DELTAR: Depth Estimation from a Light-weight ToF Sensor and RGB Image

Yijin Li[1], Xinyang Liu[1], Wenqi Dong[1], Han Zhou[1],
Hujun Bao[1], Guofeng Zhang[1], Yinda Zhang[2⋆], and Zhaopeng Cui[1⋆]

[1] State Key Lab of CAD&CG, Zhejiang University
[2] Google

Abstract. Light-weight time-of-flight (ToF) depth sensors are small, cheap, low-energy and have been massively deployed on mobile devices for the purposes like autofocus, obstacle detection, etc. However, due to their specific measurements (depth distribution in a region instead of the depth value at a certain pixel) and extremely low resolution, they are insufficient for applications requiring high-fidelity depth such as 3D reconstruction. In this paper, we propose DELTAR, a novel method to empower light-weight ToF sensors with the capability of measuring high resolution and accurate depth by cooperating with a color image. As the core of DELTAR, a feature extractor customized for depth distribution and an attention-based neural architecture is proposed to fuse the information from the color and ToF domain efficiently. To evaluate our system in real-world scenarios, we design a data collection device and propose a new approach to calibrate the RGB camera and ToF sensor. Experiments show that our method produces more accurate depth than existing frameworks designed for depth completion and depth super-resolution and achieves on par performance with a commodity-level RGB-D sensor. Code and data are available on the project webpage.

Keywords: Light-weight ToF Sensor, Depth Estimation.

1 Introduction

The depth sensor is a game changer in computer vision, especially with commodity-level products being widely available [22,29,53,33,6]. As the main player, time-of-flight (ToF) sensors have competitive features , e.g., compact and less sensitive to mechanical alignment and environmental lighting conditions. and thus have become one of the most popular classes in the depth sensor market. However, the price and power consumption, though already significantly lower than other technologies such as structured light (Microsoft Kinect V1) , are still one to two orders of magnitudes higher than a typical RGB camera when reaching a similar resolution due to a large number of photons needs to be emitted, collected, and processed. On the other hand, light-weight ToF sensors are designed to be low-cost, small, and low-energy, which have been massively deployed on mobile

⋆ Corresponding authors

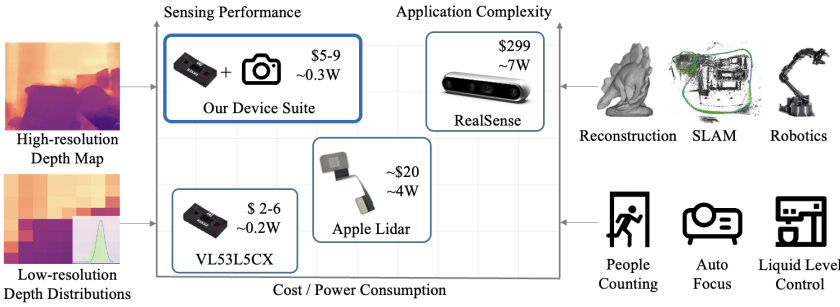


Fig. 1: Comparison between different depth sensors. Low-cost and low-power-consumed sensors like VL53L5CX are designed for simple applications such as people counting and autofocus. In this paper, we show how to improve the depth quality to be on par with a commodity-level RGB-D sensor by our DELTAR algorithm.[1]

devices for the purposes like autofocus, obstacle detection, etc [41]. However, due to the light-weight electronic design, the depth measured by these sensors has more uncertainty (i.e., in a distribution instead of single depth value) and low spatial resolution (e.g., $< 10 \times 10$), and thus cannot support applications like 3D reconstruction or SLAM [22,6], that require high-fidelity depth (see Fig. 1).

In contrast, RGB cameras are also widely deployed in modern devices with the advantage of capturing rich scene context at high resolution, but they are not able to estimate accurate depth with a single capture due to the inherent scale ambiguity of monocular vision. We observe that these two sensors sufficiently complement each other and thus propose a new setting, i.e., estimating accurate dense depth maps from paired sparse depth distributions (by the light-weight ToF sensor) and RGB image. The setting is essentially different from previous depth super-resolution and completion in terms of the input depth signal. Specifically, the task of super-resolution targets relatively low-resolution consumer depth sensors, (e.g., 256×192 for the Apple LIDAR and 240×180 for the ToF sensor on the Huawei P30). In contrast, our task targets light-weight ToF sensors with several orders of magnitude lower resolution (e.g., 8×8), but provides a depth distribution per zone (see Fig. 2). Depth completion, on the other hand, aims to densify incomplete dense high-resolution maps (e.g., given hundreds of depth samples), which is not available for light-weight ToF sensors. Therefore, our task is unique and challenging due to the extremely low resolution of the input depth but accessibility to the rich depth distribution.

To demonstrate, we use ST VL53L5CX [42] (denoted as L5) ToF sensor, which outputs 8×8 zones, each with a depth distribution, covering a total of 63° diagonal field-of-view (FoV) and runs at a power consumption of about 200mW (vs. 4W of an Apple Lidar). To fully exploit the L5 depth signals, we

[1] Icon credit: Iconfinder [20]

design DELTAR (**D**epth **E**stimation from **L**ight-weight **T**oF **A**nd **R**GB image), a neural network architecture tailored with respect to the underlying physics of the L5 sensors. Specifically, we first build the depth hypothesis map sampled from the distribution reading of L5, and then use cross-domain attention to exchange the information between the RGB image and the depth hypothesis. A self-attention is also run on image domain to exchange the information between regions covered by L5 and beyond, hence the output aligns with the RGB image and covers the whole FoV. Experiments show that DELTAR outperforms existing architectures designed for depth completion and super-resolution, and improve the raw depth readings of L5 to maintain the quality on par with commodity-level depth sensors, such as Intel RealSense D435i.

Moreover, as no public datasets are available for this new task, we build a capturing system by mounting an L5 sensor and a RealSense RGB-D sensor on a frame-wire with reasonable field-of-view overlap. To align the RGB image and L5's zones, we need to calibrate the sensors, which is challenging as the correspondence cannot be trivially built between two domains. To this end, we propose a new calibration method. An EM-like algorithm is first designed to estimate the plane from L5 signals and then the extrinsic parameters between the L5 sensor and the color camera are optimized by solving point-to-plane alignment in a natural scene with multiple planes. With this capturing system, we create a dataset called ZJU-L5, which includes about 1000 L5-image pairs from 15 real-world scenes with pixel-aligned RGB and ToF signals for training and evaluation purposes. Besides the real-world data, we also simulate synthetic L5 signals using depth from NYU-Depth V2 dataset and use them to augment the training data. The dataset is publicly available to facilitate and inspire further research in the community.

Our contributions can be summarized as follows. First, we demonstrate that light-weight ToF sensors designed for autofocus can be empowered for high-resolution and accurate depth estimation by cooperating with a color image. Second, we prove the concept with a hardware setup and design a cross-domain calibration method to align RGB and low-resolution depth distributions, which enables us to collect large-scale data. The dataset is released to motivate further research. Third, we propose DELTAR, a novel end-to-end transformer-based architecture, based on the sensors' underlying physics, can well utilize the captured depth distribution from the sensor and the color image for dense depth estimation. Experiments show that DELTAR performs better than previous architectures designed for depth completion or super-resolution and achieves more accurate depth prediction results.

2 Related Work

Monocular Depth Estimation. These methods predict a dense depth map for each pixel with a single RGB image. Early approaches [38,37,39,40] use hand-crafted features or graphical models to estimate a depth map. More recent methods employ deep CNN [9,23,48,13,49,15] due to its strong feature representation. Among them, some methods [24,19] exploit assumptions about indoor environments, e.g., plane constraints, to regularize the network. Other methods [36,35]

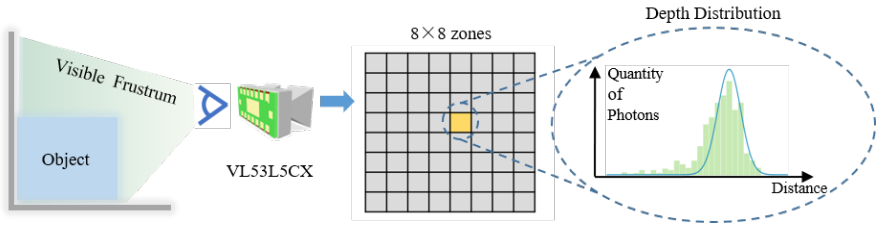


Fig. 2: L5 Sensing Principle. L5 has an extremely low resolution (8x8 zones) and provides depth distribution per zone.

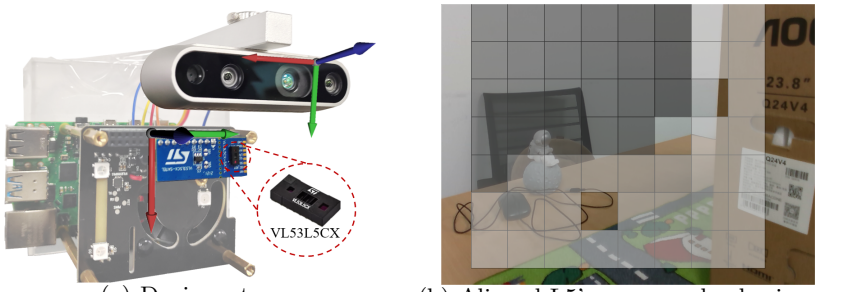
try to benefit from more large-scale and diverse data samples by designing loss functions and mixing strategies. Besides, [11,1] propose to model depth estimation as a classification task or hybrid regression to improve accuracy and generalization. Nonetheless, these methods cannot generalize well on different scenes due to their lack of metric scale innately.

Depth Completion. Depth completion aims to recover a high-resolution depth map given some sparse depth samples and an RGB image. Spatial propagation network(SPN) series methods [27,5,4,31] are one of the most popular methods which learned local affinities to refine depth predictions. Recently, some works [3,50,34] attempt to introduce 3D geometric cues in the depth completion task, e.g., by introducing surface normals as the intermediate representation, or learn a guided network [44] to utilize the RGB image better. More recently, PENet [16] propose an elaborate two-branch framework, which reaches the state-of-the-art. This type of method, however, is not suitable for our task because it assumes the pixel-wise depth-to-RGB alignment, while light-weight ToF sensors only provide a coarse depth distribution in each zone area without exact pixel-wise correspondence.

Depth Super-Resolution. This task aims to boost the consumer depth sensor to a higher spatial resolution to match the resolution of RGB images. Most early works are based on filtering [26,51] or formulate depth super-resolution as an optimization problem [30,7]. Later researches focus more on learning-based method [46,47,45,18]. Among them, Xia et al. [47] propose a task-agnostic network which can be used to process depth information from different sources. Wang et al. [45] iteratively updates the intermediate feature map to be consistent with the given low-resolution depth. In contrast to these methods which usually take a depth map with more than 10 thousand pixels as input, our task targets light-weight ToF sensors with several orders of magnitude lower resolution (e.g., 8×8), but provides a depth distribution per region.

3 Hybrid Sensor Setup

This paper aims to predict a high-resolution depth image from a light-weight ToF sensor (e.g., L5) guided by a color image. While no public datasets are available,



(a) Device setup (b) Aligned L5's zones and color image

Fig. 3: Hybrid sensor setup. (a) We mount a L5 with an Intel RealSense 435i on a metal frame. (b) Blending color images with L5' depth. White color represents close range, black color represents long range. According to the valid status returned by L5, we hide all invalid zones which may receive too less photons or fail in measurement consistency.

we build a device with hybrid sensors and propose the calibration method for this novel setup.

3.1 L5 Sensing Principle

L5 is a light-weight ToF-based depth sensor. In conventional ToF sensors, the output is typically in a resolution higher than 10 thousand pixels and measures the per-pixel distance along the ray from the optical center to the observed surfaces. In contrast, L5 provides multiple depth distributions with an extremely low resolution of 8×8 zones, covering 63° diagonal FoV in total. The distribution is originally measured by counting the number of photons returned in each discretized range of time, and then fitted with a Gaussian distribution (see Fig. 2) in order to reduce the broadband load and energy consumption since only mean and variance needs to be transmitted. Due to the low resolution and high uncertainty of L5, it cannot be directly used for indoor dense depth estimation. Please refer to supplementary materials for more details about L5.

3.2 Device Setup

Fig. 3-(a) shows our proposed device suite. An L5 and an Intel RealSense D435i are mounted on a metal frame facing in the same direction. It is worth noting that we only used RealSense's color camera along with L5 when estimating depth, and the depth output by RealSense is used as the ground truth to measure the quality of our estimation. The horizontal and vertical FoV of the L5 are both 45°, while the RealSense's color camera has 55° horizontal FoV and 43° vertical FoV. As a result, the L5 sensor and the color camera share most of the FoV but not all in our setup.

3.3 Calibration

In order to align the L5 outputs with the color image, we need to calibrate the multi-sensor setup, i.e., computing the relative rotation and translation between the color camera and the L5 sensor. Similar to the calibration between LIDAR and camera [12], we also calibrate our device suite by solving a point-to-plane fitting problem. However, it is not trivial to fit a plane with raw L5 signals since it does not provide the pixel position of the depth value. We observe that, when facing a plane, in each zone $k \in Z$, there must be a location (x_k, y_k), though unknown, whose depth is equal to the mean of the corresponding distribution m_k returned by L5. Therefore we can optimize both the plane parameter $\{n, d\}$ (the frame subscript is omitted for brevity) and the pixel position (x_k, y_k) through:

$$\begin{aligned} \{n, d, x_k, y_k \mid k \in Z\} = \arg\min \sum_{k \in Z} \|n \cdot K^{-1}(x_k, y_k, m_k)^T + d\|^2 \\ \textbf{s.t. } x^k_{\min} \le x_k \le x^k_{\max}, y^k_{\min} \le y_k \le y^k_{\max}, \end{aligned} \tag{1}$$

where $(x, y)^k_{(\min,\max)}$ is the boundary of the zone k in L5 coordinates, and K is the intrinsic matrix. Clearly, Eq. 1 is non-convex thus we solve it by an EM-like algorithm. Specifically, we first initialize all 2D positions at the center of the zone. In the E-step, we back-project these 2D points with measured mean depth, and then fit a 3D plane. In the M-step, we adjust the 2D positions within each zone by minimizing the distance of the 3D points to the plane. The steps run iteratively until convergence. During the iteration, the points that are too far from the plane are discarded from the optimization.

We then obtain the extrinsic transformation matrix by solving a point-to-plane fitting problem. We use our device suite to scan three planes that are not parallel to each other, and ensure that we only observe one plane most of the time. We employ an RGB-D SLAM [28], which recovers from color images a set of camera poses and point cloud P in real-world metric scale, and each point belongs to a certain plane.

For each time stamp $i \in F$, we use P_i to represent the subset of P that are visible in frame i and transformed from the world coordinate system to current RGB camera's. We also have the planar parameters $\{n_i \in \mathbb{R}^3, d_i \in \mathbb{R}\}$ (normal and offset to the origin) in the current L5's coordinate system by solving Eq. 1, then the extrinsic parameters can be solved by minimizing the point to plane distance:

$$\{R, t\} = \arg\min \sum_{i \in F} \sum_{p \in P_i} \|n_i \cdot (R \cdot p + t) + d_i\|^2, \tag{2}$$

where $[R, t]$ are the transformation that map 3D points from the RGB camera's coordinate system to L5's.

For the device setup shown in Fig. 3-(a), we are able to recover the rotation transformation of the two sensors close to 90 degrees. The mean distances between the L5 measurement and the 3D point cloud before and after calibration are 7.5 cm and 1.5 cm, respectively. An example of aligned L5's zones and color image is shown in Fig. 3-(b).

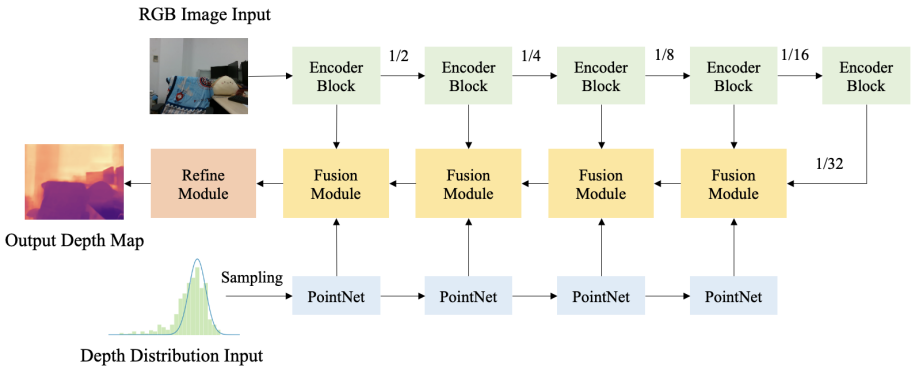


Fig. 4: Overview architecture of the fusion network. Our model takes depth distributions and color image as input, and fuse them at multiple-scale with attention based module before predicting the final depth map.

4 The DELTAR model

With the calibration, we are able to align the L5'zones with the color image. Based on the characteristic of each modality, we propose a novel attention-based network to predict a high-resolution depth image given the aligned L5 signals and color image. We first design a module to extract features from the distribution (Sec. 4.1), and then propose a cross-domain Transformer-based module to fuse with the color image features at different resolutions (Sec. 4.2). Finally, we predict the final depth values through a refinement module (Sec. 4.3). An overview of the proposed method is shown in Fig. 4.

4.1 Hybrid Feature Extraction

Many works have been designed for fine-grained observations, such as RGB images, depth and point clouds. In contrast, how to extract features from distributions is barely studied. A straightforward idea is to encode the mean and variance directly. However, the depth variance is often smaller than the mean by several magnitudes, which may make it difficult to train the network because of internal covariate shift [21]. In Section 5.3 we show that directly encoding the mean and the variance does not work well in our experiments. Therefore, we propose to discretize the distribution by sampling depth hypotheses. Instead of uniform sampling [2,14], we uniformly sample on the inverse cumulative distribution function of the distribution, so the density of the sampling follows the distribution. We utilize PointNet [32] without T-Net to extract features from the sampled depth hypotheses. Multiple pointnets are stacked to extract multi-level distribution features. We use a standard convolutional architecture for the color image, i.e., Efficient B5 [43], to extract multi-level features. Unlike image feature extraction, we do not conduct a down-sample operation when distilling multi-level distribution features.

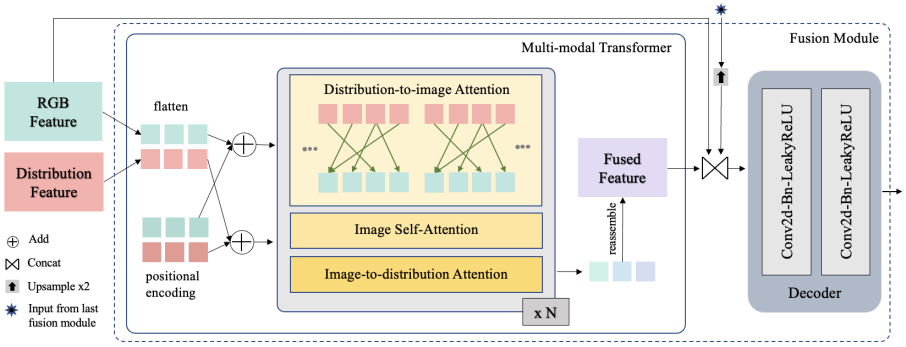


Fig. 5: Details of fusion module. The image and distribution features are flattened into 1-D vectors and added with the positional encoding. The added features are then fused by three different attention mechanisms and concatenated with last layer's feature and the skip-connected image features. Finally, a decoder decodes the concatenated feature and outputs it to the next fusion layer.

4.2 Transformer-based Fusion Module

The fusion module takes two multi-modal data, including image features and distribution features as input, and outputs fused features. In the task of depth completion and super-resolution, depth features and RGB features are usually concatenated or summed in the fusion step [34,16]. This may be sufficient for fine-grained observations which provide the pixel-wise depth-to-RGB alignment, but it is not suitable for our task since pixel-wise alignment is not available between the depth hypothesis map and the RGB features. Inspired by the recent success of Transformer [25,8,17], we adopt attention mechanisms, which process the entire input all at once and learn to focus on sub-components of the cross-modal information and retrieve useful information from each other.

Cross-attention Considering Patch-distribution Correspondence. The Transformer adopts an attention mechanism with the Query-Key-Value model. Similar to information retrieval, the query vector Q retrieves information from the value vector V, according to the attention scores computed from the dot product of queries Q and keys K corresponding to each value. The vanilla version of Transformer contains only self-attention, in which the key, query, and value vectors are from the same set. In multi-modal learning, researchers use cross-attention instead, in which the key and value vectors are from one modal data, and the query vectors are from the other. We first conduct Distribution-to-image attention, that is, taking the key and value vectors from the distribution's feature, and the query vector from the image features, so that the network learns to retrieve information from the candidate depth space. Considering that each distribution from L5 signals corresponds to a specific region in the image, we only conduct cross-attention between the corresponding patch image and the distribution (see Fig. 3-(b)). In Sec. 5.3, we show that conducting cross-attention

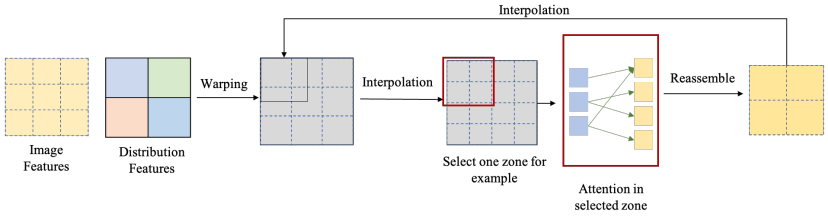


Fig. 6: Interpolation for solving misalignment. When the boundary of the L5's zones and the image feature cannot be aligned precisely, simply quantizing the floating-number boundary could introduce a large negative effect, so we propose to fuse feature after interpolation.

without considering the patch-distribution correspondence leads to severe performance degradation. Empirically, we find that adding image-to-distribution attention leads to better performance.

Propagation by Self-attention. It is not enough to conduct cross-attention as many regions on the image are not covered by the L5's FoV, and these regions cannot benefit from the distribution features. To propagate the depth information further, we also include image self-attention. This step helps the learned depth information propagate to a global context. Besides, the fused feature can be blended to make the feature map smoother. We conduct cross-attention between the two modal data and self-attention over the image feature alternatively for N times, as shown in Fig. 5. In our experiment, we set $N = 2$.

Solving Misalignment by Interpolation. Misalignment occurs when warping L5 zones to an image. See Fig. 6 for a toy example. Simply quantizing the floating-number boundary could introduce a largely negative effect, especially when the fusion is operated on low-resolution feature maps. Moreover, the image resolution corresponding to each zone should be the same to facilitate putting them into a batch. To this end, we fuse on the interpolated feature and then interpolate the fused image features back.

4.3 Refinement Module

We employ the mViT proposed in Adabins [1] as our refinement module to generate the final depth map. Unlike directly regressing depth, the refinement module predicts the depth as a linear combination of multiple depth bins. Specifically, the refinement module predicts a bin-width vector b per image and linear coefficient l at each pixel. The depth-bin's centers $c(b)$ can be calculated from b. Suppose the depth range is divided into N bins, the depth at pixel k can be formulated as:

$$d_k = \sum_{n=1}^{N} c(b_n) l_n. \tag{3}$$

4.4 Supervision

Following [1,24], we use a scaled version of the Scale-Invariant loss (SI) introduced by [10]:

$$\mathcal{L} = \alpha \sqrt{\frac{1}{T}\sum_i g_i^2 - \frac{\lambda}{T^2}(\sum_i g_i)^2}, \tag{4}$$

where $g_i = \log \tilde{d}_i - \log d_i$ defined by the estimated depth $\tilde{d}_i$ and ground truth depth $\log d_i$, and T denotes the number of pixels with valid ground truth values. We use $\lambda = 0.85$ and $\alpha = 10$ for all our experiments.

5 Experiment

5.1 Datasets and Evaluation Metrics

NYU-Depth V2 for Training. We use the NYU-Depth V2 dataset to simulate and generate the training data containing L5 signals and color images, from which we select a 24K subset following [24,1]. For each image, we select a set of zones and according to the L5 sensing principle, we count the histograms of the ground true depth map in each zone and fit them with Gaussian distributions. The fitted mean and variance are used together with the color images as the input for network training. We exclude the depths beyond the L5 measurement range during the histogram statistics.

ZJU-L5 dataset for Testing. Since the current datasets do not contain the L5 signals, we create an indoor depth dataset using the device suit in Fig. 3-(a) to evaluate our method. This dataset contains 1027 L5-image pairs from 15 scenes, of which the test set contains 527 pairs and the other 500 pairs are used for fine-tuning network. We show the results after fine-tuning in the supplementary material.

Evaluation Metrics. We report the results in terms of standard metrics including thresholded accuracy (δ_i), mean absolute relative error (REL), root mean square error (RMSE) and average ($\log_{10}$) error. The detailed definitions of the metrics are provided in the supplementary material.

5.2 Comparison with State-of-the-Art

Since we are the first to utilize L5 signals and color images to predict depth, there is no existing method for a direct comparison. Therefore, we pick three types of existing methods and let them make use of information from L5 as fully as possible. The first method is monocular depth estimation, where we use the depth information of L5 to align the predicted depth globally. The second method is depth completion, where we assume that each zone's mean depth lies at the zone's centroid to construct a sparse depth map as the input. The third method is depth super-resolution, where we consider the L5 signals as an 8×8 low-resolution depth map, with each pixel (zone) corresponding to a region of the image. Since the state-of-the-art RGB-D method is sensitive to the sparsity of the input points, we re-trained these methods for a fair comparison.

Comparison with Monocular Depth Estimation						
Methods	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	REL$\downarrow$	RMSE$\downarrow$	$\log_{10} \downarrow$
VNL [52]	0.661	0.861	0.928	0.225	0.653	0.104
BTS [24]	0.739	0.914	0.964	0.174	0.523	0.079
AdaBins [1]	0.770	0.926	0.970	0.160	0.494	0.073
Ours	**0.853**	**0.941**	**0.972**	**0.123**	**0.436**	**0.051**
Comparison with Depth Completion						
Methods	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	REL$\downarrow$	RMSE$\downarrow$	$\log_{10} \downarrow$
PrDepth [47]	0.161	0.395	0.660	0.409	0.937	0.249
NLSPN [31]	0.583	0.784	0.892	0.345	0.653	0.120
PENet [16]	0.807	0.914	0.954	0.161	0.498	0.065
Ours	**0.853**	**0.941**	**0.972**	**0.123**	**0.436**	**0.051**
Comparison with Depth Super Resolution						
Methods	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	REL$\downarrow$	RMSE$\downarrow$	$\log_{10} \downarrow$
PnP-Depth [45]	0.805	0.904	0.948	0.144	0.560	0.068
PrDepth [47]	0.800	0.926	0.969	0.151	0.460	0.063
Ours	**0.853**	**0.941**	**0.972**	**0.123**	**0.436**	**0.051**

Table 1: Quantitative evaluation on the ZJU-L5 dataset. Our method outperforms all baselines for monocular depth estimation, depth completion, and depth super-resolution.

Table 1 summarizes the comparison between ours and these three types of methods. For all metrics, our method achieves the best performance among all methods, which indicates that our network customized with respect to the underlying physics of the L5 is effective in learning from the depth distribution.

Fig. 7 shows the qualitative comparison of our method with other solutions for our task [1,47,16]. Overall, our method produces the most accurate depth as reflected by the error map. The monocular estimation method [1] can produce sharp object boundaries, however tends to make mistakes for regions with ambiguous textures. Guided depth super-resolution [47] and completion [16] by design are easier to maintain plausible depth measurement, but the output depths are often overly blurry and lack geometry details. In contrast, our method learns to leverage the high-resolution color image and low-quality L5 readings, producing the most accurate depths that are rich of details.

5.3 Ablation Studies

To understand the impact of each model component on the final performance, we conduct a comprehensive ablation study by disabling each component respectively. The quantitative results are shown in Table 2. There is a reasonable drop in performance with each component disabled, while the full model works the best.

Learning Directly from the Mean/Variance. We implement two baseline methods which learn directly from the mean and variance of the depth distribution. For the first one, we change the input to a five-dimensional tensor, which consists of RGB, mean depth and depth variance, named “Five-channel Input”.

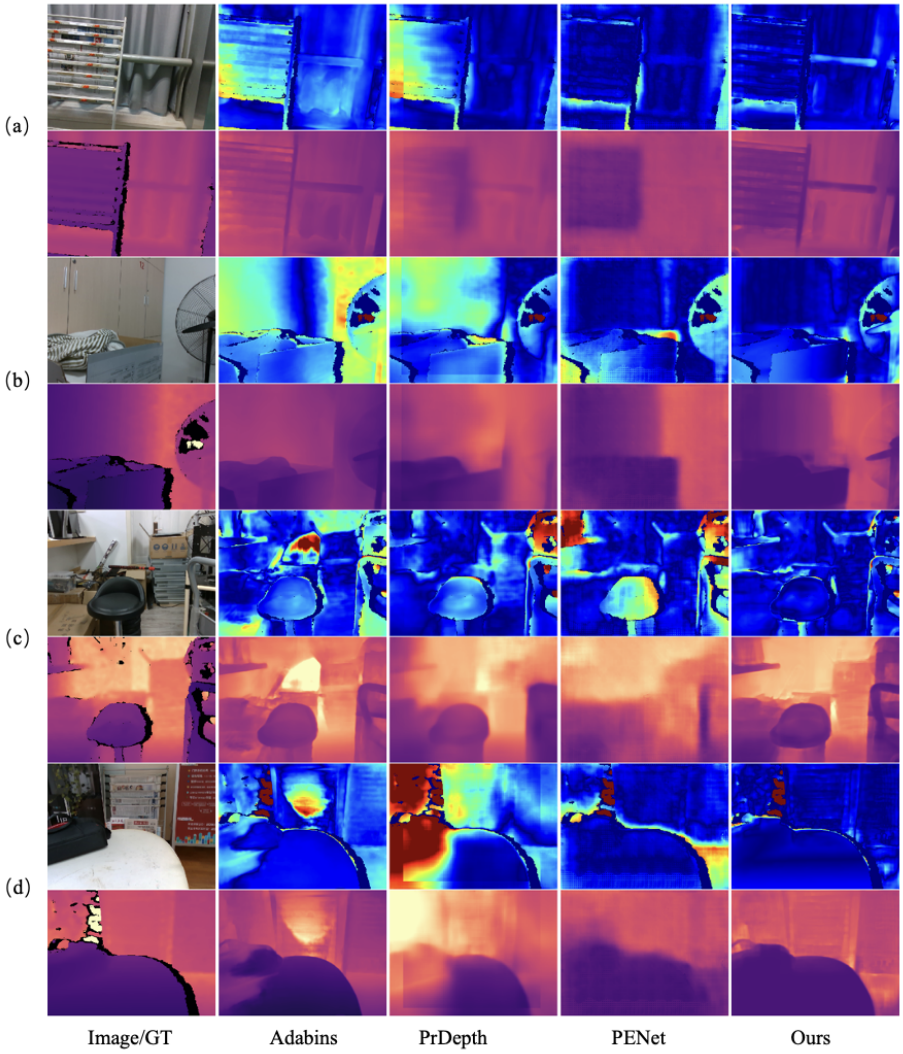


Fig. 7: Qualitative comparison on ZJU-L5 dataset with error map. Monocular estimation method [1] tends to make mistakes on some misleading textures. Guided depth super-resolution [47] and completion [16] produce overly blurry depths that are lack of geometry details. In contrast, our method learns to leverage the high resolution color image and low quality L5 reading, and produces the most accurate depths with sharp object boundaries.

For the second one, we extract features directly from the mean and variance instead of sampled depth, named "Mean-Var PointNet". The performance of these

Models	$\delta_1 \uparrow$	REL↓	RMSE↓
Mean-Var PointNet	0.434	0.298	0.669
Five-channel Input	0.619	0.251	0.583
Feature Concat	0.825	0.140	0.454
w/o Patch-Dist-Corr	0.749	0.182	0.512
w/o Img-Self-Attn	0.835	0.133	0.456
w/o Img-Dist-Attn	0.840	0.135	0.446
Uniform Sampling	0.849	0.127	0.439
w/o Refine	0.850	0.126	0.462
Full	**0.853**	**0.123**	**0.436**

Table 2: Ablation studies. We evaluate our method with each design or network component turned off. Overall, our full model achieves the best performance, which indicates the positive contribution from all design choices.

two baselines drops significantly compared to our full model, which indicates the effectiveness of our distribution feature extractor and the fusion module.

Compared with direct feature concatenation. We also replace our Transformer-based fusion module with direct feature concatenation (but retain our proposed feature extractor). It shows that our fusion module performs better than the direct concatenation, which benefits from the fact that our strategy can better gather features from totally different modalities and boost the overall accuracy by propagating features in a global receptive field.

Cross-attention without Considering Patch-distribution Correspondence. We re-train a model by relaxing the constrain on cross-attention, name "w/o Patch-Dist-Corr". Specifically, we conduct cross-attention between all distribution features and image features without considering patch-distribution correspondence. The performance degradation shows the importance of considering this correspondence.

Impact of multiple attention mechanisms. We train models without image self-attention ("w/o Img-Self-Attn") or image-to-distribution attention ("w/o Img-Dist-Attn") respectively that are proposed in Section. 4.2. The performance drop indicates that the attention modules positively contribute to our fusion model.

Impact of probability-driven sampling. We compare our methods trained with uniform sampling and probability-related sampling. The experiment indicates it brings an improvement of 0.3cm in terms of RMSE by considering the distribution probability. We report the impact of sampling points' number in the supplementary.

Impact of Refinement Module We also study the impact of the refinement module by replacing it with a simpler decoder consisting of two convolutional layers that output bin-widths vector and linear coefficient respectively. The overall performance drops but not much, which indicates, though the refiner helps, the majority improvements are brought by our distribution feature extractor and the fusion network.

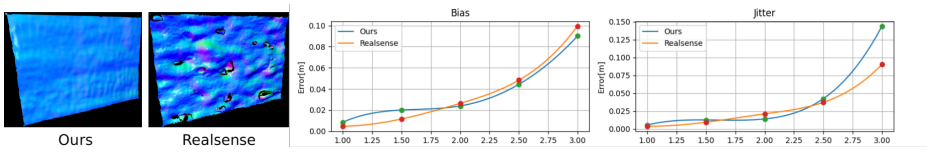


Fig. 8: Quantitative comparison with RealSense. Our method improves the raw L5 reading to a quality on par with commodity-level RGB-D sensor as reflected by the bias and jitter.

5.4 Quantitative comparison with RealSense

In this section, we compare our methods with RealSense quantitatively using traditional metrics in the area of stereo matching [54], such as jitter and bias. Specifically, we recorded multiple frames with the device in front of a flat wall at distances ranging from 1000 mm to 3000 mm. In this case, we evaluate by comparing to "ground truth" obtained with robust plane fitting. We compute bias as the average L1 error between the predicted depth and the ground truth plane to characterize the precision and compute the jitter as the standard deviation of the depth error to characterize the noise. Fig. 8 shows the comparison between our method and RealSense, together with visualizations of point clouds colored by surface normals. It can be seen that at a close range (less than three meters), our method achieves a similar and even better performance than RealSense. But as it approaches the upper range limit of L5, the jitter of our method increases dramatically. Overall, it indicates that our method improves the raw depth readings of the L5 to a quality (both resolution and accuracy) on par with a commodity-level depth sensor (i.e., the Intel RealSense D435i) in the working range of L5.

6 Conclusion and Future Work

In this work, we show that it is feasible to estimate high-quality depth, on par with commodity-level RGB-D sensors, using a color image and low-quality depth from a light-weight ToF depth sensor. The task is non-trivial due to the extremely low resolution and specific measurements of depth distribution, thus requiring a customized model to effectively extract features from depth distribution and fuse them with RGB image. One limitation of our method is that it is not fast enough for real-time performance. It is promising to further optimize the network complexity such that the system can run without much extra cost of energy consumption, or to further extend the system for more applications such as 3D reconstruction or SLAM.

Acknowledgment. This work was partially supported by Zhejiang Lab (2021PE 0AC01) and NSF of China (No. 62102356).

References

1. Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4009–4018 (2021) 4, 9, 10, 11, 12
2. Bleyer, M., Rhemann, C., Rother, C.: Patchmatch stereo-stereo matching with slanted support windows. In: British Machine Vision Conference. vol. 11, pp. 1–11 (2011) 7
3. Chen, Y., Yang, B., Liang, M., Urtasun, R.: Learning joint 2d-3d representations for depth completion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10023–10032 (2019) 4
4. Cheng, X., Wang, P., Guan, C., Yang, R.: Cspn++: Learning context and resource aware convolutional spatial propagation networks for depth completion. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 10615–10622 (2020) 4
5. Cheng, X., Wang, P., Yang, R.: CSPN: Depth Estimation via Affinity Learned with Convolutional Spatial Propagation Network. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) Computer Vision – ECCV 2018, vol. 11220, pp. 108–125. Springer International Publishing (2018) 4
6. Dai, A., Nießner, M., Zollhöfer, M., Izadi, S., Theobalt, C.: Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. ACM Transactions on Graphics (ToG) **36**(4), 1 (2017) 1, 2
7. Diebel, J., Thrun, S.: An application of markov random fields to range sensing. Advances in Neural Information Processing Systems **18** (2005) 4
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020) 8
9. Eigen, D., Puhrsch, C., Fergus, R.: Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. In: Advances in Neural Information Processing Systems. vol. 27. Curran Associates, Inc. (2014) 3
10. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in Neural Information Processing Systems **27** (2014) 10
11. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2002–2011 (2018) 4
12. Geiger, A., Moosmann, F., Car, Ö., Schuster, B.: Automatic camera and range sensor calibration using a single shot. In: 2012 IEEE International Conference on Robotics and Automation. pp. 3936–3943. IEEE (2012) 6
13. Hao, Z., Li, Y., You, S., Lu, F.: Detail preserving depth estimation from a single image using attention guided networks. In: 2018 International Conference on 3D Vision (3DV). pp. 304–313. IEEE (2018) 3
14. Hirschmuller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Transactions on Pattern Analysis and Machine Intelligence **30**(2), 328–341 (2007) 7
15. Hu, J., Ozay, M., Zhang, Y., Okatani, T.: Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1043–1051. IEEE (2019) 3

16. Hu, M., Wang, S., Li, B., Ning, S., Fan, L., Gong, X.: Penet: Towards precise and efficient image guided depth completion. In: 2021 IEEE International Conference on Robotics and Automation (ICRA). pp. 13656–13662. IEEE (2021) 4, 8, 11, 12
17. Huang, Z., Shi, X., Zhang, C., Wang, Q., Cheung, K.C., Qin, H., Dai, J., Li, H.: Flowformer: A transformer architecture for optical flow. arXiv preprint arXiv:2203.16194 (2022) 8
18. Hui, T.W., Loy, C.C., Tang, X.: Depth map super-resolution by deep multi-scale guidance. In: European Conference on Computer Vision. pp. 353–369. Springer (2016) 4
19. Huynh, L., Nguyen-Ha, P., Matas, J., Rahtu, E., Heikkilä, J.: DAV: Guiding monocular depth estimation using depth-attention volume. In: European Conference on Computer Vision. pp. 581–597. Springer (2020) 3
20. Iconfinder: Iconfinder. https://www.iconfinder.com/ (2022), Accessed 19-Jul-2022 2
21. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. pp. 448–456. PMLR (2015) 7
22. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A., et al.: Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In: Proceedings of the 24th annual ACM symposium on User interface software and technology. pp. 559–568 (2011) 1, 2
23. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 239–248 (2016). https://doi.org/10.1109/3DV.2016.32 3
24. Lee, J.H., Han, M.K., Ko, D.W., Suh, I.H.: From big to small: Multi-scale local planar guidance for monocular depth estimation. arXiv preprint arXiv:1907.10326 (2019) 3, 10, 11
25. Lin, T., Wang, Y., Liu, X., Qiu, X.: A survey of transformers. arXiv preprint arXiv:2106.04554 (2021) 8
26. Liu, M.Y., Tuzel, O., Taguchi, Y.: Joint geodesic upsampling of depth images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 169–176 (2013) 4
27. Liu, S., De Mello, S., Gu, J., Zhong, G., Yang, M.H., Kautz, J.: SPN: Learning Affinity via Spatial Propagation Networks. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017) 4
28. Mur-Artal, R., Tardós, J.D.: Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. IEEE Transactions on Robotics **33**(5), 1255–1262 (2017) 6
29. Newcombe, R.A., Fox, D., Seitz, S.M.: Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 343–352 (2015) 1
30. Park, J., Kim, H., Tai, Y.W., Brown, M.S., Kweon, I.: High quality depth map upsampling for 3d-tof cameras. In: 2011 International Conference on Computer Vision. pp. 1623–1630. IEEE (2011) 4
31. Park, J., Joo, K., Hu, Z., Liu, C.K., So Kweon, I.: Non-local spatial propagation network for depth completion. In: European Conference on Computer Vision. pp. 120–136. Springer (2020) 4, 11

32. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 652–660 (2017) 7
33. Qian, C., Sun, X., Wei, Y., Tang, X., Sun, J.: Realtime and robust hand tracking from depth. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1106–1113 (2014) 1
34. Qiu, J., Cui, Z., Zhang, Y., Zhang, X., Liu, S., Zeng, B., Pollefeys, M.: Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3313–3322 (2019) 4, 8
35. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12179–12188 (2021) 3
36. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence (2020) 3
37. Ranftl, R., Vineet, V., Chen, Q., Koltun, V.: Dense monocular depth estimation in complex dynamic scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4058–4066 (2016) 3
38. Saxena, A., Chung, S., Ng, A.: Learning Depth from Single Monocular Images. In: Advances in Neural Information Processing Systems. vol. 18. MIT Press (2005) 3
39. Saxena, A., Sun, M., Ng, A.Y.: Make3d: Learning 3d scene structure from a single still image. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(5), 824–840 (2008) 3
40. Shi, J., Tao, X., Xu, L., Jia, J.: Break ames room illusion: depth from general single images. ACM Transactions on Graphics (TOG) **34**(6), 1–11 (2015) 3
41. STMicroelectronics: STMicroelectronics Ships 1 Billionth Time-of-Flight Module. https://www.st.com/content/st_com/en/about/media-center/press-item.html/t4210.html, Accessed 19-Jul-2022 2
42. STMicroelectronics: Time-of-Flight 8x8 multizone ranging sensor with wide field of view. https://www.st.com/en/imaging-and-photonics-solutions/vl53l5cx.html, Accessed 19-Jul-2022 2
43. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning. pp. 6105–6114. PMLR (2019) 7
44. Tang, J., Tian, F.P., Feng, W., Li, J., Tan, P.: Learning guided convolutional network for depth completion. IEEE Transactions on Image Processing **30**, 1116–1129 (2020) 4
45. Wang, T.H., Wang, F.E., Lin, J.T., Tsai, Y.H., Chiu, W.C., Sun, M.: Plug-and-play: Improve depth prediction via sparse data propagation. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 5880–5886. IEEE (2019) 4, 11
46. Wang, Z., Ye, X., Sun, B., Yang, J., Xu, R., Li, H.: 40 Depth upsampling based on deep edge-aware learning. Pattern Recognition **103**, 107274 (2020) 4
47. Xia, Z., Sullivan, P., Chakrabarti, A.: Generating and exploiting probabilistic monocular depth estimates. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 65–74 (2020) 4, 11, 12
48. Xu, D., Ricci, E., Ouyang, W., Wang, X., Sebe, N.: Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5354–5362 (2017) 3

49. Xu, D., Wang, W., Tang, H., Liu, H., Sebe, N., Ricci, E.: Structured attention guided convolutional neural fields for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3917–3925 (2018) 3
50. Xu, Y., Zhu, X., Shi, J., Zhang, G., Bao, H., Li, H.: Depth completion from sparse lidar data with depth-normal constraints. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2811–2820 (2019) 4
51. Yang, Q., Yang, R., Davis, J., Nistér, D.: Spatial-depth super resolution for range images. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE (2007) 4
52. Yin, W., Liu, Y., Shen, C.: Virtual normal: Enforcing geometric constraints for accurate and robust depth prediction. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021) 11
53. Zhang, Y., Funkhouser, T.: Deep depth completion of a single rgb-d image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 175–185 (2018) 1
54. Zhang, Y., Khamis, S., Rhemann, C., Valentin, J., Kowdle, A., Tankovich, V., Schoenberg, M., Izadi, S., Funkhouser, T., Fanello, S.: Activestereonet: End-to-end self-supervised learning for active stereo systems. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 784–801 (2018) 14