Best of Both Sides: Integration of Absolute and Relative Depth Sensing Modalities Based on iToF and RGB Cameras

I-Sheng Fang[1][0009−0001−8347−5938]⋆, Wei-Chen Chiu[2][0000−0001−7715−8306], and Yong-Sheng Chen[2][0000−0002−5581−850X]

[1] Research Center for Information Technology Innovation, Academia Sinica, Taiwan
[2] National Yang Ming Chiao Tung University, Taiwan

Abstract. LiDAR sensors have become one of the most popular active depth sensing devices nowadays with their wide applications in autonomous driving and robotics. Among various types of LiDARs, indirect time of flight (iToF) has been ubiquitously applied on smartphones and consumer-level imagining devices due to its affordable price. Based on the common camera configuration on nowadays smartphones of having an iToF sensor and multiple RGB cameras with different focal lengths (thus leading to different fields of view), in this work, we investigate the integration between two opposite but complementary sensing modalities to achieve better depth estimation: 1) The active sensing modality based on iToF provides absolute and metric depths but suffers from noises caused by environmental lighting and heat; 2) The passive sensing modality based on monocular RGB cameras produces high-resolution but relative depth estimation. Our proposed integration is built upon a weakly-supervised learning framework where the learning objective mainly stems from the inter-camera geometric consistency with the help of iToF depth estimates. Moreover, we adopt the structure distillation technique for preserving structure details from the passive sensing method. We conduct experiments on both synthetic and real-world datasets and demonstrate that the depth estimation produced by the proposed integration model has a comparable quantitative performance with respect to the supervised learning baselines. Besides, the qualitative evaluation of our model shows that it utilizes the advantages and further overcomes the limitations of both sensing modalities.

Keywords: multiple view geometry · multi-modal and multi-view learning · stereo and 3D vision.

1 Introduction

Depth estimation is an essential task in computer vision. Among various depth sensors, RGB-D camera modules attract attention because of their capability of multimodal perception from the environment, providing the depth and the RGB images simultaneously. For the RGB-D camera module of consumer-level mobile phones, time-of-flight (ToF)

⋆ Work done at NYCU as graduate student.

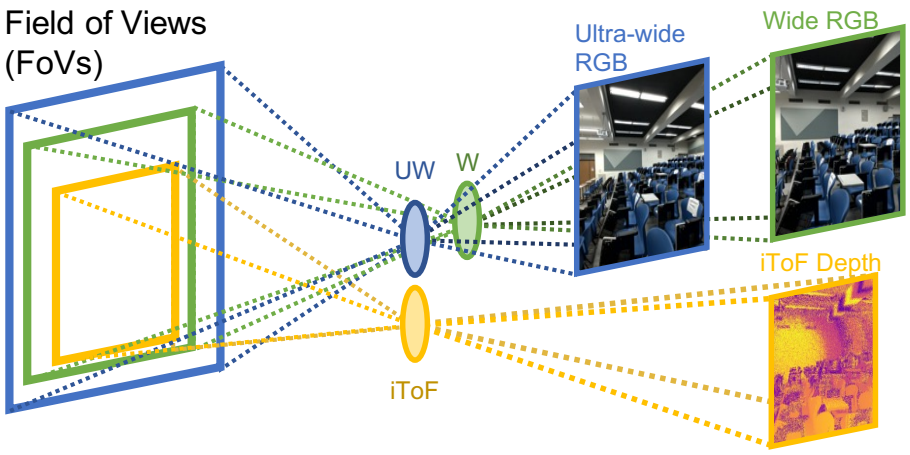

Fig. 1. Illustration of our RGB-D camera module with specific emphasis on the differences in terms of focal lengths and fields of view (FoV). Compared with the RGB cameras, the iToF camera typically has smaller resolution and longer focal length, leading to the narrower FoV.

cameras are the more affordable solution. As shown in Figure 1, the camera module used in this study comprises an indirect time-of-flight (iToF) depth camera, an ultra-wide-angle RGB camera, and a wide-angle RGB camera. Our objective is to obtain accurate metric depth with the same field of view (FoV) as that of the RGB image.

As shown in Figure 2, we have the active sensing depths measured by the iToF camera and the passive sensing depths estimated from the RGB image by the off-the-shelf vision-based monocular depth estimation model [19]. The iToF depth camera measures the phase shift between the emitted and reflected infrared light [10] for depth calculation. As a result, the depth measured by iToF is accurate in short range and has metric (absolute) values. However, its resolution and FoV are relatively lower than those of the depth maps estimated from RGB images. As shown in the right column of Figure 2, the iToF depths warped onto the RGB image plane have a large invalid part with void values (yellow region with depth value 0). Moreover, the iToF depths suffer from different types of noises and errors, such as multi-path interference errors, periodic noises, and low reflection of the infrared signal, causing inaccurate warping results. On the other hand, vision-based monocular depth estimation models [31,23] have shown impressive performance in the depth estimation with high-resolution results [19]. These models benefit from the variety of large datasets and the learned depth cues of objects, such as edges and vanishing points [12]. However, the obtained depths are relative values and may suffer from incorrect depth cues due to the domain gap. In short, the active and passive depth sensing modalities are complementary to each other and their integration stands a good chance in the combination of advantages from both sides. Our goal is to obtain a metric depth map with high resolution and less noise by utilizing both iToF depths and RGB images.

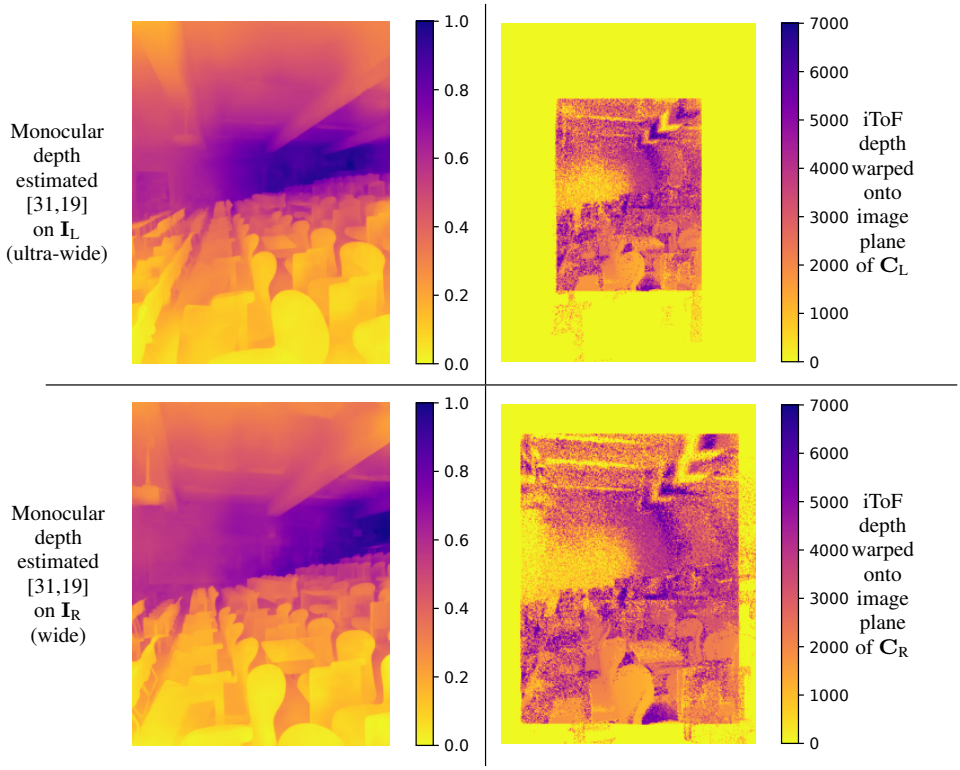

Fig. 2. An example set of the monocular depth estimation [31,19] on $\mathbf{I}_{\mathrm{L}}$ and $\mathbf{I}_{\mathrm{R}}$, with the corresponding iToF depth maps $\{\mathbf{D}^{\mathrm{L}}_{\mathrm{iToF}}, \mathbf{D}^{\mathrm{R}}_{\mathrm{iToF}}\}$ being warped onto the image planes of their respective cameras (*i.e.* $\mathbf{C}_{\mathrm{L}}$ with ultra-wide-angle lens and $\mathbf{C}_{\mathrm{R}}$ with wide-angle lens). Notice that $\{\mathbf{D}^{\mathrm{L}}_{\mathrm{iToF}}, \mathbf{D}^{\mathrm{R}}_{\mathrm{iToF}}\}$ stemmed from iToF sensor have metric depth values with smaller FoVs and contain more noises, whereas the depth maps computed by the off-the-shelf monocular depth estimation model [31,19] have higher resolutions but only relative depth values.

The straightforward idea for cross-modal depth integration is to utilize the confidence map of the metric depth, filter out the unreliable depth measurements, and train the model with supervised learning as a depth completion task. However, our iToF depth camera lacks the information for uncertainty, making it difficult to expose the confident regions in the iToF depth map. Moreover, it is difficult to reduce the influence of noise using RANSAC [30] because of the large amounts of noises in iToF depth map. Therefore, iToF depths cannot be used as ground truths for supervised learning. Furthermore, although the structured light [28] could obtain the ground-truth depths, it is labor-intensive and sensitive to noise. Another way for supervised learning is to adopt synthetic data [21]. Unfortunately, the problem of the domain gap between the real and synthetic images is difficult to overcome. As shown in Figure 3, iToF depths taken by our device have high-frequency and periodic noises, which are not typical in

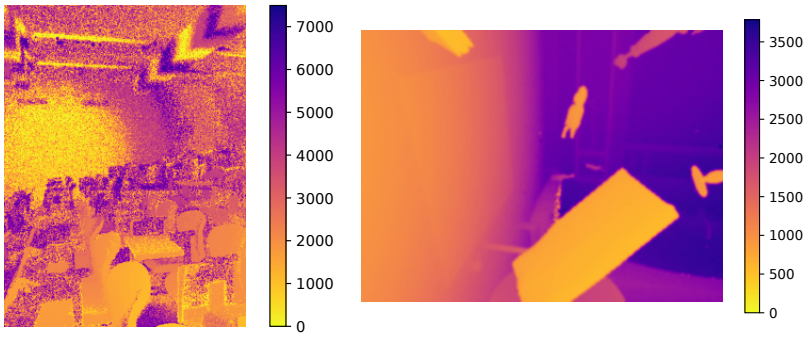

Fig. 3. Comparison on the iToF depth maps between our collected iToF-RGB1k dataset and the synthetic ToF-FlyingThings3D dataset [21]. The noises of iToF depth maps in iToF-RGB1k are high-frequency and periodic while those in ToF-FlyingThings3D are low-frequency and non-periodic, leaving a large domain gap to deploy the model trained on synthetic data.

the synthetic dataset ToF-FlyingThings3D [21], causing the issue of deployment in the real world.

To tackle these challenges, we propose a cross-modal depth estimation model to integrate passive sensing RGB and active sensing iToF images as well as its weakly supervised learning method. Instead of direct supervision with ground-truth depths, the training of our model is self-supervised with the consistency of multi-view geometry by computing the similarity between the captured RGB image and the warped one according to the estimated depths. Moreover, our model leverages the off-the-shelf monocular depth estimation model to extend the original limited FoV of iToF and distill the knowledge of depth structure.

In summary, contributions of this work include:

1. We propose a cross-modal depth estimation model and its weakly-supervised learning framework containing the **cross-warp consistency** and **depth structure distillation**. This model integrates the active iToF depths with the passive RGB image to obtain the metric depth map having the same FoV as the RGB image.
2. We collect the real-world dataset iToF-RGB1k with 1074 sets of triplet data for the training and testing of the cross-modal depth estimation model. Each triplet contains an ultra-wide RGB image, a wide RGB image, and an iToF depth map.
3. Quantitative evaluation using the synthetic dataset ToF-FlyThings3D[21] as training data shows that our model gains competitive results compared with other supervised learning methods, even though our model is a weakly supervised learning method. Our model also qualitatively performs well when trained and tested on real-world dataset iToF-RGB1k.

2 Related Works

2.1 Depth Completion

The objective of depth completion is to estimate a dense and accurate depth map from a sparse or incomplete one by recovering missing or invalid depth values. Ma *et al*. [17] propose the Sparse-to-Dense method to predict the dense depth map from a sparse set of depth measurements and a single RGB image. In their following work, Ma *et al*. [16] further improve Sparse-to-Dense by utilizing photometric consistency and camera poses calculated by PnP with RANSAC. Wong *et al*. [29] and Choi *et al*. [3] utilize temporal photometric consistency with pose estimation network and L_1 loss. DFuseNet [26] utilizes stereo photometric consistency in depth completion task. While these approaches fill in missing depth values based on confident measurements, our method extends the FoV of depth images without confidence filtering. Moreover, our model tackles the problem of large FoV differences among three cameras without additional pose estimation or stereo image rectification.

2.2 iToF Depth Refinement and Cross-modal Depth Estimation

Because of the success of deep learning [14] in various machine learning tasks, many network models have been proposed to refine iToF depth, requiring synthetic data for supervised learning [18,27,9,5]. As another modality, RGB has been used for iToF refinement or depth estimation with supervised learning for model training [21,13]. CroMo method [28] utilizes geometric consistency for self-supervised learning from the cross-modal dataset with iToF and stereo polarization images. Instead of depth data used in our method, CroMo uses iToF correlation images. Moreover, their stereo RGB cameras are with the same focal length, but ours are different. Furthermore, our method distills the knowledge of depth structure from other off-the-shelf monocular depth estimation models.

2.3 Monocular Depth Estimation and Knowledge Distillation

Monocular depth estimation models use the visual depth cue to estimate the spatial relationship between objects [12] from a single image. Godard *et al*. [7] introduce a self-supervised-learning method with left-right consistency. Recent supervised-learning works, such as MiDaS [23], DPT [22], and LeReS [31], leverage neural networks with advanced model structures and large diverse datasets. Miangoleh *et al*. [19] discover the trade-off between scene structure and high-frequency details and mix the estimated depths with low and high resolutions to boost the performance of the off-the-shelf model. Inspired by knowledge distillation, DistDepth [30] distills depth-domain structure knowledge from the off-the-shelf model into its monocular depth estimator. In contrast with our method which integrates RGB and iToF modalities to estimate absolute depths, these works use single modality (RGB) and most of them estimate relative depths.

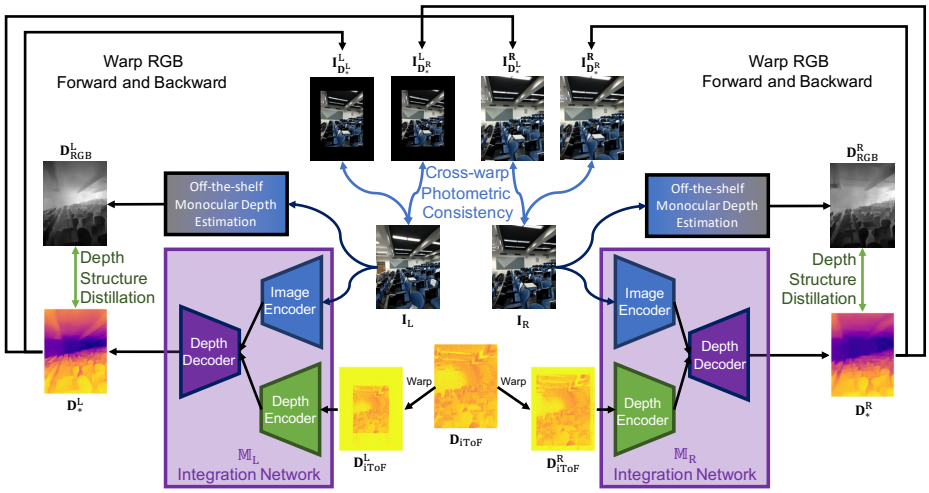

Fig. 4. Computational flow of our framework of integrating active and passive depth sensing modalities (*i.e.* iToF sensor and RGB cameras respectively). We warp the iToF depth map onto the image planes of the RGB cameras for rough alignment. Then, we input the RGB image and the warped iToF depth map into the integration network to integrate modalities and to estimate metric depth on the perspective of the RGB image. Integration Network is weakly supervised by cross-warp consistency and depth structure distillation. See Sec.3 for details.

3 Methods

3.1 Problem Statement

Our cross-modal integration scenario for depth estimation is built upon a RGB-D camera module composed of:

1. Left RGB camera $\mathbf{C}_{\mathrm{L}}$ with an ultra-wide-angle lens, where the image captured by $\mathbf{C}_{\mathrm{L}}$ is denoted as $\mathbf{I}_{\mathrm{L}}$;
2. Right RGB camera $\mathbf{C}_{\mathrm{R}}$ with a wide-angle lens, where the image captured by $\mathbf{C}_{\mathrm{R}}$ is denoted as $\mathbf{I}_{\mathrm{R}}$;
3. iToF depth camera $\mathbf{C}_{\mathrm{iToF}}$, in which $\mathbf{C}_{\mathrm{iToF}}$ produces the iToF depth map $\mathbf{D}_{\mathrm{iToF}}$.

In this RGB-D camera module, the FoV of $\mathbf{I}_{\mathrm{L}}$ is larger than $\mathbf{I}_{\mathrm{R}}$, and the iToF camera is typically with the smallest FoV. Without loss of generality, we assume that the camera with ultra-wide-angle lens is placed on the left of the one with wide-angle lens. The objective of our cross-modal integration is to acquire the depth maps $\mathbf{D}_{\mathrm{L}}$ and $\mathbf{D}_{\mathrm{R}}$ respectively for both $\mathbf{C}_{\mathrm{L}}$ and $\mathbf{C}_{\mathrm{R}}$, with well taking the complementary properties between $\{\mathbf{C}_{\mathrm{L}}, \mathbf{C}_{\mathrm{R}}\}$ and $\mathbf{C}_{\mathrm{iToF}}$ to achieve the better depth perception results. The depth map $\mathbf{D}_{\mathrm{L}}$ is expected to consist of absolute-metric and less-noisy depths from the same perspective of the left RGB camera $\mathbf{C}_{\mathrm{L}}$.

3.2 Camera Calibration

Prior to realizing cross-modal integration of the RGB-D camera module, we calibrate all the cameras to get their geometric characteristics (*i.e.* intrinsic parameters $\mathbf{K}_{\mathrm{L}}$, $\mathbf{K}_{\mathrm{R}}$, and $\mathbf{K}_{\mathrm{iToF}}$ for $\mathbf{C}_{\mathrm{L}}$, $\mathbf{C}_{\mathrm{R}}$, and $\mathbf{C}_{\mathrm{iToF}}$ respectively) as well as their geometric relationship (*i.e.* extrinsic parameters $\mathbf{T}_{\mathrm{iToF}\to\mathrm{L}}$, $\mathbf{T}_{\mathrm{L}\to\mathrm{iToF}}$, $\mathbf{T}_{\mathrm{iToF}\to\mathrm{R}}$, $\mathbf{T}_{\mathrm{R}\to\mathrm{iTof}}$, $\mathbf{T}_{\mathrm{L}\to\mathrm{R}}$, and $\mathbf{T}_{\mathrm{R}\to\mathrm{L}}$ between cameras, where $\mathbf{T}_{\mathrm{iToF}\to\mathrm{L}}$ denotes the transformation from $\mathbf{C}_{\mathrm{iToF}}$ to $\mathbf{C}_{\mathrm{L}}$ and the others are defined analogously). We adopt the calibration toolkit of OpenCV [2] and a 7×9 metric chessboard pattern to conduct calibration, where the intensity maps of RGB images $\{\mathbf{I}_{\mathrm{L}}, \mathbf{I}_{\mathrm{R}}\}$ and the infrared amplitude map of the iToF camera $\mathbf{C}_{\mathrm{iToF}}$ are taken as inputs.

3.3 Warping iToF Depths and RGB Images

With the extrinsic and intrinsic parameters among RGB and iToF cameras, the **warping grid** for building the pixel-wise correspondence across their image planes now becomes available, under the following computation procedure:

Given the intrinsic parameters $\{\mathbf{K}_{\mathrm{A}}, \mathbf{K}_{\mathrm{B}}\}$ of two cameras $\{\mathbf{C}_{\mathrm{A}}, \mathbf{C}_{\mathrm{B}}\}$, the extrinsic transformation $\mathbf{T}_{\mathrm{A}\to\mathrm{B}}$ between them, and the depth map D_{A} related to the image plane of $\mathbf{C}_{\mathrm{A}}$, the corresponding pixel p_{B} on the image plane of $\mathbf{C}_{\mathrm{B}}$ for a specific pixel p_{A} on the image plane of $\mathbf{C}_{\mathrm{A}}$ is computed by

$$\mathrm{p}_{\mathrm{B}} = \mathbf{K}_{\mathrm{B}}\mathbf{T}_{\mathrm{A}\to\mathrm{B}} z_{\mathrm{p}_{\mathrm{A}}} \mathbf{K}_{\mathrm{A}}^{-1} \mathrm{p}_{\mathrm{A}} \tag{1}$$

where $z_{\mathrm{p}_{\mathrm{A}}} = \mathrm{D}_{\mathrm{A}}(\mathrm{p}_{\mathrm{A}})$. Based on computing the corresponding pixels across cameras, the forward warping grid from camera $\mathbf{C}_{\mathrm{A}}$ to camera $\mathbf{C}_{\mathrm{B}}$ with the help of depth map $\mathbf{D}_{\mathrm{A}}$ is denoted as $\langle \mathtt{proj}(\mathbf{D}_{\mathrm{A}}, \mathbf{T}_{\mathrm{A}\to\mathrm{B}}, \mathbf{K}_{\mathrm{A}}, \mathbf{K}_{\mathrm{B}})\rangle$, indicating how the pixels on camera $\mathbf{C}_{\mathrm{A}}$'s image plane should move in order to be aligned with the content on the image plane of camera $\mathbf{C}_{\mathrm{B}}$ (following the similar notations as DistDepth [30]). Moreover, we denote the backward warping grid from camera $\mathbf{C}_{\mathrm{B}}$ to camera $\mathbf{C}_{\mathrm{A}}$ (*i.e.* the inverse mapping with respect to the forward warping grid) as $\langle \mathtt{proj}(\mathbf{D}_{\mathrm{A}}, \mathbf{T}_{\mathrm{A}\to\mathrm{B}}, \mathbf{K}_{\mathrm{A}}, \mathbf{K}_{\mathrm{B}})\rangle^{-1}$.

Based on such technique of warping grid, if we treat the iToF depth map $\mathbf{D}_{\mathrm{iToF}}$ itself as a grayscale image on the image plane of iToF camera $\mathbf{C}_{\mathrm{iToF}}$, we then are able to warp it onto the image planes of $\{\mathbf{C}_{\mathrm{L}}, \mathbf{C}_{\mathrm{R}}\}$ thus obtaining $\mathbf{D}^{\mathrm{L}}_{\mathrm{iToF}}$ and $\mathbf{D}^{\mathrm{R}}_{\mathrm{iToF}}$ respectively:

$$\mathbf{D}^{\mathrm{L}}_{\mathrm{iToF}} = \mathbf{D}_{\mathrm{iToF}} \langle \mathtt{proj}(\mathbf{D}_{\mathrm{iToF}}, \mathbf{T}_{\mathrm{iToF}\to\mathrm{L}}, \mathbf{K}_{\mathrm{iToF}}, \mathbf{K}_{\mathrm{L}})\rangle \tag{2}$$

$$\mathbf{D}^{\mathrm{R}}_{\mathrm{iToF}} = \mathbf{D}_{\mathrm{iToF}} \langle \mathtt{proj}(\mathbf{D}_{\mathrm{iToF}}, \mathbf{T}_{\mathrm{iToF}\to\mathrm{R}}, \mathbf{K}_{\mathrm{iToF}}, \mathbf{K}_{\mathrm{R}})\rangle \tag{3}$$

in which $\{\mathbf{D}^{\mathrm{L}}_{\mathrm{iToF}}, \mathbf{D}^{\mathrm{R}}_{\mathrm{iToF}}\}$ seem to already provide the depth perception from the perspective of $\{\mathbf{C}_{\mathrm{L}}, \mathbf{C}_{\mathrm{R}}\}$. However, as iToF cameras typically have a longer focal length than the RGB ones thus leading to the narrower FoV, the warped depth maps (*i.e.* $\{\mathbf{D}^{\mathrm{L}}_{\mathrm{iToF}}, \mathbf{D}^{\mathrm{R}}_{\mathrm{iToF}}\}$) from iToF camera $\mathbf{C}_{\mathrm{iToF}}$ to RGB ones $\{\mathbf{C}_{\mathrm{L}}, \mathbf{C}_{\mathrm{R}}\}$ would unfortunately have large void regions. Moreover, the noise on iToF depth map caused by environmental lighting and heat would also lead to the incorrect warping results. Figure 2 shows the void region due to the difference in terms of focal length as well as the wrong warped results caused by iToF noise. Despite these limitations, the benefits of iToF depth, such as active sensing and metric/absolute value, should be preserved after the following integration of RGB and iToF cameras.

3.4 Off-the-shelf Monocular Depth Estimation

In addition to the warped iToF depth maps $\{\mathbf{D}^{\mathrm{L}}_{\mathrm{iToF}}, \mathbf{D}^{\mathrm{R}}_{\mathrm{iToF}}\}$, another plausible and popular way of acquiring depth upon the image planes of RGB cameras $\{\mathbf{C}_{\mathrm{L}}, \mathbf{C}_{\mathrm{R}}\}$ is to use the off-the-shelf monocular depth estimation model f, thanks to the recent development of (deep-)learning-based techniques. The depth maps $\{\mathbf{D}^{\mathrm{L}}_{\mathrm{RGB}} = f(\mathbf{I}_{\mathrm{L}}), \mathbf{D}^{\mathrm{R}}_{\mathrm{RGB}} = f(\mathbf{I}_{\mathrm{R}})\}$ contribute the largest FoV with respect to $\{\mathbf{C}_{\mathrm{L}}, \mathbf{C}_{\mathrm{R}}\}$ (as all the pixels of $\{\mathbf{I}_{\mathrm{L}}, \mathbf{I}_{\mathrm{R}}\}$ have their depth estimates produced by f, while $\{\mathbf{D}^{\mathrm{L}}_{\mathrm{iToF}}, \mathbf{D}^{\mathrm{R}}_{\mathrm{iToF}}\}$ have quite some void regions) but only produce relative depth perception.

3.5 Integration of RGB and iToF

Given both the active and passive depth sensing components (*i.e.* $\{\mathbf{D}^{\mathrm{L}}_{\mathrm{iToF}}, \mathbf{D}^{\mathrm{R}}_{\mathrm{iToF}}\}$ and $\{\mathbf{D}^{\mathrm{L}}_{\mathrm{RGB}}, \mathbf{D}^{\mathrm{R}}_{\mathrm{RGB}}\}$ respectively) upon the image planes of $\{\mathbf{C}_{\mathrm{L}}, \mathbf{C}_{\mathrm{R}}\}$, we now proceed to integrate them to produce better depth perception. Instead of directly taking $\mathbf{D}^{\mathrm{L}}_{\mathrm{iToF}}$ and $\mathbf{D}^{\mathrm{L}}_{\mathrm{RGB}}$ as input to the fusion model for producing the final depth estimation where their difference in terms of the depth-scale change would lead to problematic learning, we propose a novel integration framework based on the following learning scheme composed of three important aspects and shown in Figure 4. Please note that here we take $\mathbf{C}_{\mathrm{L}}$ as an example while $\mathbf{C}_{\mathrm{R}}$ follows the analogous process. 1) An integration network $\mathbb{M}$ (as indicated by the region shaded by light purple color in Figure 4) adopts the passive sensing RGB image $\mathbf{I}_{\mathrm{L}}$ for refining the active sensing depth component $\mathbf{D}^{\mathrm{L}}_{\mathrm{iToF}}$ to obtain the refined depth $\mathbf{D}^{\mathrm{L}}_{*}$. The basic idea behind it is leveraging the rich appearance and structure information of the RGB image to help denoising $\mathbf{D}^{\mathrm{L}}_{\mathrm{iToF}}$ as well as enlarging its FoV; 2) To address the lack of ground-truth depth for supervised learning the integration, we leverage the geometric relationship across two RGB cameras $\{\mathbf{C}_{\mathrm{L}}, \mathbf{C}_{\mathrm{R}}\}$ and build the photometric and depth consistency loss to realize the unsupervised learning of $\mathbf{D}^{\mathrm{L}}_{*}$; 3) We adopt the passive component $\mathbf{D}^{\mathrm{L}}_{\mathrm{RGB}}$ as structural guidance for $\mathbf{D}^{\mathrm{L}}_{*}$ during the training of the integration network $\mathbb{M}$. In other words, we distill the knowledge of depth structure from $\mathbf{D}^{\mathrm{L}}_{\mathrm{RGB}}$. These three important aspects in our framework are driven by two main objectives: **cross-warp consistency** and **depth structure distillation**, which we detailed sequentially in the following.

Cross-warp Consistency. As we tend to maximize the practical usage and the flexibility of our proposed framework, we do not require the training of $\mathbf{D}^{\mathrm{L}}_{*}$ to rely on the ground-truth labels. In other words, the learning of $\mathbf{D}^{\mathrm{L}}_{*}$ is not supervised. Instead, we are inspired by the unsupervised objective built upon the geometric relations between cameras and photometric reconstruction, as proposed by Godard *et al*. [7], where the accurate depth estimate of the left camera should enable the reconstruction of the right image by warping the left image via the geometric transformation between them. Following the similar idea, we introduce the ***cross-warp photometric consistency loss*** $L^{\mathbf{D}^{\mathrm{L}}_{*}}_{\mathrm{xwarp\text{-}I}}$ for the refined depth $\mathbf{D}^{\mathrm{L}}_{*}$:

$$
\begin{aligned}
L_{\text{xwarp-I}}^{\mathbf{D}_*^{\text{L}}} &= L_{\text{xwarp-I}}^{\mathbf{D}_*^{\text{L}}\text{-fwd}} + L_{\text{xwarp-I}}^{\mathbf{D}_*^{\text{L}}\text{-bwd}} \\
&= \mathbb{S}(\mathbf{I}_{\mathbf{D}_*^{\text{L}}}^{\text{R}}, \mathbf{I}_{\text{R}}) + \mathbb{S}(\mathbf{I}_{\mathbf{D}_*^{\text{L}}}^{\text{L}}, \mathbf{I}_{\text{L}}) \,, \\
\text{where } \mathbb{S}(a,b) &= \alpha \frac{1 - \texttt{SSIM}(a,b)}{2} + (1-\alpha)\,|a-b|_1 \\
\text{and } \quad \mathbf{I}_{\mathbf{D}_*^{\text{L}}}^{\text{R}} &= \mathbf{I}_{\text{L}} \left\langle \texttt{proj}(\mathbf{D}_*^{\text{L}}, \mathbf{T}_{\text{L}\to\text{R}}, \mathbf{K}_{\text{L}}, \mathbf{K}_{\text{R}}) \right\rangle \\
\mathbf{I}_{\mathbf{D}_*^{\text{L}}}^{\text{L}} &= \mathbf{I}_{\text{R}} \left\langle \texttt{proj}(\mathbf{D}_*^{\text{L}}, \mathbf{T}_{\text{L}\to\text{R}}, \mathbf{K}_{\text{L}}, \mathbf{K}_{\text{R}}) \right\rangle^{-1} .
\end{aligned} \tag{4}
$$

in which function $\mathbb{S}(a,b)$ evaluates the $\texttt{SSIM}$ structural distance as well as L_1 pixel errors between a and b (noting that we follow the common practice as [30] to set $\alpha = 0.85$). $\mathbf{I}_{\mathbf{D}_*^{\text{L}}}^{\text{R}}$ denotes the reconstructed right image, using $\mathbf{D}_*^{\text{L}}$ to perform the forward warping from $\mathbf{C}_{\text{L}}$ to $\mathbf{C}_{\text{R}}$; $\mathbf{I}_{\mathbf{D}_*^{\text{L}}}^{\text{L}}$ denotes the reconstructed left image, using $\mathbf{D}_*^{\text{L}}$ to perform the backward warping from $\mathbf{C}_{\text{R}}$ to $\mathbf{C}_{\text{L}}$. Noting that $L_{\text{xwarp-I}}^{\mathbf{D}_*^{\text{R}}}$ follows the similar procedure to evaluate $\mathbb{S}(\mathbf{I}_{\mathbf{D}_*^{\text{R}}}^{\text{L}}, \mathbf{I}_{\text{L}}) + \mathbb{S}(\mathbf{I}_{\mathbf{D}_*^{\text{R}}}^{\text{R}}, \mathbf{I}_{\text{R}})$.

In addition to the cross-warp photometric consistency loss $\{L_{\text{xwarp-I}}^{\text{L}}, L_{\text{xwarp-I}}^{\text{R}}\}$ for $\{\mathbf{D}_*^{\text{L}}, \mathbf{D}_*^{\text{R}}\}$, we also modify the well-known left-right depth consistency loss [7,8] into ***cross-warp depth consistency loss*** $L_{\text{xwarp-D}}$ for our training of integration network $\mathbb{M}$, making the warped depth map of right camera equal to the depth map of left camera and vice versa, regardless of forward or backward warping:

$$
\begin{aligned}
L_{\text{xwarp-D}} &= L_{\text{xwarp-D}}^{\mathbf{D}_*^{\text{L}}} + L_{\text{xwarp-D}}^{\mathbf{D}_*^{\text{R}}} \\
&= L_{\text{xwarp-D}}^{\mathbf{D}_*^{\text{L}}\text{-fwd}} + L_{\text{xwarp-D}}^{\mathbf{D}_*^{\text{L}}\text{-bwd}} + L_{\text{xwarp-D}}^{\mathbf{D}_*^{\text{R}}\text{-fwd}} + L_{\text{xwarp-D}}^{\mathbf{D}_*^{\text{R}}\text{-bwd}} \\
&= \left| \mathbf{D}_{\mathbf{D}_*^{\text{L}}}^{\text{R}} - \mathbf{D}_*^{\text{R}} \right|_1 + \left| \mathbf{D}_{\mathbf{D}_*^{\text{L}}}^{\text{L}} - \mathbf{D}_*^{\text{L}} \right|_1 \\
&\quad + \left| \mathbf{D}_{\mathbf{D}_*^{\text{R}}}^{\text{L}} - \mathbf{D}_*^{\text{L}} \right|_1 + \left| \mathbf{D}_{\mathbf{D}_*^{\text{R}}}^{\text{R}} - \mathbf{D}_*^{\text{R}} \right|_1 , \\
\text{where } \mathbf{D}_{\mathbf{D}_*^{\text{L}}}^{\text{R}} &= \mathbf{D}_*^{\text{L}} \left\langle \texttt{proj}(\mathbf{D}_*^{\text{L}}, \mathbf{T}_{\text{L}\to\text{R}}, \mathbf{K}_{\text{L}}, \mathbf{K}_{\text{R}}) \right\rangle , \\
\mathbf{D}_{\mathbf{D}_*^{\text{L}}}^{\text{L}} &= \mathbf{D}_*^{\text{R}} \left\langle \texttt{proj}(\mathbf{D}_*^{\text{L}}, \mathbf{T}_{\text{L}\to\text{R}}, \mathbf{K}_{\text{L}}, \mathbf{K}_{\text{R}}) \right\rangle^{-1} , \\
\mathbf{D}_{\mathbf{D}_*^{\text{R}}}^{\text{L}} &= \mathbf{D}_*^{\text{L}} \left\langle \texttt{proj}(\mathbf{D}_*^{\text{R}}, \mathbf{T}_{\text{R}\to\text{L}}, \mathbf{K}_{\text{R}}, \mathbf{K}_{\text{L}}) \right\rangle , \\
\mathbf{D}_{\mathbf{D}_*^{\text{R}}}^{\text{R}} &= \mathbf{D}_*^{\text{L}} \left\langle \texttt{proj}(\mathbf{D}_*^{\text{R}}, \mathbf{T}_{\text{R}\to\text{L}}, \mathbf{K}_{\text{R}}, \mathbf{K}_{\text{L}}) \right\rangle^{-1} .
\end{aligned} \tag{5}
$$

Depth Structure Distillation. As motivated previously that our third aspect is to adopt the passive component (*e.g.* $\mathbf{D}_{\text{RGB}}^{\text{L}}$) as a structural guidance for the output of our integration model, we choose to adapt the ***structure distillation loss*** proposed by [30] into our framework for realizing such aspect, which is defined as

$$
\begin{aligned}
L_{\text{distill}} = \;& L_{\text{distill}}^{\mathbf{D}_*^{\text{L}}} + L_{\text{distill}}^{\mathbf{D}_*^{\text{R}}} \\
=\;& 1 - \texttt{SSIM}(\bar{\mathbf{D}}_*^{\mathbf{L}}, \bar{\mathbf{D}}_{\text{RGB}}^{\mathbf{L}}) \\
&+ 1 - \texttt{SSIM}(\bar{\mathbf{D}}_*^{\mathbf{R}}, \bar{\mathbf{D}}_{\text{RGB}}^{\mathbf{R}}) \,,
\end{aligned} \tag{6}
$$

where $\bar{\mathbf{D}}$ denotes the operation of normalizing depth $\mathbf{D}$ with respect to its own mean value. The depth structure distillation loss L_{distill} relies on the off-the-shelf pre-trained monocular depth estimation model f to provide the passive depth perception. Therefore, although our cross-warp consistency objective is self-supervised, we categorize our method as a weakly-supervised learning framework.

Smoothness Loss [7]: Lastly, similar to other self-supervised depth estimation methods [7,8], we also adopt the ***smoothness loss*** L_{sm} to regulate the estimated depth $\{\mathbf{D}^{\text{L}}_{*}, \mathbf{D}^{\text{R}}_{*}\}$ for making them locally smooth and edge-aware:

$$L_{\text{sm}} = \left|\partial \mathbf{D}^{\text{L}}_{*}\right| e^{-\|\partial \mathbf{I}_{\text{L}}\|} + \left|\partial \mathbf{D}^{\text{R}}_{*}\right| e^{-\|\partial \mathbf{I}_{\text{R}}\|}. \tag{7}$$

The derivative operation ∂ in L_{sm} includes both the horizontal and vertical gradients.

Total Loss. The overall objective is summarized as:

$$\begin{aligned} L_{\text{total}} = {} & \lambda_{\text{xwarp-I}} L^{\mathbf{D}^{\text{L}}_{*}}_{\text{xwarp-I}} + \lambda_{\text{xwarp-I}} L^{\mathbf{D}^{\text{R}}_{*}}_{\text{xwarp-I}} \\ & + \lambda_{\text{xwarp-D}} L_{\text{xwarp-D}} + \lambda_{\text{distill}} L_{\text{distill}} + \lambda_{\text{sm}} L_{\text{sm}}\,, \end{aligned} \tag{8}$$

where λ hyper-parameters are the weights to balance among the aforementioned losses.

3.6 Integration Network $\mathbb{M}$

Our integration network $\mathbb{M}$ is based on an U-Net [25] architecture which is also similar to the one in monodepth [7]. It contains an image feature encoder, an iToF depth feature encoder, and a feature fusion decoder. For both RGB image feature and iToF depth feature encoders, they adopt ResNet18 [11] as their backbone while the former takes the pretrained weight from ImageNet [4] classification task as warm start. The multi-scale features extracted by layers of both encoders are concatenated in a layer-wise manner and further fed to the corresponding convolutional blocks (of the same scale) in the fusion decoder.

4 Experiments

4.1 Datasets

The experiments are conducted on two datasets: the synthetic *ToF-FlyingThings3D* [21] dataset and the real-world *iToF-RGB1k* dataset collected by ourselves.

ToF-FlyingThings3D [21] As such dataset is synthetic to have full access to the groundtruth depth, we mainly adopt it for our quantitative evaluation. Two different camera configurations are used in our experiments to synthesize the dataset: 1) **pseudo camera parameters** as used in its original paper [21] for ensuring a fair comparison with other methods, where all the cameras are with the same focal length (thus nearly the same FoV) and the extrinsic transformation is simplified (*i.e*. no rotation and only 2D orthogonal translation); 2) **device camera parameters**, where we adopt the calibration parameters obtained from the RGB-D camera module (*i.e*. the device that we use for collecting our iToF-RGB1k dataset, which has different focal lengths for all three cameras and the extrinsic transformations are more complicated), making the synthesized dataset more challenging for the integration between iToF and RGB cameras.

Methods	SL	Training Ground Truth		Input		MAE(cm)
		Depth	RGB	RGB	Depth	
		Metric Relative			FoV	
DeepToF [18]	✓	✓			ToF	4.69
ToF-Net [27]	✓	✓			ToF	4.90
TOF-KPN w/o RGB [21]	✓	✓			ToF	2.44
SHARP-Net [5]	✓	✓			ToF	1.19
TOF-KPN [21]	✓	✓		✓	ToF	1.51
Our network w/ TOF-KPN loss [21]	✓	✓		✓	ToF	1.50
Cross-warp			✓	✓	RGB	3.16
Cross-warp + Structure Distillation		✓	✓	✓	RGB	3.01

Table 1. Comparison with competitive methods on the ToF-FlyingThings3D dataset [21]. SL: Supervised learning.

iToF-RGB1k We collect such iToF-RGB1k dataset by using the mobile-phone device of RGB-D camera in the natural world, in which it comprises 1074 scenes that have been randomly split into 960 sets for training and 114 sets for testing. The iToF depth has resolution of 640×480, while the RGB images have a resolution of 1280×960. As iToF is better suited for indoor environments, the majority of scenes in the dataset are indoor ones. We also consider the social impact of privacy to avoid capturing the human being.

4.2 Quantitative Experiments

Comparison with iToF Refinement Methods To ensure a fair comparison with supervised learning methods, we first train our integration network $\mathbb{M}$ using the supervised learning objective proposed by Qiu *et al*. and follow the same evaluation protocol [21]. This objective is also used in SHARP-Net [5].

As shown in the row "our network w/ TOF-KPN loss" in Table 1, we successfully reproduce the performance of [21]. We then evaluate the performance of our full model, as shown in the last row in Table 1. Our full model outperforms DeepToF [18] and ToF-Net [27] (both supervised ones) without requiring the strong supervision of ground truth depths and can achieve full FoV of RGB image. Although SHARP-Net [5] has the best performance of mean absolute error (MAE), it is limited to refining the FoV of iToF. Considering the inherited performance gap between the supervised and self-supervised learning methods [7,8], our method performs well as a weakly supervised method.

Ablation Studies on Objectives and Modalities To investigate how the objectives and input modalities affect the performance of our model, we conduct ablation studies with two camera configurations. As shown in Table 2, our ablation study of objectives starts with the supervised learning baselines (in (a) and (b) rows) and self-supervised learning baselines (in (c) and (d) rows). Then, we use a single RGB image (as shown in (c) rows) or stereo RGB images (as shown in (d) rows) as input for the integration network trained with cross-warp consistency. In pseudo camera configuration, the model using

#	Input RGB	Input iToF	SL	Weak-sup. CW	Weak-sup. SD	Cam.	Eval. Region	MAE(cm) ↓	AbsRel ↓	SqRel ↓	RMSE ↓	$\text{RMSE}_{\log}$ ↓	$\delta < 1.25 \uparrow$	$\delta < 1.25^2 \uparrow$	$\delta < 1.25^3 \uparrow$
(1) Pseudo Camera Parameters															
(a)		✓	KPN			L	RGB	2.130	0.04425	0.3086	4.475	0.04165	0.9394	0.9802	0.9957
						R		1.550	10.56	215.5	4.392	0.04170	0.9747	0.9937	0.9969
(b)	M	✓	KPN			L	RGB	1.686	0.03585	0.2254	3.541	0.03512	0.9550	0.9793	0.9958
						R		1.212	0.03614	1.0120	3.275	0.02790	0.9770	0.9938	0.9985
(c)	M			✓		L	RGB	7.267	0.08433	3.427	16.30	0.12330	0.9063	0.9397	0.9543
						R		5.965	0.08375	2.835	13.46	0.11350	0.9105	0.9446	0.9581
(d)	S			✓		L	RGB	4.059	0.05112	1.444	11.04	0.08002	0.9492	0.9672	0.9753
						R		5.151	0.07691	2.828	13.37	0.09934	0.9250	0.9508	0.9640
(e)	M	✓		✓		L	RGB	3.156	0.05717	0.760	7.775	0.06248	0.9519	0.9674	0.9762
						R		3.236	0.05914	1.215	7.941	0.06750	0.9448	0.9650	0.9762
(f)	M	✓		✓	✓	L	RGB	3.006	0.04710	0.7509	7.731	0.06208	0.9506	0.9651	0.9739
						R		3.213	0.05669	1.255	7.930	0.06841	0.9454	0.9639	0.9764
(g)						L	iToF	3.005	0.04502	0.7399	7.786	0.06098	0.9536	0.9676	0.9764
						R		3.206	0.05477	1.215	7.935	0.06717	0.9475	0.9657	0.9784
(h)						L	Ext.	3.536	0.07709	1.259	7.561	0.07777	0.9140	0.9344	0.9445
						R		3.891	0.09745	2.536	8.328	0.08513	0.9121	0.9350	0.9497
(i)	M	✓	S2D			L	RGB	54.17	1.383	118.8	64.32	0.4071	0.1132	0.2456	0.4110
						R		56.76	1.565	141.3	66.62	0.4281	0.1013	0.2216	0.3746
(j)	M	✓	S2D		✓	L	RGB	20.52	0.2246	7.323	27.92	0.1687	0.5968	0.8467	0.9168
						R		17.77	0.2234	7.141	22.99	0.1738	0.6270	0.8390	0.9082
(2) Device Camera Parameters															
(a)		✓	KPN			UW	RGB	5.454	0.06630	1.639	13.03	0.07756	0.9333	0.9664	0.9784
						W		2.473	0.03923	0.5655	6.923	0.04683	0.9586	0.9861	0.9941
(b)	M	✓	KPN			UW	RGB	1.921	0.03816	0.2513	3.655	0.03778	0.9644	0.9800	0.9893
						W		1.604	0.02757	0.1765	3.516	0.02531	0.9774	0.9927	0.9991
(c)	M			✓		UW	RGB	6.740	0.08767	2.508	14.45	0.10030	0.9021	0.9846	0.9650
						W		10.60	0.14480	3.976	17.39	0.11650	0.8185	0.9257	0.9576
(d)	S			✓		UW	RGB	11.62	0.1270	4.356	22.25	0.12260	0.8578	0.9222	0.9513
						W		21.78	0.3129	13.97	31.36	0.17940	0.5835	0.8098	0.8982
(e)	M	✓		✓		UW	RGB	6.115	0.07910	2.116	13.42	0.08497	0.9175	0.9578	0.9747
						W		9.211	0.13140	2.924	15.20	0.09361	0.8456	0.9469	0.9782
(f)	M	✓		✓	✓	UW	RGB	5.616	0.07305	1.773	12.62	0.07644	0.9232	0.9607	0.9764
						W		9.376	0.12280	2.995	15.57	0.09220	0.8579	0.9494	0.9752
(g)						UW	iToF	7.453	0.09368	2.189	13.42	0.08332	0.9027	0.9564	0.9764
						W		8.944	0.11670	2.615	14.60	0.08985	0.8718	0.9525	0.9759
(h)						UW	Ext.	4.715	0.06289	1.570	12.03	0.07208	0.9333	0.9628	0.9764
						W		10.62	0.13980	4.064	17.84	0.09721	0.8194	0.9395	0.9731

Table 2. Quantitative studies for different supervision and the input. SL: Supervised learning. Weak-sup.: Weakly supervised learning. CW: Cross-warp. SD: Structure distillation. Cam.: Camera. M: Monocular RGB. S: Stereo RGB. KPN: Using supervised depth refinement loss of TOF-KPN[21]. S2D: Using Sparse-to-Dense[17] for supervised learning. L: Left rgb camera. R: right rgb camera. UW: Ultra-wide RGB camera. W: Wide RGB camera. Eval. region: Evaluated region. Ext.: extended FoV of iToF.

stereo RGB input outperforms the one using monocular RGB input. In device camera configuration, however, the opposite results may be associated with the challenge of warping images and depths with different FoVs. Next, we evaluate the performance of our model with cross-modal stereo input (iToF and RGB), as shown in the (e) rows. The results indicate that the model using cross-modal stereo input outperforms those models using single RGB modality because the iToF depths provide metric information for absolute depth estimation. Lastly, as shown in the (f) rows, model training with structure distillation from passive depths improves most performance metrics (excluding the threshold-based ones), indicating the advantages of better structure guidance and knowledge from the off-the-shelf monocular depth estimation model.

Comparison between original and extended FoVs As shown in the (g) and (h) rows of Table 2, we evaluate the performance of our estimated depths within the original iToF FoV and the extended region outside the FoV of iToF, as illustrated in Figure 2. The results with pseudo camera parameters indicate that the estimated depths in the original iToF FoV on both RGB image planes are better than those in the extended FoVs. With device camera parameters, however, this case holds only on the wide-angle image plane, suggesting that the model for the wide-angle cameras could more depend on the metric information from iToF depths than the model for the ultra-wide camera because of the larger overlapping region between iToF and RGB cameras.

Comparison with Depth Completion Method To align with the setting of Sparse-to-Dense [17], we randomly sample 750 points from D^{L}_{iToF} to generate the sparse depth maps. These sparse depth maps are paired with I_L as the training pairs for Sparse-to-Dense [17]. The results in the (i) rows of Table 2, where the worse performance indicates that Sparse-to-Dense [17] is less effective for tackling the noise in iToF depth and for leveraging the complementary properties across modalities. Even being further regularized by the structure distillation loss (as shown in the (j) rows), the performance is still much worse than ours. In contrast, our proposed method leveraging geometric constraints for cross-warp consistency is more effective in alleviating the interference from noisy iToF and gets better fusion results.

4.3 Qualitative Experiments

We conduct experiments using our iToF-RGB1k to qualitatively evaluate our model in the real world. As shown in Figure 5, our model is capable of extending the original FoV of iToF depths to the FoV of RGB images. Moreover, our model is able to remedy the errors or noises of the iToF sensor. For example, as shown in the first and second row of Figure 5, the iToF depth values within the circled regions are largely deviated due to the reflection of the wall or the transparent umbrella. Our model refines the results by leveraging the rich appearance and structure information from the RGB image. Other examples shown in the third and fourth row in Figure 5 demonstrate our model's capability to correct depth errors from the off-the-shelf monocular depth estimation model [31,19]. Monocular depth estimation models, reliant on passive sensing RGB cameras, often misinterpret visual cues [12] from TV screens and walls because of misleading or absent textures. In this case, the depth information from active sensing iToF proves beneficial in resolving the ambiguity. To sum up, our cross-warp and depth structure distillation model successfully integrates the passive sensing RGB image and the active sensing iToF depth to estimate the full FoV metric depth map of the scene.

5 Conclusions

We introduce a weakly-supervised framework to tackle the task of cross-modal depth estimation, driven by cross-warp consistency and depth structure distillation. Our proposed cross-warp consistency adopts iToF depth estimates to build the inter-camera

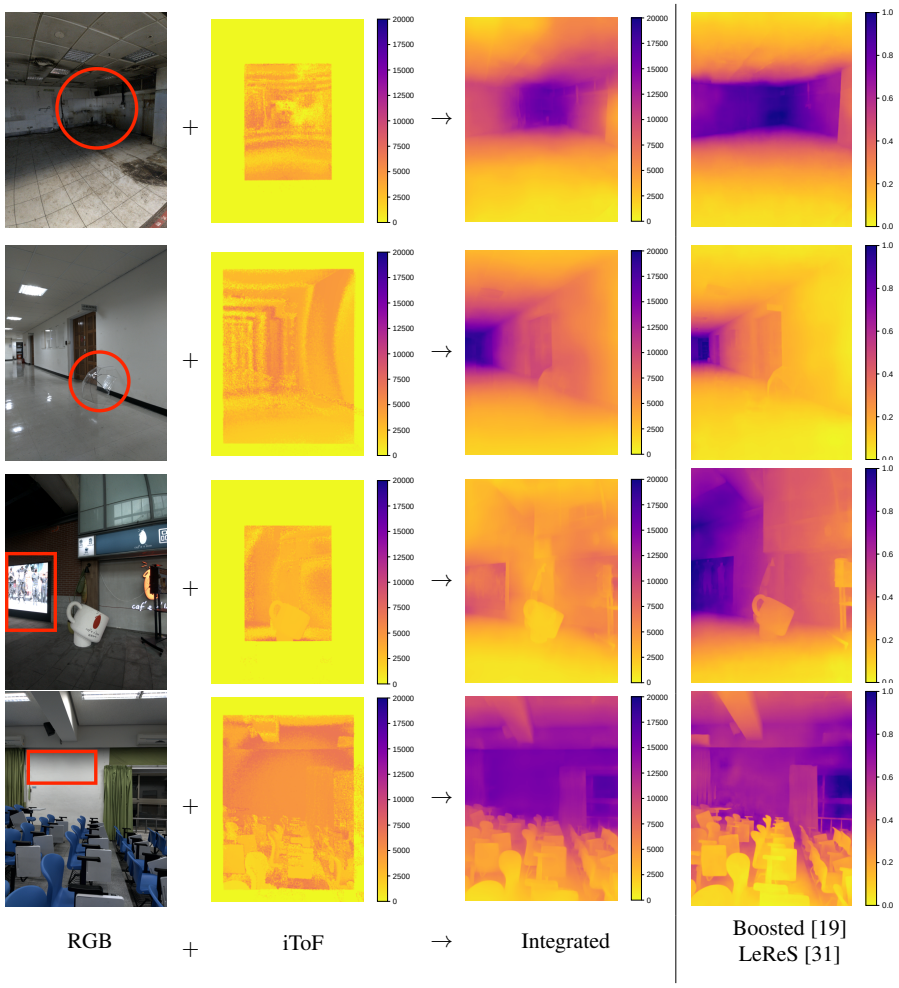

Fig. 5. Qualitative evaluation of estimated depths for the real-world dataset, iToF-RGB1k. Our model overcomes the limitations of the iToF camera and the off-the-shelf monocular depth estimation model, such as the reflective objects and transparent objects (which are red-circled in the first and second row), and the wrong visual depth cues (which are framed by red rectangles in the third and fourth row). Boosted [19] LeReS [31] is the off-the-shelf monocular depth estimation model, a relative depth model. The unit of absolute depth is millimeters.

photometric consistency for guiding the model training, and the depth structure distillation preserves the structure of RGB images under the help of an off-the-shelf monocular depth estimation model. Our quantitative experiment on ToF-FlyThings3D [21] shows that our method is able to achieve comparable performance with several supervised learning methods despite the lack of depth domain ground truths. Moreover, we collect an iToF-RGB1k dataset for performing qualitative evaluation in the real world, in which the corresponding experimental results verify the efficacy of our method in extending

the FoV of iToF as well as fixing the incorrect/noisy depth estimate where neither iToF camera nor off-the-shelf monocular depth estimation model can perform well.

References

1. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023)
2. Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools (2000)
3. Choi, J., Jung, D., Lee, Y., Kim, D., Manocha, D., Lee, D.: Selfdeco: Self-supervised monocular depth completion in challenging indoor environments. In: ICRA (2021)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009)
5. Dong, G., Zhang, Y., Xiong, Z.: Spatial hierarchy aware residual pyramid network for time-of-flight depth denoising. In: ECCV (2020)
6. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)
7. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: CVPR (2017)
8. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: ICCV (2019)
9. Guo, Q., Frosio, I., Gallo, O., Zickler, T., Kautz, J.: Tackling 3D ToF artifacts through learning and the flat dataset. In: ECCV (2018)
10. Hansard, M., Lee, S., Choi, O., Horaud, R.P.: Time-of-flight cameras: principles, methods and applications. Springer Science & Business Media (2012)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
12. Hu, J., Zhang, Y., Okatani, T.: Visualization of convolutional neural networks for monocular depth estimation. In: ICCV (2019)
13. Jung, H., Brasch, N., Leonardis, A., Navab, N., Busam, B.: Wild ToFu: Improving range and quality of indirect time-of-flight depth with rgb fusion in challenging environments. In: 3DV (2021)
14. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)
15. Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J., Han, J.: On the variance of the adaptive learning rate and beyond. arXiv preprint arXiv:1908.03265 (2019)
16. Ma, F., Cavalheiro, G.V., Karaman, S.: Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. In: ICRA (2019)
17. Ma, F., Karaman, S.: Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In: ICRA (2018)
18. Marco, J., Hernandez, Q., Munoz, A., Dong, Y., Jarabo, A., Kim, M.H., Tong, X., Gutierrez, D.: Deeptof: off-the-shelf real-time correction of multipath interference in time-of-flight imaging. ACM TOG (2017)
19. Miangoleh, S.M.H., Dille, S., Mai, L., Paris, S., Aksoy, Y.: Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In: CVPR (2021)
20. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. NeurIPS (2019)

21. Qiu, D., Pang, J., Sun, W., Yang, C.: Deep end-to-end alignment and refinement for time-of-flight RGB-D modules. In: ICCV (2019)
22. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV (2021)
23. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE PAMI (2020)
24. Riba, E., Mishkin, D., Ponsa, D., Rublee, E., Bradski, G.: Kornia: an open source differentiable computer vision library for pytorch. In: WACV (2020)
25. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
26. Shivakumar, S.S., Nguyen, T., Miller, I.D., Chen, S.W., Kumar, V., Taylor, C.J.: Dfusenet: Deep fusion of rgb and sparse depth information for image guided dense depth completion. In: ITSC (2019)
27. Su, S., Heide, F., Wetzstein, G., Heidrich, W.: Deep end-to-end time-of-flight imaging. In: CVPR (2018)
28. Verdié, Y., Song, J., Mas, B., Busam, B., Leonardis, A., McDonagh, S.: Cromo: Cross-modal learning for monocular depth estimation. In: CVPR (2022)
29. Wong, A., Soatto, S.: Unsupervised depth completion with calibrated backprojection layers. In: ICCV (2021)
30. Wu, C.Y., Wang, J., Hall, M., Neumann, U., Su, S.: Toward practical monocular indoor depth estimation. In: CVPR (2022)
31. Yin, W., Zhang, J., Wang, O., Niklaus, S., Mai, L., Chen, S., Shen, C.: Learning to recover 3D scene shape from a single image. In: CVPR (2021)