieves superior results. For methods that do not reconstruct depth or reflectivity, the corresponding entries are denoted as '-'.

Method	Reflectivity		Depth
	PSNR↑	SSIM↑	RMSE↓
SP-LiDAR Reconstruction Algorithms			
FPAMU (TCI 17) [12]	9.2556	0.0685	0.0346
PEINN (ECCV 20) [14]	-	-	0.0121
SPDSF w/ Int (SIGGRAPH 18) [15]	-	-	0.0136
SPDSF w/o Int (SIGGRAPH 18) [15]	-	-	0.0141
BPRNN (TPAMI 23) [16]	15.6759	0.4171	0.0147
Video Reconstruction Algorithms			
QUIVER (ECCV 24) [42]	22.1224	0.6698	-
RVRT (NeurIPS 22) [67]	21.8685	0.5445	-
FloRNN (ECCV 22) [68]	20.1131	0.5675	-
MemDeblur (CVPR 22) [69]	19.8106	0.4766	-
Spk2ImgNet (CVPR 21) [70]	20.1490	0.5722	-
SP-LiDeR	**23.2452**	**0.6916**	**0.0077**

The results in Fig. 12 demonstrate the performance of the different reconstruction methods. It is evident that the proposed method yields superior reflectivity reconstructions, capturing finer and sharper details that other methods strug-

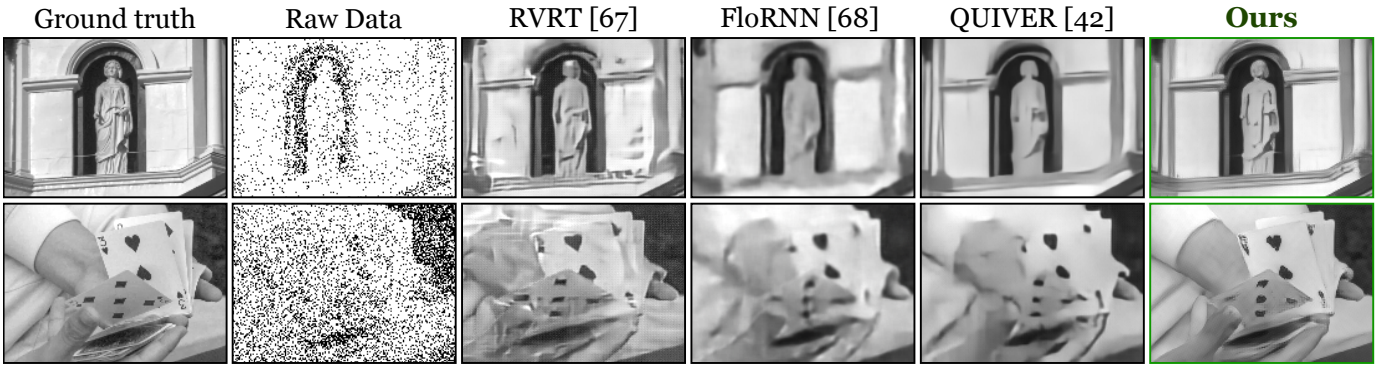


Fig. 12: **Reflectivity Estimation on Simulated Data.** Visual comparisons with existing video reconstruction algorithms demonstrate better reflectivity from our method, revealing finer details. To ensure a fair comparison, all methods receive 11 timestamp frames as input.

gle to reproduce. Quantitative results—specifically the PSNR and SSIM values presented in Tab. II—corroborate the observed improvements in reconstruction quality and similarity to the ground truth. Interestingly, the proposed method outperforms the state-of-the-art quanta video restoration algorithm, QUIVER, by more than 1 dB in reconstruction quality.

C. Real Data Experiments

We capture real timestamp frames using the SPAD sensor detailed in [63], which provides a spatial resolution of 128×192 and a frame rate of 1000 frames per second. A Picoquant LDH series 670 nm picosecond pulsed laser, with 1 nJ pulse energy, is used to achieve uniform illumination of indoor dynamic scenes having a white background. The laser operates at 25 MHz, producing pulses with an effective width of approximately 1 ns. The system's time-to-digital converter resolution is approximately 35 ps.

As shown in 13, we compare reconstruction results for two scenarios: a rotating fan and moving hands. Existing 3D reconstruction techniques, when they are given a histogram cube from a single acquisition cycle, produce depth maps affected by spurious estimations due to low photon levels. In contrast, our proposed method generates precise depth reconstructions. Regarding reflectivity results, existing methods suffer from a loss of reflectivity details. While BPRNN exhibits pronounced smoothing, the other video reconstruction algorithms also struggle to capture sharp details, notably failing to accurately reconstruct the lower hand in the moving-hand scenario. In contrast, our method consistently delivers enhanced reflectivity reconstructions.

D. Ablation

In our ablation studies, we prioritize examining the effects of the feature-sharing mechanism (CCAM), as it represents the key module, along with the optical flow feature alignment module. To ensure a fair comparison, we increased the dimensions of the feature sets in the initial feature extraction

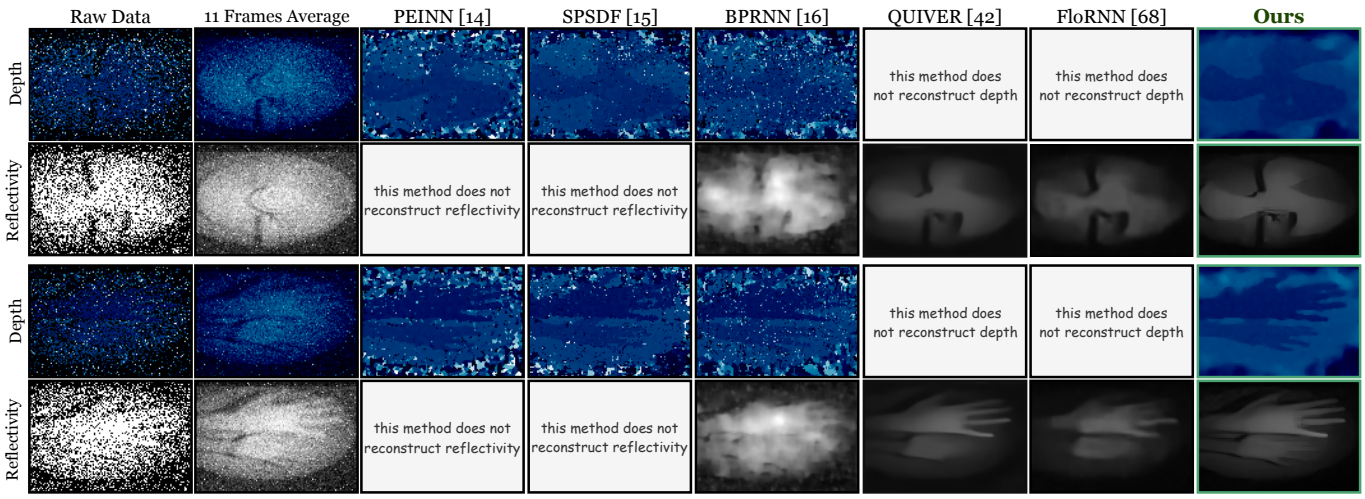


Fig. 13: **Reconstructed Results for Real Data.** Due to the low photon count, existing 3D reconstruction algorithms often produce depth maps with spurious values—a challenge addressed by our proposed method. Additionally, our approach recovers reflectivity with finer detail compared to existing methods.

stage to minimize the impact of varying numbers of trainable parameters across different variations.

Based on the quantitative comparison presented in Tab. III, the performance of both depth and reflectivity estimation improves when both modules are included. A comparison of the qualitative results with and without CCAM in Fig. 14 highlights that the feature-sharing module exchanges depth feature sets that complement reflectivity features, leading to improved results. However, under these SBR levels, the effect of the feature-sharing mechanism on depth accuracy is marginal, merely a 2.5% increase. This observation aligns with findings from the SPDSF method [14], which shows an average gain of 3.5% from intensity fusion when SBR is above one in low-photon scenarios. Our own comparison in Tab. II further supports this, indicating just a 3.67% accuracy increase with ground-truth reflectivity fusion.

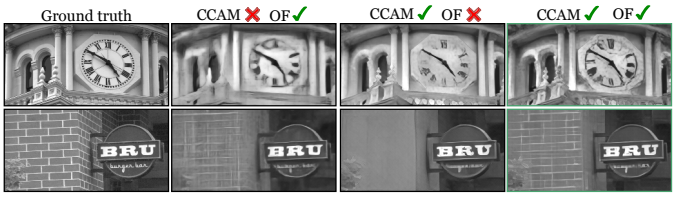


Fig. 14: **Ablation study results.** Visual comparisons of reflectivity estimation illustrate the effectiveness of the CCAM and Optical Flow modules. *Best viewed with zoom.*

TABLE III: Quantitative results of experiments to emphasize the role of optical flow alignment, and feature sharing modules.

Optical-Flow:OF	CCAM	Reflectivity		Depth
		PSNR↑	SSIM↑	RMSE↓
✗	✗	21.6954	0.6610	0.0092
✗	✓	22.2360	0.6715	0.0084
✓	✗	21.9869	0.6678	0.0079
✓	✓	23.2452	0.6916	0.0077

VI. Conclusion

We propose SP-LiDeR, an end-to-end deep learning framework for the simultaneous reconstruction of depth and reflectivity. By directly processing individual timestamp frames, our method overcomes the limitations of traditional techniques in reconstructing 3D scenes under dynamic conditions. Unlike independent or single-modality estimation methods, SP-LiDeR introduces a novel feature-sharing mechanism that enables seamless cross-modal information exchange. We also provide theoretical justification and experimental evidence to demonstrate the complementary relationship between depth and reflectivity. Experiments conducted on both synthetic and real-world datasets demonstrate the superior performance of our approach in producing accurate depth and reflectivity results. An ablation study further highlights the critical role of the feature-sharing module in enhancing reconstruction quality through cross-modal interaction. Overall, SP-LiDeR provides an effective solution for real-world SP-LiDAR applications, significantly advancing the accuracy of joint depth and reflectivity reconstruction.

VII. Appendix

In this section, we present detailed derivations of the theorems.

A. Proof of Theorem 1

The goal is to prove that the joint density of the number of photon detections M and the relative timestamps $\mathbf{t}_M = \{t_k\}_{k=1}^{M} (0 \le t_k < t_r)$ during $[0, N_r t_r)$ is

$$p\left[\mathbf{t}_M, M = m\right] = \frac{e^{-N_r \Lambda(\alpha)}}{m!} \prod_{k=1}^{m} N_r \lambda(t_k; \alpha, \tau),$$

when $M \ge 1$.

Method 1:

As mentioned in Sec. III-B, M is a Poisson random variable with a mean $N_r \Lambda(\alpha)$. Thus, its probability mass function (PMF) is

$$p_M(m; \alpha) = \frac{e^{-N_r \Lambda(\alpha)} \left[N_r \Lambda(\alpha)\right]^m}{m!}, \ m = 0, 1, \ldots \tag{20}$$

Conditioned on $M = m$, the m relative timestamps $\mathbf{t}_m = \{t_k\}_{k=1}^{m}, t_k \in [0, t_r)$ are independent and identically distributed according to the normalized photon arrival flux function as in Eq. (1), i.e.

$$f_{\tilde{t}_k | M=m}(t_k | M = m) = \frac{\lambda(t_k; \alpha, \tau)}{\Lambda(\alpha)}, \ t_k \in [0, t_r). \tag{21}$$

Therefore, when $M \ge 1$, the joint density is

$$\begin{aligned} p\left[\mathbf{t}_M, M = m\right] &= p_M(m) \prod_{k=1}^{m} f_{\tilde{t}_k | M=m}(t_k | M = m) \\ &= \frac{e^{-N_r \Lambda} (N_r \Lambda)^m}{m!} \prod_{k=1}^{m} \frac{\lambda(t_k)}{\Lambda} \\ &= \frac{e^{-N_r \Lambda}}{m!} \prod_{k=1}^{m} N_r \lambda(t_k). \end{aligned}$$

Additionally, when $M = 0$, the joint density reduces to the marginal PMF, i.e. $p\left[\mathbf{t}_M, M = 0\right] = p_M(0) = e^{-N_r \Lambda}$.

Method 2:

From the prior literature [31], [32], if we assume $\mathbf{t}_M = \{t_j\}_{j=1}^{M}$ such that $0 \le t_1 < t_2 < \ldots < t_M < t_r$, then for $M \ge 1$,

$$p\left[\mathbf{t}_M, M = m\right] = e^{-N_r \Lambda(\alpha)} \prod_{j=1}^{m} N_r \lambda(t_j).$$

We identify that the statistics are ordered here. If we transform the density of ordered statistics to unordered ones as we need, the density should shrink by $m!$ due to the permutation. The desired result will then emerge.

B. Proof of Theorem 2

The goal is to prove

$$\left[N_r\eta^2\int_0^{t_r}\frac{s^2(t-\tau)}{\eta\alpha s(t-\tau)+b_\lambda}\,dt\right]^{-1}\leq\frac{\eta S\alpha+B}{N_r\eta^2S^2},$$

where the equality holds if and only if $b_\lambda=0$. Note that

$$s(t)=S\cdot\frac{1}{\sqrt{2\pi\sigma_t^2}}e^{-\frac{t^2}{2\sigma_t^2}}=S\cdot g(t),$$

where $g(t)$ is defined as the Gaussian probability density function (PDF) and we assume $g(t-\tau)$ is fully supported on the interval $[0,t_r]$, i.e. $\int_0^{t_r}g(t-\tau)\,dt=1$. Thus, the left-hand side (LHS) becomes

$$\text{LHS}=\frac{1}{N_r\eta^2S^2}\left[\int_0^{t_r}\frac{g^2(t-\tau)}{\eta S\alpha g(t-\tau)+b_\lambda}\,dt\right]^{-1}.$$

Then, it is equivalent to prove

$$\left[\int_0^{t_r}\frac{g^2(t-\tau)}{\eta S\alpha g(t-\tau)+b_\lambda}\,dt\right]^{-1}\leq\eta S\alpha+B.$$

Starting from $\int_0^{t_r}g(t-\tau)\,dt=1$,

$$\begin{aligned}
1&\overset{(i)}{=}\int_0^{t_r}g(t-\tau)\,dt=\left[\int_0^{t_r}g(t-\tau)\,dt\right]^2\\
&\overset{(ii)}{=}\left[\int_0^{t_r}\underbrace{\frac{g(t-\tau)}{\sqrt{\eta S\alpha g(t-\tau)+b_\lambda}}}_{f(t)}\underbrace{\sqrt{\eta S\alpha g(t-\tau)+b_\lambda}}_{h(t)}\,dt\right]^2\\
&\overset{(iii)}{\leq}\int_0^{t_r}\frac{g^2(t-\tau)}{\eta S\alpha g(t-\tau)+b_\lambda}\,dt\int_0^{t_r}\left[\eta S\alpha g(t-\tau)+b_\lambda\right]\,dt\\
&\overset{(iv)}{=}\int_0^{t_r}\frac{g^2(t-\tau)}{\eta S\alpha g(t-\tau)+b_\lambda}\,dt\cdot(\eta S\alpha+B),
\end{aligned}$$

where the Cauchy-Schwarz inequality is applied from line (ii) to (iii). In the third (iii), equality holds if and only if $h(t)=k\cdot f(t)$ for $k\neq0$, or equivalently $b_\lambda=0$.

Rearrange the terms at the last line, we obtain

$$\left[\int_0^{t_r}\frac{g^2(t-\tau)}{\eta S\alpha g(t-\tau)+b_\lambda}\,dt\right]^{-1}\leq\eta S\alpha+B,$$

and the proof is complete.

REFERENCES

[1] A. Boretti, "A Perspective on Single-Photon LiDAR Systems," *Microwave and Optical Technology Letters*, vol. 66, no. 1, p. e33918, 2024.

[2] K. Morimoto, A. Ardelean, M.-L. Wu, A. C. Ulku, I. M. Antolovic, C. Bruschini, and E. Charbon, "Megapixel Time-gated SPAD Image Sensor for 2D and 3D Imaging Applications," *Optica*, vol. 7, no. 4, pp. 346–354, 2020.

[3] Z. Li, H. Pan, G. Shen, D. Zhai, W. Zhang, L. Yang, and G. Wu, "Single-photon LiDAR for Canopy Detection with a Multi-Channel Si SPAD at 1064nm," *Optics & Laser Technology*, vol. 157, p. 108749, 2023.

[4] J. Rapp, J. Tachella, Y. Altmann, S. McLaughlin, and V. K. Goyal, "Advances in Single-Photon Lidar for Autonomous Vehicles: Working Principles, Challenges, and Recent Advances," *IEEE Signal Processing Magazine*, vol. 37, no. 4, pp. 62–71, 2020.

[5] A. McCarthy, N. J. Krichel, N. R. Gemmell, X. Ren, M. G. Tanner, S. N. Dorenbos, V. Zwiller, R. H. Hadfield, and G. S. Buller, "Kilometer-range, High Resolution Depth Imaging via 1560 nm Wavelength Single-Photon Detection," *Opt. Express*, vol. 21, no. 7, pp. 8904–8915, 2013.

[6] G. Mora-Martín, A. Turpin, A. Ruget, A. Halimi, R. Henderson, J. Leach, and I. Gyongy, "High-speed Object Detection with a Single-photon Time-of-Flight image sensor," *Opt. Express*, vol. 29, no. 21, pp. 33 184–33 196, 2021.

[7] Y. Li and J. Ibanez-Guzman, "LiDAR for Autonomous Driving: The Principles, Challenges, and Trends for Automotive LiDAR and Perception Systems," *IEEE Signal Processing Magazine*, vol. 37, no. 4, pp. 50–61, 2020.

[8] W. Becker, *Advanced Time-Correlated Single Photon Counting Techniques*, 2nd ed. Springer, 2005.

[9] E. Charbon, M. Fishburn, R. Walker, R. K. Henderson, and C. Niclass, *SPAD-Based Sensors in TOF Range-Imaging Cameras*. Springer, 2013, pp. 11–38.

[10] P. Padmanabhan, C. Zhang, and E. Charbon, "Modeling and Analysis of a Direct Time-of-Flight Sensor Architecture for LiDAR Applications," *MDPI, Sensors*, vol. 19, no. 24, p. 5464, 2019.

[11] T. Neumann and F. Kallage, "Simulation of a Direct Time-of-Flight LiDAR-System," *IEEE Sensors Journal*, vol. 23, no. 13, pp. 14 245–14 252, 2023.

[12] J. Rapp and V. K. Goyal, "A Few Photons Among Many: Unmixing Signal and Noise for Photon-Efficient Active Imaging," *IEEE Transactions on Computational Imaging (TCI)*, vol. 3, no. 3, pp. 445–459, 2017.

[13] D. Shin, A. Kirmani, V. K. Goyal, and J. H. Shapiro, "Photon-Efficient Computational 3-D and Reflectivity Imaging With Single-Photon Detectors," *IEEE Transactions on Computational Imaging (TCI)*, vol. 1, no. 2, pp. 112–125, 2015.

[14] J. Peng, Z. Xiong, X. Huang, Z.-P. Li, D. Liu, and F. Xu, "Photon-Efficient 3D Imaging with A Non-local Neural Network," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 225–241.

[15] D. B. Lindell, M. O'Toole, and G. Wetzstein, "Single-Photon 3D Imaging with Deep Sensor Fusion," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 113.1–113.12, 2018.

[16] J. Peng, Z. Xiong, H. Tan, X. Huang, Z.-P. Li, and F. Xu, "Boosting Photon-Efficient Image Reconstruction With A Unified Deep Neural Network," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 45, no. 4, pp. 4180–4197, 2023.

[17] A. Maccarone, K. Drummond, A. McCarthy, U. K. Steinlehner, J. Tachella, D. A. Garcia, A. Pawlikowska, R. A. Lamb, R. K. Henderson, S. McLaughlin, Y. Altmann, and G. S. Buller, "Submerged Single-Photon LiDAR Imaging Sensor Used for Real-time 3D Scene Reconstruction in Scattering Underwater Environments," *Opt. Express*, vol. 31, no. 10, pp. 16 690–16 708, 2023.

[18] A. Halimi, A. Maccarone, A. McCarthy, S. McLaughlin, and G. S. Buller, "Object Depth Profile and Reflectivity Restoration From Sparse Single-Photon Data Acquired in Underwater Environments," *IEEE Transactions on Computational Imaging (TCI)*, vol. 3, no. 3, pp. 472–484, 2017.

[19] Y. Zhang, S. Li, J. Sun, X. Zhang, D. Liu, X. Zhou, H. Li, and Y. Hou, "Three-dimensional Single-Photon Imaging through Realistic Fog in an Outdoor Environment During the Day," *Opt. Express*, vol. 30, no. 19, pp. 34 497–34 509, 2022.

[20] D. Shin, F. Xu, F. N. C. Wong, J. H. Shapiro, and V. K. Goyal, "Computational Multi-depth Single-Photon Imaging," *Opt. Express*, vol. 24, no. 3, pp. 1873–1888, 2016.

[21] Y. Altmann, S. McLaughlin, and M. E. Davies, "Fast Online 3D Reconstruction of Dynamic Scenes From Individual Single-Photon Detection Events," *IEEE Transactions on Image Processing (TIP)*, vol. 29, pp. 2666–2675, 2020.

[22] A. Kirmani, D. Venkatraman, D. Shin, A. Colaço, F. N. C. Wong, J. H. Shapiro, and V. K. Goyal, "First-Photon Imaging," *Science*, vol. 343, no. 6166, pp. 58–61, 2014.

[23] Y. Altmann, X. Ren, A. McCarthy, G. S. Buller, and S. McLaughlin, "LiDAR Waveform-Based Analysis of Depth Images Constructed Using Sparse Single-Photon Data," *IEEE Transactions on Image Processing (TIP)*, vol. 25, no. 5, pp. 1935–1946, 2016.

[24] J. Tachella, Y. Altmann, X. Ren, A. McCarthy, G. S. Buller, S. McLaughlin, and J.-Y. Tourneret, "Bayesian 3D Reconstruction of Complex Scenes from Single-Photon LiDAR Data," *SIAM Journal on Imaging Sciences*, vol. 12, no. 1, pp. 521–550, 2019.

[25] Y. Altmann, X. Ren, A. McCarthy, G. S. Buller, and S. McLaughlin, "Robust Bayesian Target Detection Algorithm for Depth Imaging From Sparse Single-Photon Data," *IEEE Transactions on Computational Imaging (TCI)*, vol. 2, no. 4, pp. 456–467, 2016.

[26] J. Lee, A. Ingle, J. V. Chacko, K. W. Eliceiri, and M. Gupta, "CASPI: Collaborative Photon Processing for Active Single-Photon Imaging," *Nature Communications*, vol. 14, no. 1, p. 3158, 2023.

[27] J. Rapp, Y. Ma, R. M. A. Dawson, and V. K. Goyal, "Dead Time Compensation for High-Flux Ranging," *IEEE Transactions on Signal Processing*, vol. 67, no. 13, pp. 3471–3486, 2019.

[28] A. K. Pediredla, A. C. Sankaranarayanan, M. Buttafava, A. Tosi, and A. Veeraraghavan, "Signal Processing Based Pile-up Compensation for Gated Single-Photon Avalanche Diodes," *arXiv preprint arXiv: 1806.07437*, 2018.

[29] A. Gupta, A. Ingle, A. Velten, and M. Gupta, "Photon-Flooded Single-Photon 3D Cameras," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6763–6772.

[30] W. Zhang, H. K. Weerasooriya, P. Chennuri, and S. H. Chan, "Parametric Modeling and Estimation of Photon Registrations for 3D Imaging," in *IEEE 26th International Workshop on Multimedia Signal Processing (MMSP)*, 2024, pp. 1–6.

[31] I. Bar-David, "Communication under the Poisson regime," *IEEE Transactions on Information Theory*, vol. 15, no. 1, pp. 31–37, 1969.

[32] S. H. Chan, H. K. Weerasooriya, W. Zhang, P. Abshire, I. Gyongy, and R. K. Henderson, "Resolution Limit of Single-Photon LiDAR," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 25 307–25 316.

[33] W. C. Yau, W. Zhang, H. K. Weerasooriya, and S. H. Chan, "Analysis and Improvement of Rank-Ordered Mean Algorithm in Single-Photon LiDAR," in *IEEE 26th International Workshop on Multimedia Signal Processing (MMSP)*, 2024, pp. 1–6.

[34] G. Yao, Y. Chen, C. Jiang, Y. Xuan, X. Hu, Y. Liu, and Y. Pan, "Dynamic Single-Photon 3D Imaging with a Sparsity-Based Neural Network," *Opt. Express*, vol. 30, no. 21, pp. 37 323–37 340, 2022.

[35] X. Zhao, X. Jiang, A. Han, T. Mao, W. He, and Q. Chen, "Photon-efficient 3D Reconstruction Employing a Edge Enhancement Method," *Opt. Express*, vol. 30, no. 2, pp. 1555–1569, 2022.

[36] Z. Zang, D. Xiao, and D. D.-U. Li, "Non-fusion Time-resolved Depth Image Reconstruction using a Highly Efficient Neural Network Architecture," *Opt. Express*, vol. 29, no. 13, pp. 19 278–19 291, 2021.

[37] A. Ruget, S. McLaughlin, R. K. Henderson, I. Gyongy, A. Halimi, and J. Leach, "Robust Super-Resolution Depth Imaging via a Multi-feature Fusion Deep Network," *Opt. Express*, vol. 29, no. 8, pp. 11 917–11 937, 2021.

[38] Z. Sun, D. B. Lindell, O. Solgaard, and G. Wetzstein, "SPADnet: Deep RGB-SPAD Sensor Fusion Assisted by Monocular Depth Estimation," *Opt. Express*, vol. 28, no. 10, pp. 14 948–14 962, 2020.

[39] Y. Altmann, R. Aspden, M. Padgett, and S. McLaughlin, "A Bayesian Approach to Denoising of Single-Photon Binary Images," *IEEE Transactions on Computational Imaging (TCI)*, vol. 3, no. 3, pp. 460–471, 2017.

[40] E. R. Fossum, J. Ma, S. Masoodian, L. Anzagira, and R. Zizza, "The Quanta Image Sensor: Every Photon Counts," *MDPI, Sensors*, vol. 16, no. 8, p. 1260, 2016.

[41] J. Ma, S. Chan, and E. R. Fossum, "Review of Quanta Image Sensors for Ultralow-Light Imaging," *IEEE Transactions on Electron Devices*, vol. 69, no. 6, pp. 2824–2839, 2022.

[42] P. Chennuri, Y. Chi, E. Jiang, G. M. D. Godaliyadda, A. Gnanasambandam, H. R. Sheikh, I. Gyongy, and S. H. Chan, "Quanta Video Restoration," in *European Conference on Computer Vision (ECCV)*, 2024, pp. 152–171.

[43] S. Ma, S. Gupta, A. C. Ulku, C. Brushini, E. Charbon, and M. Gupta, "Quanta Burst Photography," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 79:1 – 79:16, 2020.

[44] Y. Liu, A. Krull, H. Basevi, A. Leonardis, and M. Jenkins, "Bit2Bit: 1-bit Quanta Video Reconstruction via Self-Supervised Photon Prediction," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024, pp. 88 443–88 485.

[45] Y. Chi, A. Gnanasambandam, V. Koltun, and S. H. Chan, "Dynamic Low-Light Imaging with Quanta Image Sensors," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 122–138.

[46] S. H. Chan, O. A. Elgendy, and X. Wang, "Images from Bits: Non-iterative Image Reconstruction for Quanta Image Sensors," *MDPI, Sensors*, vol. 16, no. 11, p. 1961, 2016.

[47] J. H. Choi, O. A. Elgendy, and S. H. Chan, "Image Reconstruction for Quanta Image Sensors using Deep Neural Networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 6543–6547.

[48] T. Zhang, M. Dutson, V. Boominathan, M. Gupta, and A. Veeraraghavan, "Streaming Quanta Sensors for Online, High-Performance Imaging and Vision," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 47, no. 3, pp. 1564–1577, 2025.

[49] R. Kitichotkul, J. Rapp, and V. K. Goyal, "The Role of Detection Times in Reflectivity Estimation with Single-Photon LiDAR," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 30, no. 1, pp. 1–14, 2024.

[50] K. Drummond, S. McLaughlin, Y. Altmann, A. Pawlikowska, and R. Lamb, "Joint Surface Detection and Depth Estimation from Single-Photon LiDAR Data using Ensemble Estimators," in *Sensor Signal Processing for Defence Conference (SSPD)*, 2021, pp. 1–5.

[51] A. Halimi, Y. Altmann, A. McCarthy, X. Ren, R. Tobin, G. S. Buller, and S. McLaughlin, "Restoration of Intensity and Depth Images Constructed using Sparse Single-Photon Data," in *European Signal Processing Conference (EUSIPCO)*, 2016, pp. 86–90.

[52] I. Gyongy, S. W. Hutchings, A. Halimi, M. Tyler, S. Chan, F. Zhu, S. McLaughlin, R. K. Henderson, and J. Leach, "High-speed 3D Sensing via Hybrid-mode Imaging and Guided Upsampling," *Optica*, no. 10, pp. 1253–1260, 2020.

[53] D. L. Snyder and M. I. Miller, *Random Point Processes in Time and Space*. Springer, 1991.

[54] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor Segmentation and Support Inference from RGBD Images," in *European Conference on Computer Vision (ECCV)*, 2012, pp. 746–760.

[55] Y. Zhang and T. Funkhouser, "Deep Depth Completion of a Single RGB-D Image," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 175–185.

[56] Y. Zhang, X. Guo, M. Poggi, Z. Zhu, G. Huang, and S. Mattoccia, "Completionformer: Depth Completion with Convolutions and Vision Transformers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 18 527–18 536.

[57] S. Marschner and P. Shirley, *Fundamentals of Computer Graphics*, 4th ed. A. K. Peters, Ltd., 2016.

[58] M. P. do Carmo, *Differential Geometry of Curves and Surfaces*, 2nd ed. Dover Publications, 2016.

[59] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional Block Attention Module," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[60] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.

[61] K. Zhou, W. Li, L. Lu, X. Han, and J. Lu, "Revisiting Temporal Alignment for Video Restoration," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 6053–6062.

[62] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "MaxViT: Multi-axis Vision Transformer," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 459–479.

[63] R. K. Henderson, N. Johnston, F. Mattioli Della Rocca, H. Chen, D. Day-Uei Li, G. Hungerford, R. Hirsch, D. Mcloskey, P. Yip, and D. J. S. Birch, "A 192×128 Time Correlated SPAD Image Sensor in 40-nm CMOS Technology," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 7, pp. 1907–1916, 2019.

[64] S. Scholes, G. Mora-Martin, F. Zhu, I. Gyongy, P. Soan, and J. Leach, "Fundamental Limits to Depth Imaging with Single-photon Detector Array Sensors," *Nature Scientific Reports*, vol. 13, no. 1, p. 176, 2023.

[65] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth Anything V2," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2024, pp. 21 875–21 911.

[66] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *International Conference on Learning Representations, (ICLR)*, 2015.

[67] J. Liang, Y. Fan, X. Xiang, R. Ranjan, E. Ilg, S. Green, J. Cao, K. Zhang, R. Timofte, and L. V. Gool, "Recurrent Video Restoration Transformer with Guided Deformable Attention," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022, pp. 378–393.

[68] J. Li, X. Wu, Z. Niu, and W. Zuo, "Unidirectional Video Denoising by Mimicking Backward Recurrent Modules with Look-Ahead Forward Ones," in *European Conference on Computer Vision (ECCV)*, 2022, pp. 592–609.

[69] B. Ji and A. Yao, "Multi-Scale Memory-Based Video Deblurring," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 1919–1928.

[70] J. Zhao, R. Xiong, H. Liu, J. Zhang, and T. Huang, "Spk2ImgNet: Learning to Reconstruct Dynamic Scene from Continuous Spike Stream," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 11 991–12 000.