# ST-DepthNet: A Spatio-Temporal Deep Network for Depth Completion Using a Single Non-Repetitive Circular Scanning Lidar

Örkény Zováthi, Balázs Pálffy, Zsolt Jankó, and Csaba Benedek

***Abstract*—In this letter, we propose a novel depth image completion technique based on sparse consecutive measurements of a non-repetitive circular scanning (NRCS) Lidar, demonstrating the capabilities of a new, compact, and accessible sensor technology for dense range mapping of highly dynamic scenes. Our deep network called *ST-DepthNet* is composed of a spatio-temporally (ST) extended U-Net architecture, which accepts a very sparse range data sequence as input and produces a dense depth image stream of the same field-of-view ensuring a high level of spatial details and accuracy. For evaluation, we have constructed a new urban dataset, that – to our best knowledge as the first open Benchmark in this field – comprises various simulated and real-world NRCS Lidar data samples, allowing us to simultaneously train our model on synthetic data with Ground Truth, and to validate the result via real NRCS Lidar measurements. Using this new dataset, we have shown the superiority of our method against a densified depth map obtained from the raw sensor stream, and against two independent state-of-the-art deep-learning based Lidar-only depth completion methods.**

***Index Terms*—Deep learning for visual perception, visual learning, range sensing.**

## I. Introduction

ACCURATE and dense depth map prediction is an essential problem in 3D scene understanding. In applications such as dynamic environment analysis, 3D mapping, and virtual city generation, depth information is often obtained from commercially available Lidar (light detection and ranging) sensors as they are able to map their environment in real-time by emitting multiple laser beams and receiving their returns. However, the data captured by Lidars is often very sparse, while its characteristics may vary depending on the sensors' scanning technology. In this context, depth completion algorithms need to estimate dense depth images from the Lidar-acquired sparse range measurements.

For dynamic environment perception and recognition tasks such as advanced scene analysis and understanding, repetitive, typically rotating multi-beam (RMB) Lidar sensors (e.g., Ouster OS1 or Velodyne Puck models) [1] are commonly utilized devices. RMB Lidars can produce real-time point cloud streams (300 $k$-2 $M$ points/s), however, their measurements have low spatial density, and their field of view (FoV) coverage is constant through the whole scanning process: Their vertical resolution is fixed by the number of the laser beams (16-128), while their horizontal resolution depends on the sensor's rotation frequency (5–20 Hz).

Alternatively to RMB Lidars, recent non-repetitive circular scanning (NRCS) Lidar sensors are also capable of providing measurements for real-time scene analysis in robotics and autonomous driving, at a significantly lower cost compared to the RMB technology [2], by using single- or multi-line lasers combined with high-speed scanning on a circular path. Unlike RMB Lidars, NRCS Lidars (e.g., the Livox AVIA sensor) are able to densely map large areas from a given scanning position due to their special scanning technology which follows non-repetitive e.g., rosetta patterns (Fig. 1). The main challenge is here to efficiently balance between the spatial and the temporal resolution of the recorded range data using a suitable integration window [3].

On one hand, as shown in Fig. 2(a), allowing larger integration time ($t_{\Delta} > 1\,\mathrm{s}$), the laser beams cover a higher proportion (around 90%) of the FoV yielding high spatial measurement resolution. However, the potential ego-motion of the Lidar's platform (e.g., vehicle or robot) and the dynamic objects in the surrounding area induce various artifacts, such as blurred shapes of the observed vehicles, pedestrians or buildings, which phenomena complicate dynamic event analysis. On the other hand, if the measurements are collected within a narrow time window (e.g., in 200 ms) they are spatially more precise, however, the resulting point clouds are notably sparse (around 48$k$ points, up to 40% FoV coverage), which fact yields a significant loss of details across the spatial dimension of the FoV (see Fig. 2(b)).

Manuscript received 30 November 2022; accepted 1 April 2023. Date of publication 12 April 2023; date of current version 24 April 2023. This letter was recommended for publication by Associate Editor X. Huang and Editor C. Cadena Lerma upon evaluation of the reviewers' comments. The work of Örkény Zováthi was supported by KDP-2020 Cooperative Doctoral Program under Grant KDP-977852. This work was supported in part by the Artificial Intelligence National Laboratory under Grant RRF-2.3.1-21-2022-00004, in part by the National Laboratory for Autonomous Systems under Grant RRF-2.3.1-21-2022-00002, and in part by the Hungarian NRDI Office under Grants TKP2021-NVA-01 and OTKA 143274. *(Corresponding author: Örkény Zováthi.)*

Örkény Zováthi, Balázs Pálffy, and Csaba Benedek are with the Institute for Computer Science and Control (SZTAKI), Eötvös Loránd Research Network, 1111 Budapest, Hungary, and also with the Péter Pázmány Catholic University, 1083 Budapest, Hungary (e-mail: zovathi.orkeny@sztaki.hu; palffy.balazs@sztaki.hu; benedek.csaba@sztaki.hu).

Zsolt Jankó is with the Institute for Computer Science and Control (SZTAKI), Eötvös Loránd Research Network, 1111 Budapest, Hungary (e-mail: janko.zsolt@sztaki.hu).

The datasets and the codes are publicly available at: www.github.com/sztaki-geocomp/ST-DepthNet

This letter has supplementary downloadable material available at https://doi.org/10.1109/LRA.2023.3266670, provided by the authors.

Digital Object Identifier 10.1109/LRA.2023.3266670

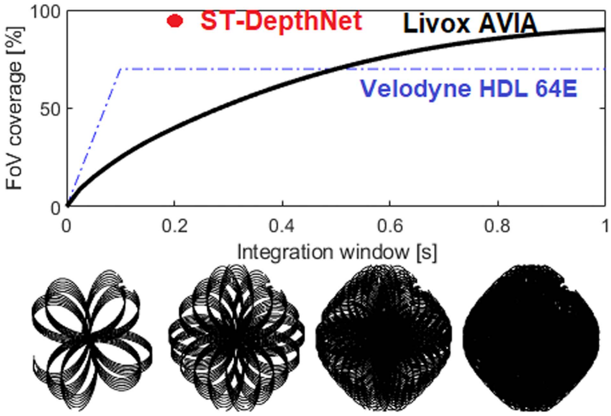
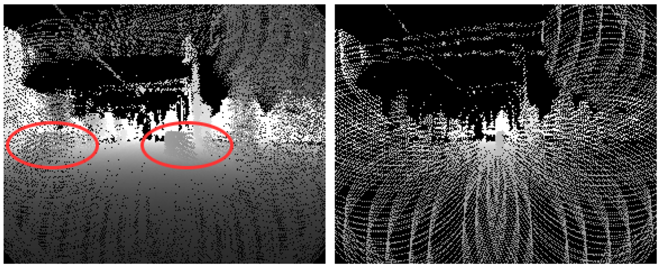


Fig. 1. Non-repetitive sampling strategy of the Livox AVIA NRCS Lidar. The circular scanning produces typical rosetta patterns which are varying across different time frames.

In this letter, we aim to overcome the above-mentioned challenges caused by the spatio-temporal trade-off of the NRCS Lidar based perception, and propose a novel deep learning based approach for densifying sparse NRCS Lidar data while keeping its spatial accuracy high. Our proposed spatio-temporal (ST) deep network called *ST-DepthNet* (Fig. 3) operates in the range domain and expects as input multiple sparse depth maps captured by a NRCS Lidar consecutively in time, with using a narrow (i.e., 200 ms) integration window for each frame. As output, the network provides a dense, high-quality range image of the same FoV, which does not reflect the sensor's original scanning artifacts (i.e., visible trails of the circular scanning pattern). The architecture of *ST-DepthNet* was directly designed to exploit both spatial and temporal patterns in the input NRCS data for depth completion, by extending a U-Net-like architecture [4], [5] with Conv2DLSTM [6] layers.

We define the main contributions of the paper as follows:

- We propose a novel deep learning based solution called *ST-DepthNet*, which extends the classical U-Net architecture with a spatio-temporal downscaling branch for utilizing consecutive sparse measurements captured by NRCS Lidars. Our model produces spatially precise high-density depth data using a spatial upscaling branch following effective temporal pooling steps.
- We provide a new synthetic urban dataset called *LivoxCarla*, which contains simulated NRCS Lidar data with corresponding dense depth Ground Truth (GT) information. We demonstrate that with the *LivoxCarla* dataset, we are able to simulate realistic urban NRCS Lidar measurements, which can provide a basis for training and evaluation of methods developed for processing real NRCS sensor data.
- We provide a real-life dataset called *LivoxBudapest*, which contains real NRCS Lidar measurement data collected in Budapest, Hungary, both in downtown and speedway areas, by a sensor mounted on a moving vehicle. Testing with the real measurements allows us to clearly demonstrate the usability of our method trained on synthetic data in real-life urban scenarios.
- We qualitatively and quantitatively evaluate the proposed algorithm, and experimentally demonstrate its advantages against two state-of-the-art methods. We also share the datasets and the source code of the proposed method with the community.



(a) $t_\Delta = 1$ s (b) $t_\Delta = 200$ ms

Fig. 2. A dynamic scene captured by a NRCS Lidar with different $t_\Delta$ integration windows. A large integration time (a) induce several blurring artifacts, while a narrow integration window (b) yields the loss of details. Blurred pedestrians are marked by red ellipses.

## II. Related Work

In this section, we present a state-of-the-art study of depth completion techniques and challenges. In the past few years, research on Lidar-based approaches emerged as a hot topic in the literature, due to the availability of popular public datasets like the KITTI Depth Completion Benchmark [7].

The majority of the recent methods focus on completing depth maps obtained from RMB Lidars fused with optical images as guidance to recover the pixels with missing depth measurements [8], [9]. However, optical images may not provide eligible information in cases of sudden illumination changes [9] or in low-light environments [10]. In these cases, depth completion must be performed solely based on sparse Lidar range measurement samples, which includes significantly harder challenges [11].

First, without relying on external sources (e.g., high-resolution RGB images), edges and other finely textured structures on the generated depth images are often missing, blurred or distorted [11], [12]. In [11], global and local depth variations are separated based on the fact that in the wavelet representation of the images, the fine structures mainly appear in the high-frequency domain while the global regions are defined by the low-frequency coefficients. In order to exploit this phenomenon, they introduce a frequency-based recurrent depth coefficient refinement scheme. The difficulty of data upsampling near the edges also appears in [12], where feature extraction by an edge convolution layer is used to strengthen the precision at fine 3D structures. In our approach, we recover the fine structures by adding an appropriate edge-loss term [13] to our loss function, instead of performing edge enhancement by a dedicated sub-network.

Second, a limitation of many existing depth completion networks is that they generate new range values for all image pixels, instead of filling only the missing information [14]. Therefore the Implicit Lidar Network [14] learns the weights of an interpolation function for 3D point cloud completion, thus the original measurements are not modified and only the missing points are estimated. For similar reasons, our solution connects the last sparse input image to the output by a direct

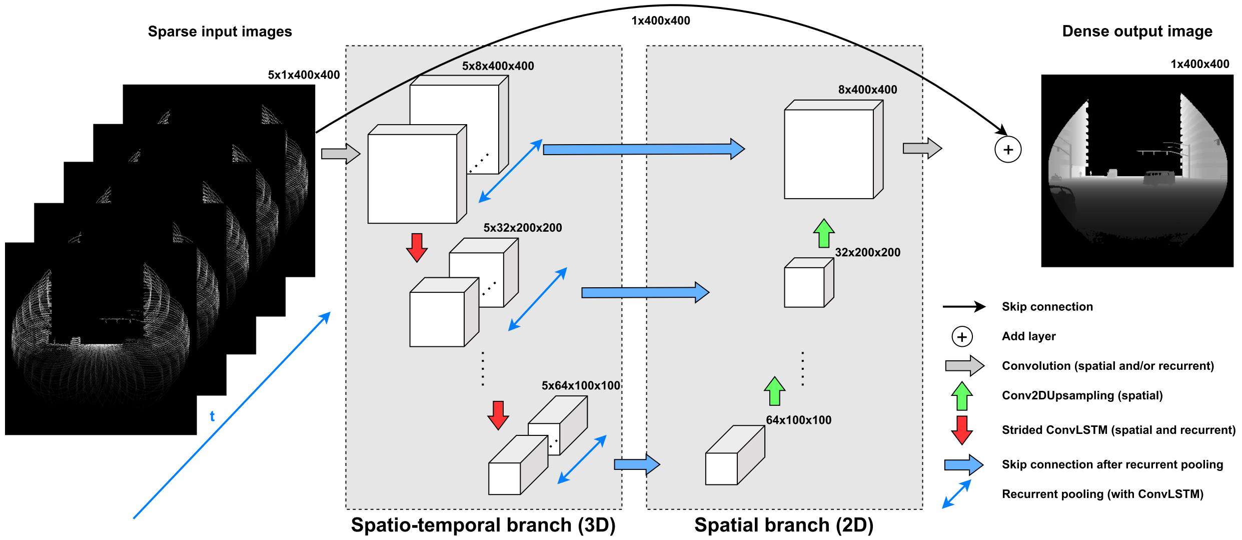

Sparse input images
5x1x400x400
t
1x400x400
5x8x400x400
5x32x200x200
5x64x100x100
8x400x400
32x200x200
64x100x100
+
Dense output image
1x400x400
Skip connection
Add layer
Convolution (spatial and/or recurrent)
Conv2DUpsampling (spatial)
Strided ConvLSTM (spatial and recurrent)
Skip connection after recurrent pooling
Recurrent pooling (with ConvLSTM)
Spatio-temporal branch (3D)
Spatial branch (2D)

Fig. 3. The architecture of the proposed ST-DepthNet network.

skip connection to force our proposed model to complete the original sparse, but precise range map instead of overwriting it with completely new values.

The most closely related methods to our approach that focus on Lidar-only depth completion are [15] which effectively combines morphological operations and bilateral filtering, and [10] that investigates different sampling strategies for training a generative adversarial network. However, as our experiments show in Section IV, both approaches are highly sensitive to the measurement characteristics of the applied Lidar sensor and fail to accurately compensate for the irregular, non-repetitive sampling pattern of NRCS Lidars. As NRCS Lidars are relatively new to the market, to the best of our knowledge, this letter is the first to provide a dataset and method utilizing information for depth completion propagated from their measurements.

## III. The Proposed Method

The goal of the proposed solution is to produce a high-quality, dense and spatially precise point cloud stream from measurements of a single NRCS Lidar sensor. Our approach consists of two main steps: First, the consecutive measurements of the NRCS Lidar are grouped to form discrete time frames, using a narrow, 200 ms integration window (up to 40% FoV coverage in each frame). Thereafter, within a frame the distances of the measured 3D field points from the sensor are assigned to corresponding pixels in a high-resolution range image. By each actual time frame, the last five collected depth images (covering together around 95% of the FoV) are fed to the *ST-DepthNet* depth completion network, which composes a high-quality range image as output, with almost 100% FoV coverage, also eliminating the motion blurring artifacts. The output high-quality range image can be backprojected to the 3D space as well.

### A. Range Image Generation

Range images are widely used, compact representations of Lidar-based depth measurements [1], [3], [16], which enable to adopt 2D convolution operations and effective image-based neural network architectures [4], [6] during processing.

In our approach, the captured sparse point clouds (collected within $t_\triangle = 200$ ms) are converted from the Cartesian (x, y, z) to the spherical (distance, azimuth, elevation) polar coordinate system. Then, a 2D pixel lattice is generated by quantizing the horizontal (azimuth) and vertical (elevation) FoVs. In the resulting range images, the horizontal and vertical pixel coordinates represent the polar azimuth and elevation angles, while the pixel's depth value encodes the distance of the corresponding point.

In our experiments, we exploit the parameters of the Livox AVIA state-of-the-art NRCS Lidar sensor [3]. The sensor's FoV is mapped onto a $400 \times 400$ pixel lattice, which resolution (5.6 px/°) yields both high spatial accuracy and reasonable computational requirements. As experienced, the density of the recorded valid range values is decreasing towards the peripheral regions of the range image due to the nature of the circular scanning technique: the scanning pattern crosses the optical center of the sensor significantly more frequently, than the FoV's perimeter, making the central regions of the range images densely filled, and leaving peripheral areas notably sparse (see Figs. 1 and 2). As a result of using an integration time window of 200 ms for collecting the consecutive time frames, around 60% of the range image pixels receive undefined range values. Such a level of sparseness of the range image makes it difficult to efficiently visualize the data or to perform scene analysis, emerging the need for the proposed depth estimation approach.

### B. ST-DepthNet Architecture

Next, we use a range image sequence acquired by the NRCS Lidar as input to the proposed *ST-DepthNet* deep network (Fig. 3). As discussed earlier, sparse measurement frames collected in 200 ms time windows cover only a low proportion of the defined $400 \times 400$ range image lattice. On the other hand, using a 1 s time frame, the collected point set covers almost fully the sensor's FoV (Fig. 1), but it is affected by motion blur.

Nevertheless, we can expect that the measurements from the last 1 s time interval always contain dense range information from the scene. Thus to also prevent blurring, we take five consecutive, "sparse" range images (each one recorded in 200 ms) as our network's input.

Since the main goal is to generate a high-quality output image from the sparse range image inputs, we have adopted an image-to-image U-Net [4] like architecture: More specifically, we extended the downscaling part of a U-Net network enabling to exploit temporal connections where the input is an image sequence, by utilizing Conv2DLSTM layers presented first in [6]. Let us introduce a regular Long-Short Term Memory (LSTM) cell, which has a memory state $C_t$ and a final state $H_t$. At each timestep $t$, the memory is updated as a function of its current input $X_t$, previous final state $H_{t-1}$ based on an input gate $i_t$, while the propagation of its previous value $C_{t-1}$ depends on a forget gate $f_t$. The propagation of the memory state $C_t$ to the final state $H_t$ depends on the output gate $o_t$. In each dependency, from state $\alpha$ to $\beta$, there is a weight term $W_{\alpha\beta}$ and a bias term $b_\alpha$. A Conv2DLSTM cell operates similarly to a regular LSTM cell, with an extension that the input $X_t$, memory state $C_t$ and final state $H_t$ with their respective gates ($i_t, f_t, o_t$) are 3D tensors – with one temporal and two spatial dimensions – and both the spatial and recurrent transformations are convolutional (marked by $*$ in Equation system (1)) and not element-wise (marked by $\circ$), making it able to propagate spatio-temporal features:

$$\begin{aligned}
C_t &= f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc} * X_t + W_{hc} * H_{t-1} + b_c)\\
i_t &= \sigma(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} C_{t-1} + b_i)\\
f_t &= \sigma(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f\\
o_t &= \sigma(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_o)\\
H_t &= o_t \circ \tanh(C_t)
\end{aligned} \tag{1}$$

Hence, in our proposed approach, we keep a spatio-temporal three-dimensional (two spatial and one temporal) downscaling branch at the whole left side of the U-Net structure. On the other hand, the upscaling branch of our proposed network is purely two-dimensional, in order to accurately restore the single output image of our interest. Skip connections at each level are performed by recurrent pooling utilizing the last output of a Conv2DLSTM layer which represents features of the last 200 ms measurement.

Last but not least, we directly connect the last input image to our output. With this modification, which is also supported by our ablation experiments (see Section IV), we can exploit that the last and most up-to-date input image contains spatially precise points, and therefore, our network only has to learn the missing regions of the range image [14].

### C. Training Process

The proposed *ST-DepthNet* network is responsible for learning and predicting a high-density range image using a sparse input range image sequence. To deal with the challenging artifacts presented in Section II, our *loss function* $\mathcal{L}$ is composed of three main components.

First, we calculate the L1 Loss ($\mathcal{L}_{\text{L1}}$) as the mean absolute error between the generated and the GT depth images to force detailed, pixel-level accurate predictions. Second, we adopt the Structure Similarity Index Measure ($\mathcal{L}_{\text{SSIM}}$) proposed by [17], which quantifies the perceived difference in luminance, contrast and structural information between the predicted and GT depth images using a variety of known properties of the human visual system. Third, we also utilize a smoothness or edge loss term ($\mathcal{L}_{\text{EDGE}}$) specifically proposed for depth images by [13], which induces sharp contours on the generated images, thus spatially precise boundaries are enforced between objects in the 3D space. Our final *loss function* can be expressed therefore as follows:

$$\mathcal{L} = \alpha_1 \mathcal{L}_{\text{SSIM}} + \alpha_2 \mathcal{L}_{\text{L1}} + \alpha_3 \mathcal{L}_{\text{EDGE}}. \tag{2}$$

Following a parameter optimization step (see Section IV), $\alpha_1 = 0.7$, $\alpha_2 = 1.4$ and $\alpha_3 = 1.5$ were used in the final model. The loss function was minimized by the Adam optimizer. The learning rate was set to 0.0002 and the decay rate of the first moment to 0.5. We have trained our model on 10 epochs which took around 27 hours on a NVIDIA GTX 1080 Ti graphical processing unit (GPU).

### D. Datasets

While the main goal of our work is to propose an algorithm which can accurately deal with real NRCS Lidar measurement sequences, it is challenging to provide dense, spatially precise GT depth information for real data due to the independent movements of dynamic objects of the scene including the ego robot or vehicle (Fig. 2). Instead, we constructed a synthetic range image dataset called *LivoxCARLA* from a realistic virtual world using the CARLA simulator [14], [18], where the behaviour of the Livox AVIA NRCS Lidar sensor (Fig. 1) was simulated. The virtual world allows us to extract dense, spatially precise depth information, used as GT for the Lidar's sparse, rosetta patterned samples. During data extraction, the synthetic NRCS sensor was placed by default on the front-top of the capturing vehicle and was pointing forwards. The vehicle was dynamically moving during the whole data recording. To augment on the extractable information (e.g., due to varying ground level), the sensor's position was randomly rotated along the up axis by $[-22.5°, 22.5°]$, and its height was randomly adjusted between [1.5 m , 2.5 m].

Our *LivoxCARLA* dataset consists of 11726 randomly sampled input-output range image pairs, from which 10000 were used for training, 500 as validation and 1226 for testing. Each pair consists of $400 \times 400$ images: the input range images were generated with NRCS-characteristics by a Livox AVIA sensor model, in 200 ms integration windows (with ca. 40% FoV coverage), while a high-resolution ground truth range image was sampled by each fifth input frame.

Besides the *LivoxCARLA* dataset, we also collected real measurement sequences from Budapest. In these experiments, we used the Livox AVIA sensor mounted on the front-top of our test vehicle on a driving path of total 5.5 kilometers in both speedways and in the city center. Similarly to synthetic data generation, the real test vehicle was continuously moving during the measurements, while many different traffic participants were captured. Although this real dataset, referred as *LivoxBudapest*, does not include GT data, it enables us to validate the effectiveness of the proposed algorithm in real environment, despite the fact that the network is purely trained on synthetic data.

## IV. Experiments

We have trained and quantitatively evaluated the proposed method using the *LivoxCarla* dataset, exploiting its sparse input–dense output range image pairs generated by a simulated car-mounted NRCS Lidar sensor during virtual drives in dense city environments with several dynamic traffic participants (humans, vehicles, bikes). Besides quantitative validation, we also evaluated qualitatively the performance of the proposed method on real data using the *LivoxBudapest* dataset. Both datasets are introduced earlier in Section III-D.

### *A. Evaluation Metrics*

During quantitative analysis, we performed evaluation in both 2D and 3D, analysing the generated range images, and the backprojected 3D point clouds, respectively.

*1) 2D Errors:* For measuring the similarity between the generated range images to the GT, we adopted the following metrics from the KITTI Depth Completion Benchmark [7]:

- RMSE: Root mean squared error [mm]

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(I_{P_i} - I_{GT_i}\right)^2} \tag{3}$$

- MAE: Mean absolute error [mm]

$$\mathrm{MAE} = \sum_{i=1}^{N}\left|I_{P_i} - I_{GT_i}\right| \tag{4}$$

- iRMSE: RMSE of the inverse depth [1/km]

$$\mathrm{iRMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{I_{P_i}} - \frac{1}{I_{GT_i}}\right)^2} \tag{5}$$

- iMAE: MAE of the inverse depth [1/km]

$$\mathrm{iMAE} = \sum_{i=1}^{N}\left|\frac{1}{I_{P_i}} - \frac{1}{I_{GT_i}}\right| \tag{6}$$

In the above (3)–(6), $I_{P_i}$ denotes the $i$th pixel of the image generated by the actual method, while $I_{GT_i}$ is the $i$th pixel of the corresponding GT image. $N$ denotes the number of pixels, in our case $N = 400 \times 400$.

*2) 3D Errors:* Besides range image based evaluation, we also compared the generated point clouds to the reference model in the 3D space. Let us denote the GT and a predicted point cloud by $P_{\mathrm{GT}}$ and $P_{\mathrm{P}}$, and the number of points in $P_{\mathrm{GT}}$ and $P_{\mathrm{P}}$ by $\#P_{\mathrm{GT}}$ and $\#P_{\mathrm{P}}$, respectively. We evaluate the quality of the predicted point cloud with respect to the GT data using the symmetric Normalized Chamfer Distance (NCD) and Normalized Median Distance (NMD) [1], while these evaluation measures are also used to compare the performance of different baseline algorithms in the 3D space:

$$S_{P_1,P_2} = \sum_{p\in P_1}\min_{q\in P_2}||p-q||^2 \tag{7}$$

$$M_{P_1,P_2} = \operatorname*{Med}_{p\in P_1}\min_{q\in P_2}\sqrt{||p-q||^2} \tag{8}$$

$$Q_{\mathrm{NCD}}(P_{\mathrm{P}},P_{\mathrm{GT}}) = \sqrt{\frac{1}{2}\left(\frac{S_{P_P,P_{GT}}}{\#P_P} + \frac{S_{P_{GT},P_P}}{\#P_{GT}}\right)} \tag{9}$$

$$Q_{\mathrm{NMD}}(P_{\mathrm{P}},P_{\mathrm{GT}}) = \frac{1}{2}\left(M_{P_P,P_{GT}} + M_{P_{GT},P_P}\right) \tag{10}$$

TABLE I
An Ablation Study of the Proposed ST-DepthNet Architecture

| Temporal fusion | Skip connections | RMSE ↓ | MAE ↓ |
|---|---|---|---|
| X | X | 3512.95 | 1549.36 |
| X | Inner levels | 3367.75 | 1353.49 |
| X | Inner levels+Output | 2869.84 | 1383.10 |
| Early | X | 2969.22 | 883.92 |
| Early | Inner levels | 2767.87 | 832.87 |
| Early | Inner levels+Output | 2129.39 | 686.34 |
| Late | X | 2672.20 | 800.16 |
| Late | Inner levels | 1897.99 | 523.57 |
| **Late** | **Inner levels+Output** | **1799.16** | **440.42** |

TABLE II
Different Hyperparameter Setups for the Final Model

| Output skip layer | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | RMSE ↓ | MAE ↓ |
|---|---|---|---|---|---|
| X | 0.85 | 1.00 | 0.90 | 4267.88 | 1494.07 |
| X | 0.85 | 1.00 | 1.00 | 3871.45 | 1366.16 |
| X | 0.60 | 2.50 | 2.50 | 2938.23 | 1512.53 |
| X | 0.70 | 1.00 | 1.20 | 2000.33 | 541.06 |
| X | 0.70 | 1.40 | 1.50 | 1897.99 | 523.57 |
| ✓ | **0.70** | **1.40** | **1.50** | **1799.16** | **440.42** |

### *B. Ablation Study and Hyperparameters*

For optimizing the network structure, we investigated the effect of how deeply we integrate temporal information in the network architecture. In the first setup (No fusion), we trained the network without utilizing temporal data and considering only the measurements from the last 200 ms. In the second setup (Early fusion), we fused the multitemporal information only in the first Conv2DLSTM layer and the remaining layers remained pure spatial convolutions. Finally, as proposed, we propagated the temporal information through the whole feature downscaling branch (Late fusion). Furthermore, in each setup, we examined the effect of including/excluding U-Net-like skip connections in the network (Inner levels) and to directly bind the last input and the output depth image (Output). According to our comparative results displayed in Table I, the proposed late fusion approach produced the less RMSE and MAE rates, while allowing direct skip connections between the latest sparse input frame and the predicted output significantly improved on the results at each temporal setup. Using these connections, the network learns to *complete* the missing regions of the sensor's sparse range map, while keeping high fidelity to the accurate range measurements from the last 200 ms time frame.

Next, we also performed hyperparameter optimization steps in the final late fusion based model where we compared different weight combinations of the $\mathcal{L}$ *loss function*'s subterms. The most relevant configurations are summarized in Table II. First, using a relatively higher weight for the $\mathcal{L}_{\mathrm{SSIM}}$ loss term results in smoothed edges and blurred fine structures and therefore it produces higher RMSE and MAE errors. On the other hand, if

TABLE III
COMPARATIVE RESULTS BETWEEN 2D RANGE IMAGES

| Method | iRMSE↓ | iMAE↓ | RMSE↓ | MAE↓ |
|---|---|---|---|---|
| Large integration | 74.745 | 21.12 | 4259.83 | 1119.41 |
| IP-Basic++ [15] | 170.02 | 24.31 | 2918.33 | 574.99 |
| Sparse-to-Dense [10] | 493.65 | 151.55 | 4583.75 | 1672.79 |
| **ST-DepthNet** | **59.46** | **15.94** | **1799.16** | **440.42** |

the weight of $\mathcal{L}_{\mathrm{SSIM}}$ is significantly smaller than the weight of $\mathcal{L}_{\mathrm{L1}}$, image regions with uniform depth remain noisy, resulting again in higher RMSE and MAE rates. As a good balance, we experienced that an optimal ratio between the weights of $\mathcal{L}_{\mathrm{SSIM}}$ and $\mathcal{L}_{\mathrm{L1}}$ is around $1:2$. Second, based on experiments, the point level $\mathcal{L}_{\mathrm{L1}}$ and edge based $\mathcal{L}_{\mathrm{EDGE}}$ loss terms are in the best balance with a weight ratio of around $1:1$.

### C. Reference Methods

We have compared the results of the proposed *ST-DepthNet* model to related approaches published in the recent years. Note that the majority of existing methods [7], [8], [9] relies on fused Lidar based sparse depth maps and dense RGB images, therefore we cannot directly compare the proposed method to them, as we address Lidar-only scenarios [10]. As the first baseline for comparison, we investigated how the sensor itself can produce high density images, by allowing a large integration window ($t_{\triangle} = 1$ s) to cover a high proportion ($>95\%$) of the FoV. We refer to this method from now on as *Large integration*. As the second reference, we adopted an improved version of the method presented in [15], called hereafter as *IP-Basic++*, by optimizing its morphological operations to our irregular NRCS data and extending it with bilateral blurring. We have chosen as the third reference the approach of [10], called henceforward *Sparse-to-Dense*, which is proposed directly for Lidar-only perception, and we trained it on our *LivoxCarla* dataset, with the parameters described in [10]. To adopt the latter method to our dataset, we changed the size of the input layer from $480 \times 480$ to our range image lattice $400 \times 400$.

### D. Comparative Results

Next, we compare the *ST-DepthNet* to the above three reference methods on the *LivoxCarla* test set, in both 2D range image based and 3D point cloud based representations.

The overall mean values of the calculated 2D error rates are displayed in Table III. Regarding all numerical quality measures, the performances of the *Large integration* and the *Sparse-to-Dense* [10] approaches are quite similar, the *IP-Basic++* [15] works better in average, while the proposed *ST-DepthNet* significantly outperforms all of them, reducing their RMSE errors by more than 1 m.

First, we can observe that the main sources for large errors of the *Large integration* method are the movement of the capturing platform and the presence of dynamic objects of the scene. Second, the *IP-Basic++* [15] approach predicts missing depth values more robustly on large, homogeneous surfaces, but fails estimating the fine details. Third, the depth image estimation by the *Sparse-to-Dense* [10] method keeps the trails of the circular scanning pattern of the NRCS sensor still visible, while scene objects and finely textured regions are often merged with their background. Fig. 4 demonstrates these limitations of [10] and [15] on a range image sample, where the proposed method performs significantly better.

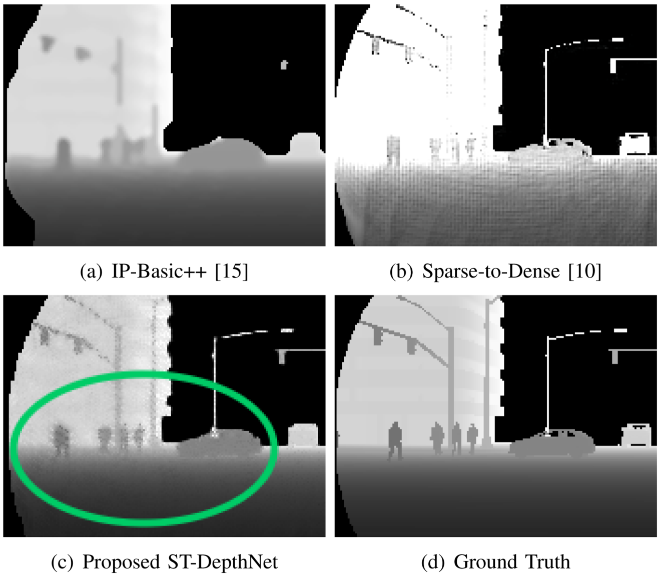

(a) IP-Basic++ [15]
(b) Sparse-to-Dense [10]
(c) Proposed ST-DepthNet
(d) Ground Truth

Fig. 4. Fine structures (marked by green ellipses) recognized by the proposed ST-DepthNet approach, but remained partly or fully unrecognized (merged to wall or background) by the reference IP-Basic++ [15] and Sparse-to-Dense [10] approaches with respect to the Ground Truth data.

TABLE IV
COMPARATIVE RESULTS IN THE 3D SPACE

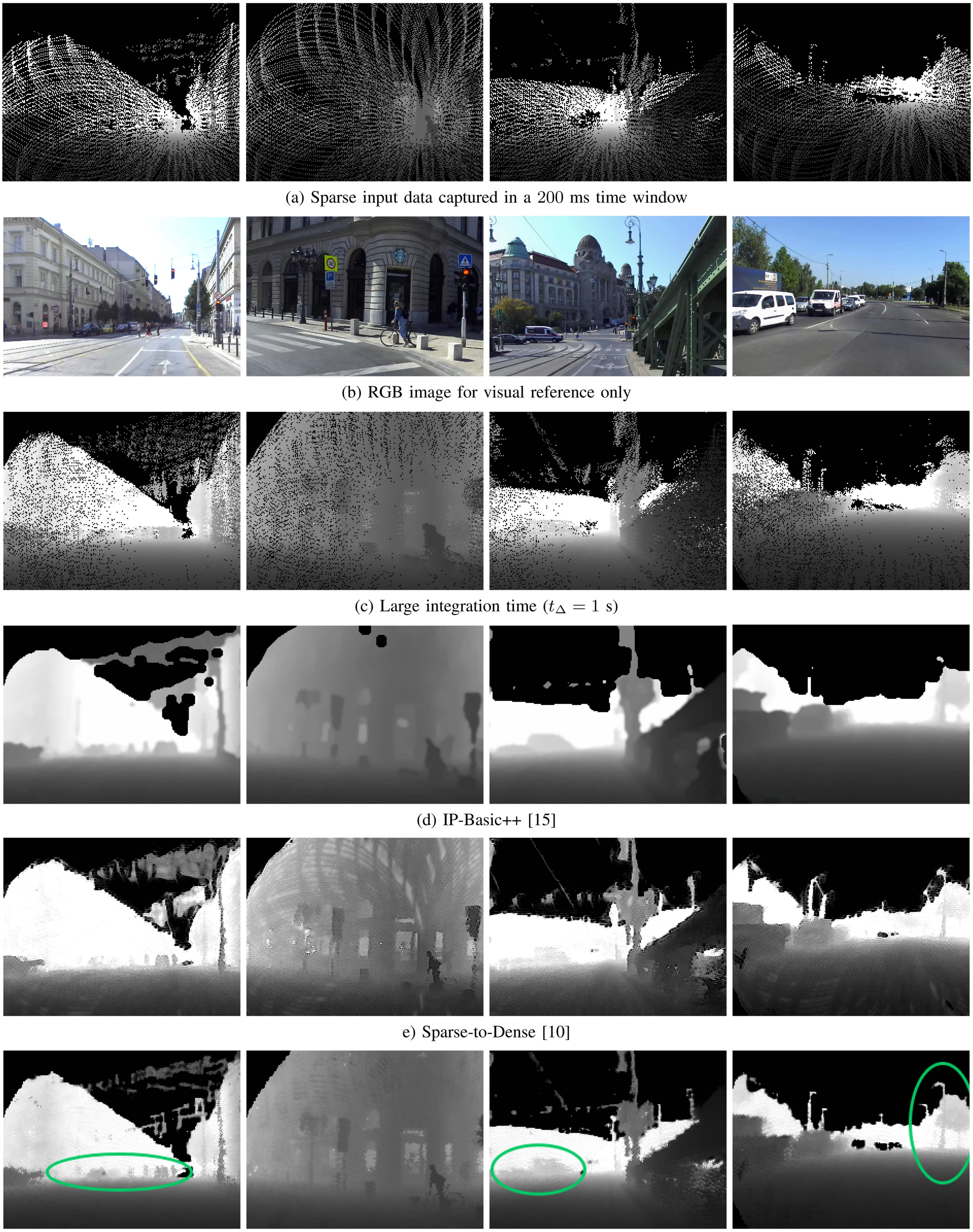
| Method | $Q_{\mathrm{NMD}}$[mm]↓ | $Q_{\mathrm{NCD}}$[mm]↓ |
|---|---|---|
| Large integration | 1754.12 | 3830.62 |
| IP-Basic++ [15] | 1072.60 | 2241.69 |
| Sparse-to-Dense [10] | 4466.14 | 6065.44 |
| **ST-DepthNet** | **687.36** | **1718.53** |

We conducted further analysis in the 3D domain, by comparing the 3D Ground Truth scene models to point clouds back-projected from the range images generated by the proposed and reference methods. Errors obtained by calculating the symmetric Normalized Chamfer and Median Distance metrics are displayed in Table IV. As shown, the error of the proposed method is the smallest, by a margin of around a half meter regarding both metrics. Note that, while *Large integration* seems to work better than *Sparse-to-Dense* in 3D, this observation is mainly the consequence of the fact that object regions affected by motion blur can still have points close to GT in the 3D space, and vice versa.

### E. Analysis on Real Measurements

Beyond a comprehensive numerical evaluation on our synthetic *LivoxCarla* dataset, we also validated the proposed method on real Lidar measurement sequences of the *LivoxBudapest* set, supporting its future real-life application.

The *LivoxBudapest* test set contains three different scenarios: two pathway recordings from the city center (a *boulevard* and a *narrow street*), both around 1 km long, and a *speedway* section



(a) Sparse input data captured in a 200 ms time window

(b) RGB image for visual reference only

(c) Large integration time ($t_\Delta = 1$ s)

(d) IP-Basic++ [15]

e) Sparse-to-Dense [10]

(f) Proposed ST-DepthNet

Fig. 5. Results on real measurements from the *LivoxBudapest* test set. (a) Sparse measurements, (b) visual reference image, (c)–(f) predicted depth maps by different methods. Accurately predicted fine object structures by *ST-DepthNet* are highlighted with green ellipses.

TABLE V
COMPARATIVE SURVEY RESULTS ON REAL MEASUREMENTS

| Method/ score↑ | Boulevard | Narrow st. | Speedway | Mean |
|---|---|---|---|---|
| Sparse input data | 2.95 | 2.75 | 2.60 | 2.76 |
| Large integration | 4.30 | 4.45 | 3.75 | 4.17 |
| IP-Basic++ [15] | 5.75 | 5.65 | 5.60 | 5.67 |
| Sparse-to-Dense [10] | 6.10 | 6.35 | 6.40 | 6.28 |
| **ST-DepthNet** | **7.15** | **7.20** | **6.65** | **7.00** |

near the city, recorded on a path of around 3.5 km. Fig. 5 displays selected relevant sample frames from the three scenarios. We can observe that similarly to the experiments with synthetic data, the *IP-Basic++* [15] approach robustly completes missing values in case of larger surfaces (walls, ground areas and even vehicles), but fails to accurately estimate fine structures. For example, pedestrians in the first column of Fig. 5(d) are blurred into one object, while traffic lights and signs are partly merged to the background in the second and fourth columns of Fig. 5(d). The *Sparse-to-Dense* method cannot eliminate the rosetta patterns of the input Lidar measurements, which are typically visible on ground and wall areas (e.g., second column in Fig. 5(e)). The tendency of merging fine structures into larger surfaces is also notable: In the first and third columns of Fig. 5(e), vehicles and pedestrians are falsely merged to the wall behind them. Such artefacts can mean critical problems for urban scene understanding tasks, while as shown, they are handled better by the proposed *ST-DepthNet* approach (see regions marked by green). The *Sparse-to-Dense* method heavily blurs other fine structures (columns, traffic signs and lights, etc.) as well, as displayed in Fig. 5(e). Moreover, while objects close to the sensor are usually well recognizable for the human eye, they are often predicted at inaccurate distances with this method (e.g., cyclist in the second column of Fig. 5(e)). As for the *Large integration* method, it performs significantly worse on real data than on the simulated samples and its generated range images are extremely noisy. The regions of moving street objects become blurred even if the platform is static, while if the platform is moving, all structures are barely recognizable.

Besides the above qualitative analysis, we conducted a survey for visual verification of the generated depth image streams, by asking 20 computer vision related experts to rate the quality of the input and the output of each method in all three videos with scores between 1 and 9, where 9 is the best possible score. The results provided in Table V confirm that the test subjects found the proposed method significantly better than the reference techniques.

In summary, the proposed *ST-DepthNet* method can better compensate for both the noisiness and the sampling pattern of the sensor data, while the predicted distance values of dynamic objects or background scene structures are more accurate than in the depth maps of the reference approaches. Regarding the computation time, for the prediction of a single depth frame, the *ST-DepthNet* method needs 100 ms.

## V. Conclusion

In this letter we proposed a novel depth completion method called *ST-DepthNet*, which is capable of creating high-density depth images from sparse consecutive depth maps acquired by a NRCS Lidar. For training and quantitative evaluation, we constructed a new synthetic Benchmark set called *LivoxCarla*, and we shown that our approach outperforms two state-of-the-art reference methods. The usability of the proposed method on real NRCS measurement data has also been demonstrated using our recorded *LivoxBudapest* real-life dataset. In the future, we aim to use the proposed method in intelligent robot and vehicle platforms, for improving the limited spatial resolution of NRCS Lidars.

## References

[1] Ö. Zováthi, B. Nagy, and C. Benedek, "Point cloud registration and change detection in urban environment using an onboard LiDAR sensor and MLS reference data," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 110, 2022, Art. no. 102767.

[2] C. Glennie and P. Hartzell, "Accuracy assessment and calibration of low-cost autonomous LiDAR sensors," *ISPRS Arch. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. XLIII-B1-2020, pp. 371–376, 2020.

[3] L. Kovács, M. Kégl, and C. Benedek, "Real-time foreground segmentation for surveillance applications in NRCS LiDAR sequences," *ISPRS Arch. Photogrammetry Remote Sens. Spatial Inf. Sci.*, vol. XLIII-B1-2022, pp. 45–51, 2022.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.

[5] T. Shan, J. Wang, F. Chen, P. Szenher, and B. Englot, "Simulation-based LiDAR super-resolution for ground vehicles," *Robot. Auton. Syst.*, vol. 134, 2020, Art. no. 103647.

[6] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 802–810.

[7] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant CNNs," in *Proc. Int. Conf. 3D Vis.*, 2017, pp. 11–20.

[8] S. Zhao, M. Gong, H. Fu, and D. Tao, "Adaptive context-aware multimodal network for depth completion," *IEEE Trans. Image Process.*, vol. 30, pp. 5264–5276, 2021.

[9] D. Nazir, A. Pagani, M. Liwicki, D. Stricker, and M. Z. Afzal, "SemAttNet: Towards attention-based semantic aware guided depth completion," *IEEE Access*, vol. 10, pp. 120781–120791, 2022.

[10] M. F. F. Khan, N. D. Troncoso Aldas, A. Kumar, S. Advani, and V. Narayanan, "Sparse to dense depth completion using a generative adversarial network with intelligent sampling strategies," in *Proc. ACM Int. Conf. Multimedia*, New York, NY, USA, 2021, pp. 5528–5536.

[11] R. Li, D. Xue, Y. Zhu, H. Wu, J. Sun, and Y. Zhang, "Self-supervised monocular depth estimation with frequency-based recurrent refinement," *IEEE Trans. Multimedia*, early access, Aug. 08, 2022, doi: 10.1109/TMM.2022.3197367.

[12] A. Savkin, Y. Wang, S. Wirkert, N. Navab, and F. Tombari, "LiDAR upsampling with sliced Wasserstein distance," *IEEE Robot. Automat. Lett.*, vol. 8, no. 1, pp. 392–399, Jan. 2023.

[13] C. Godard, O. Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis.*, 2019, pp. 3828–3838.

[14] Y. Kwon, M. Sung, and S. Yoon, "Implicit LiDAR network: Lidar super-resolution via interpolation weight prediction," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2022, pp. 8424–8430.

[15] J. Ku, A. Harakeh, and S. L. Waslander, "In defense of classical image processing: Fast depth completion on the CPU," in *Proc. 15th Conf. Comput. Robot. Vis.*, 2018, pp. 16–22.

[16] B. Nagy, L. Kovács, and C. Benedek, "Change GAN: A deep network for change detection in coarsely registered point clouds," *IEEE Robot. Automat. Lett.*, vol. 6, no. 4, pp. 8277–8284, Oct. 2021.

[17] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[18] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. Annu. Conf. Robot. Learn.*, 2017, pp. 1–16.