iToF-flow-based High Frame Rate Depth Imaging

Yu Meng[1], Zhou Xue[2], Xu Chang[3], Xuemei Hu[1] *, Tao Yue[1]
[1] Nanjing University, Nanjing, China, [2] Li Auto, [3] Bytedance Inc.
mengyu@smail.nju.edu.cn, xuezhou08@gmail.com, changxu.21@bytedance.com,
{xuemeihu,yuetao}@nju.edu.cn

Abstract

iToF is a prevalent, cost-effective technology for 3D perception. While its reliance on multi-measurement commonly leads to reduced performance in dynamic environments. Based on the analysis of the physical iToF imaging process, we propose the iToF flow, composed of cross-mode transformation and uni-mode photometric correction, to model the variation of measurements caused by different measurement modes and 3D motion, respectively. We propose a local linear transform (LLT) based cross-mode transfer module (LCTM) for mode-varying and pixel shift compensation of cross-mode flow, and uni-mode photometric correct module (UPCM) for estimating the depth-wise motion caused photometric residual of uni-mode flow. The iToF flow-based depth extraction network is proposed which could facilitate the estimation of the 4-phase measurements at each individual time for high framerate and accurate depth estimation. Extensive experiments, including both simulation and real-world experiments, are conducted to demonstrate the effectiveness of the proposed methods. Compared with the SOTA method, our approach reduces the computation time by 75% while improving the performance by 38%. The code and database are available at https://github.com/ComputationalPerceptionLab/iToF_flow.

1. Introduction

Time of flight (ToF) imaging is a cornerstone technology for depth imaging, renowned for its broad application across numerous fields [17, 20, 36, 38]. The basic principle of ToF is to estimate the time difference between emitting and receiving signals to retrieve depth information [22]. Commonly, ToF imaging technology can be categorized into direct-ToF (dToF) imaging and indirect-ToF (iToF) imaging. In dToF imaging, single photon avalanche diodes (SPAD) or avalanche photodiode arrays (APD) are commonly utilized and the time delay can be directly measured [35]. However, the dToF system is constrained by two main limitations, i.e. high hardware costs and low spatial resolution. In iToF imaging with amplitude-modulated continuous wave (AMCW), depth information is encoded in multiple modes (e.g., 4 modes with 90° phase shifts respectively), which correspond to the cross-correlation integrals of the receiving and emitting signal with different phase shifts [22]. The indirect working principle allows for lower hardware costs and higher spatial resolution [38].

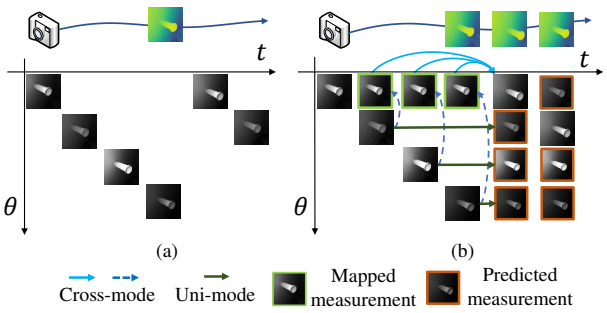


Figure 1. Overview of the proposed iToF flow-based high frame rate depth reconstruction. (a) Captured alternating iToF measurements, (b) iToF flow-based propagation of full-mode measurements, and reconstructed depths.

Nevertheless, the reliance on multiple measurements leads to errors in the depth estimation of dynamic scenes and limited framerate of depth imaging. To overcome the problem, manually designed constraint rules and supplementary information were proposed to correct pixel values for specific motion scenes [5, 10, 21, 29]. Besides, motion compensation methods based on optical flow estimation were also proposed for iToF measurements alignment [11, 28]. However, the realistic 3D motion can not be represented losslessly with 2D optical flow (OF) [33]. Such a dilemma can be tackled with an explicit 3D motion compensation (e.g. scene flow [33]) on the iToF measurement while overcoming the photometric inconsistencies caused by 3D motion and different modes. However, direct estimation of scene flow usually relies on explicit 3D reconstruction [32,33], and the 3D motion estimation is also highly ill-posed and computationally complex. Besides, photometric inconsistency due to different modes of iToF measurements also exacerbates the ill-posedness of OF estimation.

*Corresponding author.

Based on the analysis of the iToF imaging process, we propose iToF-flow to model the variation caused by the 3D motion and alternation of measurement mode, and develop an iToF-flow-based depth extraction neural network for high frame rate depth estimation. Specifically, from a physics-inspired perspective, as shown in Fig. 1, we decompose the variation of iToF measurements into cross-mode flow, which models the photometric variation and pixel shift among different modes, and the uni-mode flow, which models the photometric residuals caused by depth-wise motion of the same mode. As for the cross-mode flow, we observe the motion-insensitive local linear transformation between different modes of the measurements and propose the LLT-based cross-mode transfer module (LCTM). As for uni-mode flow, we derive the depth-dependent photometric residual formulation and propose the uni-mode photometric compensation module (UPCM). With the end-to-end processing, the 3D motion is separated into the 2D plane of OF due to space shift and luminance residuals due to depth-wise motion, which can be extracted sequentially and separately. Compared with other methods, the advantages in runtime and accuracy of our method are demonstrated with extensive experiments. In all, our contributions are concluded as:

- We propose an iToF flow model, composed of cross-mode flow and uni-mode flow, to comprehensively model the variation of measurements caused by different demodulation and 3D motion in iToF imaging.
- We build an iToF flow-based depth extraction network based upon LCTM and UPCM, for high-frame-rate and accurate depth imaging from iToF measurements.
- We provide an extension database with different modulation parameters and scenes to augment the existing iToF databases [27, 28].
- Extensive experiments with simulation and real-world data are conducted and demonstrate the effectiveness of the proposed method.

2. Related Work

iToF for Time-varying Scene. Depth estimation of iToF depends on multiple exposure measurements [22]. Scene and camera motions can induce misalignment between measurements, resulting in errors in depth estimation. Modeling motion-induced pixel misalignment as a noisy time series, Kalman filtering was employed to mitigate the effects of transverse motion [29]. Jan *et al*. [30] proposed a model-based tracking approach using iToF raw measurement to obtain a $10\times$ higher depth frame rate. Chen *et al*. [5] proposed an alignment method based on extra data from a highly dynamic sensor using a short exposure. Gao *et al*. [10] proposed finer categorizations of motion-introduced errors and designed different correction methods respectively based on the neighboring pixel. Lee *et al*. [21] designed rules for detecting moving regions with regular electric charge relations and proposed a replacement method with adjacent pixels.

Database	Type	GT	Size	Motion
FLAT [11]	Syn.	Yes	1.2k	Yes
CB-ToF [27]	Syn.	Yes	21.4k	No
CB-ToF-Extension [28]	Syn.	Yes	2.1k	Yes
MF-ToF [12]	Syn.	Yes	155k	No
Ours	Syn.	Yes	2k	Yes

Table 1. Summary of public iToF datasets.

Furthermore, learning-based method [3, 4] are proposed to recover depth information from extremely low signal-to-noise measurements with short exposure and reduced motion-induced error. Guo *et al*. [11] proposed a neural network-based encoder-decoder architecture to output velocity maps, which can be used to align the raw iToF measurements. Similarly, Michael *et al*. [28] used the optical flow estimation network as their baseline and proposed multiple loss regularization terms to help overcome the photometric gap between iToF measurements. It's worth noting that the majority of the mentioned methods primarily focus on aligning pixel positions, neglecting the photometric errors induced by depth-wise motion.

Dynamic Cross Modality Imaging. To capture multi-modality information with a single camera, imaging techniques that alternately capture different modalities at different times are proposed. Specifically, in the field of real-time hyperspectral imaging, Hu *et al*. [14] propose a complex optical flow (COF)-based method to reconstruct hyperspectral video information from spectral-sweep video sequences. In the field of high dynamic range (HDR) video imaging, capturing and fusion low dynamic ranges (LDR) frames with different exposures is commonly adopted [2, 6, 19]. In [19], a CNN-based method was proposed to estimate the motion flow between two frames explicitly. Chen *et al*. [2] proposed a multi-stage spatial temporal pixel alignment for HDR video reconstruction. In addition to pixel domain alignment, feature domain alignment was proposed [6]. Pu *et al*. [26] proposed a Pyramidal Alignment and Masked merging network (PAMnet) to align the pyramid multi-scale feature of LDR images to synthesize HDR images. These works, based on the physical principle of different measurement modes, showing elegant performance in reconstructing multi-mode video from cross-mode measurements.

iToF Simulation and Database. Data-driven approaches typically rely on extensive, high-quality datasets paired with accurate ground truth. Obtaining detailed 3D ground truth data from real-world iToF photography is both challenging and expensive. Therefore, simulation has become the primary way of acquiring iToF dataset. Mitsuba is a physically-based rendering system developed for research in computer graphics and physically-based modeling [16]. Based on Mitsuba, MitsubaToFRenderer [25] was proposed for rendering ToF data specifically. Besides, common development tools, such as Blender [7] or Unity [13], are able to support importing ToF cameras in simulation. Tab. 1 summarized the publicly accessible large datasets of iToF

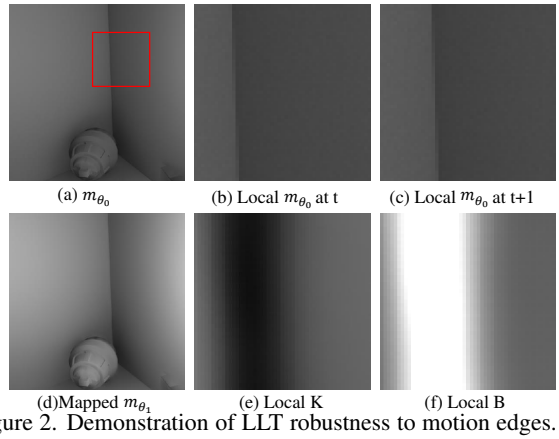


Figure 2. Demonstration of LLT robustness to motion edges. Local area in (b), (c), (e) and (f) is marked with red box in (a).

imaging. As shown, there is still a shortage of adequate datasets featuring motion. We propose to supplement the iToF dataset with around 2K motion samples.

3. iToF Flow-Based Depth Extraction Network

In alternating 4-mode iToF imaging, measurements at different times exhibit varying phase-shifts. Groups of four measurements, each with a 90° phase-shift, are conventionally used to extract depth, reflectance, and environmental illumination. For dynamic scenes with fast-moving objects, the asynchrony of these 4-mode measurements introduces ghost and blur artifacts in the moving edges, compromising the accuracy of the extracted depth. To solve the problem, it is natural to propagate the three absent mode information of a single moment from previous measurements with optical flow [28]. However, the photometric inconsistency introduced by the heterogeneous phase shift and depth-wise motion at different times prevents the direct utilization of the conventional optical flow-based methods, leading to errors and artifacts which is hard-to-ignore in practice.

In this paper, we propose an iToF flow model that builds the relationship between the measurement at different times and phase shifts, and decomposes the iToF flow depth estimation network into two modules: the cross-mode transfer and uni-mode photometric correction modules. Building upon these two modules, we propose the iToF flow-based depth estimation neural network that accurately estimates depth of each moment in an end-to-end manner by propagating absent phase information from previous measurements to the target time and correcting depth-wise motion-induced photometric-intensity-bias with the uni-mode flow. Details of the iToF model and the depth extraction network are illustrated in this section.

3.1. LLT-based Cross-mode Transformation

In order to tackle the measurement absence problem, we propose an LLT-based cross-mode transformation for propagating the measurements with the corresponding mode at the previous moments to the target times. In the following, we give a detailed analysis of the LLT-based cross-mode transfer and the corresponding network structure design.

Cross-mode LLT Mapping. The challenging aspect of information propagation lies in accurately aligning the 2D motion flow between two modes at different times, known as cross-mode flow estimation. Thus we propose a motion-insensitive local linear transfer model that could formulate the mapping between different modes. Thanks to its motion-insensitive property, the model remains invariant for slightly moved frames within a short time range, such as a single iToF acquisition period with 4-mode frames. The LLT model relies on a universal image property: for each sufficiently small local region around the edge, there are two types of features corresponding to the piece-wise smooth regions on the two sides of the edge. The features of pixels on the edges can be represented as a linear combination of the features of the two sides [37], which still holds for iToF measurements regardless of the measured depth. Based on this property, the 4-mode iToF measurements within a local region could be represented by the points along the line determined by the two 4-mode iToF measurement vectors of two regions around the edge. Denoting the 4-mode vectors on the two sides of the edges as $\mathbf{m}' = [m'_{\theta_1}, m'_{\theta_2}, m'_{\theta_3}, m'_{\theta_4}]^T$ and $\mathbf{m}'' = [m''_{\theta_1}, m''_{\theta_2}, m''_{\theta_3}, m''_{\theta_4}]^T$, the vector $\mathbf{m}^p$ of a certain pixel p in the local region $\mathcal{S}$ could be represented by

$$\mathbf{m}^p = \beta^p \mathbf{m}' + (1 - \beta^p)\mathbf{m}'', \tag{1}$$

where $\mathbf{m}^p = [m^p_{\theta_1}, m^p_{\theta_2}, m^p_{\theta_3}, m^p_{\theta_4}]$, β^p denotes the combination factor of pixel p. Thus, the formulation among elements of the 4-mode vectors could be represented as linear mapping based model, i.e.

$$\begin{aligned} m^p_{\theta_i} &= k_{j\to i} m^p_{\theta_j} + b_{j\to i}, \forall p \in \mathcal{S}, \\ k_{j\to i} &= \frac{m'_{\theta_i} - m''_{\theta_i}}{m'_{\theta_j} - m''_{\theta_j}}, \\ b_{j\to i} &= m''_{\theta_i} - k_{j\to i} m''_{\theta_j},\ i, j \in \{\theta_1, \theta_2, \theta_3, \theta_4\}. \end{aligned} \tag{2}$$

It is obvious that for a certain region where the two-feature property is valid, the parameters of LLT, i.e., $k_{j\to i}$ and $b_{j\to i}$, are identical for all the pixels in the local region. In other words, for a small local region, the mapping parameters $k_{i\to j}$ and $b_{i\to j}$ at different pixels are uniform and do NOT change in edge position. Therefore, if the edge movement in a relatively small time period does not exceed the range of the region, the LLT could remain unchanged, demonstrating a valuable insensitive property in our scenario. As shown in the Fig. 2, we calculated the LLT maps between alternating modes at different times, and it can be seen that k and b present an extended area around the edges, which conforms to our assessment of the motion insensitive property of the LLT transformation in the edge region. Benefiting from this motion-insensitive characteristic, we could easily comple-

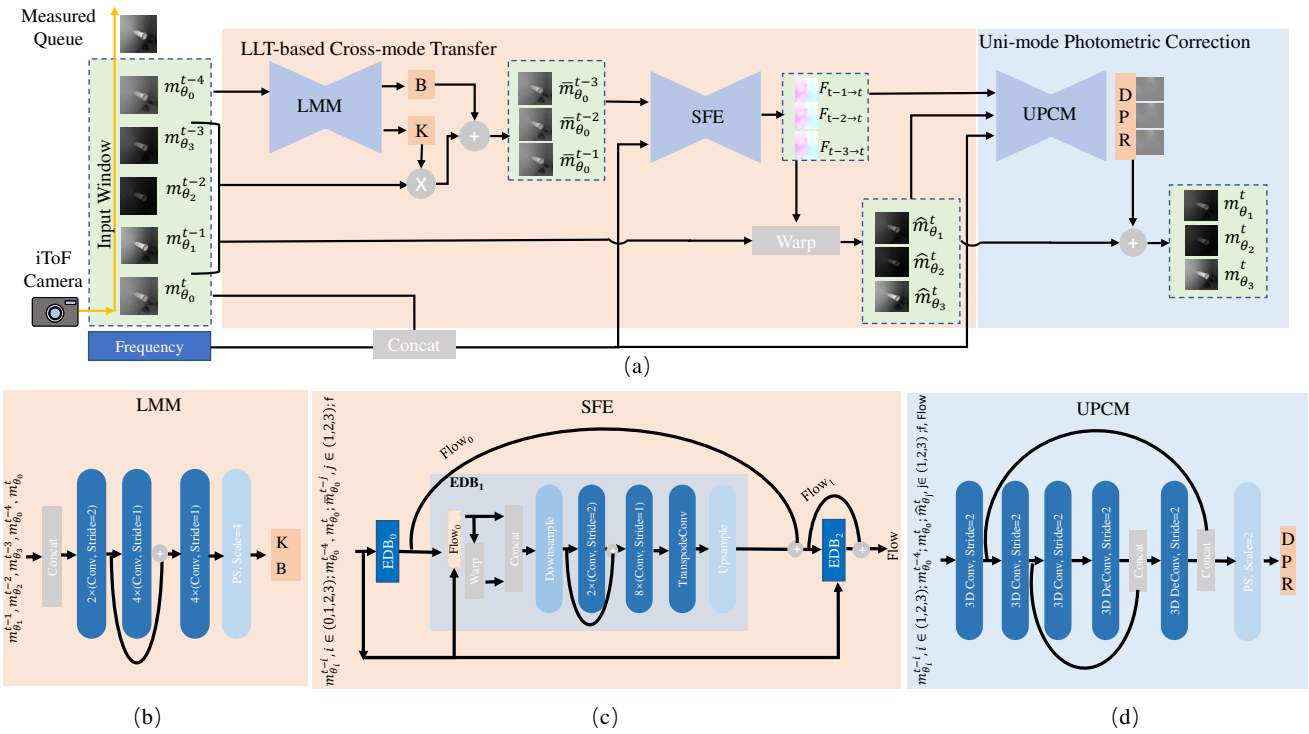


Figure 3. (a) Framework of the proposed network, composed of cross-mode transfer module and uni-mode correction module. (b) Structure of submodule LMM. (c) Cascade Schematic of submodule SPE. (d) Illustration of the structure of UPCM.

ment the absent modes from the current measurement and the LLT maps estimated from the previous frames.

LLT-based Cross-mode Transfer Module. We propose an LLT-based Cross-mode Transfer Module (LCTM) to compute the LLT maps from the previous measurements and transfer the absent modes at the current moment from previous measurements. Instead of computing the depth by using the measurements computed by LLT directly, we introduce an optical flow-based measurement propagation framework to prevent the imperfections caused by the invalid cases of the two-feature property in very few regions, i.e., estimate the optical flow between the LLT mapped measurements at time $t-i, i \in 1,2,3$ and the real captured measurements at current time t, and then compute the absent measurements of time t by warping the previous measurements according to the optical flows. As shown in Fig. 3 (a) and (b), the measurements at the aligned target moment t is $m^t_{\theta_0}$. The submodule LLT-based mapping module (LMM) computes the $\boldsymbol{K} = [K_{\theta_1\to\theta_0}, K_{\theta_2\to\theta_0}, K_{\theta_3\to\theta_0}]$ and $\boldsymbol{B} = [B_{\theta_1\to\theta_0}, B_{\theta_2\to\theta_0}, B_{\theta_3\to\theta_0}]$ from the previous alternating measurements by

$$\mathrm{LMM}(m^{t-4}_{\theta_0}, m^{t-3}_{\theta_3}, m^{t-2}_{\theta_2}, m^{t-1}_{\theta_1}, m^{t}_{\theta_0}) = [\boldsymbol{K}; \boldsymbol{B}]. \quad (3)$$

As shown in Fig. 3 (b), we use a pithy Convolutional Neural Network(CNN) structure to construct the $\mathrm{LMM}(\cdot)$. Two convolutional layers with a stride of 2 downsample the input to a quarter of the original resolution. Eight convolutional layers with 1-pixel stride extract features. The output is upsampled by pixelshuffle. Since we do NOT use the simple patch-based closed form LLT in Eq. (2), there is no explicit hyper-parameters or constraints on motion sizes are required. To supervise the LMM, we evaluate the performance of the LMM with mean absolute error (MAE).

$$\mathcal{L}_{\mathrm{LMM}} = \frac{1}{3}\sum_{i=1}^{3} \mathrm{MAE}(m^{t-i}_{\theta_i} K_{\theta_i\to\theta_0} + B_{\theta_i\to\theta_0}, m^{t-i}_{\theta_0}). \quad (4)$$

After generating the absent mode $\bar{m}^{t-i}_{\theta_0}, i \in \{1,2,3\}$ at three previous moments by LMM, we use a spatial flow estimation (SFE) submodule with the hierarchical encoder-decoder block (EDB) cascaded architecture [9, 15] to transfer the corresponding modes to current time t, complet-menting the three absent modes. The structure of the SFE is shown in Fig. 3 (c), we used three EDB cascaded in our approach. The proposed SFE can be formed as,

$$\mathrm{SFE}(m^{t-4}_{\theta_0}, m^{t-3}_{\theta_3}, m^{t-2}_{\theta_2}, m^{t-1}_{\theta_1}, m^{t}_{\theta_0}, f) = Flow, \quad (5)$$

where f is the modulation frequency, $Flow = [F_{t-3\to t}, F_{t-2\to t}, F_{t-1\to t}]$ is the optical flows from time $t-i, i \in 1,2,3$ to t. Our cross-mode and uni-mode framework is preserved in SFE by feeding the $m^{t-4}_{\theta_0}$ into the SFE. The long-term extra information gives additional characteristics of motion and texture for better short-term motion estimation, which has been proven in video processing [8].

At each EDB, the input is downsampled with different scale ratios, (i.e., [0.25, 0.5, 1] for EDB0-2). Such a

pyramid-like process can utilize coarse-to-fine bias estimation. As shown in Fig. 3 (c), we also use the output of each block as the input to the next blocks. The MAE loss between warped measurements and ground truth is used to supervise the optical flow estimation.

$$\mathcal{L}_{\text{SFE}} = \frac{1}{3}\sum_{i=1}^{3} \text{MAE}(warp(m_{\theta_i}^{t-i}), m_{\theta_i}^{t}), \tag{6}$$

where $warp(\cdot)$ is the warping function with optical flow.

3.2. iToF Uni-mode Photometric Correction

Based on the LCTM, the absent modes of time t could be transferred from the three previous measurements $m_{\theta_j}^{t-i}, i \in \{1,2,3\}$. Then, the photometrical inconsistency of the same modes introduced by the depth-wise motion needs to be correct for accurate depth estimation. Here, we construct the iToF uni-mode photometric compensation model to correct the inconsistency. Specifically, we approximate the iToF measurement model with its first-order Taylor expansion, formulating the relationship between the measurement disturbance to the phase disturbance. Based on this, we build a Uni-mode Photometric Correction Module (UPCM) to correct the uni-mode photometric inconsistency from a set of previous uni-mode measurement disturbances. Some theoretical analysis of the iToF uni-mode photometric compensation model and the UPCM is described in the following.

iToF Uni-mode Photometric Compensation Model. For the iToF measurement with the same demodulation phase of dynamic scenes, the measurement model is

$$m_\theta(\boldsymbol{u};t) = I(\boldsymbol{u};t) + A(\boldsymbol{u};t)\cos\left(\varphi(\boldsymbol{u};t) + \theta\right), \tag{7}$$

where $\boldsymbol{u}$ denotes the pixel coordinate (x, y). $I(\boldsymbol{u};t)$ and $A(\boldsymbol{u};t)$ represent the environment illumination and scene reflectance ratio respectively. $\varphi(\boldsymbol{u};t) = \frac{4\pi f d}{c}$ is the phase difference corresponding to the time of flight. d is depth. c denotes the light speed. $\theta \in \{0°, 90°, 180°, 270°\}$ is the phase of demodulation function.

When measuring a 3D scene, a moving 3D point $\boldsymbol{P}(t) = (X, Y, Z)$ with a 3D displacement $[\Delta X, \Delta Y, \Delta Z]$ from t to $t + \Delta t$ can be projected to iToF measurement with the pixel location shift $\Delta\boldsymbol{u} = (\Delta x, \Delta y)$ and the variation of depth map Δd. Further, Δd leds to the change $\Delta\varphi$,

$$\Delta\varphi = \varphi(\boldsymbol{u} + \Delta\boldsymbol{u}; t + \Delta t) - \varphi(\boldsymbol{u};t) = \frac{4\pi f \Delta d}{c}. \tag{8}$$

As for the uni-mode cases, although the photometric inconsistency caused by slight depth-wise motion exists, the optical flow $\Delta\boldsymbol{u}$ between the uni-mode measurements still can be estimated with relatively high precison. Besides, the environment illumination keep constant after warping with the optical flow, i.e., $I(\boldsymbol{u};t) = I(\boldsymbol{u} + \Delta\boldsymbol{u}; t + \Delta t)$. For reflection-dependent parameter $A(u;t)$, considering the light travel distance and the projected size on sensor, the relationship between $A(u;t)$ and the depth d is $A(u;t) = R(u;t)\frac{S}{d^2}$ with the inverse-square law [23], where S is the amplitude of emitted signal, $R(u;t)$ is reflection rate. When depth changes, we can get the approximate expression by first-order Taylor equation,

$$\Delta A(u;t) = -2R(u;t)\frac{S}{d^3}\Delta d, \tag{9}$$

Considering the fact that $d \gg \Delta d$, we can simplify such tiny variation and get $A(\boldsymbol{u};t) = A(\boldsymbol{u} + \Delta\boldsymbol{u}; t + \Delta t)$. Then, the photometric residual could be represented by

$$\begin{aligned} &m_{\theta_i}(\boldsymbol{u} + \Delta\boldsymbol{u}; t + \Delta t) - m_{\theta_j}(\boldsymbol{u};t) = \\ &A(\boldsymbol{u};t)(cos(\varphi(\boldsymbol{u};t) + \Delta\varphi + \theta_i) - cos(\varphi(\boldsymbol{u};t) + \theta_j)). \end{aligned} \tag{10}$$

For uni-mode with same θ, to simplify the model, we expand Eq. (10) by first-order Taylor equation as

$$\Delta m_\theta = -\Delta\varphi A(\boldsymbol{u};t) sin\left(\varphi(\boldsymbol{u};t) + \theta\right), \tag{11}$$

then we build the relationship between the depth-wise motion caused photometric bias Δm_θ and the phase variation $\Delta\varphi$ with a linear equation. Considering that the modes are different as time varying, we can get $\Delta m_{\theta_{i\to j}}^{(t+\Delta t)\to t} = m_{\theta_i}(\boldsymbol{u} + \Delta\boldsymbol{u}; t + \Delta t) - m_{\theta_j}(\boldsymbol{u};t)$ with the estimated optical flow $\Delta\boldsymbol{u}$. According to Eq. (10), based on linear motion assumption of $\Delta\varphi$ within 4 alternating frames, we can get the mapping $(\varphi(\boldsymbol{u};t), A(\boldsymbol{u};t), \Delta t\Delta\varphi) \to \Delta m_{\theta_{i\to j}}^{(t+\Delta t)\to t}$. Based on at least three consecutive pixel-aligned frames from LCTM, we get unknown independent variables $(\varphi(\boldsymbol{u};t), A(\boldsymbol{u};t), \Delta\varphi)$. According to Eq. (11), the uni-mode photometric compensation variations for depth-wise motion could be estimated.

In brief conclusion, from the above derivation, we find that from previous alternating 4-mode measurements, the uni-mode measurement variation for correcting the depth-wise motion caused photometric residual of the complemented modes at time t could be recovered. In the following, we propose the UPCM network to correct the photometric residual of the absent modes transferred from previous measurements so that the accurate depth at time t could be extracted.

Uni-mode Photometric Correction Module. Here, we introduce the UPCM to compensate the depth-wise motion-induced photometric residuals. The UPCM module is designed based upon 3D CNN [18]. As shown in Fig. 3 (d), our UPCM module uses 3D CNN only for three downsampling and two upsampling, which is much simpler compared to other 3D CNN modules or Unet-like structures.

Based upon the above discussion, we take the previous consecutive measurements $m_{\theta_0}^{t-4}, m_{\theta_3}^{t-3}, m_{\theta_2}^{t-2}, m_{\theta_1}^{t-1}$, warped measurements $\hat{m}_{\theta_3}^{t}, \hat{m}_{\theta_2}^{t}, \hat{m}_{\theta_1}^{t}$ and optical flows $Flow$ from SFE and the intermediate measurements $m_{\theta_0}^{t}$ as the inputs. The UPCM can be presented by

$$\text{UPCM}(m_{\theta_0}^{t-4}, m_{\theta_0}^{t}, m_{\theta_i}^{t-i}, \hat{m}_{\theta_i}^{t}, Flow, f) = R_{\text{dp}}, \tag{12}$$

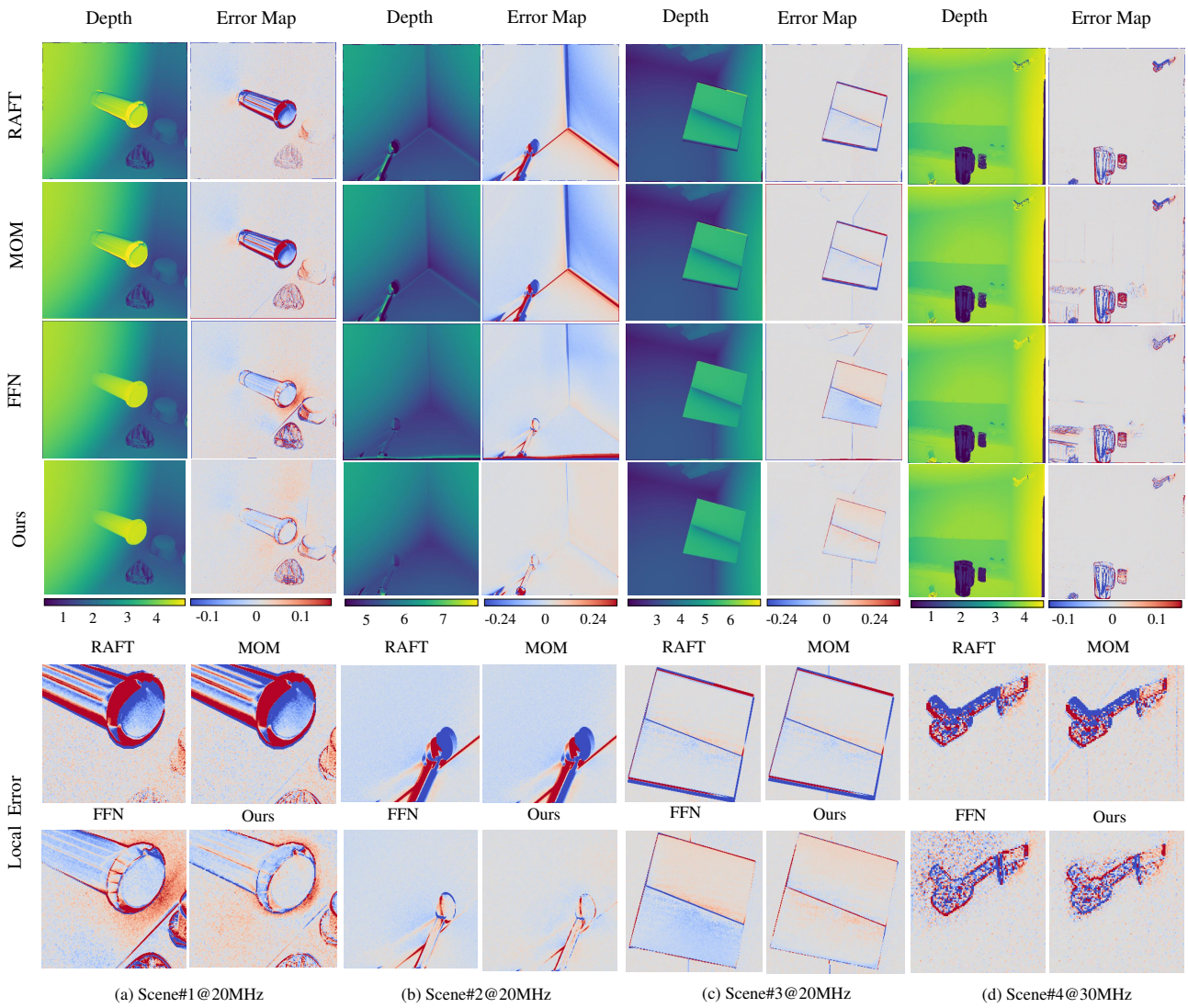


Figure 4. Qualitative comparison. The error map of local details is cropped out and enlarged for visualization below.

where $i \in (1, 2, 3)$, $R_{\text{dp}} = [r_1, r_2, r_3]$ is the output of the UPCM, i.e., the estimated depth-dependent photometric residual (DPR). We add residual to corresponding SFE-aligned measurements to generate the final retrieved measurements at time t. For supervision, we utilize MAE loss to supervise the DPR estimation,

$$\mathcal{L}_{\text{UPCM}} = \frac{1}{3}\sum_{i=1}^{3} \text{MAE}(\hat{m}^t_{\theta_i} + r_i, m^t_{\theta_i}). \quad (13)$$

Loss Function. In our method, we train all the modules in an end-to-end way, and the total loss function is,

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{LMM}} + \lambda_{\text{SFE}}\mathcal{L}_{\text{SFE}} + \lambda_{\text{UPCM}}\mathcal{L}_{\text{UPCM}}, \quad (14)$$

where λ_{SFE} and λ_{UPCM} is the balancing coefficient and empirically chosen as 1.

4. Experiment

In this section, we first introduce the proposed supplemental dataset. Then, we present the results of the proposed method in comparison with the State-of-the-Art (SOTA) methods [11, 28] and a representative optical flow estimation method RAFT [31]. Besides, ablation experiments are conducted to analyze the effectiveness of the proposed LCTM, and UPCM modules. Lastly, we further validate the generalizability of the proposed method to higher speeds and noisey real-world data.

4.1. Supplement Database

For training, testing and validation, the utilized database includes both the database proposed in the [27, 28] and our proposed database. Our proposed supplement contains 20 LiDAR-scanned real scenes of Matterport3D [1] with complexer background textures compared with the Cornell-

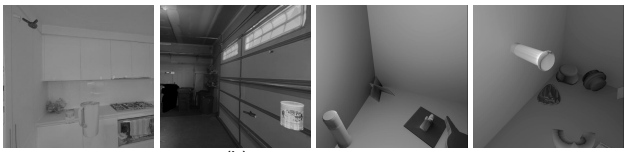


Figure 5. Measurements comparison of our proposed supplement (a) and (b) with the previous database [27, 28] (c) and (d).

Box-like scene in previous databases [27, 28] as shown in Fig. 5. At least 5 objects of YCB-V dataset [34] are imported to each scene at varying distances. Each object has 100 random 6 degrees of freedom (6DoF) moving steps. Further, modulation frequency that differ from the previous works [27, 28], 30 MHz is added. Our supplement database provides additional features to construct the data-driven methods. Specifically, we implement the iToF imaging model with ray tracing in Blender [7] for simulation.

4.2. Comparisons with State-of-the-Art Methods

In this subsection, we show the comparison of our proposed method with FFN [28] and MOM [11]. Besides, we compare our methods with the pre-trained RGB flow net RAFT [31]. We adopt the mean absolute error (MAE) of depth MAE_d and MAE of predicted four mode measurement value MAE_p to evaluate the performance. The unit of depth error in this section is centimeters (cm). We remap all the measurements to $[0, 1024]$ to truncate the overexposure pixel value before error estimation.

Dataset. For training, validation, and test datasets, 20 MHz data in [27, 28] and the proposed 30 MHz data are utilized. We construct the dynamic scenes by extracting measurements of alternating phases at successive viewpoints corresponding to successive timestamps. Specifically, 124 scenes at 20MHz and 16 scenes at 30MHz are used for training. 15 scenes at 20MHz and 2 scenes at 30MHz are used for validation. For testing, 15 scenes at 20MHz and 2 scenes at 30MHz are utilized. Data enhancement with randomized cropping and rotation is used during the training process.

Implementation Details. We train the proposed network with a batch size of 24 for 60 epochs. Cosine annealing [24] is used to decay the learning rate from 2×10^{-4} to 2×10^{-6}. For fair comparison, we retrain FFN [28] and MOM [11] with the same settings in [28] on the same database. All runtime tests are performed on RTX 2080 Ti. We set the batch size to 1, record the forward runtime, and average to the mean runtime in seconds (s). Number of parameters is recorded in millions (M). The patch spatial resolution for training, validation, and test is 448×448.

Overall Performance Comparison. As the baseline to all the evaluations, the standard depth estimation (SDE) with unaligned measurements is demonstrated. The quantitative comparison results are shown in Tab. 2. Our method achieves the best performance in both photometric and depth reconstruction. Compared with the SOTA method

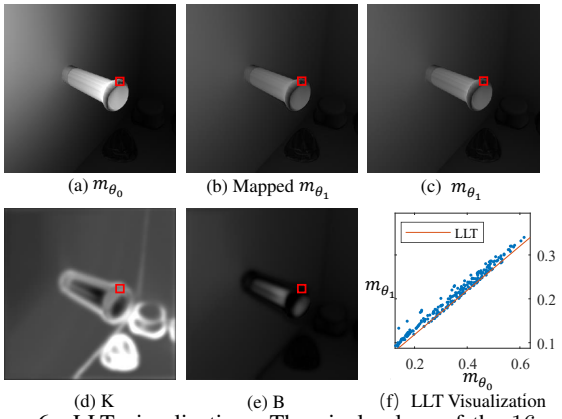
Metric	SDE	RAFT [31]	FFN [28]	MOM [11]	ours
MAE_p	7.31	6.96	3.79	6.72	1.51
MAE_d	16.72	15.40	7.58	14.28	4.72
Mask rate	-	0.82%	0.80%	0.45%	0%
Para. (M)	-	5.26	1.37	9.03	7.75
Time (s)	-	0.27	0.32	0.028	0.081

Table 2. Quantitative comparison results.

FNN [28], the depth reconstruction error of our method is reduced by 37% and the runtime of our method is reduced to a quarter. The mask rate [28] is introduced, which can indicate the number of warping-failed pixels (e.g., out of the image coordinate plane). As shown, our method presents the smallest mask rate. Qualitative comparisons are shown in Fig. 4, and the reconstructed depth maps and error maps for each method are shown. From the comparison, our method shows excellent performance in motion compensation and eliminates most of the artifacts in depth reconstruction. As for runtime comparison, our method has a distinct advantage in terms of speed, being on the same order of magnitude as the best methods MOM [11], which demonstrates the efficiency of our approach. Note that although our method has a relatively higher number of parameters, compared to RAFT [31] and FFN [28], the efficiency of the algorithm is not affected due to the high degree of parallelism inherent in the proposed network.

4.3. Ablation Study

In this subsection, we conduct ablation experiments to further demonstrate the effectiveness of the proposed LCTM, and UPCM. The submodule LMM of LCTM is proposed to predict the LLT maps, which facilitate the estimation of the optical flow. As shown in the Tab. 2, through comparing the method with solely SFE for optical flow estimation, the performance improvement introduced through combining LMM and SFE demonstrates the efficiency of introducing LLT-based cross-mode photometric correction. With SFE and UPCM, the introduction of LMM could significantly improve the performance, further demonstrating the pivotal role of LMM in the overall framework. As shown in Fig. 6 (a) and (b), the coefficients K and B predicted by LMM effectively map the pixel intensity. Further, we verify the local validity of the linear transformation relationship predicted by LMM. As shown in Fig. 6 (f), the linear mapping relationship between pixel intensities within the 16×16 sized region marked by the red box can all be formed by k and b from the center pixel. The UPCM is proposed to estimate the photo residual caused by depth-wise motion. As shown in Tab. 3, UPCM demonstrates significant performance gains in the final MAE_d and MAE_p. Note that we introduce a few 3D CNNs in UPCM, which is the primary source of computational complexity, and achieve the performance benefits from 3D CNNs while not introducing an excessive computational load.



(a) m_{θ_0} (b) Mapped m_{θ_1} (c) m_{θ_1} (d) K (e) B (f) LLT Visualization

Figure 6. LLT visualization. The pixel values of the 16 × 16 region, marked by the red box in (a), are plotted in (f) along with the LLT predicted by LCTM for the center of that region.

LMM	SFE	UPCM	MAE_p	MAE_d	Para.(M)	Time (s)
	✓		1.94	6.24	6.83	0.014
✓	✓		1.82	5.83	7.54	0.017
	✓	✓	1.74	5.51	7.04	0.073
✓	✓	✓	1.51	4.72	7.75	0.081

Table 3. Ablation experiments of different modules, showing the effectiveness of the proposed LCTM and UPCM.

In summary, through the sophisticated integration of LCTM and UPCM based on the iToF flow model, the proposed depth extraction network architecture can efficiently eliminate the influences of mode change and depth-wise motion, achieving high-precision depth estimation.

4.4. Generalization

Validation of Higher-speed Motion. We first validate how the performance of each method changes over different motion speeds. The FFN [28], MOM [11], and our method are trained only on the original database with a maximum step span of 4 corresponding to the 4-mode measurements of successive moving step in the simulated measurement sequence. We select three scenes corresponding to ego motion and scene motion at 20 MHz and scene motion at 30 MHz. The maximum step span in the selected scenes is increased for simulating different multiplicative speeds, i.e., maximum step span 5, 6, 7 and 8 for speed ratio 1.25×, 1.5×, 1.75× and 2×. As shown in Fig. 7, as the speed increases, the error of each method becomes larger. Our method maintains the best performance in the tests with different speeds, fully demonstrating the advantages of our method in high-speed motion scenarios.

Validation of Real-world Data. To further demonstrate our method, we compare the depth reconstruction performance on real-world data. We capture real-world data at 30MHz from the Sony IMX518 with spatial resolution at QVGA. The exposure time for each measurement is 500 μs. As shown in Fig. 8, two scenes are captured, i.e. the hand gesture and box-throwing scene. The "Raw data" image is synthesized by fusing two consecutive measurements, which can visualize the real motion blur and noise. The results of MOM [11], FFN [28] and ours are shown in Fig. 8. For gesture scenes, our method successfully eliminates most of the motion blur and realistically restores the gap between fingers. For box-throwing scenes at a faster speed, our method locates the position and contour of the box more accurately than the other methods. This proves the generalizability of our method.

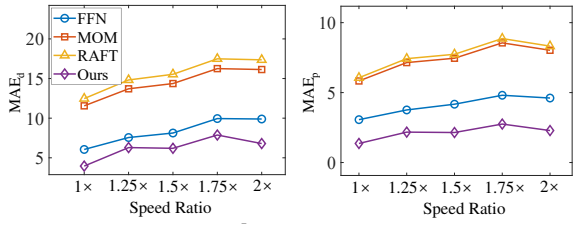


Figure 7. Effect of the speed on the performance.

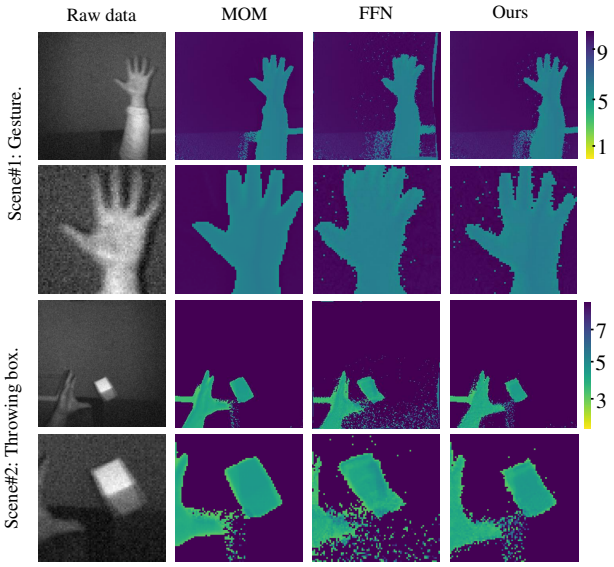


Figure 8. Performance comparison on real-world data.

5. Conclusion

In this paper, we introduced the iToF-flow to model the raw iToF measurement variation due to the measurement mode change and 3D motion, which can be categorized into uni-mode and cross-mode flow. Based on this model, we proposed the iToF-flow-based depth extraction network, comprising LCTM and UPCM. With extensive experiments on both simulated and real-world data, we demonstrate the efficacy of the proposed method.

6. Acknowledgements

This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFA1207200, in part by the NSFC Projects under Grant 61971465.

References

[1] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *2017 International Conference on 3D Vision (3DV)*, pages 667–676, 2017. 6

[2] Guanying Chen, Chaofeng Chen, Shi Guo, Zhetong Liang, Kwan-Yee K Wong, and Lei Zhang. Hdr video reconstruction: A coarse-to-fine network and a real-world benchmark dataset. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2502–2511, 2021. 2

[3] Yan Chen, Keyuan Qian, Xuanye Cheng, and Jingkun Zhou. A learning method to optimize depth accuracy and frame rate for time of flight camera. In *2019 IOP Conference Series: Materials Science and Engineering*, volume 563, page 042067, 2019. 2

[4] Yan Chen, Jimmy Ren, Xuanye Cheng, Keyuan Qian, Luyang Wang, and Jinwei Gu. Very power efficient neural time-of-flight. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2257–2266, 2020. 2

[5] Zhuo Chen, Peilin Liu, Fei Wen, Jun Wang, and Rendong Ying. Restoration of motion blur in time-of-flight depth image using data alignment. In *2020 International Conference on 3D Vision (3DV)*, pages 820–828, 2020. 1, 2

[6] Haesoo Chung and Nam Ik Cho. Lan-hdr: Luminance-based alignment network for high dynamic range video reconstruction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12760–12769, 2023. 2

[7] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 2, 7

[8] Duolikun Danier, Fan Zhang, and David Bull. St-mfnet: A spatio-temporal multi-flow network for frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3521–3531, 2022. 4

[9] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *2015 IEEE international conference on computer vision (ICCV)*, pages 2758–2766, 2015. 4

[10] Jing Gao, Xueqiang Gao, Kaiming Nie, Zhiyuan Gao, and Jiangtao Xu. A deblurring method for indirect time-of-flight depth sensor. *IEEE Sensors Journal*, 23(3):2718–2726, 2023. 1, 2

[11] Qi Guo, Iuri Frosio, Orazio Gallo, Todd Zickler, and Jan Kautz. Tackling 3d tof artifacts through learning and the flat dataset. In *2018 Proceedings of the European Conference on Computer Vision (ECCV)*, pages 368–383, 2018. 1, 2, 6, 7, 8

[12] Felipe Gutierrez-Barragan, Huaijin Chen, Mohit Gupta, Andreas Velten, and Jinwei Gu. itof2dtof: A robust and flexible representation for data-driven time-of-flight imaging. *IEEE Transactions on Computational Imaging*, 7:1205–1214, 2021. 2

[13] John K Haas. A history of the unity game engine. 2014. 2

[14] Xuemei Hu, Xing Lin, Tao Yue, and Qionghai Dai. Multispectral video acquisition using spectral sweep camera. *Optics Express*, 27(19):27088–27102, 2019. 2

[15] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2462–2470, 2017. 4

[16] Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, and Delio Vicini. Dr.jit: A just-in-time compiler for differentiable rendering. *IEEE Transactions on Graphics (TOG)*, 41(4), July 2022. 2

[17] Daniel S Jeon, Andréas Meuleman, Seung-Hwan Baek, and Min H Kim. Polarimetric itof: Measuring high-fidelity depth through scattering media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12353–12362, 2023. 1

[18] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(1):221–231, 2012. 5

[19] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep hdr video from sequences with alternating exposures. In *Computer graphics forum*, volume 38, pages 193–205. Wiley Online Library, 2019. 2

[20] Araya Kongpecth, Natavut Kwankeo, and Visuttha Manthamkarn. 360 degrees object detection using multiple tof sensors for educational robot. In *2022 International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTICON)*, pages 1–4. IEEE, 2022. 1

[21] Seungkyu Lee. Time-of-flight depth camera motion blur detection and deblurring. *IEEE Signal Processing Letters*, 21(6):663–666, 2014. 1, 2

[22] Larry Li et al. Time-of-flight camera—an introduction. *Technical white paper*, (SLOA190B), 2014. 1, 2

[23] Chung Ping Liu, Bo Han Cheng, Pei Ling Chen, and Tsun Ren Jeng. Study of three-dimensional sensing by using inverse square law. *IEEE Transactions on Magnetics*, 47(3):687–690, 2011. 5

[24] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2016. 7

[25] Adithya Pediredla, Ashok Veeraraghavan, and Ioannis Gkioulekas. Ellipsoidal path connections for time-gated rendering. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 2

[26] Zhiyuan Pu, Peiyao Guo, M Salman Asif, and Zhan Ma. Robust high dynamic range (hdr) imaging with complex motion and parallax. In *2020 Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 134–149, 2020. 2

[27] Michael Schelling, Pedro Hermosilla, and Timo Ropinski. Radu: Ray-aligned depth update convolutions for tof data denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 671–680, 2022. 2, 6, 7

[28] Michael Schelling, Pedro Hermosilla, and Timo Ropinski. Weakly-supervised optical flow estimation for time-of-flight.

In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2135–2144, 2023. 1, 2, 3, 6, 7, 8

[29] Lee Streeter. Time-of-flight range image measurement in the presence of transverse motion using the kalman filter. *IEEE Transactions on Instrumentation and Measurement (TIM)*, 67(7):1573–1578, 2018. 1, 2

[30] Jan Stuhmer, Sebastian Nowozin, Andrew Fitzgibbon, Richard Szeliski, Travis Perry, Sunil Acharya, Daniel Cremers, and Jamie Shotton. Model-based tracking at 300hz using raw time-of-flight observations. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3577–3585, 2015. 2

[31] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *2020 Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–419, 2020. 6, 7

[32] Zachary Teed and Jia Deng. Raft-3d: Scene flow using rigid-motion embeddings. In *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 8375–8384, 2021. 1

[33] Sundar Vedula, Simon Baker, Peter Rander, Robert Collins, and Takeo Kanade. Three-dimensional scene flow. In *1999 IEEE International Conference on Computer Vision (ICCV)*, volume 2, pages 722–729, 1999. 1

[34] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 7

[35] Augusto Ronchini Ximenes, Preethi Padmanabhan, Myung-Jae Lee, Yuichiro Yamashita, Dun-Nian Yaung, and Edoardo Charbon. A modular, direct time-of-flight depth sensor in 45/65-nm 3-d-stacked cmos technology. *IEEE Journal of Solid-State Circuits*, 54(11):3203–3214, 2019. 1

[36] Tao Yang, You Li, Cheng Zhao, Dexin Yao, Guanyin Chen, Li Sun, Tomas Krajnik, and Zhi Yan. 3d tof lidar in mobile robotics: A review. *arXiv preprint arXiv:2202.11025*, 2022. 1

[37] Tao Yue, Ming-Ting Sun, Zhengyou Zhang, Jinli Suo, and Qionghai Dai. Deblur a blurred rgb image with a sharp nir image through local linear mapping. In *2014 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2014. 3

[38] Simone Zennaro, Matteo Munaro, Simone Milani, Pietro Zanuttigh, Andrea Bernardi, Stefano Ghidoni, and Emanuele Menegatti. Performance evaluation of the 1st and 2nd generation kinect for multimedia applications. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2015. 1