A HISTOGRAM IS WORTH A THOUSAND PIXELS: SINGLE-PHOTON METRIC ANCHORING FOR MONOCULAR DEPTH ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-pixel measurements acquired with a single-photon avalanche diode (SPAD) provide accurate metric range information but lack explicit spatial correspondence with image pixels, making their fusion with dense monocular depth predictions inherently ambiguous. We introduce **Hist2Depth**, a global-to-local framework that uses monocular depth itself to associate spatially unresolved SPAD measurements with pixel-wise depth hypotheses. Global D2 calibration first aligns the monocular prediction with the measured depth distribution through an order-preserving metric calibration. The multi-scale gated recurrent unit (MS-GRU) then performs depth-conditioned recurrent refinement, dynamically querying SPAD evidence according to the evolving pixel-wise depth estimates and integrating it with spatial image context. On NYU-Depth-v2, Hist2Depth with AdaBins achieves 0.0649 AbsRel and 0.2750 m RMSE, and consistently improves RGB-only predictions across six monocular backbones. When transferred directly from synthetic training to captured RGB–SPAD measurements without fine-tuning, Hist2Depth also retains strong metric accuracy. Under photon thinning, performance remains nearly unchanged at a 60 s equivalent acquisition, using only 7.4% of the full photon count. These results demonstrate that global SPAD measurements can effectively guide dense metric depth reconstruction through depth-conditioned global-to-local fusion.

1 INTRODUCTION

Monocular depth estimation (MDE) provides dense spatial geometry from a single RGB image, but even modern metric depth estimators can retain global and depth-dependent errors in their predictions (Bhat et al., 2021; Bochkovskiy et al., 2025). Photon time-of-flight measurements provide direct metric range information by measuring the round-trip travel time of light. However, acquiring spatially resolved range measurements typically requires sequential scanning or time-resolved detector arrays, increasing acquisition or hardware complexity (Nishimura et al., 2020). A more compact alternative is single-pixel SPAD sensing, which integrates photon returns over the entire field of view into a global photon arrival-time histogram (Nishimura et al., 2020). Combining this global measurement with a monocular prediction brings together measured range information and image-derived spatial structure, allowing the transient to guide the correction of a dense depth map.

Existing RGB–SPAD methods incorporate global range measurements into monocular depth estimation. SSD aligns the predicted depth distribution with the measured SPAD transient through histogram matching (Nishimura et al., 2020), while PhoMoNet learns to fuse global SPAD information with spatial image features (Lee et al., 2023). Although these methods improve depth reconstruction, the modality mismatch underlying RGB–SPAD fusion warrants further investigation. Specifically, the SPAD transient provides global metric range evidence, whereas MDE provides dense spatial geometry but can retain metric errors. This mismatch gives rise to two coupled challenges. First, direct dense pixel-wise correction from a single global transient is fundamentally underconstrained, as the measurement provides only global distributional information without spatially resolved constraints. Second, the spatially unresolved transient provides no explicit correspondence between depth responses and image pixels, which limits spatially varying correction.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We propose Hist2Depth, a two-stage global-to-local framework that progressively transfers SPAD range information to dense monocular depth prediction through global metric calibration and local depth refinement. To address the underconstrained nature of direct dense correction, Global D2 restricts the correction to a bounded, order-preserving global mapping estimated from the discrepancy between predicted and measured depth distributions. To resolve the missing spatial correspondence, MS-GRU uses each pixel's evolving depth estimate to query depth-relevant SPAD evidence and combines it with image context for spatially varying refinement.

We evaluate Hist2Depth with six monocular backbones on NYU-Depth-v2 (Silberman et al., 2012), improving upon every RGB-only baseline. Our best configuration, using DA3Metric-L, achieves an AbsRel of 0.0385 and an RMSE of 0.1798 m. Under the same AdaBins backbone, our method outperforms histogram matching and all six evaluated fusion alternatives, reducing AbsRel by 18.5% relative to PhoMoNet. On captured RGB–SPAD data, Hist2Depth maintains strong metric accuracy even after photon thinning to approximately 7.4% of the original photon count. Four ablation studies further validate the fusion design, component contributions, calibration order, and recurrent iterations.

Our contributions are:

- We propose Hist2Depth, a global-to-local framework for single-pixel SPAD-guided monocular depth estimation. It addresses the RGB–SPAD modality mismatch by progressively incorporating global range measurements into dense monocular predictions through metric calibration and local refinement.
- We develop a two-stage correction scheme consisting of Global D2 and MS-GRU. Global D2 uses the discrepancy between SPAD and MDE depth distributions to estimate a quadratic, order-preserving global calibration. MS-GRU uses evolving pixel-wise depth estimates to query depth-conditioned SPAD features and progressively refines local depth through multi-scale recurrent updates.
- We demonstrate the effectiveness and robustness of Hist2Depth across different MDE models and fusion strategies, including under substantially reduced photon counts.

2 Related Work

Monocular depth estimation. Earlier methods improve monocular depth estimation through task-specific designs, including ordinal regression in DORN, local planar guidance in BTS, and adaptive depth bins in AdaBins (Fu et al., 2018; Lee et al., 2019; Bhat et al., 2021). More recent work has shifted toward depth foundation models that leverage large-scale and diverse training data and strong pretrained visual representations to improve cross-scene generalization and recover detailed geometric structures (Ranftl et al., 2022; Yang et al., 2024a;b; Ranftl et al., 2021; Ke et al., 2024; Xu et al., 2025; He et al., 2025; Lin et al., 2026).

Metric supervision and dedicated prediction architectures further enable the recovery of absolute scene depth (Bhat et al., 2023; Bochkovskiy et al., 2025). Relative-depth pretraining followed by metric-depth fine-tuning transfers learned geometric representations to the prediction of distances in metric units. Although these methods produce detailed metric depth maps, errors in absolute distance can remain.

We therefore use SPAD measurements as a global metric range reference to correct monocular depth predictions. Hist2Depth exploits this metric reference while retaining the dense spatial structure recovered by MDE.

Spatial depth priors. Depth completion and depth super-resolution combine RGB images with spatially indexed depth measurements. Marigold-DC uses sparse depth observations to guide a pretrained diffusion model at inference (Viola et al., 2025). Prompt Depth Anything injects low-resolution LiDAR depth into a depth foundation model through multi-scale feature fusion (Lin et al., 2025). Prior Depth Anything uses predicted geometry to pre-fill incomplete metric priors, followed by conditioned refinement (Wang et al., 2026).

These methods share our motivation of correcting monocular depth with measured range. Sparse depth measurements provide limited constraints relative to the pixel-wise degrees of freedom of

dense correction, making the problem underdetermined (Rim et al., 2026). However, their measurements still retain spatial correspondence with the image. A single-pixel SPAD transient further removes this correspondence by collapsing spatial information into a global depth distribution. We address these two challenges with Global D2 and MS-GRU: Global D2 constrains dense correction through low-dimensional global calibration, while MS-GRU restores pixel-wise measurement guidance through depth-conditioned recurrent refinement.

Single-pixel SPAD guidance. Nishimura et al. introduce SSD to refine monocular depth through histogram matching (Nishimura et al., 2020). Their method first transforms the reflectance-weighted monocular depth histogram to match the distribution of the processed SPAD transient. It then maps the matched depth values back to image pixels using the pixel-to-bin correspondences established from the original monocular depth map. However, SSD does not account for local spatial context during this step. This limits the correction of spatially varying errors, since pixels with similar predicted depths may require different adjustments. PhoMoNet instead uses single-pixel guidance modules to fuse shared SPAD and image features through cosine similarity, allowing the guidance to vary across spatial locations (Lee et al., 2023). However, the correspondence between pixel-wise depth hypotheses and depth-axis SPAD evidence remains implicit in this fusion.

Hist2Depth instead uses the evolving pixel-wise depth estimate to explicitly query SPAD evidence at the corresponding depth locations. MS-GRU combines the queried evidence with image context and progressively refines local depth through recurrent updates. This depth-conditioned lookup provides more targeted SPAD guidance for local correction by restricting each query to evidence around the current depth hypothesis.

3 METHOD

3.1 PROBLEM SETUP

Given an RGB image $I \in \mathbb{R}^{3\times H\times W}$ and a global SPAD transient $h \in \mathbb{R}^{1024}$, our goal is to refine the MDE prediction $D_{\text{base}} \in \mathbb{R}^{H\times W}$ into a more accurate metric depth map. The SPAD transient provides metric range information over $[D_{\min}, D_{\max}]$, but only as a global depth distribution without pixel-wise correspondence.

3.2 DISTRIBUTION-CONSTRAINED GLOBAL CALIBRATION

Depth distribution construction. The global SPAD transient provides metric information only at the distribution level, leaving direct dense correction underconstrained. Global D2 therefore constrains the correction to a global parametric mapping inferred from the discrepancy between MDE- and SPAD-derived depth distributions. To make the two distributions comparable, we apply preprocessing, including compensation for measurement-related effects and rebinning onto a common depth grid. SPAD photon counts depend on surface reflectance and inverse-square range attenuation (Nishimura et al., 2020). We therefore weight pixel contributions by RGB grayscale and compensate range attenuation in the measured transient. For the monocular prediction, hard depth binning is not differentiable with respect to depth, so we use Gaussian soft assignment to construct the predicted distribution.

Specifically, we divide $[D_{\min}, D_{\max}]$ into $N_h = 64$ equal-width bins, with width $\Delta_h = (D_{\max} - D_{\min})/N_h$ and centers $c_j = D_{\min} + (j - \frac{1}{2})\Delta_h$. For pixel depth d_p and grayscale weight ρ_p, the soft assignment $a_{p,j}$ and normalized MDE-derived depth distribution $q_{b,j}$ are

$$a_{p,j} = \frac{\exp\left[-\frac{(d_p - c_j)^2}{2\Delta_h^2}\right]}{\sum_{l=1}^{N_h} \exp\left[-\frac{(d_p - c_l)^2}{2\Delta_h^2}\right]}, \qquad q_{b,j} = \frac{\sum_p \rho_p a_{p,j}}{\max\left(\sum_p \rho_p, 10^{-6}\right)}. \tag{1}$$

For the SPAD measurement, we first subtract background counts to reduce ambient-light and dark-count contributions. Let z_t denote the metric range corresponding to transient bin t and $\bar{h}_t$ the background-subtracted count. We compensate inverse-square attenuation using $\tilde{h}_t = z_t^2 \bar{h}_t$, then rebin and normalize the transient on the same depth grid to obtain the SPAD-derived depth distribution $q_{t,j}$. We further compute the cumulative distribution difference $\Delta Q_j = \sum_{m\le j}(q_{t,m} - q_{b,m})$, which retains the ordered accumulation of mismatch along depth.

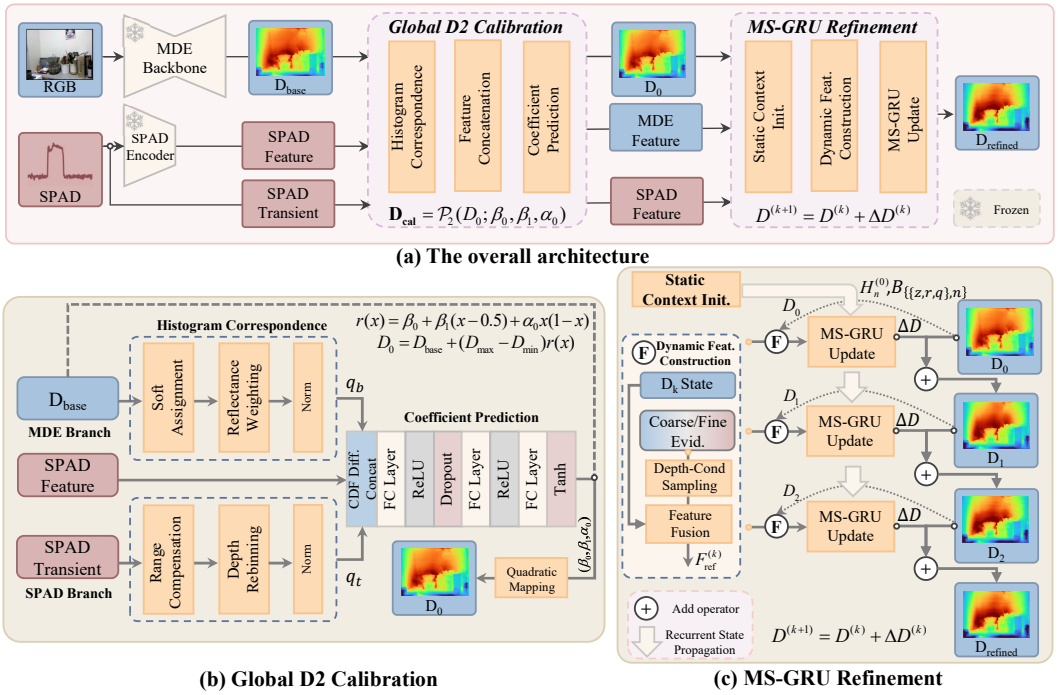

Figure 1: **Hist2Depth.** Given an RGB image, a global SPAD transient, and the metric depth prediction from a frozen MDE model, Global D2 first performs coarse metric calibration by aligning the predicted and measured depth distributions. MS-GRU then combines depth-conditioned SPAD evidence with image context for fine spatial refinement, producing the final metric depth map through iterative updates.

Bounded quadratic calibration. Unlike histogram matching, where distribution alignment directly determines corrected depths, Global D2 uses the distribution discrepancy only to estimate compact global calibration parameters. This low-dimensional design follows compact parametric calibration under limited range constraints (Rim et al., 2026). A quadratic mapping provides additional nonlinear flexibility beyond scale and shift while retaining only three parameters.

A three-layer MLP predicts the coefficients β_0, β_1, and α_0 from q_b, q_t, ΔQ, and a projected shared SPAD feature. A scaled $\tanh$ constrains them to $|\beta_0| \leq 0.10$, $|\beta_1| \leq 0.10$, and $|\alpha_0| \leq 0.05$. These bounds were chosen empirically in preliminary experiments and kept fixed across all datasets and backbones. Writing $L = D_{\max} - D_{\min}$ and $x = \mathrm{clip}((D_{\text{base}} - D_{\min})/L, 0, 1)$, the calibrated depth is

$$D_0 = D_{\text{base}} + L\left[\beta_0 + \beta_1(x - 0.5) + \alpha_0 x(1 - x)\right]. \tag{2}$$

For $x \in [0, 1]$, the coefficient bounds give $\partial D_0/\partial D_{\text{base}} = 1 + \beta_1 + \alpha_0(1 - 2x) \geq 0.85$, so the mapping remains strictly increasing and preserves the initial depth ordering. Depth values are clipped to $[D_{\min}, D_{\max}]$ when constructing subsequent distribution and SPAD lookup evidence.

In summary, this design offers two benefits: 1) **Low-DoF Metric Calibration**: three parameters correct global offset, scale, and nonlinear depth-dependent errors; and 2) **Geometry Preservation**: the monotonic mapping adjusts metric depth while retaining the relative ordering recovered by MDE. Its globally shared correction cannot model spatially varying residuals, motivating the subsequent MS-GRU refinement.

3.3 Depth-Conditioned Recurrent Refinement

Static context initialization. Before iteration, MS-GRU encodes selected MDE decoder features, RGB, normalized D_0, grayscale reflectance, the shared SPAD feature, and the CDF mismatch recomputed from D_0. This static context initializes hidden states at $H/2 \times W/2$, $H/4 \times W/4$, and

$H/8 \times W/8$, together with fixed ConvGRU gate biases. During refinement, only the hidden states evolve; the static context and gate biases remain fixed.

Depth-conditioned evidence lookup. The global SPAD transient provides no explicit spatial correspondence with image pixels. We therefore use each pixel's current depth estimate to associate it with relevant evidence along the SPAD depth axis. At each iteration, we rebuild the depth-axis evidence from the current prediction $D^{(k)}$. A coarse 64-bin axis captures distribution mismatch, while a fine 1024-bin axis retains detailed metric-depth structure. Each axis combines the SPAD target distribution, the current predicted distribution, and their bin-wise and cumulative differences. The resulting sequences are encoded by separate one-dimensional encoders.

At the fine recurrent resolution, each $D^{(k)}_{\text{low}}(u, v)$ is mapped to both encoded depth axes. We sample around the mapped locations with radii $r_c = 2$ and $r_f = 4$, obtaining 40- and 72-D features that form a 112-D lookup. The lookup is combined with the normalized current depth and its displacement from D_0 to form $F^{(k)}_{\text{ref}}$.

Spatial update and feedback. MS-GRU updates the recurrent states from coarse to fine. Each coarser state conditions the next finer scale, and $F^{(k)}_{\text{ref}}$ is injected at the fine scale for local correction. The scale-specific ConvGRUs share parameters across iterations, while the static gate biases remain fixed. The updated fine state $H^{(k+1)}_f$ predicts the low-resolution depth update and convex upsampling mask:

$$\begin{aligned} D^{(k+1)}_{\text{low}} &= D^{(k)}_{\text{low}} + 0.2 \tanh\Big(\phi_d(H^{(k+1)}_f)\Big), \\ D^{(k+1)} &= D_0 + \mathcal{U}_{\text{cvx}}\Big(D^{(k+1)}_{\text{low}} - D^{(0)}_{\text{low}}; M^{(k)}\Big). \end{aligned} \tag{3}$$

Here ϕ_d is the depth-update head, and $D^{(0)}_{\text{low}}$ is D_0 at the fine recurrent resolution. The mask head predicts 36 channels, which are Softmax-normalized to form local 3×3 convex weights for 2×2 high-resolution subpixels. Each recurrent depth update is bounded to ± 0.2 m. Only the accumulated residual is upsampled and added to D_0, preserving the calibrated full-resolution initialization. The updated prediction then rebuilds the depth-axis evidence and determines the query locations for the next iteration. Detailed GRU and upsampling configurations are provided in Appendix A.

In summary, MS-GRU offers two main advantages: 1) **Explicit Pixel–Measurement Association**: depth-conditioned lookup uses each pixel's current depth hypothesis to establish an explicit link to the corresponding region of the global SPAD depth axis; and 2) **Spatially Adaptive Refinement**: multi-scale recurrent context allows pixels with similar depth hypotheses to receive different corrections.

3.4 Training Objective

Masked SILog losses supervise D_0 and all three recurrent predictions, with recurrent-loss weighting $\gamma = 0.8$. To explicitly preserve consistency with the global SPAD measurement, Wasserstein-1 losses align the SPAD-derived depth distribution with the depth distributions constructed from D_0 and the final refined prediction $D^{(K)}$. Additional regularizers constrain the magnitude and spatial smoothness of the recurrent residual as well as the Global D2 coefficients. Detailed loss definitions and weights are provided in Appendix A.6.

4 Experiments

4.1 Experimental Setup

Datasets and measurement protocol. We evaluate on NYU-Depth-v2 using the standard 654-image test split and a valid depth range of 0.001–10 m. Synthetic SPAD transients follow the standard RGB–SPAD forward model (Sun et al., 2020), with 10^6 signal photons and SBR sampled from $[5, 100]$ during training. Unless otherwise specified, quantitative evaluation uses SBR$= 100$. We also evaluate zero-shot transfer on the captured RGB–SPAD dataset released with SSD (Nishimura et al., 2020), including photon-count reduction by nested binomial thinning. Additional simulation and preprocessing details are provided in Appendix A.

Table 1: NYU-Depth-v2 accuracy and inference cost. Bold marks the best depth metrics in this comparison.

Method	Input	AbsRel↓	RMSE↓	δ_1 ↑	Params. (M)	Time (ms)↓
DORN	RGB	0.1095	0.4050	0.8720	110.28	13.66
BTS	RGB	0.1096	0.3918	0.8855	47.00	22.25
AdaBins	RGB	0.1037	0.3653	0.9016	78.26	19.99
PixelFormer	RGB	0.0940	0.3376	0.9188	270.90	21.59
SSD (DORN)	RGB+SPAD	0.0889	0.3693	0.9084	110.28	78.63
SSD (BTS)	RGB+SPAD	0.0864	0.3553	0.9136	47.00	92.03
SSD (AdaBins)	RGB+SPAD	0.0806	0.3341	0.9235	78.26	88.09
SSD (PixelFormer)	RGB+SPAD	0.0711	0.2953	0.9395	270.90	91.38
PhoMoNet (AdaBins)	RGB+SPAD	0.0796	0.3013	0.9420	79.49	34.39
Hist2Depth (AdaBins)	RGB+SPAD	**0.0649**	**0.2750**	**0.9538**	86.66	41.50

Implementation and evaluation. Unless otherwise specified, the MDE backbones are fixed and only the proposed correction modules are trained for eight epochs using AdamW with a learning rate of 5×10^{-5} and 416×544 crops. The shared SPAD encoder follows the PhoMoNet design (Lee et al., 2023) and is initialized from its pretrained weights, remaining fixed during training. We report AbsRel, RMSE, and δ_1 in metric depth. For all controlled comparisons and ablation studies, we use AdaBins as the common MDE backbone to isolate the effects of SPAD guidance, fusion design, and model components. For the cross-backbone study, Hist2Depth is trained separately with each evaluated MDE backbone. Additional optimization settings and metric definitions are provided in Appendix A.

4.2 Accuracy and Computational Cost

Baselines. Following prior single-pixel SPAD-guided depth studies, we include DORN (Fu et al., 2018), BTS (Lee et al., 2019), AdaBins (Bhat et al., 2021), and PixelFormer (Agarwal & Arora, 2023) as RGB-only MDE baselines, together with their SSD-guided variants (Nishimura et al., 2020). PhoMoNet (Lee et al., 2023), using AdaBins as its MDE backbone, is included as the representative learned single-pixel SPAD fusion method.

Quantitative comparison. Table 1 compares Hist2Depth with existing single-pixel SPAD-guided depth methods on NYU-Depth-v2. To isolate the effect of measurement guidance from differences in monocular prediction, we focus first on methods using the common AdaBins backbone. Hist2Depth improves over PhoMoNet in AbsRel (0.0796→0.0649, -18.5%), RMSE (0.3013→0.2750 m, -8.7%). Compared with SSD under the same backbone, the improvements are larger in AbsRel (0.0806→0.0649, -19.5%) and RMSE (0.3341→0.2750 m, -17.7%). Overall, Hist2Depth achieves the best performance across all three metrics among the compared methods. This improvement comes with moderate additional cost: Hist2Depth uses 86.66 M parameters and 41.50 ms per image, compared with 79.49 M and 34.39 ms for the AdaBins-based PhoMoNet setting.

The results also indicate that the final reconstruction accuracy depends on the underlying monocular predictor, whose baseline performance varies across architectures. We therefore next evaluate Hist2Depth with multiple MDE backbones to examine whether the proposed SPAD guidance remains effective for different initial monocular depth predictions.

Qualitative comparison. Figure 2 shows representative predictions and error maps on NYU-Depth-v2. SSD's global distribution-based correction does not guarantee pixel-wise improvement. In some scenes, discrepancies between the initial monocular prediction and the SPAD measurement can even amplify local errors. PhoMoNet partially alleviates this spatial inconsistency through learned fusion. Because PhoMoNet lacks explicit measurement-to-pixel correspondence, its limitations are more clearly reflected in the corresponding error maps than in the visually similar depth reconstructions. In contrast, Hist2Depth produces more spatially consistent corrections and reduces errors over broader image regions.

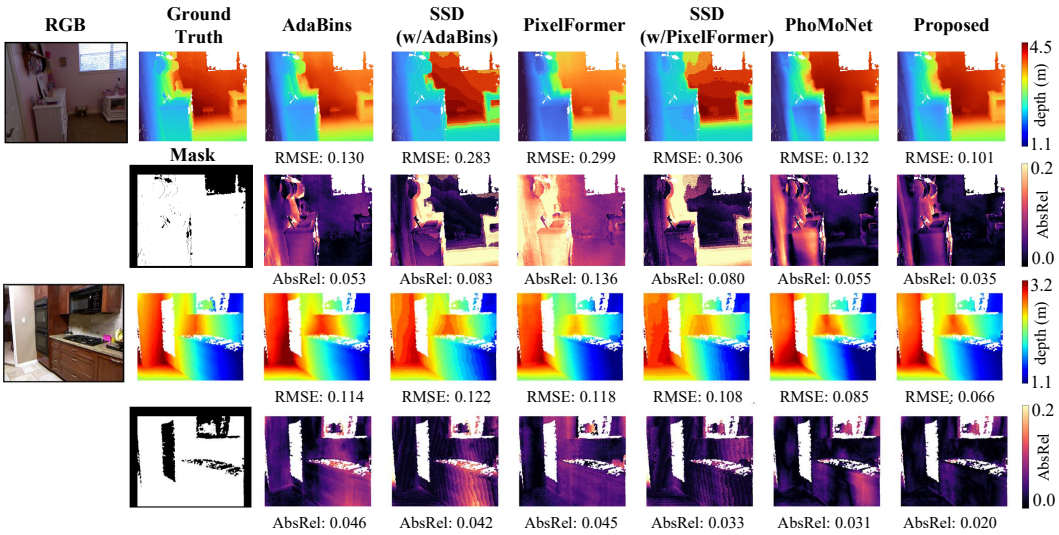

Figure 2: Qualitative comparison on two NYU examples. Predicted metric depth and pixel-wise AbsRel are shown with common display ranges within each example. The proposed method retains the principal scene structure while reducing the displayed metric errors.

Table 2: Cross-backbone evaluation on NYU-Depth-v2.

Metric	Variant	BTS arXiv'19	AdaBins CVPR'21	PixelFormer WACV'23	ZoeDepth-N arXiv'23	Depth Pro ICLR'25	DA3Metric-L ICLR'26
AbsRel↓	Base	0.1096	0.1037	0.0940	0.0755	0.0930	0.0722
	SSD	**0.0864**	0.0806	0.0711	0.0623	0.0639	0.0661
	Ours	0.0934	**0.0649**	**0.0663**	**0.0621**	**0.0408**	**0.0385**
RMSE↓	Base	0.3918	0.3653	0.3376	0.2715	0.3337	0.2845
	SSD	0.3553	0.3341	0.2953	0.2591	0.2778	0.2847
	Ours	**0.3532**	**0.2750**	**0.2572**	**0.2278**	**0.1918**	**0.1798**
δ_1 ↑	Base	0.8855	0.9016	0.9188	0.9545	0.9347	0.9628
	SSD	**0.9136**	0.9235	0.9395	0.9527	0.9469	0.9445
	Ours	0.9073	**0.9538**	**0.9557**	**0.9657**	**0.9821**	**0.9842**

4.3 Across Monocular Backbones

Compared with in-network feature fusion, Hist2Depth retains a common SPAD-guided correction architecture across MDE backbones by adapting the decoder features. Table 2 evaluates six backbones: BTS (Lee et al., 2019), AdaBins (Bhat et al., 2021), PixelFormer (Agarwal & Arora, 2023), ZoeDepth-N (Bhat et al., 2023), Depth Pro (Bochkovskiy et al., 2025), and DA3Metric-L (Lin et al., 2026). For each backbone, the MDE is fixed and the same Global D2 and MS-GRU architecture is trained separately under a common protocol. For spatial context, we construct a 256-channel representation from intermediate-to-late decoder features; PixelFormer combines two spatially aligned decoder scales. The exact extraction points are provided in Appendix A.

Hist2Depth consistently improves the RGB-only prediction across all six backbones in AbsRel, RMSE, and δ_1. The gains remain substantial for recent metric-depth models; with DA3Metric-L, Hist2Depth improves AbsRel (0.0722→0.0385, -46.7%) and RMSE (0.2845→0.1798 m, -36.8%). By contrast, SSD shows less consistent gains across MDE models; on DA3Metric-L, for example, it leaves the RMSE essentially unchanged (0.2845→0.2847 m), whereas Hist2Depth reduces it to 0.1798 m. Across all 18 metric–backbone combinations, Hist2Depth outperforms SSD in 16, with the two exceptions occurring on BTS in AbsRel and δ_1. The weaker performance on BTS may be related to its spatial context being extracted from a relatively earlier decoder stage than those used for several other backbones. Overall, these results show that the proposed learned guidance adapts well to heterogeneous MDE representations and provides more consistent improvements than histogram matching.

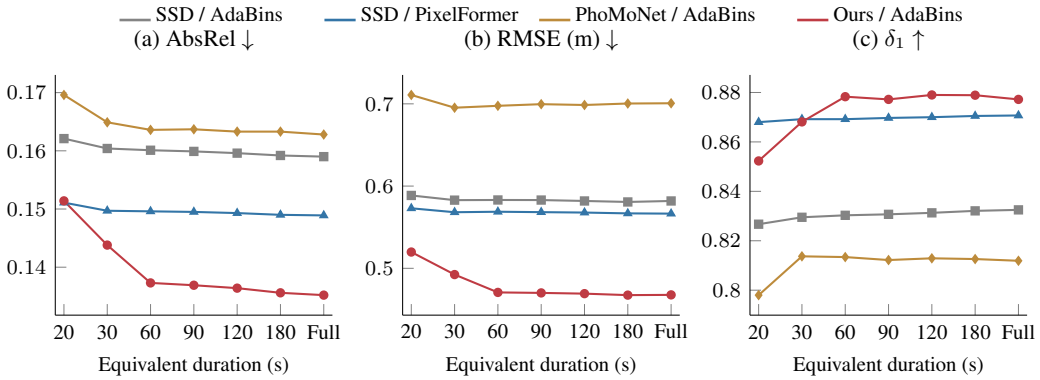

Figure 3: Sensitivity to photon counts on captured RGB–SPAD data. Shorter equivalent durations are generated by nested binomial thinning of the full measured histograms (average duration 816 s).

Table 3: Ablation studies with AdaBins on NYU-Depth-v2.

Strategy	AbsRel↓	RMSE↓	δ_1 ↑
RGB only	0.1037	0.3653	0.9016
Concat-conv	0.0852	0.3156	0.9352
Cross-attention	0.0825	0.3032	0.9362
Modulation	0.1046	0.3465	0.9096
CGAFusion	0.0842	0.3105	0.9348
MFM	0.0776	0.3039	0.9465
PhoMoNet-SPGM	0.0796	0.3013	0.9420
Hist2Depth	**0.0649**	**0.2750**	**0.9538**

(a) Fusion strategy. Modulation (Zhao et al., 2023); CGAFusion (Chen et al., 2024); MFM (Zamir et al., 2020); PhoMoNet-SPGM (Lee et al., 2023).

Order	AbsRel↓	RMSE↓	δ_1 ↑
Base	0.1037	0.3653	0.9016
0	0.0706	0.2840	0.9526
1	0.0705	0.2834	0.9529
2	**0.0702**	**0.2830**	**0.9530**

(c) Effect of polynomial order on Global calibration. The quadratic mapping performs best.

	AbsRel↓	RMSE↓	δ_1 ↑
AdaBins	0.1037	0.3653	0.9016
D2 only	0.0702	0.2830	0.9530
MS-GRU only	0.0663	0.2757	0.9535
Full	**0.0649**	**0.2750**	**0.9538**

(b) Effect of using Global D2 and MS-GRU individually and jointly. Both components improve the baseline, and their combination gives the best overall accuracy.

K	AbsRel↓	RMSE↓	δ_1 ↑	Time (ms)↓
0	0.0702	0.2830	0.9530	24.35
1	0.0679	0.2787	0.9533	32.45
2	0.0673	0.2774	0.9534	37.49
3	**0.0649**	**0.2750**	**0.9538**	42.93

(d) Effect of inference iterations on MS-GRU accuracy and runtime. Runtime is benchmarked independently, and small run-to-run variations may arise from normal GPU execution variability.

4.4 Captured RGB–SPAD Measurements

Zero-shot transfer to measured data. We evaluate both zero-shot transfer and photon-count sensitivity on captured RGB–SPAD measurements. Hist2Depth achieves the best performance under the AdaBins setting, improving over SSD in AbsRel (0.1590→0.1352, -15.0%) and RMSE (0.5819→0.4677 m, -19.6%), while also outperforming PhoMoNet. To assess robustness across photon budgets, we progressively reduce the photon count from the full measurement (average duration: 816 s), as shown in Figure 3. At 60 s, only 7.4% of the photons are retained, yet performance remains nearly unchanged (AbsRel: 0.1352→0.1373; RMSE: 0.4677→0.4707 m). Degradation becomes noticeable at 30 and 20 s, but Hist2Depth still achieves the lowest RMSE at 20 s (0.5198 m). Complete quantitative results are provided in Tables 8 and 9 in the Appendix.

4.5 Ablation Studies

Fusion strategy. Table 3(a) compares alternative RGB–SPAD fusion strategies with AdaBins. More complex fusion does not consistently improve accuracy: modulation underperforms concatenation in AbsRel (0.1046 vs. 0.0852), while CGAFusion is comparable to cross-attention (0.0842 vs. 0.0825). Hist2Depth performs best, reducing the best competing AbsRel from 0.0776 to 0.0649

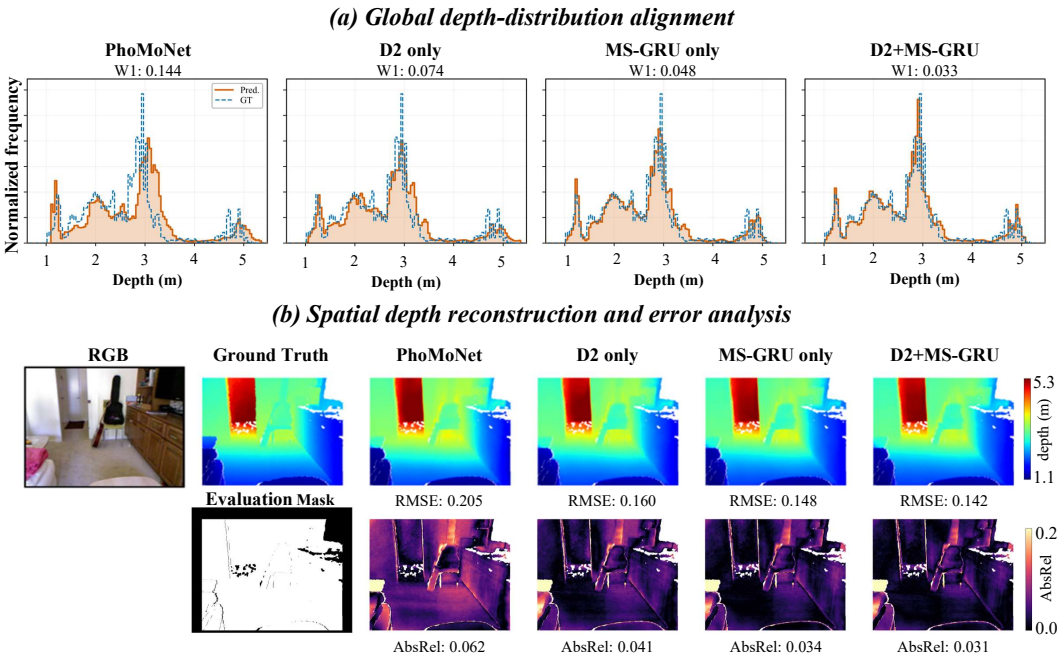

Figure 4: Component visualization on NYU-Depth-v2. (a) Predicted and ground-truth depth distributions with Wasserstein-1 (W1) distance. (b) Reconstructed depth and pixel-wise AbsRel maps.

(-16.4%). These results show that increasing fusion complexity alone does not guarantee better performance and support the effectiveness of Hist2Depth's global-to-local formulation.

Component analysis. Table 3(b) shows that both stages are effective. Global D2 already provides a strong improvement (AbsRel: 0.1037→0.0702), while MS-GRU yields the larger standalone gain (0.1037→0.0663). Combining both further reduces AbsRel to 0.0649 and gives the lowest RMSE of 0.2750 m, supporting their complementary roles in global metric calibration and spatially adaptive refinement. Figure 4 visualizes the global-to-local refinement: Global D2 reduces the distribution mismatch, while MS-GRU further improves spatial accuracy, with the full model performing best.

Design sensitivity. Table 3(c) examines the polynomial order in Global D2. The low-order mappings yield similar accuracy, indicating limited sensitivity to this choice. We retain the quadratic mapping for its additional flexibility in modeling nonlinear depth-dependent errors.

Table 3(d) varies the number of inference updates using the same checkpoint trained with three recurrent iterations. Accuracy improves as the refinement proceeds from one to three updates. Since the recurrent weights are shared, the inference depth can be adjusted without retraining or adding parameters, providing an accuracy–runtime trade-off. We use three updates by default.

5 Conclusion

We present Hist2Depth, a global-to-local framework for monocular metric depth estimation from spatially unresolved SPAD measurements. Global D2 addresses the underconstrained dense correction through distribution-based metric calibration, while MS-GRU addresses the missing pixel–measurement correspondence by querying SPAD evidence with evolving depth hypotheses. Experiments on NYU-Depth-v2 and captured RGB–SPAD data show consistent gains across MDE backbones and reduced photon budgets; future work will extend evaluation to more diverse captured datasets.

AI USE STATEMENT

AI-assisted tools were used to help restructure and polish the author-provided manuscript, assist with reference retrieval and related-work discovery, and support the drafting and condensation of selected sections. The method specifications, experimental protocols, numerical results, and original qualitative figures were provided by the authors. No new experimental results were generated by AI-assisted tools.

REFERENCES

Ashutosh Agarwal and Chetan Arora. Attention attention everywhere: Monocular depth prediction with skip attention. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5861–5870, 2023. doi: 10.1109/WACV56688.2023.00581.

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. AdaBins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4009–4018, 2021. doi: 10.1109/CVPR46437.2021.00400.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth Pro: Sharp monocular metric depth in less than a second. In *International Conference on Learning Representations (ICLR)*, 2025.

Zixuan Chen, Zewei He, and Zhe-Ming Lu. DEA-Net: Single image dehazing based on detail-enhanced convolution and content-guided attention. *IEEE Transactions on Image Processing*, 33: 1002–1015, 2024. doi: 10.1109/TIP.2024.3354108.

Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2002–2011, 2018. doi: 10.1109/CVPR.2018.00214.

Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In *International Conference on Learning Representations (ICLR)*, 2025.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9492–9502, 2024. doi: 10.1109/CVPR52733.2024.00907.

Hongjae Lee, Jinbum Park, Wooseok Jeong, and Seung-Won Jung. Monocular depth estimation network with single-pixel depth guidance. *Optics Letters*, 48(3):594–597, 2023. doi: 10.1364/OL.478375.

Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.

Haotong Lin, Sida Peng, Jingxiao Chen, Songyou Peng, Jiaming Sun, Minghuan Liu, Hujun Bao, Jiashi Feng, Xiaowei Zhou, and Bingyi Kang. Prompting depth anything for 4k resolution accurate metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17070–17080, 2025. doi: 10.1109/CVPR52734.2025.01591.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth Anything 3: Recovering the visual space from any views. In *International Conference on Learning Representations (ICLR)*, 2026.

Mark Nishimura, David B. Lindell, Christopher A. Metzler, and Gordon Wetzstein. Disambiguating monocular depth estimation with a single transient. In *European Conference on Computer Vision (ECCV)*, pp. 139–155, 2020. doi: 10.1007/978-3-030-58589-1_9.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12159–12168, 2021. doi: 10.1109/ICCV48922.2021.01196.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2022. doi: 10.1109/TPAMI.2020.3019967.

Patrick Rim, Hyoungseob Park, Vadim Ezhov, Jeffrey Moon, and Alex Wong. Radar-guided polynomial fitting for metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26941–26952, 2026.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *European Conference on Computer Vision (ECCV)*, pp. 746–760, 2012. doi: 10.1007/978-3-642-33715-4_54.

Zhanghao Sun, David B. Lindell, Olav Solgaard, and Gordon Wetzstein. SPADnet: Deep RGB-SPAD sensor fusion assisted by monocular depth estimation. *Optics Express*, 28(10):14948–14962, 2020. doi: 10.1364/OE.392386.

Massimiliano Viola, Kevin Qu, Nando Metzger, Bingxin Ke, Alexander Becker, Konrad Schindler, and Anton Obukhov. Marigold-DC: Zero-shot monocular depth completion with guided diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5359–5370, 2025. doi: 10.1109/ICCV51701.2025.00509.

Zehan Wang, Siyu Chen, Lihe Yang, Jialei Wang, Ziang Zhang, Hengshuang Zhao, and Zhou Zhao. Depth anything with any prior. In *International Conference on Learning Representations (ICLR)*, 2026.

Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. What matters when repurposing diffusion models for general dense perception tasks? In *International Conference on Learning Representations (ICLR)*, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10371–10381, 2024a. doi: 10.1109/CVPR52733.2024.00987.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, volume 37, pp. 21875–21911, 2024b. doi: 10.52202/079017-0688.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Learning enriched features for real image restoration and enhancement. In *European Conference on Computer Vision (ECCV)*, pp. 492–511, 2020. doi: 10.1007/978-3-030-58595-2_30.

Qiankun Zhao, Yingcai Wan, Jiqian Xu, and Lijin Fang. Cross-modal attention fusion network for RGB-D semantic segmentation. *Neurocomputing*, 548:126389, 2023. doi: 10.1016/j.neucom.2023.126389.

A Implementation Details

We provide additional details required to reproduce Hist2Depth. We first specify the Global D2 and MS-GRU architectures, followed by the decoder features used for different MDE backbones. We then describe the preprocessing of captured RGB–SPAD measurements and implementation settings omitted from the main paper.

A.1 Global D2 Architecture

Global D2 predicts three calibration coefficients from the discrepancy between the monocular and SPAD-derived depth distributions. The SPAD transient is first encoded by a pretrained SPAD encoder following the PhoMoNet design (Lee et al., 2023); this encoder remains frozen throughout Hist2Depth training. The resulting representation is combined with the base-depth histogram, target histogram, and cumulative distribution discrepancy. Table 4 summarizes the corresponding parameter-prediction branch.

Table 4: Architecture of the Global D2 parameter-prediction branch.

Layer Index	Layer Type	Input	Output Channels
0	Input	SPAD transient (1024)	–
		Base-depth histogram (64)	–
		SPAD-derived target histogram (64)	–
		CDF discrepancy (64)	–
1	FC + Leaky ReLU	SPAD transient	256
2	FC + Leaky ReLU	Layer 1	128
3	FC	Layer 2	128
4	FC	Layer 3	64
5	Concatenation	Layer 4 + three 64-D distribution descriptors	256
6	FC + ReLU + Dropout	Layer 5	128
7	FC + ReLU	Layer 6	64
8	FC + Tanh	Layer 7	3

Layers 1–3 form the shared SPAD encoder. Layer 4 projects the SPAD representation to 64 dimensions before fusion with the three distribution descriptors. The dropout rate is 0.1. Layer 8 predicts the bounded coefficients β_0, β_1, and α_0.

A.2 Static Context Initialization

MS-GRU separates static scene context from the measurement evidence that changes with the current depth estimate. The static branch combines decoder features, RGB appearance, calibrated depth, reflectance, the shared SPAD representation, and the post-calibration distribution discrepancy. It initializes the recurrent states and context-dependent gate biases once before iterative refinement.

Table 5: Architecture of the static context initialization in MS-GRU.

Layer Index	Layer Type	Input	Output Channels
0	Input	Selected MDE decoder representation $\mathbf{F}_d$ (256)	–
		RGB image $\mathbf{I}$ (3)	–
		Normalized D2 depth $\bar{D}_0$ (1)	–
		Grayscale reflectance R (1)	–
		Shared SPAD feature $\mathbf{f}_s$ (128)	–
		Post-calibration CDF discrepancy $\Delta\mathbf{Q}_0$ (64)	–
1	Conv2D + GN + ReLU	$[\mathbf{F}_d, \mathbf{I}]$	128
2	Conv2D + GN + ReLU	Layer 1	128

Layer Index	Layer Type	Input	Output Channels
3	FC + ReLU	$[\mathbf{f}_s, \Delta\mathbf{Q}_0]$	64
4	FC + ReLU	Layer 3	32
5	Concatenation	Layer 2 + broadcast Layer 4 + $\bar{D}_0 + R$	–
6	Conv2D + GN + ReLU	Layer 5	128
7	Stride-2 Conv2D + GN + ReLU	Current-scale context	128
8	Conv2D + GN + ReLU	Layer 7	128
9	Conv2D + Tanh	Layer 8	128
10	Conv2D	Layer 8	384

Layers 7–10 are applied successively at the fine, middle, and coarse resolutions. Layer 9 initializes $H_n^{(0)}$, while Layer 10 is divided into the fixed context-bias maps $B_{z,n}$, $B_{r,n}$, and $B_{q,n}$. The static context is computed once and remains fixed during recurrent refinement.

A.3 Recurrent Refinement

The recurrent branch updates its measurement evidence according to the current depth prediction. At each iteration, coarse and fine depth-axis evidence is encoded independently and queried around each pixel's current depth hypothesis. The resulting lookup feature is integrated through coarse-to-fine ConvGRU updates.

Table 6: Architecture of the iterative refinement branch in MS-GRU.

Layer Index	Layer Type	Input	Output Channels
0	Input	Coarse evidence $\mathbf{E}_c^{(k)}$	4
		Fine evidence $\mathbf{E}_f^{(k)}$	4
		Depth-state maps $[\bar{D}^{(k)}, \bar{R}^{(k)}]$	2
11	Conv1D + GN + ReLU	$\mathbf{E}_c^{(k)}$	8
12	Conv1D + GN + ReLU	Layer 11	8
13	Conv1D + GN + ReLU	$\mathbf{E}_f^{(k)}$	8
14	Conv1D + GN + ReLU	Layer 13	8
S	Depth-conditioned lookup	Layers 12 and 14 + $D_{\text{low}}^{(k)}$	112
15	Conv2D + GN + ReLU	Lookup feature	64
16	Conv2D + GN + ReLU	Layer 15	64
17	Conv2D + GN + ReLU	Depth-state maps	32
18	Conv2D + GN + ReLU	Layer 17	32
19	Concatenation	Layers 16 and 18	96
20	Conv2D + ReLU	Layer 19	128
21-c	Context-biased ConvGRU	Coarse recurrent input	128
21-m	Context-biased ConvGRU	Middle recurrent input	128
21-f	Context-biased ConvGRU	Fine recurrent input + $\mathbf{F}_{\text{ref}}^{(k)}$	128
22	Conv2D + ReLU	$H_f^{(k+1)}$	256
23	Conv2D	Layer 22	1
24	Conv2D + ReLU	$H_f^{(k+1)}$	64
25	Conv2D	Layer 24	36

The coarse and fine evidence axes contain 64 and 1024 bins, respectively. With lookup radii $r_c = 2$ and $r_f = 4$, the two branches provide 40-D and 72-D features, yielding the 112-D lookup. Layer 23 predicts the recurrent depth update, while Layer 25 predicts the 36-channel convex-upsampling mask.

A.4 BACKBONE FEATURE EXTRACTION

The same Global D2 and MS-GRU architecture is used across all six evaluated MDE backbones. The only backbone-dependent interface is the decoder representation used to provide spatial context. Table 7 lists the specific model variants and feature extraction points used in our experiments.

Table 7: Model variants and decoder feature extraction used for spatial context across MDE backbones.

Backbone	Model Variant	Extraction Point	Channels	Construction
BTS	BTS-DenseNet161 (`bts_size=512`, NYU)	`bn4_2`	256	Intermediate decoder feature before DenseASPP.
AdaBins	AdaBins (EfficientNet-B5 + mViT, 256 bins)	`decoder.up3`	256	Decoder feature from the PhoMoNet-derived AdaBins branch.
PixelFormer	PixelFormer-Large07 (Swin-Large, window size 7)	`q0` + `q2`	$128 + 128$	Spatial alignment followed by channel concatenation to 256 channels.
ZoeDepth-N	ZoeDepth-N	`r1` (`refinenet1`)	256	MiDaS/DPT refinement feature.
Depth Pro	Apple Depth Pro (DINOv2-L/16-384)	Decoder output	256	First MultiresConvDecoder output before the inverse-depth head.
DA3Metric-L	Depth Anything 3 Metric-Large	`refinenet1`	256	DPT refinement feature before metric-depth prediction.

All selected features are spatially aligned with the MS-GRU refinement grid before context encoding. PixelFormer is the only evaluated backbone for which two decoder features are combined, with the 128-channel `q0` and spatially aligned 128-channel `q2` concatenated into the common 256-channel context representation. For the AdaBins setting in the six-backbone comparison, we use the PhoMoNet-derived AdaBins checkpoint and its `decoder.up3` feature branch.

A.5 CAPTURED RGB–SPAD PROCESSING

The captured measurements differ from the synthetic training input mainly in temporal discretization and measurement range. We therefore convert the released measurements to the network representation used during training without modifying the trained model.

The released global transient contains 3584 temporal bins covering approximately 0.4–9.0 m. To align this measured support with the wider depth interval used by the network, we extend the temporal axis by adding 126 background-level bins before the measured interval and 860 bins after it. The added bins are assigned the minimum transient count and therefore extend only the temporal support without introducing additional return peaks. The resulting 4570-bin transient is then linearly resampled to the fixed 1024-bin SPAD axis used by Hist2Depth.

Two representations are constructed from the aligned 1024-bin transient. A min–max normalized copy is passed to the frozen SPAD encoder to obtain the shared SPAD feature. In parallel, the unnormalized aligned transient is processed using the same physical correction described in Sec. 3.2, including background subtraction, range-squared compensation, depth rebinning, and normalization. The physically corrected SPAD distribution is used to construct the 64-bin target distribution q_t and the cumulative discrepancy ΔQ for Global D2, and also provides the depth-axis evidence used by MS-GRU. Global D2 therefore combines the distribution descriptors q_b, q_t, and ΔQ with the projected feature from the frozen SPAD encoder to predict the bounded polynomial coefficients β_0, β_1, and α_0.

All captured-data experiments use the model trained only on synthetic NYU-Depth-v2 measurements. No captured samples are used for fine-tuning, and the learned network parameters are un-

changed. The synthetic-to-real adaptation therefore consists only of converting the measured transient to the temporal and depth-axis representations expected by Hist2Depth.

For photon-budget evaluation, the full measured integer-count histogram is treated as the reference acquisition. The full captured measurement has an average acquisition duration of 816 s, which we use as the nominal reference duration for the photon-budget analysis. Shorter equivalent acquisitions are generated by seeded nested binomial thinning on the original 3584-bin integer-count transient before temporal-support extension and linear resampling. For two successive acquisition durations $T_i < T_{i-1}$, each temporal-bin count is sampled as

$$h_t^{(i)} \sim \mathrm{Binomial}\left(h_t^{(i-1)}, \frac{T_i}{T_{i-1}}\right).$$

Starting from the full measurement, this procedure is applied successively to obtain the 180, 120, 90, 60, 30, and 20 s equivalents using a fixed random seed. Each thinned 3584-bin transient then undergoes the same temporal-support extension and 4570-to-1024 linear resampling described above.

A.6 Additional Training and Evaluation Details

Unless otherwise specified, all backbone experiments use the same training objective for Global D2 and MS-GRU. The pretrained MDE backbone and shared SPAD encoder remain frozen, while Global D2 and MS-GRU are optimized jointly. MS-GRU is unrolled for $K = 3$ iterations during training, with recurrent weighting factor $\gamma = 0.8$.

Training objective. The total objective is

$$\begin{aligned}\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{GRU}} + \lambda_{\mathrm{hist}}\mathcal{L}_{\mathrm{W1}}^{(K)} + \lambda_{\mathrm{smooth}}\mathcal{L}_{\mathrm{smooth}} + \lambda_{\mathrm{res}}\mathcal{L}_{\mathrm{res}} \\ + \lambda_0\mathcal{L}_0 + \lambda_{\mathrm{hist}}\mathcal{L}_{\mathrm{W1}}^{(0)} + \lambda_{\mathrm{coeff}}\mathcal{L}_{\mathrm{coeff}},\end{aligned} \tag{4}$$

where

$$\lambda_{\mathrm{hist}} = 0.2, \qquad \lambda_{\mathrm{smooth}} = 0.02, \qquad \lambda_{\mathrm{res}} = 0.001,$$
$$\lambda_0 = 1.0, \qquad \lambda_{\mathrm{coeff}} = 10^{-4}.$$

Depth supervision. Let D_{gt} denote the ground-truth depth. For valid pixels $p \in \Omega$, define

$$g_p = \log D_p - \log D_{\mathrm{gt},p}.$$

The SILog loss is

$$\mathcal{L}_{\mathrm{SILog}}(D, D_{\mathrm{gt}}) = 10\sqrt{\mathrm{Var}_{p\in\Omega}(g_p) + 0.15\,[\mathrm{Mean}_{p\in\Omega}(g_p)]^2}. \tag{5}$$

The Global D2 output D_0 is directly supervised by

$$\mathcal{L}_0 = \mathcal{L}_{\mathrm{SILog}}(D_0, D_{\mathrm{gt}}). \tag{6}$$

All recurrent predictions are also supervised. With

$$w_k = \gamma^{K-k},$$

the recurrent depth loss is

$$\mathcal{L}_{\mathrm{GRU}} = \frac{\sum_{k=1}^{K} w_k\,\mathcal{L}_{\mathrm{SILog}}(D^{(k)}, D_{\mathrm{gt}})}{\sum_{k=1}^{K} w_k}. \tag{7}$$

For $K = 3$ and $\gamma = 0.8$, the three weights are 0.64, 0.8, and 1.0.

Distribution consistency. The depth distributions constructed from D_0 and $D^{(K)}$ use the differentiable histogram representation introduced in Sec. 3.2. Let $\mathbf{q}_0$ and $\mathbf{q}_K$ denote the corresponding 64-bin distributions, and let $\mathbf{q}_t$ denote the SPAD-derived target distribution.

For a predicted distribution $\mathbf{q}$ and target distribution $\mathbf{q}_t$, define

$$Q_m = \sum_{j=1}^{m} q_j, \qquad Q_{t,m} = \sum_{j=1}^{m} q_{t,j}.$$

The discrete one-dimensional Wasserstein-1 loss is

$$\mathcal{L}_{\mathrm{W1}}(\mathbf{q}, \mathbf{q}_t) = \Delta \sum_{m=1}^{M} \left| Q_m - Q_{t,m} \right|, \tag{8}$$

where $M = 64$ and

$$\Delta = \frac{D_{\max} - D_{\min}}{M}.$$

The two distribution losses are

$$\mathcal{L}_{\mathrm{W1}}^{(0)} = \mathcal{L}_{\mathrm{W1}}(\mathbf{q}_0, \mathbf{q}_t), \qquad \mathcal{L}_{\mathrm{W1}}^{(K)} = \mathcal{L}_{\mathrm{W1}}(\mathbf{q}_K, \mathbf{q}_t). \tag{9}$$

Both terms use $\lambda_{\mathrm{hist}} = 0.2$.

Residual regularization. Let

$$\Delta D^{(K)} = D^{(K)} - D_0$$

denote the final spatial residual relative to the Global D2 initialization. Its edge-aware smoothness loss is

$$\begin{aligned} \mathcal{L}_{\mathrm{smooth}} = \ & \mathrm{Mean}\left[|\partial_x \Delta D^{(K)}| \exp(-\alpha G_x(I)) \right] \\ & + \mathrm{Mean}\left[|\partial_y \Delta D^{(K)}| \exp(-\alpha G_y(I)) \right], \end{aligned} \tag{10}$$

where $G_x(I)$ and $G_y(I)$ denote the mean absolute RGB gradients in the horizontal and vertical directions, respectively, and $\alpha = 10$.

The residual-magnitude regularizer is

$$\mathcal{L}_{\mathrm{res}} = \frac{\mathrm{Mean}\left(|\Delta D^{(K)}| \right)}{D_{\max} - D_{\min}}. \tag{11}$$

Coefficient regularization. The three Global D2 coefficients are regularized by

$$\mathcal{L}_{\mathrm{coeff}} = \frac{1}{3}\left(\beta_0^2 + \beta_1^2 + \alpha_0^2 \right). \tag{12}$$

This term suppresses unnecessarily large global calibration parameters.

Optimization and augmentation. Training uses a batch size of 2 and an AdamW weight decay of 10^{-4}. Data augmentation consists of random horizontal flipping, rotation within $\pm 2.5^\circ$, and mild gamma, brightness, and color perturbations. Spatial augmentation is performed before synthetic SPAD generation so that the transient corresponds to the same RGB/depth field of view used by the network.

Runtime and evaluation protocol. Parameter counts and inference times are measured on an NVIDIA RTX 5090 using FP32 with batch size 1. Runtime is averaged over 100 synchronized forward passes after 20 warm-up iterations. Data loading, metric computation, and result saving are excluded. For SSD, the reported runtime includes its CPU/NumPy histogram-transport procedure.

B Additional Quantitative Results

We provide the complete quantitative results for the captured RGB–SPAD experiments discussed in the main paper. Results already reported for NYU-Depth-v2, cross-backbone evaluation, and ablation studies are not repeated.

B.1 CAPTURED RGB–SPAD COMPARISON

Table 8 reports the complete set of depth metrics under the full captured measurement. This complements the AbsRel and RMSE values highlighted in the main paper.

Table 8: Quantitative comparison on captured RGB–SPAD measurements under the full-acquisition setting.

Method	Input	AbsRel↓	SqRel↓	RMSE↓	$\text{RMSE}_{\log}$↓	$\log_{10}$↓	δ_1↑	δ_2↑	δ_3↑
DORN	RGB	0.2249	0.2376	0.9067	0.2962	0.1095	0.4377	0.9065	0.9946
BTS	RGB	0.2428	0.2597	0.9545	0.3015	0.1159	0.3862	0.9117	**0.9974**
AdaBins	RGB	0.2069	0.2315	0.9135	0.2710	0.1003	0.5175	0.9241	0.9953
PixelFormer	RGB	0.2356	0.2624	0.9537	0.3045	0.1172	0.3999	0.8876	0.9917
SSD (DORN)	RGB+SPAD	0.1676	0.2198	0.6051	0.2075	0.0612	0.8217	0.9543	0.9848
SSD (BTS)	RGB+SPAD	0.1669	0.1975	0.5743	0.2049	0.0615	0.8059	0.9639	0.9908
SSD (AdaBins)	RGB+SPAD	0.1590	0.1987	0.5819	0.1967	0.0586	0.8325	0.9657	0.9890
SSD (PixelFormer)	RGB+SPAD	0.1489	0.2042	0.5665	0.1890	0.0542	0.8707	0.9711	0.9859
PhoMoNet (AdaBins)	RGB+SPAD	0.1628	0.2596	0.7007	0.2376	0.0725	0.8119	0.9214	0.9517
Hist2Depth (AdaBins)	RGB+SPAD	**0.1352**	**0.1221**	**0.4677**	**0.1715**	**0.0522**	**0.8772**	**0.9809**	0.9933

B.2 PHOTON-BUDGET EVALUATION

Table 9 gives the complete values underlying the photon-budget curves in Fig. 3. The full measurement has an average acquisition duration of 816 s.

Table 9: Captured RGB–SPAD performance under different equivalent acquisition durations.

Metric	Duration	AdaBins	PixelFormer	SSD (AdaBins)	SSD (PixelFormer)	PhoMoNet (AdaBins)	**Hist2Depth (AdaBins)**
AbsRel↓	Full	0.2069	0.2356	0.1590	0.1489	0.1628	**0.1352**
	180 s	0.2069	0.2356	0.1592	0.1490	0.1633	**0.1356**
	120 s	0.2069	0.2356	0.1596	0.1493	0.1633	**0.1364**
	90 s	0.2069	0.2356	0.1599	0.1495	0.1637	**0.1369**
	60 s	0.2069	0.2356	0.1601	0.1496	0.1636	**0.1373**
	30 s	0.2069	0.2356	0.1604	0.1497	0.1649	**0.1438**
	20 s	0.2069	0.2356	0.1621	**0.1511**	0.1696	0.1514
RMSE (m)↓	Full	0.9135	0.9537	0.5819	0.5665	0.7007	**0.4677**
	180 s	0.9135	0.9537	0.5807	0.5668	0.7004	**0.4674**
	120 s	0.9135	0.9537	0.5819	0.5678	0.6986	**0.4692**
	90 s	0.9135	0.9537	0.5830	0.5683	0.6996	**0.4701**
	60 s	0.9135	0.9537	0.5831	0.5688	0.6976	**0.4707**
	30 s	0.9135	0.9537	0.5829	0.5682	0.6953	**0.4924**
	20 s	0.9135	0.9537	0.5886	0.5730	0.7108	**0.5198**
δ_1↑	Full	0.5175	0.3999	0.8325	0.8707	0.8119	**0.8772**
	180 s	0.5175	0.3999	0.8321	0.8705	0.8126	**0.8789**
	120 s	0.5175	0.3999	0.8313	0.8700	0.8129	**0.8790**
	90 s	0.5175	0.3999	0.8307	0.8697	0.8122	**0.8772**
	60 s	0.5175	0.3999	0.8303	0.8692	0.8134	**0.8783**
	30 s	0.5175	0.3999	0.8295	**0.8692**	0.8137	0.8681
	20 s	0.5175	0.3999	0.8267	**0.8680**	0.7980	0.8523

Full denotes the complete captured measurement with an average acquisition duration of 816 s. Shorter durations are generated from the same measured histograms using nested binomial thinning.

C Additional Qualitative Results

We provide additional qualitative examples covering different monocular backbones, SPAD signal levels, and captured RGB–SPAD measurements. For each example, the corresponding depth predictions and pixel-wise AbsRel maps are shown for visual comparison.

DA3Metric-L at SBR$= 100$. Figure 5 shows two representative indoor scenes from NYU-Depth-v2. The comparison includes the RGB input, ground-truth depth and valid mask, PhoMoNet, the RGB-only DA3Metric-L prediction, SSD with DA3Metric-L, and Hist2Depth with DA3Metric-L.

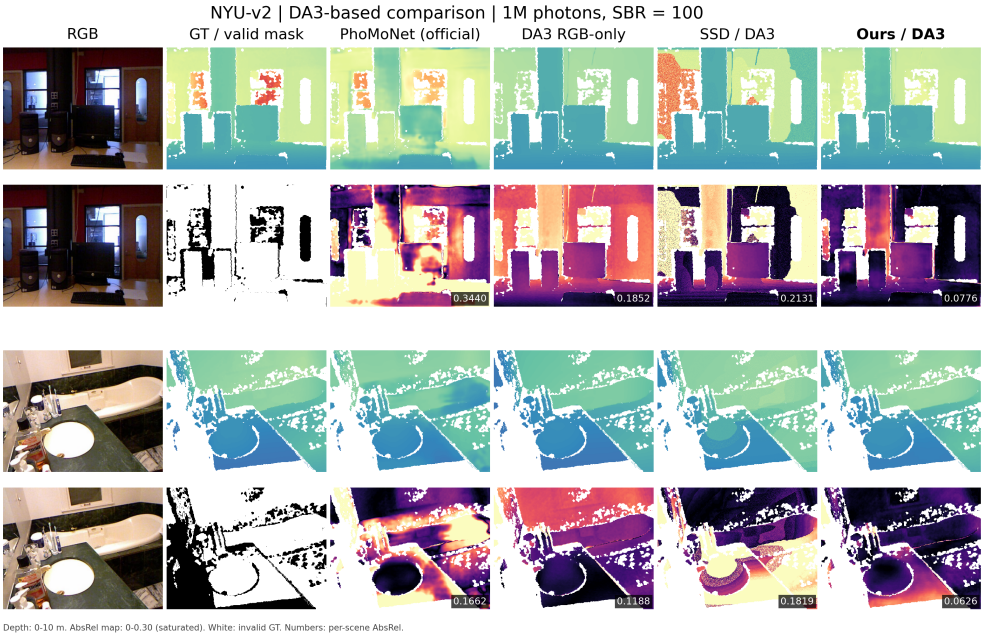

Figure 5: Qualitative comparison on NYU-Depth-v2 using DA3Metric-L at SBR$= 100$. Depth predictions and pixel-wise AbsRel maps are shown for two representative scenes.

AdaBins at SBR$= 10$. Figure 6 shows two NYU-Depth-v2 examples under the low-SBR setting, including an indoor workspace scene and a kitchen scene. The displayed methods correspond to the AdaBins-based comparison used in the main evaluation.

Captured measurements under different photon budgets. Figure 7 shows a representative classroom scene from the captured RGB–SPAD dataset under the full acquisition and equivalent acquisition durations of 60 s and 20 s. Figure 8 shows more captured RGB–SPAD examples under equivalent acquisition durations of 60 s and 20 s. The displayed scenes include an indoor corridor and a countertop scene. The shorter-duration measurements are generated from the corresponding captured transients using the photon-thinning protocol described in Appendix A.5.

Results with different monocular backbones. Figures 9 and 10 provide qualitative comparisons of RGB-only predictions and the corresponding Hist2Depth results using ZoeDepth-N, DA3Metric-L, and Depth Pro on four NYU-Depth-v2 scenes. Pixel-wise AbsRel maps and per-scene AbsRel values are shown for reference.

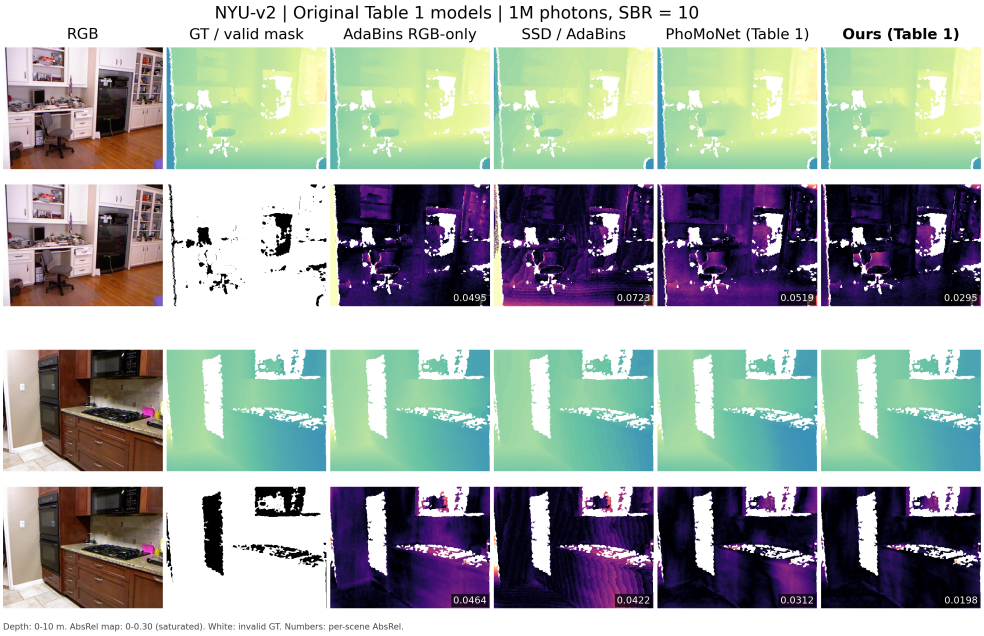

Figure 6: Qualitative comparison on NYU-Depth-v2 using AdaBins at SBR= 10. Results are shown for an indoor workspace scene and a kitchen scene.

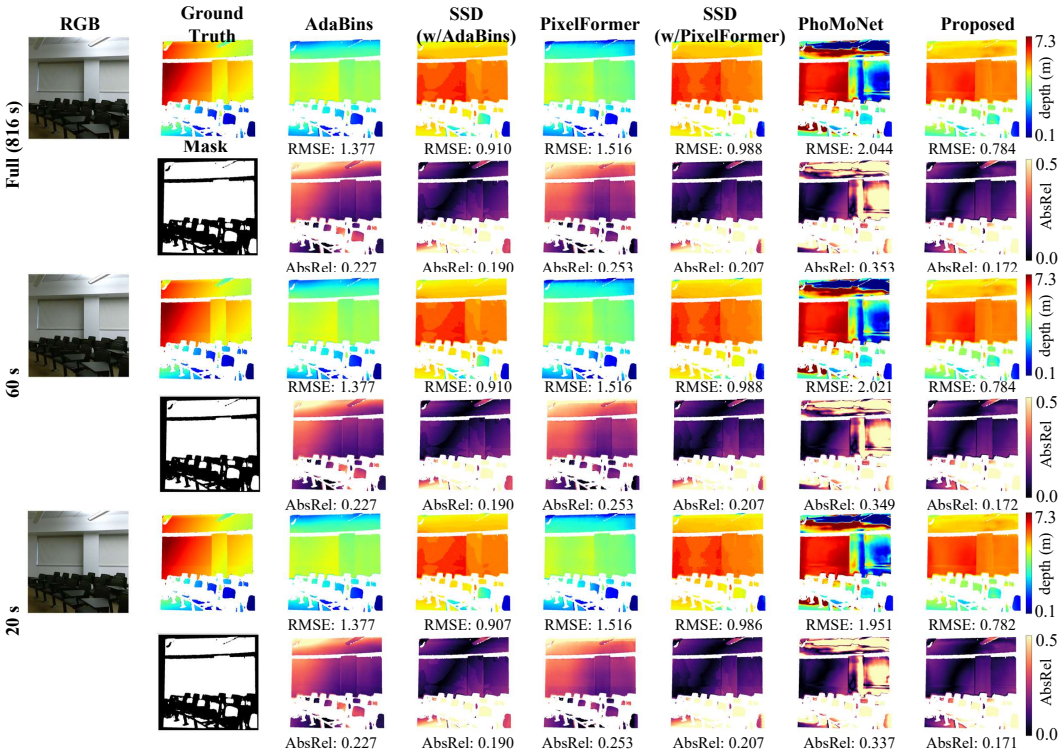

Figure 7: Qualitative comparison on a captured classroom scene using the full measured transient and photon-thinned equivalents corresponding to 60 s and 20 s acquisition durations. Full denotes the original, unthinned captured measurement.

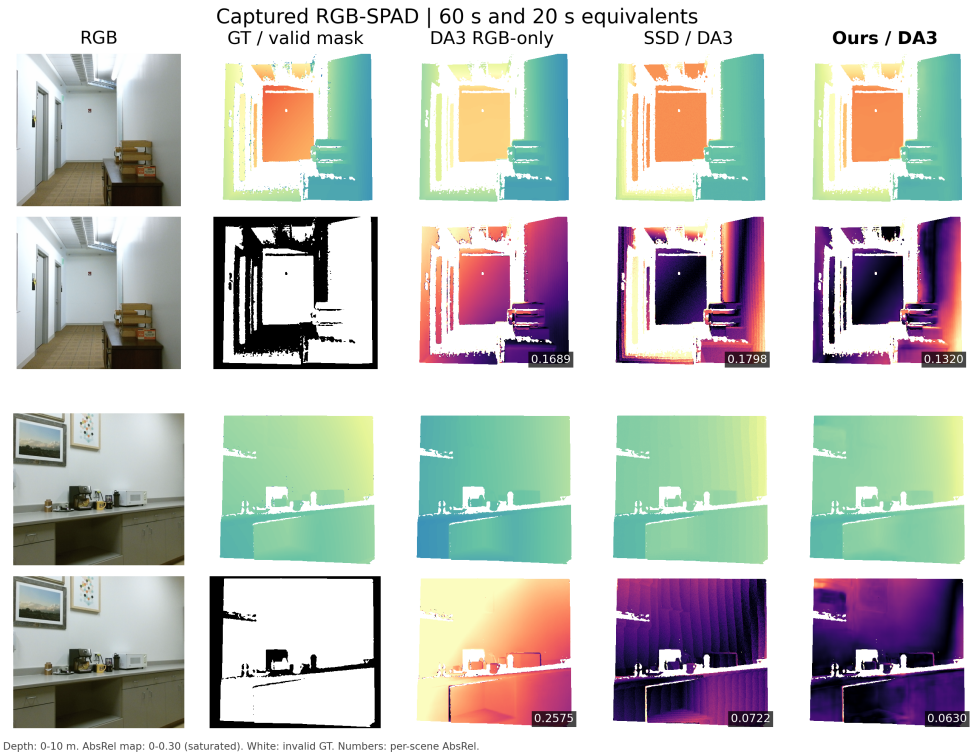

Figure 8: Additional qualitative examples on captured RGB–SPAD data under reduced photon budgets. The corridor scene (top) uses the 60 s equivalent acquisition, while the countertop scene (bottom) uses the 20 s equivalent acquisition. Depth predictions and pixel-wise AbsRel maps are shown for the evaluated methods.

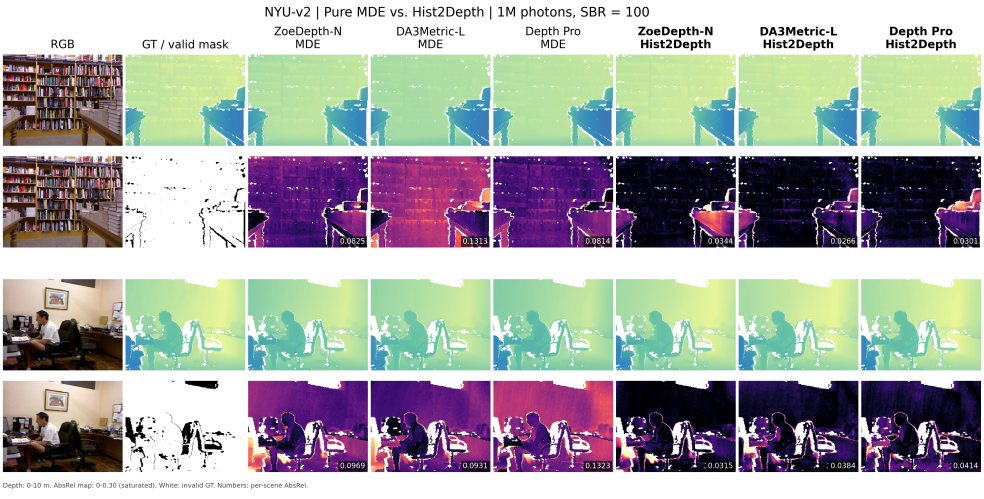

Figure 9: Qualitative comparison with ZoeDepth-N, DA3Metric-L, and Depth Pro on two NYU-Depth-v2 scenes. RGB-only predictions and the corresponding Hist2Depth results are shown.

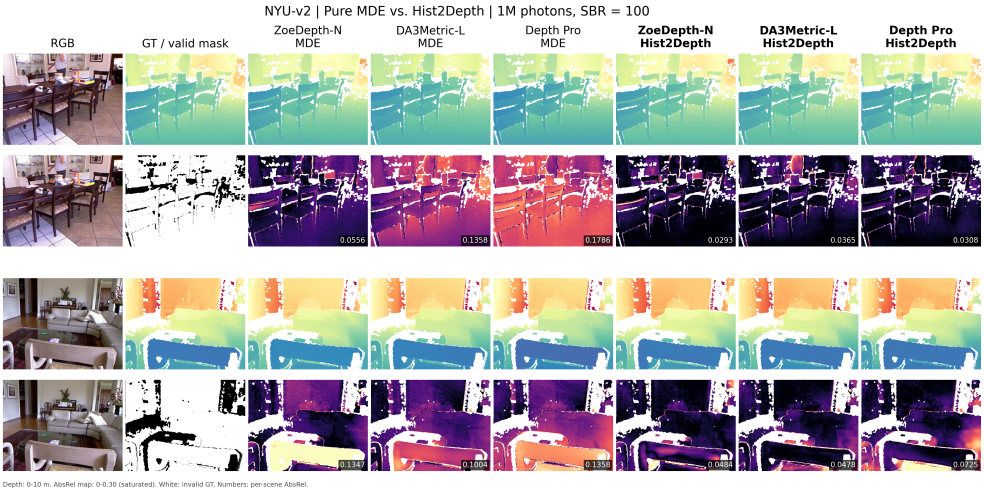

Figure 10: Additional qualitative comparison with ZoeDepth-N, DA3Metric-L, and Depth Pro on two NYU-Depth-v2 scenes. RGB-only predictions and the corresponding Hist2Depth results are shown.