Joint Depth and Reflectivity Estimation using Single Photon LiDAR

Hashan K. Weerasooriya[†1] Prateek Chennuri[*1] Weijian Zhang[*1] Istvan Gyongy[2]
Stanley H. Chan[1] [1] Purdue University, [2] University of Edinburgh

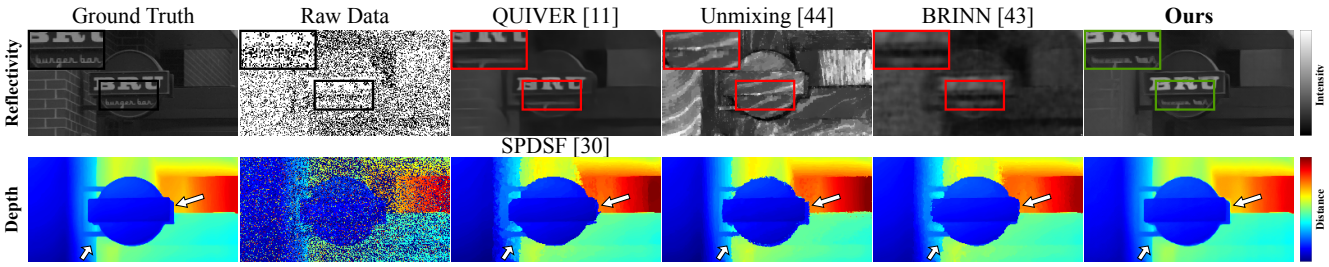


Figure 1. **Objective of this paper.** We propose SP-LiDeR, a new method for joint estimation of depth and reflectivity from single-photon LiDAR data in dynamic, high-noise scenarios. Unlike previous approaches that either focus on a single modality (e.g., QUIVER [11] for reflectivity, SPDSF [30] for depth) or attempt joint estimation with limited success (e.g., Unmixing [44], BRINN [43]), SP-LiDeR leverages the correlation between depth and reflectivity to achieve sharper and more accurate reconstructions for both modalities.

Abstract

Single-photon light detection and ranging (SPL) is becoming the predominant technology for long-range high-precision 3D vision tasks. In SPL, the timestamps carry two pieces of complementary information: The time it takes for the pulse to travel (i.e., the depth), and the number of photons reflected by the object (i.e., the reflectivity). Existing SPL image reconstruction methods often focus on recovering either the depth or reflectivity. However, for dynamic scenes involving fast-moving objects, it is more desirable to directly process the timestamps and use one signal to help the other. In this paper, we present an estimation method to ***simultaneously*** *recover the depth and reflectivity, for fast-moving scenes. Our contributions are: (1) We theoretically show that depth and reflectivity are mutually correlated and conditions under which the joint estimation would become useful. (2) Based on the theory, we develop a new reconstruction method, SP-LiDeR, to reconstruct the signal by leveraging the shared information. On synthetic and real SPL data, our method demonstrates superior image reconstruction results compared to existing methods.*

1. Introduction

With the rapid growth of 3D applications from autonomous driving to virtual reality, single-photon light detection and ranging is becoming one of the most important technologies for range-related products [7, 26–28, 35, 36, 46]. The sensing principle of modern SPL is to send a laser pulse train to the object and use a single-photon avalanche diode (SPAD) [6, 10] with a time-correlated single-photon counting module (TCSPC) [6] to record the time it takes to travel [38, 39].

The measured timestamps of SPL contain two pieces of information: (1) **Depth**: the time it takes for the pulse to travel, which tells us how far the object is; (2) **Reflectivity**: the number of photons reflected by the object, e.g., a shiny object reflects photons while a dull object absorbs them. Since both depth and reflectivity are generated from the same timestamp stream, simultaneously recovering them is feasible, at least in theory.

However, joint estimation of depth and reflectivity is not very common in the literature. When the signal-to-background (SBR) is sufficiently high, empirically people observe that the individual estimations of depth and reflectivity are often good enough. Even if the SBR is not too high and if the scene is static, one can aggregate multiple timestamps to construct a 3D histogram cube using a principle similar to temporal averaging. The problem comes when the object is moving so the number of photons per pixel (PPP) per second needs to be strictly limited to avoid motion blur. In this case, independently estimating depth and reflectivity will be sub-optimal. See Fig. 1.

The goal of this paper is to explore the benefit of simultaneous estimation of depth and reflectivity. To this end, we aim to answer two questions:

1. *Does recovering depth help recover reflectivity, and vice versa?* To answer this question, we theoretically derive the maximum-likelihood estimator (MLE) of the joint reflectivity and depth recovery problem. By analyzing the Cramér–Rao lower bound (CRLB), we identify conditions under which the information sharing is effective. This provides a foundation for joint estimation.
2. *Can we build an efficient neural network to jointly recover depth and reflectivity, for fast-moving scenes?* Building upon our theoretical findings, we develop a dual-channel joint estimation network (SP-LiDeR) integrating quanta video restoration, motion field estimation, and feature sharing. We conduct extensive comparisons with existing methods on both synthetic and real SPAD data to demonstrate the effectiveness of SP-LiDeR.

Most results of this paper are evaluated on videos. For better visualization, we invite readers to check our videos in the supplementary material.

2. Prior Work

Depth Estimation. Assuming that the scene is static, depth reconstruction can generally be done by maximum-likelihood and its variants [1, 2, 21, 23, 50]. Under a photon-limited regime where the average number of PPP is $\mathcal{O}(1)$, dead time becomes negligible [17, 41, 45, 59] and so the depth estimation is simplified to a matched filter matched to the photon arrival flux function [5, 9]. The matched filter reduces to the timestamp sample mean when the SBR is sufficiently high. When SBR is low, however, the sample mean estimator is no longer accurate. The matched filter also does not perform well because the search is computationally heavy and it is prone to be trapped by local optima due to nonconvexity. Therefore, additional pre-processing [21, 44, 48, 57] needs to be employed to reject outliers before applying the sample mean estimation. Advanced depth recovery methods using neural networks that incorporate the noise censoring idea are also available [30, 42].

Reflectivity Estimation. Reflectivity reconstruction is typically formulated via another MLE of photon counts [2, 3, 21, 48], as it is proportional to the number of photons reflected by the object. Because of the photon-counting nature, the estimation problem is largely formulated by Poisson distribution and/or Binomial distribution for a sum of binary photon detections, similar to quanta image sensors [8, 11–13, 16, 31–33, 58]. Most previous work on reflectivity estimation is based on photon counts rather than the timestamps suggested by [22, 44]. To avoid the exhausting grid search and the knowledge of depth required in solving the timestamp MLE, an alternate approach [44] is proposed by conducting censoring first and then computing the photon count MLE. Although the result achieves the state-of-the-art (SOTA), we recognize that depth information is not utterly utilized because the censoring merely relies on reflectivity estimation.

Joint Estimation. Joint estimation of depth and reflectivity is theoretically beneficial but highly challenging due to the continuous, alternating search between the two parameters. Nevertheless, in the absence of background noise, the joint MLE is *separable* [44], inspiring the two-path algorithms that censor noise first and then solve two independent problems. However, we observe that the SOTA censoring methods depend on the reflectivity estimation. Although it is easy to show that reflectivity helps the downstream depth estimation, the reverse is not clear. To our knowledge, the only deep learning method that considers a joint feature extraction from depth and reflectivity is [43], but its decoders are designed independently. In contrast, we propose an information-sharing module where both feature extraction and recovery are shared between depth and reflectivity, fully exploiting the mutual guidance.

Dynamic Scenes. Depth and reflectivity estimation for dynamic scenes are significantly more difficult because we can no longer aggregate multiple frames without compensating for the motion. Tobin *et al*. [51] and Legros *et al*. [24] utilize temporal correlations among video frames for dynamic scenes, but they are only suitable for slow motions due to their histogram cube data collection. Altmann *et al*. [4] proposed a different data acquisition mechanism based on individual photon detections instead of histogram cubes, which is more suitable for fast-moving objects.

3. Depth and Reflectivity Information Sharing

In this section, we discuss the dependency between depth and reflectivity under the per-pixel MLE framework and demonstrate how and when they benefit each other with a toy problem. Proofs of theorems and corollaries can be found in the supplement.

3.1. Notation

The notations of this paper follow the literature, e.g., [9, 44, 48]. We assume that the SPL 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 [5, 49]

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

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. 1 provides a summary of these notations.

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
τ	true time delay	η	quant. eff.
α	true reflectivity	λ_d	dark current

Table 1. Notations in this paper.

Given $\lambda_{i,j}(\ell, t)$, the core quantity we are interested in is the distribution of the timestamps. Following prior work such as [5, 9, 22, 44], the distribution is defined according to the theorem below.

Theorem 1 (Joint density of M timestamps $\mathbf{t}_M$ [5, 22])**.** *Let* $\mathbf{t}_M = \{t_k\}_{k=1}^M$. *For* $M \geq 1$,

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

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 SBR as $SBR = \eta\alpha S/B$.

The core signal estimation problem is formulated as a constrained maximum likelihood (CML):

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

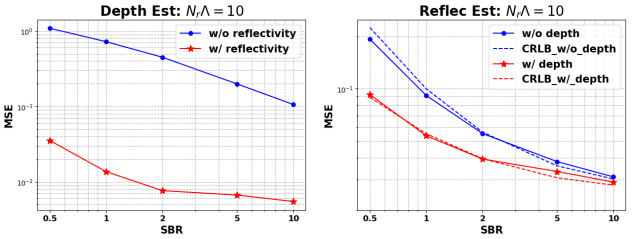
where evidently depth and reflectivity rely on each other as long as $b_\lambda > 0$. In many prior works such as [44], the joint estimation is solved via two separable problems when one eliminates skeptical outliers and then assumes *zero* background noise:

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

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

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

meaning that they can be solved separately.



(a) (b)

Figure 2. **Per-pixel information sharing under MLE.** (a). Reflectivity helps the depth estimation. (b). Depth helps the reflectivity estimation. The performance gap becomes larger when noise is more dominant.

When α and τ are separable, recovering one will *not* offer any help to recover the other. This can perhaps explain why joint estimation is not common in the literature.

3.2. Information sharing under MLE

In this subsection, we discuss how depth and reflectivity can help each other when $b_\lambda \neq 0$.

How reflectivity helps depth? This direction is straightforward because, without the reflectivity knowledge, the most we can do is run a log-matched filter assuming no noise exists, as indicated in Eq. (5). This is a simple but less accurate approximation compared to solving Eq. (3) with prior knowledge of α and a search algorithm [9].

Fig. 2a shows that the depth estimator with the knowledge of reflectivity consistently outperforms the one without, demonstrating reflectivity helps depth estimation.

How depth helps reflectivity? The reverse is non-trivial since people typically utilize photon counts to estimate reflectivity and never examine the extra information embedded in timestamps for reflectivity estimation until [22]. The former does not need depth information while the latter does. We compare the 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 reflectivity estimator that does not depend on depth

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

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*

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

✠ *Remark: The estimation becomes more certain if we have more data, higher SBR, and larger system energy (ηS) compared to the object (α).*

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. (A numerical algorithm 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 unconstrained reflectivity estimator $\widehat{\alpha}_t^*$ that relies on depth constantly surpasses $\widehat{\alpha}_c^*$ 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 supplementary materials. Theorem 2 provides a theoretical justification for why depth can help reflectivity 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. This again confirms our hypothesis that depth benefits the reflectivity estimation whenever there is noise, as predicted in Eq. (10). The reason why sometimes the experimental MSE is lower than the CRLB is that the positive constraint of α pulls negative estimations to 0, which is closer to the ground truth.

The numerical results in Fig. 2b are consistent with those from [22], although Theorem 2 and the numerical algorithm are new. More importantly, the context of our results is slightly different: In [22], a sequential estimation was developed whereas in our work, we aim for a joint estimation.

3.3. Information sharing in the feature space

While we have a mathematical explanation of how and when depth and reflectivity estimations can help each other, one can still question whether the same conclusion will help when we implement algorithms using deep neural networks.

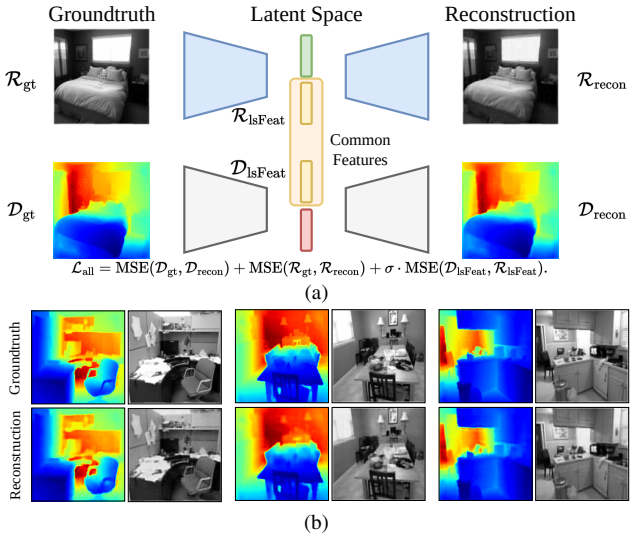


Figure 3. **Information Sharing Pilot Study.** (a). The setup involves two autoencoders attempting to reconstruct the inputs while isolating the common features in the latent space. (b). Reconstruction results for three scenes from the toy experiment verify the claims about shared features.

To this end, we conduct a toy experiment to assess how much *common features* are shared across the depth and reflectivity.

Our toy experiment is constructed as follows. We have two identical convolutional autoencoder structures [14]: one for reconstructing reflectivity and the other for reconstructing depth. In the latent space, we minimize the gap between specific latent features by setting up three loss terms: (i) Depth estimate MSE($\mathcal{D}_{\mathrm{gt}}, \mathcal{D}_{\mathrm{recon}}$), (ii) Reflectivity estimate MSE($\mathcal{R}_{\mathrm{gt}}, \mathcal{R}_{\mathrm{recon}}$), and (iii) Feature Information Sharing loss MSE($\mathcal{D}_{\mathrm{lsFeat}}, \mathcal{R}_{\mathrm{lsFeat}}$) which measures the deviations of the features in the latent space. A pictorial representation of our toy experiment is shown in Fig. 3a.

We trained this toy autoencoder network using the NYU V2 RGB-D dataset [37]. Fig. 3b displays the results of the experiment. If reflectivity and depth were *not* helping each other, we should expect to see degradation in the respective estimates. The fact that the estimates are reasonably good even if we force them to share features provides evidence that some features in the latent space are indeed common. In the supplementary materials, we visualize the low-dimensional representations of these shared features.

The feature sharing result we present here is consistent with several prior work in the literature which showed how fusing reflectivity in the feature domain enhances depth accuracy [30, 40, 55, 60, 61]. Additionally, surface properties such as curvature and normals are naturally shared in the depth and reflectivity feature space [15, 34].

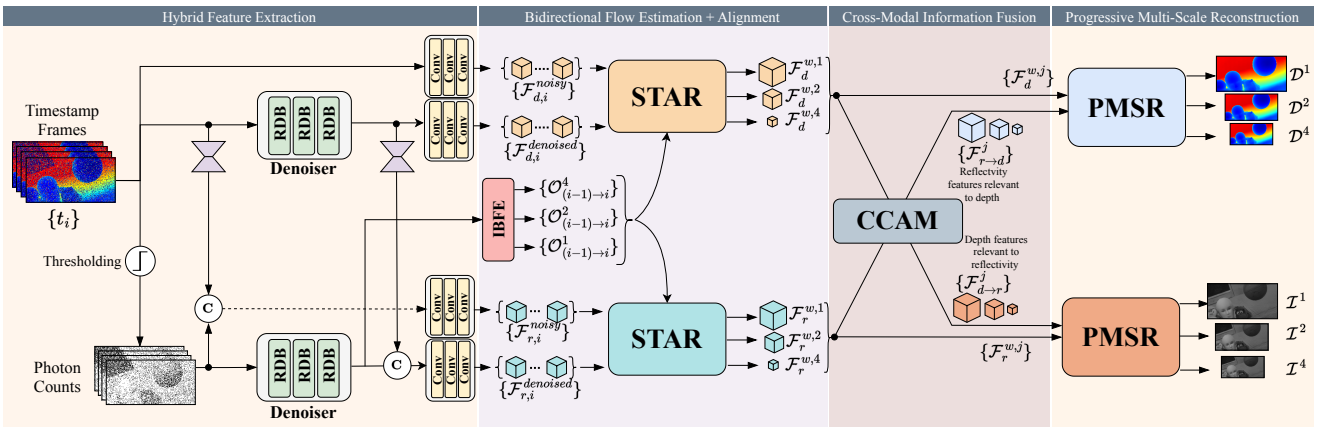


Figure 4. **The SP-LiDeR Network Architecture**. The proposed SP-LiDeR network consists of four main components. This network produces multiscale reconstructions for both depth and reflectivity simultaneously. PMSR - Progressive Multi-Scale Reconstruction, STAR - Spatio-Temporal Alignment with Residual Refinement, IBFE - Iterative Multi-Scale Bi-directional Flow Estimation, CCAM - Convolutional Cross Attention Module

4. SP-LiDeR Network

The proposed method is called **S**ingle-**P**hoton **LiDAR** joint **De**pth and **R**eflectivity estimator. SP-LiDeR is an end-to-end deep learning model for ***joint estimation*** of depth and reflectivity using SPL data. Building on the insights from prior research and the analyses we presented in the previous sections, we build a cross-information sharing mechanism to realize the synergy between depth and reflectivity. By designing to simultaneously output both modalities, SP-LiDeR achieves highly accurate, noise-robust, and blur-free results even in challenging conditions. The schematic diagram of SP-LiDeR is shown in Fig. 4. SP-LiDeR adopts two parallel branches for the two modalities (depth and reflectivity), each enhancing its own characteristics and features. The process can be divided into four main stages.

4.1. Hybrid Feature Extraction

The photon timestamp slices from SPL serve as inputs to SP-LiDeR for depth and reflectivity estimation. Due to noise from dark current, background photons, pulse width, and shot noise, captured raw data often suffers from low SNR. To address this, as shown in Fig. 4, we denoise both the timestamps and timestamp-quantized photon counts, enhancing the SNR for depth and reflectivity estimates. Although the denoised data may lack fine details such as textures and subtle motion cues, it provides a structural view of dominant scene components. By combining high-SNR structural features with the hidden fine-grained details from noisy data, we can achieve enhanced reconstruction.

A shallow feature extraction block with 2D convolutions, batch normalization, and ReLU activation extracts features from both the denoised and noisy inputs, as shown in Fig. 4. While it is straightforward to extract depth features from the denoised and noisy timestamp data, reflectivity feature extraction requires additional information due to the binarized nature of photon counts. Hence, we use noisy binary frames along with noisy timestamp data to capture fine-grained reflectivity details. Denoised depth and reflectivity data are also integrated to provide mutual structural cues, enabling SP-LiDeR's cross-modal feature sharing to enhance depth and reflectivity extraction.

4.2. Bidirectional Flow Estimation + Alignment

An accurate estimation of the scene's motion is necessary to effectively align and fuse pixel information present in the neighboring frames. Low-bit noisy sensor data poses challenges in flow estimation due to its noise vulnerability. We address this by leveraging denoised reflectivity frames as inputs for motion estimation, relying on spatio-temporal motion consistency to compensate for missing fine-grained details.

Inspired by [63], we design and propose a multi-scale, bidirectional flow estimation and refinement module, which calculates flow progressively from the coarsest to the finest scales. Given N-frame input, with the last frame as a reference, at each level l, starting from the $(N-l)^{\text{th}}$ frame f_{N-l}, an initial optical flow is estimated and refined iteratively, utilizing each previous level's flow for enhanced temporal consistency. The final result is a spatio-temporally consistent flow maps $\{\mathcal{O}^j_{(i-1)\to i}\}$ across scales.

For alignment and spatio-temporal feature extraction across frames, a custom multi-scale temporal recurrent network leverages bidirectional flow estimates to align features across past and future frames. As shown in Fig. 5, at each scale, a hidden feature map is initialized and refined using residual channel-attention U-Net, capturing detailed structural information. This refined map is progressively warped

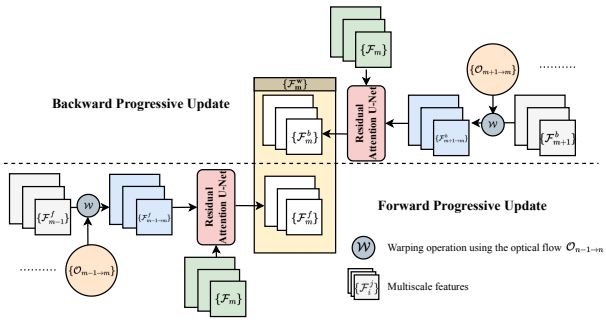


Figure 5. **Bi-directional feature extraction module** progressively propagates features from both directions to obtain the reference frame m warped features $\mathcal{F}_m^w$.

and adjusted with each subsequent frame's details, accumulating both spatial and temporal information. This process allows us to capture spatio-temporal features in both directions with respect to every input frame.

4.3. Cross-Modal Information Fusion

To enable efficient cross-information sharing (Sec. 3), we introduce the Convolutional Cross Attention Module (CCAM), inspired by CBAM [54]. CCAM facilitates multi-scale feature sharing between depth and reflectivity branches using parallelized channels and spatial cross-attention. For simplicity, we provide an overview of CCAM for a single scale. The modules of CBAM generate a channel attention map $\phi_c(\mathcal{F}^w) \in \mathbb{R}^{C\times 1}$ and a spatial attention map $\phi_s(\mathcal{F}^w) \in \mathbb{R}^{H\times W}$:

$$\phi_c(\mathcal{F}) = \sigma_s(\text{MLP}(\mathcal{F}_c^{\text{avg}} + \mathcal{F}_c^{\text{max}})) \quad (11)$$

$$\phi_s(\mathcal{F}) = \sigma_s(\text{Conv}([\mathcal{F}_s^{\text{avg}}; \mathcal{F}_s^{\text{max}}])) \quad (12)$$

where $\mathcal{F}_i^{\text{avg}}$ and $\mathcal{F}_i^{\text{max}}$ are $i \in \{c, s\}$ are the channel or spatial average and max-pooled maps of warped features $\mathcal{F}^w$, and σ_s is the sigmoid function. The 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 $\Phi_s(\mathcal{F}^w) \in \mathbb{R}^{P\times HW}$.

These channel and spatial attention maps are generated for both depth $\mathcal{F}_{\text{d}}^w$ and reflectivity $\mathcal{F}_{\text{r}}^w$ warped features independently, yielding $\Phi_s(\mathcal{F}_{\text{d}}^w)$, $\Phi_s(\mathcal{F}_{\text{r}}^w)$, $\Phi_c(\mathcal{F}_{\text{d}}^w)$, and $\Phi_c(\mathcal{F}_{\text{r}}^w)$. Cross-attention is then implemented using Multi-Head Attention (MHA) [53], enabling reflectivity-to-depth feature sharing. To simplify, we explain the feature-sharing mechanism from reflectivity to depth as follows:

$$\begin{aligned} \text{head}_i^c &= \text{Attention}(Q_c^d W_{i,c}^Q, K_c^r W_{i,c}^K, V_c^r W_{i,c}^V) \\ \text{MHA}_c^d(Q_c^d, K_c^r, V_c^r) &= \text{Concat}(\text{h}_1^c, \ldots, \text{h}_n^c) W_c^o \end{aligned} \quad (13)$$

where $Q_c^d = \Phi_c(\mathcal{F}_{\text{d}}^w)$, and $K_c^r = V_c^r = \Phi_c(\mathcal{F}_{\text{r}}^w)$. Spatial cross-attention is similarly defined.

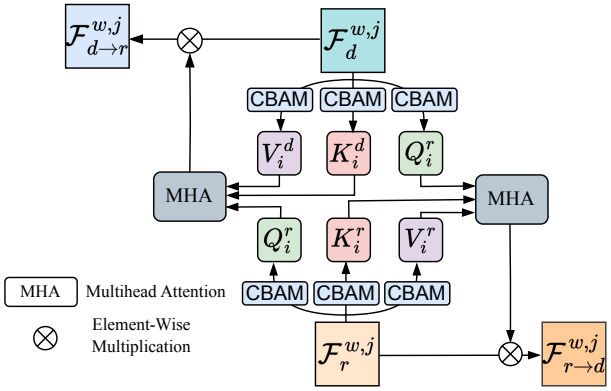


Figure 6. **Convolutional Cross Attention Module** (CCAM) demonstrates how we use warped features to identify the most relevant features for depth and reflectivity through a multi-head attention model.

The final depth-feature-infused reflectivity feature map $\mathcal{F}_{r\to d}^w$ is produced by element-wise multiplication of the cross-attention maps with warped features. This same process generates the reflectivity-feature-infused depth feature map $\mathcal{F}_{d\to r}^w$, achieving cross-modality fusion. The process of generating infused feature maps is illustrated in Fig. 6.

4.4. Progressive Multi-Scale Reconstruction

Along with the multi-scale reflectivity $\{\mathcal{F}_{\text{r}}^{w,j}\}$ and depth $\{\mathcal{F}_{\text{d}}^{w,j}\}$ warped feature maps, we use the multi-scale infused features $\{\mathcal{F}_{d\to r}^{w,j}\}$ for estimating reflectivity and the multi-scale features $\{\mathcal{F}_{r\to d}^{w,j}\}$ for depth estimation, thereby supporting a cross-information sharing approach.

Starting with the lowest scale, 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 [52] 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. 7.

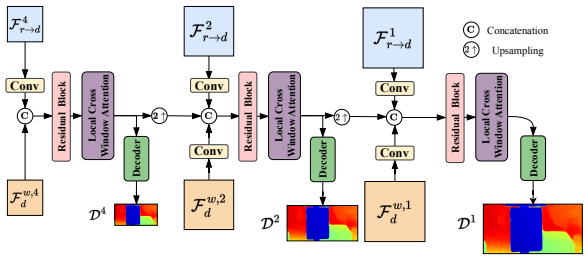


Figure 7. **Progressive reconstruction module** demonstrates how to reconstruct the multiscale depth maps $\{\mathcal{D}^i\}$.

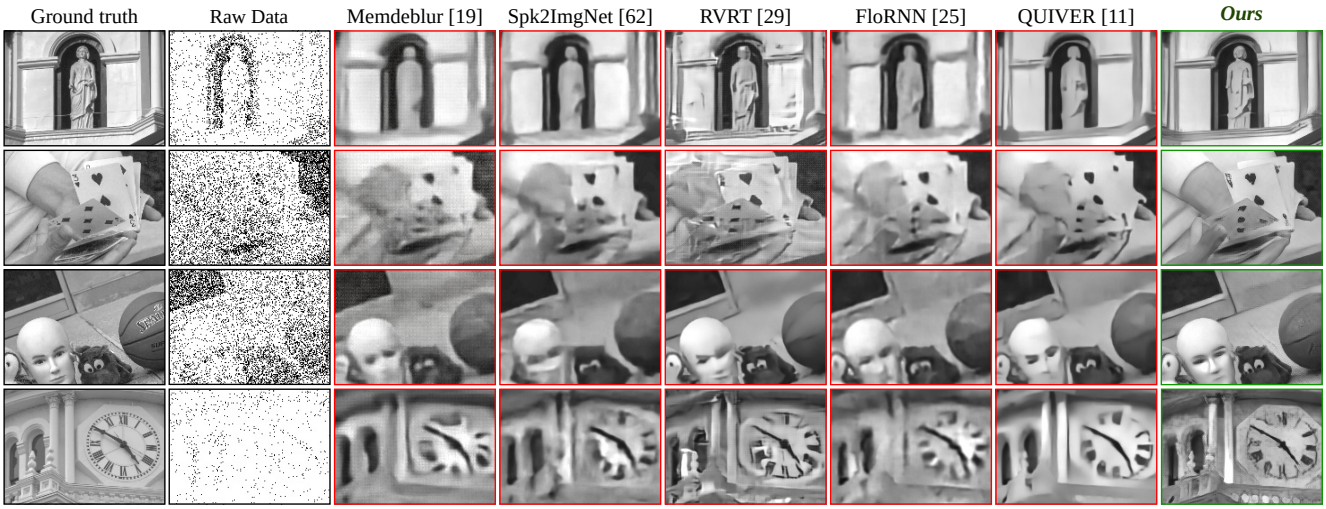


Figure 8. **Reflectivity Estimation on Simulated Data.** Visual comparisons with existing state-of-the-art reconstruction algorithms show that our method achieves sharper and more stable reflectivity, enhancing intricate details. For a fair comparison, we input 11 frames for all methods.

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

$$\mathcal{L}_{\mathcal{H}} = \lambda_1^{\mathcal{H}}\mathcal{L}(\mathcal{H}_{gt}^1, \mathcal{H}_{den}) + \lambda_2^{\mathcal{H}}\mathcal{L}(\mathcal{H}_{gt}^1, \mathcal{H}^1) + \lambda_3^{\mathcal{H}}\mathcal{L}(\mathcal{H}_{gt}^2, \mathcal{H}^2) + \lambda_4^{\mathcal{H}}\mathcal{L}(\mathcal{H}_{gt}^4, \mathcal{H}^4) + \lambda_5^{\mathcal{H}}\mathcal{P}(\mathcal{H}_{gt}^1, \mathcal{H}^1), \tag{14}$$

where $\mathcal{H}^i \in \{\mathcal{D}^i, \mathcal{I}^i\}$ denotes the depth or reflectivity output respectively, while i denotes the scale of the output. $\mathcal{P}$ represents the LPIPS loss, and $\mathcal{L}(\mathcal{A}, \mathcal{B}) = \|\mathcal{A} - \mathcal{B}\|_1 + \|\nabla_x\mathcal{A} - \nabla_x\mathcal{B}\|_1 + \|\nabla_y\mathcal{A} - \nabla_y\mathcal{B}\|_1$.

5. Experiments

In this section, we describe the experimental setup used to evaluate the proposed network. We simulate the SPL timestamp frame as outlined in [47], using an SPL setup with a 1000 frame rate, where each SPAD pixel registers the first photon in each frame, providing the time-of-flight value [18]. The timestamp follows the distribution in Eq. (2) with $M = 1$.

Since publicly available depth datasets operate within a range of 30 Hz to 60 Hz, which is unsuitable for our simulation pipeline, we use the I2-2000FPS high-speed video dataset [11], which contains 280 videos—249 for training and 31 for testing—featuring various types of high-speed motion. To generate depth maps, we use the pre-trained network provided in [56]. Further details of the simulation pipeline are available in the supplementary material.

Training of SP-LiDER. We train the network by minimizing the loss function given in Eq. (14), 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 [20] is used with an initial learning rate of 10^{-4} and a learning rate scheduler. Training is conducted on an *NVIDIA A100* Tensor GPU for two days.

5.1. Synthetic Data Experiments

Reflectivity Estimation. For the reflectivity benchmarking experiments, we use MemDeblur [19], Spk2ImgNet [62], RVRT [29], FloRNN [25], and QUIVER [11], which are SOTA dynamic reflectivity reconstruction algorithms. According to both visual Fig. 8 and quantitative Tab. 2 results, our method outperforms all benchmark algorithms. We find that joint estimation with the feature-sharing module enhances output quality and improves reconstruction stability.

Method	Metrics	
	PSNR↑	SSIM↑
QUIVER [11]	21.1108	0.6372
RVRT [29]	21.8685	0.5445
FloRNN [25]	20.1131	0.5675
MemDeblur [19]	19.8106	0.4766
Spk2ImgNet [62]	20.1490	0.5722
SP-LiDeR	**22.6552**	**0.6856**

Table 2. Quantitative evaluation of the proposed method against other state-of-the-art dynamic reflectivity reconstruction methods shows that our method achieves superior results.

Depth Estimation. We compare the depth output results with three existing SPL reconstruction networks: BRINN [43], SPDSF with intensity fusion [30], and DDFN [42]. Here, we emphasize the advantage of utilizing timestamp frames over conventional methods, especially in dynamic

scenarios as shown in Fig. 9.

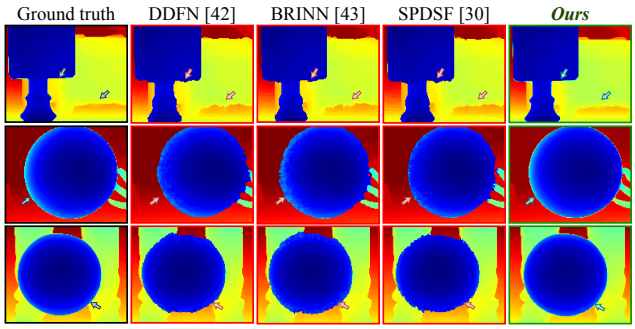


Figure 9. **Depth Estimation on Simulated Data.** Visual comparisons with state-of-the-art depth reconstruction algorithms show that SP-LiDeR produces sharper, blur-free results, while all the baselines distort object shapes. For a fair comparison, we maintain constant exposure time across all methods.

5.2. Real Data Experiments

We collect real timestamp data using [18], which is capable of reaching 1000 fps with a spatial resolution of 128×192, capturing various shapes, sizes, and motions. The system has a time-to-digital converter (TDC) resolution of approximately 35 ps. The laser operates at 25 MHz, with an effective laser pulse width of around 1 ns. As shown in Fig. 10 and Fig. 11, our method outperforms depth and reflectivity estimation, consistent with simulated data results.

5.3. Ablation

For the ablation studies, we prioritize the effect of the feature-sharing mechanism, CCAM, as it is the key contribution of this paper. Based on the qualitative results in Fig. 12 and the qualitative comparison in Tab. 3, the performance of both depth and reflectivity estimation improves. Visually, the reflectivity results show that the feature-sharing module is pivotal for achieving superior results with intricate details.

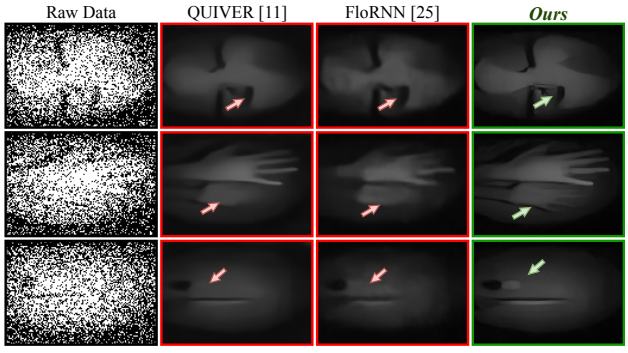


Figure 10. **Reflectivity Estimation on Real Data.** SP-LiDeR achieves improved reflectivity estimation compared to baseline methods.

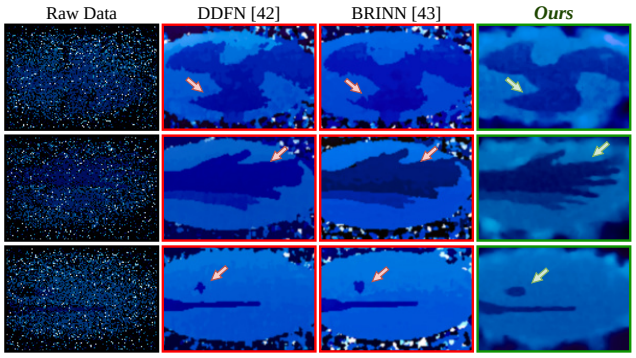


Figure 11. **Depth Estimation on Real Data.** Compared to the baselines, SP-LiDeR produces blur-free sharp results.

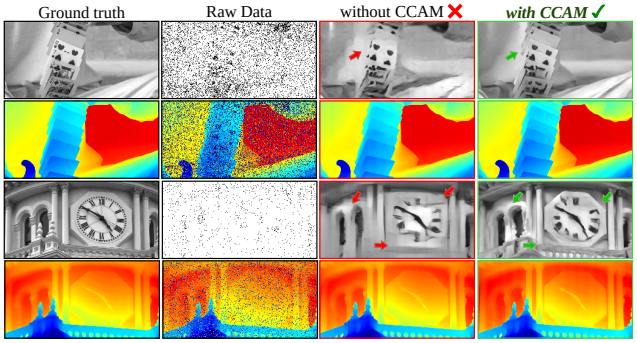


Figure 12. **Ablation Study.** Effect of Cross-Modal Information Fusion i.e., CCAM on depth and reflectivity estimation.

Method	Reflectivity		Depth
	PSNR↑	SSIM↑	RMSE↓
w/o CCAM	21.8713	0.6452	0.0095
with CCAM	22.2609	0.6706	0.0087

Table 3. Comparison of quantitative metric values for depth and reflectivity, both without and with cross-modal information fusion.

6. Conclusion

We propose SP-LiDeR, an end-to-end framework for simultaneous depth and reflectivity reconstruction from noisy SPL data of fast-moving objects. By leveraging cross-information sharing, SP-LiDeR effectively addresses high noise levels and motion blur. Extensive experiments demonstrate that SP-LiDeR surpasses SOTA methods, producing high-quality depth and reflectivity outputs, while ablation studies validate the importance of information sharing. This work introduces a novel solution to dynamic SPL reconstruction, advancing the field significantly.

References

[1] Yoann Altmann, Ximing Ren, Aongus McCarthy, Gerald S. Buller, and Steve McLaughlin. Robust bayesian target detection algorithm for depth imaging from sparse single-photon data. *IEEE Transactions on Computational Imaging*, 2(4): 456–467, 2016. 2

[2] Yoann Altmann, Ximing Ren, Aongus McCarthy, Gerald S. Buller, and Steve McLaughlin. LiDAR waveform-based analysis of depth images constructed using sparse single-photon data. *IEEE Transactions on Image Processing*, 25 (5):1935–1946, 2016. 2

[3] Yoann Altmann, Reuben Aspden, Miles Padgett, and Steve McLaughlin. A bayesian approach to denoising of single-photon binary images. *IEEE Transactions on Computational Imaging*, 3(3):460–471, 2017. 2

[4] Yoann Altmann, Stephen McLaughlin, and Michael E. Davies. Fast online 3D reconstruction of dynamic scenes from individual single-photon detection events. *IEEE Transactions on Image Processing*, 29:2666–2675, 2020. 2

[5] I. Bar-David. Communication under the poisson regime. *IEEE Transactions on Information Theory*, 15(1):31–37, 1969. 2, 3

[6] Wolfgang Becker. *Advanced Time-Correlated Single Photon Counting Techniques*. Springer, 2005. 1

[7] Alberto Boretti. A perspective on single-photon LiDAR systems. *Microwave and Optical Technology Letters*, page e33918, 2024. 1

[8] Stanley H Chan, Omar A Elgendy, and Xiran Wang. Images from bits: Non-iterative image reconstruction for quanta image sensors. *Sensors*, 16(11):1961, 2016. 2

[9] Stanley H. Chan, Hashan K. Weerasooriya, Weijian Zhang, Pamela Abshire, Istvan Gyongy, and Robert K. Henderson. Resolution limit of single-photon LiDAR. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 25307–25316, 2024. 2, 3

[10] Edoardo Charbon, Matt Fishburn, Richard Walker, Robert K. Henderson, and Cristiano Niclass. SPAD-Based Sensors. In *TOF Range-Imaging Cameras*, pages 11–38. Springer, 2013. 1

[11] Prateek Chennuri, Yiheng Chi, Enze Jiang, G. M. Dilshan Godaliyadda, Abhiram Gnanasambandam, Hamid R. Sheikh, Istvan Gyongy, and Stanley H. Chan. Quanta video restoration. In *Computer Vision – ECCV 2024*, pages 152–171, Cham, 2024. Springer Nature Switzerland. 1, 2, 7

[12] Yiheng Chi, Abhiram Gnanasambandam, Vladlen Koltun, and Stanley H Chan. Dynamic low-light imaging with quanta image sensors. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 122–138. Springer, 2020.

[13] Joon Hee Choi, Omar A Elgendy, and Stanley H Chan. Image reconstruction for quanta image sensors using deep neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6543–6547. IEEE, 2018. 2

[14] Luke Ditria and Tom Drummond. Long-term prediction of natural video sequences with robust video predictors. *arXiv preprint arXiv:2308.11079*, 2023. 4

[15] Manfredo P. do Carmo. *Differential Geometry of Curves and Surfaces*. Dover Publications, 2nd edition, 2016. 4

[16] Eric R. Fossum, Jiaju Ma, Saleh Masoodian, Leo Anzagira, and Rachel Zizza. The quanta image sensor: Every photon counts. *Sensors*, 16(8):1260, 2016. 2

[17] Anant Gupta, Atul Ingle, Andreas Velten, and Mohit Gupta. Photon-flooded single-photon 3D cameras. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6763–6772, 2019. 2

[18] Robert K. Henderson, Nick Johnston, Francescopaolo Mattioli Della Rocca, Haochang Chen, David Day-Uei Li, Graham Hungerford, Richard Hirsch, David Mcloskey, Philip Yip, and David J. S. Birch. A 192×128 time correlated SPAD image sensor in 40-nm CMOS technology. *IEEE Journal of Solid-State Circuits*, 54(7):1907–1916, 2019. 7, 8

[19] Bo Ji and Angela Yao. Multi-scale memory-based video deblurring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1919–1928, 2022. 7

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 7

[21] Ahmed Kirmani, Dheera Venkatraman, Dongeek Shin, Andrea Colaço, Franco N. C. Wong, Jeffrey H. Shapiro, and Vivek K Goyal. First-photon imaging. *Science*, 343(6166): 58–61, 2014. 2

[22] Ruangrawee Kitichotkul, Joshua Rapp, and Vivek K Goyal. The role of detection times in reflectivity estimation with single-photon lidar. *IEEE Journal of Selected Topics in Quantum Electronics*, 30(1: Single-Photon Technologies and Applications), 2024. 2, 3, 4

[23] Jongho Lee, Atul Ingle, Jenu V. Chacko, Kevin W. Eliceiri, and Mohit Gupta. CASPI: collaborative photon processing for active single-photon imaging. *Nature Communications*, 14(1):3158, 2023. 2

[24] Quentin Legros, Julian Tachella, Rachael Tobin, Aongus Mccarthy, Sylvain Meignen, Gerald S. Buller, Yoann Altmann, Stephen Mclaughlin, and Michael E. Davies. Robust 3D reconstruction of dynamic scenes from single-photon LiDAR using beta-divergences. *IEEE Transactions on Image Processing*, 30:1716–1727, 2021. 2

[25] Junyi Li, Xiaohe Wu, Zhenxing Niu, and Wangmeng Zuo. Unidirectional video denoising by mimicking backward recurrent modules with look-ahead forward ones. In *Computer Vision – ECCV 2022*, pages 592–609. Springer Nature Switzerland, 2022. 7

[26] You Li and Javier Ibanez-Guzman. LiDAR for autonomous driving: The principles, challenges, and trends for automotive LiDAR and perception systems. *IEEE Signal Processing Magazine*, 37(4):50–61, 2020. 1

[27] Zhaohui Li, Haifeng Pan, Guangyue Shen, Didi Zhai, Weihua Zhang, Lei Yang, and Guang Wu. Single-photon LiDAR for canopy detection with a multi-channel Si SPAD at 1064 nm. *Optics & Laser Technology*, 157:108749, 2023.

[28] Zheng-Ping Li, Jun-Tian Ye, Xin Huang, Peng-Yu Jiang, Yuan Cao, Yu Hong, Chao Yu, Jun Zhang, Qiang Zhang,

Cheng-Zhi Peng, Feihu Xu, and Jian-Wei Pan. Single-photon imaging over 200km. *Optica*, pages 344–349, 2021. 1

[29] Jingyun Liang, Yuchen Fan, Xiaoyu Xiang, Rakesh Ranjan, Eddy Ilg, Simon Green, Jiezhang Cao, Kai Zhang, Radu Timofte, and Luc Van Gool. Recurrent video restoration transformer with guided deformable attention. *arXiv preprint arXiv:2206.02146*, 2022. 7

[30] David B. Lindell, Matthew O'Toole, and Gordon Wetzstein. Single-photon 3D imaging with deep sensor fusion. *ACM Trans. Graph.*, 37(4):113.1–113.12, 2018. 1, 2, 4, 7

[31] Yehe Liu, Alexander Krull, Hector Basevi, Ales Leonardis, and Michael W. Jenkins. bit2bit: 1-bit quanta video reconstruction via self-supervised photon prediction. *arXiv preprint arXiv: 2410.23247*, 2024. 2

[32] Jiaju Ma, Stanley Chan, and Eric R. Fossum. Review of quanta image sensors for ultralow-light imaging. *IEEE Transactions on Electron Devices*, 69(6):2824–2839, 2022.

[33] Sizhuo Ma, Shantanu Gupta, Arin C. Ulku, Claudio Brushini, Edoardo Charbon, and Mohit Gupta. Quanta burst photography. *ACM Transactions on Graphics (TOG)*, 39(4), 2020. 2

[34] Steve Marschner and Peter Shirley. *Fundamentals of Computer Graphics, Fourth Edition*. A. K. Peters, Ltd., USA, 4th edition, 2016. 4

[35] Aongus McCarthy, Nils J. Krichel, Nathan R. Gemmell, Ximing Ren, Michael G. Tanner, Sander N. Dorenbos, Val Zwiller, Robert H. Hadfield, and Gerald S. Buller. Kilometer-range, high resolution depth imaging via 1560 nm wavelength single-photon detection. *Opt. Express*, 2013. 1

[36] Kazuhiro Morimoto, Andrei Ardelean, Ming-Lo Wu, Arin Can Ulku, Ivan Michel Antolovic, Claudio Bruschini, and Edoardo Charbon. Megapixel time-gated SPAD image sensor for 2D and 3D imaging applications. *Optica*, 7(4): 346–354, 2020. 1

[37] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 4

[38] Torben Neumann and Franz Kallage. Simulation of a direct time-of-flight LiDAR-system. *IEEE Sensors Journal*, 23(13):14245–14252, 2023. 1

[39] Preethi Padmanabhan, Chao Zhang, and Edoardo Charbon. Modeling and analysis of a direct time-of-flight sensor architecture for LiDAR applications. *Sensors*, 19(24), 2019. 1

[40] Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *Proc. of European Conference on Computer Vision (ECCV)*, 2020. 4

[41] Adithya K. Pediredla, Aswin C. Sankaranarayanan, Mauro Buttafava, Alberto Tosi, and Ashok Veeraraghavan. Signal processing based pile-up compensation for gated single-photon avalanche diodes. *arXiv preprint arXiv: 1806.07437*, 2018. 2

[42] Jiayong Peng, Zhiwei Xiong, Xin Huang, Zheng-Ping Li, Dong Liu, and Feihu Xu. Photon-efficient 3D imaging with a non-local neural network. In *Computer Vision – ECCV 2020*, 2020. 2, 7

[43] Jiayong Peng, Zhiwei Xiong, Hao Tan, Xin Huang, Zheng-Ping Li, and Feihu Xu. Boosting photon-efficient image reconstruction with a unified deep neural network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 2, 7

[44] Joshua Rapp and Vivek K Goyal. A few photons among many: Unmixing signal and noise for photon-efficient active imaging. *IEEE Transactions on Computational Imaging*, 3 (3):445–459, 2017. 1, 2, 3

[45] Joshua Rapp, Yanting Ma, Robin M. A. Dawson, and Vivek K Goyal. Dead time compensation for high-flux ranging. *IEEE Transactions on Signal Processing*, 67(13):3471–3486, 2019. 2

[46] Joshua Rapp, Julian Tachella, Yoann Altmann, Stephen McLaughlin, and Vivek K Goyal. Advances in single-photon lidar for autonomous vehicles: Working principles, challenges, and recent advances. *IEEE Signal Processing Magazine*, 37(4):62–71, 2020. 1

[47] Stirling Scholes, German Mora-Martin, Feng Zhu, Istvan Gyongy, Phil Soan, and Jonathan Leach. Fundamental limits to depth imaging with single-photon detector array sensors. *Nature Scientific Reports*, 13:176, 2023. 7

[48] Dongeek Shin, Ahmed Kirmani, Vivek K Goyal, and Jeffrey H. Shapiro. Photon-efficient computational 3-d and reflectivity imaging with single-photon detectors. *IEEE Transactions on Computational Imaging*, 1(2):112–125, 2015. 2

[49] D. L. Snyder and M. I. Miller. *Random Point Processes in Time and Space*. Springer, 1991. 2

[50] Julián Tachella, Yoann Altmann, Ximing Ren, Aongus McCarthy, Gerald S. Buller, Stephen McLaughlin, and Jean-Yves Tourneret. Bayesian 3D reconstruction of complex scenes from single-photon lidar data. *SIAM Journal on Imaging Sciences*, 12(1):521–550, 2019. 2

[51] Rachael Tobin, Abderrahim Halimi, Aongus McCarthy, Philip J. Soan, and Gerald S. Buller. Robust real-time 3D imaging of moving scenes through atmospheric obscurant using single-photon LiDAR. *Scientific Reports*, 11(1):11236, 2021. 2

[52] Zhengzhong Tu, Hossein Talebi, Han Zhang, Feng Yang, Peyman Milanfar, Alan Bovik, and Yinxiao Li. Maxvit: Multi-axis vision transformer. In *European conference on computer vision*, pages 459–479. Springer, 2022. 6

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 6

[54] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *Computer Vision – ECCV 2018*. Springer International Publishing, 2018. 6

[55] Zhiqiang Yan, Kun Wang, Xiang Li, Zhenyu Zhang, Jun Li, and Jian Yang. Rignet: Repetitive image guided network for depth completion. In *Computer Vision – ECCV 2022*, pages 214–230, 2022. 4

[56] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv: 2406.09414*, 2024. 7

[57] William C. Yau, Weijian Zhang, Hashan Kavinga Weerasooriya, and Stanley H. Chan. Analysis and improvement of rank-ordered mean algorithm in single-photon LiDAR. *arXiv preprint arXiv: 2407.20399*, 2024. 2

[58] Tianyi Zhang, Matthew Dutson, Vivek Boominathan, Mohit Gupta, and Ashok Veeraraghavan. Streaming quanta sensors for online, high-performance imaging and vision. *arXiv preprint arXiv: 2406.00859*, 2024. 2

[59] Weijian Zhang, Hashan K. Weerasooriya, Prateek Chennuri, and Stanley H. Chan. Parametric modeling and estimation of photon registrations for 3D imaging. *arXiv preprint arXiv: 2407.02712*, 2024. 2

[60] Yinda Zhang and Thomas Funkhouser. Deep depth completion of a single RGB-D image. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 175–185. IEEE Computer Society, 2018. 4

[61] Youmin Zhang, Xianda Guo, Matteo Poggi, Zheng Zhu, Guan Huang, and Stefano Mattoccia. Completionformer: Depth completion with convolutions and vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18527–18536, 2023. 4

[62] Jing Zhao, Ruiqin Xiong, Hangfan Liu, Jian Zhang, and Tiejun Huang. Spk2imgnet: Learning to reconstruct dynamic scene from continuous spike stream. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 7

[63] Kun Zhou, Wenbo Li, Liying Lu, Xiaoguang Han, and Jiangbo Lu. Revisiting temporal alignment for video restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6053–6062, 2022. 5