Noise-Aware Adaptation of Pre-trained Foundation Models for Single-photon Image Classification

Anonymous authors
Paper under double-blind review

Abstract

Adapting pre-trained foundation models to novel sensor modalities is a fundamental challenge. These models are pre-trained on large RGB datasets that typically lack exposure to the imaging characteristics of other modalities. Physical acquisition effects, such as photon statistics and sensor-specific noise, produce appearance shifts that are underrepresented in pre-training and can degrade transfer performance. We propose a noise-aware adaptation framework that conditions model adaptation on sensor-specific acquisition statistics. Central to our approach is a lightweight Noise Adapter that modulates pre-trained visual features using summary statistics of the sensor's outputs, to decouple acquisition-induced appearance variation from semantics and improve robustness in low-label regimes. We instantiate this idea as a case study on single-photon LiDAR depth images by designing a Noise Adapter that leverages summary statistics computed from raw single-photon histograms for few-shot classification. We also present an exploratory analysis showing how learned modulation patterns correspond to noise-induced feature shifts, providing insight into the adapter's role in feature robustness. Experiments on both synthetic and real single-photon datasets show that our method improves accuracy over baselines, with an average improvement of 3% over the best baseline. These results suggest that explicitly conditioning adaptation on physical acquisition factors is a practical and promising strategy that may generalize to other non-standard modalities.

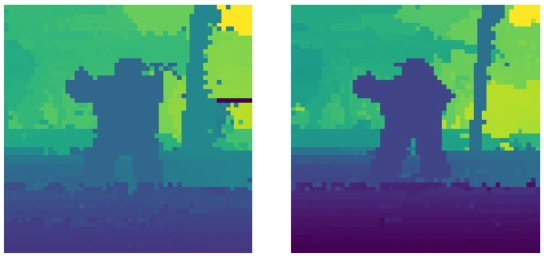
1 Introduction

Foundation models such as CLIP (Radford et al., 2021) and DINO (Caron et al., 2021; Oquab et al., 2024) have demonstrated impressive generalization across a wide range of vision tasks. However, their pre-training is grounded in large-scale RGB datasets, which do not systematically cover the sensing characteristics of non-standard modalities such as thermal cameras or medical imaging, nor of new modalities such as single-photon LiDAR. These domains introduce appearance variations driven by physical acquisition processes (e.g., photon statistics, sensor noise) that are largely absent from pre-training. Additionally, datasets in such modalities are typically small, making it often insufficient to rely solely on fine-tuning to bridge the domain gap. This raises a central challenge: how can we adapt foundation models to new sensing modalities in a way that explicitly incorporates the variability introduced by their physical acquisition conditions?

Single-photon avalanche diode (SPAD) LiDAR provides a representative case study of this challenge. SPAD detectors enable high-precision depth imaging in photon-starved environments by capturing individual photons with picosecond timing resolution (Hadfield, 2009; Halimi et al., 2019; McCarthy et al., 2025; Chan et al., 2024). This makes them particularly attractive for long-range and low-albedo scenarios such as autonomous navigation, robotics, and environmental sensing (Degnan, 2016; Rapp et al., 2020; Shangguan et al., 2023). While existing research has primarily focused on improving geometric reconstruction (Tachella et al., 2019; Malik et al., 2023), there is growing interest in extending SPAD imaging toward semantic tasks such as classification and scene understanding (Suonsivu et al., 2025; Axelsson, 2024; Zhang et al., 2025). However, achieving reliable semantic recognition from SPAD images remains difficult due to both the scarcity of labeled data (Li et al., 2024; Hong et al., 2023) and their strong sensitivity to photon-level noise. Variations in photon counts and signal-to-background ratio (SBR) can drastically alter image appearance, as shown in

Figure 1, introducing imaging condition-driven shifts that models struggle to disentangle from true semantic differences (Suonsivu et al., 2025).

Adapting foundation models to SPAD images offers a promising direction to address these challenges. These models learn transferable representations from massive RGB datasets and can be adapted to downstream tasks with only a few labeled samples (Radford et al., 2021; Li et al., 2023). Despite the modality gap, SPAD images often preserve coarse geometric patterns that can support transferable representations (Auty & Mikolajczyk, 2023; Huang et al., 2023). Several recent adaptation methods have proposed lightweight mechanisms to adapt pre-trained foundation models to downstream tasks (Zhou et al., 2022b; Gao et al., 2024; Zhang et al., 2022). However, these methods primarily adjust features or prompts without explicitly incorporating physical acquisition conditions. As a result, they remain vulnerable to the large intra-class appearance variations induced by SPAD imaging noise. This motivates us to develop a noise-aware adaptation framework tailored to the characteristics of SPAD data.



(a) Low photon count (b) High photon count

Figure 1: Examples of SPAD depth images under varying imaging conditions. Images are drawn from the SPAD real dataset (Zhang et al., 2025) and correspond to the same scene captured with different average photon counts per pixel.

We argue that effective adaptation to new sensing modalities requires making the adaptation process explicitly aware of physical acquisition conditions. We instantiate this principle in SPAD imaging, where photon statistics and SBR dominate appearance variability. To this end, we propose a noise-aware adaptation framework that incorporates physically interpretable descriptors of imaging conditions as priors into the adaptation process. [1] Specifically, we extract the average detected photons per pixel and the estimated SBR from raw histograms and feed them into a Noise Adapter module that modulates frozen pre-trained visual features. This design enables the model to attenuate noise-sensitive feature dimensions while preserving robust semantic channels. Furthermore, we observe that the learned gating pattern correlates with feature sensitivity to imaging noise, which motivates an exploratory feature-level augmentation strategy. Although the gains from this augmentation are modest, it suggests a potential direction for leveraging gating behavior to improve robustness.

Our contributions are three-fold: (1) We propose a modality-aware adaptation framework that conditions feature modulation on imaging condition descriptors, bridging a key gap in adapting foundation models to new sensing modalities. We instantiate this framework in the case of SPAD LiDAR depth images classification, introducing a Noise Adapter that leverages photon statistics and SBR to guide feature modulation and improve robustness. (2) We provide an exploratory analysis of how the learned gating vector captures per-dimension noise sensitivity, and show that this property can be used to guide feature-level augmentation that simulates noise-induced shifts. (3) We validate our approach on both synthetic and real SPAD datasets, where it consistently outperforms existing adapter tuning methods.

2 Related Works

2.1 Semantic Understanding of Single-photon Lidar Data

Single-photon LiDAR systems have primarily been studied for depth estimation and 3D reconstruction under extreme conditions (Legros et al., 2020; Malik et al., 2023; Luo et al., 2025). However, the semantic understanding of SPAD data remains largely unexplored. To promote research in this area, the Time-Resolved MNIST dataset was introduced as a simulated SPAD dataset for single-photon recognition, demonstrating how varying photon flux levels impact CNN recognition accuracy (Suonsivu et al., 2025).

[1] In this paper, "physically interpretable" refers to the input noise descriptors (e.g., photon count and SBR), which have clear physical meaning. The Noise Adapter itself is a learned module that transforms these descriptors into a modulation signal and is not claimed to be fully interpretable.

Many existing methods adapt architectures originally developed for RGB images to noisy, low-resolution SPAD data, addressing tasks like object detection and human activity recognition under photon-starved conditions (MoraMartin et al., 2024; Li et al., 2024). A recent study further integrates active learning to enhance model performance using fewer labeled examples (Zhang et al., 2025). However, these approaches rely on reconstructed depth images, neglecting richer noise information in raw SPAD histograms. Some recent works directly utilize raw SPAD data for classification and detection without intermediate depth reconstruction (Hong et al., 2023; Zhu et al., 2020). Yet data structures of raw SPAD data significantly differ from RGB images, making it difficult to leverage models pre-trained on RGB images.

To bridge this gap, we propose a lightweight noise adapter that incorporates physically interpretable noise descriptors as priors to capture rich information from raw SPAD histograms, while enabling effective adaptation of powerful vision-language models pre-trained on large-scale RGB datasets.

2.2 Lightweight Adaptation of Foundation Models

Recent large-scale foundation models have demonstrated impressive zero-shot and few-shot generalization capabilities by learning robust image-text alignment or self-supervised vision tasks from massive datasets (Radford et al., 2021; Li et al., 2023; Oquab et al., 2024). Existing lightweight adaptation approaches typically fall into two main categories: prompt-tuning methods and adapter tuning.

Prompt tuning methods optimize a small set of learnable parameters to adapt the input of a frozen backbone, either through learnable textual prompts (Zhou et al., 2022b;a) or through visual prompts (Jia et al., 2022; Zeng et al., 2024; Han et al., 2023). Adapter tuning, in contrast, modifies the representations produced by the pre-trained encoder. Some approaches train lightweight modules on top of the frozen backbone (Gao et al., 2024; Huang et al., 2024). Other methods adopt cache-based models that store key-value representations derived from the few-shot training set. These caches can be used either as fixed non-parametric memories or further optimized as learnable components to enhance matching performance at test time (Zhang et al., 2022; Song et al., 2023; Zhu et al., 2023; Udandarao et al., 2023).

Despite these advances, current adaptation methods are predominantly tailored toward RGB imagery. Extending these adaptation frameworks to novel sensor modalities, such as SPAD depth images, poses unique challenges due to the distinct physical noise characteristics. Addressing these challenges requires methods capable of explicitly modeling sensor-specific variability alongside semantic content.

3 Preliminary: Physical Modeling of Single-Photon Noises

3.1 The Physics of Single-photon Imaging

Single-photon LiDAR operates by emitting pulsed laser signals, employs a single-photon avalanche diode (SPAD) to capture echo signals, combines with time-correlated single-photon counting to record temporal information of reflected photons, and derives target depth and reflectivity through analyzing photon statistical characteristics (Hadfield, 2009; McCarthy et al., 2025).

Specifically, for each pixel (i, j) of the target scene, we emit a pulsed laser beam with temporal waveform $s(t)$. The transient photon flux $\Phi_{i,j}(t)$ at the corresponding coordinate can be detected by SPAD arrays, given by Eq. 1:

$$\Phi_{i,j}(t) = \eta\alpha_{i,j}[s(t - \frac{2z_{i,j}}{c}) + b] + d, \tag{1}$$

where η is the quantum efficiency of the single photon detector, $\alpha_{i,j}$ is the reflectivity of the target object, $z_{i,j}$ is the depth of the target object, c is the speed of light, b is the ambient light intensity, and d is the dark count of the detector.

When the light flux is very low, the response of the SPAD detector (photon count histogram) can be considered as a non-uniform Poisson process as Eq. 2:

$$H_{i,j}(k) \sim \mathcal{P}\left(N\Phi_{i,j}(t)\right), \tag{2}$$

where $k = 1, \ldots, K$ represents the time-bins, N is the pulse repetition period (Rapp & Goyal, 2017). Each time bin corresponds to a fixed photon arrival interval determined by the system's bin duration Δt. Upon completing the full scan of the target scene, the histogram data from all pixel positions collectively form a matrix H. Then, single-photon imaging algorithms are used to recover the target's depth and reflectivity from the H (Tachella et al., 2019; Yao et al., 2022).

3.2 Exploiting Physical Noise for Image Understanding

Single-photon LiDAR captures depth information by recording the arrival times of photons at each pixel location as a histogram H. The observed photon counts are influenced by multiple imaging factors, including signal strength, background illumination, and total exposure. These factors vary across conditions and scenes, leading to significant visual changes in the reconstructed depth images. While conventional imaging algorithms (Tachella et al., 2019; Yao et al., 2022) recover depth estimates from H, downstream recognition models that operate solely on these depth maps typically overlook the detailed noise statistics inherent in the raw photon histograms. This limits their ability to distinguish between semantic variations and appearance shifts induced by different noise conditions.

To address this issue, we extract a compact noise embedding from H, capturing two physically interpretable descriptors: the average detected photon count and the estimated signal-to-background ratio (SBR). These descriptors reflect meaningful properties of the imaging process: photon count relates to signal sparsity, while SBR quantifies the proportion of useful signal relative to ambient noise. Crucially, both can be derived from the raw histogram data without requiring additional supervision or calibration. By integrating these descriptors into the downstream adaptation process, we enable the model to condition its predictions on imaging context—helping it decouple semantic information from noise-induced variations. This forms the basis of our noise-aware adaptation framework, detailed in the following sections.

4 Method

In this section, we formalize a framework for adapting a pre-trained foundation model to the unseen SPAD-LiDAR depth image modality. Our goal is to transfer rich semantic knowledge from the pre-training domain (RGB images) to SPAD data, while explicitly accounting for the unique noise characteristics of SPAD imaging. We address two key challenges: (1) the substantial distribution shift between the pre-training domain (natural RGB images) and the target SPAD depth images, and (2) strong intra-class appearance variations in SPAD images caused by photon-level noise under varying imaging conditions.

To tackle these challenges, we propose a noise-aware adaptation framework consisting of three main components: (1) Noise-Embedding Extraction, (2) Noise Adapter, and (3) Gate-Guided Feature Augmentation (GGFA). In the following subsections, we first formalize the problem setup and present the overall framework, and then describe each component in detail.

4.1 Overall Framework

We denote the available SPAD dataset as $\mathcal{D} = (H^i, I^i, y^i)_{i=1}^N$, where $H^i \in \mathbb{R}^{M\times W\times B}$ is the raw SPAD histogram, $I^i \in \mathbb{R}^{M\times W}$ is the corresponding depth image reconstructed from H^i via a SPAD imaging algorithm (Yao et al., 2022), and y^i is the associated label. Each histogram records photon counts over B time bins at each of the $M \times W$ pixels. The reconstructed depth image I^i serves as the input to the visual encoder, while the raw histogram H^i is used to extract noise embeddings.

Our method leverages the strong semantic priors encoded in the frozen pre-trained visual encoder $f_V(\cdot)$, while explicitly modeling the impact of noise. The overall framework is illustrated in Figure 2. For each SPAD image, we first extract a global noise embedding z from the raw histogram H^i, capturing key imaging

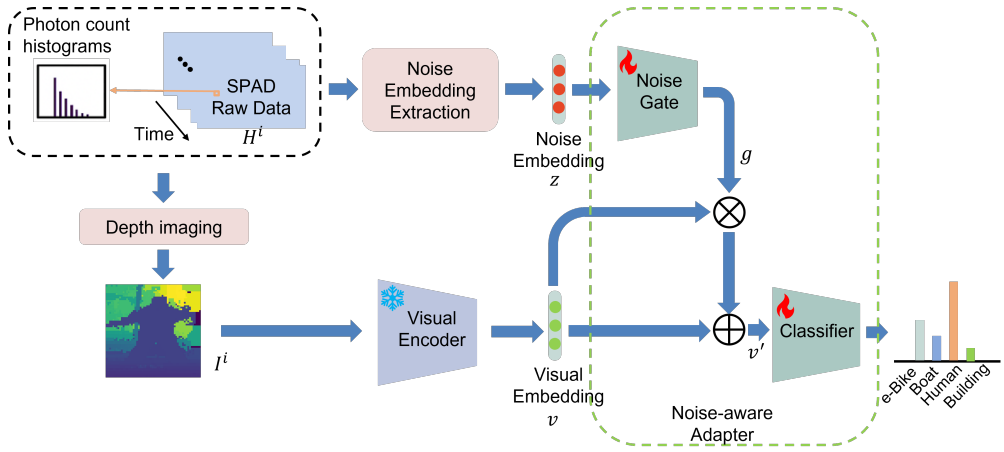


Figure 2: Overview of our noise-aware adaptation framework for SPAD depth images. A noise embedding extracted from the SPAD raw data modulates visual features obtained from the depth image via a noise-gated adapter. The modulated features are used for classification.

condition statistics. This noise embedding is then used to modulate the visual feature $v = f_V(I^i)$ through a learnable gating mechanism, producing a noise-aware feature representation. Additionally, we introduce a Gate-Guided Feature Augmentation (GGFA) strategy that generates realistic feature perturbations guided by the learned gating behavior, further improving model robustness to noise-induced variation. In the following sections, we describe each component of our method in detail.

4.2 Noise Embedding Extraction

We first describe how to extract a noise embedding from the raw SPAD histogram, summarizing each sample's imaging condition as input to the noise adapter. While many statistics can be derived from a SPAD histogram (e.g., variance, skewness, higher-order moments), we use two global descriptors: average photon count per pixel and signal-to-background ratio (SBR). In SPAD imaging, once the scene geometry, reflectance, and LiDAR acquisition parameters (e.g., pulse repetition period, bin duration, detector efficiency, and system impulse response) are fixed, the per-pixel photon arrival statistics are largely determined by the detected photon number and the signal-to-background ratio (Rapp & Goyal, 2017). These two quantities thus provide a compact and physically grounded summary of the dominant noise conditions. Other descriptors are possible, but we select these statistics for their concise coverage of key variability factors.

Given a SPAD histogram H^i, we compute a global noise embedding $z \in \mathbb{R}^2$ consisting of two scalar statistics: the average number of detected photons per pixel, and an estimated SBR. The average photon count per pixel, $\bar{c}_{\text{pix}}$, is computed as the total number of detected photons divided by the number of pixels ($M \times W$), providing a coarse measure of signal strength and imaging quality. Throughout this paper, we use the overline notation (e.g., $\bar{c}_{\text{pix}}$) to indicate spatial averaging across all pixels.

To estimate SBR, we leverage the depth image I^i reconstructed from H^i using a SPAD imaging algorithm (Yao et al., 2022). For each pixel (x, y), we first identify the expected signal arrival bin as $p(x, y) = \text{round}(2I^i(x, y)/(c\Delta t))$, where c denotes the speed of light and Δt is the bin duration. We then define a signal window of size w centered around $p(x, y)$ and compute the number of signal photons as Eq. 3:

$$c_{\text{sig}}(x, y) = \sum_{b=p(x,y)-w}^{p(x,y)+w} H^i(x, y, b). \tag{3}$$

The background photon count is computed by subtracting the estimated signal photons from the total count at each pixel, i.e., $c_{\text{bg}}(x, y) = \sum_{b=1}^{B} H^i(x, y, b) - c_{\text{sig}}(x, y)$. Then, the per-pixel SBR is computed as Eq. 4:

$$\mathrm{SBR}(x,y) = \frac{c_{\mathrm{sig}}(x,y)}{c_{\mathrm{bg}}(x,y) + \epsilon}, \tag{4}$$

where ϵ is a small constant to prevent division by zero. The global SBR is obtained by averaging SBR(x,y) across all pixels, denoted as $\overline{\mathrm{SBR}}$. Finally, the global noise embedding is defined as $z = \left[\bar{c}_{\mathrm{pix}};\quad \overline{\mathrm{SBR}}\right]$. The resulting noise embedding z enables the noise adapter to modulate visual features according to the imaging conditions.

4.3 Noise Adapter

SPAD depth images can be viewed as a superposition of two factors with different generative origins: a stable semantic structure determined by scene geometry and object identity, and stochastic variations induced by photon-limited imaging conditions. Vision encoders pre-trained on natural RGB images capture semantic structure but have never been exposed to SPAD-specific variations. As a result, they lack robustness to the condition-induced shifts characteristic of photon-limited imaging. Our goal is therefore to disentangle the two components, preserving semantics while mitigating noise-driven variability.

To this end, we introduce a Noise Adapter that conditions feature adaptation on a noise embedding z. Given a pre-trained visual feature $v \in \mathbb{R}^D$ and $z \in \mathbb{R}^2$, we compute a per-dimension gating vector $g \in [0,1]^D$ via a multilayer perceptron (MLP): $g = \mathrm{MLP}(z)$. The adapted representation is then obtained as Eq. 5:

$$v' = v \odot g + v, \tag{5}$$

where $\odot$ denotes element-wise multiplication. This formulation splits the representation into two complementary paths: an identity path v that preserves the pre-trained semantic baseline, and a residual path $v \odot g$ that applies a condition-specific correction. Intuitively, the gating vector learns to attenuate feature dimensions sensitive to noise while preserving robust semantic channels, and the residual connection prevents the adapter from overwriting useful representations. Consequently, the model learns lightweight, noise-aware adjustments rather than rediscovering an appropriate representation.

Finally, the modulated feature v' is passed through a classification head to produce semantic predictions. By conditioning feature modulation on physically grounded descriptors, the Noise Adapter enhances robustness to appearance variations and improves the generalization of pre-trained models to the previously unseen SPAD modality.

4.4 Gate-Guided Feature Augmentation

We observe that the learned gating vector g implicitly captures the sensitivity of different feature dimensions to variations in imaging conditions. Specifically, dimensions with larger average gate values tend to exhibit smaller variations across noise levels. This observation is supported by empirical analysis in Sec. 5.3.2, where we show a negative correlation between the average gate value and the standard deviation of CLIP features across different noise levels.

Motivated by this, we propose a Gate-Guided Feature Augmentation (GGFA) strategy to further improve model robustness under limited supervision. The key idea is to inject feature perturbations in a gate-informed manner, where the perturbation strength for each feature dimension is proportional to $(1 - g)$, encouraging the model to be robust to variations from imaging conditions.

The overall GGFA workflow is as follows. We first train the Noise Adapter without GGFA, obtaining a learned noise gate. In a second training stage, we freeze the noise gate and use it to guide feature augmentation. Specifically, for each original visual feature v, we add controlled random noise to generate an augmented feature as Eq. 6:

$$v_{\mathrm{aug}}[i] = v[i] + \alpha \cdot (1 - g[i]) \cdot \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0,1), \tag{6}$$

where α is a global scaling factor that controls the overall perturbation strength, and $(1 - g)$ determines the dimension-wise noise scaling based on the gate. Since the gating network will further affect the noise distribution, we pass the perturbed feature v_{aug} through the noise gate to obtain the final augmented modulated feature, $v'_{\text{aug}} = v_{\text{aug}} \odot g + v_{\text{aug}}$.

During this augmentation phase, we freeze the noise gate parameters and only train the classifier on both original and augmented features. For each augmented feature, we randomly sample a noise embedding to generate the corresponding gate vector g. This design prevents the degradation of the previously learned noise-aware modulation and ensures that the classifier is exposed to a broader range of noise-conditioned feature variations. The training objective is a standard cross-entropy loss computed over both the original and augmented samples in each iteration. This two-stage strategy enables the model to learn robust decision boundaries that generalize better across a spectrum of realistic noise-induced feature variations, while preserving the noise-aware feature modulation introduced by the Noise Adapter.

5 Experiments

5.1 Experimental Setup

Dastaset: We evaluate our method on both synthetic and real SPAD datasets comprising 11 categories introduced in the prior work (Zhang et al., 2025). The **synthetic SPAD dataset** is generated by simulating single-photon histograms from RGB-D images. It contains about 33,000 samples across 11 object classes. To align the spatial sparsity of SPAD imaging, all simulated images are downsampled to a resolution of 128×72 pixels. The **real SPAD dataset** contains 4,400 reconstructed depth images spanning the same 11 categories as the synthetic dataset. Each class includes 400 single-photon samples. The depth image is reconstructed using the SSPI algorithm (Yao et al., 2022) with a resolution of 64×64 pixels.

Implementation Details: We evaluate our method on both synthetic and real single-photon datasets, splitting each dataset into training, validation, and test subsets with a ratio of 0.6:0.1:0.3. All methods use the pre-trained CLIP ViT-B/32 model as the visual feature encoder. To ensure a fair comparison, all trainable models are optimized using the AdamW optimizer with an initial learning rate of 0.001, batch size of 64, and a cosine annealing learning rate schedule over 100 epochs.

Following prior work, we evaluate performance under different data-scarce scenarios by selecting 1, 2, 4, 8, and 16 labeled samples per class for training (Zhang et al., 2022; Gao et al., 2024). Each experiment is repeated with 5 different random seeds. For each run, the data split is regenerated following the specified ratio. We report the mean classification accuracy and standard deviation across these runs.

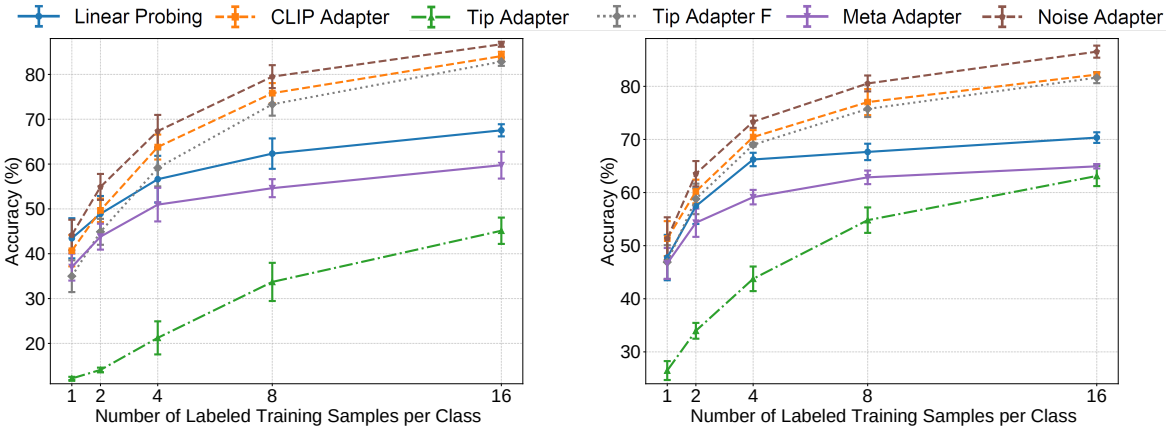
In our proposed Noise Adapter, both the noise-gating module and the classification head are implemented as two-layer MLPs. The hidden dimension of each MLP is set equal to the feature dimension of the CLIP visual encoder output.

Baseline: (1) Linear-probing: trains a linear classifier on top of the frozen visual encoder using the few-shot labeled samples; (2) Tip-Adapter: a training-free adapter that builds a key-value cache from support features and combines visual-textual similarities by tuning their weights on the validation set (Zhang et al., 2022); (3) Tip-Adapter-F: fine-tunes the cached visual embeddings while also searching for the optimal combination of visual and textual similarity weights using the validation set (Zhang et al., 2022); (4) CLIP-Adapter: trains a two-layer MLP to adapt visual features and combines them with CLIP text features, weights are tuned on the validation set for best fusion (Gao et al., 2024); (5) Meta-Adapter utilizes meta-testing mechanism and a lightweight adapter (Song et al., 2023).

5.2 Results and Analysis

Figure 3 shows the few-shot classification performance of various methods on both the synthetic SPAD dataset (Figure 3a) and the real SPAD dataset (Figure 3b). Our proposed Noise Adapter consistently outperforms all baseline methods across different numbers of labeled samples per class, especially when more

training examples are available. Due to the domain shift between the RGB data used to pre-train CLIP and the SPAD depth images in our task, the zero-shot accuracy of CLIP is relatively low (Table 4 and Table 5). This domain gap also negatively impacts methods that rely heavily on CLIP's text embeddings, such as Meta-Adapter, which shows limited performance improvements even as the number of labeled samples increases. In addition, the large modality gap makes it difficult for training-free approaches like Tip-Adapter or methods that only apply minimal changes to the pre-trained features (e.g., Linear Probing), to adapt effectively, leading to suboptimal accuracy. In contrast, CLIP-Adapter and Tip-Adapter-F, which allow for more flexible adaptation of CLIP features, achieve relatively better performance.



(a) Syn (b) Real

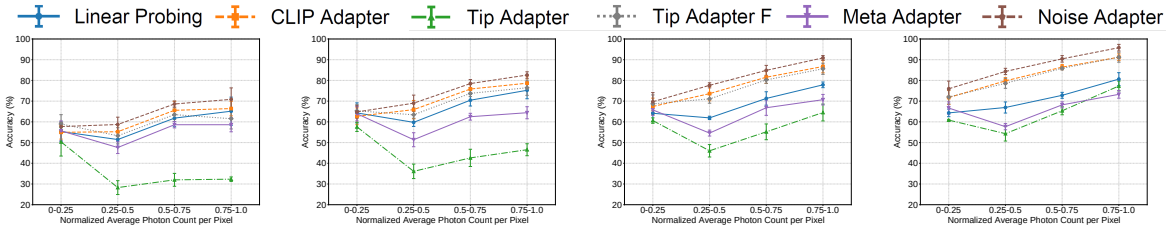
Figure 3: Few-shot classification accuracy comparison on (a) synthetic and (b) real SPAD datasets. Each curve shows the average accuracy and standard deviation over 5 random trials.

Our Noise Adapter adopts a similar structure to CLIP-Adapter, utilizing a two-layer MLP to adapt visual features. However, it further enhances adaptation by incorporating noise embeddings extracted from SPAD raw data. These noise embeddings encode information about imaging conditions, which helps the adapter distinguish between noise-induced and semantics-induced variations in the images. As a result, the model achieves higher accuracy.

As shown in Figure 3, Noise-Adapter performs comparably to other trainable adapters in the 1-shot setting, as limited data makes it challenging to learn meaningful noise-conditioned representations. However, as more training data becomes available, Noise-Adapter demonstrates significant performance gains, highlighting the benefit of modeling noise explicitly in this domain.

Performance of Different Noise Levels We further evaluate the robustness of different methods under varying noise levels on the real SPAD depth image dataset. Since the quality of SPAD depth images is highly correlated with the average number of detected photons per pixel, we partition the test samples into four groups based on their normalized average photon count. Lower photon counts (e.g., 0–0.25) correspond to lower imaging quality and higher noise levels. Figure 4 shows the performance of all methods across different photon count ranges under 2-shot, 4-shot, 8-shot, and 16-shot settings. Across all methods, prediction accuracy generally improves as photon count increases, confirming that photon noise significantly impacts classification performance in SPAD depth images.

In the low-shot settings (2-shot and 4-shot), our proposed Noise Adapter exhibits clear advantages over baselines in the higher photon count ranges (0.5–1.0), but performs comparably to baselines in the lowest photon count range (0–0.25). We attribute this to the limited number of labeled samples constraining the model's ability to fully learn the complex interaction between noise characteristics and semantic structure in very noisy samples. However, the explicit incorporation of noise embeddings provides the adapter with valuable information about imaging conditions, helping to improve its overall accuracy.



(a) 2 shots (b) 4 shots (c) 8 shots (d) 16 shots

Figure 4: Classification accuracy of our method and baselines under different noise levels of the real SPAD dataset. Each subfigure corresponds to a different few-shot setting ((a) 2, (b) 4, (c) 8, and (d) 16 shots), while the x-axis indicates four photon count intervals that represent increasing image quality. Results are averaged over 5 random trials.

As the number of labeled samples increases (8-shot and 16-shot), Noise Adapter consistently achieves the best accuracy across all photon count ranges. With more supervision, the model can learn more effectively how to disentangle noise-induced variations from semantic content, enabling better generalization across both low- and high-noise samples. These results validate our hypothesis that incorporating explicit noise-awareness into the adaptation process is critical for recognizing SPAD depth images that are highly sensitive to variations in imaging conditions.

5.3 Ablation Study

To better understand the contributions of different components in our method, we conduct a series of ablation studies on the syn/real SPAD depth image datasets. The ablations are organized into two parts: model architectures and feature augmentations. In addition, we provide extended analysis in the appendix, including results across different pre-trained backbones (Appendix C) and a deeper investigation of GGFA (Appendix D).

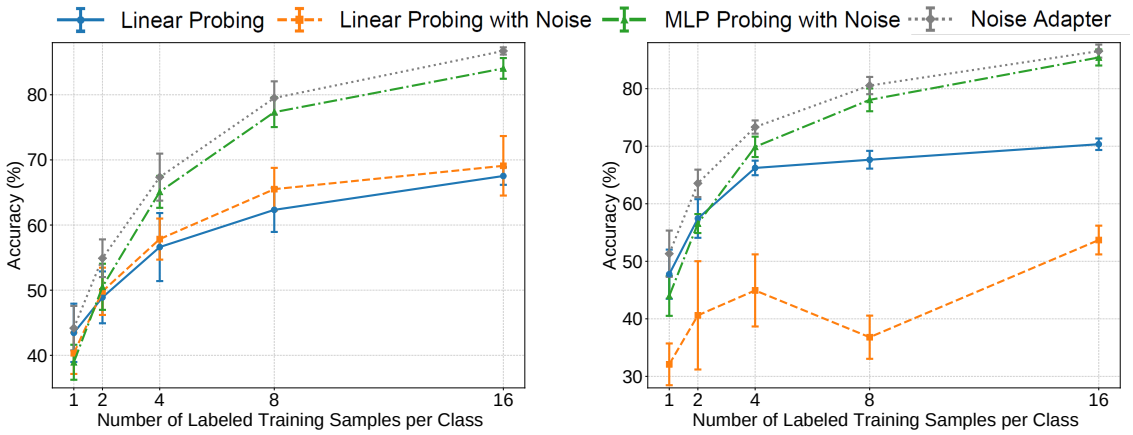
5.3.1 Analysis of Noise Adapter architecture

First, we analyze the impact of different model architectures for incorporating noise information. We compare (1) a linear classifier trained on the CLIP visual features v, (2) a linear classifier trained on concatenated visual features and noise embeddings $[v, z]$, (3) a MLP classifier trained on the same concatenated features, and (4) our proposed Noise Adapter, which uses a noise gate to map the noise embedding into a gating vector that modulates the visual features.

As shown in Figure 5, the worst-performing variant is linear probing with noise embedding, which even underperforms standard linear probing. This suggests that directly concatenating noise embeddings to CLIP features can harm the representation, especially when the classifier lacks sufficient capacity to compensate for this disturbance (e.g. Linear probing).

MLP probing with noise performs better and is able to exploit the additional information provided by noise embeddings. However, under few-shot settings (e.g., 1-shot and 2-shot), it tends to overfit the specific noise conditions in the training samples, leading to suboptimal generalization. This result likely stems from its reliance on learning complex mappings from limited data, without a structured mechanism to separate noise-related variations from semantic cues.

In contrast, our Noise Adapter achieves consistently strong results. Under limited labeled samples, it performs on par with or better than the linear probing baseline, indicating that it preserves the semantic structure of CLIP features. As the number of labeled samples increases, the Noise Adapter shows more substantial improvements over all other variants. This demonstrates the benefit of its design: by using noise embeddings to modulate the visual features through a learnable gating mechanism, it allows the model to suppress noise-induced variations while retaining relevant semantic information. Overall, these results high-

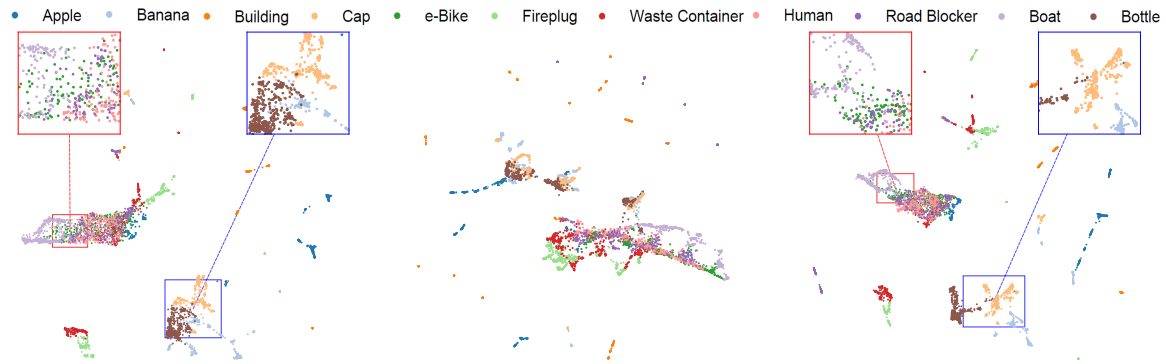


(a) Syn (b) Real

Figure 5: Ablation study comparing different ways of incorporating noise information into the classification pipeline. We compare: (1) Linear Probing, (2) Linear Probing with Noise, which concatenates noise embeddings to visual features before linear classification; (3) MLP Probing with Noise, which uses an MLP on concatenated features; and (4) our proposed Noise Adapter on (a) Syn SPAD dataset and (b) Real SPAD dataset.

light the importance of both incorporating noise-aware signals and using them in a structured, learnable way to enhance model robustness under varying imaging conditions.

Feature Visualization To further analyze how different ways of integrating noise information impact feature representations, we visualize the features produced by different variants using PaCMAP (Wang et al., 2021) in the real SPAD depth image dataset, as shown in Figure 6.



(a) CLIP visual feature v (b) Concatenated feature $[v, z]$ (c) Modulated features v'

Figure 6: PaCMAP visualization of different feature representations on the real SPAD dataset. (a) Original CLIP visual features v, (b) features obtained by concatenating CLIP features with noise embeddings $[v, z]$, and (c) features v' produced by our Noise Adapter.

In Figure 6a, the original CLIP visual features demonstrate good generalization despite being pre-trained on natural images, which are substantially different from SPAD depth images. Distinct classes form well-separated clusters, indicating that the pre-trained CLIP encoder captures transferable semantic information even under the SPAD modality.

However, as shown in Figure 6b, directly concatenating noise embeddings with CLIP visual features introduces degradation in features. The added noise embedding increases sensitivity to imaging condition

variations, causing features of the same semantic class to fragment into multiple sub-clusters corresponding to different noise levels. Moreover, certain classes become less separable and exhibit overlap with other categories, suggesting that naive concatenation can amplify nuisance variation rather than helping the model to disambiguate it.

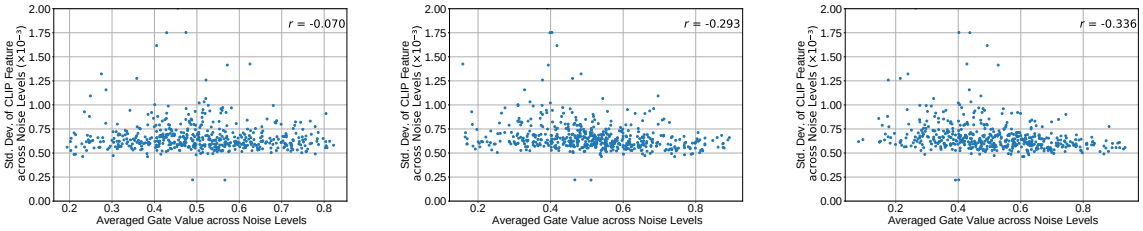
In contrast, Figure 6c shows that our Noise Adapter effectively leverages the noise embeddings through a learned gating mechanism, improving the quality of the feature space. The modulated features exhibit improved inter-class separation, indicating that the adapter successfully helps to disentangle noise-induced variation from semantic structure. This supports our design choice of using a gating-based modulation instead of direct feature concatenation, allowing the model to better preserve semantic information while adapting to varying imaging conditions.

5.3.2 Analysis of Gate-Guided Feature Augmentation

Second, we evaluate the effect of different feature augmentation strategies. We compare (1) no feature augmentation, (2) adding random noise with fixed standard deviation to the modulated features v', (3) adding random noise with fixed standard deviation to the original visual features v, and then applying the learned noise gate to produce augmented modulated features, and (4) our full method, where the noise gate's mean values are used to determine dimension-wise noise scaling for feature augmentation.

Correlation between Gate and Feature across various noise levels We first analyze how the learned gating vector g relates to the noise sensitivity of different feature dimensions. Figure 7 shows the correlation between the average gate value and the standard deviation of CLIP visual features across different noise levels, under 1-shot, 4-shot, and 16-shot settings. Each point corresponds to one feature dimension.

A negative correlation is observed, indicating that dimensions with lower gate values tend to exhibit higher variability across noise conditions. This suggests that the gating mechanism implicitly captures per-dimension noise sensitivity, with g modulating feature robustness accordingly.



(a) 1 shot (b) 4 shots (c) 16 shots

Figure 7: Correlation between the averaged gate value and the standard deviation of CLIP features across different noise levels. Each point corresponds to one feature dimension. A negative correlation is observed, indicating that the gating mechanism tends to assign lower gate values to dimensions with higher feature variability. (a), (b), and (c) show the results under the 1-shot, 4-shot, and 16-shot settings, respectively. The correlation coefficient is computed using Spearman's rank correlation.

As shown in Figure 7, this correlation (Spearman's rank correlation (Zwillinger & Kokoska, 1999)) becomes stronger as the number of labeled samples increases. In the 1-shot setting, the correlation is weak ($r = -0.070$), likely due to insufficient supervision. However, with 4-shot and 16-shot settings, the negative correlation strengthens ($r = -0.293$ and $r = -0.336$, respectively), suggesting that the model progressively learns to align the gating pattern with feature sensitivity as more data becomes available.

These results provide empirical support for the design of our Gate-Guided Feature Augmentation (GGFA) strategy, where $(1 - g)$ is used to modulate the strength of feature perturbations. By leveraging the learned gate pattern, GGFA introduces noise-consistent variations during training, making the augmented features more closely resemble the actual feature variations observed under different imaging conditions. This encourages the classifier to generalize better to real noise-induced feature shifts.

Effect of gate-guided feature augmentation We next analyze the impact of different feature augmentation strategies on the performance of the Noise Adapter. The experiments compare variants with and without feature augmentation, as well as different noise injection schemes designed to improve the robustness of the modulated features. As shown in Table 1 and Table 2, the benefits of GGFA are more pronounced in low-shot scenarios (1-shot and 2-shot), where limited supervision makes it more important to expose the model to a diverse range of noise-conditioned feature variations. In higher-shot settings, the model can already learn such variations from the data itself, diminishing the impact of the feature augmentation.

Table 1: Ablation study of Gate-Guided Feature Augmentation (GGFA) on the SPAD real dataset. Results show average accuracy with standard deviation (%) over 5 trials. The best results are shown in blue.

Few-shot Setup	1	2	4	8	16
w/o GGFA	50.39 ± 4.75	62.95 ± 2.31	73.09 ± 0.95	80.14 ± 1.68	86.61 ± 1.19
Rand	49.55 ± 5.11	62.68 ± 2.26	72.68 ± 0.62	80.36 ± 2.08	86.26 ± 1.18
Rand with Gate	51.02 ± 4.37	63.27 ± 2.53	72.97 ± 0.98	80.42 ± 2.64	86.21 ± 1.38
GGFA	51.24 ± 4.48	63.18 ± 2.70	73.15 ± 0.52	80.45 ± 2.42	86.68 ± 1.28

Results show that adding random noise directly to the modulated features (Rand) does not yield noticeable performance gains. We attribute this to the fact that such augmented features deviate substantially from the true distribution of noise-modulated features. In contrast, injecting noise at the original feature level followed by gating (noise gate) produces augmented features that better match the characteristics of the modulated feature space, leading to more consistent performance improvements across both synthetic and real datasets.

Using gate-aware scaling (GGFA) further refines the augmentation process by adapting the noise strength based on the learned sensitivity of each feature dimension. However, this brings only modest gains. One reason is that the correlation between the gate values and the actual feature variance across noise levels is relatively weak (Spearman correlation ≈ -0.33), as shown in Figure 7, and stronger relationships require more training data to emerge. Another factor is that the range of observed per-dimension feature variance across gate value is relatively small (typically varying from approximately 0.5×10^{-3} to 0.75×10^{-3}), which limits the potential benefit of fine-grained noise scaling. Overall, these results suggest that the primary contribution of our framework lies in the Noise Adapter, while GGFA should be viewed as an exploratory extension. Although its gains are modest, GGFA offers useful insights into how the learned gate relates to feature sensitivity and provides a promising direction for future work on noise-aware augmentation.

6 Conclusion

Foundation models have demonstrated strong generalization in conventional vision domains, but adapting them to non-standard sensing modalities remains challenging. These modalities, such as thermal cameras, medical imaging, or SPAD LiDAR, exhibit appearance variations tied to underlying physical acquisition conditions that are absent from large-scale RGB pre-training. Addressing this gap calls for adaptation strategies that are explicitly aware of modality-specific noise.

In this work, we proposed a noise-aware adaptation framework and instantiated it in the context of SPAD imaging. The central component is a Noise Adapter that incorporates physically interpretable noise descriptors as priors to modulate pre-trained visual features, thereby attenuating noise-sensitive channels while preserving semantic structure. This yields improved robustness to condition-induced variability and consistent performance gains across both synthetic and real SPAD datasets. Beyond this main design, we also examined how the learned gating pattern reflects feature sensitivity and explored its use for feature-level augmentation. While the improvements from this augmentation are modest, the analysis provides additional insight into the interplay between noise-aware modulation and feature-level data augmentation.

Taken together, our findings illustrate that explicitly conditioning adaptation on acquisition factors can be an effective strategy for transferring foundation models to sensing modalities unseen during pre-training. SPAD serves here as a case study, demonstrating the value of lightweight, noise-aware modules in bridging

the gap between large-scale pre-training and physics-driven data domains. More broadly, this perspective highlights the potential of modality-aware adaptation as a practical route for extending foundation models beyond natural RGB imagery into photon-limited or otherwise non-standard modalities.

7 Limitations and Future Directions

While our approach demonstrates strong performance improvements for SPAD depth image understanding, several limitations remain. First, the SPAD histogram data inherently contains richer information beyond what is captured by the two simple global statistics, average photon count and signal-to-background ratio, used in our current noise embedding. More sophisticated representations of the raw histogram could provide additional cues about imaging quality and noise characteristics, and may further improve adaptation performance. Exploring how to effectively incorporate this richer information and better align it with pre-trained RGB-based foundation models is an important direction for future research.

Second, the quality of SPAD depth images at different noise levels is also influenced by the choice of the SPAD imaging algorithm. In this work, we rely on the depth images provided by the public dataset and do not analyze how different reconstruction methods affect the learned features or the adaptation process. Studying the interaction between SPAD imaging algorithms and adaptation remains an open area for further exploration.

References

Dylan Auty and Krystian Mikolajczyk. Learning to prompt clip for monocular depth estimation: Exploring the limits of human language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2039–2047, 2023.

Maria Axelsson. Semantic segmentation of persons in point clouds from photon counting lidar. In *2024 IEEE International Conference on Imaging Systems and Techniques (IST)*, pp. 1–6. IEEE, 2024.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Stanley H Chan, Hashan K Weerasooriya, Weijian Zhang, Pamela Abshire, Istvan Gyongy, and Robert K Henderson. Resolution limit of single-photon lidar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25307–25316, 2024.

John J Degnan. Scanning, multibeam, single photon lidars for rapid, large scale, high resolution, topographic and bathymetric mapping. *Remote Sensing*, 8(11):958, 2016.

Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision*, 132(2):581–595, 2024.

Robert H Hadfield. Single-photon detectors for optical quantum information applications. *Nature Photonics*, 3(12):696–705, 2009.

Abderrahim Halimi, Rachael Tobin, Aongus McCarthy, Jose Bioucas-Dias, Stephen McLaughlin, and Gerald S Buller. Robust restoration of sparse multidimensional single-photon lidar images. *IEEE Transactions on Computational Imaging*, 6:138–152, 2019.

Cheng Han, Qifan Wang, Yiming Cui, Zhiwen Cao, Wenguan Wang, Siyuan Qi, and Dongfang Liu. E2vpt: An effective and efficient approach for visual prompt tuning. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17445–17456, 2023.

Yu Hong, Yuxiao Li, Chen Dai, Jun-Tian Ye, Xin Huang, and Feihu Xu. Image-free target identification using a single-point single-photon lidar. *Optics Express*, 31(19):30390–30401, 2023.

Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson WH Lau, Wanli Ouyang, and Wangmeng Zuo. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22157–22167, 2023.

Yunshi Huang, Fereshteh Shakeri, Jose Dolz, Malik Boudiaf, Houda Bahig, and Ismail Ben Ayed. Lp++: A surprisingly strong linear probe for few-shot clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23773–23782, 2024.

Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pp. 709–727. Springer, 2022.

Quentin Legros, Sylvain Meignen, Stephen McLaughlin, and Yoann Altmann. Expectation-maximization based approach to 3D reconstruction from single-waveform multispectral lidar data. *IEEE Transactions on Computational Imaging*, 6:1033–1043, 2020.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pp. 19730–19742. PMLR, 2023.

Xiaozhe Li, Jinyi Liu, Guoyang Zhao, Lijun Liu, Weiping Zhang, Xiaomin Hu, and Shuming Cheng. High precision single-photon object detection via deep neural networks. *Optics Express*, 32(21):37224–37237, 2024.

Weihan Luo, Anagh Malik, and David B Lindell. Transientangelo: Few-viewpoint surface reconstruction using single-photon lidar. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8723–8733. IEEE, 2025.

Anagh Malik, Parsa Mirdehghan, Sotiris Nousias, Kyros Kutulakos, and David Lindell. Transient neural radiance fields for lidar view synthesis and 3d reconstruction. *Advances in Neural Information Processing Systems*, 36:71569–71581, 2023.

Aongus McCarthy, Gregor G Taylor, Jorge Garcia-Armenta, Boris Korzh, Dmitry V Morozov, Andrew D Beyer, Ryan M Briggs, Jason P Allmaras, Bruce Bumble, Marco Colangelo, et al. High-resolution long-distance depth imaging lidar with ultra-low timing jitter superconducting nanowire single-photon detectors. *Optica*, 12(2):168–177, 2025.

German MoraMartin, Stirling Scholes, Robert K Henderson, Jonathan Leach, and Istvan Gyongy. Human activity recognition using a single-photon direct time-of-flight sensor. *Optics Express*, 32(10):16645–16656, 2024.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024, 2024. URL `https://openreview.net/forum?id=a68SUt6zFt`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

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

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

Mingjia Shangguan, Zhifeng Yang, Zaifa Lin, Zhongping Lee, Haiyun Xia, and Zhenwu Weng. Compact long-range single-photon underwater lidar with high spatial–temporal resolution. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023.

Lin Song, Ruoyi Xue, Hang Wang, Hongbin Sun, Yixiao Ge, Ying Shan, et al. Meta-adapter: An online few-shot learner for vision-language model. *Advances in Neural Information Processing Systems*, 36: 55361–55374, 2023.

Aleksi Suonsivu, Lauri Salmela, Edoardo Peretti, Leevi Uosukainen, Radu Ciprian Bilcu, and Giacomo Boracchi. Time-resolved mnist dataset for single-photon recognition. In *European Conference on Computer Vision*, pp. 127–143. Springer, 2025.

Julián Tachella, Yoann Altmann, Nicolas Mellado, Aongus McCarthy, Rachael Tobin, Gerald S Buller, Jean-Yves Tourneret, and Stephen McLaughlin. Real-time 3D reconstruction from single-photon lidar data using plug-and-play point cloud denoisers. *Nature Communications*, 10(1):4984, 2019.

Vishaal Udandarao, Ankush Gupta, and Samuel Albanie. Sus-x: Training-free name-only transfer of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2725–2736, 2023.

Yingfan Wang, Haiyang Huang, Cynthia Rudin, and Yaron Shaposhnik. Understanding how dimension reduction tools work: an empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization. *Journal of Machine Learning Research*, 22(201):1–73, 2021.

Gongxin Yao, Yiwei Chen, Chen Jiang, Yixin Xuan, Xiaomin Hu, Yong Liu, and Yu Pan. Dynamic single-photon 3D imaging with a sparsity-based neural network. *Optics Express*, 30(21):37323–37340, 2022.

Runjia Zeng, Cheng Han, Qifan Wang, Chunshu Wu, Tong Geng, Lifu Huangg, Ying Nian Wu, and Dongfang Liu. Visual fourier prompt tuning. *Advances in Neural Information Processing Systems*, 37:5552–5585, 2024.

Renrui Zhang, Wei Zhang, Rongyao Fang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free adaption of clip for few-shot classification. In *European Conference on Computer Vision*, pp. 493–510. Springer, 2022.

Zili Zhang, Ziting Wen, Yiheng Qiang, Hongzhou Dong, Wenle Dong, Xinyang Li, Xiaofan Wang, and Xiaoqiang Ren. Label-efficient single photon images classification via active learning. *arXiv preprint arXiv:2505.04376*, 2025.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13001–13008, 2020.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pp. 16816–16825, 2022a.

Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022b.

Xiangyang Zhu, Renrui Zhang, Bowei He, Aojun Zhou, Dong Wang, Bin Zhao, and Peng Gao. Not all features matter: Enhancing few-shot clip with adaptive prior refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2605–2615, 2023.

Yan Zhu, Jianhong Shi, Xiaoyan Wu, Xialin Liu, Guihua Zeng, Jun Sun, Lulu Tian, and Feng Su. Photon-limited non-imaging object detection and classification based on single-pixel imaging system. *Applied Physics B*, 126(1):21, 2020.

Daniel Zwillinger and Stephen Kokoska. *CRC standard probability and statistics tables and formulae*. Crc Press, 1999. See Section 14.7.

A Broader Impact Statement

This work aims to enable semantic understanding of SPAD LiDAR depth images by adapting large-scale pre-trained vision-language models, such as CLIP, using a novel noise-aware adapter. SPAD LiDAR systems offer unique capabilities for depth imaging under extremely low-light or long-range conditions, making them attractive for applications such as autonomous navigation and remote sensing. However, the semantic interpretation of SPAD data remains a significant challenge due to photon-level noise and the scarcity of data. Our proposed method improves robustness and label efficiency in this setting.

The broader impact of this work is twofold:

Positive Impact: By bridging the gap between pre-trained vision-language models and photon-limited sensing, our method could facilitate intelligent perception in safety-critical environments where conventional cameras or LiDARs fail (e.g., nighttime robotics or rescue missions under adverse conditions). It also contributes toward reducing the data collection burden in new imaging modalities.

Potential Concerns: Like many general-purpose vision models, SPAD-based perception systems adapted from foundation models may inherit dataset biases from the pre-training dataset, and could be misused in surveillance or military systems. Although our method is technically agnostic to downstream applications, developers and practitioners should ensure that such systems are deployed responsibly, with careful consideration of fairness, accountability, and privacy.

We encourage the community to continue exploring reliable, interpretable, and ethically sound methods for adapting foundation models to new sensing modalities.

B Visualizations

To complement the quantitative results in the main paper, we provide additional visualizations that illustrate (1) how different adaptation methods behave under varying SPAD imaging conditions, and (2) representative failure cases that highlight the challenges of classification when photon-level noise severely degrades imaging quality.

B.1 Classification Across Imaging Conditions

Figure 8 presents several representative examples. For each class, we visualize three samples captured under different imaging conditions, with corresponding signal-to-background ratio (SBR) and average photon count per pixel (PPP). Below each image we report the predictions of two methods: CLIP-Adapter and our Noise Adapter. We observe that CLIP-Adapter often misclassifies samples under degraded imaging conditions, but recovers when SBR/PPP improves. In contrast, the Noise Adapter achieves stable predictions across various imaging conditions, demonstrating robustness to appearance shifts induced by imaging conditions. These visualizations provide direct evidence of how incorporating noise descriptors improves consistency across imaging conditions.

B.2 Failure Case Analysis

Figure 9 shows representative failure cases from the 16-shot setting using the proposed Noise Adapter. Most errors occur under poor imaging conditions (e.g. low SBR and low photon counts), which cause the SPAD imaging algorithm to produce depth maps with significant structural distortions. For example, the human sample in (h) is missing the head, apples and bananas show strong geometric deformation (a-b), the e-bike appears fragmented (c), and the boat region contains many voids (j). These artifacts hinder reliable semantic recognition and explain the observed misclassifications. In contrast, failure cases under relatively clean imaging (e.g., the building example) are rare, suggesting that classification errors are predominantly linked to degraded imaging conditions rather than limitations of the noise adapter itself.

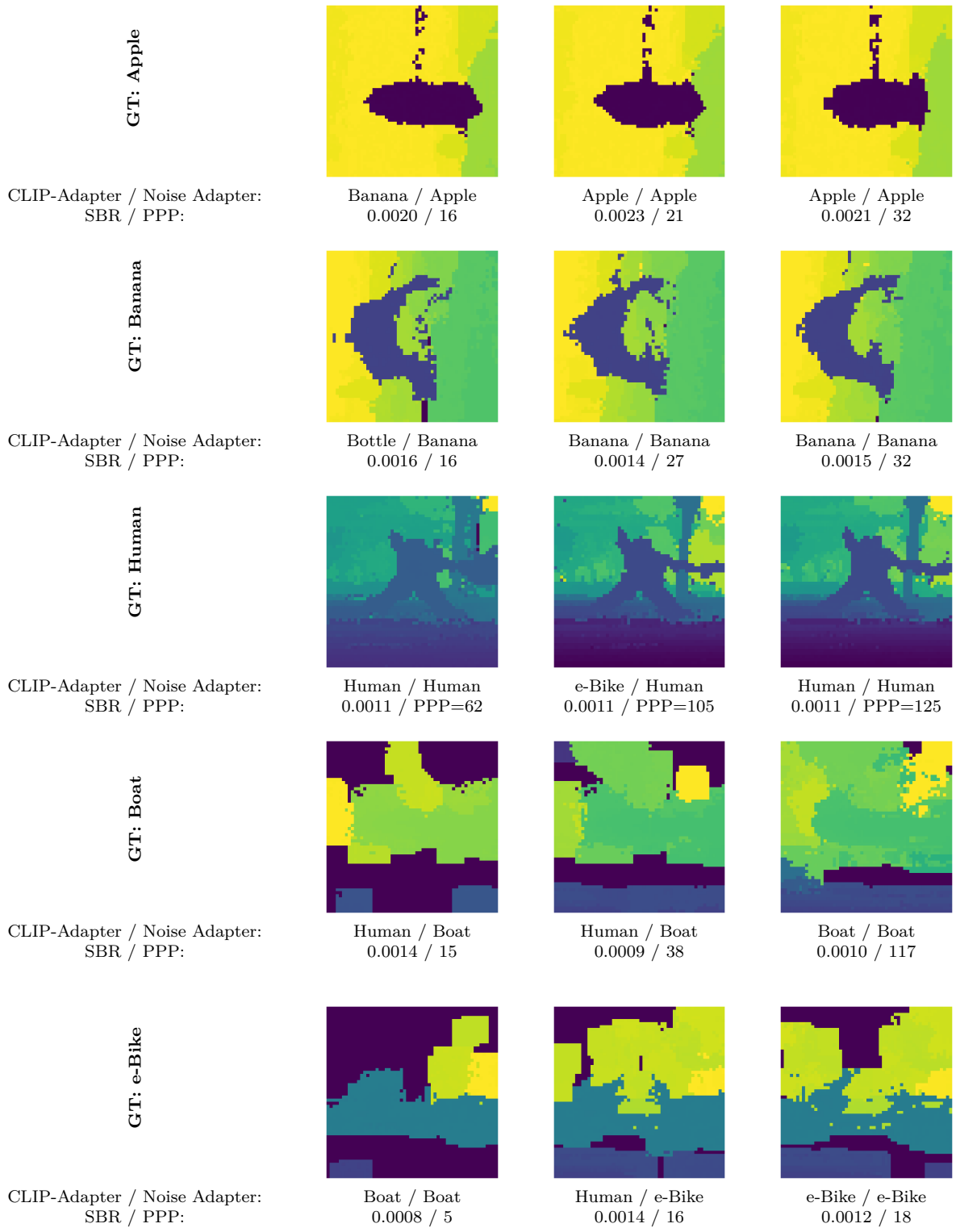
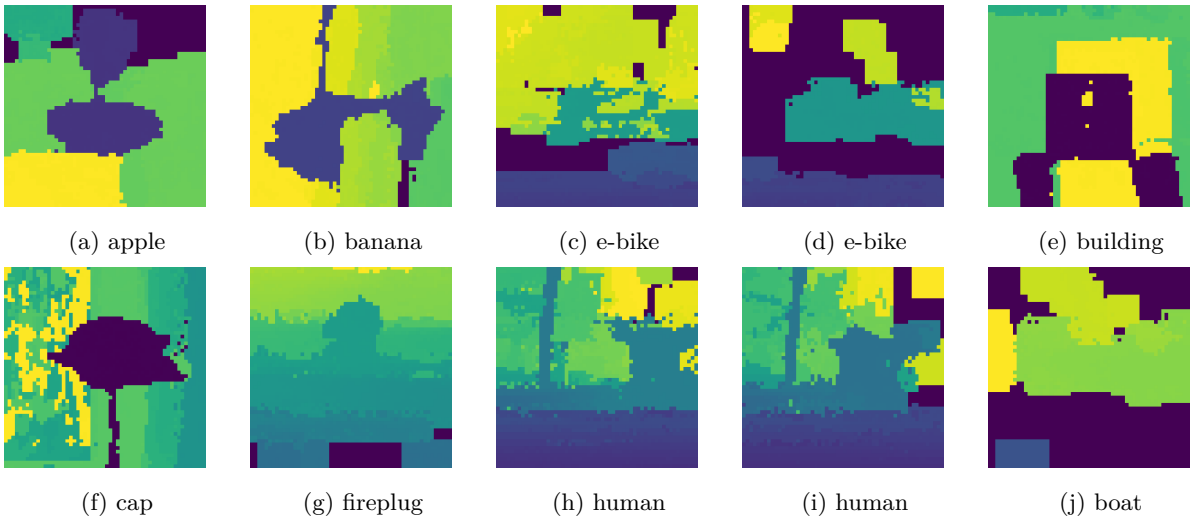


Figure 8: Qualitative comparisons. Each row corresponds to one ground-truth class (left label). For each class, three SPAD depth images captured under different imaging conditions are shown with predictions from CLIP-Adapter and the proposed Noise-Adapter, along with signal-to-background ratio (SBR) and photon count per pixel (PPP).



(a) apple (b) banana (c) e-bike (d) e-bike (e) building

(f) cap (g) fireplug (h) human (i) human (j) boat

Figure 9: Examples of failure cases from the real SPAD dataset. Our noise-aware adapter fails to correctly classify these samples despite training under the 16-shot setting.

C Influence of Pre-trained Backbone

To assess the generality of our method, we further evaluate its performance across a variety of visual backbones. In addition to CLIP ViT-B/16 and ResNet-50, we further experiment with recent self-supervised models, including DINOv2 ViT-S/14, DINOv2 ViT-B/14, DINOv3 ViT-S/16, and DINOv3 ViT-B/16 (distilled versions). As shown in Figure 10, our Noise-Adapter consistently outperforms all baselines across various numbers of labeled samples per class, regardless of the backbone architecture.

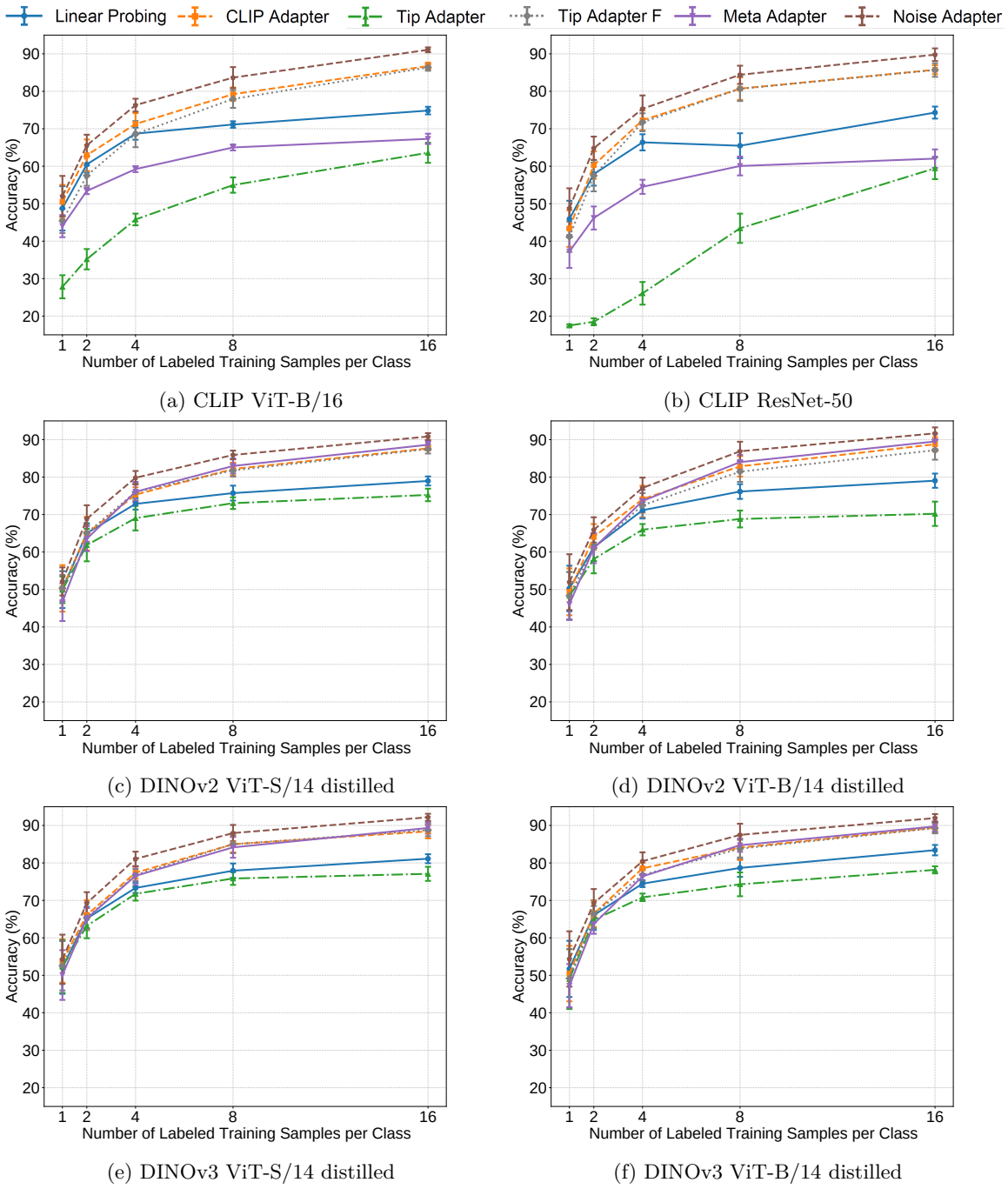
While the absolute accuracies vary slightly due to differences in encoder capacity, architecture and pre-training strategy. Noise-Adapter maintains a clear performance advantage across both low-shot and higher-shot regimes. Interestingly, several baselines (e.g., Tip-Adapter, Meta-Adapter, and Linear Probing) exhibit large sensitivity to the choice of backbone, whereas our method remains comparatively stable. This suggests that our approach is not tied to a particular encoder design and can robustly adapt to SPAD depth images across a range of vision backbones. These results demonstrate the flexibility and generalizability of Noise-Adapter, confirming that incorporating noise-aware representations remains effective even when the underlying visual feature extractor changes.

D Additional Analysis of GGFA

The numerical results of the ablation study in the synthetic SPAD dataset are shown in table 2, where we report the average accuracy and standard deviation over 5 runs.

Table 2: Ablation study of Gate-Guided Feature Augmentation (GGFA) on the SPAD synthetic dataset. Results show average accuracy with standard deviation (%) over 5 trials. The best results are shown in blue.

Few-shot Setup	1	2	4	8	16
w/o GGFA	43.67 ± 3.61	54.53 ± 2.62	67.13 ± 3.51	79.34 ± 2.62	86.52 ± 0.60
Rand	43.49 ± 3.44	53.75 ± 3.14	67.04 ± 4.13	79.34 ± 2.55	86.71 ± 0.49
Rand with Gate	44.47 ± 3.66	54.90 ± 2.99	67.29 ± 3.66	79.46 ± 2.59	86.78 ± 0.63
GGFA	44.41 ± 3.58	54.96 ± 3.07	67.50 ± 3.71	79.48 ± 2.54	86.68 ± 0.52



(a) CLIP ViT-B/16

(b) CLIP ResNet-50

(c) DINOv2 ViT-S/14 distilled

(d) DINOv2 ViT-B/14 distilled

(e) DINOv3 ViT-S/14 distilled

(f) DINOv3 ViT-B/14 distilled

Figure 10: Performance comparison of different methods using various pre-trained visual backbones: (a) CLIP ViT-B/16, (b) CLIP ResNet-50, (c) DINOv2 ViT-S/14 distilled, (d) DINOv2 ViT-B/14 distilled, (e) DINOv3 ViT-S/16 distilled, and (f) DINOv3 ViT-B/16 distilled. Results are reported on the real SPAD dataset across different few-shot settings, averaged over 5 random seeds.

D.1 Effect of Hyper-parameter α

We further analyze the sensitivity of GGFA to the perturbation strength α, which controls the magnitude of feature perturbations during augmentation. The choice of α is guided by the empirical observation that the mean standard deviation of feature dimensions in the training set is approximately 0.02. Based on this,

we vary α in the range $[0.015, 0.05]$ and evaluate the performance on the real SPAD dataset across 1-shot, 2-shot, 4-shot, 8-shot, and 16-shot settings.

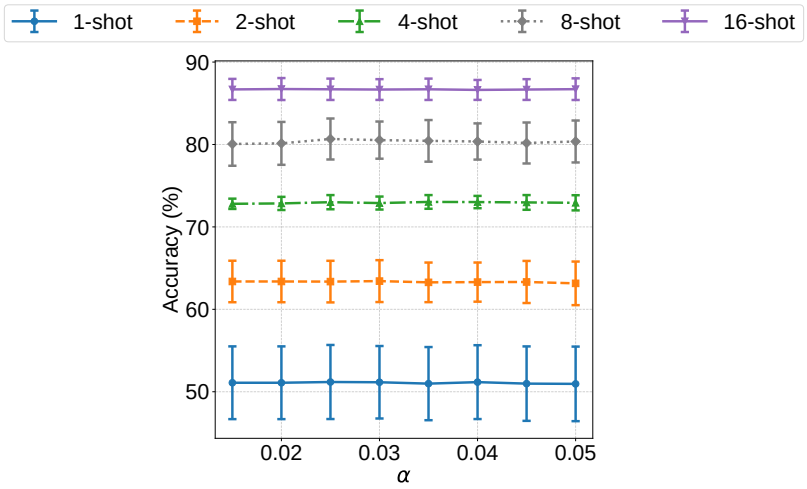


Figure 11: Sensitivity of GGFA to perturbation strength α on the real SPAD dataset. Accuracy remains stable across different α values, indicating robustness to the choice of this hyper-parameter.

Figure 11 shows that the classification accuracy remains stable across different values of α, indicating that GGFA is not sensitive to the precise choice of this hyperparameter. This robustness simplifies practical deployment, as α can be set to roughly match the average feature standard deviation without the need for extensive tuning.

D.2 Visualization Analysis of Gate-Guided Feature Augmentation

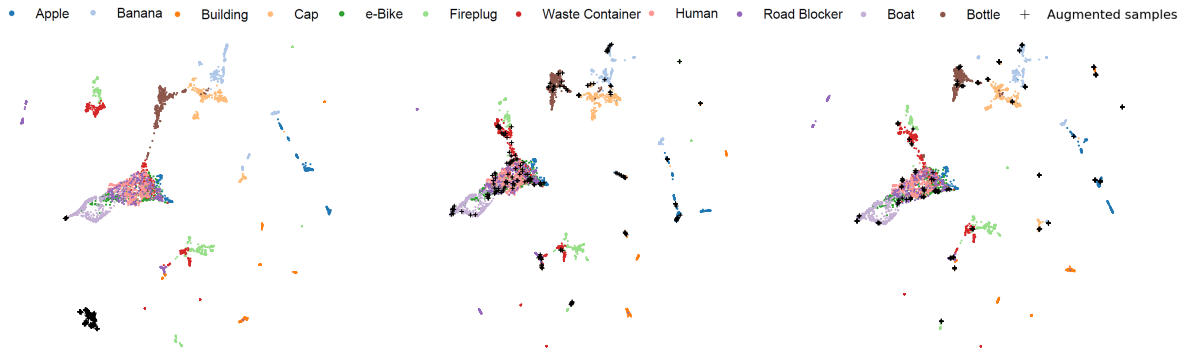
To better understand how different augmentation strategies affect feature distributions, we visualize the augmented features produced by various methods using PaCMAP on the real SPAD dataset. Specifically, we compare (a) adding uniform random noise (with $\alpha = 0.02$) directly to the modulated features, (b) adding the same random noise to the CLIP features and applying the learned noise gate to obtain augmented modulated features (Rand with Gate), and (c) our proposed GGFA method. For each method, we generate 200 augmented samples; black points indicate the augmented features in the visualizations shown in Figure 12.

As observed, directly adding random noise to the modulated features results in augmented samples that are distributed far from the original feature clusters. This likely explains why this approach fails to improve performance in Table 1, as the resulting augmented features do not resemble realistic noise-induced variations.

In contrast, both Rand with Gate and GGFA produce augmented features that are well aligned with the distribution of real features, with samples naturally scattered around the corresponding class clusters. This indicates that applying perturbations before the gating operation better preserves the underlying feature structure. Compared to Rand with Gate, GGFA tends to generate fewer augmented samples in ambiguous regions where features from multiple classes overlap, which likely reduces the introduction of unrealistic or label-ambiguous samples. This behavior may explain why GGFA achieves slightly better performance than Rand with Gate.

D.3 Comparison with Image-level Augmentation

In addition to the analyses on hyperparameter sensitivity and different injection strategies, we further investigate the relationship between Gate-Guided Feature Augmentation (GGFA) and image-level augmentation methods. Image-level augmentation, such as flipping or random erasing (Zhong et al., 2020), is widely adopted in vision tasks to increase data diversity. However, the variations it introduces are different from those induced by SPAD imaging conditions. In SPAD data, appearance shifts are strongly governed by



(a) Rand (b) Rand with Gate (c) GGFA

Figure 12: PaCMAP visualization of augmented features generated by different augmentation strategies on the real SPAD dataset. Black markers denote 200 augmented samples. (a) **Rand**: random noise added directly to the modulated features, (b) **Rand with Gate**: random noise added to the original CLIP features followed by gate modulation, (c) **GGFA**: noise strength is scaled by $(1 - g)$ and the resulting perturbed features are passed through the noise gate.

physical factors such as photon count and signal-to-background ratio (SBR), which affect the reliability and completeness of depth reconstruction. These effects cannot be easily replicated through simple image-level augmentation. For example, while one can degrade a high-quality SPAD capture into a low-SBR version by down-sampling photon counts or injecting additional background noise, the reverse process is not feasible: a low-SBR or low-photon-count depth image often contains erroneous or missing regions that prevent reconstructing a plausible high-quality counterpart. This key difference highlights the need for feature-level augmentation strategies, such as GGFA, that explicitly model the impact of imaging conditions on learned representations.

To this end, we directly compare GGFA with common image-level augmentation techniques, including random erasing and Gaussian noise injection. Feature visualization in Figure 13 demonstrates that GGFA produces augmented samples that are distributed more closely to real imaging variations, whereas image-level augmentations tend to generate features that deviate from the true distribution. This indicates that GGFA is able to mimic realistic feature shifts induced by changes in imaging conditions.

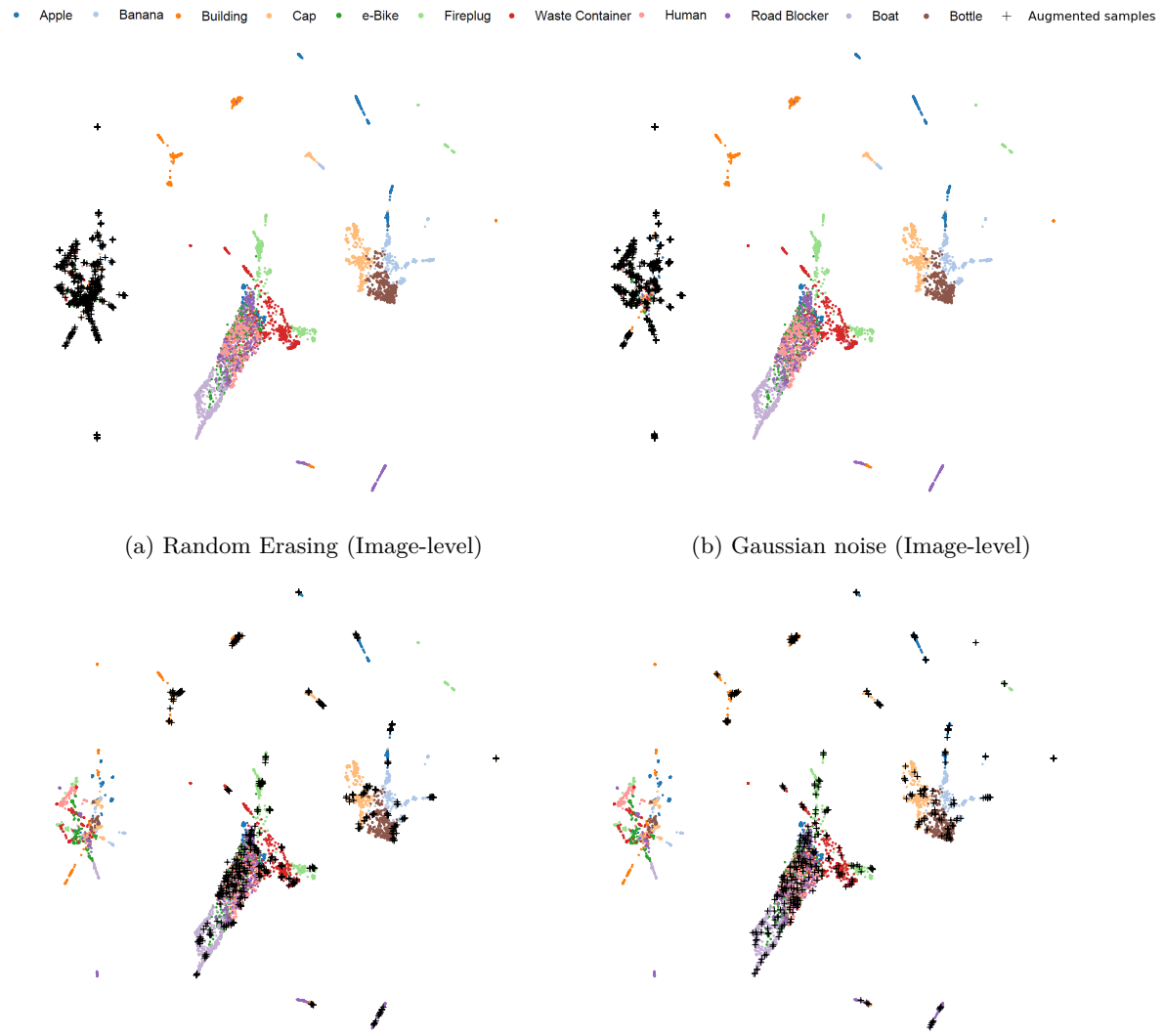
We further evaluate the effect of combining GGFA with image-level augmentation. Experimental results in Table 3 show that while each augmentation strategy brings moderate improvements individually, combining GGFA with image-level methods consistently achieves better performance, particularly in the low-shot regime. This suggests that GGFA and image-level augmentation are complementary: image-level transformations enrich diversity, while GGFA introduces variability that more directly reflects the impact of physical imaging conditions.

In summary, GGFA not only provides a way of imaging condition-induced feature variability in SPAD data, but also complements image-level augmentation strategies, together offering a richer and more realistic augmentation pipeline.

E SBR Estimation Accuracy

We further evaluated the accuracy of our SBR estimation on the synthetic dataset, where the ground-truth SBR values (0.06, 0.5, and 2.0) are known. The accuracy was measured using the relative estimation error, ε,

$$\varepsilon = \frac{\hat{\mathrm{SBR}} - \mathrm{SBR}_{\mathrm{gt}}}{\mathrm{SBR}_{\mathrm{gt}}}, \tag{7}$$



(a) Random Erasing (Image-level)

(b) Gaussian noise (Image-level)

(c) GGFA (Feature-level)

(d) Imaging variations (Ground truth)

Figure 13: PaCMAP visualization of augmented features generated by different augmentation strategies on the real SPAD dataset. Black markers denote 500 augmented samples. (a) Random Erasing, (b) Add Gaussian noise, (c) GGFA and (d) Imaging variations

Table 3: Performance comparison of Noise Adaptor versus baselines on Real dataset. Results show average classification accuracy with standard deviation (%) across 5 trials. The best results are shown in red and the second best results are shown in blue. single augmentation

Few-shot Setup	1	2	4	8	16
W/o Augmentation	50.39 ± 4.75	62.95 ± 2.31	73.09 ± 0.95	80.14 ± 1.68	86.61 ± 1.19
Flipping	51.52 ± 3.87	63.58 ± 2.71	73.61 ± 0.98	80.09 ± 1.74	87.08 ± 1.40
Gaussian noise	51.52 ± 3.92	63.11 ± 2.83	73.65 ± 0.96	79.61 ± 2.26	86.95 ± 1.23
Random erasing	51.35 ± 4.81	64.17 ± 3.35	75.00 ± 2.20	80.91 ± 1.30	87.74 ± 1.09
Flipping + Gaussian noise	51.48 ± 3.93	62.65 ± 2.34	73.42 ± 1.38	79.65 ± 1.75	86.39 ± 1.04
Flipping + Random erasing	52.77 ± 4.38	63.77 ± 2.96	74.14 ± 2.11	80.26 ± 2.06	87.73 ± 1.21
Random erasing + Gaussian noise	52.20 ± 4.05	62.98 ± 2.50	73.09 ± 2.27	80.29 ± 2.38	87.12 ± 1.04
Random erasing + GGFA	53.15 ± 3.88	64.58 ± 3.45	75.47 ± 2.74	80.74 ± 2.01	87.44 ± 1.59

where $\hat{\text{SBR}}$ is the estimated value and SBR_{gt} is the ground-truth SBR in the simulation. The mean relative error is 0.093, and the overall distribution of errors is shown in Figure 14. These results indicate that our estimation procedure provides a reliable approximation of SBR, sufficient for representing imaging conditions in the proposed framework.

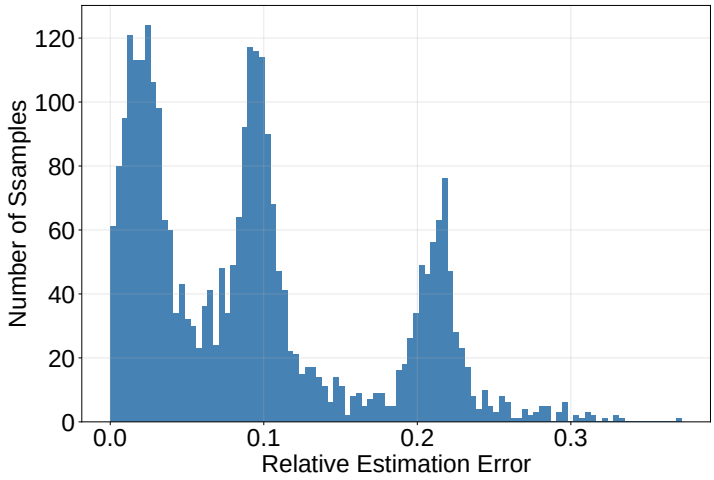


Figure 14: Histogram of relative estimation errors ε on simulated single-photon LiDAR data.

F Numerical Results

The numerical results of the fig. 3 in the main text are shown in table 4 and table 5, where we report the average accuracy and standard deviation over 5 runs.

Table 4: Performance comparison of Noise Adaptor versus baselines on Synthetic dataset. Results show average classification accuracy with standard deviation (%) across 5 trials. The best results are shown in red and the second best results are shown in blue.

Few-shot Setup	1	2	4	8	16
CLIP Zero-shot			11.18 ± 0.14		
Tip-Adapter	12.18 ± 0.36	14.09 ± 0.51	21.23 ± 3.69	33.71 ± 4.26	45.13 ± 2.94
Tip-Adapter-F	34.99 ± 3.54	44.89 ± 2.89	59.15 ± 4.13	73.33 ± 2.53	82.85 ± 0.94
CLIP-Adapter	40.66 ± 3.48	49.68 ± 2.62	63.80 ± 2.78	75.81 ± 2.25	84.08 ± 0.96
Meta-Adapter	37.03 ± 3.01	43.81 ± 2.89	50.95 ± 3.75	54.62 ± 2.01	59.75 ± 2.99
Linear Probing	43.45 ± 4.48	48.90 ± 3.98	56.62 ± 5.22	62.33 ± 3.39	67.53 ± 1.36
Noise-Aware Adapter (ours)	44.41 ± 3.58	54.96 ± 3.07	67.50 ± 3.71	79.48 ± 2.54	86.68 ± 0.52

Table 5: Performance comparison of Noise Adaptor versus baselines on Real dataset. Results show average classification accuracy with standard deviation (%) across 5 trials. The best results are shown in red and the second best results are shown in blue.

Few-shot Setup	1	2	4	8	16
CLIP Zero-shot			18.99 ± 0.85		
Tip-Adapter	26.48 ± 1.85	33.97 ± 1.38	43.76 ± 2.84	54.80 ± 2.58	63.12 ± 1.77
Tip-Adapter-F	46.91 ± 3.18	58.80 ± 2.87	69.00 ± 0.36	75.74 ± 1.52	81.64 ± 1.04
CLIP-Adapter	51.33 ± 3.28	60.20 ± 2.23	70.45 ± 1.30	77.03 ± 2.45	82.21 ± 0.45
Meta-Adapter	46.65 ± 2.91	54.32 ± 2.66	59.14 ± 1.36	62.86 ± 1.27	64.94 ± 0.40
Linear Probing	47.76 ± 4.27	57.44 ± 3.36	66.23 ± 1.26	67.65 ± 1.54	70.35 ± 1.00
Noise-Aware Adapter (ours)	51.24 ± 4.48	63.18 ± 2.70	73.15 ± 0.52	80.45 ± 2.42	86.68 ± 1.28