# Exploiting Dual-Correlation for Multi-frame Time-of-Flight Denoising

Guanting Dong, Yueyi Zhang✉, Xiaoyan Sun, and Zhiwei Xiong

MoE Key Laboratory of Brain-inspired Intelligent Perception and Cognition,
University of Science and Technology of China
gtdong@mail.ustc.edu.cn, {zhyuey, sunxiaoyan, zwxiong}@ustc.edu.cn

**Abstract.** Recent advancements in Time-of-Flight (ToF) depth denoising have achieved impressive results in removing Multi-Path Interference (MPI) and shot noise. However, existing methods only utilize a single frame of ToF data, neglecting the correlation between frames. In this paper, we propose the first learning-based framework for multi-frame ToF denoising. Different from existing methods, our framework leverages the correlation between neighboring frames to guide ToF noise removal with a confidence map. Specifically, we introduce a Dual-Correlation Estimation Module, which exploits both intra- and inter-correlation. The intra-correlation explicitly establishes the relevance between the spatial positions of geometric objects within the scene, aiding in depth residual initialization. The inter-correlation discerns variations in ToF noise distribution across different frames, thereby locating the regions with strong ToF noise. To further leverage dual-correlation, we introduce a Confidence-guided Residual Regression Module to predict a confidence map, which guides the residual regression to prioritize the regions with strong ToF noise. The experimental evaluations have consistently shown that our framework outperforms existing ToF denoising methods, highlighting its superior performance in effectively reducing strong ToF noise. The source code is available at https://github.com/gtdong-ustc/multi-frame-tof-denoising.

**Keywords:** Dual-correlation · ToF denoising · Multi-frame

## 1 Introduction

Time-of-Flight (ToF) cameras capture depth images over long distances and are commonly used across various applications. Among different types of ToF cameras, indirect ToF (iToF) cameras continually illuminate the scene with a periodically modulated light signal and aim to ascertain the phase offset between the emitted and received signal, providing information about the signal's travel time. iToF cameras stand as the most prevalent in the market, and are the primary focus of this paper. Despite the prevalence of ToF cameras in the market, they continue to struggle with noise, which hinders their performance in high-level tasks [9,25,26,28,35,46,47]. The ToF camera noise encompasses two primary types: shot noise and MPI. Shot noise constitutes a form of random noise that is

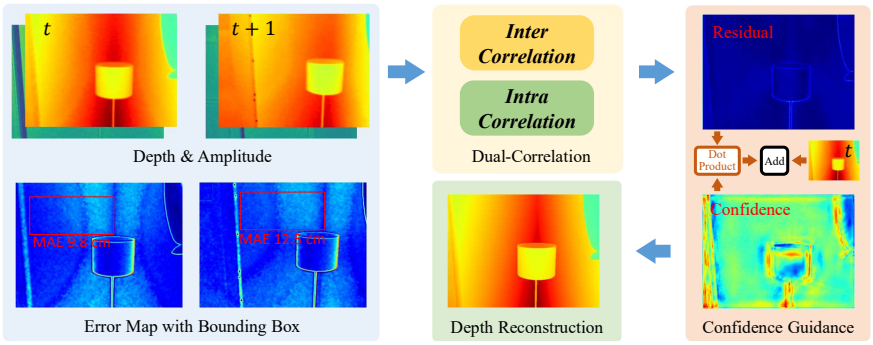


**Fig. 1:** The overview of our proposed multi-frame ToF denoising framework.

pervasive in nearly all depth sensors and arises from the electronic accumulation of received light signals in the sensor [14, 45]. MPI is a noise associated with the geometric structure of the scene, stemming from the multi-bounce reflection of the received light signal during the exposure time. Several previous studies have attempted to eliminate MPI for ToF depth images [3, 7, 21, 27, 30, 31, 37]. However, current works only apply single-frame processing and ignore utilizing the correlation from multi-frame depth images to improve performance during ToF denoising.

In this paper, our objective is to harness the correlation inherent in multi-frame ToF depth images. Referring to existing methods, we find that multi-frame processing frameworks in other tasks [34, 38, 41, 42], both flow-based and flow-free, typically involve three steps: *Alignment*, *Fusion*, and *Reconstruction*. However, given the particular challenges present in multi-frame ToF denoising, existing frameworks are not directly applicable. This is because the presence of MPI leads to the spatial distribution of ToF noise to be *regional*. As shown in the error map of Fig. 1, ToF noise is regionally distributed in the depth image space. Therefore, an explicit alignment and fusion process struggles to capture additional geometric information from neighboring frames, leading to poor performance for multi-frame ToF denoising. By observing continuous ToF data, we discover that changes in camera perspectives can cause a notable difference in the spatial distribution of ToF noise. This phenomenon is clearly demonstrated in Fig. 1, where the red box highlights the area exhibiting an increase in the Mean Absolute Error (MAE) due to a shift in the camera perspective. This special property of ToF noise offers the potential to predict its spatial distribution. Based on this observation, we introduce a novel framework specifically tailored for multi-frame ToF denoising, including *Correlation*, *Guidance* and *Reconstruction*. Our framework stands apart from existing ones by not incorporating an explicit alignment and fusion process. Instead, we aggregate temporal information, specifically the inter-correlation between two neighboring frames, to predict a confidence map that guides the denoising process to focus on removing strong ToF noise.

Based on our proposed framework shown in Fig. 1, we design a Multi-frame ToF Denoising Network (MTDNet), which takes in two frames of amplitude images and ToF depth images, and generates a denoised depth image. Moreover, we introduce a Dual-Correlation Estimation Module (DCEM) to estimate the intra- and inter-correlation, respectively. The intra-correlation explicitly establishes the relationship between the spatial positions of geometric objects within the scene, which is beneficial for regressing an initial depth residual to remove MPI. The inter-correlation is calculated to capture the variations in the spatial distribution of ToF noise across multiple frames, which aids our MTDNet in the removal of strong ToF noise, such as MPI at edges and in corner areas. In addition, we propose a Confidence-guided Residual Regression Module (CRRM) as the backbone of our MTDNet. Building upon the capabilities of dual-correlation, the CRRM predicts a confidence map, which guides residual regression to prioritize the regions with strong ToF noise. Through deploying multiple CRRMs at different scales, our MTDNet constructs a confidence-guided residual pyramid in a coarse-to-fine manner. Finally, we introduce a residual channel attention block [36] and a depth refinement module [7] to combine the residual pyramid together, producing a comprehensive depth residual that enables accurate removal of noise in ToF depth images, particularly MPI and shot noise. In brief, our contributions can be summarized as follows:

- We first propose a multi-frame processing framework for ToF denoising that estimates dual-correlation between neighboring frames to guide the ToF noise removal with a confidence map.
- We propose a Dual-Correlation Estimation Module, including the inter- and intra-correlations, to initialize the MPI estimation and capture the variations in ToF noise distribution across multiple frames.
- We propose a Confidence-guided Residual Regression Module to utilize the dual-correlation, obtaining a confidence map that guides residual regression to focus on the regions with strong ToF noise.
- Our proposed MTDNet outperforms existing methods in the quantitative and qualitative comparisons for ToF denoising on both synthetic and realistic datasets.

## 2 Related Work

The sensitivity of ToF imaging to both shot noise and MPI has been extensively documented in prior research [17,44]. While shot noise is ubiquitous in all sensors and originates from sensor electronics, it has been the subject of extensive study within the context of ToF sensors. Traditional filtering techniques, such as bilateral filtering, have demonstrated effectiveness in eliminating shot noise [2,18]. In contrast, MPI removing poses a more intricate challenge in ToF denoising.

**Transient imaging-based ToF denoising.** Conventional indirect ToF systems extract temporal frequencies from the Fourier Transform of the transient image of the scene [20]. It is essential to capture a wide range of frequencies to

accurately determine the depth from the transient image [5]. Gupta *et al.* [12] explored the impact of modulation frequencies on MPI and proposed a phasor imaging technique involving the emission of two signals with significantly differing frequencies. Freedman *et al.* [10] introduced a model utilizing a compressible backscattering representation to tackle challenges associated with multipath scenarios involving more than two paths, to achieve real-time processing speed. Buratto *et al.* [6] predicted the intensity and arrival time of the first two peaks of the impulse response by assuming that the direct reflection reaches the ToF sensor early.

**Learning-based ToF denoising.** Recently, learning-based techniques have significantly advanced the field of 3D information processing, leading to the development of numerous learning-based MPI removal methods. Marco *et al.* [15] simulated the light transport model of MPI in the ToF imaging process and generated a substantial dataset for ToF denoising. Additionally, they introduced the first two-stage deep neural network to refine the coarse estimation of ToF depth images [21]. Su *et al.* [32] proposed a deep end-to-end network that directly inputs raw correlation measurements. Guo *et al.* [11] offered a suite of advanced transient rendering tools and created a large-scale ToF dataset called FLAT. They also innovatively applied the divide and conquer concept, designing a residual-based U-Net [29] and a kernel prediction network [4, 22] for the removal of MPI and shot noise, respectively. Agresti *et al.* [2] devised an adversarial learning strategy to address the domain shift between unlabeled realistic scenes and synthetic training datasets, employing a generative adversarial network for unsupervised domain adaptation. Qiu *et al.* [27] proposed a deep end-to-end network for camera alignment and ToF depth refinement, explicitly leveraging corresponding RGB images provided by the RGB-D camera. Dong *et al.* [7] focused on leveraging the scene's spatial hierarchical structure by constructing a depth residual pyramid with multiple scales. Gutierrez *et al.* [13] introduced the iToF2dToF method, which generates interpolated frequency measurements to estimate dToF images, providing an alternative output representation. The dToF representation benefits from the denoising abilities of the data-driven model and aids in removing MPI by separating direct and indirect illumination. RADU [30] extended 2D ToF data denoising to 3D and employed 3D point CNNs for ray-aligned depth updating.

**Multi-frame depth processing.** In practical scenarios, depth processing is commonly conducted on multiple images rather than a single frame. Many prior methods have focused on extracting multi-view geometry from monocular RGB videos or on self-supervised depth estimation [8, 24, 39, 40, 43]. Meanwhile, the efficient utilization of inter-correlation has yet to be explored. Patil *et al.* [24] employed a ConvLSTM structure to fuse concatenated frames without alignment. Li *et al.* [19] explicitly aligned multiple frames using a pre-trained scene flow estimator in a stereo video. Sun *et al.* [34] devised a dToF video super-resolution framework with a more flexible and error-tolerant multi-frame alignment to better leverage multi-frame correlations. However, the performance of these methods

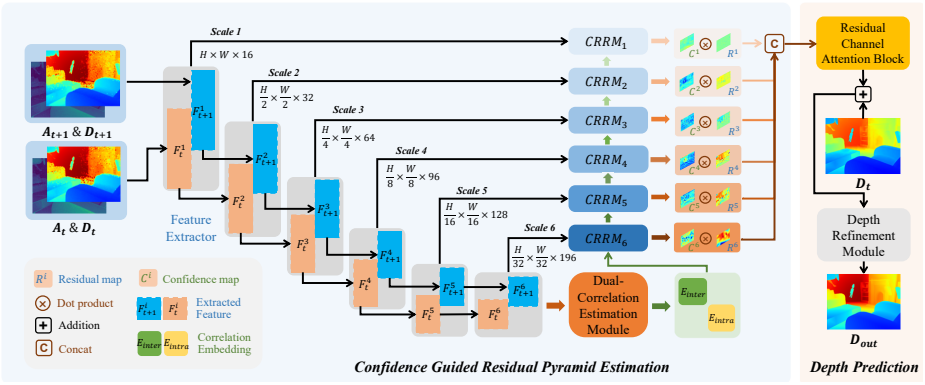


**Fig. 2:** The architecture of our proposed MTDNet. Here, $CRRM_i$ represents the confidence-guided residual regression module at the pyramid network's $i^{th}$ scale.

is primarily constrained by the inefficient or inaccurate multi-frame alignment module.

## 3 Method

### 3.1 Multi-Path Interference Model

In the case of an AMCW iToF sensor, the received light signal $r(t)$ is not solely from direct reflection but rather a combination of both directly and indirectly received signals. Let $\hat{r}(t)$ represent the direct light signal, and $r_p(t)$ denote the indirectly received signals that undergo multiple bounces before being captured by the camera. Thus, the received signal $r(t)$ can be modelled as $r(t) = \hat{r}(t) + \int_{p \in P} r_p(t)$, where $P$ represents the set of all light paths followed by the indirectly received signals. The difference between $r(t)$ and $\hat{r}(t)$ introduces a deviation in ToF depth, commonly known as MPI.

### 3.2 Network Overview

Our proposed MTDNet consists of two phases, as shown in Fig. 2. The first phase, named Confidence-guided Residual Pyramid Estimation, consists of three modules: a weight-sharing Feature Extractor that provides multi-scale features from the concatenation of depth and amplitude maps, a dual-correlation estimation module for estimating the intra- and inter-correlation, and a confidence-guided residual regression module serving as the backbone for predicting multi-scale residuals. The second phase, Depth Prediction, consists of two modules: a residual channel attention block to obtain a coarse depth residual for removing MPI and shot noise, and a depth refinement module to eliminate the rest of the noise (mainly shot noise). The following subsections explain these five modules respectively.

### 3.3 Confidence-guided Residual Pyramid Estimation

**Feature Extractor.** To achieve stable feature representations, we construct the feature extractor with $3 \times 3$ convolutions and residual channel attention blocks [36]. And then, the feature extractor is utilized to extract two multi-scale feature pyramids, denoted as $\{F_t^i\}_{i=1}^L$ and $\{F_{t+1}^i\}_{i=1}^L$, from the combination of depth and amplitude images, *i.e.*, $D_t\&A_t$ and $D_{t+1}\&A_{t+1}$. Here, $L$ denotes the total number of levels in the pyramid. At the $i^{th}$ level, $F_t^i$ represents the feature map extracted from the $t^{th}$ frame and pooled $i$ times. We denote the size of input images as $W \times H$. Therefore, the dimension of the feature map at the $i^{th}$ level is $\frac{W}{2^{i-1}} \times \frac{H}{2^{i-1}} \times CN_i$, where $CN_i$ signifies the number of output channels. We fix $L = 6$ in our network to ensure optimal denoising performance. The corresponding values for $CN_i$ are 16, 32, 64, 96, 128, and 192, respectively. The feature pyramid performs a hierarchical encoding of geometric information, progressing from simpler to more complex scene structures.

**Dual-Correlation Estimation Module.** For constructing dual-correlation, we introduce a cost volume layer [33] that utilizes the extracted features to construct the cost volume as shown in Fig. 3(a). The cost volume stores the matching costs for associating a pixel with its corresponding pixels at next frame as follows:

$$cost(x_t) = \frac{1}{d^2}(F_t^6(x_t))^T F_{t+1}^6(x_t^{'}), \tag{1}$$

where $x_t$ and $x_t^{'}$ represent coordinates and its neighboring ones in pixel space, $T$ is the transpose operator. We use a limited range of $d$ pixels to compute the cost volume, *i.e.*, $|x_t - x_t^{'}|_\infty \leq d$. The dimension of the 3D cost volume is $d^2 \times \frac{H}{32} \times \frac{W}{32}$, where we set $d = 3$. Then, we calculate the intra-cost, which refers to the matching cost of the relative correspondences for the geometric objects in the $t^{th}$ frame. In addition, we define the inter-cost as the matching cost between the features of $t^{th}$ and $(t+1)^{th}$ frames, *i.e.*, $F_t^6$ and $F_{t+1}^6$. Next, we use the combination of the inter-cost, $F_t^6$ and $F_{t+1}^6$ as inputs for two $3 \times 3$ convolutions to generate the feature embedding of the inter-correlation called $E_{inter}$. Similarly, we combine the intra-cost and $F_t^6$ as inputs for two $3 \times 3$ convolutions to produce the feature embedding of intra-correlation known as $E_{intra}$.

**Confidence-guided Residual Regression Module (CRRM).** At each level, we introduce a CRRM, as illustrated in Fig. 3(b), to regress a depth residual map and a confidence map. Firstly, the depth residual map and the confidence map from the lower level denoted as $R^{i+1}$ and $C^{i+1}$, respectively, are upsampled by a factor of 2 using bicubic interpolation. It is then concatenated with the feature map $F_t^i$ of the $t^{th}$ frame and the upsampled feature embedding of the dual-correlation at the current level. The resulting concatenated feature is fed into five sequential convolutional layers, which output the residual map $R^i$ for the current level. Similarly, we concatenate the confidence map with the feature maps extracted from the $t^{th}$ and $(t+1)^{th}$ frames, along with the feature embedding of inter-correlation at the current level. Again, the concatenated feature is passed through five sequential convolutional layers, resulting in the residual map $R^i$ for the current level. For the bottom level, the input of CRRM consists only of the

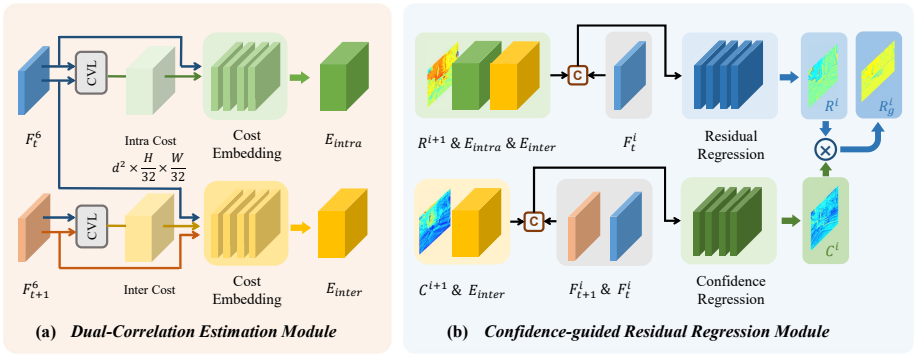


**Fig. 3:** The architecture of our proposed Dual-Correlation Estimation Module and Confidence-guided Residual Regression Module. Here, $R_g^i$ denotes the confidence-guided depth residual map at $i^th$ level.

feature map with a size of $\frac{W}{32} \times \frac{H}{32} \times 192$, as there are no depth residual map and confidence map from the lower level. Finally, we obtain a residual pyramid $\{R^i\}_{i=1}^L$ consisting of depth residual maps with different scales, and a confidence pyramid $\{C^i\}_{i=1}^L$ containing corresponding confidence maps.

### 3.4 Depth Prediction

It is worth noting that information from the lower levels of the pyramid can be lost during the layer-by-layer convolution and upsampling operations. We draw inspiration from SHARP-Net to address this concern and introduce a residual pyramid combination operation that explicitly concatenates the confidence-guided depth residual maps from all scales [7]. We first upsample the confidence-guided depth residual to the original resolution using bicubic interpolation to achieve this. These upsampled depth residual maps are then concatenated to obtain a concatenated residual volume that serves as the input for a channel attention block. Following the convolutional operation, we obtain a coarse depth residual map. We add this depth residual map to the original input depth image to recover the depth image, resulting in the reconstructed depth image.

Following the existing processing steps, MPI noise is significantly reduced. Although shot noise is also mitigated to some extent, it is not as effectively removed as MPI. Shot noise still poses a challenge to applying ToF depth sensing. To tackle this issue, we introduce a depth refinement module that leverages a kernel prediction network to output a final denoised ToF depth image denoted as $D_{out}$ [4, 22]. Details about the kernel prediction network can be found in the supplementary material. This module is crucial in effectively removing shot noise and refining the depth images, enabling enhanced accuracy and quality.

### 3.5 Loss Function

To train our proposed MTDNet effectively, we compare the predicted depth image $D_{out}$ with the ground truth depth image $D_{gt}$ by calculating their differences. Our main goal is to remove depth noise accurately while preserving important geometric details. The loss function we utilized, inspired by [27], consists of two components: the $L_1$ loss and the gradients of the refined depth image. The loss function can be expressed as:

$$L = \frac{1}{N} \sum ||D_{out} - D_{gt}||_1 + \lambda ||\nabla D_{out} - \nabla D_{gt}||_1, \tag{2}$$

where the $||\cdot||_1$ represents the $L_1$ norm, and $N$ denotes the total number of pixels. The gradients are computed using the discrete Sobel operator. In our experimental setup, we set the value of $\lambda$ to 10, as suggested in [7, 27].

## 4 Experiments

### 4.1 Datasets

MTDNet is a neural network employing supervised learning to effectively remove noise from ToF depth images. We require ToF datasets that provide ground truth depth information to train all the network parameters. The typical approach for generating suitable datasets is to employ transient rendering technology to simulate the ToF imaging process while introducing MPI and shot noise [15]. Existing CNN-based methods for ToF denoising have created synthetic datasets with thousands of scenes. For our experiments, we select two large-scale synthetic datasets: ToF-FlyingThings3D (TFT3D) [27] and Cornell Box [30] for training and evaluation. The TFT3D dataset consists of 4000 scenes, including living rooms and kitchens. We utilize the ToF amplitude images and ToF depth images with a $640 \times 480$ resolution as input for our proposed method. The Cornell Box dataset comprises 21.3K scenes, consisting of raw measurements with different frequencies and corresponding ground truth depths. In our experiments, we convert the raw measurements at 50MHz frequency into ToF depth and amplitude images with a resolution of $600 \times 600$ [30]. Additionally, to assess our MTDNet's performance on realistic scenes, we also adopt the HAMMER dataset developed by Jung *et al.* [16], which offers iToF measurements and encompasses 13 different scenes with a resolution of $384 \times 576$. For the HAMMER dataset, we exclusively utilize the real data collected with Lucid Helios (iToF) in the HAMMER dataset.

### 4.2 Data Pre-processing

Firstly, the input depth images are normalized based on the provided depth value range in the dataset. Any pixels with depth values outside the range (0, 1] are filtered out Next, for the convenience of experiments, we crop the images

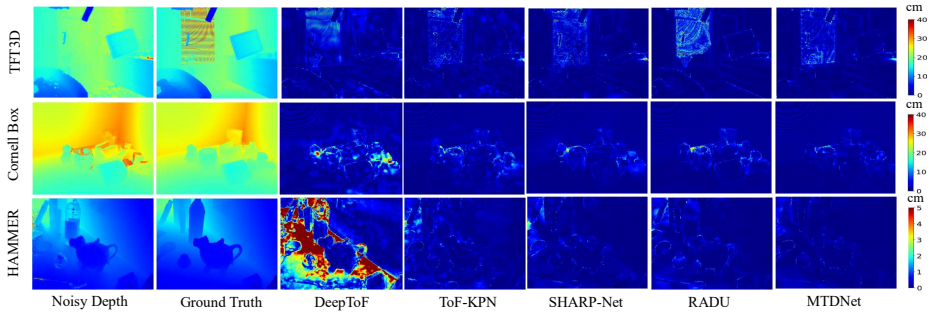


**Fig. 4:** Qualitative comparison on the TFT3D, Cornell-Box, and HAMMER datasets for ToF denoising. The colour bars on the right show the colour scale for error maps with the unit in cm.

from the TFT3D dataset, Cornell-Box dataset and HAMMER dataset to a size of $384 \times 512$. Considering the different camera movement speeds in each dataset, we use different numbers of intermediate frames to distinguish between the $t^{th}$ and $(t+1)^{th}$ frames for each dataset: three for the TFT3D dataset, three for the Cornell Box dataset, and seven for the HAMMER dataset. Finally, for all three datasets, we randomly select 10% scenes as the test set while the rest for training.

### 4.3 Training Settings

The methods involved in the quantitative and qualitative comparison include DeepToF, ToF-KPN, SHARP-Net, RADU and our proposed MTDNet. For the TFT3D dataset, the learning rate is $4 \times 10^{-4}$, which is reduced by 20% after every 8 epochs. We train all the methods for 100 epochs with a batch size of 2. We set the learning rate for the Cornell Box dataset as $1 \times 10^{-3}$ and the decay rate to 0.1 every 25 epochs. We train all the methods for 120 epochs with a batch size of 4. We set the HAMMER dataset's learning rate as $4 \times 10^{-4}$ and the decay rate to 0.8 every 8 epochs. We train all the methods for 70 epochs with a batch size of 2. The network is implemented using the TensorFlow framework [1] and trained using Adam optimizer. With four NVIDIA 1080Ti graphics cards, training takes about 30 hours for all TFT3D, Cornell Box and HAMMER datasets.

### 4.4 Results on Synthetic Datasets

To conduct a quantitative comparison, we utilize MAE to evaluate the performance of different denoising methods. The MAE is calculated by measuring the absolute difference between the denoised depth image and the ground truth depth image. To provide a comprehensive evaluation at various error levels, we adopt a similar evaluation method as ToF-KPN [27] and SHARP-Net [7] in our experiments. Specifically, we partition the pixels in the test set into four sets based on distinct error levels, with each set corresponding to a quantile of the

**Table 1:** Quantitative comparison with competitive ToF denoising methods on TFT3D, Cornell Box and HAMMER datasets. Here, 'Original' denotes the MAE between the original and the ground truth depth images.

| Model | TFT3D Dataset: MAE(cm) | | | | |
|---|---|---|---|---|---|
| | 1st Quan. | 2nd Quan. | 3rd Quan. | 4th Quan. | Overall |
| Original | 1.55 | 5.88 | 11.55 | 29.98 | 12.24 |
| DeepToF [21] | 0.78 | 0.91 | 1.09 | 5.10 | 1.97 |
| ToF-KPN [27] | 0.61 | 0.76 | 0.97 | 3.26 | 1.40 |
| SHARP-Net [7] | 0.45 | 0.49 | 0.59 | 2.55 | 1.02 |
| RADU [30] | 0.42 | 0.48 | 0.63 | 3.27 | 1.20 |
| MTDNet | **0.36** | **0.40** | **0.49** | **2.03** | **0.82** |
| Model | Cornell Box Dataset: MAE(cm) | | | | |
| | 1st Quan. | 2nd Quan. | 3rd Quan. | 4th Quan. | Overall |
| Original | 2.47 | 6.79 | 11.99 | 36.75 | 14.50 |
| DeepToF [21] | 0.95 | 0.96 | 0.96 | 1.85 | 1.18 |
| ToF-KPN [27] | 0.57 | 0.56 | 0.56 | 1.91 | 0.90 |
| SHARP-Net [7] | 0.48 | 0.47 | 0.45 | 0.96 | 0.59 |
| RADU [30] | 0.45 | 0.53 | 0.55 | 1.51 | 0.76 |
| MTDNet | **0.42** | **0.43** | **0.43** | **0.68** | **0.49** |
| Model | HAMMER Dataset: MAE(cm) | | | | |
| | 1st Quan. | 2nd Quan. | 3rd Quan. | 4th Quan. | Overall |
| Original | 0.194 | 0.656 | 2.089 | 11.709 | 3.662 |
| DeepToF [21] | 0.473 | 0.641 | 0.991 | 2.739 | 1.211 |
| ToF-KPN [27] | 0.091 | 0.105 | 0.141 | 0.355 | 0.173 |
| SHARP-Net [7] | 0.059 | 0.065 | 0.090 | 0.250 | 0.116 |
| RADU [30] | 0.062 | 0.070 | 0.104 | 0.356 | 0.148 |
| MTDNet | **0.050** | **0.055** | **0.076** | **0.207** | **0.097** |

total number of pixels in the ToF depth image. Employing this evaluation approach, we can thoroughly assess our MTDNet's performance across different error levels. It is worth noting that different denoising methods may exhibit varying performance levels depending on the specific error level.

Our MTDNet demonstrates outstanding performance in terms of MAE at the overall error level on both synthetic and realistic datasets, as shown in Table 1. Specifically, on the TFT3D dataset, the MAE between the input and ground truth depth is significantly reduced from 12.24 cm to an impressive 0.82 cm. Similarly, the MAE is reduced from 14.50 cm to 0.49 cm on the Cornell-Box dataset at 50MHz frequency. Our MTDNet significantly improves the denoising performance compared to baseline methods, especially at high error levels (third and fourth quantiles).

To further validate the denoising performance of our MTDNet, Fig. 4 presents a qualitative comparison between DeepToF, ToF-KPN, SHARP-Net, RADU, and our MTDNet. Overall, our MTDNet provides more accurate depth images while preserving the geometric structures within the scene. Furthermore, RADU outperforms SHARP-Net regarding noise removal in low-error-level regions, such as those with large-scale geometric shapes like walls and desktops. However, RADU struggles with high-error-level regions containing the geometric details of objects, such as the edge regions of cabinets and teacups. In contrast, our

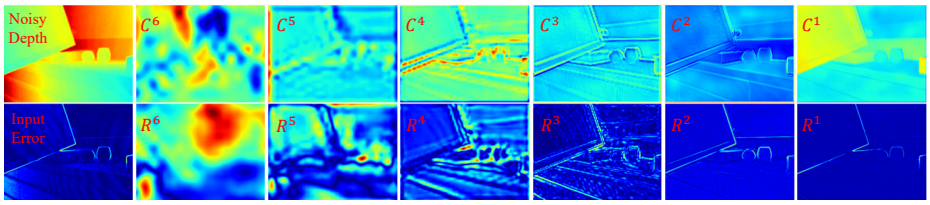


**Fig. 5:** Visualization of the immediate results produced by our MTDNet at different scales on the TFT3D dataset. Here, $C^i$ and $R^i$ denote the confidence map and the depth residual map at $i^{th}$ scale, respectively.

MTDNet simultaneously achieves superior results for low-error-level and high-error-level regions.

Furthermore, as depicted in Fig. 5, we present visualizations of the depth residual map and the confidence map at each scale. Notably, our coarse-to-fine framework progressively refines the confidence maps, directing the corresponding residual maps to target on eliminating the prominent ToF noises attributed to complex scene geometry. Furthermore, by incorporating intra-correlation and utilizing the residual pyramid architecture, our MTDNet effectively removes ToF noise, especially the MPI artifacts, at lower error levels. This integration allows our MTDNet to effectively remove large-scale MPI artifacts while maintaining consistent performance in handling strong ToF noises.

### 4.5 Results on the Realistic Dataset

In addition, we test our MTDNet along with existing methods on a realistic dataset. All the tested models are trained on the training set of the HAMMER dataset. The experimental results for all the methods on the HAMMER dataset are also shown in Table 1. Here, DeepToF faces difficulties in handling challenging scenes, especially for transparent and highly reflective objects that are common in the HAMMER dataset, which leads to DeepToF seeming to be completely unusable. However, DeepToF encounters difficulties in handling challenging scenes, especially for transparent and highly reflective objects common in the HAMMER dataset, leading to poor denoising performance. Compared with SHARP-Net and RADU, our MTDNet demonstrates strong performance across both low and high error levels on the real judgmental, which is generally consistent with the experimental results observed on synthetic datasets. At the bottom of Fig. 4, we present the qualitative comparison on a scene selected from the HAMMER dataset. MTDNet demonstrates the best visual effects.

### 4.6 Ablation Studies

MTDNet is a CNN-based method with a 6-level confidence-guided residual regression module as the backbone. It incorporates a dual-correlation estimation module, as well as combination and refinement modules. To assess the effectiveness of our proposed modules, we conduct experiments to compare MTDNet

**Table 2:** Quantitative comparison with the variants on the HAMMER dataset.

| Model | HAMMER Dataset:MAE(cm) | | | | |
|---|---|---|---|---|---|
| | 1st Quan. | 2nd Quan. | 3rd Quan. | 4th Quan. | Overall |
| w/o CRRM | 0.053 | 0.058 | 0.082 | 0.251 | 0.111 |
| w/o DCEM | 0.059 | 0.065 | 0.093 | 0.271 | 0.122 |
| w/o RCAB | 0.151 | 0.175 | 0.252 | 0.674 | 0.313 |
| w/o DRM | 0.052 | 0.058 | 0.080 | 0.230 | 0.105 |
| w/o Inter | 0.053 | 0.059 | 0.082 | 0.270 | 0.116 |
| w/o Intra | 0.057 | 0.063 | 0.089 | 0.259 | 0.117 |
| MTDNet | **0.050** | **0.055** | **0.076** | **0.207** | **0.097** |

against its variants. Additionally, we study the impact of varying the number of input frames on the denoising performance of the MTDNet and evaluate the temporal consistency and computational efficiency of the MTDNet. Please refer to the supplementary material for more details.

**Illustrating the significance of our proposed modules.** Our MTDNet consists of a dual-correlation estimation module (DCEM), a confidence-guided residual regression module (CRRM), a residual channel attention block (RCAB) and a depth refinement module (DRM). To validate the effectiveness of our proposed modules, we conduct experiments to compare our MTDNet against its variants. First, we devise four variants by removing the corresponding modules from our network: 'w/o DCEM', 'w/o CRRM', 'w/o RCAB', and 'w/o DRM'. Notably, 'w/o CRRM' only means removing the confidence map from our network to confirm its effectiveness for regressing the depth residual. To verify the effects of our inter- and intra-correlation, we design an experiment without inter-correlation and intra-correlation, respectively ('w/o Inter' and 'w/o Intra'). For a fair comparison, we adjust the number of convolution kernel channels of the variants to ensure that the number of parameters of these variants is nearly the same as MTDNet.

Incorporating a dual-correlation estimation module into our MTDNet significantly improves performance. This enhancement reduces the overall MAE from 0.122 cm to 0.097 cm compared to the 'w/o DCEM' approach. Our analysis also identifies the critical role of inter- and intra-correlation in accurately removing ToF noise within high-error and low-error levels, respectively. This observation is supported by the results obtained from excluding inter-correlation ('w/o Inter') and intra-correlation ('w/o Intra'). Moreover, the introduction of a confidence map into our MTDNet has proven beneficial, leading to a remarkable 12.6% reduction in the overall MAE compared to the 'w/o CRRM' approach. This finding underscores the importance of constructing a confidence map based on inter-correlation in the MPI removal.

Comparing MTDNet with 'w/o DRM' and 'w/o RCAB', it is evident that incorporating the depth refinement module and residual pyramid combination decreases the overall MAE by 7.6% and 69%, respectively. This highlights the necessity of those two modules in improving performance. The removal of the

**Table 3:** Quantitative comparison with competitive multi-frame processing strategies on the HAMMER dataset.

| Model | HAMMER Dataset: MAE(cm) | | | | |
|---|---|---|---|---|---|
| | 1st Quan. | 2nd Quan. | 3rd Quan. | 4th Quan. | Overall |
| SHARP-Net w/o align. | 0.070 | 0.077 | 0.111 | 0.330 | 0.147 |
| RADU w/o align. | 0.067 | 0.077 | 0.116 | 0.392 | 0.163 |
| MTDNet w/o align. | 0.053 | 0.058 | 0.081 | 0.244 | 0.109 |
| Deformable based alignment | 0.053 | 0.059 | 0.081 | 0.239 | 0.108 |
| Flow-based alignment | 0.054 | 0.060 | 0.085 | 0.257 | 0.114 |
| Our framework | **0.050** | **0.055** | **0.076** | **0.207** | **0.097** |

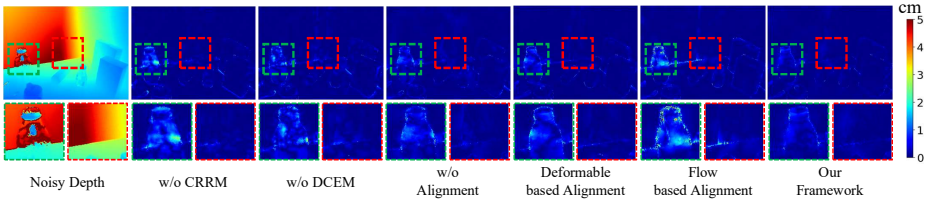


**Fig. 6:** Visualization of the results produced by the variants on the HAMMER dataset.

RCAB prevents the network from integrating depth residuals across various scales, disrupting the spatial hierarchy assumption and leading to a notable decrease in performance. In Fig. 6, we present visualizations of the error maps, including MTDNet, methods with different multi-frame processing strategies, 'w/o CRRM' and 'w/o DCEM'. Zoomed figures are incorporated to illustrate and affirm the effectiveness of our MTDNet in mitigating strong MPI noise.

**Comparing different multi-frame processing strategies.** We begin by comparing various multi-frame processing strategies as shown in Table 3 and Fig. 6. In the simplest scenario, features from multiple frames are concatenated without alignment. This approach significantly reduces performance as it fuses features from unrelated spatial locations. Flow-based alignment utilizes a pre-trained (fixed) optical flow estimator to align features at each scale of pyramids across frames. However, this approach is plagued by inaccurate flow estimations and the fundamental issue of foreground-background mixing [23]. Based on EDVR [38], deformable alignment involves incorporating deformable convolutions at every pyramid scale to align extracted features. While this yields a slight performance enhancement, it also considerably increases computational complexity. Following EDVR [38], deformable-based alignment involves deformable convolutions at various scales to align extracted features. Our dual-correlation-based framework avoids these issues and allows the network to pick out accurate geometric information from the features of multi-frame ToF data. Therefore, our proposed framework achieves superior results compared with other multi-frame processing strategies.

**Table 4:** Quantitative comparison with different number of depth images between $t^{th}$ and $(t+1)^{th}$ frames on the TFT3D dataset.

| Model | TFT3D Dataset: MAE(cm) | | | | |
|---|---|---|---|---|---|
| | 1st Quan. | 2nd Quan. | 3rd Quan. | 4th Quan. | Overall |
| 0 frame | 0.39 | 0.43 | 0.54 | 2.16 | 0.88 |
| 1 frame | 0.39 | 0.43 | 0.55 | 2.11 | 0.87 |
| 3 frame | **0.36** | **0.40** | **0.49** | **2.03** | **0.82** |
| 5 frame | 0.38 | 0.42 | 0.52 | 2.08 | 0.85 |
| 7 frame | 0.38 | 0.41 | 0.51 | 2.14 | 0.86 |

**Changing the number of intermediate frames the between $t^{th}$ and $(t+1)^{th}$ frames.** It's worth noting that, unlike alignment-based frameworks, the performance of our MTDNet is not linearly related to the increase or decrease of the number of intermediate frames between the $t^{th}$ and $(t+1)^{th}$ frames. In practical applications, the $(t+1)^{th}$ frame denotes the subsequent frame following $t^{th}$ frame. The inclusion of intermediate frames serves to simulate the impact of different camera movement speeds on the performance of our framework during real-world usage. As shown in Table 4, both an insufficient and an excessive number of intermediate frames result in sub-optimal performance. The scarcity of intermediate frames hampers the discernment of noise distribution across the $t^{th}$ and $(t+1)^{th}$ frames, whereas an overabundance of intermediate frames complicates the computation of matching costs.

## 5 Conclusion

In this paper, we first propose a novel multi-frame ToF denoising framework. Unlike existing alignment-based multi-frame processing methods, our proposed framework does not adopt an alignment-based architecture. Instead, it aggregates the dual correlation in multiple frames to guide the removal of ToF noise with a confidence map. We design a multi-frame ToF denoising network based on our proposed framework, *i.e.*, MTDNet. Our MTDNet consists of the Dual-Correlation Estimation Module and the Confidence-guided Residual Regression Module. The former constructs the inter- and intra-correlations to initialize the MPI estimation and capture the variations in ToF noise distribution across multiple frames. The latter utilizes the dual correlation to obtain a confidence map, guiding residual regression to focus on regions with strong ToF noise. Extensive experiments on the synthetic and realistic datasets show that our MTDNet surpasses the state-of-the-art methods across all error levels.

## Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 62032006, 62131003 and 62021001.

## References


1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: A system for large-scale machine learning. In: Symposium on Operating Systems Design and Implementation. pp. 265–283 (2016)
2. Agresti, G., Schaefer, H., Sartor, P., Zanuttigh, P.: Unsupervised domain adaptation for tof data denoising with adversarial learning. In: CVPR. pp. 5584–5593 (2019)
3. Agresti, G., Schäfer, H., Sartor, P., Incesu, Y., Zanuttigh, P.: Unsupervised domain adaptation of deep networks for tof depth refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(12), 9195–9208 (2021)
4. Bako, S., Vogels, T., McWilliams, B., Meyer, M., Novák, J., Harvill, A., Sen, P., Derose, T., Rousselle, F.: Kernel-predicting convolutional networks for denoising monte carlo renderings. ACM Transactions on Graphics (TOG) **36**(4), 97 (2017)
5. Bhandari, A., Feigin, M., Izadi, S., Rhemann, C., Schmidt, M., Raskar, R.: Resolving multipath interference in kinect: An inverse problem approach. In: Sensors. pp. 614–617. IEEE (2014)
6. Buratto, E., Simonetto, A., Agresti, G., Schäfer, H., Zanuttigh, P.: Deep learning for transient image reconstruction from tof data. Sensors **21**(6), 1962 (2021)
7. Dong, G., Zhang, Y., Xiong, Z.: Spatial hierarchy aware residual pyramid network for time-of-flight depth denoising. In: ECCV. pp. 35–50. Springer (2020)
8. Duzceker, A., Galliani, S., Vogel, C., Speciale, P., Dusmanu, M., Pollefeys, M.: Deepvideomvs: Multi-view stereo on video with recurrent spatio-temporal fusion. In: CVPR. pp. 15324–15333 (2021)
9. Fan, L., Xiong, X., Wang, F., Wang, N., Zhang, Z.: Rangedet: In defense of range view for lidar-based 3d object detection. In: ICCV. pp. 2918–2927 (2021)
10. Freedman, D., Smolin, Y., Krupka, E., Leichter, I., Schmidt, M.: Sra: Fast removal of general multipath for tof sensors. In: ECCV. pp. 234–249. Springer (2014)
11. Guo, Q., Frosio, I., Gallo, O., Zickler, T., Kautz, J.: Tackling 3d tof artifacts through learning and the flat dataset. In: ECCV. pp. 368–383 (2018)
12. Gupta, M., Nayar, S.K., Hullin, M.B., Martin, J.: Phasor imaging: A generalization of correlation-based time-of-flight imaging. ACM Transactions on Graphics (TOG) **34**(5), 156 (2015)
13. Gutierrez-Barragan, F., Chen, H., Gupta, M., Velten, A., Gu, J.: itof2dtof: A robust and flexible representation for data-driven time-of-flight imaging. IEEE Transactions on Computational Imaging **7**, 1205–1214 (2021)
14. Illade-Quinteiro, J., Brea, V.M., López, P., Cabello, D., Doménech-Asensi, G.: Distance measurement error in time-of-flight sensors due to shot noise. Sensors **15**(3), 4624–4642 (2015)
15. Jarabo, A., Marco, J., Muñoz, A., Buisan, R., Jarosz, W., Gutierrez, D.: A framework for transient rendering. ACM Transactions on Graphics (TOG) **33**(6), 177 (2014)
16. Jung, H., Ruhkamp, P., Zhai, G., Brasch, N., Li, Y., Verdie, Y., Song, J., Zhou, Y., Armagan, A., Ilic, S., et al.: Is my depth ground-truth good enough? hammer–highly accurate multi-modal dataset for dense 3d scene regression. arXiv preprint arXiv:2205.04565 (2022)
17. Jung, J., Lee, J.Y., Jeong, Y., Kweon, I.S.: Time-of-flight sensor calibration for a color and depth camera pair. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(7), 1501–1513 (2014)

18. Lenzen, F., Schäfer, H., Garbe, C.: Denoising time-of-flight data with adaptive total variation. In: International Symposium on Visual Computing. pp. 337–346. Springer (2011)
19. Li, Z., Ye, W., Wang, D., Creighton, F.X., Taylor, R.H., Venkatesh, G., Unberath, M.: Temporally consistent online depth estimation in dynamic scenes. In: WACV. pp. 3018–3027 (2023)
20. Lin, J., Liu, Y., Hullin, M.B., Dai, Q.: Fourier analysis on transient imaging with a multifrequency time-of-flight camera. In: CVPR. pp. 3230–3237 (2014)
21. 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 Transactions on Graphics (TOG) **36**(6), 219 (2017)
22. Mildenhall, B., Barron, J.T., Chen, J., Sharlet, D., Ng, R., Carroll, R.: Burst denoising with kernel prediction networks. In: CVPR. pp. 2502–2510 (2018)
23. Niklaus, S., Liu, F.: Softmax splatting for video frame interpolation. In: CVPR. pp. 5437–5446 (2020)
24. Patil, V., Van Gansbeke, W., Dai, D., Van Gool, L.: Don't forget the past: Recurrent depth estimation from monocular video. IEEE Robotics and Automation Letters **5**(4), 6813–6820 (2020)
25. Piao, Y., Ji, W., Li, J., Zhang, M., Lu, H.: Depth-induced multi-scale recurrent attention network for saliency detection. In: ICCV. pp. 7254–7263 (2019)
26. Qiao, S., Zhu, Y., Adam, H., Yuille, A., Chen, L.C.: Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation. In: CVPR. pp. 3997–4008 (2021)
27. Qiu, D., Pang, J., Sun, W., Yang, C.: Deep end-to-end alignment and refinement for time-of-flight rgb-d module. In: ICCV. pp. 9994–10003 (2019)
28. Reading, C., Harakeh, A., Chae, J., Waslander, S.L.: Categorical depth distribution network for monocular 3d object detection. In: CVPR. pp. 8555–8564 (2021)
29. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention. pp. 234–241. Springer (2015)
30. Schelling, M., Hermosilla, P., Ropinski, T.: Radu: Ray-aligned depth update convolutions for tof data denoising. In: CVPR. pp. 671–680 (2022)
31. Simonetto, A., Agresti, G., Zanuttigh, P., Schäfer, H.: Lightweight deep learning architecture for mpi correction and transient reconstruction. IEEE Transactions on Computational Imaging **8**, 721–732 (2022)
32. Su, S., Heide, F., Wetzstein, G., Heidrich, W.: Deep end-to-end time-of-flight imaging. In: CVPR. pp. 6383–6392 (2018)
33. Sun, D., Yang, X., Liu, M.Y., Kautz, J.: Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In: CVPR. pp. 8934–8943 (2018)
34. Sun, Z., Ye, W., Xiong, J., Choe, G., Wang, J., Su, S., Ranjan, R.: Consistent direct time-of-flight video depth super-resolution. In: CVPR. pp. 5075–5085 (2023)
35. Tu, J., Ren, M., Manivasagam, S., Liang, M., Yang, B., Du, R., Cheng, F., Urtasun, R.: Physically realizable adversarial examples for lidar object detection. In: CVPR. pp. 13716–13725 (2020)
36. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. In: CVPR. pp. 3156–3164 (2017)
37. Wang, X., Zhou, W., Jia, Y.: Attention gan for multipath error removal from tof sensors. IEEE Sensors Journal **22**(20), 19713–19721 (2022)
38. Wang, X., Chan, K.C., Yu, K., Dong, C., Change Loy, C.: Edvr: Video restoration with enhanced deformable convolutional networks. In: CVPRW (2019)

39. Watson, J., Mac Aodha, O., Prisacariu, V., Brostow, G., Firman, M.: The temporal opportunist: Self-supervised multi-frame monocular depth. In: CVPR. pp. 1164–1174 (2021)
40. Wei, Y., Liu, S., Rao, Y., Zhao, W., Lu, J., Zhou, J.: Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In: ICCV. pp. 5610–5619 (2021)
41. Xiao, Z., Liu, Y., Gao, R., Xiong, Z.: Cutmib: Boosting light field super-resolution via multi-view image blending. In: CVPR. pp. 1672–1682 (2023)
42. Xiao, Z., Weng, W., Zhang, Y., Xiong, Z.: Eva2: Event-assisted video frame interpolation via cross-modal alignment and aggregation. IEEE Transactions on Computational Imaging **8**, 1145–1158 (2022)
43. Yin, Z., Shi, J.: Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: CVPR. pp. 1983–1992 (2018)
44. Zanuttigh, P., Marin, G., Dal Mutto, C., Dominio, F., Minto, L., Cortelazzo, G.M.: Time-of-flight and structured light depth cameras. Technology and Applications, ISSBN pp. 978–3 (2016)
45. Zhang, X., Yan, H., Zhou, Q.: Overcoming the shot-noise limitation of three-dimensional active imaging. Optics Letters **36**(8), 1434–1436 (2011)
46. Zhu, S., Brazil, G., Liu, X.: The edge of depth: Explicit constraints between segmentation and depth. In: CVPR. pp. 13116–13125 (2020)
47. Zhu, X., Zhou, H., Wang, T., Hong, F., Ma, Y., Li, W., Li, H., Lin, D.: Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In: CVPR. pp. 9939–9948 (2021)