

Joint Depth and Normal Estimation from Real-world Time-of-flight Raw Data

Rongrong Gao*, Na Fan*, Changlin Li, Wentao Liu, and Qifeng Chen

Abstract—We present a novel approach to joint depth and normal estimation for time-of-flight (ToF) sensors. Our model learns to predict the high-quality depth and normal maps jointly from ToF raw sensor data. To achieve this, we meticulously constructed the first large-scale dataset (named ToF-100) with paired raw ToF data and ground-truth high-resolution depth maps provided by an industrial depth camera. In addition, we also design a simple but effective framework for joint depth and normal estimation, applying a robust Chamfer loss via jittering to improve the performance of our model. Our experiments demonstrate that our proposed method can efficiently reconstruct high-resolution depth and normal maps and significantly outperforms state-of-the-art approaches.

I. INTRODUCTION

Time-of-flight (ToF) cameras are popular depth sensors for robots to obtain dense depth estimation by measuring the time difference between emitting light and receiving the returned light at the camera. ToF cameras have been widely used on robots and smartphones for near-range depth estimation, but their depth maps often have limited resolution and depth accuracy, and no normal map directly is provided by the sensor [16]. In this work, we are interested in estimating high-quality depth and normal maps from ToF raw sensor data, which are fundamental representations of a 3D scene for various robotic tasks such as 3D reconstruction, 3D object detection, and autonomous driving [33], [44]–[46].

Various methods have been proposed for ToF depth enhancement by designing new hardware for sensor data acquisition or post-processing techniques. New hardware coding strategies have been designed to get the high-performance of ToF sensing [23]–[25], [38], but these hardware designs require non-trivial energy and long exposure time, which is not desirable in practice. Deep learning approaches have also been applied directly to derive depth information from ToF raw sensor data [21], [28], [40], [41]. However, these learning based ToF depth sensing approaches are limited to training with simulated synthetic data. To address this issue, we follow a similar data collection protocol by [8] to build a large-scale real-world dataset (*ToF-100*) with paired raw ToF data and ground-truth depth maps for ToF depth sensing.

Our dataset contains 100 scenes captured by a ToF sensor and a high-end industrial depth camera, with ToF raw data at the resolution 240×180 and high-resolution 1280×1024 depth maps from the industrial stereo camera IDS Ensenso N35 [1]. Fig. 1 shows our data capture equipment, ToF raw

data, ToF confidence map, ToF/IDS depth maps, and computed normal maps from depth. Looking at the confidence map, we believe that high-resolution depth and normal maps can be obtained from the ToF sensor as there are fine details in the raw ToF data and its confidence map.

With the collected dataset, we design a new convolutional neural network architecture to jointly estimate the high-resolution depth and normal maps from ToF raw data. The architecture first recovers the initial depth and normal maps from ToF raw data and then refines the estimated depth map by joint refinement module, generating the high-resolution depth map. To achieve the best performance, we utilize a Chamfer loss with jittering that accounts for the imperfect ICP alignment between two point clouds during training. Experiments demonstrate that our approach significantly outperforms current state-of-the-art methods on both depth and normal estimation. In summary, our contribution is threefold:

- We constructed the first large-scale real-world dataset named *ToF-100* with each data sample consisting of ToF raw measurements, confidence map, sparse point cloud, depth map, normal map of resolution 240×180 , and the corresponding ground-truth dense point cloud, depth map and normal map of resolution 1280×1024 . There are 100 scenes in our dataset, and each scene has 25 depth map pairs from different viewpoints. The dataset will be publicly available online soon.
- Based on *ToF-100* dataset, we design a novel end-to-end network model to estimate the high-resolution depth and normal maps from ToF raw data. This is, to the best of our knowledge, the first attempt to generate high-resolution normal maps of ToF sensors directly.
- To compensate for the effects of the misalignment between two point clouds, we introduce a robust Chamfer loss via jittering to account for the imperfect alignment measurement.

II. RELATED WORK

Traditional ToF imaging. Depth maps decoded directly from ToF sensors often suffer from artifacts caused by Multi-Path Inference (MPI). To alleviate these artifacts, researchers have proposed traditional methods with multipath light transport models with different assumptions, including the Lambertian surfaces assumption [13], introducing reflectivity as Cauchy distribution [20], focusing on the photometric cause [14], or extending to general multipath [7], [39].

Then an optimization problem with either iterative optimization [15] or a closed-form solution [11], [20] is formulated to solve for depth estimation.

*Joint first authors. Rongrong Gao, Na Fan, Changlin Li, and Qifeng Chen (cqf@ust.hk) are with the Department of Computer Science and Engineering, HKUST. Wentao Liu is with Sensetime.

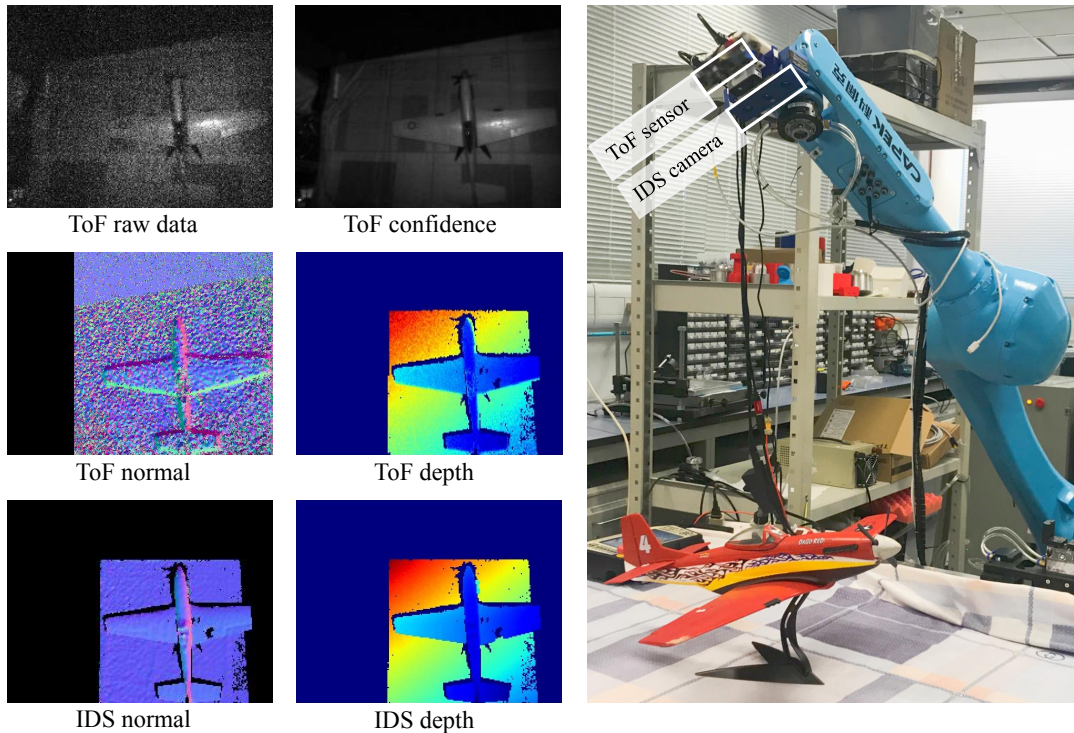


Fig. 1: **The data acquisition process in our *ToF-100* dataset.** The ToF camera and the IDS stereo camera are fixed on a robot arm to acquire the paired data simultaneously. The normal maps and IDS depth map in the lower-left framed part are computed given the point clouds. It can be seen that the ToF camera produces depth and normal maps with limited resolution and accuracy when compared with the IDS camera.

To improve ToF data acquisition, some researchers tried to combine a ToF sensor with structured light [2], [32], then separated disturbing lights in the frequency domain or fused depth obtained from a structured light principle [5]. Similarly, another line of approaches [3], [17], [18], [29], [31] studies depth fusion of a ToF-stereo system, as we can use a high-resolution stereo camera for ground-truth acquisition. To improve the depth resolution, recently, coding functions are redesigned [22], [23] from a hardware perspective. Despite the development of traditional methods, cumulative error and information loss are inevitable due to the hand-crafted pipeline processing.

Learning based ToF imaging. Learning-based methods have recently been applied to ToF imaging [3], [9], [21], [28], [36], [40]. [40] first trained an MPI range-recovery network with a structured light camera for obtaining ground truth. [3] targeted the fusion of ToF and stereo depth by learning a confidence map for local consistency. [28] proposed a two-stage training strategy with an autoencoder to extract general low-level depth features and a decoder to remove MPI. [21] first solved noise, scene motion, and MPI simultaneously. Using ToF raw data as input, [41] realized an end-to-end network to directly output depth maps. However, all these previous methods were only trained on synthetic ToF data due to difficulties in obtaining sufficient real-world data with ground truth.

To bridge the domain gap between synthetic and real ToF measurements, [4] applied unsupervised adversarial learning

on real scenes as a complement to the training on a synthetic dataset, based on a Coarse-Fine CNN [6]. However, as real-world scenes are generally complicated, improving performance on real scenes remains an essential but challenging task. Some other real ToF datasets were presented, either with limited scale [19] or lack of ground truth [36], or the limited resolution of ground-truth depth [9].

Joint depth and normal estimation. To improve depth estimation, researchers have considered joint depth and normal estimation [10], [27], [35], [43], [47]. Early work by [10] was a multi-scale convolutional architecture to predict depth, normal, and semantic maps simultaneously. [43] introduced a dense conditional random field to jointly regularize the four streams, namely depth, normal, planar likelihood, and boundary of a CNN. [35] proposed the two-branch depth-to-normal and normal-to-depth networks, using geometric consistency as supervision. Nevertheless, these frameworks mostly take a single RGB image as input, which would severely degrade the performance in dark or texture-less environments, while ToF data with totally different characteristics are not affected.

To the best of our knowledge, we are the first to use a model that jointly learns high-resolution depth and normal maps from real-world ToF raw sensor data. Compared to past methods, our *ToF-100* dataset is much larger with 2500 real-world images pairs with high-resolution ground-truth depth and raw ToF data. Moreover, trained on this large-scale real-world dataset with raw ToF data as input, our supervised end-to-end training for joint depth and normal enhancement

TABLE I: Comparison between our *ToF-100* dataset and other real-world ToF datasets.

Dataset	Scenes	Raw	ToF Resolution	GT Resolution
[30]	140	no	320×240	No
[19]	3	no	176×144	No
[40]	900	no	320×240	yes
[28]	10	no	200×200	yes
[6]	8	no	320×239	yes (320×239)
[21]	15	yes	320×240	yes
[41]	5	yes	320×240	no
[36]	400	yes	640×480	no
[4]	113	yes	320×239	yes (320×239)
[9]	200	yes	320×240	yes (320×240)
Ours	2500	yes	240×180	yes (1280×1024)

can generalize well to real-world scenes without the need for domain adaptation.

III. DATASETS

There are a few real-world ToF datasets provided in prior work. However, the raw ToF data that contains rich information is unavailable in these dataset [6], [19], [28], [30], [40]. All of them are either limited in scale or without ground truth depth maps. To this end, we propose a real-world dataset of 2,500 image pairs, named *ToF-100* of 100 objects captured by a ToF sensor, for depth and normal enhancement tasks. Our dataset contains raw ToF data, low-resolution 240×180 ToF depth maps, confidence images decoded from the traditional pipeline of the ToF sensor, and the corresponding ground-truth high-resolution depth maps and normal maps. Here is a comparison between our dataset and other real ToF datasets, as shown in Table I.

a) Dataset overview: Our *ToF-100* dataset contains 100 different objects (or object combinations), including models, household objects, and lab materials, with a wide range of scales. The materials vary from plastic, wood, paper to sponges, cloth, rubber, etc., and avoid black, transparent, and reflective objects. Fig. 2 shows some examples of the objects in the *ToF-100* dataset. For each object, we capture 25 scenes by either sending the cameras to different viewing distances and directions or changing the pose of the object. While IDS camera records 1 frame, ToF camera records 10 frames per scene. Therefore, the total amount of data is 100 objects × 25 scenes/object × (1 frame per scene for IDS, 10 frames per scene for ToF) data pairs.

b) Hardware setup and preprocessing: The ToF camera configuration parameters are set to 20MHz and 100MHz double frequency, 400 μ s exposure time, and 30 FPS as its common setting in commercial use. To generate the high-quality ground truth, we use an industrial stereo camera, IDS Ensensio N35 [1], with 1280×1024 resolution, which can output a dense point cloud of high precision in a working distance of up to 3m. The two cameras are then firmly mounted to the end effector of a robot arm, as shown in Fig. 1. Shooting the same scene, they can generate an overlapped 3D data pair of low and high quality, respectively.

To align the data pairs, we introduced an average ICP



Fig. 2: Some examples in the *ToF-100* dataset.

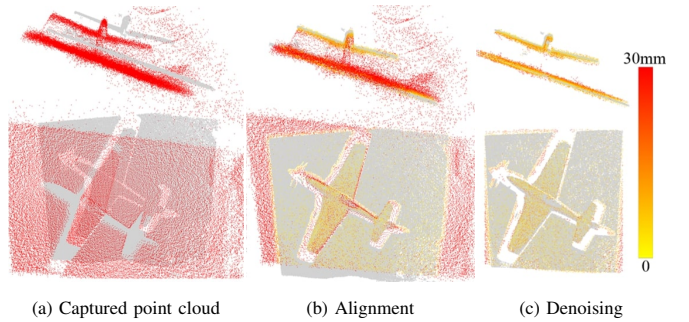


Fig. 3: **Pre-processing steps.** The point cloud pairs collected by the IDS stereo camera (dense in grey) and ToF camera (sparse in red) are aligned by the averaged-ICP method and denoised to obtain clean high-quality ground truth, focusing on the target object.

method to calibrate using point cloud pairs of some object with distinctive geometric features (*e.g.*, board with holes [34]). After alignment, the IDS point clouds are cropped and denoised to obtain high-quality ground truth, focusing on the target objects in the neighborhood of ToF data. Finally, point clouds are projected onto the ToF image plane to generate ground truth depth and masks. Fig. 3 shows the workflow of data pre-processing.

IV. METHOD

Our model for high-resolution joint depth and normal estimation from ToF raw data is illustrated in Fig. 4. The input of our model contains multiple-channel images $\{I_1, I_2, I_3, I_4, I_5, I_6, I_7, I_8\}$ that are decoded from ToF raw data for different phase offsets. As additional input, we also use their corresponding confidence and depth maps obtained from the ToF sensor.

Our model consists of three modules. The first network is designed for estimating an initial high-resolution depth map (*raw2depth*), and the second one is used for normal map estimation (*raw2normal*). Then the two outputs from these two networks will be refined jointly to recover a final high-quality and high-resolution depth map.

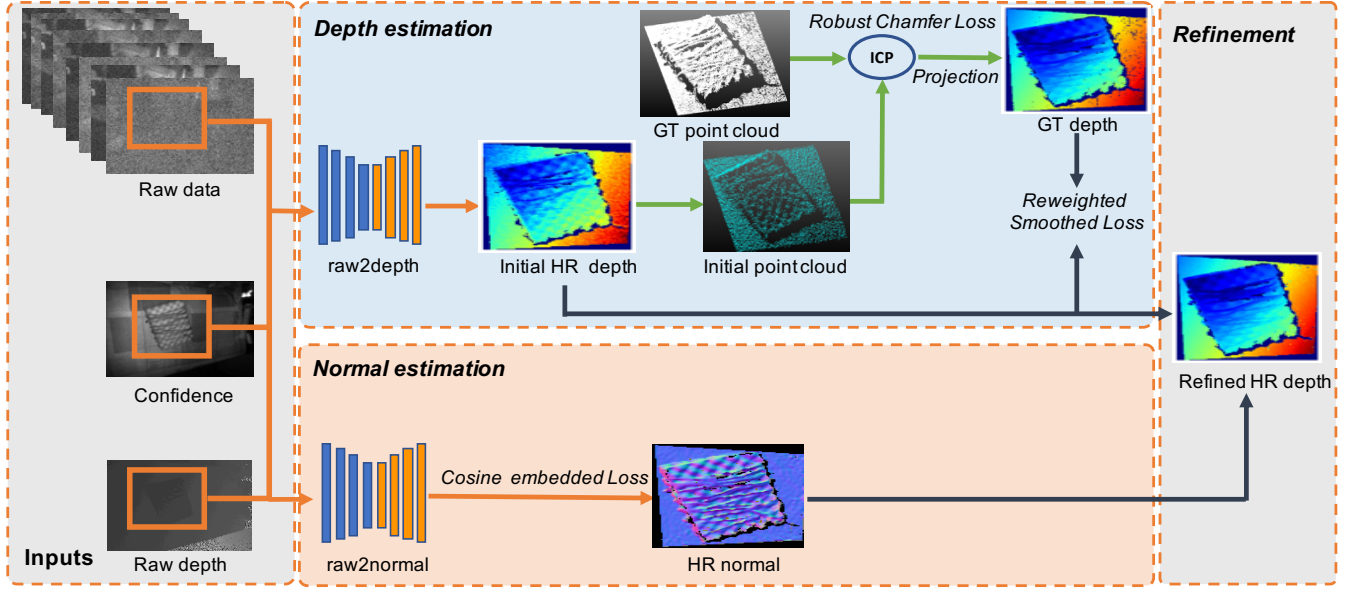


Fig. 4: Our model take 8-channel raw measurements, a confidence map, and a raw depth map as input. During the depth estimation process, an initial depth is first estimated through a *raw2depth* network. Then by projecting the initial depth into an initial point cloud, a robust Chamfer loss and a reweighted smoothness ℓ_1 loss are utilized during training. In parallel, the normal estimation stage generates a HR normal map. At last, refined HR depth is recovered by the joint refinement with estimated depth and normal maps.

A. Raw2depth network

As a straightforward idea, we first train the Raw2depth network end-to-end to generate an initial depth map at high resolution with the IDS depth maps as ground truth. We use a traditional U-Net network [37] as the backbone for the Raw2depth network, whose structure is proven to be efficient in various dense prediction tasks. Different from the structure designed based on physical characters of input channels in [41], our model uses an end-to-end network that can make the best use of high-resolution ground truth. Other kinds of network architectures, such as modified encoder-decoder networks, may also be efficient in improving the performance, which is not the key in this work.

a) *Loss functions*: Due to the imperfect alignment and resolution mismatch between the input raw data and the ground-truth depth map, a simple ℓ_1 or ℓ_2 loss for training may lead to blurry depth estimation. So we design a combined loss in both 2D and 3D domains. In the 2D domain, a smoothed ℓ_1 loss is used to minimize the error between the generated depth d and ground-truth (IDS) depth $\tilde{d}$. In the 3D dimension, we use a robust Chamfer loss to compare the similarity between our point cloud (derived from the generated depth) and the ground-truth point cloud.

b) *Reweighted smoothed ℓ_1 loss*: While our method produces a complete dense depth map, some pixels may have higher confidence than others. Can we quantify this by estimating an error map (difference from the ground-truth) of the estimated depth? We can even use this estimated error map to enhance the depth estimation by re-weighting each pixel. We train a U-net [37] that uses ToF raw data and the estimated initial depth as input to output the expected error map. Then the inverse of the error map is taken as the weighting confidence to further improve the depth

estimation. Suppose $e_p, d_p, \tilde{d}_p$ are the corresponding error, predicted depth and ground truth depth, the reweighted smoothed loss ℓ_{rs} is:

$$\ell_{rs} = \begin{cases} \sum_p \frac{\lambda}{2} \left| \frac{d_p - \tilde{d}_p}{\delta} \right|^2 & \text{if } |d_p - \tilde{d}_p| < \delta \\ \sum_p \lambda \left(\left| \frac{d_p - \tilde{d}_p}{\delta} \right| - \frac{1}{2} \right) & \text{otherwise} \end{cases} \quad (1)$$

where $\lambda = \frac{\tilde{d}_p}{e_p + \epsilon}$, $\epsilon = 0.001$, δ is the threshold and $\delta = 20$. Note that the error map can be also used to indicate the confidence of depth map in practice.

c) *Robust Chamfer loss*: To compare our 3D point cloud and the ground-truth point cloud, we adopt the Chamfer loss that computes the sum of the distance of the nearest neighbor of each point [42]. Let P and Q be the point clouds recovered from our output depth and the ground-truth depth. Assume that x, y are corresponding matched points in each point cloud after the ICP matching process. Then the standard Chamfer loss is

$$\ell_{ch} = \sum_{x \in P} \min_{y \in Q} \|x - y\|_2^2 + \sum_{y \in Q} \min_{x \in P} \|x - y\|_2^2. \quad (2)$$

Robustness via jittering. Although the two point clouds are calibrated as well as possible, slight misalignment still exists, which may cause blurry results in learning-based methods. In order to further alleviate this negative effect by slight misalignment, we move P along the $x/y/z$ dimension for both backward and forward directions by 1 centimeter, resulting in six slightly moved point clouds. We then compute the Chamfer loss between each slightly moved point cloud and Q , resulting in six more Chamfer loss scores. Then we choose the lowest one, which implies the most accurate point cloud, among the seven scores we computed for training.

B. Raw2normal Network

Similarly, we design a normal estimation framework to generate a high-resolution normal map for the ToF sensor, which takes ToF raw data together with the confidence map and ToF sensor depth as input. By training a U-net [37] end-to-end, we learn a model to generate a high-resolution normal map directly.

Cosine embedded loss: To learn normal vectors from ToF raw data, we use the cosine embedded loss ℓ_{cos} to measure the angle similarity between the leaned normal vector $\mathbf{n}_1$ and its ground-truth normal $\mathbf{n}_2$. The cosine embedded loss is formulated as follows:

$$\ell_{cos} = 1 - \frac{\mathbf{n}_1 \cdot \mathbf{n}_2}{\|\mathbf{n}_1\| \|\mathbf{n}_2\|}. \quad (3)$$

Thanks to our *ToF-100* dataset with dense ground-truth point clouds, our work can predict high-resolution normal maps from ToF sensor raw measurements directly.

C. Joint Refinement

There is a strong geometric correlation between depth and the surface normal that a normal map can be calculated from a depth map, and a normal can inversely help refine a depth, as the normal reveals the high-frequency details of the geometry feature. Similar to GeoNet [35], we perform a depth refinement post-processing step that involves the interactive refinement of depth-to-normal and normal-to-depth mapping. We take depth and normal maps provided by the previous modules as the initial inputs and perform the iterative process to refine them.

V. EXPERIMENTS

In this section, we validate our method for the depth and normal maps reconstruction on the *ToF-100* dataset. In order to quantitatively and qualitatively evaluate the performance, we compare our results with other existing approaches, such as [41] and [21]. Moreover, some complex scenes in the wild such as office environments, also are tested to demonstrate the generalization capability of our model. Then an ablation study is provided to study the effects of different modules and inputs in our model.

A. Implementation details

We firstly train the depth and normal estimation network for about 230 and 460 epochs separately, with the learning rate $1e^{-4}$ and Adam optimizer [26]. Then the estimated depth and normal maps are taken as inputs for the refinement module, and the final results are obtained after several epochs. The inference time for our whole framework is 48 ms when running on an Nvidia RTX 2080 Ti GPU. Both the depth and normal estimation models are trained on the input size 240×180 , and we have recovered depth and normal maps both at 2x resolution 480×360 , which is also the resolution for depth and normal evaluation. In principle, we could also reconstruct higher resolution depth and normal maps because the absolute resolution of ground truth is 1080×1024 .

TABLE II: Quantitative evaluation for depth and normal estimation on the *ToF-100* dataset.

Method	Depth				Normal	
	ABS	SQ	RMSE	MAE	MAE	20°
ToF sensor	0.07	53.4	312.9	113.5	0.17	16.3
Su et al. [41]	0.06	46.5	256.4	84.6	0.12	10.5
Guo et al. [21]	0.05	43.1	230.7	84.7	0.11	9.9
Ours	0.03	12.9	242.3	77.9	0.08	4.0

B. Results on depth estimation

We compare the depth map decoded from the ToF traditional pipeline, the method of [41], the method of [21] and our full model. Some visual results are shown in Fig. 5. Compared with the depth map from the ToF traditional pipeline, we can see that our method greatly enhances the depth map quality with more fine-grained details. Comparing the error maps of [41], [21] and our method, our generated depth map also is the most accurate one. Regarding object boundaries, our method alleviates the misalignment effect with the help of robust Chamfer loss.

C. Results on normal estimation

To visually compare with the normal map derived from the ToF depth map, we upsample the original ToF sensor depth map to a high resolution one, and then transform it to a dense point cloud for normal computation using Cloud Compare Software¹. The results are shown in Fig. 6. It is clear that our method produces normal maps with much fewer noises than the one from a traditional pipeline. The normal around the object edges can be better preserved in our results, which motivates us to use it to refine depth map.

D. Quantitative evaluation

In this part, we compare the depth prediction performance with some quantitative metrics: the relative absolute error (ABS), the relative square error (SQ), the root mean square error (RMSE), and the mean absolute error (MAE). The evaluation unit is a millimeter. It can be seen from Table II that our method significantly improves the quality of the generated depth maps. Both visual and quantitative results can further confirm our conclusion. In summary, our method achieves an average 3% relative error in depth estimation, which is much better than traditional methods.

Moreover, we evaluate the predicted normal maps with two metrics: the mean absolute error (MAE) and the average per-pixel angle distance between the prediction and ground-truth normals less than 20° as used in [12]. As shown in Table II, our method significantly outperforms state-of-the-art methods on ToF depth estimation.

E. Analysis

Ablation study. For the analysis, we first conduct the ablation study to evaluate the effectiveness of each module of our approach, including multiple data source input (confidence map and decoded depth map), the proposed Chamfer loss,

¹<https://cloudcompare.org>

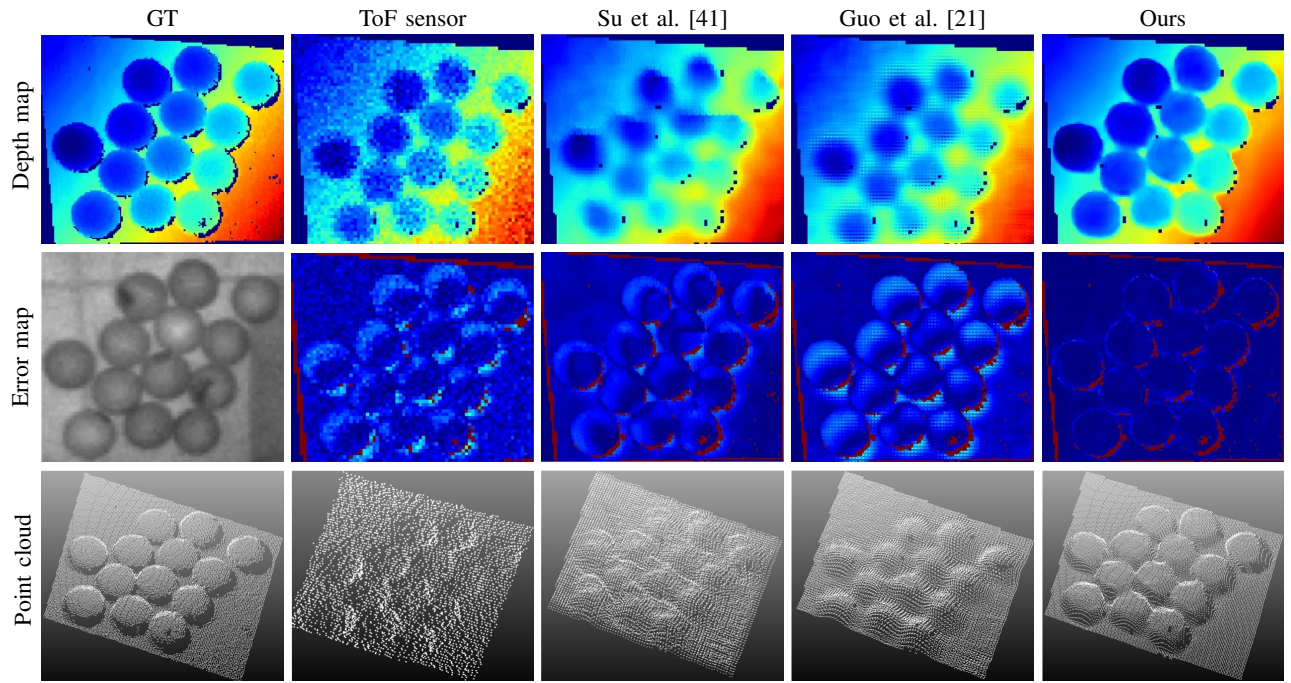


Fig. 5: **Depth estimation results on the *ToF-100* dataset.** From left to right are the depth maps of ground truth, the ToF sensor, [41], results by [21], and our proposed model. The second row below each depth map indicates its corresponding error map(the first one is the confidence map). The third row shows the corresponding estimated point clouds. More results are shown in the supplement.

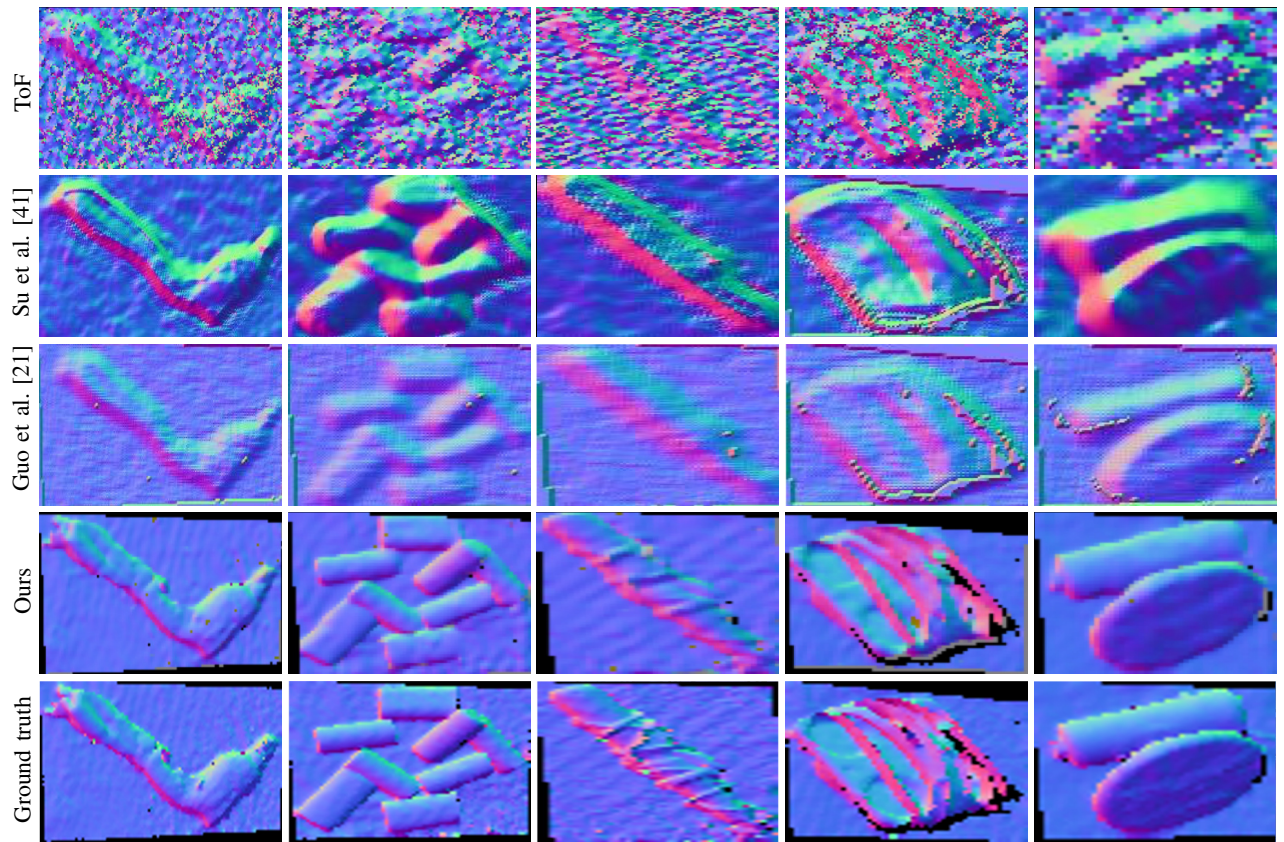


Fig. 6: **Normal estimation results on the *ToF-100* dataset.** The first to fifth rows shows the normal maps derived from the ToF sensor, Su et al. [41], Guo et al [21], our method and the ground truth.

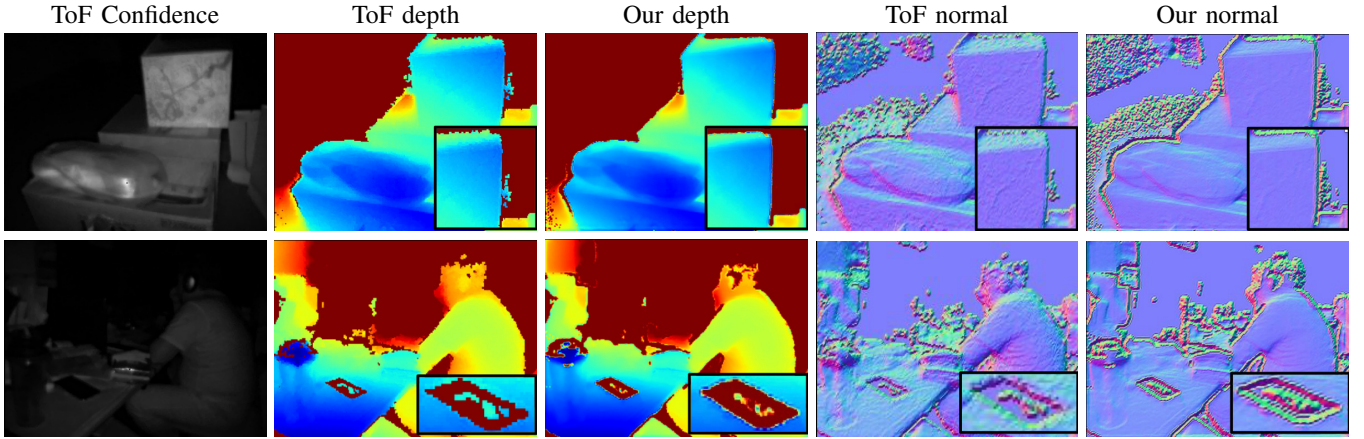


Fig. 7: **Depth and normal estimation in complex indoor scenes.** From left to right are the confidence map, depth map from the ToF sensor, our depth map, ToF normal map, and our normal map. The right bottom corners are zoomed-in regions.

TABLE III: Ablation study for analyzing different modules of our model in depth estimation.

Method	ABS	SQ	RMSE	MAE
w/o multi-input	0.06	44.0	277.7	103.7
w/o Chamfer loss	0.04	14.9	244.9	80.3
w/o error map	0.04	13.7	243.9	80.1
w/o refinement	0.03	13.9	244.8	79.3
Ours (full model)	0.03	12.9	242.3	77.9

TABLE IV: Experiments with different input data combinations.

Method	ABS	SQ	RMSE	MAE
Conf.	0.13	70.9	338	218.3
Raw	0.06	22.1	278.5	95.3
Depth	0.05	18.8	269.1	87.2
Depth + Raw	0.03	13.1	245.9	78.4
Depth + Conf. + Raw	0.03	12.9	242.3	77.9

error map, and the normal map-based refinement. As shown in Table III, the proposed modules significantly improve the accuracy of depth estimation, showing their respective effectiveness. Even only adopting a simple U-Net framework as the backbone, our method is able to recover a high-resolution and high-quality depth map.

Besides, to demonstrate the effects of multiple inputs sources. We compare the results of 5 kinds of input combinations, and the qualitative results are shown in Table IV. From the results, we can find that even based on raw data only, our approach can recover the depth map, which indicates that an end-to-end deep learning based method can replace the traditional depth reconstruction pipeline. Among all these results, the result with all the raw data, decoded depth maps, and confidence maps generates the best result, indicating that decoding cues from the traditional method can help to guide the depth estimation from ToF raw measurements.

MPI removal analysis. Multi-path interference causes distortion in the recovered ToF depth map, which is inherent to the working principle of extracting depth from raw phase-shifted measurements with respect to emitted modulated

infrared signals. Traditional methods usually explicitly model the correction of the MPI error to further enhance the results, which is not suitable for real-world scenes. In our proposed method, we take double-frequency ToF raw measurements as inputs, each of them is a 4-channel data with different phase shifts, from which indirect light may help improve the performance by leveraging additional sources of information. Moreover, the ground-truth depth captured from the IDS camera has little MPI error because it is a stereo-based camera. These ideas help to compensate for the MPI effect, and the visual results show that our estimated high-resolution depth map preserves many details.

Noise removal analysis. In the traditional pipeline of ToF depth recovering, denoising is usually based on arbitrary rules and assumptions, which often lose effectiveness with changes in intensity and scenes of the received signals. This causes serious noise in areas with low reflection for weak input signals. In contrast to conventional methods, our proposed algorithm provides a simple but strong learning framework to translate the noisy ToF raw data to the high-quality depth map, avoiding the above strict assumptions. From Fig. 7, we can see that our estimated depth maps are much smoother than ToF depth maps in the plane areas. Two ways are used to remove the effect of noise - we take 10 shots of the raw measurements for each scene to reduce the shot noise, and an average depth of the 10 shots is used as ground truth to alleviate random noise.

F. Results in complex indoor scenes

Our *ToF-100* dataset primarily focuses on single or multiple objects, which are relatively small in scale. Here we validate the generalization capability of our method on some other complex indoor scenes, including office room and laboratory. As can be seen in Fig. 7, our method also provides competitive visual results performance, demonstrating its generalization capability.

VI. CONCLUSION

In this paper, we collect the first large-scale real-world *ToF-100* dataset consisting of ToF sensor raw measurements,

confidence map, sparse point cloud, normal map, depth map, and the corresponding ground-truth. Based on *ToF-100*, we design a novel learning framework to jointly estimate the high-resolution depth and normal maps from ToF raw data. Experiments show that our approach significantly outperforms the current state-of-the-art methods and achieves outstanding performance on our *ToF-100* dataset for the estimation of depth and normal maps jointly.

REFERENCES

- [1] 3D Camera Ensenso N35. <https://en.ids-imaging.com/ensenso-n35.html>. 1, 3
- [2] Supreeth Achar, Joseph R. Bartels, William L. Whittaker, Kiriakos N. Kutulakos, and Srinivasa G. Narasimhan. Epipolar time-of-flight imaging. *ACM Trans. Graph.*, 36(4):37:1–37:8, 2017. 2
- [3] Gianluca Agresti, Ludovico Minto, Giulio Marin, and Pietro Zanuttigh. Deep learning for confidence information in stereo and tof data fusion. In *ICCV Workshops*, 2017. 2
- [4] Gianluca Agresti, Henrik Schäfer, Piergiorgio Sartor, and Pietro Zanuttigh. Unsupervised domain adaptation for tof data denoising with adversarial learning. In *CVPR*, 2019. 2, 3
- [5] Gianluca Agresti and Pietro Zanuttigh. Combination of spatially-modulated tof and structured light for mpi-free depth estimation. In *ECCV Workshops*, 2018. 2
- [6] Gianluca Agresti and Pietro Zanuttigh. Deep learning for multi-path error removal in tof sensors. In *ECCV Workshops*, 2018. 2, 3
- [7] Ayush Bhandari, Achuta Kadambi, Refael Whyte, Christopher Barsi, Micha Feigin, Adrian Dorrington, and Ramesh Raskar. Resolving multipath interference in time-of-flight imaging via modulation frequency diversity and sparse regularization. *Optics letters*, 39(6):1705–1708, 2014. 1
- [8] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *CVPR*, 2018. 1
- [9] Yan Chen, Jimmy S. J. Ren, Xuanye Cheng, Keyuan Qian, Luyang Wang, and Jinwei Gu. Very power efficient neural time-of-flight. In *WACV*, 2020. 2, 3
- [10] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, 2015. 2
- [11] Micha Feigin, Ayush Bhandari, Shahram Izadi, Christoph Rhemann, Mirko Schmidt, and Ramesh Raskar. Resolving multipath interference in kinect: An inverse problem approach. *IEEE Sensors*, 16(10):3419–3427, 2015. 1
- [12] David F Fouhey, Abhinav Gupta, and Martial Hebert. Data-driven 3d primitives for single image understanding. In *ICCV*, 2013. 5
- [13] Stefan Fuchs. Multipath interference compensation in time-of-flight camera images. In *ICPR*, 2010. 1
- [14] Stefan Fuchs, Michael Suppa, and Olaf Hellwich. Compensation for multipath in tof camera measurements supported by photometric calibration and environment integration. In *ICVS*, 2013. 1
- [15] David Fuentes-Jiménez, Daniel Pizarro, Manuel Mazo, and Sira E. Palazuelos. Modeling and correction of multipath interference in time of flight cameras. *Image Vis. Comput.*, 32(1):1–13, 2014. 1
- [16] Peter Fürsattel, Simon Placht, Michael Balda, Christian Schaller, Hannes G. Hofmann, Andreas K. Maier, and Christian Riess. A comparative error analysis of current time-of-flight sensors. *IEEE Trans. Computational Imaging*, 2(1):27–41, 2016. 1
- [17] Vineet Gandhi, Jan Cech, and Radu Horaud. High-resolution depth maps based on tof-stereo fusion. In *ICRA*, 2012. 2
- [18] Yuan Gao, Sandro Esquivel, Reinhard Koch, and Joachim Keinert. A novel self-calibration method for a stereo-tof system using a kinect V2 and two 4k gopro cameras. In *3DV*, 2017. 2
- [19] Valeria Garro, Carlo Dal Mutto, Pietro Zanuttigh, and Guido M. Cortelazzo. Edge-preserving interpolation of depth data exploiting color information. *Ann. des Télécommunications*, 68(11-12):597–613, 2013. 2, 3
- [20] John P. Godbaz, Michael J. Cree, and Adrian A. Dorrington. Closed-form inverses for the mixed pixel/multipath interference problem in AMCW lidar. In *Computational Imaging*, 2012. 1
- [21] Qi Guo, Iuri Frosio, Orazio Gallo, Todd E. Zickler, and Jan Kautz. Tackling 3D ToF artifacts through learning and the FLAT dataset. In *ECCV*, 2018. 1, 2, 3, 5, 6
- [22] Mohit Gupta, Andreas Velten, Shree K. Nayar, and Eric Breibach. What are optimal coding functions for time-of-flight imaging? *ACM Trans. Graph.*, 37(2):13:1–13:18, 2018. 2
- [23] Felipe Gutierrez-Barragan, Syed Azer Reza, Andreas Velten, and Mohit Gupta. Practical coding function design for time-of-flight imaging. In *CVPR*, 2019. 1, 2
- [24] Jungong Han, Ling Shao, Dong Xu, and Jamie Shotton. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE Trans. Cybern.*, 43(5):1318–1334, 2013. 1
- [25] Achuta Kadambi, Refael Whyte, Ayush Bhandari, Lee V. Streeter, Christopher Barsi, Adrian A. Dorrington, and Ramesh Raskar. Coded time of flight cameras: sparse deconvolution to address multipath interference and recover time profiles. *ACM Trans. Graph.*, 32(6):167:1–167:10, 2013. 1
- [26] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 5
- [27] Bo Li, Chunhua Shen, Yuchao Dai, Anton van den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *CVPR*, 2015. 2
- [28] Julio Marco, Quercus Hernandez, Adolfo Muñoz, Yue Dong, Adrián Jarabo, Min H. Kim, Xin Tong, and Diego Gutierrez. Deeptof: off-the-shelf real-time correction of multipath interference in time-of-flight imaging. *ACM Trans. Graph.*, 36(6):219:1–219:12, 2017. 1, 2, 3
- [29] Giulio Marin, Pietro Zanuttigh, and Stefano Mattoccia. Reliable fusion of ToF and stereo depth driven by confidence measures. In *ECCV*, 2016. 2
- [30] Simone Milani and Giancarlo Calvagno. Joint denoising and interpolation of depth maps for ms kinect sensors. In *ICASSP*, 2012. 3
- [31] Carlo Dal Mutto, Pietro Zanuttigh, and Guido Maria Cortelazzo. Probabilistic tof and stereo data fusion based on mixed pixels measurement models. *PAMI*, 37(11):2260–2272, 2015. 2
- [32] Nikhil Naik, Achuta Kadambi, Christoph Rhemann, Shahram Izadi, Ramesh Raskar, and Sing Bing Kang. A light transport model for mitigating multipath interference in time-of-flight sensors. In *CVPR*, 2015. 2
- [33] Trong-Nguyen Nguyen, Huu-Hung Huynh, and Jean Meunier. 3d reconstruction with time-of-flight depth camera and multiple mirrors. *IEEE Access*, 6:38106–38114, 2018. 1
- [34] Jaesik Park, Hyeonwoo Kim, Yu-Wing Tai, Michael S. Brown, and In-So Kweon. High quality depth map upsampling for 3D-TOF cameras. In *ICCV*, 2011. 3
- [35] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. GeoNet: Geometric neural network for joint depth and surface normal estimation. In *CVPR*, 2018. 2, 5
- [36] Di Qiu, Jiahao Pang, Wenxiu Sun, and Chengxi Yang. Deep end-to-end alignment and refinement for time-of-flight RGB-D module. In *ICCV*, 2019. 2, 3
- [37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4, 5
- [38] Michael Schober, Amit Adam, Omer Yair, Shai Mazor, and Sebastian Nowozin. Dynamic time-of-flight. In *CVPR*, 2017. 1
- [39] Daniel Freedman Yoni Smolin, Eyal Krupka, Ido Leichter, and Mirko Schmidt. SRA: fast removal of general multipath for tof sensors. In *ECCV*, 2014. 1
- [40] Kilho Son, Ming-Yu Liu, and Yuichi Taguchi. Learning to remove multipath distortions in time-of-flight range images for a robotic arm setup. In *ICRA*, 2016. 1, 2, 3
- [41] Shuochen Su, Felix Heide, Gordon Wetzstein, and Wolfgang Heidrich. Deep end-to-end time-of-flight imaging. In *CVPR*, 2018. 1, 2, 3, 4, 5, 6
- [42] Arasanathan Thayananthan, Bjoern Stenger, Philip H. S. Torr, and Roberto Cipolla. Shape context and chamfer matching in cluttered scenes. In *CVPR*, 2003. 4
- [43] Peng Wang, Xiaohui Shen, Bryan C. Russell, Scott Cohen, Brian L. Price, and Alan L. Yuille. SURGE: surface regularized geometry estimation from a single image. In *NeurIPS*, 2016. 2
- [44] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark E. Campbell, and Kilian Q. Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *CVPR*, 2019. 1
- [45] Jiaxin Xie, Chenyang Lei, Zhuwen Li, Li Erran Li, and Qifeng Chen. Video depth estimation by fusing flow-to-depth proposals. In *IROS*, 2020. 1
- [46] Kai Zhang, Jiaxin Xie, Noah Snavely, and Qifeng Chen. Depth sensing beyond lidar range. In *CVPR*, 2020. 1
- [47] Zhenyu Zhang, Zhen Cui, Chunyan Xu, Yan Yan, Nicu Sebe, and Jian Yang. Pattern-affinitive propagation across depth, surface normal and semantic segmentation. In *CVPR*, 2019. 2