Separating the Wheat from the Chaff: Spatio-Temporal Transformer with View-interweaved Attention for Photon-Efficient Depth Sensing

Letian Yu[1], Jiaxi Yang[1], Bo Dong[2,*], Qirui Bao[1], Yuanbo Wang[1], Felix Heide[3], Xiaopeng Wei[1,*], Xin Yang[1,*]

[1]Key Laboratory of Social Computing and Cognitive Intelligence, Dalian University of Technology
[2] Cephia AI
[3] Princeton University
{letianyu, 1612000589, wangyuanbo}@mail.dlut.edu.cn, 517542583@qq.com, bo.dong@cephia.ai, fheide@cs.princeton.edu, {xpwei, xinyang}@dlut.edu.cn

Abstract

Time-resolved imaging is an emerging sensing modality that has been shown to enable advanced applications, including remote sensing, fluorescence lifetime imaging, and even non-line-of-sight sensing. Single-photon avalanche diodes (SPADs) outperform relevant time-resolved imaging technologies thanks to their excellent photon sensitivity and superior temporal resolution on the order of tens of picoseconds. The capability of exceeding the sensing limits of conventional cameras for SPADs also draws attention to the photon-efficient imaging area. However, photon-efficient imaging under degraded conditions with low photon counts and low signal-to-background ratio (SBR) still remains an inevitable challenge. In this paper, we propose a spatio-temporal transformer network for photon-efficient imaging under low-flux scenarios. In particular, we introduce a view-interweaved attention mechanism (VIAM) to extract both spatial-view and temporal-view self-attention in each transformer block. We also design an adaptive-weighting scheme to dynamically adjust the weights between different views of self-attention in VIAM for different signal-to-background levels. We extensively validate and demonstrate the effectiveness of our approach on the simulated Middlebury dataset and a specially self-collected dataset with real-world-captured SPAD measurements and well-annotated ground truth depth maps.

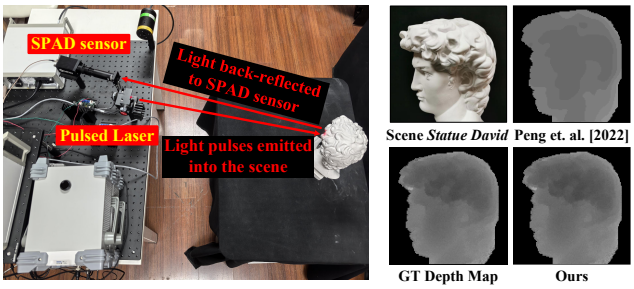

Figure 1: (a) Single-photon depth sensing system, which contains a pulsed laser source, a TCSPC module and a SPAD sensor and measures a spatio-temporal photon histogram when photons are emitted by the pulsed laser into the scene and reflected back to the SPAD sensor. (b) Our approach can extract spatio-temporal correlations and outperform the state-of-the-art algorithms for depth sensing tasks in real-world applications.

Introduction

Time-resolved imaging, or the time-tagging of the optical response of a scene to transient illumination, allows it to exceed the sensing boundaries beyond conventional frame-based intensity imaging and analyze the temporal information of light transport. Time-resolved measurements provide superior temporal resolutions and rich temporal cues for scene understanding and are engaged in practical applications, *e.g.*, health care and life sciences (*e.g.* fluorescence lifetime imaging), robotics and autonomous driving, military defenses, and even non-line-of-sight imaging or "looking around a corner" (Velten et al. 2012) .

A notable approach available for recording time-resolved measurements is the emerging technology of single-photon avalanche diodes (SPADs). SPADs are single-photon sensitive devices working in active depth sensing due to their high sensitivity to individual photons and ability to record time-tagged photon arrivals with picosecond timing resolution (Rochas 2003). By employing a pulsed laser with picosecond duration, a single photon depth sensing system illuminates the targets with abundant pulses and collects the scattered photons to the SPAD detector. A spatio-temporal histogram is then constructed via the time-of-arrival of multiple collected photons to infer the target depth in a time-of-flight format. The histogram approximates the amounts of the returning photons and enables recovery distance, reflectance, and 3D geometry.

However, robustly sensing depth from raw photon histograms under degraded conditions remains a significant challenge. The back-reflected signal photons are weak and contaminated with strong noise photons from ambient light, resulting in low photon counts and low signal-to-background ratio (SBR). This phenomenon brings severe distortions into the raw single-photon measurements, leading to poor depth reconstruction quality, especially in long-

*Corresponding Authors.

distance scenarios. Therefore, high-performance depth reconstruction algorithms are essential for promoting the practicability of single-photon depth sensing systems under low SBR scenarios.

Several heuristic algorithms have been proposed to process single-photon measurements with noisy photon counts (Kirmani et al. 2014; Shin et al. 2015, 2016; Rapp and Goyal 2017). However, such approaches demand strict restrictions, need many user-defined parameters, and meet performance drops when applied to diverse sensing conditions in real-world scenarios. Recent learning-based works (Lindell, O'Toole, and Wetzstein 2018; Peng et al. 2020, 2022) achieve quite promising results on the single-photon depth sensing task, they still fail to recover single-photon histogram cubes under photon-starved regimes, and the network structures lack specific designs to correlate with the characteristics of photon-efficient measurements.

Recent works (Peng et al. 2020, 2022) explore long-range correlations on both spatial and temporal dimensions, when the spatial-view correlations are more inclined to seek structure similarities and depth discontinuities, and the temporal-view correlations focus more on the clustering characteristics in the temporal domain. However, sparse binary signal photons suppressed by massive noisy background photons make it challenging to find spatio-temporal correlations as SBR decreases. Exploring the relationship between long-range correlations in different views and SBR levels and designing a reasonable adaptive-fusion strategy would naturally benefit more for the single-photon depth sensing tasks.

Based on such motivation, we propose a 3D spatio-temporal transformer for single photon depth sensing (SPSD-STFormer) to extract information from the input measurements and their inter-relation to construct the spatial and temporal correlations globally. To exploit the differences in long-range correlations between different SBR levels, we propose a view-interweaved attention mechanism (VIAM) in every 3D transformer block to learn both spatial-view and temporal-view long-range correlations among single photon measurements and adaptively fuse the outputs from both spatial and temporal self-attention blocks. We also propose an adaptive-weighting scheme to learn the global feature of SBR conditions to regulate the fusion weights in a self-modulated strategy. As a result, the proposed components can work together to effectively explore and excavate global and local cues for the depth sensing task. We perform extensive experiments to demonstrate the superiority of our method over previous state-of-the-art methods under different SBR conditions and validate that our approach can generalize on a real-world single-photon depth sensing system (shown in Fig. 1).

Our contributions can be summarized as:

- We propose a 3D spatial-temporal transformer for single photon depth sensing (SPDS-STFormer), to address the degraded performance for 3D single-photon measurements under low photon counts and low SBR conditions.
- We introduce a view-interweaved attention mechanism to adaptively fuse spatial-view self-attention and temporal-view self-attention in each transformer block, which can learn global features of SBR conditions by an adaptive-weighting scheme.
- We validate and demonstrate that our approach outperforms state-of-the-art methods. Our ablation study evidences the effectiveness of each key component of our approach.
- We validate the generalizability and robustness on a novel single-photon depth sensing system, which captures 128×128 SPAD-based measurements with corresponding well-annotated depth maps.

Related Work

Single-photon Sensors. Single-photon avalanche diodes (SPADs) are an emerging pixel technology with ultra-high sensitivity down to individual photons (Cova et al. 1996; Fossum et al. 2016). SPADs are reverse-biased photodiodes that operate above their breakdown voltage, *i. e.*, in Geiger-mode (Aull et al. 2002).

Every photon incident on a SPAD has a probability of triggering an electron avalanche and records a timestamped event, providing a temporal resolution from ten to hundreds of picoseconds. By recording and combining the timestamped photon events returning from a pulsed illumination source, which often operates at MHz rates, a spatio-temporal photon counts histogram can be accumulated to characterize the reflectance and 3D geometry to be recovered. SPADs have received wide attention due to their excellent single-photon sensitivity and have been commonly used for a wide range of applications in optical telecommunication, fluorescence lifetime imaging (Castello et al. 2019), and remote sensing systems (e.g., LIDAR) (Kirmani et al. 2014; Tinsley et al. 2016) for their superior timing resolution and excellent photon efficiency. The supplemental materials include the imaging principles of SPAD sensors.

Depth Sensing. Depth sensing is a technology that measures the distance between a sensor and a target surface and is commonly employed in virtual reality, augmented reality, autonomous driving, and other computer vision applications (Geiger et al. 2013; Cabon, Murray, and Humenberger 2020; Mei et al. 2021; Zhang et al. 2023a). Below is an overview of the most important methods of depth sensing, including passive depth sensing, active depth sensing, and a combination of passive and active sensing.

Passive Depth Sensing. Passive depth sensing systems utilize available ambient light or artificial lighting to illuminate the object and infer depth information by capturing and analyzing naturally occurring environmental cues or ambient light variations. Representative technologies about passive depth sensing include stereo vision (Hirschmuller 2007; Furukawa et al. 2009) and structure from motion (Han et al. 2015; Lowe 2004). Stereo vision methods leverage multiple cameras to capture a scene from slightly different viewpoints and rely upon an optimal correspondence matching between pixels on epipolar lines in the left and right images to concatenate depth features and compute/aggregate the matching cost (Li et al. 2021; Guo et al. 2022; Zhang

et al. 2023b). Structure from motion recovers 3D structure from a sequence of 2D images taken from different viewpoints as the scene or camera moves and exploits multi-view photometric or feature-metric constraints to enforce the relationship between dense depth and the camera pose in network while simultaneously estimating scene depth, camera poses, and intrinsic parameters (Wei et al. 2020; Chawla et al. 2022; Chen, Kumar, and Yu 2023). However, with advantages such as less power consumption, passive depth sensing technology has limitations in low-light conditions. It achieves low accuracy, especially in complex scenes with varying lighting conditions and surface textures.

Active Depth Sensing. Active depth sensing systems employ an active light source, projecting random patterns onto the visible scene, measuring the time it takes for the signals to return, and calculating the distances of the objects in the scene. Representative technologies about active depth sensing include Time-of-flight (ToF) (Lange and Seitz 2001; Zhang 2012) and light detection and ranging (LiDAR) (Wandinger 2005; Baek and Heide 2022). ToF imaging measures the time it takes for a signal, typically a pulse of light, to travel to an object and back to the sensor. Direct ToF (dToF) (Padmanabhan, Zhang, and Charbon 2019; Sun et al. 2023) refers to emitting a single pulse and calculating the distance based on the time difference between the emitted pulse and the received reflection. In contrast, indirect ToF (iToF) (Qiu et al. 2019; Meuleman et al. 2022) uses a continuous modulated/ coded stream of light. LiDAR is usually used to describe a unique scanning-based ToF technology. The scanning process is repeated up to millions of times per second, producing a precise 3D point cloud of the environment (Jiang et al. 2021; Liu et al. 2022). However, with advantages such as higher accuracy, active depth sensing technology has spatial resolution limitations and a key prerequisite for accumulating enough photons under various adversary conditions.

In this paper, our SPAD-based depth sensing measurements belong to active sensing technology. With superior photon efficiency and excellent timing resolution, SPAD-based systems have the capability to sense depth information in low-flux scenarios.

SPAD-based Depth Sensing. SPADs are highly sensitive to low levels of light, allowing for the detection of single photons and enabling depth sensing in low-light conditions. Additionally, SPADs have fast response times, making them suitable for real-time depth sensing applications. Recent works on SPAD-based LiDAR are also rapidly becoming commercially available (Corporation 2021; Europe 2023). However, SPADs are susceptible to noise, including dark counts and afterpulsing, which can affect the accuracy of depth measurements, especially in low-light conditions. A significant amount of related work is currently dedicated to addressing this problem for SPADs. Shin *et. al.* (Shin et al. 2013) provide a spatiotemporally regularized estimation of the depth maps based on accurate physical modeling of the time-inhomogeneous photon detection process. Kirman *et. al.* (Kirmani et al. 2014) introduce a low-flux imaging technique which exploits spatial correlations in real-world scenes and recovers 3D structure and reflectivity from the first detected photon per pixel. Shin *et. al.* (Shin et al. 2015, 2016) utilize both the transverse smoothness and longitudinal sparsity of natural scenes to reconstruct depth and intensity in low-light environments. Rapp *et. al.* (Rapp and Goyal 2017) recover depth and intensity by unmixing the contributions from signal and noise sources with an adaptive temporal window. Lindell *et. al.* (Lindell, O'Toole, and Wetzstein 2018) propose a deep-learning based method to reconstruct depth maps with high-resolution intensity images. Recent works (Peng et al. 2020, 2022) introduce long-range correlations in spatial and temporal views, where the spatially neighbourhoods have similar geometry would have similar depth values with high possibilities, and signal photons are supposed to cluster together near the true depth. Lee *et. al.* (Lee et al. 2023) collaboratively exploit both local and non-local correlations in the spatio-temporal photon measurements and estimate scene properties reliably even under very challenging lighting conditions. However, these spatial and temporal correlations change drastically when the number of signal photons decreases, and the number of background photons increases. The non-local blocks (Peng et al. 2020, 2022) can only exploit the global information in the spatial dimensions but ignore the long-range correlations in the temporal domain. These fixed non-local architectures also cannot effectively describe spatio-temporal correlated features according to different SBR levels, and their performance degrades under low photon counts and SBR. We also focus on spatial and temporal information among photon-efficient measurements, and our network designs an adaptive-weighting strategy to study the proportion between long-range correlations in spatial and temporal dimensions in different SBR conditions.

Methodology

Single-photon measurements have superior temporal information and photon sensitivity, which also results in degraded performance when signal photons are flooded by background noise photons under low SBR scenarios.

To address this challenge, we propose SPDS-STFormer, a 3D spatio-temporal transformer to reconstruct depth maps from 3D single photon histogram cubes ($128 \times 128 \times 1024$ for our depth sensing system) utilizing ViT (Dosovitskiy et al. 2020) structures. In the following, we first show the overview of the network architecture and then describe the details of the main components.

Network Overview

As illustrated in Fig. 2, the architecture of SPDS-STFormer consists of a 3D feature extractor, a U-shaped encoder-decoder transformer structure , and a 3D decoder. We follow (Peng et al. 2022) to utilize a 3D version of dilated dense fusion block as the 3D local feature extractor.

We then design a hierarchical multi-scale 3D spatio-temporal transformer to not only reduce the computational burden but also perceive contextual information in a larger field of view. Each transformer block is composed of a 3D convolution layer for tokenizing, a view-interweaved

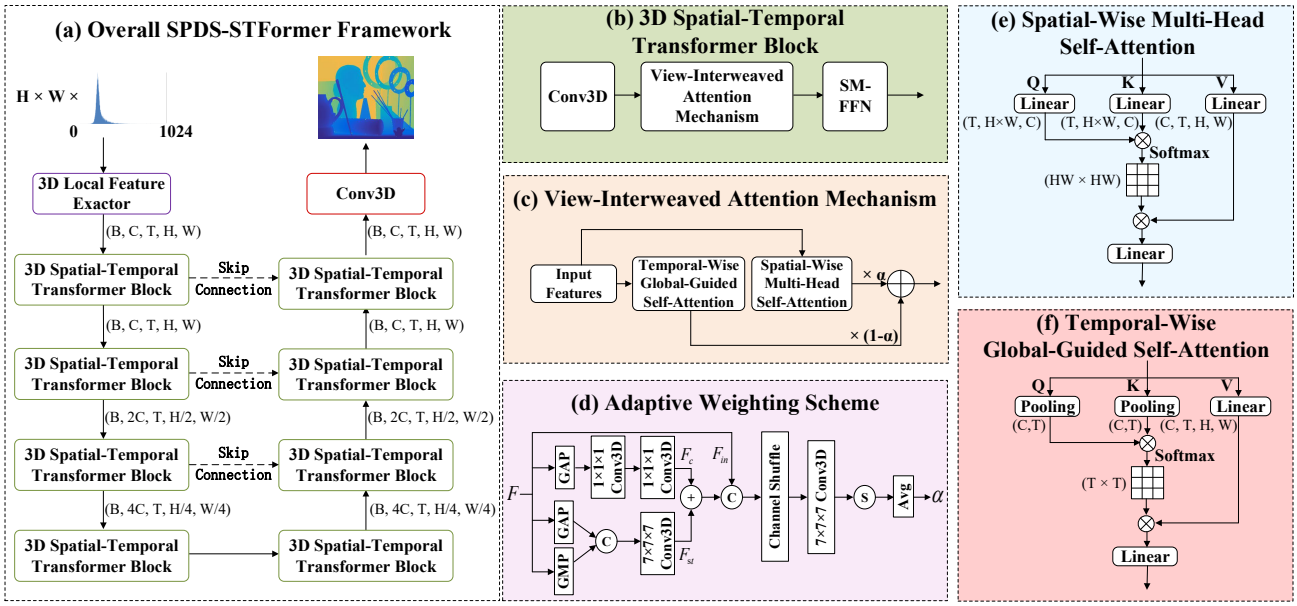

Figure 2: (a) Overview of our SPDS-STFormer and its main building blocks: (b) a 3D Spatio-Temporal Transformer Block, (c) a View-interweaved Attention Mechanism, (d) an Adaptive-weighting Scheme, (e) a Temporal-View Global-Guided Self-Attention Block and (f) a Spatial-View Multi-Head Self-Attention Block.

attention mechanism (VIAM) using an adaptive-weighting scheme for adaptively fusing spatial-view and temporal-view self-attention, and a self-modulated feed-forward network (SM-FFN) (Lai, Yan, and Fu 2023) for amplifying the activation of regions with high information density and improve the reconstruction quality for high-frequency components. The overall process of each transformer block can be expressed as:

$$\begin{aligned} \hat{X} &= \text{BN}(\text{Conv3D}(X)), \\ Y &= \text{SM-FFN}(\text{VIAM}(\hat{X})), \end{aligned} \tag{1}$$

where $X \in \mathbb{R}^{C\times L\times H\times W}$ and $Y \in \mathbb{R}^{C\times L\times H\times W}$ refer to the input features and output features, BN denotes batch normalization, and Conv3D denotes the 3D convolution operation.

In the decoder branch, we simply leverage several 3D deconvolutions to recover the denoised histograms from features and generate prediction maps.

View-interweaved Attention Mechanism

In each 3D transformer block of SPDS-STFormer, the core component is the View-interweaved Attention Mechanism (VIAM), which not only models a Spatial-View Multi-Head Self-Attention (SV-MHSA) module to achieve long-range spatial-correlated features, but also introduces a Temporal-View Global-Guided Self-Attention (TV-GGSA) module to extract features with long-range temporal correlations. Both features are integrated into different views.

In our SV-MHSA module, we apply original multi-head self-attention process (Vaswani et al. 2017) in a 3D version to obtain the heads $\mathbf{Q}_S, \mathbf{K}_S, \mathbf{V}_S \in \mathbb{R}^{C\times L\times N}$ of query, key and value with the same dimensions by linear projection operations, respectively, where $N = H \times W$. We further apply a linear projection $\mathbf{W}_S \in \mathbb{R}^{C\times C}$ to obtain the spatial-correlated long-range features $F_S \in \mathbb{R}^{C\times L\times H\times W}$:

$$\begin{aligned} \text{Attn}(\mathbf{Q}_S, \mathbf{K}_S, \mathbf{V}_S) &= \text{Softmax}(\frac{\mathbf{Q}_S \cdot \mathbf{K}_S^T}{\sqrt{d_{head}}})\mathbf{V}_S, \\ F_S &= \mathbf{W}_S \cdot \text{Attn}(\mathbf{Q}_S, \mathbf{K}_S, \mathbf{V}_S). \end{aligned} \tag{2}$$

In our TV-GGSA module, for computational simplicity, we perform the global average pooling on input features $F_{in} \in \mathbb{R}^{C\times L\times H\times W}$ to obtain the global features, *i.e.* $\mathbf{Q}_T, \mathbf{K}_T \in \mathbb{R}^{C\times L}$ instead of linear projections for query and key features in conventional self-attention block (Vaswani et al. 2017). We only linearly project value $\mathbf{V}_T \in \mathbb{R}^{C\times L\times H\times W}$ from F_{in} in our TV-GSSA module.

$$\begin{aligned} \text{Attn}(\mathbf{Q}_T, \mathbf{K}_T, \mathbf{V}_T) &= \mathbf{V}_T \cdot \text{Softmax}(\mathbf{K}_T \cdot \mathbf{Q}_T), \\ F_T &= \mathbf{W}_T \cdot \text{Attn}(\mathbf{Q}_T, \mathbf{K}_T, \mathbf{V}_T), \end{aligned} \tag{3}$$

where we perform another linear projection $\mathbf{W}_T \in \mathbb{R}^{C\times C}$ to better transform the features to the output $F_T \in \mathbb{R}^{C\times L\times H\times W}$.

In our proposed method, the two operations are implemented in parallel, and the output features F_S and F_T are combined as follows to obtain the output of the VIAM:

$$F_{fuse} = \alpha \cdot F_S + (1-\alpha) \cdot F_T, \tag{4}$$

where the weighting parameter $\alpha \in [0, 1]$ are learned from an adaptive-weighting scheme. Details about the adaptive-weighting scheme will be introduced below.

Adaptive-weighting Scheme

To alleviate the contamination by noisy background photons and keep awareness of illumination changes, we propose an adaptive weighting scheme to extract the re-calibrated features of global illumination information and signal-to-background levels. The detailed procedures of the adaptive weighting scheme are depicted in Fig. 2. Given the input 3D features $F \in \mathbb{R}^{C\times L\times H\times W}$, our adaptive-weighting scheme aims to mix global attention weights to guarantee that our learnable weights are concrete illumination-aware to suppress the influence from noisy photons.

We first calculate the global attention weights in channel and spatio-temporal dimensions. The channel attention calculates a channel-wise vector $F_c \in \mathbb{R}^{C\times 1\times 1\times 1}$ by a 3D version of the channel recalibration (Hu, Shen, and Sun 2018). The spatio-temporal attention calculates a spatio-temporal importance map $F_{st} \in \mathbb{R}^{C\times L\times H\times W}$ to indicate the importance levels of different volumes adaptively.

We compute the corresponding weights F_c and F_{st} by following,

$$\begin{aligned} F_{st} &= \mathcal{C}_{7\times 7\times 7}([F^{st}_{GAP}, F^{st}_{GMP}]), \\ F_c &= \mathcal{C}_{1\times 1\times 1}(\max(\mathcal{C}_{1\times 1\times 1}(F^{c}_{GAP}))), \end{aligned} \tag{5}$$

where F^{st}_{GAP}, F^{st}_{GMP}, F^{c}_{GAP} denote the features extracted from global average pooling across the spatio-temporal dimensions, global max pooling across the spatio-temporal dimensions, and global average pooling across the channel dimensions, respectively.

Then we fuse F_{st} and F_c via an addition operation to obtain coarse global illumination-aware features $F_{ia} \in \mathbb{R}^{C\times L\times H\times W}$. In order to achieve finer features to model illumination and SBR information, we then utilize a 3D version of channel shuffle operation (Zhang et al. 2018) on F_{ia} and F. Finally, we calculate the average and obtain the learnable weights α.

$$\begin{aligned} F_{ia} &= F_{st} + F_c, \\ \alpha &= \frac{1}{N}\sum_{n=1}^{N}(\sigma(CS([F_{ia}(n), F(n)]))), \end{aligned} \tag{6}$$

where σ denotes the sigmoid operation, $CS(\cdot)$ denotes the channel shuffle operation and N denotes the total element number, respectively.

Loss Function

For the depth sensing task, we apply a combination of Kullback-Leibler (KL) divergence, L2 loss, and total variation (TV) loss as the loss function for supervised training of SPDS-STFormer. The supplemental materials include the details of loss functions.

Assessment

Implementation Details. We implement our method in PyTorch (Paszke et al. 2019). Following previous literature (Lindell, O'Toole, and Wetzstein 2018; Peng et al. 2020, 2022), for training set, we simulate SPAD histogram measurements using RGB-D images from the NYU v2 dataset by sampling the inhomogeneous Poisson process. Each measurement has 1,024 temporal bins to construct the histogram with a bin size of 80 ps and a detected illumination pulse with a full width at half maximum (FWHM) of 400 ps. To vary the signal and background noise levels across the dataset, we simulate an average of 2, 5, and 10 signal photons detected per pixel, with 2, 10, and 50 background photons at each signal level. A total of 13,500 measurements are produced for training and 2,800 for validation using the NYU v2 dataset, respectively.

Evaluation Metrics. The evaluation metric for depth sensing tasks is the commonly used root-mean-square-error (RMSE) between the reconstruction depth map and the ground truth.

Qualitative and Quantitative Evaluation

We first compare the effectiveness of our approach to existing photon-efficient depth sensing methods relying on SPAD measurements only from the simulated Middlebury stereo dataset (Scharstein and Pal 2007). We report the RMSE values averaged across 8 Middlebury test scenes with a large spatial resolution of 576 $\times$ 704 and a uniform temporal resolution of 1024 over a number of simulated signal and noise levels (10:2, 5:2, 2:2, 10:10, 5:10, 2:10, 10:50, 5:50, 2:50) generally reported in previous literatures to ensure a fair comparison with baseline approaches. It is worth noting that our method is designed independently of temporal resolution. As shown in Tab. 1, our SPDS-STFormer outperforms the state-of-the-art methods on the simulated Middlebury Dataset. Compared with other learning-based methods (Lindell, O'Toole, and Wetzstein 2018; Peng et al. 2022) degrade dramatically when SBR levels change, our SPDS-STFormer behaves much better with an elegant degradation, which indicates the effectiveness of our network under extremely low photon counts and low SBR.

Fig. 3 further qualitatively illustrates the effectiveness of SPDS-STFormer. The existing methods recover decent maps under a high SBR of 10:2, but their performances degrade severely when the SBR gets as low as 2:50. The log-matched filter method produces a noisy depth map, and Shin *et. al.* (Shin et al. 2016) estimate an out-of-range result due to this overwhelming background noise levels. Rapp *et. al.* (Rapp and Goyal 2017) and Lindell *et. al.* (Lindell, O'Toole, and Wetzstein 2018) cannot even recover the main structures of the scene when signal photons are flooded with background photons. Peng *et. al.* (Peng et al. 2022) has the capability to reconstruct main structures; however,it cannot avoid prediction errors caused in background regions. Our proposed SPDS-STFormer not only succeeds in reconstructing main structures and fine details but also suppresses the contamination by the noisy background photons.

Ablation Study

We validate the impact of each component in SPDS-Former by disabling one or more components and comparing the performance on the Middlebury test set (Tab. 2).

Effectiveness of VIAM components. Experiments (A), (B) gauge the impact of the VIAM components. Experiment

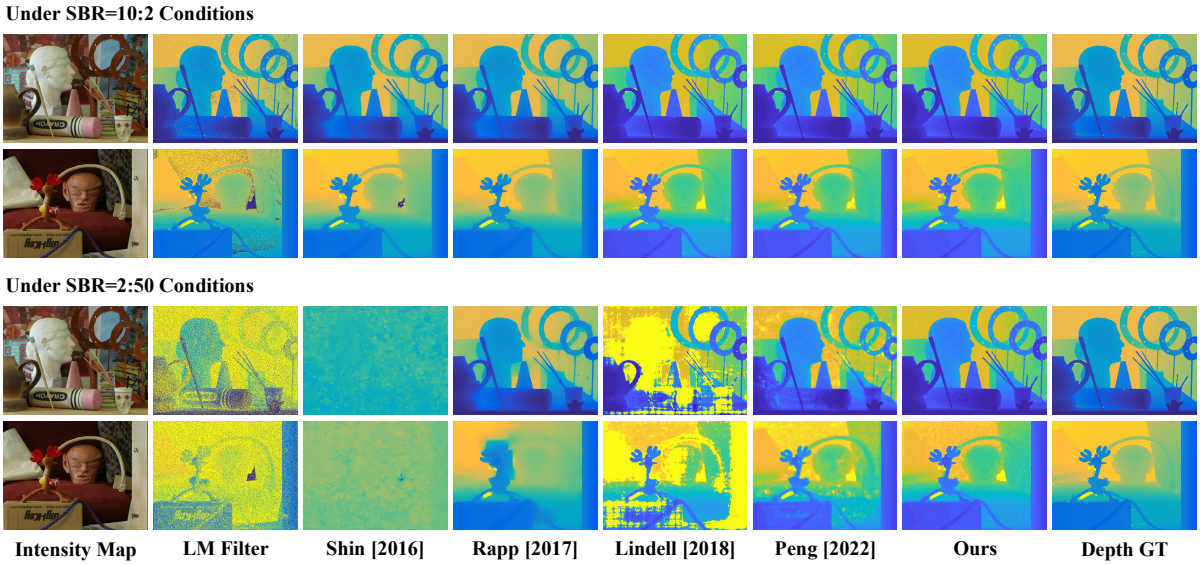

Figure 3: The visual comparisons of different methods in single-photon depth sensing for the *Art* (the upper image) and *Reindeer* (the lower image) on different SBR levels: 10:2 and 2:50. Over these previous methods, Lindell *et. al.* [2018] (Lindell, O'Toole, and Wetzstein 2018) and Peng *et. al.* [2022] (Peng et al. 2022) are learning-based methods for estimating the depth maps, while all other methods are heuristic methods that need manually-defined SBR parameters. Our approach outperforms competing methods with less noisy artifacts and better visual performance.

Avg. Photons	Avg. BG (SBR)	LM Filter	Shin [2016]	Rapp [2017]	Lindell [2018]	Peng [2022]	Ours
10	2 (5)	0.8362	0.0637	0.0579	0.0658	0.0634	**0.0543**
5	2 (2.5)	1.8912	0.0638	0.0629	0.0837	0.0631	**0.0551**
2	2 (1)	3.7243	0.2520	0.0668	0.1853	0.1190	**0.0607**
10	10 (1)	1.3173	0.2108	0.0527	0.2057	0.0900	**0.0481**
5	10 (0.5)	2.6531	2.1886	0.0628	0.3656	0.1063	**0.0607**
2	10 (0.2)	4.7607	4.3054	0.0602	1.2437	0.0786	**0.0581**
10	50 (0.5)	1.8511	4.3085	0.0555	0.1942	0.0790	**0.0516**
5	50 (0.1)	3.5602	5.0394	0.0566	0.4334	0.0678	**0.0528**
2	50 (0.05)	5.7798	5.4816	0.0716	1.7969	0.0935	**0.0711**

Table 1: Quantitative comparison of our SPDS-STFormer with state-of-the-art methods on Middlebury Dataset with different SBR levels. All results are calculated as RMSE metrics over the test set containing 8 scenes. We highlight the best results in **bold** and demonstrate that our method achieves the best performance over all SBR conditions.

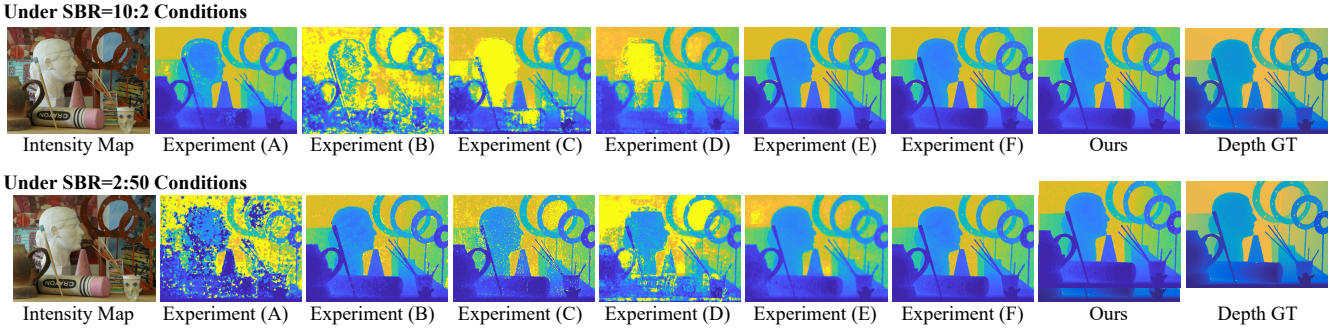

Figure 4: Visual ablation comparison of different SPDS-STFormer variants.

Experiments	High SBR (10:2)	Low SBR (2:50)
A w/o TV-GSSA	0.0919	0.5005
B w/o SV-MHSA	0.5557	0.1240
C α fixed to 0.2	0.4740	0.3696
D α fixed to 0.5	0.1705	0.7378
E w/o ad-weighting	0.0705	0.0714
F w/o channel shuffle	0.0641	0.0713
G Ours	**0.0543**	**0.0711**

Table 2: Quantitative ablation studies on the effectiveness of each component in our SPDS-STFormer.

RMSE	LM Filter	Shin [2016]	Peng [2022]	Ours
[mm]	96.77	45.60	11.92	**9.59**

Table 3: Comparison of the proposed method with state-of-the-art methods on Real-Captured Dataset. Our method achieves the state-of-the-art among comparison methods.

(A) only introduces the spatial-view self-attention mechanism in each transformer block, experiment (B) only contains the temporal-view self-attention mechanism in each transformer block. Compared to experiment (A) and experiment (B), the results of our approach in the experiment (F) achieve a better performance, demonstrating that spatial-view and temporal-view attentions are both critical for single-photon depth sensing in different SBR levels.

From qualitative comparisons in Fig. 4, we notice that only using spatial-view attention (Fig. 4 2nd column) can recover details in high SBR levels but will lead to checker artifacts in low SBR levels. Meanwhile, only using temporal-view attention (Fig. 4 3rd column) will bring out degraded noise in high SBR conditions but perform well in low SBR environments. Based on our reasoning, signal photons cluster better than the background photons and have quantitative dominance in the temporal domain under high SBR levels. We infer that spatial-wise long-range correlations play a leading role in high SBR scenarios. However, this assumption collapses in the sub-photon regime, where it is hard to locate signal photon clusters reliably, and in the high background flux regime, where noisy background photons may appear clustered. At the same time, temporal-wise attention leads to superior performance under low SBR scenarios.

Effectiveness of adaptive-weighting scheme. A key contribution of our approach is the adaptive-weighting scheme to adjust the fusion weights between different views of attention. To demonstrate its importance, we remove the adaptive-weighting scheme and fix $\alpha = 0.2$, fix $\alpha = 0.5$, use a simple dynamic parameter α and utilize a different weighting structure without channel shuffle operation (Zhang et al. 2018) (shown in Fig. 4(4*th*-7*th* column) and Tab. 2(C-F)). Compared to the original SPDS-STFormer, none of the variants achieves the same quality. Notably, our adaptive-weighting scheme extracts a global perceptive for illumination and SBR perception, which is beneficial for robust single-photon depth sensing in different SBR levels.

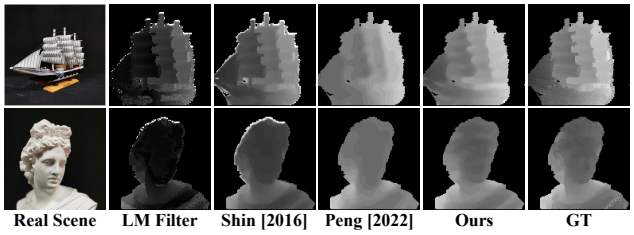

Figure 5: The visual comparisons of different methods in real-world depth sensing for Scene *Boat* (the upper image) and Scene *Statue Apollo* (the lower image).

Real-world Results

We also evaluate our approach on different scenes captured by a self-developed single-photon depth sensing system. We first collect a real-world captured dataset to quantitatively validate the performance of our network on SPAD measurements captured with our single-photon depth sensing system. Our self-collected dataset contains 35 pairs of SPAD measurements and corresponding ground-truth depth. Each SPAD measurement contains 1,024 temporal bins, each with a resolution limited to 4 ps. (More details about the prototype system and our dataset can be seen in Supplemental Material.) Different from evaluating indoor scenes captured by a single-photon imaging prototype in Lindell *et. al.* (Lindell, O'Toole, and Wetzstein 2018) as real-world results in previous works, our self-captured single-photon test set has carefully annotated ground truth depth maps to provide quantitative comparisons for different algorithms. Tab. 3 and Fig. 5 show the quantitative and qualitative comparisons in real-captured scenes. For LM Filter and Shin *et. al.*, they fail to reconstruct fine structures, and for Peng *et. al.* (Peng et al. 2022), they will result in error predictions even in main structure when sensing depth. Our approach can achieve the minimum RMSE metrics down to a millimeter scale,demonstrating that our approach has generalizability and robustness when single-photon measurements are captured with finer temporal resolution to even 4 ps.

Conclusion

In this paper, we propose a 3D spatial-temporal transformer for single photon depth sensing (SPDS-STFormer) to address degraded performance under low photon counts and low SBR conditions. SPDS-STFormer features a view-interweaved attention mechanism and an adaptive-weighting scheme. The attention mechanism adaptively fuses spatial and temporal self-attention to model long-range correlations, while the weighting scheme extracts detailed features for dynamic fusion. Our extensive experimental results show that SPDS-STFormer can effectively estimate the depth information from single-photon measurements. We also validate the generalizability and robustness on our single-photon depth sensing system. When the input measurements are captured at a much finer temporal resolution (4 ps), our approach still has the capability to recover depth with millimeter-to-centimeter accuracy.

Acknowledgements

This work is supported in part by the National Key Research and Development Program of China (No. 2022ZD0210500), the National Natural Science Foundation of China (No. 62332019), the Distinguished Young Scholars Funding of Dalian (No. 2022RJ01), and the Ningbo Major Research and Development Plan Project of China (No. 2023Z225).

References

Aull, B. F.; Loomis, A. H.; Young, D. J.; Heinrichs, R. M.; Felton, B. J.; Daniels, P. J.; and Landers, D. J. 2002. Geiger-mode avalanche photodiodes for three-dimensional imaging. *Lincoln laboratory journal*, 13(2): 335–349.

Baek, S.-H.; and Heide, F. 2022. All-photon polarimetric time-of-flight imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17876–17885.

Cabon, Y.; Murray, N.; and Humenberger, M. 2020. Virtual kitti 2.

Castello, M.; Tortarolo, G.; Buttafava, M.; Deguchi, T.; Villa, F.; Koho, S.; Pesce, L.; Oneto, M.; Pelicci, S.; Lanzanó, L.; et al. 2019. A robust and versatile platform for image scanning microscopy enabling super-resolution FLIM. *Nature methods*, 16(2): 175–178.

Chawla, H.; Varma, A.; Arani, E.; and Zonooz, B. 2022. Transformers in Unsupervised Structure-from-Motion. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics*, 281–303. Springer.

Chen, W.; Kumar, S.; and Yu, F. 2023. Uncertainty-driven dense two-view structure from motion. *IEEE Robotics and Automation Letters*, 8(3): 1763–1770.

Corporation, S. S. S. 2021. Sony to Release a Stacked SPAD Depth Sensor for Automotive LiDAR Applications, an Industry First Contributing to the Safety and Security of Future Mobility with Enhanced Detection and Recognition Capabilities for Automotive LiDAR Applications.

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.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Europe, C. 2023. Canon developing world-first ultra-high-sensitivity ILC equipped with SPAD sensor, supporting precise monitoring through clear color image capture of subjects several km away, even in darkness.

Fossum, E. R.; Ma, J.; Masoodian, S.; Anzagira, L.; and Zizza, R. 2016. The quanta image sensor: Every photon counts. *Sensors*, 16(8): 1260.

Furukawa, Y.; Curless, B.; Seitz, S. M.; and Szeliski, R. 2009. Manhattan-world stereo. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 1422–1429. IEEE.

Geiger, A.; Lenz, P.; Stiller, C.; and Urtasun, R. 2013. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11): 1231–1237.

Guo, W.; Li, Z.; Yang, Y.; Wang, Z.; Taylor, R. H.; Unberath, M.; Yuille, A.; and Li, Y. 2022. Context-enhanced stereo transformer. In *European Conference on Computer Vision*, 263–279. Springer.

Han, X.; Leung, T.; Jia, Y.; Sukthankar, R.; and Berg, A. C. 2015. Matchnet: Unifying feature and metric learning for patch-based matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3279–3286.

Hirschmuller, H. 2007. Stereo processing by semiglobal matching and mutual information. *IEEE Transactions on pattern analysis and machine intelligence*, 30(2): 328–341.

Hu, J.; Shen, L.; and Sun, G. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7132–7141.

Jiang, L.; Shi, S.; Tian, Z.; Lai, X.; Liu, S.; Fu, C.-W.; and Jia, J. 2021. Guided point contrastive learning for semi-supervised point cloud semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6423–6432.

Kirmani, A.; Venkatraman, D.; Shin, D.; Colaço, A.; Wong, F. N.; Shapiro, J. H.; and Goyal, V. K. 2014. First-photon imaging. *Science*, 343(6166): 58–61.

Lai, Z.; Yan, C.; and Fu, Y. 2023. Hybrid Spectral Denoising Transformer with Guided Attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13065–13075.

Lange, R.; and Seitz, P. 2001. Solid-state time-of-flight range camera. *IEEE Journal of quantum electronics*, 37(3): 390–397.

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.; Liu, X.; Drenkow, N.; Ding, A.; Creighton, F. X.; Taylor, R. H.; and Unberath, M. 2021. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6197–6206.

Lindell, D. B.; O'Toole, M.; and Wetzstein, G. 2018. Single-photon 3D imaging with deep sensor fusion. *ACM Trans. Graph.*, 37(4): 113–1.

Liu, J.; Chen, Y.; Ye, X.; Tian, Z.; Tan, X.; and Qi, X. 2022. Spatial pruned sparse convolution for efficient 3d object detection. *Advances in Neural Information Processing Systems*, 35: 6735–6748.

Lowe, D. G. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60: 91–110.

Mei, H.; Dong, B.; Dong, W.; Peers, P.; Yang, X.; Zhang, Q.; and Wei, X. 2021. Depth-aware mirror segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3044–3053.

Meuleman, A.; Kim, H.; Tompkin, J.; and Kim, M. H. 2022. Floatingfusion: Depth from tof and image-stabilized stereo cameras. In *European Conference on Computer Vision*, 602–618. Springer.

Padmanabhan, P.; Zhang, C.; and Charbon, E. 2019. Modeling and analysis of a direct time-of-flight sensor architecture for LiDAR applications. *Sensors*, 19(24): 5464.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Peng, J.; Xiong, Z.; Huang, X.; Li, Z.-P.; Liu, D.; and Xu, F. 2020. Photon-efficient 3d imaging with a non-local neural network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, 225–241. Springer.

Peng, J.; Xiong, Z.; Tan, H.; Huang, X.; Li, Z.-P.; and Xu, F. 2022. Boosting Photon-Efficient Image Reconstruction with A Unified Deep Neural Network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Qiu, D.; Pang, J.; Sun, W.; and Yang, C. 2019. Deep end-to-end alignment and refinement for time-of-flight RGB-D module. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9994–10003.

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.

Rochas, A. 2003. Single photon avalanche diodes in CMOS technology. Technical report, Citeseer.

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.

Shin, D.; Kirmani, A.; Colaço, A.; and Goyal, V. K. 2013. Parametric Poisson process imaging. In *2013 IEEE Global Conference on Signal and Information Processing*, 1053–1056. IEEE.

Shin, D.; Kirmani, A.; Goyal, V. K.; and Shapiro, J. H. 2015. Photon-efficient computational 3-D and reflectivity imaging with single-photon detectors. *IEEE Transactions on Computational Imaging*, 1(2): 112–125.

Shin, D.; Xu, F.; Venkatraman, D.; Lussana, R.; Villa, F.; Zappa, F.; Goyal, V. K.; Wong, F. N.; and Shapiro, J. H. 2016. Photon-efficient imaging with a single-photon camera. *Nature communications*, 7(1): 1–8.

Sun, Z.; Ye, W.; Xiong, J.; Choe, G.; Wang, J.; Su, S.; and Ranjan, R. 2023. Consistent Direct Time-of-Flight Video Depth Super-Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5075–5085.

Tinsley, J. N.; Molodtsov, M. I.; Prevedel, R.; Wartmann, D.; Espigulé-Pons, J.; Lauwers, M.; and Vaziri, A. 2016. Direct detection of a single photon by humans. *Nature communications*, 7(1): 12172.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Velten, A.; Willwacher, T.; Gupta, O.; Veeraraghavan, A.; Bawendi, M. G.; and Raskar, R. 2012. Recovering three-dimensional shape around a corner using ultrafast time-of-flight imaging. *Nature communications*, 3(1): 745.

Wandinger, U. 2005. Introduction to lidar. In *Lidar: range-resolved optical remote sensing of the atmosphere*, 1–18. Springer.

Wei, X.; Zhang, Y.; Li, Z.; Fu, Y.; and Xue, X. 2020. Deepsfm: Structure from motion via deep bundle adjustment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 230–247. Springer.

Zhang, X.; Zhou, X.; Lin, M.; and Sun, J. 2018. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6848–6856.

Zhang, Z. 2012. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2): 4–10.

Zhang, Z.; Dong, B.; Li, T.; Heide, F.; Peers, P.; Yin, B.; and Yang, X. 2023a. Single depth-image 3d reflection symmetry and shape prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8896–8906.

Zhang, Z.; Peng, R.; Hu, Y.; and Wang, R. 2023b. GeoMVSNet: Learning multi-view stereo with geometry perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21508–21518.