Learning to Suppress SPAD-based LiDAR Flare

Xuanya Zhu[1]*, Linghao Shen[2(✉)]

[1] Centre for Vision, Speech and Signal Processing, University of Surrey, UK
[2] Advanced Technology Center, Sony (China) Ltd., China
marka.shen@sony.com

Abstract. Single-Photon Avalanche Diode (SPAD)-based Light Detection and Ranging (LiDAR) is emerging for autonomous vehicles due to its high sensitivity and precise depth sensing capabilities. However, flare caused by excessive photon returns or pile-up effects can lead to incorrect depth estimation and exaggerated boundaries in point clouds, resulting in severe distortions of geometric measurements, making flare suppression essential for safety-critical applications. Existing flare mitigation methods primarily operate at the hardware or signal-processing levels. While effective under specific configurations, they are largely rule-based and configuration-dependent, lacking learnable representations that generalize across diverse sensing scenarios. In this work, we reformulate flare suppression as a semantic segmentation problem, enabling data-driven learning of geometric and photometric cues directly from SPAD measurements. We first benchmark representative segmentation models on the newly introduced SPAD flare dataset and observe that they struggle to exploit the intrinsic multi-echo characteristics of SPAD signals. Motivated by this observation, we propose **Physically-Informed segmentation for LiDAR Flare (PILF)**, a learning-based approach that treats the first and second echoes, together with ambient illumination, as distinct modalities, aggregating cross-echo information while jointly encoding geometric and photometric features. Experiments across multiple real-world scenes demonstrate that PILF significantly outperforms compared segmentation models, achieving up to **79.32% mIoU**, and providing an effective solution for SPAD-based LiDAR flare suppression.

Keywords: Flare Suppression · SPAD LiDAR · Semantic Segmentation

1 Introduction

LiDAR has become a cornerstone sensor for autonomous driving and robotics. Among various implementations, SPAD-based LiDAR systems are emerging due to their high sensitivity and precise depth sensing capabilities [20, 32]. However, these systems are highly susceptible to flare caused by excessive photon arrivals or pile-up effects [10, 17], which can distort geometric shapes and produce spurious obstacles, as illustrated in Figure 1. Existing methods for suppressing flare in SPAD LiDAR primarily operate at the hardware or signal-processing

* Work done during an internship at Advanced Technology Center, Sony (China) Ltd.

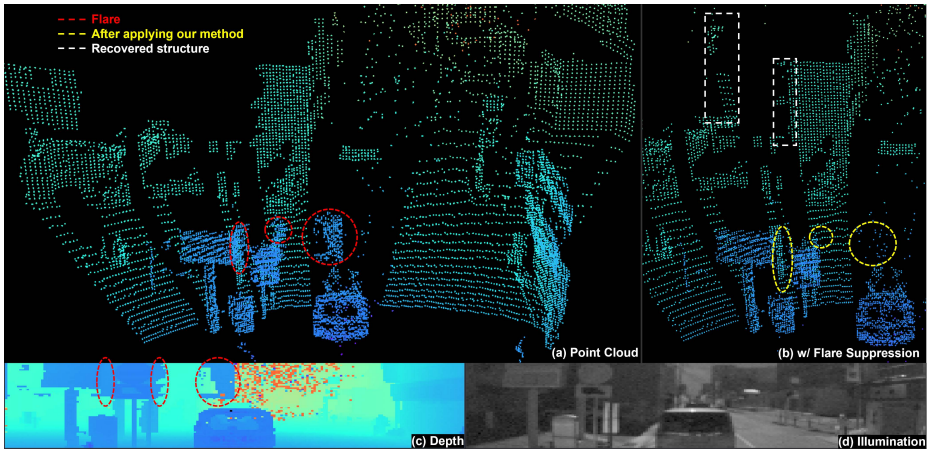

Fig. 1: Illustration of SPAD LiDAR flare effects in point clouds. (a) The captured 3D point cloud. (b) The point cloud after flare suppression. (c) The depth map. (d) The ambient illumination image. Excessive photon returns from retro-reflective surfaces, such as traffic signs, can produce strong flare that leads to false or exaggerated structures in the point cloud, potentially distorting geometric measurements. This highlights the importance of flare suppression for safety-critical autonomous systems.

levels [17, 29]. These non-learnable approaches, while effective in controlled settings, rely on rule-based procedures and are largely dependent on specific system configurations, limiting their generalization across different sensing scenarios.

To address these limitations, we reformulate flare suppression as a semantic segmentation task, enabling data-driven learning of geometric and photometric cues directly from SPAD measurements. We first benchmark representative segmentation models on the newly introduced `FLARE` dataset, revealing that they struggle to accurately distinguish flare from valid structures, particularly along boundaries where flare closely adjoins real objects. We attribute this limitation primarily to their inability to leverage the unique multi-echo characteristics of SPAD signals. Specifically, in typical processing strategies [13, 33], the first-echo depth and intensity are retained when converting raw SPAD signals into point clouds, while later echoes, which often provide complementary depth information for regions corrupted in the first echo, are discarded, thereby limiting effective discrimination between flare and valid structures.

Building on this insight, we propose **PILF**, a learning-based approach that directly leverages raw SPAD measurements to exploit their inherent multi-echo and photometric information. It treats the first and second echoes, together with ambient illumination, as distinct modalities, aggregating cross-echo information while jointly encoding geometric and photometric features.

Our **contributions** are threefold. *First*, we reformulate SPAD LiDAR flare suppression as a semantic segmentation task. *Second*, we design a physically informed multi-echo representation and fusion framework for distinguishing flare

from valid structures. *Third*, we release `FLARE`, a SPAD LiDAR dataset covering diverse real-world flare scenes, and provide systematic evaluation against representative segmentation baselines.

2 Related Work

SPAD-based LiDAR Flare Suppression. Flare in SPAD-based LiDAR systems primarily arises from excessive photon returns and pile-up effects, which induce optical cross-talk among neighboring SPAD pixels and produce saturated, spatially spread false returns that distort depth estimation and scene geometry. Traditional flare suppression methods primarily operate at the hardware or signal-processing levels [16, 36]. Hardware approaches mitigate detector non-idealities through sensor design and calibration, such as optical isolation, trench structures, on-chip time gating, and advanced readout architectures [9,10,17,35]. Signal-processing approaches suppress distortions from raw SPAD measurements or histograms using techniques such as peak suppression, temporal gating, photon modeling, and asynchronous acquisition [3, 11, 12, 26]. While effective under controlled conditions, these approaches are largely configuration-dependent, offering limited adaptability and lacking a unified framework for integrating physical cues [3]. Recent learning-based waveform methods study related artifacts, such as ghosts [15], assuming dense temporal signals [31], whereas we target a practical bandwidth- and latency-constrained setting where full histograms are typically unavailable to downstream perception modules.

Learning-based Semantic Segmentation. Recent advances in semantic segmentation have significantly improved performance in both 2D and 3D scene understanding. In the 2D domain, segmentation has evolved from convolution-based architectures [4,23,30] to transformer-based frameworks [21,38], achieving strong spatial reasoning and contextual modeling. However, these methods are typically designed for RGB or single-modality images, lacking access to geometric or temporal cues that are crucial in LiDAR sensing. In the 3D domain, segmentation approaches learn spatial semantics across various representations, including point-based [27,28,37], voxel-based [6,22,41], and range-view formats [1,8,18,24]. While effective at capturing geometric context, these methods generally overlook the physical principles underlying point cloud formation, such as multi-echo photon distributions, which are crucial for distinguishing flare from valid structures. This limits their direct applicability to SPAD flare suppression.

3 Methodology

3.1 Physics-Guided Input Representation

Conventional Data Processing. In a typical LiDAR scan, time-resolved SPAD measurements are collected for each laser pulse. Commonly, the first echo is retained and converted into a 3D point cloud [3], which can then be projected

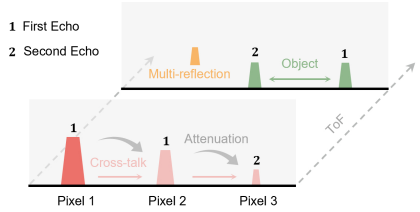

Fig. 2: Illustration of SPAD flare formation. Excessive photons cause flare to occupy the first echo in neighboring pixels, while valid returns appear in later echoes.

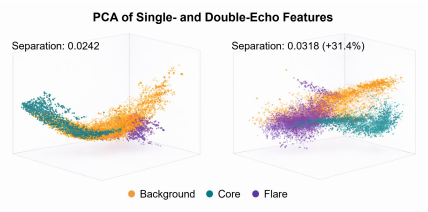

Fig. 3: PCA-based 3D scatter plots showing up to 10,000 points per class. Utilizing both first and second echoes improves class separability by approximately 30%.

onto a 2D cylindrical range view of resolution $H \times W$, where H and W denote the height and width of the range view. Each point (p_x, p_y, p_z) is mapped to its corresponding pixel (i, j) as follows:

$$\begin{pmatrix} i \\ j \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\left[1 - \arctan(p_y, p_x)\pi^{-1}\right] W \\ \left[1 - \left(\arcsin(p_z, r^{-1}) + \phi_{down}\right)\xi^{-1}\right] H \end{pmatrix}, \tag{1}$$

where $r = \sqrt{p_x^2 + p_y^2 + p_z^2}$ denotes the point depth, and $\xi = |\phi_{down}| + |\phi_{up}|$ represents the vertical field-of-view of the sensor. By retaining only the first echo, conventional processing discards subsequent echoes that may contain valuable information related to the flare. This limitation motivates leveraging multi-echo signals to form a richer, multi-modal representation.

Multi-Echo Characteristics. Figure 2 illustrates that excessive photons from a retro-reflective surface, defined as the core region, are received by SPAD pixel 1, inducing optical crosstalk that oversaturates neighboring pixels 2. In this region, the flare intensity is significantly higher than that of valid object returns, forcing the true signal to appear in the second echo. As the distance from the reflective source increases to pixel 3, the flare intensity gradually decays below the object signal, allowing the valid response to reappear in the first echo. Excessive returns at pixel 1 may also trigger multi-reflections, which generate secondary interference echoes at approximately twice the path length.

These effects corrupt the first-echo measurements, especially near core regions, while later echoes often preserve complementary geometric-photometric cues for distinguishing flare from valid structures. Figure 3 shows PCA-based 3D scatter plots, where using both echoes improves class separability by around 30%, highlighting the benefit of multi-echo information.

Multi-Modal Input Construction. Based on these observations, we leverage the first two echoes of the SPAD LiDAR signal, each providing three physical measurements: *intensity*, *depth*, and *pulse width*. Intensity represents reflected photon

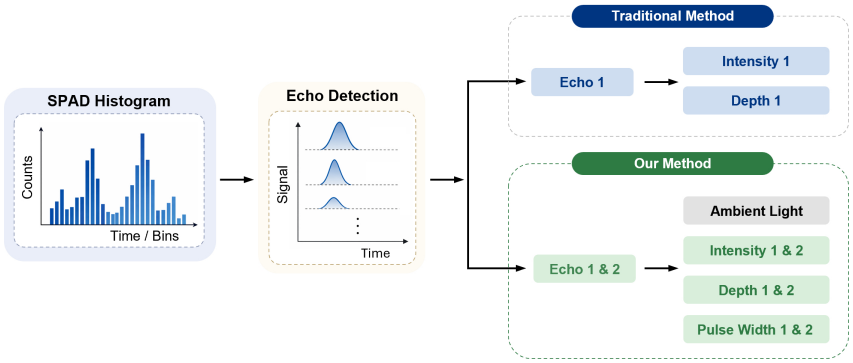

Fig. 4: Comparison of conventional and proposed input construction pipelines. Traditional processing retains only the first echo. Our method selects two echoes and incorporates ambient illumination to construct a 7-channel representation.

counts, depth corresponds to time of flight, and pulse width measures the temporal spread of the returned pulse. We further approximate flare as an excessive illumination component and introduce an ambient channel to help separate flare responses from intrinsic surface reflectivity [19].

Accordingly, our input is a multi-modal 7-channel range-view image. As illustrated in Figure 4, conventional SPAD LiDAR processing often relies on selected first-echo attributes for downstream representation. In contrast, our method uses the first two echoes together with ambient illumination and pulse-width information to form a richer multi-modal representation. Due to the fixed SPAD scanning layout, the measurements naturally form a dense 2D grid aligned with angular coordinates, preserving native spatial resolution without additional resampling. The resulting representation remains physically interpretable while providing richer signal diversity for learning-based flare suppression.

3.2 Physically-Informed Flare Segmentation

As shown in Figure 5, PILF consists of two branches incorporating physical priors. The single-echo branch focuses on the first echo x_1, while the multi-echo branch takes both the first x_1 and second x_2 echoes as input by applying Depth Decomposition (DD). Both branches process the same underlying physical properties and therefore share the Physics-Aware Encoder (PAE), which extracts geometric-photometric representations separately for each modality. With ambient illumination x_0, the resulting features are aggregated by the Echo-Aware Fusion (EAF), which adaptively balances complementary information from the first and second echoes to form a unified representation. This representation is then fed into the backbone to predict a mask with three classes: flare, core (its corresponding source), and background.

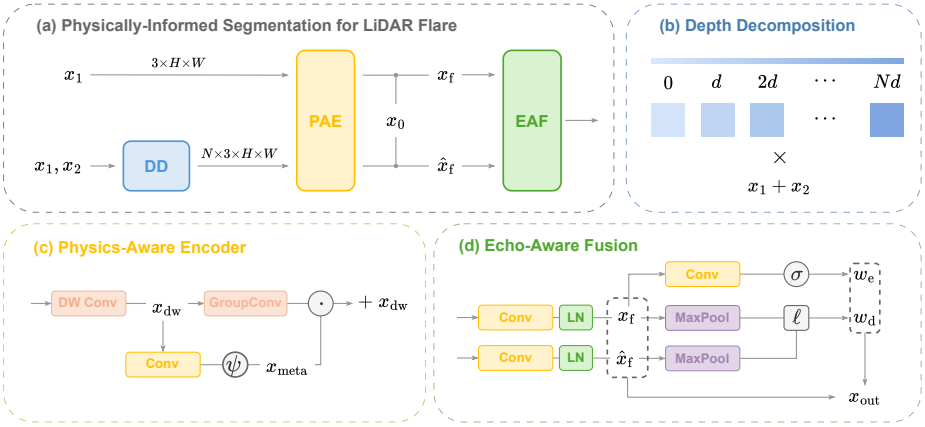

Fig. 5: Overview of PILF. (a) The overall architecture processes first echo x_1, second echo x_2, and ambient illumination x_0 through two branches. (b) Depth Decomposition (DD) organizes multi-echo signals into depth-aware bins. (c) Physics-Aware Encoder (PAE) extracts modality-aware geometric-photometric features. (d) Echo-Aware Fusion (EAF) adaptively aggregates echo and illumination cues for segmentation.

Depth Decomposition. Multi-echo SPAD signals often contain substantial noise along the depth dimension, which is randomly distributed and lacks spatial continuity, whereas valid measurements are primarily obtained from the first echo. Fusing such noisy inputs directly complicates model learning. In contrast, flare and its corresponding core exhibit strong spatial continuity within specific depth ranges, and objects partially occluded in the first echo often reappear in subsequent echoes while maintaining consistent depth.

We discretize the depth space into N bins according to the depth range D derived from the first echo. This operation not only aggregates spatially coherent flare and object features across echoes but also disperses the randomly distributed noise, as illustrated in Figure 6. A binary mask for each bin n with depth range $d_n = \left(\frac{(n-1)\cdot D}{N}, \frac{n\cdot D}{N}\right)$ is defined as:

$$B^{(n)}(i,j) = \begin{cases} 1, & \text{if } d(i,j) \in d_n, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

Here, $d(i,j)$ denotes the depth at pixel (i,j). Within each bin n, masks $B_1^{(n)}$ and $B_2^{(n)}$ for the first and second echoes are applied to extract depth-aware features:

$$\hat{x}^{(n)} = B_1^{(n)} \odot x_1 + B_2^{(n)} \odot x_2, \quad n = 1, \ldots, N. \tag{3}$$

The resulting tensor $\hat{x} \in \mathbb{R}^{N\times 3\times H\times W}$ aggregates features within each depth bin.

Physics-Aware Encoder. In our range-view representation, intensity, depth, and pulse width share pixel coordinates but describe different aspects of the photon-return distribution along the same projection ray. Their responses are therefore

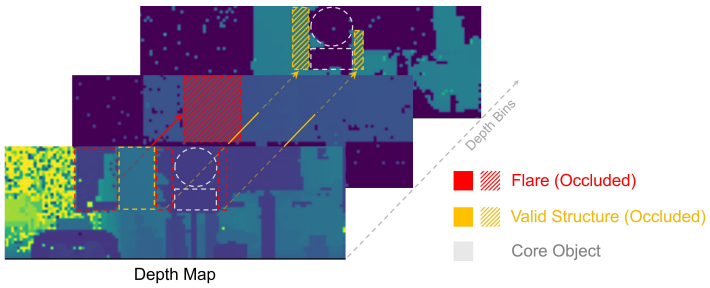

Fig. 6: Example of depth decomposition. Within a specific depth range, objects occluded in the second echo, as well as potential flare regions, appear more completely after being merged with the first echo.

not strictly aligned in 3D: intensity may integrate multiple reflections, depth may be biased by strong returns, and pulse width captures the temporal spread within a depth interval. Such modality-dependent responses introduce local geometric inconsistencies: neighboring pixels may correspond to different depths or 3D structures, and even the same surface point can produce inconsistent responses across modalities. For example, strong retro-reflections may broaden the pulse width or bias the estimated depth. Consequently, directly sharing convolutional weights across modalities can cause physical aliasing, entangle modality-specific geometric cues, and produce physically inconsistent features.

To address this, we propose PAE, which learns modality-specific spatial representations while maintaining cross-modal consistency. Given a three-modal input x, it first applies depth-wise convolution to extract independent modality-wise features, yielding x_{dw}. A suppressive modulation map is then computed by applying a convolution and the meta function $\psi(\cdot)$:

$$x_{\mathrm{meta}} = \psi\left(\underset{3\times3}{\mathrm{Conv}}\left(x_{\mathrm{dw}}\right)\right) \tag{4}$$

where $\psi(\cdot)$ mitigates physically unstable responses, such as over-strong reflections or depth outliers. Table 3 compares performance under different ψ formulations. Finally, modality-wise features are updated by a group convolution and combined with the original features through a residual gated transformation:

$$\mathrm{PAE}(x) = x_{\mathrm{dw}} + \underset{3\times3}{\mathrm{GroupConv}}\left(x_{\mathrm{dw}}\right) \odot x_{\mathrm{meta}}, \tag{5}$$

The outputs are denoted as $x_{\mathrm{echo}} \in \mathbb{R}^{C\times H\times W}$ for the first echo x_1, and $\{x_{\mathrm{d}}^{(n)}\}_{n=1}^{N}$ for the depth-decomposed features $\hat{x}$, which are stacked for following processing.

Echo-Aware Fusion. In flare scenes, multi-modal cues exhibit different reliability and statistics across spatial and depth dimensions, making direct concatenation suboptimal. First, the large dynamic range between flare and normal regions,

together with varying illumination and different valid depth bins, can lead to unstable batch-level feature statistics. Second, the first echo is reliable in non-flare regions, whereas later echoes provide complementary cues in flare-corrupted areas, making static fusion insufficient. Third, flare is spatially localized, typically emerging from high-intensity cores and spreading outward, which requires spatially adaptive suppression while preserving valid geometry.

To address these challenges, we propose EAF, which adaptively fuses first-echo and depth-aware features. Specifically, the illumination map x_0 is first embedded into x_{amb}, which provides contextual guidance for both streams:

$$\begin{aligned} x_{\text{f}} &= \text{Concat}(x_{\text{echo}}, x_{\text{amb}}), \\ \hat{x}_{\text{f}} &= \left[\text{Concat}(x_{\text{d}}^{(1)}, x_{\text{amb}}), \ldots, \text{Concat}(x_{\text{d}}^{(N)}, x_{\text{amb}})\right]. \end{aligned} \tag{6}$$

A convolution followed by LayerNorm is applied to stabilize per-sample feature statistics across samples, which is important for robustly fusing first-echo and depth-bin features. This operation updates $x_{\text{f}} \in \mathbb{R}^{C\times H\times W}$ and $\hat{x}_{\text{f}} \in \mathbb{R}^{N\times C\times H\times W}$. For simplicity, the following operations are described for each depth bin n.

To selectively integrate multi-echo features, EAF estimates both a depth-bin coefficient and a spatial gate. Adaptive max pooling first extracts global descriptors from updated first-echo feature x_{f} and each depth-aware feature $\hat{x}_{\text{f}}^{(n)}$.

$$z_{\text{f}} = \text{MaxPool}(x_{\text{f}}), \quad \hat{z}_{\text{f}}^{(n)} = \text{MaxPool}(\hat{x}_{\text{f}}^{(n)}), \tag{7}$$

In the depth dimension, these descriptors are combined to compute a bin-wise relevance coefficient, allowing the model to emphasize reliable depth bins. In the spatial dimension, a gate is generated from the first-echo feature x_{f}, which carries strong geometric cues, to adaptively modulate each location and reduce over-reliance on corrupted responses. The two coefficients are computed as:

$$w_{\text{d}}^{(n)} = \ell\left(\text{Concat}(z_{\text{f}}, \hat{z}_{\text{f}}^{(n)})\right), \quad w_{\text{e}} = \sigma\left(\underset{3\times3}{\text{Conv}}\,(x_{\text{f}})\right), \tag{8}$$

where $\ell(\cdot)$ denotes a learnable projection and $\sigma(\cdot)$ denotes the sigmoid activation.

Finally, fusion first aggregates echoes along the depth dimension for depth consistency and then applies spatial modulation for local coherence. This approximates the physical intuition of photon integration along depth before projection onto the range-view plane. The final representation is:

$$x_{\text{out}} = x_{\text{f}} + w_{\text{e}} \odot \sum_{n=1}^{N} \left(w_{\text{d}}^{(n)}\ \hat{x}_{\text{f}}^{(n)}\right), \tag{9}$$

where the scalar $w_{\text{d}}^{(n)}$ is broadcast over the channel and spatial dimensions, and $\odot$ denotes the Hadamard product.

Backbone Architecture. The fusion results are processed by a U-Net [30] equipped with a multi-head attention decoder [34]. Since flare propagation is generally

anisotropic and exhibits a dominant spreading direction, we constrain attention to a single principal direction (horizontal in our setup) to better capture this structured pattern. This directional restriction enhances long-range feature correlation while maintaining computational efficiency.

3.3 Mask-Based Flare Suppression

Given the predicted flare mask, we perform echo-level correction on the structured multi-echo representation parsed from the raw sensor stream. Each frame contains multiple echo entries per pixel, including temporal and amplitude measurements, while shared metadata such as calibration parameters and acquisition descriptors are preserved during suppression.

Flare predominantly manifests as a premature first photon return caused by multipath or internal reflections, which disrupts the expected echo ordering and corrupts depth consistency. Let M denote the predicted flare mask. For pixels where $M(i,j) = 1$, we locally replace the flare-contaminated first echo with the corresponding second echo:

$$E'_{i,j,0,:} = \begin{cases} E_{i,j,1,:}, & M(i,j) = 1, \\ E_{i,j,0,:}, & M(i,j) = 0. \end{cases} \tag{10}$$

After correction, the representation is written back to the original raw data structure while preserving the header and non-echo metadata, ensuring compatibility with downstream processing pipelines.

This strategy modifies only the flare-contaminated primary echo while retaining other metadata and auxiliary echo information. It is guided by photon arrival ordering in SPAD-based time-of-flight sensing rather than sensor-specific heuristics, allowing integration with hardware-level mitigation techniques.

4 Experiment

4.1 Experimental Setup

Dataset. We evaluate our method on the `FLARE` dataset, collected using the `IMX459` SPAD LiDAR. The dataset covers a wide range of driving and roadside scenes, including urban roads, tunnels, highways, road signs, vehicle rear reflectors, and traffic lights, providing varied flare patterns under different illumination conditions. All input channels are derived from raw SPAD measurements, without requiring any external imaging sensor.

Each frame is pixel-wise annotated by multiple trained annotators using synchronized intensity, depth, and pulse-width range-view maps. Flare regions are identified based on depth discontinuities and abnormal intensity patterns, while associated retro-reflective cores are marked as stable high-intensity regions. A final manual cross-check refines region boundaries, especially around depth transitions, to separate true object surfaces from flare artifacts. By explicitly labeling

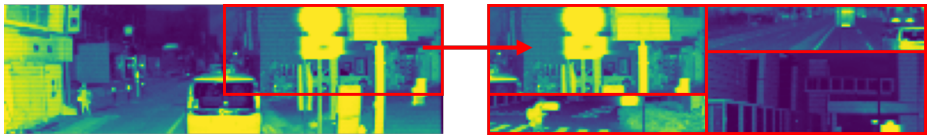

Fig. 7: A flare-core region is cropped from the source image and fused into another image while preserving the correspondence between flare and its associated core.

both flare and valid core regions, the dataset supports evaluation of flare suppression in terms of both semantic segmentation and structural fidelity.

For training, we use 90% of the main-road scenes, about 2,000 frames. The test set is a mixed set containing the remaining 10% of main-road frames and additional frames from other scenes, about 500 frames in total, to evaluate generalization under unseen conditions.

Normalization. We apply modality-specific normalization to the multi-modal input. For SPAD LiDAR, the measured intensity I, after scaling to $[0, 1)$, reflects photon counts rather than reflected energy E. We model its nonlinear response and apply the inverse mapping to approximate the corresponding energy scale:

$$E = -\frac{1}{\lambda}\log(1 - I). \tag{11}$$

Here, λ is a sensor-dependent response coefficient set by the hardware configuration and fixed in our experiments. The transformed intensity, depth, and pulse-width channels are min–max normalized using first-echo statistics, as both echoes observe the same spatial region with attenuation differences. The illumination channel is globally min–max normalized for consistent scaling.

Augmentation. The dataset exhibits severe class imbalance, with an approximate pixel ratio of background, flare, and core of 100 : 2 : 1. To address this, we adopt *FlareAug*, a SPAD-aware augmentation strategy consisting of two operations, crop and fuse, each applied with a probability of 75% during training.

The crop operation extracts local regions containing both flare and its corresponding core, increasing the frequency and diversity of flare patterns. The fuse operation resizes the cropped regions to a fixed resolution and inserts them into another training sample in a mosaic-style manner [2], increasing contextual variability, as illustrated in Figure 7. Based on SPAD array characteristics, flare rarely crosses the image midline and typically extends from its associated core along a dominant horizontal direction in our dataset. Accordingly, fusion is performed separately on the left and right halves, and each fused flare region is paired with its corresponding core to avoid ambiguous supervision.

Implementation Details. PILF is built based on `MMSegmentation` [7] and trained for 2×10^4 iterations using Adam, with a batch size of 16 and a learning rate of 1×10^{-3}. The module configurations are set as follows. DD uses $N = 16$

Table 1: Quantitative comparison of PILF with 2D, 3D, and RV segmentation methods. Each baseline follows its native input setting.

Type	Method	Input	Flare	Core	mIoU
2D	FCN [23]	$(I_{1,2}, D_{1,2}, P_{1,2}, L)$	57.08	69.22	63.15
	DeepLabV3+ [4]	$(I_{1,2}, D_{1,2}, P_{1,2}, L)$	67.72	76.56	72.14
	OCRNet [39]	(I_1, D_1, P_1)	63.21	75.34	69.28
		$(I_{1,2}, D_{1,2}, P_{1,2}, L)$	65.69	73.48	69.59
		(I_1, D_1, P_1)	63.85	72.89	68.37
	KNet [40]	$(I_{1,2}, D_{1,2}, P_{1,2}, L)$	57.36	68.22	62.79
	SegFormer [38]	$(I_{1,2}, D_{1,2}, P_{1,2}, L)$	52.26	69.92	61.09
	Mask2Former [5]	$(I_{1,2}, D_{1,2}, P_{1,2}, L)$	52.95	69.21	61.08
3D	PVCNN [22]	(x, y, z, I_1)	35.40	42.10	38.75
	RandLANet [14]	(x, y, z, I_1)	65.40	68.20	66.80
	Cylinder3D* [41]	(x, y, z, I_1)	70.17	72.57	71.37
	PTv3 [37]	(x, y, z, I_1)	57.01	54.17	55.59
RV	RangeNet++ [24]	(D_1, x, y, z, I_1)	48.83	46.69	47.76
	SalsaNext [8]	(D_1, x, y, z, I_1)	65.66	76.74	71.20
	RangeFormer† [18]	–	–	–	–
	RangeViT [1]	(D_1, x, y, z, I_1)	56.34	57.38	56.86
Ours	PILF	$(I_{1,2}, D_{1,2}, P_{1,2}, L)$	78.58	80.06	79.32

* Voxel predictions are interpolated to point-level labels for fair comparison.
† Indicates no official code released.

depth bins. PAE adopts a base channel dimension of $C = 16$, with $\psi(\cdot)$ set to its best-performing form. EAF uses $C = 64$ to match the U-Net backbone, and its bin-wise coefficient $\ell(\cdot)$ is implemented as a linear layer. The network is optimized using an equally weighted combination of cross-entropy and Dice loss [25]. We evaluate segmentation performance using flare IoU, core IoU, and mean IoU (mIoU), capturing corrupted-region localization, valid-core preservation, and overall segmentation quality, respectively.

4.2 Flare Segmentation and Suppression

Segmentation Performance. We compare PILF with representative 2D, 3D, and range-view (RV) segmentation methods. Here, (I, D, P) denotes the intensity, depth, and pulse width of a single echo with subscripts indicating echo indices, L denotes ambient illumination, and (x, y, z) denotes 3D point coordinates. For fairness, each method is evaluated with its preferred input representation: 2D models use the proposed 7-channel input, while 3D methods operate on point clouds with intensity, and RV methods use depth maps with their native attributes. All methods are trained and evaluated under the same settings. Table 1 reports overall mIoU for flare and core regios. Our method achieves the best performance, providing a reliable mask for subsequent flare suppression.

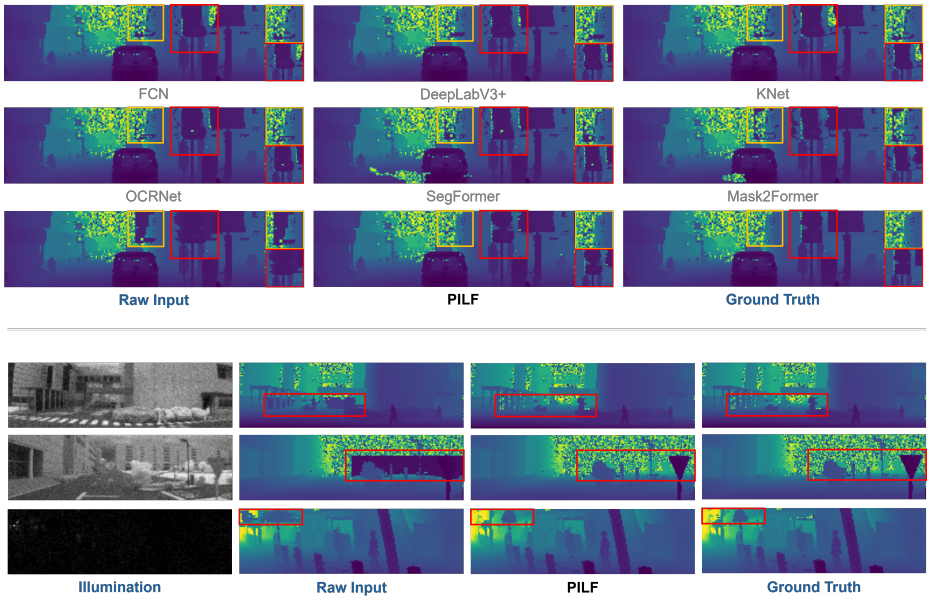

Fig. 8: Qualitative visualization of depth maps after flare suppression. The upper block compares different 2D segmentation methods, while the lower block shows additional scenarios under diverse illumination conditions and flare patterns. PILF achieves more complete flare suppression and better structural preservation.

Input Representation. To isolate the effect of the proposed input formation, we additionally evaluate OCRNet and DeepLabV3+ with first-echo input (I_1, D_1, P_1), shown as gray rows in Table 1. Our multi-modal input consistently improves both baselines, confirming the benefit of incorporating second-echo and ambient illumination cues for flare segmentation.

Suppression Results. The predicted mask is used to suppress flare-contaminated echoes, yielding more plausible depth maps and improved scene geometry. Figure 8 presents qualitative suppression results. The upper block compares depth maps obtained using segmentation masks from different 2D methods, where PILF more completely removes flare artifacts while preserving valid object structures. The lower block shows additional scenarios under diverse illumination conditions and flare patterns, demonstrating consistent suppression quality.

Depth Correction. We further evaluate the effect of flare suppression on depth recovery using an annotation-guided echo correction as the reference. This reference is obtained by applying the same echo replacement rule with the annotated flare mask. We compare the original first-echo depth and the predicted corrected depth against this reference in flare and foreground regions, where foreground denotes the union of flare and core pixels. MSE and MAE are computed on normalized depth values, as shown in Table 2.

Table 2: Depth-level evaluation before and after flare suppression. Metrics are computed against annotation-guided echo correction. Lower is better.

Setting	Flare		Foreground	
	MSE ↓	MAE ↓	MSE ↓	MAE ↓
Before	$0.2275_{\pm 0.1535}$	$0.3849_{\pm 0.1601}$	$0.1447_{\pm 0.1014}$	$0.2447_{\pm 0.1165}$
After	$0.0321_{\pm 0.0398}$	$0.0613_{\pm 0.0608}$	$0.0219_{\pm 0.0211}$	$0.0413_{\pm 0.0341}$
Reduction	85.89%	84.08%	84.85%	83.11%

Practical Discussion. Failure cases mainly arise from imperfect flare masks, which may cause under- or over-correction. Multiple corrupted echoes or missing reliable later echoes can also limit recovery quality, although these cases are relatively rare in our observations. By performing suppression directly in the 2D range-view domain, PILF avoids computationally intensive 3D convolutional processing. The end-to-end latency is 37 ms per frame on a single `RTX 3090`, which is compatible with 10 fps acquisition. This suggests that deployment on edge devices may be feasible with further optimization.

4.3 Ablation Study

Hyperparameters. We ablate the number of depth bins N in DD and the meta function $\psi(\cdot)$ in PAE. Each factor is varied independently while keeping the other at its best-performing setting, and each configuration is trained and tested five times. As shown in Table 3, $N = 16$ and $\psi(x) = 1/e^{x^2}$ achieve the best mIoU, while performance remains relatively stable across different settings.

Modalities. We ablate the second echo x_2 and ambient illumination x_0 by zeroing out the corresponding inputs while keeping the architecture unchanged. As shown in the first block of Table 4, removing either modality degrades performance, confirming that multi-echo and ambient cues are complementary for distinguishing flare from valid structures.

Augmentations. We evaluate FlareAug by selectively disabling the crop and fuse operations. The second block of Table 4 shows that both components improve generalization, with their combination achieving the best performance.

Modules. We conduct stepwise ablations to assess each architectural component. Starting from the full model, we first remove the attention decoder while retaining the U-Net backbone. We then replace EAF with a shared convolutional layer to preserve channel compatibility. Finally, we simplify PAE to a single convolutional block and set $N = 1$ to match the expected input dimension from DD. As shown in the third block of Table 4, each component improves performance, with the complete model achieving the best result.

Table 3: Ablation study on the number of depth bins N and meta function $\psi(\cdot)$.

Depth bins		mIoU
N	4	78.21±0.73
	8	77.87±0.46
	16	78.49±0.29
	32	77.28±0.60

Meta function		mIoU
$\psi(\cdot)$	$1/(1+x^2)$	77.62±0.30
	$1/e^{x^2}$	78.49±0.29
	$\cos(\pi/2 \cdot \tanh(x))$	77.67±0.41
	$1/(1+\lvert x\rvert)$	77.83±0.22

Table 4: Ablation studies on input modalities, augmentation operations, and architectural components. All results are averaged over five runs.

x_2	x_0	mIoU
		75.18±0.45
✓		75.56±0.81
	✓	77.67±0.22
✓	✓	78.49±0.29

Crop	Fuse	mIoU
		74.91±1.30
✓		77.41±0.53
	✓	75.58±0.56
✓	✓	78.49±0.29

PAE	EAF	Attn.	mIoU
			64.41±0.35
✓			71.01±0.49
✓	✓		75.40±0.54
✓	✓	✓	78.49±0.29

5 Conclusion

In this paper, we present PILF, a physically informed segmentation method for flare suppression in SPAD-based LiDAR systems. By reformulating flare suppression as a semantic segmentation problem, PILF learns geometric-photometric cues from raw SPAD measurements while leveraging multi-echo and ambient illumination information. This design enables effective discrimination between spurious flare artifacts and valid object structures. Extensive experiments show that PILF effectively suppresses flare while preserving the structural fidelity of true object surfaces across diverse real-world scenarios. Operating in the 2D range-view domain, PILF avoids computationally intensive 3D convolutional processing. Moreover, since suppression is performed at the output level, it does not alter the underlying hardware or signal acquisition process, making it complementary to hardware- and signal-level mitigation strategies and compatible with existing pipelines. Overall, this work demonstrates the potential of combining physically grounded modeling with learning-based segmentation for mitigating challenging flare artifacts in SPAD-based LiDAR systems.

Acknowledgements

We would like to thank Hayakawa Michio from Sony Semiconductor Solutions for his valuable assistance in data collection.

References

1. Ando, A., Gidaris, S., Bursuc, A., Puy, G., Boulch, A., Marlet, R.: Rangevit: Towards vision transformers for 3d semantic segmentation in autonomous driving. In:

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5240–5250 (2023)
2. Byun, Y.S., Jeong, R.G.: High-speed outlier removal filter for lidar sensor point cloud data. IEEE Access (2024)
3. Chen, G., Wiede, C., Kokozinski, R.: Data processing approaches on spad-based d-tof lidar systems: A review. IEEE Sensors Journal **21**(5), 5656–5667 (2020)
4. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018)
5. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1290–1299 (2022)
6. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3075–3084 (2019)
7. Contributors, M.: Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark (2020)
8. Cortinhal, T., Tzelepis, G., Erdal Aksoy, E.: Salsanext: Fast, uncertainty-aware semantic segmentation of lidar point clouds. In: Advances in Visual Computing: 15th International Symposium, ISVC 2020, San Diego, CA, USA, October 5–7, 2020, Proceedings, Part II 15. pp. 207–222. Springer (2020)
9. Cusini, I., Berretta, D., Conca, E., Incoronato, A., Madonini, F., Maurina, A.A., Nonne, C., Riccardo, S., Villa, F.: Historical perspectives, state of art and research trends of spad arrays and their applications (part ii: Spad arrays). Frontiers in Physics **10**, 906671 (2022)
10. Ficorella, A., Pancheri, L., Dalla Betta, G.F., Brogi, P., Collazuol, G., Marrocchesi, P.S., Morsani, F., Ratti, L., Savoy-Navarro, A.: Crosstalk mapping in cmos spad arrays. In: 2016 46th European Solid-State Device Research Conference (ESSDERC). pp. 101–104. IEEE (2016)
11. Gupta, A., Ingle, A., Gupta, M.: Asynchronous single-photon 3d imaging. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7909–7918 (2019)
12. Gupta, A., Ingle, A., Velten, A., Gupta, M.: Photon-flooded single-photon 3d cameras. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6770–6779 (2019)
13. Gyongy, I., Dutton, N.A., Henderson, R.K.: Direct time-of-flight single-photon imaging. IEEE Transactions on Electron Devices **69**(6), 2794–2805 (2021)
14. Hu, Q., Yang, B., Xie, L., Rosa, S., Guo, Y., Wang, Z., Trigoni, N., Markham, A.: Randla-net: Efficient semantic segmentation of large-scale point clouds. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11108–11117 (2020)
15. Ikeda, K., Hara, R., Nagata, R., Sako, O., Ding, Z., Kado, T., Fujioka, I., Beppu, T., Isogawa, M., Yoshioka, K.: Ghost-fwl: A large-scale full-waveform lidar dataset for ghost detection and removal. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17164–17173 (2026)
16. Incoronato, A., Locatelli, M., Zappa, F.: Statistical modelling of spads for time-of-flight lidar. Sensors **21**(13), 4481 (2021)
17. Jahromi, S., Kostamovaara, J.: Timing and probability of crosstalk in a dense cmos spad array in pulsed tof applications. Optics express **26**(16), 20622–20632 (2018)

18. Kong, L., Liu, Y., Chen, R., Ma, Y., Zhu, X., Li, Y., Hou, Y., Qiao, Y., Liu, Z.: Rethinking range view representation for lidar segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 228–240 (2023)
19. Land, E.H., McCann, J.J.: Lightness and retinex theory. Journal of the Optical society of America **61**(1), 1–11 (1971)
20. Li, Y., Ibanez-Guzman, J.: Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems. IEEE Signal Processing Magazine **37**(4), 50–61 (2020)
21. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
22. Liu, Z., Tang, H., Lin, Y., Han, S.: Point-voxel cnn for efficient 3d deep learning. Advances in neural information processing systems **32** (2019)
23. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
24. Milioto, A., Vizzo, I., Behley, J., Stachniss, C.: Rangenet++: Fast and accurate lidar semantic segmentation. In: 2019 IEEE/RSJ international conference on intelligent robots and systems (IROS). pp. 4213–4220. IEEE (2019)
25. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision (3DV). pp. 565–571. Ieee (2016)
26. Po, R., Pediredla, A., Gkioulekas, I.: Adaptive gating for single-photon 3d imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16354–16363 (2022)
27. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 652–660 (2017)
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 **30** (2017)
29. Rech, I., Ingargiola, A., Spinelli, R., Labanca, I., Marangoni, S., Ghioni, M., Cova, S.: Optical crosstalk in single photon avalanche diode arrays: a new complete model. Optics express **16**(12), 8381–8394 (2008)
30. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18. pp. 234–241. Springer (2015)
31. 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. pp. 28913–28924 (2025)
32. Tashiro, Y., Ito, K.: Spad depth sensor for automotive lidar systems. JSAP Review **2023**, 230402 (2023)
33. Tontini, A., Gasparini, L., Perenzoni, M.: Numerical model of spad-based direct time-of-flight flash lidar cmos image sensors. Sensors **20**(18), 5203 (2020)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

35. Veerappan, C., Charbon, E.: A substrate isolated cmos spad enabling wide spectral response and low electrical crosstalk. IEEE Journal of Selected Topics in Quantum Electronics **20**(6), 299–305 (2014)
36. Villa, F., Severini, F., Madonini, F., Zappa, F.: Spads and sipms arrays for long-range high-speed light detection and ranging (lidar). Sensors **21**(11), 3839 (2021)
37. Wu, X., Jiang, L., Wang, P.S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., Zhao, H.: Point transformer v3: Simpler, faster, stronger. In: CVPR (2024)
38. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems **34**, 12077–12090 (2021)
39. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16. pp. 173–190. Springer (2020)
40. Zhang, W., Pang, J., Chen, K., Loy, C.C.: K-net: Towards unified image segmentation. Advances in Neural Information Processing Systems **34**, 10326–10338 (2021)
41. Zhou, H., Zhu, X., Song, X., Ma, Y., Wang, Z., Li, H., Lin, D.: Cylinder3d: An effective 3d framework for driving-scene lidar semantic segmentation. arXiv preprint arXiv:2008.01550 (2020)