

# SP-TransientBench: A Real-Captured Single Photon Perception Benchmark

Hongzhou Dong<sup>1</sup>, Zili Zhang<sup>1</sup>, Ziting Wen<sup>2</sup>, Yiheng Qiang<sup>1</sup>, Runrong Deng<sup>2</sup>, Wenle Dong<sup>1</sup>, Ziwen Jiang<sup>1</sup>, Xinyang Li<sup>1</sup>, Rui Lu<sup>1</sup>, Shuoyao Sun<sup>1</sup>, Wenyu Wang<sup>3</sup>, Haitao Zheng<sup>1</sup>, Guodong Shi<sup>4</sup>, and Xiaoqiang Ren<sup>1,3</sup>\*

<sup>1</sup> School of Mechatronic Engineering and Automation, Shanghai University, Shanghai 200444, China

<sup>2</sup> School of Automation and Intelligent Manufacturing, Southern University of Science and Technology, Shenzhen 518055, China

<sup>3</sup> School of Future Technology, Shanghai University, Shanghai 200444, China

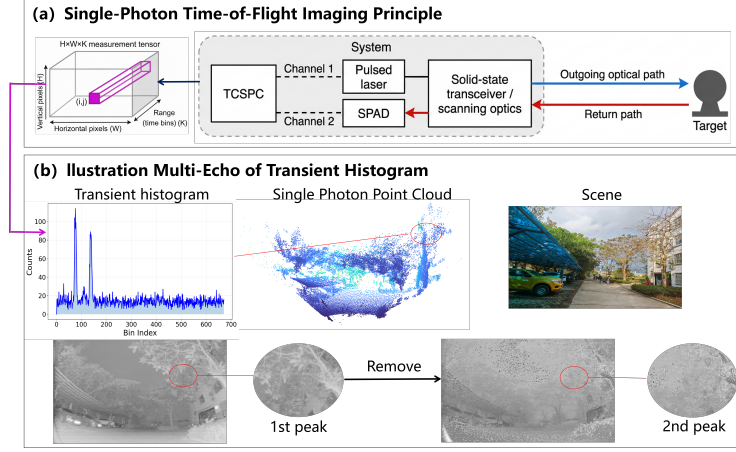
<sup>4</sup> Australian Center for Robotics, School of Aerospace, Mechanical and Mechatronics Engineering, The University of Sydney, NSW 2006, Australia

**Abstract.** Single-photon LiDAR (SPL) based on single-photon avalanche diode (SPAD) sensing enables time-resolved photon measurements with extreme sensitivity, offering unique potential for active 3D perception. However, real-world single photon perception remains fundamentally challenging due to unique measurement noise and complex multi-return transient phenomena, which jointly complicate geometric reconstruction and semantic scene understanding. Despite growing interest in SPAD-based sensing, existing studies are largely limited to simulated data or small-scale controlled captures. As a result, systematic evaluation of real-world single photon perception across depth estimation, multi-view reconstruction, and 3D semantic understanding remains underexplored. To bridge this gap, we introduce SP-TransientBench (STB), a real-captured multi-task benchmark for single photon perception. STB provides  $256 \times 192$  transient data for three tasks: 10 depth estimation samples, 9 reconstruction scenes, and 27 semantic sequences with 10,297 samples. Each view provides full time-of-flight histograms with multi-return behavior, standardized metadata, and calibrated camera poses for multi-view evaluation. We further provide 13-class 3D semantic annotations and histogram-domain multi-return annotations for selected scenes, enabling the study of raw transient cues. By providing dedicated data splits and evaluation protocols for each task, STB enables consistent and reproducible benchmarking of real-world single photon perception across multiple 3D vision problems. The dataset and code are available at: <https://huggingface.co/datasets/shuinb/SP-TransientBench>

**Keywords:** Single-photon LiDAR · Transient histograms · 3D perception · Benchmark

---

\* Corresponding author.



**Fig. 1:** Single-photon LiDAR (SPL) sensing pipeline and representative observations. (a) System overview of a SPAD-based SPL setup. A pulsed laser illuminates the scene and returned photons are detected by a SPAD and time-stamped via a TCSPC module, forming an  $H \times W \times K$  transient measurement tensor. (b) Example real-captured measurements and reconstructions. Transient histograms may exhibit multi-peak structures due to multi-path or multi-surface returns, which propagate to downstream point-cloud reconstruction and can cause noise, sparsity, and local structural ambiguities in challenging regions (highlighted).

## 1 Introduction

Single-photon LiDAR (SPL) has emerged as a powerful modality for active 3D perception due to its extreme photon sensitivity. By detecting and time-stamping individual photons, SPL systems operate under low-power illumination while achieving centimeter-level depth precision and kilometer-scale ranging [32]. Unlike conventional pulsed or amplitude-modulated LiDAR systems, which typically report a limited number of discrete echo returns per pixel, SPL records full time-resolved photon arrival histograms. This capability enables explicit modeling of transient light transport, multi-surface returns, and fine temporal structures [20, 22]. As illustrated in Fig. 1 (a), a pulsed laser illuminates the scene, and returned photons are captured by a Single Photon Avalanche Diode (SPAD) detector and time-stamped via a time-correlated single-photon counting (TCSPC) module, forming a three-dimensional transient tensor of size  $H \times W \times K$ . These properties make SPL particularly attractive for long-range perception in autonomous driving, remote sensing, and non-line-of-sight (NLOS) imaging [25, 34, 46].

The sensing principle of SPL fundamentally differs from that of conventional imaging modalities. Photon detections are discrete stochastic events governed by Poisson statistics, and measurements are accumulated as time-of-flight histograms over repeated illuminations [36]. In practice, signal photons are sparse and often contaminated by ambient illumination, resulting in a low signal-to-



background ratio (SBR) [4]. Additional factors such as pulse broadening, attenuating media, detector dead time, and temporal quantization further distort transient measurements [31, 41]. As a result, transient histograms frequently exhibit multi-peak structures arising from multi-path and multi-surface returns (Fig. 1 (b)), which propagate to downstream point-cloud reconstruction and induce noise, sparsity, and local geometric ambiguities. Robust single photon perception therefore requires both accurate physical modeling of photon statistics and strong structural priors over scene geometry. At the same time, these complex transient measurements also contain temporal cues that are lost when they are collapsed into intensity images, depth maps, or single-return point clouds, such as multi-surface and partially occluded returns.

Recent years have seen rapid progress in single photon perception, driven by advances in transient imaging models, signal processing techniques, and learning-based reconstruction frameworks [22, 24, 39, 40]. While these methods demonstrate promising results on simulated data or small-scale controlled captures, the community still lacks a large-scale, real-captured benchmark that supports systematic multi-task evaluation. This gap is becoming increasingly limiting as SPL research transitions from methodological exploration toward real-world deployment.

Constructing such a benchmark presents two major challenges. First, SPAD-based SPL hardware remains specialized and relatively inaccessible, with customized acquisition pipelines that constrain scalable data collection and standardized release. Second, annotating transient measurements is substantially more complex than labeling RGB-D images or conventional point clouds. A single pixel histogram may contain multiple temporally separated peaks corresponding to distinct surfaces along a ray [38]. These peaks can overlap or be obscured by background noise, making the mapping between temporal bins and physical surfaces inherently ambiguous. Unlike conventional labeling pipelines that operate on single-depth maps or consolidated point clouds, transient data require direct histogram-domain annotation and explicit handling of multi-return structures, which is not supported by existing tools and typically requires extensive manual disambiguation. Consequently, existing SPAD datasets remain limited in scale, realism, or task coverage.

To overcome these limitations, we introduce SP-TransientBench (STB), a real-captured multi-task SPL benchmark with full time-of-flight histograms, calibrated multi-view poses, and consistent 3D semantic annotations. Our main contributions are summarized as follows:

- (1) We introduce SP-TransientBench (STB), the first publicly available real-captured multi-task SPL dataset for 3D perception with full time-of-flight histograms, calibrated multi-view poses, and 3D semantic annotations.
- (2) We establish a standardized SPL benchmark covering depth estimation, multi-view reconstruction, and 3D semantic segmentation with unified splits and evaluation protocols.
- (3) We develop a histogram-domain semantic annotation framework and release dedicated tools for labeling multi-return transient measurements.

## 2 Related Work

### 2.1 SPAD Datasets and Benchmarks

Time-resolved SPAD sensing has led to a growing number of datasets for single-photon imaging and transient-based 3D reconstruction. Existing releases can be broadly categorized into three groups: (i) real-captured transient imaging datasets, (ii) simulation-based SPL depth datasets, and (iii) multi-view transient rendering datasets.

Early real-captured datasets demonstrate waveform-based reconstruction from SPAD measurements [11, 24], but primarily focus on single-view depth recovery and typically lack calibrated multi-view geometry or semantic annotations. Simulation-based datasets [8, 39, 40] enable controlled modeling of photon statistics, yet may not fully reproduce real-world multi-return effects and sensor non-idealities. More recent works such as TransientNeRF [22] combine real and simulated measurements for neural transient rendering. Recent datasets have also studied multi-bounce or NLOS transient perception. Shoot-Bounce-3D [14] provides simulated single-photon data for two-bounce, occlusion-aware 3D reconstruction, while DENALI [6] collects real space-time histograms for NLOS spatial reasoning.

Overall, existing datasets are designed for specific subtasks under heterogeneous acquisition protocols, limiting systematic cross-method evaluation. To date, no real-captured SPL benchmark jointly supports depth estimation, multi-view 3D reconstruction, and semantic perception under standardized splits and evaluation protocols. STB addresses this gap by providing full time-resolved histograms together with calibrated multi-view poses within a unified real acquisition framework. Table 1 summarizes representative datasets for comparison.

### 2.2 SPAD Processing, Reconstruction, and Perception

SPL processing aims to recover scene depth and reflectivity from sparse photon detections affected by background noise and temporal response effects. Early approaches relied on histogram peak detection and waveform fitting [2, 3, 26], followed by photon-efficient and probabilistic formulations that improved robustness under low-flux conditions [12, 13, 30, 35, 37, 38].

More recently, learning-based methods directly map time-resolved histograms to depth or reflectivity [27, 44, 45]. Time-resolved measurements have also been incorporated into neural implicit representations for multi-view reconstruction and transient rendering [5, 15, 20, 22], and further extended to semantic perception from SPL-derived depth maps or point clouds [17, 23, 43].

Despite rapid algorithmic progress, evaluation is typically conducted on proprietary captures or simulation-based datasets under heterogeneous protocols. STB provides a standardized real-captured benchmark that enables systematic and fair evaluation across depth estimation, multi-view reconstruction, and 3D semantic understanding.

**Table 1:** Comparison of representative SPAD datasets and benchmarks.

| Dataset                  | Annotations | Pose | Depth | Sim    | Real   | Task Focus      |
|--------------------------|-------------|------|-------|--------|--------|-----------------|
| O’Toole et al. [24]      | ✗           | ✗    | ✗     | —      | 7      | Imaging         |
| Bian et al. [8]          | ✗           | ✗    | ✗     | 2,790  | —      | Imaging         |
| SPADNet [39]             | ✗           | ✗    | ✓     | 8,266  | —      | Imaging         |
| Time-Resolved MNIST [40] | ✓           | ✗    | ✗     | 70,000 | —      | Semantics       |
| TransientNeRF [22]       | ✗           | ✗    | ✓     | 100    | 20     | Reconstruction  |
| <b>STB (Ours)</b>        | ✓           | ✓    | ✓     | —      | 10,316 | General-Purpose |

### 3 Capture Platform

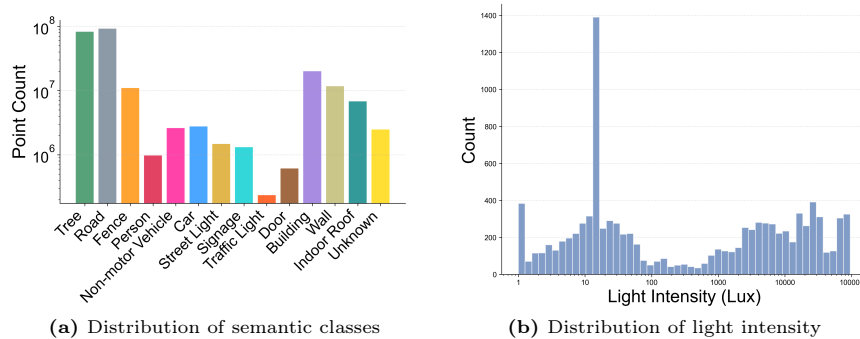
To enable real-world data collection for single photon perception, we build a multi-sensor capture platform consisting of a solid-state single-photon LiDAR (Adaps ADS6311 Hawk [1]) and an auxiliary LiDAR (Livox Avia [19]). The Livox Avia is primarily used during dataset construction for pose estimation and depth reference generation.

#### 3.1 Hardware Overview

**Single-photon LiDAR.** STB is collected using an Adaps ADS6311 Hawk, a solid-state SPL device operating under a Direct Time-of-Flight (DToF) scheme with time-correlated single-photon counting (TCSPC). A VCSEL transmitter emits ultrashort laser pulses, and returned photons are detected by a synchronized SPAD array. Photon timestamps from repeated shots are accumulated into per-pixel time-of-flight histograms, from which range is estimated based on the emission–reception delay. The native SPAD resolution is  $768 \times 576$ . With  $3 \times 3$  binning, the device outputs per-frame range measurements at  $256 \times 192$ . It provides a field of view of  $128^\circ \times 96^\circ$  (horizontal  $\times$  vertical) and operates at 10–20 Hz.

**Auxiliary LiDAR.** A Livox Avia LiDAR is mounted as an auxiliary sensor to support pose estimation and depth reference generation. The device outputs dense 3D point clouds and integrates an onboard IMU with timestamped measurements. In our pipeline, the point clouds are processed using mature LiDAR-inertial SLAM frameworks [16] to estimate drift-reduced sensor trajectories. These trajectories are transformed into the SPL coordinate frame via calibrated extrinsics and used as per-view camera poses, benefiting from the robustness and maturity of conventional LiDAR-based SLAM for geometric localization.

For the depth estimation track, Livox measurements are treated as reference ground truth. The point clouds are registered to the SPL coordinate system and projected into depth maps to provide metrically consistent supervision. Although LiDAR measurements are subject to ranging noise and cross-sensor calibration errors, they offer dense and geometrically reliable depth references for benchmarking SPL-based reconstruction methods.



**Fig. 2:** Statistical overview of semantic and illumination distributions in STB.

In SPAD-based SPL sensing, photon detections are timestamped relative to the emission clock and discretized into uniform temporal bins. Accumulating timestamps over repeated laser shots yields a per-pixel time-of-flight histogram. The temporal location of a peak encodes the round-trip travel time and thus scene depth, while its magnitude reflects return intensity; background illumination contributes an approximately uniform noise floor. Multi-surface or multi-path propagation may produce multiple peaks within a single histogram, motivating echo-level interpretation.

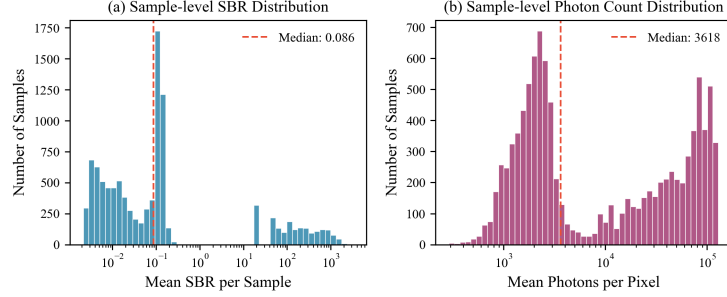
In STB, each histogram consists of  $T = 672$  bins with temporal resolution  $\Delta t = 750ps$ . We release full time-resolved waveforms rather than only peak-based depth estimates, enabling explicit modeling of multi-return transients and background statistics. This design supports algorithms that reason over peak structure, background contamination, and range-dependent sparsity directly in the histogram domain. We additionally record ambient illumination using a calibrated light meter for dataset statistics.

## 4 Dataset

### 4.1 Overview and Statistics

We introduce SP-TransientBench (STB), a real-captured benchmark for single photon perception.

The current release comprises three complementary subsets. For depth estimate, we provide 10 samples. For multi-view reconstruction, we provide 9 scenes, each captured with 20–40 transient views. For semantic understanding, we collect 27 sequences totaling 10,297 samples across indoor and outdoor environments. Each view provides full time-of-flight histograms together with standardized metadata and annotations supporting three benchmark tracks: (i) depth estimation, (ii) multi-view 3D reconstruction, and (iii) 3D semantic segmentation. Additionally, each capture is associated with measured ambient illumination intensity, enabling analysis of dataset distribution under varying lighting conditions.



**Fig. 3:** Key dataset statistics, including range distribution and measurement factors: Signal-to-Background Ratio (SBR) and Mean Photons Per Pixel (MPPP).

Figure 2 illustrates the distribution of semantic categories and recorded ambient illumination levels. Figure 3 further presents key sensing statistics, including range distribution and photon-limited measurement factors such as signal-to-background ratio (SBR) and mean photons per pixel (MPPP). Overall, STB spans a wide spectrum of photon regimes, from higher-SBR, denser-return conditions to strongly background-dominated, sparse-return scenarios. This diversity is essential for evaluating the robustness and generalization of transient-based depth recovery and downstream 3D perception models under realistic SPL operating conditions.

## 4.2 Dataset Organization

Each scene in STB is organized as a collection of views captured under a unified SPL configuration. The only shared core data across all tasks is the raw time-resolved histogram of size  $256 \times 192 \times 672$  (full waveform) for each view. The full dataset occupies approximately 168.7 GB.

**Geometric calibration.** Per-view camera poses are estimated using an auxiliary Livox Avia LiDAR via LiDAR SLAM [7, 16]. The estimated trajectories are transformed into the SPL coordinate frame through a rigid extrinsic calibration obtained via checkerboard-based alignment between the Livox and SPL sensors. SPL intrinsic parameters, including focal length, principal point, and distortion coefficients, are calibrated using a checkerboard procedure with the MATLAB Camera Calibration Toolbox, enabling consistent geometric back-projection of histogram-derived depth.

**Instrument response characterization.** The instrument response function (IRF) of the SPL system is measured by imaging a high-reflectance (99%) Enhanced Specular Reflector (ESR) target at a fixed range. The fitted temporal profile characterizes pulse broadening and temporal resolution, and is released together with the dataset. Detailed calibration procedures are provided in supp. Sec B.

**Depth reference generation.** For the depth estimation track, we provide a LiDAR-based reference depth when available. Auxiliary Livox point clouds are

**Table 2:** Per-track release contents in STB. All tracks include the raw SPAD histogram of size  $256 \times 192 \times 672$ .

| Track                     | SPAD histogram | Livox data | SPAD-Livox extrinsics | Pose | Labels | Light intensity |
|---------------------------|----------------|------------|-----------------------|------|--------|-----------------|
| Depth Estimation          | ✓              | ✓          | ✓                     | ✗    | ✗      | ✗               |
| Multi-view Reconstruction | ✓              | ✓          | ✓                     | ✓    | ✗      | ✗               |
| Semantic Understanding    | ✓              | ✗          | ✗                     | ✗    | ✓      | ✓               |

temporally accumulated to improve geometric stability and projected into the SPL viewpoint using calibrated poses to obtain metrically consistent reference depth maps.

**Task-specific releases.** In addition to the shared SPL histograms, each benchmark track provides different auxiliary data depending on its evaluation requirements. Table 2 summarizes the released components for each track. For the semantic understanding track, 3D semantic annotations are generated using a dedicated histogram-domain labeling framework tailored to multi-return transient measurements. This annotation methodology enables consistent assignment of multiple semantic labels along a single pixel ray. Details of the annotation pipeline are provided in Sec. 4.3.

### 4.3 3D Semantic Annotations

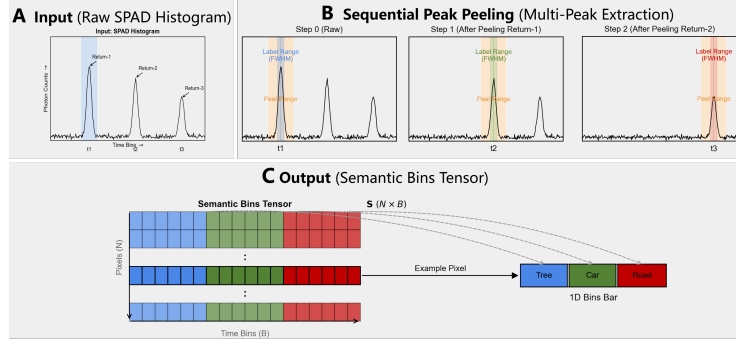
To support the semantic understanding track described above, we design a dedicated annotation pipeline tailored to multi-return SPL measurements.

**Histogram-Domain 3D Semantic Annotation via Sequential Peak Peeling.** To leverage the multi-return nature of SPAD measurements, we move beyond conventional single-depth labeling and introduce a dense annotation scheme defined directly in the histogram domain. As illustrated in Fig. 4, the ground-truth labels are represented as a semantic bin tensor  $\mathbf{S} \in \{0, \dots, C\}^{N \times B}$ , where  $N = H \times W$  denotes the number of pixels and  $B$  the number of temporal bins.

We adopt an iterative sequential peak peeling strategy to resolve overlapping and multi-peak signals (Fig. 4-B). Specifically: (1) **Peak Identification**: the dominant peak (e.g., return at  $t_1$ ) is detected in the raw histogram; (2) **Interval Assignment**: a semantic label is assigned to the peak-support interval, defined using the full width at half maximum (FWHM) to maintain temporal precision; (3) **Signal Peeling**: the signal within a broader neighborhood of the labeled peak is suppressed to expose subsequent, weaker returns (e.g.,  $t_2, t_3$ ). This procedure is repeated until all valid returns are annotated.

The resulting tensor  $\mathbf{S}$  enables a single pixel ray to encode multiple disjoint semantic entities (e.g., a semi-transparent Tree canopy followed by Road and Car, see Fig. 4-C). For geometric supervision, each labeled bin is converted from temporal index to metric range and back-projected to 3D using calibrated intrinsics, producing dense volumetric semantic supervision.

To support this process, we develop an annotation tool tailored to time-of-flight histograms with a multi-return structure. The tool is used to generate all



**Fig. 4:** Sequential peak labeling for multi-return semantic bin annotations. We label bin segments in the histogram domain and peel local peak support to reveal subsequent returns.

semantic bin annotations in STB and will be released together with the dataset and evaluation code.

## 5 Experiments

We evaluate algorithms using SPAD-only inputs, without RGB guidance or RGB-SPAD fusion. The benchmark includes three task families: (i) depth estimation, (ii) multi-view 3D reconstruction and novel-view evaluation, and (iii) 3D semantic segmentation from SPAD-derived geometry.

### 5.1 Task 1: Depth Estimation

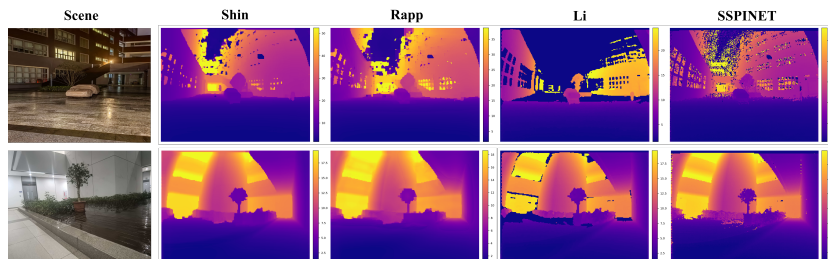
**Baselines** To evaluate depth recovery from time-resolved transient histograms, we benchmark four representative methods: Shin [37], Rapp [30], Li [18], and SSPINET [44]. All methods estimate depth directly from raw photon time-of-flight histograms. These baselines provide broad coverage of prevailing approaches in histogram-based depth estimation. Shin et al. [37] formulate photon arrivals using a probabilistic image-formation model and solve a regularized optimization problem, enabling stable reconstruction in photon-starved regimes. Rapp et al. [30] propose a statistical signal-background unmixing framework that explicitly models ambient illumination and noise. Li et al. [18] focus on long-range, extremely low-return scenarios with photon-efficient reconstruction. SSPINET [44] represents learning-based approaches, leveraging sparsity-aware neural priors for transient-to-depth inference. Together, these methods encompass model-based optimization, statistical inference, photon-efficient reconstruction, and data-driven learning, forming a comprehensive evaluation suite for assessing performance and generalization on STB.

**Evaluation Metrics** Following the protocol of Scheuble et al. [33], we evaluate depth reconstruction using Chamfer Distance (CD) and Recall. CD measures



**Table 3: Depth Estimation on STB.** Chamfer Distance (CD, m) and Recall evaluated at different temporal tolerances.

| Method       | CD ↓ (m) | Recall ↑ (%) |       |       |
|--------------|----------|--------------|-------|-------|
|              |          | 1bin         | 3bins | 5bins |
| Shin [37]    | 1.6908   | 25.97        | 74.83 | 89.14 |
| Rapp [30]    | 1.7486   | 29.17        | 68.62 | 83.88 |
| Li [18]      | 1.5506   | 34.60        | 77.20 | 88.60 |
| SSPINET [44] | 1.4978   | 44.27        | 83.72 | 93.68 |

**Fig. 5:** Qualitative results for depth estimation on STB. Columns correspond to ground truth, Shin [37], Rapp [30], Li [18], and SSPINET [44].

geometric discrepancy between predicted and reference point clouds after back-projecting depths, while Recall measures the fraction of ground-truth points recovered within a distance threshold. The thresholds are defined by the temporal resolution of the transient measurements (750 ps per bin). We report Recall at tolerances of 1, 3, and 5 temporal bins to analyze reconstruction accuracy under different geometric tolerances.

**Results and Discussion** Table 3 summarizes depth reconstruction performance on STB. SSPINET achieves the lowest CD and the highest Recall across all tolerance levels, while model-based methods remain competitive under relaxed thresholds but show reduced Recall at stricter tolerances (1 bin).

The gap between Recall (1 bin) and Recall (5 bins) highlights the effect of temporal tolerance on reconstruction accuracy. While CD values are relatively close across methods, Recall varies more noticeably, indicating differences in reconstruction completeness. Figure 5 shows representative qualitative results. Additional qualitative visualizations, together with the corresponding reference RGB images, are provided in supp. Fig. 5.

## 5.2 Task 2: Multi-View 3D Reconstruction

**Baselines** We focus on the multi-view 3D reconstruction track, which aims to recover scene geometry from multiple SPL views using camera poses for view alignment. To evaluate STB on this task, we adopt representative baselines

**Table 4: Multi-view reconstruction on STB.** We report novel-view rendering metrics for intensity (SSIM, LPIPS), depth ( $L_1$ ), and histogram (PSNR) outputs.

| Method               | Intensity       |                    | Depth              | Histogram       |
|----------------------|-----------------|--------------------|--------------------|-----------------|
|                      | SSIM $\uparrow$ | LPIPS $\downarrow$ | $L_1$ $\downarrow$ | PSNR $\uparrow$ |
| TransientNeRF [22]   | 0.7717          | 0.2315             | 0.7836             | 45.4251         |
| Transientangelo [20] | 0.7597          | 0.2501             | 0.8710             | 41.5735         |

including TransientNeRF [22], Transientangelo [20]. These methods represent typical transient-based rendering approaches that reconstruct scene geometry directly from single-photon time-of-flight histograms.

**Implementation details** Models were trained on a single RTX 4090 using Adam with an initial learning rate of  $1 \times 10^{-3}$  for 100k iterations with a batch size of 512 pixels. For each scene, we construct training sets using 3, 5, or 10 views, respectively, while reserving the remaining views for novel-view rendering and geometry evaluation. Unless otherwise specified, results reported in the main paper correspond to the 10-view setting. Results under the 3-view and 5-view configurations are included in the supp. Table 5 for completeness.

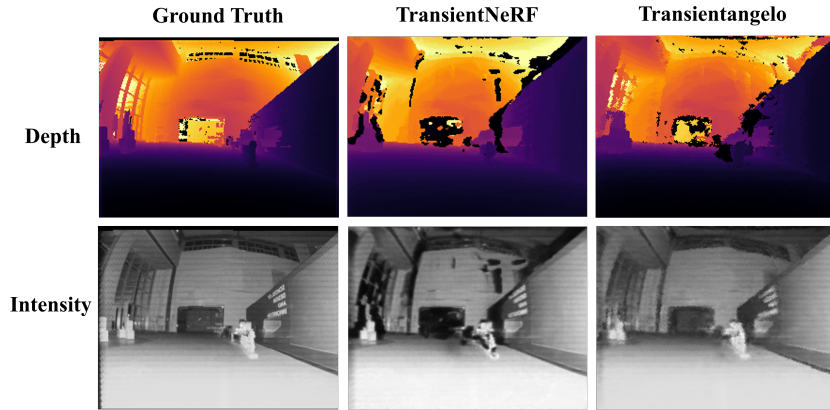
**Evaluation Metrics** Following the previous work [34], we assess reconstruction quality under three output domains, depending on the predicted representation and the rendered modality at novel views. **(i) Intensity rendering:** We report Structural Similarity Index Measure (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) between rendered intensity images and references. **(ii) Depth rendering:** We compute the per-pixel  $L_1$  error between rendered and reference depth maps over valid pixels. **(iii) Histogram rendering:** For methods that render per-pixel transient histograms, we report PSNR (Peak Signal-to-Noise Ratio) in the histogram domain.

**Results and Discussion** Table 4 reports quantitative results for multi-view reconstruction on STB. We focus on representative approaches that reconstruct scene geometry and synthesize novel views directly from single-photon transient histograms. These results demonstrate the feasibility of recovering geometry and appearance from time-resolved SPL measurements.

Figure 6 presents representative qualitative examples of depth and intensity renderings at novel viewpoints. Additional analyses and comparisons with classical reconstruction pipelines that operate on depth or intensity images are provided in supp. Sec E.2.

### 5.3 Task 3: 3D Semantic Segmentation

**Baselines** We introduce the semantic understanding track of STB, which targets 3D semantic segmentation from SPAD time-resolved measurements. To evaluate STB under this setting, we benchmark representative pipelines that combine histogram-domain preprocessing with point-cloud segmentation backbones.



**Fig. 6:** Qualitative results for multi-view reconstruction on STB. Depth (top) and intensity (bottom) renderings at a novel view. Columns correspond to Ground Truth, TransientNeRF, Transientangelo.

Since SPAD transient histograms are typically noisy along the temporal axis, we first apply preprocessing to enhance signal returns and then convert the processed histograms into single-photon point clouds through a fixed histogram-to-range projection. We consider four representative preprocessing methods: Thresholding, Matched Filtering [42], PPC [10], and SSPINET [44]. Matched filtering is a classical peak-enhancement technique, while PPC and SSPINET represent SPAD-oriented photon processing and learning-based reconstruction approaches.

For point-cloud segmentation, we adopt four widely used backbones: PointNet++ [28], PointMLP [21], Point Transformer [47], and PointNeXt [29]. As summarized in Tab. 5, we evaluate all preprocessing-backbone combinations to analyze the impact of histogram filtering and network design on semantic segmentation performance.

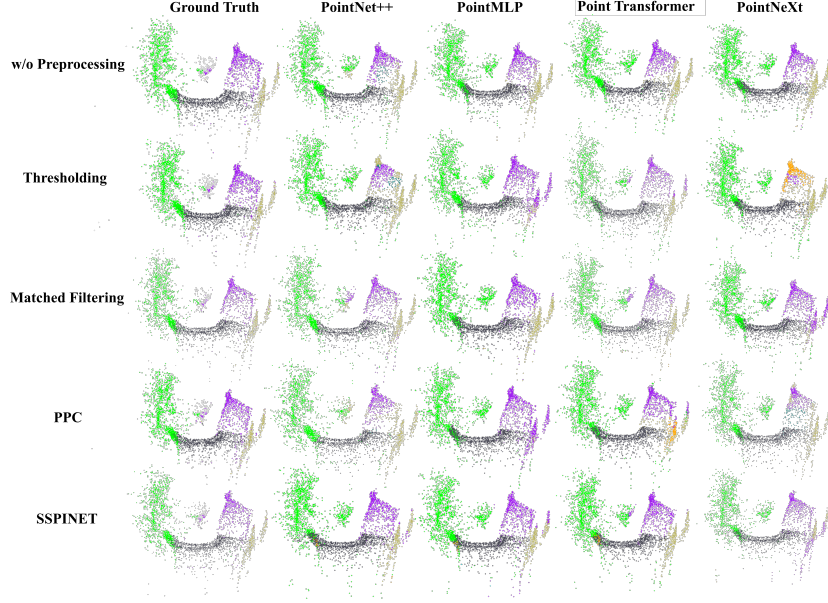
**Implementation details** Models were trained on a single RTX 4090 using 8,297 training samples (2,000 test) for 100 epochs with batch size 64 and an initial learning rate of  $1 \times 10^{-3}$ ; results are averaged over three random seeds.

**Evaluation Metrics** Consistent with prevailing 3D semantic segmentation benchmarks, our evaluation framework is designed for a thorough assessment of semantic understanding performance on STB. In particular, we focus on Overall Accuracy (OA) and mean Intersection-over-Union (mIoU) as the primary evaluation metrics, following common practice in point-cloud scene understanding. All reported results are the mean  $\pm$  standard deviation computed across three independent runs using random seeds 1, 456, 789.

**Results and Discussion** Table 5 reports 3D semantic segmentation performance under different combinations of histogram-domain preprocessing and point-cloud backbones. Applying preprocessing consistently improves results over raw transient inputs. Even simple thresholding yields noticeable gains, indicating

**Table 5:** 3D semantic segmentation from SPAD transient inputs on STB.

| Preprocessing          | PointNet++ [28]  |                  | PointMLP [21]    |                  | Point Transformer [47] |                  | PointNeXt [29]   |                  |
|------------------------|------------------|------------------|------------------|------------------|------------------------|------------------|------------------|------------------|
|                        | OA $\uparrow$    | mIoU $\uparrow$  | OA $\uparrow$    | mIoU $\uparrow$  | OA $\uparrow$          | mIoU $\uparrow$  | OA $\uparrow$    | mIoU $\uparrow$  |
| w/o Preprocessing      | $89.95 \pm 0.37$ | $49.16 \pm 1.31$ | $89.90 \pm 0.36$ | $49.11 \pm 2.40$ | $90.23 \pm 0.11$       | $49.51 \pm 0.47$ | $88.45 \pm 0.19$ | $41.05 \pm 0.75$ |
| Thresholding           | $90.53 \pm 0.21$ | $50.58 \pm 0.59$ | $90.81 \pm 0.11$ | $52.22 \pm 0.47$ | $89.96 \pm 0.88$       | $48.59 \pm 2.68$ | $89.05 \pm 0.50$ | $42.86 \pm 1.31$ |
| Matched Filtering [42] | $90.26 \pm 0.21$ | $49.71 \pm 1.77$ | $89.93 \pm 0.25$ | $50.02 \pm 0.46$ | $90.43 \pm 0.14$       | $50.37 \pm 0.30$ | $87.07 \pm 2.29$ | $40.04 \pm 1.40$ |
| PPC [10]               | $90.50 \pm 0.33$ | $50.41 \pm 1.29$ | $90.59 \pm 0.30$ | $51.50 \pm 0.13$ | $90.35 \pm 0.44$       | $48.95 \pm 1.62$ | $88.43 \pm 0.29$ | $41.26 \pm 1.10$ |
| SSPINET [44]           | $90.73 \pm 0.13$ | $51.31 \pm 0.31$ | $90.65 \pm 0.10$ | $51.78 \pm 0.70$ | $90.51 \pm 0.36$       | $50.03 \pm 0.95$ | $88.46 \pm 0.89$ | $41.54 \pm 1.52$ |

**Fig. 7:** Qualitative results for 3D semantic segmentation on STB. Columns correspond to Ground Truth and predictions from different preprocessing and backbone combinations.

that suppressing background noise before histogram-to-point projection benefits downstream semantic learning. Learning-based preprocessing further provides stable improvements across most backbones. Across architectures, performance differences are relatively modest compared to the impact of preprocessing. Despite high OA, mIoU remains around 50%, reflecting geometric sparsity and photon noise inherent to SPL measurements. Figure 7 shows representative qualitative segmentation results. Additional visualizations are provided in supp. Fig. 6.

#### 5.4 Simulation-to-Real Gap

Most learning-based methods for single-photon imaging rely on simulated data due to the difficulty of collecting large-scale real SPL measurements with annotations [33]. However, simulated transients often fail to reproduce the complex

**Table 6:** Effect of real data ratio and pretraining strategy for 3D semantic segmentation. Models are evaluated on real SPL measurements from STB.

| Training | 10%        |            | 20%        |            | 50%        |            | 100%       |            |
|----------|------------|------------|------------|------------|------------|------------|------------|------------|
|          | OA         | mIoU       | OA         | mIoU       | OA         | mIoU       | OA         | mIoU       |
| Scratch  | 83.80±0.79 | 33.19±1.37 | 86.16±0.68 | 37.98±1.09 | 89.09±0.36 | 45.94±1.37 | 91.02±0.11 | 52.24±0.36 |
| Finetune | 85.58±0.34 | 36.57±0.61 | 87.52±0.25 | 41.39±1.42 | 89.55±0.56 | 47.98±0.93 | 90.61±0.81 | 51.20±1.41 |

noise statistics, multi-return effects, and sensor artifacts present in real observations, leading to a performance gap when models are applied to real SPL data.

To study this gap, we conduct a controlled experiment on 3D semantic segmentation using PointNet++ [28]. Simulated single-photon point clouds are generated from ScanNet [9] by converting depth maps into transient histograms with a photon noise model. We compare two training strategies: (i) **Scratch**, trained only on real data, and (ii) **Finetune**, initialized from simulation pretraining and further optimized on real captures.

Table 6 reports results using different fractions of the real training set. Simulation pretraining improves performance when only limited real data is available, indicating that simulated data provides useful feature initialization. However, this advantage diminishes as more real data is introduced. With the full training set, models trained from scratch achieve comparable or slightly better performance.

These results suggest that while simulated transient data offers useful priors, it cannot fully capture the characteristics of real SPL observations. Therefore, real-captured datasets such as STB remain essential for developing robust single-photon perception algorithms.

## 6 Conclusion

We presented SP-TransientBench (STB), a real-captured benchmark for single-photon LiDAR (SPL) that enables systematic evaluation of time-resolved transient sensing for 3D perception. The dataset provides full per-pixel photon time-of-flight histograms, calibrated multi-view poses, and consistent 3D semantic annotations collected under a unified acquisition setup. By releasing raw transient measurements rather than only derived depth maps, STB supports algorithmic development that directly models photon arrival statistics, multi-return structures, and background noise characteristics inherent to SPL sensing. Beyond the dataset itself, STB establishes standardized tasks, official splits, and evaluation protocols covering depth estimation, multi-view 3D reconstruction, and 3D semantic understanding. These benchmarks enable fair and reproducible comparisons across diverse approaches, including signal-processing methods, physics-based reconstruction techniques, and learning-based perception models. To further facilitate research on transient data, we introduce a histogram-domain semantic annotation framework that supports scalable labeling of multi-return

measurements. By providing both data and tooling tailored to SPL observations, we hope STB will serve as a practical foundation for future studies on photon-limited 3D perception and transient-based scene understanding.

**Challenges and Outlook.** STB preserves intrinsic sensing characteristics of SPL, including photon sparsity, background contamination, and multi-return ambiguity, which collectively pose significant challenges for robust perception. While poses and depth references are derived from auxiliary LiDAR and cross-sensor calibration, minor SLAM drift and registration errors may introduce geometric uncertainties. Consequently, depth evaluation in STB should be interpreted relative to LiDAR-based references rather than absolute ground truth. Future extensions will expand scene diversity and explore heterogeneous sensing setups across different SPL devices and acquisition conditions. In addition, ambient illumination intensity is recorded during data capture. Although not directly incorporated into the current benchmark metrics, this metadata provides contextual information for analyzing photon-limited regimes and may support future research on illumination-aware reconstruction, sensing robustness, and adaptive perception under varying environmental conditions.

## Acknowledgments

The work was supported in part by the National Natural Science Foundation of China under Grant 62336005, 62273223, 62421004, and 62461160313, and the Shanghai Municipal Commission of Education under Grant 24SG38.

## References

1. Adaps Photonics: Ads6311 hawk solid-s. <https://www.adapsphotonics.com/en/product-55669-218658.html> (2024), accessed: 2026-01-10
2. Altmann, Y., Aspden, R., Padgett, M., McLaughlin, S.: A bayesian approach to denoising of single-photon binary images. *IEEE Transactions on Computational Imaging* **3**(3), 460–471 (2017)
3. Altmann, Y., Ren, X., McCarthy, A., Buller, G.S., McLaughlin, S.: Lidar waveform-based analysis of depth images constructed using sparse single-photon data. *IEEE Transactions on Image Processing* **25**(5), 1935–1946 (2016)
4. Altmann, Y., Ren, X., McCarthy, A., Buller, G.S., McLaughlin, S.: Robust bayesian target detection algorithm for depth imaging from sparse single-photon data. *IEEE Transactions on Computational Imaging* **2**(4), 456–467 (2016)
5. Attal, B., Laidlaw, E., Gokaslan, A., Kim, C., Richardt, C., Tompkin, J., O’ Toole, M.: Törf: Time-of-flight radiance fields for dynamic scene view synthesis. In: *Advances in Neural Information Processing Systems*. vol. 34, pp. 26289–26301 (2021)
6. Behari, N., Rivero, D., Apostolides, L., Ghosh, S., Liang, P.P., Raskar, R.: DENALI: A dataset enabling non-line-of-sight spatial reasoning with low-cost lidars. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2026)
7. Besl, P., McKay, N.D.: A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **14**(2), 239–256 (1992)

8. Bian, L., Song, H., Peng, L., Chang, X., Yang, X., Horstmeyer, R., Ye, L., Zhu, C., Qin, T., Zheng, D., et al.: High-resolution single-photon imaging with physics-informed deep learning. *Nature Communications* **14**(1), 5902 (2023)
9. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 5828–5839 (2017)
10. Goyal, B., Gutierrez-Barragan, F., Lin, W., Velten, A., Li, Y., Gupta, M.: Robust 3d object detection using probabilistic point clouds from single-photon lidars. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 28417–28427 (2025)
11. Gutierrez-Barragan, F., Ingle, A., Seets, T., Gupta, M., Velten, A.: Compressive single-photon 3d cameras. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 17854–17864 (2022)
12. Halimi, A., Tobin, R., McCarthy, A., McLaughlin, S., Buller, G.S.: Restoration of multilayered single-photon 3d lidar images. In: *European Signal Processing Conference (EUSIPCO)*. pp. 708–712 (2017)
13. Kirmani, A., Venkatraman, D., Shin, D., Colaço, A., Wong, F.N., Shapiro, J.H., Goyal, V.K.: First-photon imaging. *Science* **343**(6166), 58–61 (2014)
14. Klinghoffer, T., Somasundaram, S., Xiang, X., Fan, Y., Richardt, C., Dave, A., Raskar, R., Ranjan, R.: Shoot-bounce-3d: Single-shot occlusion-aware 3d from lidar by decomposing two-bounce light. In: *SIGGRAPH Asia* (2025)
15. Klinghoffer, T., Xiang, X., Somasundaram, S., Fan, Y., Richardt, C., Raskar, R., Ranjan, R.: Platonerf: 3d reconstruction in plato’s cave via single-view two-bounce lidar. In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 14565–14574 (2024)
16. Kümmerle, R., Grisetti, G., Strasdat, H., Konolige, K., Burgard, W.: G2o: A general framework for graph optimization. In: *2011 IEEE international conference on robotics and automation (ICRA)*. pp. 3607–3613 (2011)
17. Li, X., Liu, J., Zhao, G., Liu, L., Zhang, W., Hu, X., Cheng, S.: High precision single-photon object detection via deep neural networks. *Optics Express* **32**(21), 37224–37237 (2024)
18. Li, Z.P., Huang, X., Cao, Y., Wang, B., Li, Y.H., Jin, W., Yu, C., Zhang, J., Zhang, Q., Peng, C.Z., et al.: Single-photon computational 3d imaging at 45 km. *Photonics Research* **8**(9), 1532–1540 (2020)
19. Livox Technology: Livox avia lidar. <https://www.livoxtech.com/cn/avia> (2024), accessed: 2026-01-10
20. Luo, W., Malik, A., Lindell, D.B.: Transientangelo: Few-viewpoint surface reconstruction using single-photon lidar. In: *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. pp. 8723–8733 (2025)
21. Ma, X., Qin, C., You, H., Ran, H., Fu, Y.: Rethinking network design and local geometry in point cloud: A simple residual mlp framework. In: *International Conference on Learning Representations (ICLR)* (2022)
22. Malik, A., Mirdehghan, P., Noursias, S., Kutulakos, K., Lindell, D.: Transient neural radiance fields for lidar view synthesis and 3d reconstruction. *Advances in neural information processing systems (NeurIPS)* **36**, 71569–71581 (2023)
23. Mora-Martín, G., Turpin, A., Ruget, A., Halimi, A., Henderson, R., Leach, J., Gyongy, I.: High-speed object detection with a single-photon time-of-flight image sensor. *Optics express* **29**(21), 33184–33196 (2021)

24. O’Toole, M., Heide, F., Lindell, D.B., Zang, K., Diamond, S., Wetzstein, G.: Reconstructing transient images from single-photon sensors. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 1539–1547 (2017)
25. Ovrén, H., Holmberg, M., Henriksson, M.: Long-range time-correlated single-photon counting lidar 3d-reconstruction from a moving ground vehicle. *Journal of Field Robotics* (2025)
26. Pellegrini, S., Buller, G.S., Smith, J.M., Wallace, A.M., Cova, S.: Laser-based distance measurement using picosecond resolution time-correlated single-photon counting. *Measurement Science and Technology* **11**(6), 712–716 (2000)
27. Peng, J., Xiong, Z., Huang, X., Li, Z.P., Liu, D., Xu, F.: Photon-efficient 3d imaging with a non-local neural network. In: European Conference on Computer Vision (ECCV). pp. 225–241 (2020)
28. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems (NeurIPS)* **30** (2017)
29. Qian, G., Li, Y., Peng, H., Mai, J., Hammoud, H., Elhoseiny, M., Ghanem, B.: Pointnext: Revisiting pointnet++ with improved training and scaling strategies. In: *Advances in Neural Information Processing Systems (NeurIPS)*. vol. 35, pp. 23192–23204 (2022)
30. Rapp, J., Goyal, V.K.: A few photons among many: Unmixing signal and noise for photon-efficient active imaging. *IEEE Transactions on Computational Imaging* **3**(3), 445–459 (2017)
31. Rapp, J., Ma, Y., Dawson, R.M., Goyal, V.K.: Dead time compensation for high-flux ranging. *IEEE Transactions on Signal Processing* **67**(13), 3471–3486 (2019)
32. Rapp, J., Tachella, J., Altmann, Y., McLaughlin, S., Goyal, V.K.: Advances in single-photon lidar for autonomous vehicles: Working principles, challenges, and recent advances. *IEEE Signal Processing Magazine* **37**(4), 62–71 (2020)
33. Scheuble, D., Holzhüter, H., Peters, S., Bijelic, M., Heide, F.: Lidar waveforms are worth 40x128x33 words. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 28913–28924 (2025)
34. Scheuble, D., Ramazzina, A., Holzhüter, H., Gasperini, S., Peters, S., Tombari, F., Bijelic, M., Heide, F.: Transient lasso: Transient large-scale scene reconstruction. In: *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*. pp. 1–12 (2025)
35. Shin, D., Kirmani, A., Goyal, V.K., Shapiro, J.H.: Photon-efficient computational 3-d and reflectivity imaging with single-photon detectors. *IEEE Transactions on Computational Imaging* **1**(2), 112–125 (2015)
36. Shin, D., Shapiro, J.H., Goyal, V.K.: Single-photon depth imaging using a union-of-subspaces model. *IEEE Signal Processing Letters* **22**(12), 2254–2258 (2015)
37. Shin, D., Xu, F., Venkatraman, D., Lussana, R., Villa, F., Zappa, F., Goyal, V.K., Wong, F.N., Shapiro, J.H.: Photon-efficient imaging with a single-photon camera. *Nature communications* **7**(1), 12046 (2016)
38. Shin, D., Xu, F., Wong, F.N., Shapiro, J.H., Goyal, V.K.: Computational multi-depth single-photon imaging. *Optics express* **24**(3), 1873–1888 (2016)
39. Sun, Z., Lindell, D.B., Solgaard, O., Wetzstein, G.: Spadnet: deep rgb-spad sensor fusion assisted by monocular depth estimation. *Optics express* **28**(10), 14948–14962 (2020)
40. Suonsivu, A., Salmela, L., Peretti, E., Uosukainen, L., Bilcu, R.C., Boracchi, G.: Time-resolved mnist dataset for single-photon recognition. In: *Computer Vision – ECCV 2024 Workshops*. pp. 127–143 (2024)



41. Tobin, R., Halimi, A., McCarthy, A., Laurenzis, M., Christnacher, F., Buller, G.S.: Three-dimensional single-photon imaging through obscurants. *Optics express* **27**(4), 4590–4611 (2019)
42. Turin, G.: An introduction to matched filters. *IRE Transactions on Information Theory* **6**(3), 311–329 (1960)
43. Wen, Z., Dong, W., Zhang, Z., Qiang, Y., DING, K., Ren, X.: Noise-aware adaptation of pre-trained foundation models for single-photon image classification. *Transactions on Machine Learning Research* (2026)
44. Yao, G., Chen, Y., Jiang, C., Xuan, Y., Hu, X., Liu, Y., Pan, Y.: Dynamic single-photon 3d imaging with a sparsity-based neural network. *Optics Express* **30**(21), 37323–37340 (2022)
45. Yao, G., Chen, Y., Liu, Y., Hu, X., Pan, Y.: Robust photon-efficient imaging using a pixel-wise residual shrinkage network. *Optics Express* **30**(11), 18856–18873 (2022)
46. Young, A., Batagoda, N.M., Zhang, H., Dave, A., Pediredla, A., Negrut, D., Raskar, R.: Enhancing autonomous navigation by imaging hidden objects using single-photon lidar. In: 2025 IEEE International Conference on Robotics and Automation (ICRA). pp. 4907–4914 (2025)
47. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV). pp. 16259–16268 (2021)