A Paradigm Shift in High-Resolution Depth Estimation Using SPAD-Based LiDAR Histograms: From Signal Filtering to Lightweight Similarity Learning

Minsung Lee[*1], **Seo Hyun Kim**[*1], **Yeonsu Park**[†2], **Hyeongseok Seo**[†3], **Jongmin Lee**[†1]

[1]Department of Intelligence Semiconductor Engineering, Ajou University, Suwon, Republic of Korea
[2]Department of Computer Science and Engineering, Kangwon National University, Chuncheon, Republic of Korea
[3]Department of Electrical and Electronics Engineering, Pusan National University, Busan, Republic of Korea
david513828@ajou.ac.kr, ksh9954@ajou.ac.kr, yeonsu.park@kangwon.ac.kr, h.seo@pusan.ac.kr, jongmin@ajou.ac.kr

Abstract

Accurate and efficient depth estimation from time-of-flight (ToF) LiDAR is essential for autonomous systems operating in real-world environments. However, traditional histogram-based depth estimation (HBDE) algorithms face fundamental limitations in balancing depth performance and computational cost, and they struggle under signal-induced pile-up distortion. While deep learning has shown promise, existing neural network-based methods rely on large models that are impractical for deployment on edge hardware. To bridge this critical gap, we propose a paradigm shift in histogram-based ToF estimation, reframing depth estimation from signal filtering to lightweight similarity learning. Instead of attempting to correct the distorted signal, our approach learns a specialized metric where the measure of similarity between the distorted histogram and a reference pulse is the temporal shift itself. The resulting 57.61 KB model, over 215.2× smaller than state-of-the-art deep learning approaches, achieves real-time performance (106.27 fps) on an FPGA. It delivers superior accuracy across nearly all signal-noise conditions, including 2.21 cm RMSE at severe pile-up scenarios, significantly outperforming conventional methods while remaining practical for on-device deployment.

Code — https://github.com/knu-bigdata/litofnet

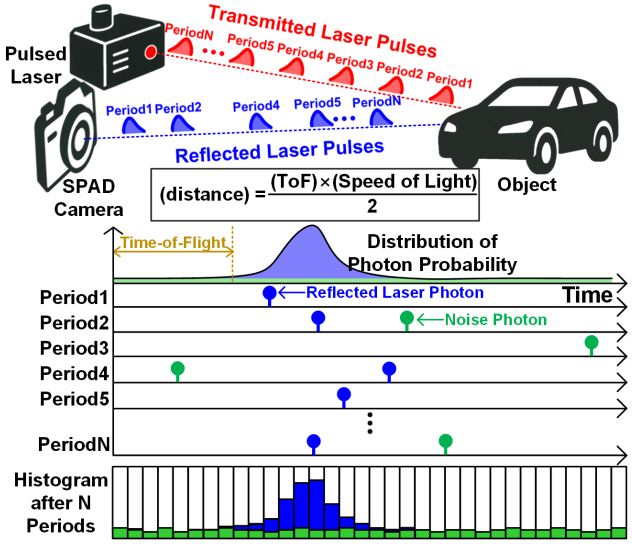


Figure 1: Time-correlated single photon counting method and an example of the built histogram. Individual photon detection events follow a Poisson process.

Introduction

3D depth sensing is crucial for applications such as autonomous driving and robotics. Among existing techniques, time-of-flight (ToF) light detection and ranging (LiDAR) provides centimeter-level accuracy at ranges of hundreds of meters. Unlike stereo vision (Li et al. 2018) or structured light, ToF reliably measures distance without textured surfaces or structured illumination.

LiDAR systems measure the round-trip time of emitted laser pulses using single-photon avalanche diodes (SPADs) (Cova et al. 1996) and accumulate timestamps into histograms through time-correlated single-photon counting (TCSPC) (Figure 1). Histogram-based depth estimation (HBDE) algorithms, such as cross-correlation using instrumental response functions (IRF) (Umasuthan et al. 1998) or center-of-mass (COM) (Niclass et al. 2005), then produce centimeter or millimeter scale resolution even under low-signal or noisy conditions.

However, traditional HBDE algorithms suffer from two key limitations. First, there is a fundamental trade-off between computational cost and accuracy. Low-cost methods (Niclass et al. 2014) generally compromise accuracy, whereas accurate methods (Heide et al. 2018) require high computational resources, making real-time hardware implementation impractical. Second, conventional algorithms degrade severely with strong return signals from nearby or highly reflective objects. In these scenarios, signal-induced pile-up distorts the histogram shape by saturating detectors with early-arriving photons and suppressing the counts of laser photons (Zhang et al. 2024). Since traditional linear methods assume a fixed histogram shape regardless of sig-

[*]These authors contributed equally.
[†]Corresponding authors.

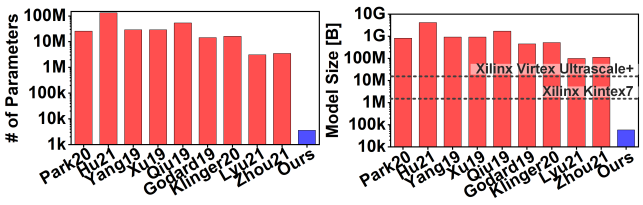


Figure 2: Comparison of parameter count (left) and model size (right) across state-of-the-art deep learning-based depth estimation models. Yu et al. (2025); Peng et al. (2023) are omitted due to unavailable code; their complex 3D architectures are presumed to far exceed typical FPGA constraints.

nal intensity, they yield incorrect cross-correlation and thus poor accuracy under such conditions.

Deep learning has emerged as a strong alternative to traditional algorithms, but it introduces a key challenge: deployability. As shown in Figure 2, state-of-the-art networks depend on large models whose size and computational cost hinder real-time inference on resource-limited hardware such as FPGAs (Park et al. 2020; Yang, Wong, and Soatto 2019; Qiu et al. 2019; Klinger et al. 2020; Lyu et al. 2021).

This leaves a critical gap for a solution that is both robust enough to handle severe distortion and efficient enough for on-device deployment. To bridge this gap, we propose a paradigm shift from signal filtering to lightweight similarity learning. Instead of attempting to correct or align distorted histograms, our approach learns a specialized metric where the measure of similarity itself becomes the temporal shift. In essence, the model is trained to answer: "Even with shape distortion present, what temporal shift value yields the highest similarity between these two signals?"

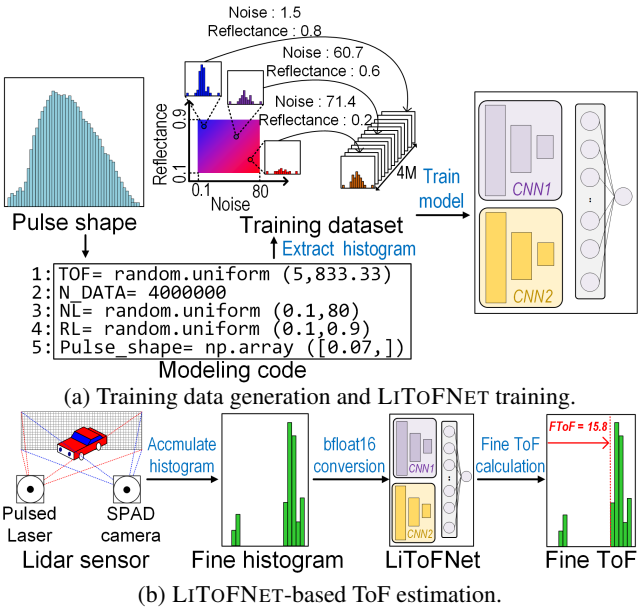
In summary, our key contributions are threefold:

- **A new paradigm:** We shift from signal filtering to lightweight similarity learning, presenting the first practical on-device AI that effectively handles pile-up.
- **Hardware-validated practicality:** Our compact 57.61 KB model achieves real-time inference at 106.27 fps on an FPGA, showing its suitability for embedded systems.
- **Demonstrated robustness:** The proposed method exhibits remarkable robustness across severe pile-up, abrupt reflectance changes, and challenging lighting conditions where existing approaches fail.

Related Work

Classical HBDE techniques. Early HBDE algorithms computed ToF via cross-correlation between photon histograms and laser pulse shapes (Umasuthan et al. 1998), requiring extensive pre-calibration and Gaussian modeling (Buller and Wallace 2007). Subsequent methods handled noise by thresholding signals (Niclass et al. 2005, 2008) or simplified computation using COM estimation assuming Gaussian pulses (Grull et al. 2011). Despite simplicity, these methods fail to maintain accuracy under severe pile-up conditions.

Pile-up compensation algorithms. Recent pile-up compensation algorithms combine theoretical and practical approaches: optimal photon-flux criteria via adaptive attenuation (Gupta et al. 2019), Markov chain modeling of SPAD dead-time (Rapp et al. 2019), training-free spatio-temporal correlation pipelines (Lee et al. 2023), probabilistic image-formation models (Heide et al. 2018), and Gaussian-uniform mixture models for dead-time correction (Zhang et al. 2024). These methods effectively handle noise-induced pile-up, but neglect severe signal-induced pile-up. Additionally, their iterative optimization steps require substantial computational resources, making real-time hardware deployment difficult.



(a) Training data generation and LiToFNet training.

(b) LiToFNet-based ToF estimation.

Figure 3: The overview of the system architecture.

Deep learning-based depth estimation. Deep learning has enhanced indirect ToF (iToF) depth sensing (Su et al. 2018; Chugunov et al. 2021) and LiDAR-based object detection (Zhang et al. 2023; Choi et al. 2023). For direct ToF (dToF), methods combining passive imagery with neural upsampling improve spatial resolution (Zhuo et al. 2023), while CNN-based sensor fusion techniques merge ToF histograms and intensity images for robust low-photon-count reconstruction (Lindell, O'Toole, and Wetzstein 2018; Hu et al. 2021; Xu et al. 2019; Godard et al. 2019; Zhou et al. 2021). However, existing networks either assume undistorted pulse shapes or require computationally heavy models impractical for on-device inference. Compact models capable of handling severe pile-up conditions for real-time FPGA implementation remain an unresolved challenge.

System Architecture

The top-level architecture of the proposed system is depicted in Figure 3. Our LiDAR module employs a two-step histogram approach: first, coarse histograms with 2-meter temporal bins (13.33 ns resolution) identify echo positions and determine the coarse ToF (CToF) value; second, starting from the CToF position, a 64-bin fine histogram with 6.25-cm resolution (416.66 ps) captures detailed depth

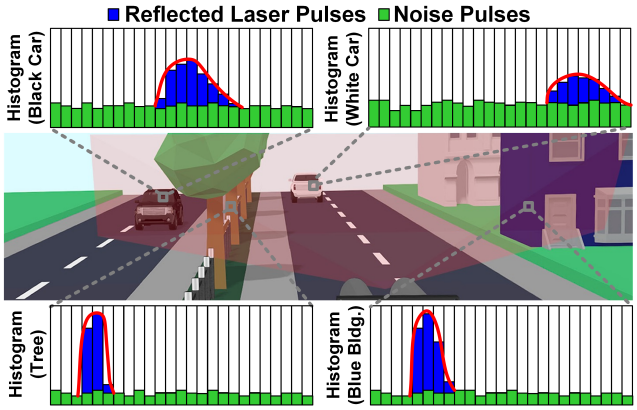


Figure 4: Accumulated ToF histogram examples for multiple objects, illustrating how object reflectance and distance affect both noise floor levels and reflected laser signal intensity in the same sunlight environment.

data over a 2-coarse-bin window, significantly reducing histogram memory usage. The LiDAR system outputs both the CToF value and the corresponding fine histogram. The proposed architecture uses the CToF value as auxiliary input, converts fine histograms into `bfloat16` format, and processes them via an FPGA-based LITOFNET accelerator to estimate the fine ToF (FToF) value.

Data augmentation. To construct the training dataset, we developed a simulation framework that models fine histograms generated by our LiDAR sensor under diverse conditions (Figure 3(a)). We first extract the laser pulse waveform from a commercial LiDAR and use it to build a pulse-modeling framework that takes four parameters—ToF, noise level, reflectance, and TDC resolution. By combining these with the TX characteristics of our system, the simulation accounts for intensity variations from inverse-square distance dependence and the interactions between noise levels, signal strength, and reflectance. The resulting photonic events are modeled as an inhomogeneous Poisson process with a time-varying rate reflecting both the signal and noise, while incorporating SPAD dead time to reproduce the pile-up distortions observed in real histograms. Since histogram shapes vary strongly with reflectance and distance—strong signals produce narrow pile-up-driven histograms, while weak signals follow the original pulse shape (Figure 4)—we uniformly sample noise levels from [0.1, 80], reflectance from [0.1, 0.9], and ToF from [5 ns, 833.33 ns]. For each configuration, the simulation generates both CToF values and fine histograms consistent with our actual LiDAR outputs. Repeating this across all parameter combinations yields four million histogram samples for training LITOFNET.

Inference architecture. Once trained, the model is deployed within the inference pipeline shown in Figure 3(b). This pipeline processes an accumulated fine histogram from the sensor, first converting it to the `bfloat16` format and then processing it through the core LITOFNET module to produce the final ToF estimate.

ToF Estimation via Similarity Learning

The conventional ToF estimation task, whether via cross-correlation (Niclass et al. 2014) or maximum likelihood (Altmann et al. 2016), is fundamentally a search for a temporal shift that maximizes the similarity between a measured histogram and a reference pulse. This problem structure is an ideal fit for a Siamese network, which excels at learning such similarity metrics (Bromley et al. 1993; Chopra, Hadsell, and LeCun 2005; Koch et al. 2015).

Our model, LITOFNET, learns a similarity metric by processing the reference pulse and the measured histogram through two identical, weight-sharing subnetworks. Instead of correcting the distorted signal, the network learns to extract robust latent features from both the reference pulse and the measured histogram, such that the difference between these features directly regresses the temporal shift (ToF). This end-to-end design yields two key advantages: it provides inherent immunity to the pile-up distortions, and the weight-sharing ensures the model remains compact and efficient for real-time FPGA deployment.

Model

To learn the temporal shift between the reference pulse and the measured histogram, LITOFNET utilizes a dual-branch CNN architecture with shared weights (Figure 5). Each branch comprises L sequential layers, with each layer l containing N_l computational units, where $\mathbf{h}_{1,l}$ denotes the hidden vector at layer l for the pulse shape branch, and $\mathbf{h}_{2,l}$ denotes the same for the fine histogram branch. In order to enable the subsequent feature difference computation, the original pulse shape is first transformed through an initial fully-connected (FC) layer to match the 64-dimensional representation of the fine histogram, ensuring dimensional compatibility at the network input stage. We use exclusively rectified linear units (ReLU) in all layers of the subnetworks.

Each subnetwork consists of a sequence of convolutional layers. The first convolutional layer utilizes a single input channel, 14 filters, a kernel size of 12, a stride of 3, and no padding. The second convolutional layer uses 14 input channels, 42 filters, a kernel size of 3, a stride of 2, and padding of 2. Following each convolutional layer, we apply average pooling (kernel size 2, stride 2). This choice is crucial for our similarity comparison, as it preserves the relative characteristics of the signal distribution. In contrast, max pooling discards this information by focusing only on peak values. The hyperparameter configuration was determined through an experimental performance optimization process. We apply a ReLU activation function to the output feature maps. Thus, the k-th feature map in each layer is given by:

$$a_{i,m}^{(k)} = \max(0, \mathbf{W}_{l-1,l}^{(k)} \circledast \mathbf{h}_{i,l-1} + b_l^{(k)}), \quad i \in \{1, 2\}$$

where $\mathbf{W}_{l-1,l}$ is the tensor representing the weights for layer l, $b_l^{(k)}$ is the scalar bias term corresponding to the kth feature map of layer l, and $a_{i,m}^{(k)}$ represents the activation output at position m for the kth feature map for input i. Here, $\circledast$ denotes the convolution operation.

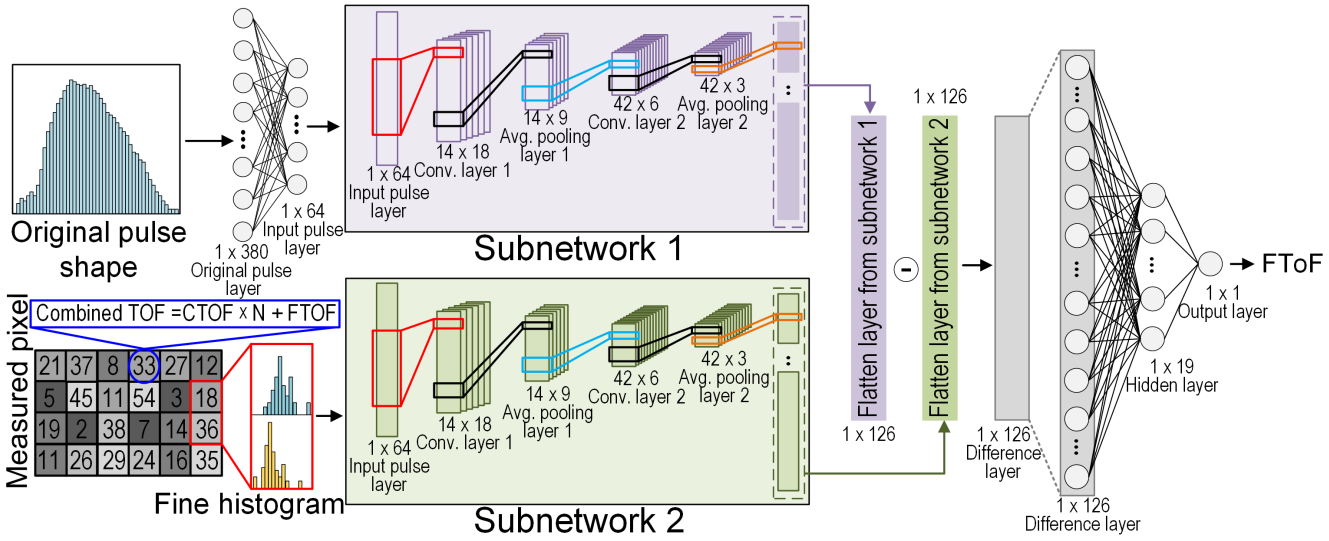


Figure 5: The proposed neural network architecture of LiTOFNet.

The flattened feature vectors from the two subnetworks, denoted $\mathbf{h}_{1,L-1}$ and $\mathbf{h}_{2,L-1}$, encapsulate the learned representations of the reference pulse and the measured histogram, respectively. The core of our estimation lies in their difference, which directly encodes ToF. This difference vector is then fed into a sequence of FC layers, $\mathcal{F}$, which acts as a regression head to produce the final ToF estimate:

$$\hat{t}_{\text{ToF}} = \mathcal{F}(\mathbf{h}_{2,L-1} - \mathbf{h}_{1,L-1})$$

Learning

Loss function. We train LiTOFNet by minimizing the mean squared error (MSE) between the predicted ToF ($\hat{t}_{\text{ToF}}$) and the ground truth (t_{ToF}), formulated as:

$$\mathcal{L} = (\hat{t}_{\text{ToF}} - t_{\text{ToF}})^2$$

Optimization. We train the network using the Adam optimizer (Kingma and Ba 2014) with a learning rate of 3.87×10^{-4} and a mini-batch size of 100. These hyperparameters were tuned using Optuna (Akiba et al. 2019). The model is trained for a maximum of 500 epochs with an early stopping patience of 100 on the validation loss to prevent overfitting.

Weight initialization. To facilitate stable learning, all convolutional and FC layer weights are initialized using Kaiming initialization (He et al. 2015), suitable for our ReLU-based network. All bias terms are initialized to zero.

Dataset. We train LiTOFNet using the 3 million simulated fine histogram samples for training and 91k samples for validation, as described in the 'Data Augmentation' subsection (page 3). This synthetic dataset comprehensively covers noise conditions, reflectance variations, and distance ranges encountered in real-world LiDAR applications.

Hardware Implementation

The practical deployability of our 57.61 KB model is validated by its hardware implementation on a Xilinx Kintex-7 FPGA. Three key optimizations ensure efficiency: (1) precomputing the static pulse's output, leveraging the Siamese structure; (2) adopting the `bfloat16` format, a choice informed by our ablation study (Table 4); and (3) employing a pipelined architecture to maximize throughput. This design achieves real-time performance, with detailed metrics presented in the Experimental Evaluation section.

	Ours	Heide18	Rapp19	Niclass14	Gyongy20	Okino20	Buller07
Avg. (m)	**-0.002**	0.273	-0.279	-0.304	0.260	<u>-0.154</u>	-0.297
Min (m)	-0.182	**0.059**	-0.468	-0.508	<u>0.106</u>	-0.687	-0.452
Max (m)	<u>0.052</u>	0.517	**-0.047**	0.517	0.470	0.107	-0.089
RMSE (m)	**0.023**	0.319	0.319	0.342	0.290	<u>0.248</u>	0.319

Table 1: Depth estimation accuracy comparison under non-uniform reflectance conditions. Best results are shown in **bold**, and second-best results are <u>underlined</u>.

Experimental Evaluation

Experimental Setup

All real-world data were collected using a commercial SPAD-based dToF LiDAR module (SV110, SolidVue inc.), which employs a bi-axial optical setup with TX and RX lenses covering the same field-of-view of $120.96° \times 32.76°$. The imaging sensor system features a 192×52 SPAD array paired with a matching laser diode array that enables row-by-row scanning (Roh et al. 2024). The laser operates at 940 nm wavelength with a repetition rate of 133.3 kHz.

The module implements the two-step multievent TDC architecture (Seo et al. 2021), generating both CToF values and fine histograms from the entire SPAD array. Each fine histogram is accumulated from 64 laser pulses for reliable depth measurements. These outputs are fed into the FPGA-embedded LiTOFNet for real-time inference. For comparison, we evaluate existing HBDE techniques, which include state-of-the-art deep learning-based methods, processing the same data on a PC platform.

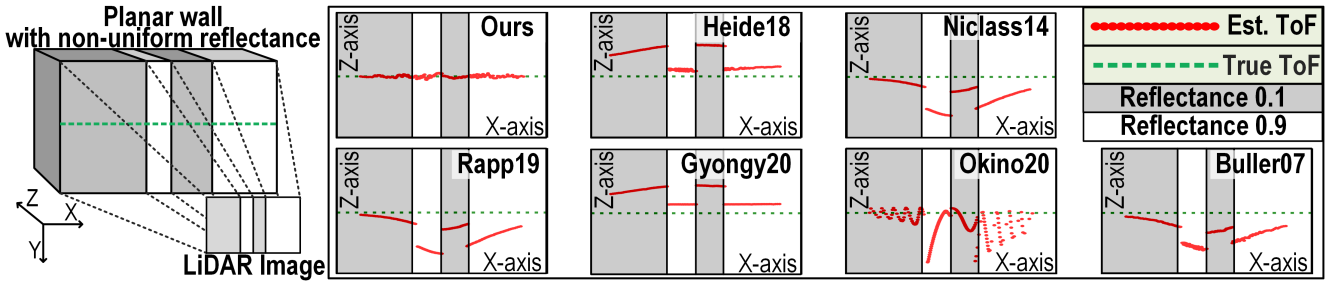


Figure 6: Depth estimation results on a planar wall with non-uniform reflectance.

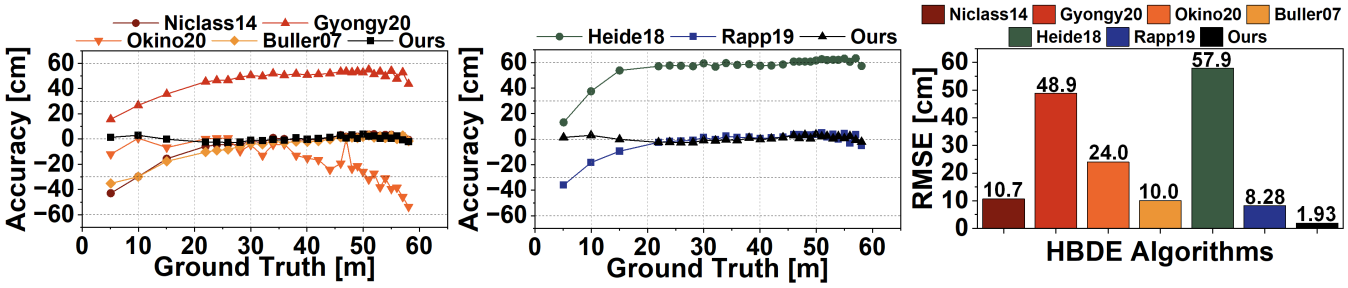


Figure 7: Performance comparison of HBDE algorithms and the proposed method: (Left) Traditional HBDE accuracy, (Middle) Pile-up correction HBDE accuracy, and (Right) RMSE comparison.

Evaluation Metrics

Depth measuring performance is evaluated using the standard root-mean-square error (RMSE) between the estimated depth and ground truth. This metric effectively captures systematic depth errors from pile-up distortion, making it suitable for measuring the resulting RMSE increase.

Pile-Up Evaluation

We simulated a LiDAR sensor facing a wall, using the central row with an angular resolution of 0.63° per horizontal pixel. The 192 simulated pixels spanned a width where the leftmost and rightmost points were 10 m from the sensor, placing the closest center point at approximately 4.93 m.

To evaluate performance under varying reflectance, we divided the wall into four sections with alternating 10% and 90% reflectance (first/third and second/fourth quarters). Given the wall's distance range and reflectance variations, the echo-photon flux spanned 1 to 37 photons/pulse, capturing moderate to severe pile-up. This platform produced 9× intensity changes at the boundaries, where abrupt transitions challenge algorithm linearity and expose pile-up distortions. Synthetic histograms were generated using the simulation framework described in the 'Data Augmentation' (page 3).

Figure 6 shows depth profiles across the wall presented as a top-view of the planar wall point cloud. When comparing algorithm performance, LiToFNet exhibited robustness against abrupt signal intensity variations, maintaining consistent accuracy without noticeable distortion at reflectance transition boundaries. The visualization shows that, unlike LiToFNet, other methods produced significant distortions in what should appear as a straight wall, with visible splitting and curvature at the reflectance boundaries and close-range high-reflectance regions where pile-up effects were most severe. These results confirm LiToFNet's learned intensity compensation capability, allowing it to maintain accuracy despite severe pile-up conditions.

Table 1 summarizes the depth estimation accuracy under non-uniform reflectance for LiToFNet and state-of-the-art HBDE algorithms. Our method achieves the lowest RMSE of 0.023 m with minimal bias -0.002 m, significantly outperforming all comparison methods that exhibit RMSE values ranging from 0.248 to 0.344 m.

Real-World Validation

Daytime single-point measurement. To assess daytime depth performance across different ranges, we conducted a single-point sweep using a 1 m×1 m Lambertian target (10% reflectance), swept from 5 to 58 m under 94.6 klux daylight. Absolute distance was referenced with a Bosch GLM-150C range finder on the same optical axis. Figure 7 shows depth accuracy versus distance over 1000 acquisitions per point.

Traditional HBDE algorithms (Figure 7 (left)) degrade rapidly below 15 m due to severe signal-induced pile-up, exceeding 20 cm error at 6 m. Even pile-up correction HBDE algorithms (Figure 7 (middle)) showed limited performance in the near-field range, as existing pile-up compensation methods primarily address noise-induced pile-up or assume narrow laser full width at half maximum (FWHM), failing to compensate for signal-induced pile-up effects from broader pulse widths. In contrast, LiToFNet maintains accuracy across all distance ranges, demonstrating robust performance, particularly in the challenging below-15 m region where other methods underestimate distances due to negative bias from uncompensated pile-up distortion. Figure 7

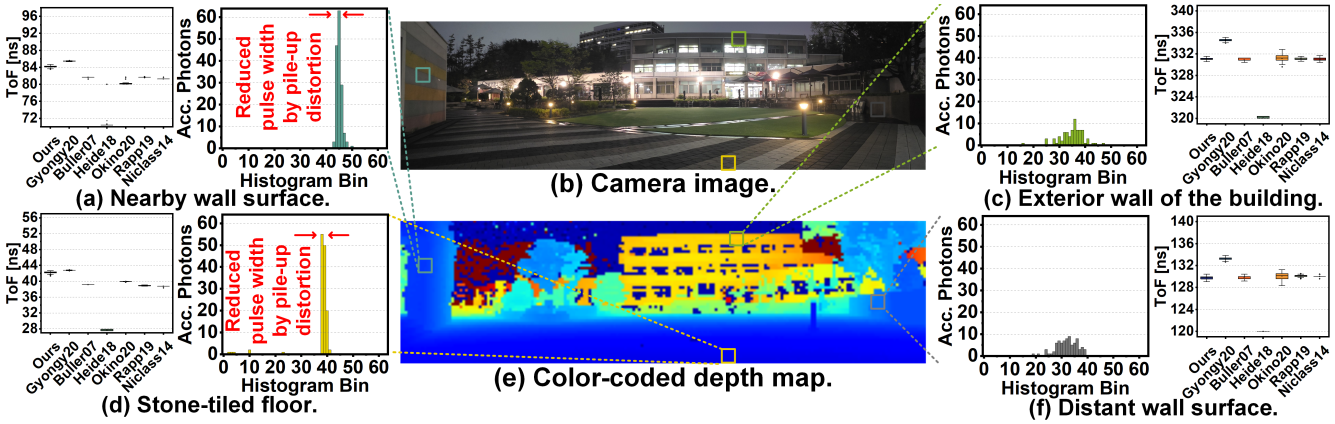


Figure 8: Commercial LiDAR sensor results with fine histograms and 1000 times ToF statistics.

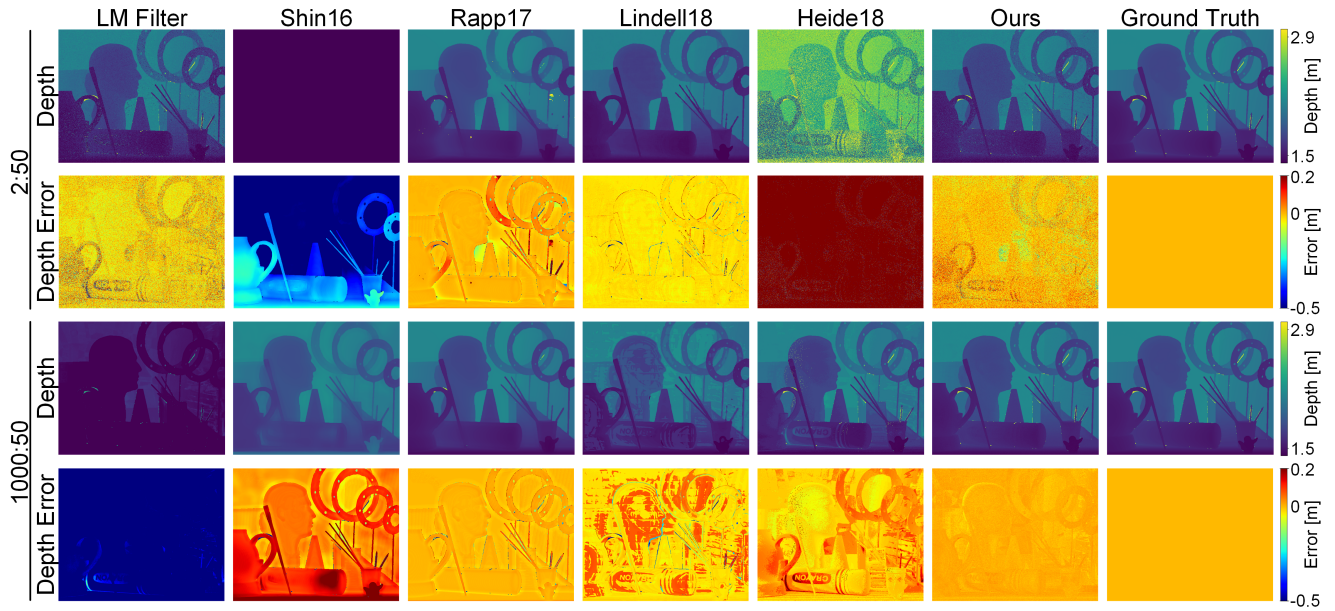


Figure 9: Depth estimation performance comparison on the Middlebury dataset: prior works versus the proposed LiToFNet with ground truth error analysis.

(right) shows that our method achieves superior RMSE of 0.0193 m compared to other approaches (0.0828–0.579 m), representing over 4.29× improvement in accuracy.

Nighttime outdoor measurement. We validated LiToFNet using real-world data captured at nighttime across varying distances (up to 60 m) and reflectance scenarios. We specifically analyzed four representative points exhibiting distinct characteristics. In Figure 8, points (a) and (d), located at shorter ranges (approximately 12 m, respectively), showed significant pulse-width compression due to severe pile-up. Most existing HBDE algorithms notably underestimated these distances, confirming their vulnerability to pile-up. In contrast, LiToFNet maintained accurate estimations under these challenging conditions. Points (c) and (f) demonstrated reflectance effects: despite having similar histogram intensities, they differed considerably in distance (approximately 50 m and 20 m, respectively), highlighting the complexity of predicting depth from signal intensity alone. Realistic testing shows that other methods exhibit the same distance-underestimation bias seen under strong pile-up in Figure 7. The consistent performance of LiToFNet in Figure 8, without such systematic bias, confirms its immunity to pile-up effects and superior accuracy in challenging near-field scenarios.

Middlebury Dataset Evaluation

We conducted a comprehensive evaluation using the Middlebury dataset (Scharstein and Pal 2007; Hirschmuller and Scharstein 2007) to compare LiToFNet against state-of-the-art HBDE methods. While previous studies model laser pulses as narrow Gaussian functions (hundreds of ps FWHM), automotive LiDAR systems (Niclass et al. 2014; Seo et al. 2021) use broad pulses (> 4 ns FWHM) to achieve high power output. Our data augmentation framework (in

Avg. Photons	Avg. BG (SBR)	LM Filter RMSE[cm]	Shin16 RMSE[cm]	Rapp17 RMSE[cm]	Lindell18 RMSE[cm]	Heide18 RMSE[cm]	**Ours** RMSE[cm]
2	2 (1)	8.71	52.4	9.61	7.80	55.4	**5.88**
5	2 (2.5)	6.23	54.2	8.31	7.75	53.7	**3.99**
10	2 (5)	8.18	46.1	7.37	7.68	51.3	**3.25**
2	10 (0.2)	9.48	52.3	20.5	7.83	55.7	**6.50**
5	10 (0.5)	6.48	54.5	6.68	7.79	53.6	**4.19**
10	10 (1)	8.35	46.1	5.83	7.72	51.1	**3.33**
2	50 (0.04)	14.7	52.1	35.9	**9.69**	71.6	9.79
5	50 (0.1)	7.54	54.2	10.9	7.95	57.3	**5.05**
10	50 (0.2)	9.11	46.4	9.08	7.79	52.3	**3.77**
20	50 (0.4)	14.7	42.8	5.96	7.48	47.1	**3.14**
50	50 (1)	27.2	31.3	5.89	7.89	36.3	**2.65**
100	50 (2)	36.8	24.0	5.97	8.48	27.1	**2.44**
1000	50 (20)	55.6	16.0	5.78	8.90	9.17	**2.21**

Table 2: Depth estimation performance comparison showing RMSE (cm) results under various signal-noise conditions. We mark the best/second-best results in **bold**/underlined.

Pulse Net.	**2nd Conv. Layer**	**Hidden Layer**	**RMSE** [cm]
✓	✓	✓	0.695
✗	✓	✓	0.756
✓	✗	✓	0.718
✓	✓	✗	2.360

Table 3: Component-wise ablation results for LITOFNET architecture. All RMSE values are in centimeters.

'Data Augmentation' (page 3)) utilizes realistic broad pulse shapes from actual TX modules. Due to data format incompatibilities, we used internally generated data for comparison: Lindell, O'Toole, and Wetzstein (2018) was trained on 1024-bin histograms while our dataset uses 64 bins, and Rapp and Goyal (2017) operates on continuous time-domain data with binary timestamps rather than sampled histograms. Other methods were evaluated using our broad pulse histograms. Unfortunately, Peng et al. (2023) and Yu et al. (2025) are excluded from our comparisons because their official code is unavailable, despite our inquiry to the authors.

To evaluate performance under varying signal and noise conditions, we tested signal-to-background noise combinations from established Yu et al. (2025) protocols (10:2, 5:2, 2:2, 10:10, 5:10, 2:10, 10:50, 5:50, 2:50) and extended to extremely high-signal scenarios (20:50, 50:50, 100:50, 1000:50). Our extended evaluation with high absolute signal levels under severe noise conditions enables proper assessment of signal-induced pile-up distortion in realistic high-intensity LiDAR scenarios.

Figure 9 shows algorithm performance under two extreme conditions on the Art scene from the Middlebury dataset: the lowest SBR and the most severe signal-induced pile-up, with error maps indicating deviations from ground truth. Since Rapp and Goyal (2017) and Lindell, O'Toole, and Wetzstein (2018) are designed for narrow-pulse histograms, we use narrow pulses for them in our evaluation—a choice that makes them appear immune to pile-up and obscures limitations that emerge with broad pulses. On broad pulse histograms, Heide et al. (2018) and Shin et al. (2016) perform better at higher signals where pile-up aligns with their narrow-pulse assumptions, while the log-matched filter performs better at lower signals without pile-up. LITOFNET achieves superior and stable results across all conditions.

	`bfloat16`	`float32`	**Fixed Point 8**
RMSE [cm]	2.364	0.695	26.393

Table 4: Impact of data precision on accuracy and precision.

HBDE algorithm	**Runtime**	**HBDE algorithm**	**Runtime**
LITOFNET (CPU)	2.72	Okino et al. (2020)	0.0217
LITOFNET (GPU)	0.333	Gyongy et al. (2020)	0.130
LITOFNET (FPGA)	8.40e-4	Niclass et al. (2014)	0.992
Buller and Wallace (2007)	1.07	Rapp et al. (2019)	28.0
Heide et al. (2018)	6.85		

Table 5: Comparison of per-pixel runtime (in milliseconds) for depth estimation.

Table 2 summarizes the comprehensive evaluation results averaged across all eight Middlebury scenes for various signal-noise combinations. LITOFNET demonstrates both consistency and superior accuracy across all tested conditions, achieving the lowest RMSE values with minimal variation between different signal scenarios. Notably, even under the most severe pile-up condition (1000:50), our method maintains excellent performance with an RMSE of 2.21 cm.

Ablation Study

The ablation study presented in Tables 3 and 4 validates our architectural design and data precision. As shown in Table 3, removing any component, such as the second convolutional layer or hidden layers, resulted in significant accuracy degradation, confirming the necessity of the full model configuration. Furthermore, Table 4 demonstrates that `bfloat16` offers the optimal trade-off between accuracy (2.364 cm RMSE) and hardware efficiency. While `float32` yields slightly higher accuracy, its overhead is prohibitive, and 8-bit quantization leads to unacceptable error levels.

Runtime Analysis

We evaluated LITOFNET's per-pixel runtime against conventional HBDE algorithms (Table 5). On a CPU (Intel i9-14900k), it ran in 2.72 ms—slightly slower than native TDC methods but faster than computationally intensive algorithms (Rapp et al. 2019; Heide et al. 2018)—while maintaining superior accuracy. GPU acceleration (NVIDIA RTX 4090) reduced runtime to 0.33 ms, and FPGA implementation (Xilinx Kintex-7) achieved 0.84 μs per pixel, confirming its suitability for real-time embedded deployment.

Conclusions

We introduced a paradigm shift from signal filtering to lightweight similarity learning for histogram-based LiDAR depth estimation. Our approach enables a compact network to learn a metric inherently robust to pile-up. Validated on an FPGA, the resulting 57.61 KB model is over 215.2$\times$ smaller than prior works and achieves real-time inference (106.27 fps). It maintains superior accuracy (2.21 cm RMSE) even in severe pile-up scenarios, successfully bridging the gap between high performance and on-device deployability.

Acknowledgments

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT) (RS-2023-00242528, RS-2024-00408040), BK21 FOUR (Department of Intelligence Semiconductor Engineering, Ajou University), and Ajou University research fund. The EDA tool was supported by the IC Design Education Center (IDEC), Korea.

References

Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; and Koyama, M. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2623–2631.

Altmann, Y.; Ren, X.; McCarthy, A.; Buller, G. S.; and McLaughlin, S. 2016. Lidar Waveform-Based Analysis of Depth Images Constructed Using Sparse Single-Photon Data. *IEEE Transactions on Image Processing*, 25(5): 1935–1946.

Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; and Shah, R. 1993. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6.

Buller, G. S.; and Wallace, A. M. 2007. Ranging and Three-Dimensional Imaging Using Time-Correlated Single-Photon Counting and Point-by-Point Acquisition. *IEEE Journal of Selected Topics in Quantum Electronics*, 13(4): 1006–1015.

Choi, D.; Cho, W.; Kim, K.; and Choo, J. 2023. iDet3D: Towards Efficient Interactive Object Detection for LiDAR Point Clouds. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI'24)*, 1433–1441. IEEE/CVF.

Chopra, S.; Hadsell, R.; and LeCun, Y. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, 539–546. IEEE.

Chugunov, I.; Baek, S.-H.; Fu, Q.; Heidrich, W.; and Heide, F. 2021. Mask-ToF: Learning Microlens Masks for Flying Pixel Correction in Time-of-Flight Imaging. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'21)*, 9112–9122. IEEE/CVF.

Cova, S.; Ghioni, M.; Lacaita, A.; Samori, C.; and Zappa, F. 1996. Avalanche photodiodes and quenching circuits for single-photon detection. *Applied Optics*, 35(12): 1956–1976.

Godard, C.; Aodha, O. M.; Firman, M.; and Brostow, G. 2019. Digging Into Self-Supervised Monocular Depth Estimation. In *Proceedings of the IEEE international conference on computer vision*, 3828–3838.

Grull, F.; Kirchgessner, M.; Kaufmann, R.; Hausmann, M.; and Kebschull, U. 2011. Accelerating Image Analysis for Localization Microscopy with FPGAs. In *2011 21st International Conference on Field Programmable Logic and Applications*, 1–5.

Gupta, A.; Ingle, A.; Velten, A.; and Gupta, M. 2019. Photon-flooded single-photon 3d cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6770–6779.

Gyongy, I.; Hutchings, S. W.; Halimi, A.; Tyler, M.; Chan, S.; Zhu, F.; McLaughlin, S.; Henderson, R. K.; and Leach, J. 2020. High-speed 3D sensing via hybrid-mode imaging and guided upsampling. *Optica*, 7(10): 1253–1260.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, 1026–1034.

Heide, F.; Diamond, S.; Lindell, D. B.; and Wetzstein, G. 2018. Sub-picosecond photon-efficient 3D imaging using single-photon sensors. *Scientific reports*, 8(1): 17726.

Hirschmuller, H.; and Scharstein, D. 2007. Evaluation of cost functions for stereo matching. In *2007 IEEE conference on computer vision and pattern recognition*, 1–8. IEEE.

Hu, M.; Wang, S.; Li, B.; Ning, S.; Fan, L.; and Gong, X. 2021. Penet: Towards precise and efficient image guided depth completion. In *IEEE International Conference on Robotics and Automation (ICRA'21)*, 13656–13662.

Kingma, D.; and Ba, J. 2014. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations*.

Klinger, M.; Termohlen, J.-A.; Mikolajczyk, J.; and Fingscheidt, T. 2020. Self-Supervised Monocular Depth Estimation: Solving the Dynamic Object Problem by Semantic Guidance. In *European Conference on Computer Vision (ECCV'20)*, 582–600.

Koch, G.; Zemel, R.; Salakhutdinov, R.; et al. 2015. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2, 1–30. Lille.

Lee, J.; Ingle, A.; Chacko, J. V.; Eliceiri, K. W.; and Gupta, M. 2023. CASPI: collaborative photon processing for active single-photon imaging. *Nature Communications*, 14(1): 3158.

Li, Z.; Dong, Q.; Saligane, M.; Kempke, B.; Gong, L.; Zhang, Z.; Dreslinski, R.; Sylvester, D.; Blaauw, D.; and Kim, H.-S. 2018. A 1920×1080 30-frames/s 2.3 TOPS/W stereo-depth processor for energy-efficient autonomous navigation of micro aerial vehicles. *IEEE Journal of Solid-State Circuits*, 53(1): 76–89.

Lindell, D. B.; O'Toole, M.; and Wetzstein, G. 2018. Single-photon 3d imaging with deep sensor fusion. *ACM Trans. Graph.*, 37(4): 113.

Lyu, X.; Liu, L.; Wang, M.; Kong, X.; Liu, L.; Liu, Y.; Chen, X.; and Yuan, Y. 2021. HR-Depth: High Resolution Self-Supervised Monocular Depth Estimation. In *Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI'21)*, 2294–2301. IEEE/CVF.

Niclass, C.; Favi, C.; Kluter, T.; Gersbach, M.; and Charbon, E. 2008. A 128×128 single-photon image sensor with column-level 10-bit time-to-digital converter array. *IEEE Journal of Solid-State Circuits*, 43(12): 2977–2989.

Niclass, C.; Rochas, A.; Besse, P.-A.; and Charbon, E. 2005. Design and characterization of a CMOS 3-D image sensor based on single photon avalanche diodes. *IEEE Journal of Solid-State Circuits*, 40(9): 1847–1854.

Niclass, C.; Soga, M.; Matsubara, H.; Ogawa, M.; and Kagami, M. 2014. A 0.18-μm CMOS SoC for a 100-m-range 10-frame/s 200×96-pixel time-of-flight depth sensor. *IEEE Journal of Solid-State Circuits*, 49(1): 315–330.

Okino, T.; Yamada, S.; Sakata, Y.; Kasuga, S.; Takemoto, M.; Nose, Y.; Koshida, H.; Tamaru, M.; Sugiura, Y.; Saito, S.; Koyama, S.; Mori, M.; Hirose, Y.; Sawada, M.; Odagawa, A.; and Tanaka, T. 2020. A 1200×900 6μm 450fps Geiger-Mode Vertical Avalanche Photodiodes CMOS Image Sensor for a 250m Time-of-Flight Ranging System Using Direct-Indirect-Mixed Frame Synthesis with Configurable-Depth-Resolution Down to 10cm. In *IEEE International Solid-State Circuits Conference (ISSCC)*, 96–98.

Park, J.; Joo, K.; Hu, Z.; Liu, C.-K.; and Kweon, I. S. 2020. Non-local spatial propagation network for depth completion. In *European Conference on Computer Vision (ECCV'20)*, 120–136.

Peng, J.; Xiong, Z.; Tan, H.; Huang, X.; Li, Z.-P.; and Xu, F. 2023. Boosting Photon-Efficient Image Reconstruction With A Unified Deep Neural Network. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4): 4180–4197.

Qiu, J.; Cui, Z.; Zhang, Y.; Zhang, X.; Liu, S.; Zeng, B.; and Pollefeys, M. 2019. DeepLiDAR: Deep Surface Normal Guided Depth Prediction for Outdoor Scene from Sparse LiDAR Data and Single Color Image. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19)*, 3313–3322.

Rapp, J.; and Goyal, V. K. 2017. A Few Photons Among Many: Unmixing Signal and Noise for Photon-Efficient Active Imaging. *IEEE Transactions on Computational Imaging*, 3(3): 445–459.

Rapp, J.; Ma, Y.; Dawson, R. M. A.; and Goyal, V. K. 2019. Dead Time Compensation for High-Flux Ranging. *IEEE Transactions on Signal Processing*, 67(13): 3471–3486.

Roh, W.; Seo, H.; Piao, C.; Lee, H.; Kim, M.; Lee, M.-J.; Kim, S.-J.; Chun, J.-H.; and Choi, J. 2024. A 400× 112 CMOS LiDAR Sensor with Reconfigurable-Resolution Histogramming Time-to-Digital Converters and Sub-cm Depth Refining Filter. In *2024 IEEE Asian Solid-State Circuits Conference (A-SSCC)*, 1–3. IEEE.

Scharstein, D.; and Pal, C. 2007. Learning conditional random fields for stereo. In *2007 IEEE conference on computer vision and pattern recognition*, 1–8. IEEE.

Seo, H.; Yoon, H.; Kim, D.; Kim, J.; Kim, S.-J.; Chun, J.-H.; and Choi, J. 2021. Direct TOF scanning LiDAR sensor with two-step multievent histogramming TDC and embedded interference filter. *IEEE Journal of Solid-State Circuits*, 56(4): 1022–1035.

Shin, D.; Xu, F.; Venkatraman, D.; Lussana, R.; Villa, F.; Zappa, F.; Goyal, V. K.; Wong, F. N. C.; and Shapiro, J. H. 2016. Photon-efficient imaging with a single-photon camera. *Nature Communications*, 7(1): 12046.

Su, S.; Heide, F.; Wetzstein, G.; and Heidrich, W. 2018. Deep End-to-End Time-of-Flight Imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 6383–6392.

Umasuthan, M.; Wallace, A. M.; Massa, J. S.; Buller, G. S.; and Walker, A. C. 1998. Processing time-correlated single photon counting data to acquire range images. *IEE Proceedings - Vision, Image and Signal Processing*, 145(4): 237–245.

Xu, Y.; Zhu, X.; Shi, J.; Zhang, G.; Bao, H.; and Li, H. 2019. Depth completion from sparse lidar data with depth-normal constraints. In *Proceedings of the IEEE international conference on computer vision*, 2811–2820.

Yang, Y.; Wong, A.; and Soatto, S. 2019. Dense depth posterior (DDP) from single image and sparse range. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'19)*, 3353–3362.

Yu, L.; Yang, J.; Dong, B.; Bao, Q.; Wang, Y.; Heide, F.; Wei, X.; and Yang, X. 2025. Separating the Wheat from the Chaff: Spatio-Temporal Transformer with View-interweaved Attention for Photon-Efficient Depth Sensing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(9): 9626–9634.

Zhang, L.; Yang, A. J.; Xiong, Y.; Casas, S.; Yang, B.; Ren, M.; and Urtasun, R. 2023. Towards Unsupervised Object Detection from LiDAR Point Clouds. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'23)*, 9317–9328. IEEE/CVF.

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

Zhou, Z.; Fan, X.; Shi, P.; and Xin, Y. 2021. R-MSFM: Recurrent Multi-Scale Feature Modulation for Monocular Depth Estimating. In *Proceedings of the IEEE international conference on computer vision*, 12777–12786.

Zhuo, S.; Xia, T.; Zhao, L.; Sun, M.; Wu, Y.; Wang, L.; Yu, H.; Xu, J.; Wang, J.; Lin, Z.; Li, Y.; Qiu, L.; Bai, R.; Chen, X.; and Chiang, P. Y. 2023. Solid-State dToF LiDAR System Using an Eight-Channel Addressable, 20-W/Ch Transmitter and a 128×128 SPAD Receiver With SNR-Based Pixel Binning and Resolution Upscaling. *IEEE Journal of Solid-State Circuits*, 58(3): 757–771.