2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

# MSDPN: Monocular Depth Prediction with Partial Laser Observation using Multi-stage Neural Networks

Hyungtae Lim[1] , Hyeonjae Gil[2], Hyun Myung[3], *Senior Member, IEEE*

***Abstract*— In this study, a deep-learning-based multi-stage network architecture called Multi-Stage Depth Prediction Network (MSDPN) is proposed to predict a dense depth map using a 2D LiDAR and a monocular camera. Our proposed network consists of a multi-stage encoder-decoder architecture and Cross Stage Feature Aggregation (CSFA). The proposed multi-stage encoder-decoder architecture alleviates the partial observation problem caused by the characteristics of a 2D LiDAR, and CSFA prevents the multi-stage network from diluting the features and allows the network to learn the inter-spatial relationship between features better. Previous works use sub-sampled data from the ground truth as an input rather than actual 2D LiDAR data. In contrast, our approach trains the model and conducts experiments with a physically-collected 2D LiDAR dataset. To this end, we acquired our own dataset called KAIST RGBD-scan dataset and validated the effectiveness and the robustness of MSDPN under realistic conditions. As verified experimentally, our network yields promising performance against state-of-the-art methods. Additionally, we analyzed the performance of different input methods and confirmed that the reference depth map is robust in untrained scenarios.**

## I. INTRODUCTION

Three-dimensional perception plays a key role in both robotics and computer vision fields, such as Simultaneous Localization and Mapping (SLAM), autonomous driving, object recognition, and scene understanding [1], [2]. Many research works have utilized a 3D LiDAR sensor [3], [4] for its high accuracy and long maximum measurement. However, since 3D LiDARs are cost-prohibitive, other existing several depth sensors, e.g. stereo cameras, structured-light-based depth sensors, are also utilized in many tasks as inexpensive alternatives [5]–[8]. Among them, 2D LiDAR sensors are often exploited on mobile robot platforms, since they can not only perceive surroundings with a wide field of view but also have high accuracy thanks to their laser-based measurement principle [9], [10].

In the meanwhile, numerous researchers in computer vision have conducted studies to use a monocular camera for depth estimation due to its low cost, lightweight, and energy-efficiency [11]. Particularly, with the rapid development of research in deep learning, many learning-based studies have been proposed for depth prediction using monocular cameras [12]–[14], producing significant results. However, these approaches have a *global scale ambiguity* issue: depth estimation from the network might be unreliable for robotics applications since monocular images cannot provide global-scale measurements directly [14], [15].

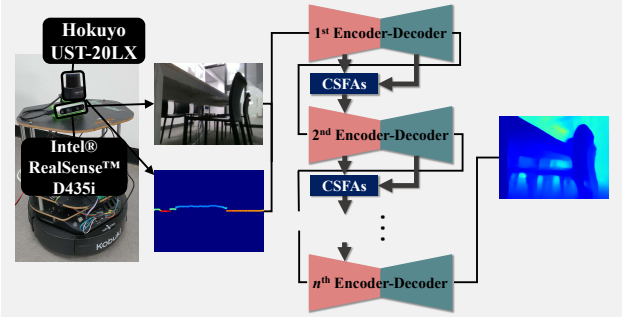



Fig. 1. An overview of the proposed learning-based framework.

To alleviate this issue, some researchers have conducted depth completion tasks, a.k.a sparse-to-dense depth conversion [16]–[18]. Using sparse depth on the image plane, they not only mitigate scale ambiguity but also show significant improvement in depth prediction. Ma and Karaman [17] conducted thorough analysis on the performance of the deep neural network with different number of sparse depth and suggested many applications. Lee *et al.* [16] and Zhang and Funkhouser [19] suggested novel deep feature-guided methods. Cheng *et al.* [18] suggested efficient spatial feature propagation module and showed promising performance improvement.

By extension, Liao *et al.* [15] and Yin *et al.* [20] suggested novel methods for predicting depth using an image and 2D LiDAR scans to mitigate the limitation of 2D LiDAR scans: extremely limited vertical field of view. Unlike sparse-to-dense works, the authors reported that the network can hardly learn useful information from the projected depth image generated by 2D laser scans due to the extremely partially sparse distribution, i.e. most range measurements are concentrated horizontally on the image plane. Besides, Liao *et al.* argued that coexistence of zero-padded values, which are kind of boolean values to denote whether the laser range data are projected or not, and valid depth values might confuse the network.

[1]Hyungtae Lim and [3]Hyun Myung are with School of Electrical Engineering, KI-AI, KI-R at KAIST (Korea Advanced Institute of Science and Technology) and [2]Hyeonjae Gil is a undergraduate intern of the laboratory at KAIST, Daejeon, 34141, South Korea. {shapelim, now9728, jungmokoo, hmyung}@kaist.ac.kr

This work was supported in part by the Industrial Convergence Core Technology Development Program (Development of Robot Intelligence Technology for Mobility with Learning Capability Toward Robust and Seamless Indoor and Outdoor Autonomous Navigation) funded by the Ministry of Trade, Industry and Energy (MOTIE), South Korea, under Grant 10063172 and in part by the Industry Core Technology Development Project (Development of Artificial Intelligence Robot Autonomous Navigation Technology for Agile Movement in Crowded Space) funded by the MOTIE, South Korea, under Grant 20005062.

For these reasons, Liao *et al.* [15] introduced a novel input-level depth information propagation method, called *reference depth map*, which is generated by extending projected 2D observations in the image plane along the gravity direction. Through the reference depth map and their proposed networks, they resolved the partial observation issues and showed prominent performance improvement.

In this paper, a multi-stage network with Cross Stage Feature Aggregation (CSFA) module called Multi-Stage Depth Prediction Network (MSDPN) is proposed as shown in Fig. 1. To the best of our knowledge, it is the first approach that applies multi-stage architecture to mitigate the partial observation problem in realistic environments and, in contrast with previous works [15], [20], it utilizes a physically-collected 2D LiDAR dataset. Unlike previous single-stage networks which downsample and upsample the feature only once, the MSDPN employs the network-level depth information propagation method to propagate partially distributed spatial information in the feature to upward and downward regions through repeated upsampling and downsampling.

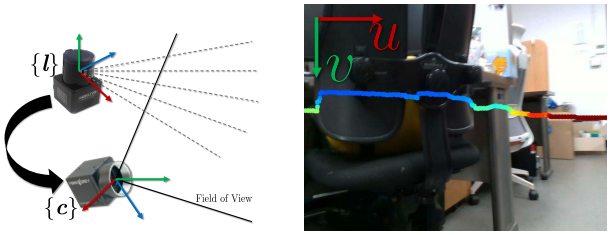
The contribution of this paper is threefold:

- The multi-stage network on which CSFAs are attached is proposed to allow the neural network to prevent saturation and to learn the inter-spatial relationship of the features so that the MSDPN yields the better and acceptable depth map compared to previous approaches with fewer number of parameters.
- Unlike previous works which use sub-sampled data as if they were obtained from an actual 2D LiDAR, this paper has trained the model and conducted research using a physically collected dataset from a 2D LiDAR. Accordingly, we established an indoor dataset using a 2D LiDAR and a RGB-D camera, called KAIST RGBD-scan dataset.
- We also verify how each component of the MSDPN affects the performance in the ablation study. Next, the performance comparison with different number of laser scans and different types of input method are investigated. Finally, we suggest an appropriate input method for each 2D LiDAR sensor in realistic applications.

The rest of the paper is organized as follows: Section II introduces the method for projecting laser scans on the image plane which is utilized in acquiring KAIST RGBD-scan dataset. Section III presents our proposed multi-stage neural network in detail. Section IV describes the experiments, and Section V examines the experimental results. Finally, Section VI summarizes our contributions and describes future works.

## II. Sensor Systems

In this section, the characteristics of the sensor system which consists of a 2D LiDAR and a monocular camera will be examined, as illustrated in Fig. 2. Specifically, the method for projecting 2D LiDAR scans onto the image plane of the RGB camera is explained, which is utilized in acquiring KAIST RGBD-scan dataset (See Section IV.*A*), and the distribution of the laser scan hits on the image plane is analyzed. For simplicity, we use `proj-d` for the projected depth image, `RGB` for the RGB image, and `ref-d` for the reference depth map [15].




(a) (b)

Fig. 2. (a) An overview of the system. (b) A visualized example where the projected depth (`proj-d`) is overlapped on the image.

To project laser scans on the image plane, the extrinsic and intrinsic parameters should be estimated. Let $\mathbf{p}^l = (x^l, y^l, z^l)^T$ and $\mathbf{p}^c = (x^c, y^c, z^c)^T$ be each point in the 2D LiDAR coordinate system and in the camera coordinate system, respectively. Let us denote a rigid body transformation as $\mathcal{T} \in SE(3)$ that consists of corresponding rotation $\mathbf{R} \in SO(3)$ and translation $\mathbf{t} \in \mathbb{R}^3$. The relationship between $\mathbf{p}^c$ and $\mathbf{p}^l$ can be formulated as:

$$\mathbf{p}^c = \mathbf{R}\mathbf{p}^l + \mathbf{t}. \tag{1}$$

Next, according to a pinhole camera model, $\mathbf{p}^c$ is projected onto the $\mathbf{p}^{pixel} = (u, v)^T$ by triangulation, where $\mathbf{p}^{pixel}$ denotes the pixel coordinate in the image plane. The corresponding equations are as follows:

$$s\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{K}\mathbf{p}^c = \begin{bmatrix} f_x & \alpha & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} x^c \\ y^c \\ z^c \end{bmatrix} \tag{2}$$

where $\mathbf{K}$ is the camera's intrinsic matrix consisting of focal lengths $f_x$, $f_y$, a principal point $(c_x, c_y)^T$, and a skew coefficient $\alpha$. $s$ denotes the scale factor of the image plane. It should be noted that the scans located beyond the field of view (FOV) of the camera are filtered out.

In short, when value of $y_c$ is not large, i.e. when a 2D LiDAR sensor is deployed close to a camera on the y-axis of the camera coordinate, $v$ value is close to $c_y$ in the majority of points. For example, in KAIST RGBD-scan dataset, the mean and the standard deviation of the minimum $v$ positions on the image plane are 101.2 and 8.8, respectively when the input size is resized to $304 \times 228$. Furthermore, the top 90% (10th percentile) exists between 89 and 117 on the $v$-axis.

## III. Network Architecture

In this section, our proposed Multi-Stage Depth Prediction Network (MSDPN) will be explained. Specifically, we introduce multi-stage encoder-decoder-based network and Cross Stage Feature Aggregation (CSFA) to propagate features to the next stage for more well-described features. The encoding part is based on ResNet-18 [21], whose last average pooling and linear transformation layer are neglected, and `UpProj`

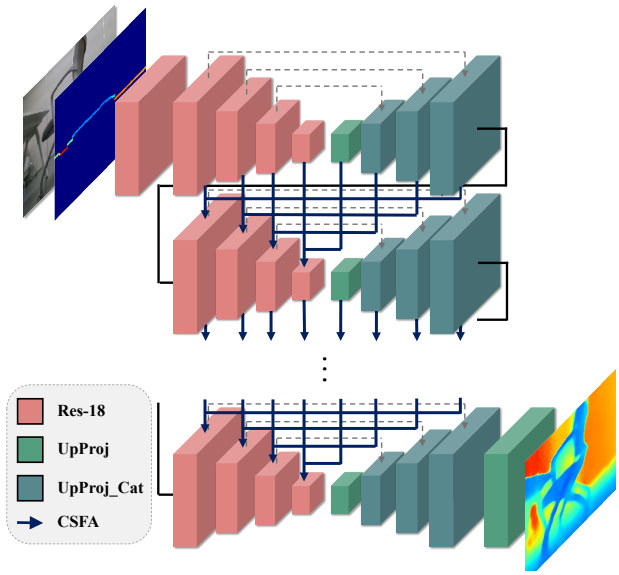



Fig. 3. Our Multi-Stage Depth Prediction Network (MS-DPN). The networks is composed of a) multi-stage U-shaped architecture whose backbone is ResNet-18 and b) Cross Stage Feature Aggregation (zoomed-in Figure 4) (best viewed in color).

[22] and `UpProj_Cat` [18] are used in the decoder part. The single module structure itself does not have novelty, but it is the first time to apply it in a multi-stage encoder-decoder network with CSFA to resolve the partial observation issues effectively. Besides, through the ablation study, it is shown that the modules not only require fewer parameters but also show better performance compared to single-stage architectures (See Section V.*B* and Table III)

### *A. Multi-stage Architecture*

As illustrated in Fig. 3, our multi-stage network consists of $N$ encoder-decoder-based neural networks. Compared to a single-stage network, which extracts representative features in the down sampling process and recovers the lost information in the upsampling procedure only once, a multi-stage architecture is able to estimate the final depth with the refined feature. The aforementioned partial observation problem, where 2D laser scans are mostly located horizontally on the center of the image, causes spatial information imbalance of the feature. In other words, only the features in the middle contain actual range information directly obtained by the laser scans, whereas the other features are extracted by `RGB`.

For example, many single-stage studies utilize the ResNet-50 [21] as the backbone [16]–[18]. When the input is fed into the ResNet-50-based architecture for KAIST RGBD-scan dataset, the direct depth cues completely propagate from the middle to the upward and downward regions of feature after passing through `conv4_x`. That is to say, the spatial information is completely propagated when the feature reaches the final layer of the encoder part.

Unlike single-stage architectures, our multi-stage networks consist of a number of encoder-decoder architectures, so the depth information within feature becomes less uneven due to the repeated upsampling and downsampling. As a consequence, this repetition alleviates the partial observation issue so that relatively well-balanced feature compared to that of single-stage networks could help the network predict refined depth.

### *B. Cross Stage Feature Aggregation*

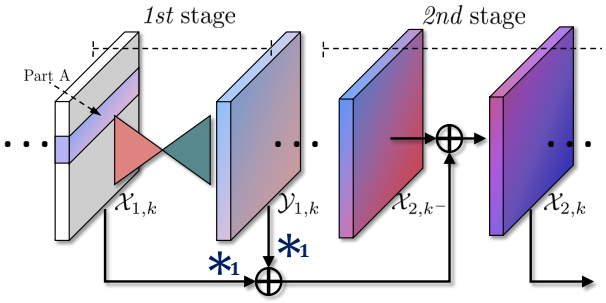



Fig. 4. The procedure of the Cross Stage Feature Aggregation, especially between the first and second stages. The colored part (Part A) denotes the region where valid laser range information is propagated directly to the feature (best viewed in color).

Despite the well-balanced feature by using the multi-stage architecture, the architecture is associated with potential risks: it might undergo saturation and thus become vulnerable to diluting the information during repeated up and down sampling [23]. Thus, we adopted the Cross Stage Feature Aggregation (CSFA), which is inspired by [21], [24] to alleviate this problem as follows:

$$\mathcal{X}_{n,k} = \mathcal{X}_{n,k^-} \oplus (\phi(\mathcal{X}_{n-1,k}) \oplus \psi(\mathcal{Y}_{n-1,k})) \tag{3}$$

where $\phi(\cdot)$ and $\psi(\cdot)$ are learnable transformations, i.e. $1\times1$ convolution, $\oplus$ is element-wise addition, $\mathcal{X}_{n,k}$ denotes the input feature of each $k$-th block of ResNet-18 in the $n$-th stage, and $\mathcal{Y}_{n,k}$ denotes the output feature of each $k$-th block of decoder in the $n$-th stage, respectively. $k^-$ is the output feature of $(k-1)$-th block of the encoder.

As shown in Fig. 4, prior features from the encoder and decoder are introduced to the encoder part of the next stage. By propagating multi-scale features from early stages to the current stage, CSFA can a) prevent the network from diluting the feature and b) allow it to learn better the inter-spatial relationship between features.

### *C. Loss Function*

To train the proposed network, different $L_1$ term is used based on the types of the input method, i.e. `proj-d` or `ref-d`. When `proj-d` is taken as an input, the loss function is defined as follows:

$$\mathcal{L}_r = \left\| \kappa_{\{\mathcal{D}>0\}} \cdot (\mathcal{D} - \hat{\mathcal{D}}) \right\|_1 \tag{4}$$

where $\mathcal{D}$ is the ground truth depth and $\hat{\mathcal{D}}$ is the estimated depth. $\kappa$ denotes a pixel-wise filter that is equal to one when the pixel value of ground truth is available or zero otherwise.

On the contrary, when ref-d is taken as an input, the loss function is defined as follows:

$$\mathcal{L}_r = \left\| \kappa_{\{\mathcal{D}>0\}} \cdot (\mathcal{D} - (\hat{\mathcal{D}}_{res} + \mathcal{R})) \right\|_1 \quad (5)$$

where $\hat{\mathcal{D}}_{res}$ is the predicted residual depth from the end part of our network and $\mathcal{R}$ indicates ref-d.

Cao *et al.* [25] argued that the softmax classification loss, $\mathcal{L}_c$, was better than $\mathcal{L}_r$ and Liao *et al.* [15] demonstrated that the tightly coupled loss term between $\mathcal{L}_r$ and $\mathcal{L}_c$ was better than $\mathcal{L}_c$ only. However, we noticed that only using $\mathcal{L}_r$ showed better performance than using combination of the classification loss terms $\mathcal{L}_c$ and $\mathcal{L}_r$ from [15] in realistic environments (see Section V.*A* for its rationale).

## IV. Experiments

### A. Dataset

**KAIST RGBD-scan Dataset** Synchronized RGB-D and 2D LiDAR scan data were collected to evaluate accuracy and applicability of the experiment in physical indoor environments. Data were collected by a mobile robot loaded with Intel RealSense D435i [26] and Hokuyo UST-20LX [27], as shown in Fig. 1. Sampling frequency of the RGB-D sensor was dropped to 15Hz followed by post-processing of the collected data to remove overlapping scenes. Our dataset consists of 14,143 training data for 99 scenes and 1,042 test data for 53 scenes, all of which were acquired from KAIST campus. Note that the distinct interior design of individual in-campus buildings makes the indoor scenes of KAIST much appropriate for a dataset.

**KITTI Odometry Dataset** Unfortunately, RealSense D435i depth sensor is too noisy to be qualified as a ground truth for evaluating depth prediction [28]. Thus, for clear comparison, our network was also trained and evaluated using KITTI [29], which is an outdoor dataset. We followed the procedure in [17] and simulated 2D laser scans were sub-sampled from the Velodyne 64E 3D scanner based on [15].

**NCLT Dataset** The University of Michigan North Campus Long-Term Vision and LiDAR (NCLT) dataset provides synchronized 3D LiDAR scans, 2D LiDAR scans, and images in both indoor and outdoor environments [30]. This dataset was utilized to check which input method type, i.e. ref-d or proj-d, is more robust under untrained scenarios. 9,733 samples from three categories ("2012-01-08", "2012-03-25", "2012-05-11") are used for the training dataset and 1,050 samples from the four categories ("2012-01-15", "2012-03-31", "2012-08-04", "2012-11-04") were used for the test dataset. Note that all training data were captured only in daytime, e.g. sunny or partly cloudy weathers, whereas the test data additionally includes some scenes with foliage and some in snowy and cloudy weathers. Because of the relatively small training set, the network was initialized by the weights learned from KITTI.

### B. Training the Network

To train our network, the Adam optimizer [31] was exploited for 30 epochs with learning rate of 0.0001, momentum of 0.9, weight decay of 0.0001, decay rate of 0.98, and batch size of 20. Note that weight decays for every epoch. Our network was modeled by PyTorch and was trained with two NVidia TITAN Xp GPUs.

### C. Error Metrics

For evaluation, we followed some metrics commonly utilized in depth prediction area [15], [17], [18]. Let $\mathcal{D}_l$ and $\hat{\mathcal{D}}_l$ be the ground truth and the estimated depth on each pixel $l$, respectively. Then metrics are as following:

- Root Mean Squared Error (RMSE): $\sqrt{\frac{1}{N}\sum_l (\hat{\mathcal{D}}_l - \mathcal{D}_l)^2}$
- Mean Absolute Relative Error (REL): $\frac{1}{N}\sum_l \frac{|\hat{\mathcal{D}}_l - \mathcal{D}_l|}{\mathcal{D}_l}$
- $\delta_n$: percentage of $\mathcal{D}_l$, such that $\max(\frac{\hat{\mathcal{D}}_l}{\mathcal{D}_l}, \frac{\mathcal{D}_l}{\hat{\mathcal{D}}_l}) < \delta^n$, $\delta$ = 1.25 and $n$ = 1, 2, 3.

where $N$ is the total number of pixels.

## V. Results and Discussion

Keep in mind that the visualization of all the sparse depth measured from both 2D and 3D LiDAR sensors has been magnified for ease of understanding. Originally, each range data is represented in only one pixel. As previous works [15]–[17] show that RGB with some sparse depth leads to better depth prediction results compared to those from RGB only, we instead gave more emphasis on the comparison of encoding methods for 2D LiDAR scans, i.e. ref-d and proj-d. Thus, the symbol RGB is omitted in front of the symbols ref-d and proj-d for simplicity in our discussion.

### A. Ablation Study

**Loss Function** Through the ablation study, it is noted that combination of the loss terms with ref-d shows more promising performance than that with proj-d. However, for our proposed networks, it is shown that only using $\mathcal{L}_r$ leads to better performance than using combination of $\mathcal{L}_r$ and $\mathcal{L}_c$ as shown in Table I.

TABLE I: Ablation study: performance with different loss functions and input method types on the KAIST RGBD-scan dataset when the number of stage $N$=2.

| Loss Function | Input method | RMSE [mm] | REL | $\delta_1$ [%] |
|---|---|---|---|---|
| $L_r + L_c$ | proj-d | 540 | 0.122 | 88.4 |
| | ref-d | 504 | 0.101 | 89.9 |
| $L_r$ only | proj-d | 496 | 0.098 | 90.0 |
| | ref-d | **491** | **0.096** | **90.5** |

**Cross Stage Feature Aggregation** After choosing the loss function as $\mathcal{L}_r$, the effect of CSFA on the performance is analyzed. As depicted in Table II, it is shown that CSFA leads to significant performance improvement.

It is shown that naïvely stacking the two networks ($N$=2) demonstrates rather worse performance than having one

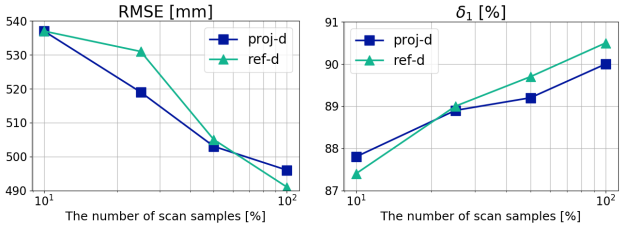
network ($N$=1). Besides, employing 1×1 convolutions enhances the depth accuracy and leads to larger performance improvement than merely connecting the networks. Thus, the result implies that CSFA mitigates the saturation issue of multi-stage networks in depth prediction tasks.

TABLE II: Ablation study: performance comparison based on the presence or absence of each component of CSFA on the KAIST RGBD-scan dataset.

| Stages | Connection | Conv. | RMSE [mm] | REL | $\delta_1$ [%] |
|---|---|---|---|---|---|
| 1×Res-18 | | | 512 | 0.107 | 89.2 |
| 2×Res-18 | | | 516 | 0.107 | 89.2 |
| | ✓ | | 506 | 0.103 | 89.9 |
| | ✓ | ✓ | 491 | **0.096** | **90.5** |
| 4×Res-18 | ✓ | ✓ | **485** | 0.098 | 90.0 |

**Multi-stage Architecture** As reported in Table II, the increasing number of $N$ leads to performance improvement. Specifically, our 2-stage network led to 21mm improvement over 1-stage network and obtained 491mm of RMSE. These experiments indicate that multi-stage architecture successfully mitigates the partial observation issues for better depth prediction in terms of errors. Additionally, estimated depth from the 4-stage network has smaller errors with the reliable geometry compared to 2-stage network, yet its performance gain is decreased.

**Impact of the number of 2D Laser Scans** As reported in Table III, taking `RGB` with `ref-d` as input led to smaller error compared to `RGB` with `proj-d`. Accordingly, performance analysis with different number of laser scans are conducted on KAIST RGBD-scan dataset. As described in Fig. 5, the `ref-d` outperforms `proj-d` with over 50% of scan dropout. On the other hand, `proj-d` has a tendency to outperform `ref-d` as the sample size decreases under 50%.

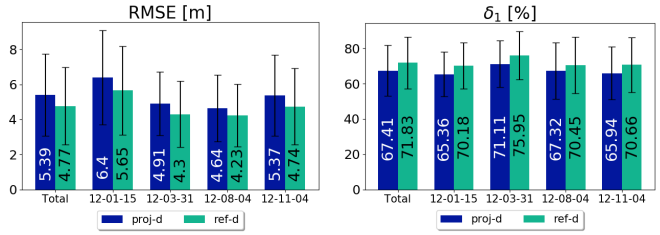



(a) (b)

Fig. 5. Performance changes with varying number of scan samples. (a) RMSE (*the lower, the better*) (b) $\delta_1$ (*the higher, the better*) with the scans dropout.

Thus, we could draw a conclusion that one should choose the encoding method depending on one's circumstances: a) `ref-d` is suitable for 2D LiDARs with high resolution, e.g. angular resolution is higher than 0.25° like Hokuyo UST-20LX [27]. b) `proj-d` is suitable for those with less dense resolution, e.g. angular resolution is lower than 1.0° like LDS-01 sensor [32] in realistic environments.

### *B. Robustness on the Untrained Scenarios*




(a) (b)

Fig. 6. Performances with respect to different input methods. (a) RMSE (*the lower, the better*) (b) $\delta_1$ (*the higher, the better*) with each date class of the test data on NCLT dataset.

In untrained scenarios, which are not included in the train dataset, accuracy of `ref-d` is better on every date as described in Fig. 6. Specifically, the `ref-d` outperforms `proj-d` with 4.77m of RMSE and 71.83% of $\delta_1$. Especially, the performance gap also increases on the "12-01-15" test data, which are captured on snowy day so that there exists huge photometric differences with the training dataset.

The visualized results are presented in Fig. 7. The tested scenes are also included in the train dataset with different weathers so that it becomes hard to predict depth precisely. Both results of `ref-d` and `proj-d` imply that they both can precisely estimate the depth of a region where the laser actually scans. However, since `ref-d` propagates geometrical information gathered by the laser scan vertically, the network with `ref-d` can estimate more refined depth of an image that has objects with vertically identical depth. Thus, the experiment shows that taking `ref-d` is more robust under untrained conditions, especially for realistic environments where geometrically structured objects are present.

### *C. Comparison with State-of-the-arts*

Our best model investigated in Section V.*A* was compared with the CNN-based state-of-the-art methods quantitatively, Liao *et al.* [15], Ma and Karaman [17], and Cheng *et al.* [18], trained on the KAIST RGBD-scan dataset and KITTI.

**KAIST RGBD-Scan Dataset** The results on KAIST RGBD-scan dataset are reported in Table III and shown in Fig. 8. We can see that results of [15] and [17] give acceptable and precise depth maps, filtering out undesirable irradiation errors caused by the characteristics of the RGB-D camera. As reported in Table III, our MSDPN exhibits most promising results with more sharp edges and refined depths for both types of input. Specifically, our method yields smaller errors with fewer parameters and FLOPs.

Furthermore, it was checked that the MSDPN propagates the depth information effectively to upward and downward regions so that it learns inter-spatial relationships better than other state-of-the-arts methods. As shown in the upper bounding box of the third row of Fig. 8, Due to its own propagation method and multi-stage architecture, our MSPDN estimates more precise depth of the region where direct depth

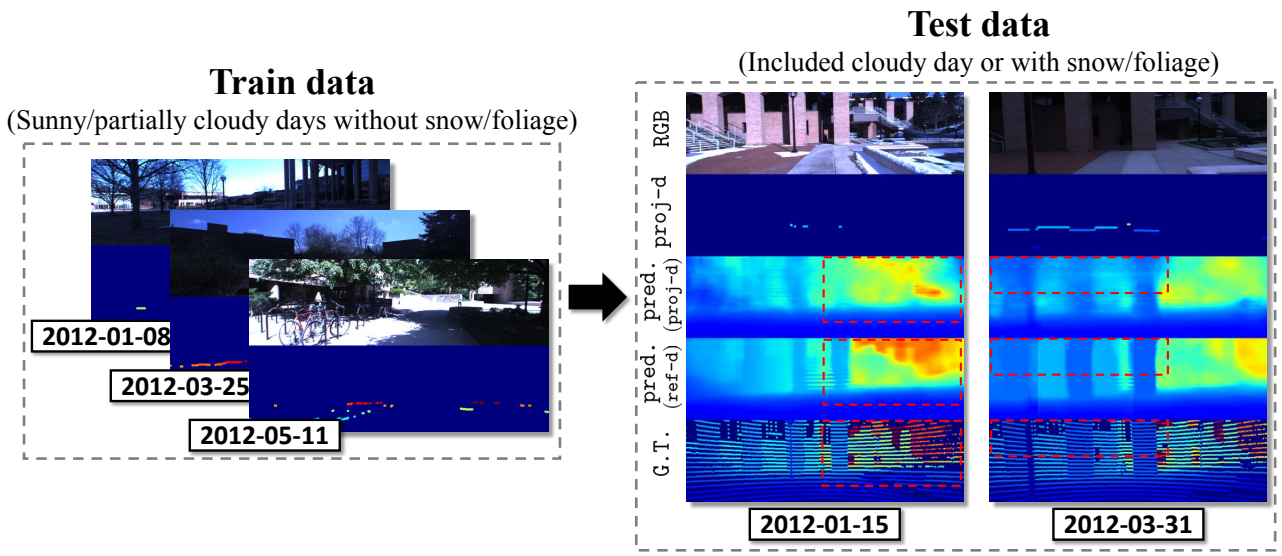
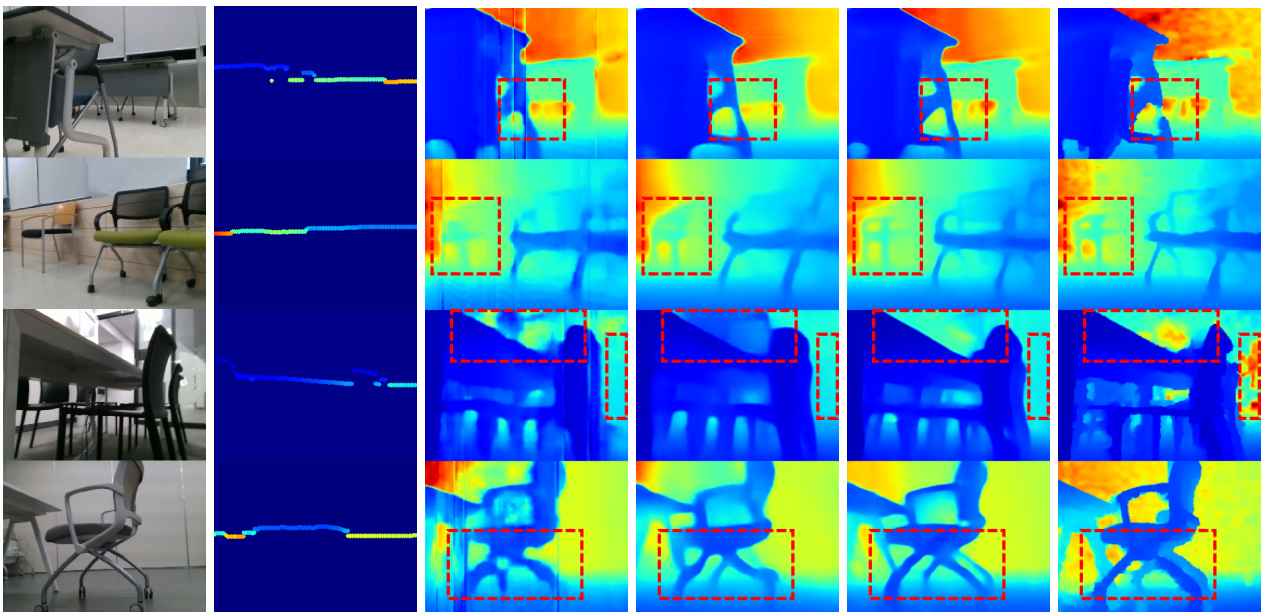



Fig. 7. An overview of robustness test with different types of input method and visualized results on NCLT dataset.



(a) RGB (b) Laser scans (c) [15] (d) [17] (e) Ours (f) Ground truth

Fig. 8. Comparison of depth estimation results with state-of-the-arts on KAIST RGBD-scan dataset.

TABLE III: Comparison with state-of-the-arts on the KAIST RGBD-scan dataset

| Input | Method | Backbone | Params | FLOPs | Error *(Lower the Better)* | | Accuracy *(Highter the better)* | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | RMSE [mm] | REL | $\delta_1$ [%] | $\delta_2$ [%] | $\delta_3$ [%] |
| RGB+proj-d | Liao *et al.* [15] | Res-50 | 48.7M | 39.8G | 640 | 0.132 | 83.8 | 93.7 | 97.0 |
| | Ma and Karaman [17] | Res-50 | 63.6M | 42.8G | 504 | 0.102 | 89.7 | 96.0 | 98.0 |
| | MSDPN (Ours) | 2×Res-18 | **47.4M** | **30.0G** | **496** | **0.098** | **90.0** | **96.1** | **98.0** |
| RGB+ref-d | Liao *et al.* [15] | Res-50 | 48.7M | 39.8G | 534 | 0.109 | 87.9 | 95.4 | 97.8 |
| | Ma and Karaman [17] | Res-50 | 63.6M | 42.8G | 516 | 0.107 | 89.2 | 95.8 | 98.0 |
| | MSDPN (Ours) | 2×Res-18 | **47.4M** | **30.0G** | **491** | **0.096** | **90.5** | **96.3** | **98.2** |

cues are missing than other single-stage networks. Therefore, it is verified that repeated upsampling and downsampling of feature in multiple encoder-decoder architecture mitigates the partial observation issue.

**KITTI Odometry Dataset** All quantitative results for KITTI dataset are reported in Table IV. We trained [17],

[18], and ours on KITTI. As shown in Fig. 9, the experiments showed that all state-of-the-arts alleviate the global ambiguity issue by utilizing simulated 2D LiDAR scans compared to those that take RGB as input. Among them, our MSDPN yields smaller RMSE and higher $\delta_1$ compared to other methods. Furthermore, it is noted that our MSDPN presents significant depth accuracy when taking RGB as input compared to other methods. Therefore, we draw a conclusion that MSDPN not only somehow addresses the partial distribution issue effectively, but also proves itself to be an effective way to predict depth in general.

TABLE IV: Comparison with state-of-the-arts on the KITTI Odometry dataset. Results of RGB are quoted from [17].

| Input | Method | RMSE [m] | REL | $\delta_1$ [%] |
|---|---|---|---|---|
| RGB | Make3D [22] | 8.734 | 0.280 | 60.1 |
| | Mancini [33] | 7.508 | - | 31.8 |
| | Eigen *et al.* [14] | 7.156 | 0.190 | 69.2 |
| | Ma and Karaman [17] | 6.266 | 0.208 | 59.1 |
| | MSDPN (Ours) | **4.656** | **0.104** | **87.3** |
| RGBd | Liao *et al.* [15] | 4.500 | 0.113 | 87.4 |
| | Ma and Karaman [17] | 4.093 | **0.084** | 91.1 |
| | Cheng *et al.* [18] | 4.017 | 0.090 | 90.9 |
| | MSDPN (Ours) | **3.834** | 0.090 | **91.5** |

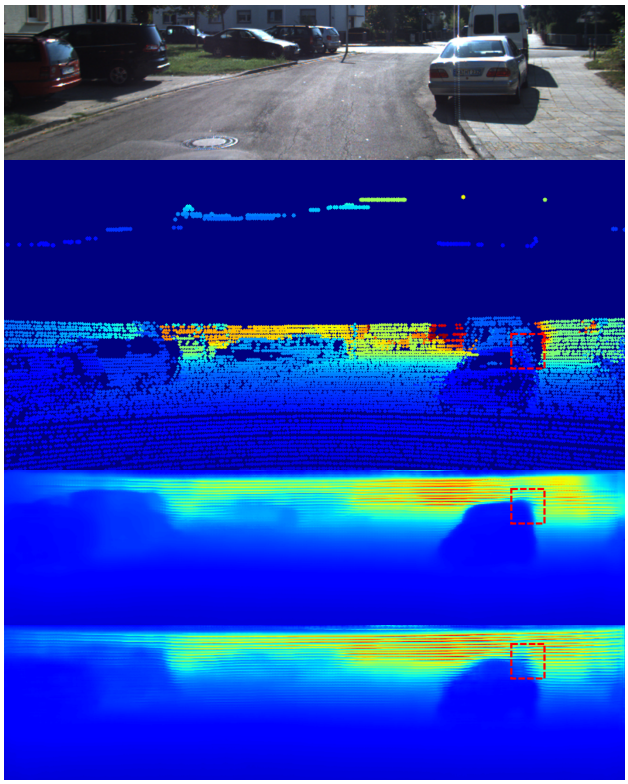


Fig. 9. Qualitative comparison with state-of-the-arts on KITTI (T-B): RGB, simulated 2D laser scans (proj-d), ground truth, Cheng *et al.* [18], and our (best viewed in color).

## VI. Conclusion

In this study, an encoder-decoder-based multi-stage deep learning architecture, namely Multi-Stage Depth Prediction Network (MSDPN), has been proposed for efficient depth prediction from 2D LiDAR scans and a monocular image. We also suggested Cross Stage Feature Aggregation to prevent the network from saturating and to allow the network to learn inter-spatial relationships within the features. Our proposed network was tested on realistic environments: using not only the sub-sampled input data but also physically collected dataset from a 2D LiDAR. Accordingly, through the ablation study and the comparison with state-of-the-art methods, the effectiveness of the MSDPN was verified and it was shown that our proposed method yielded the most precise depth estimations both quantitatively and qualitatively. Finally, the robustness of the different input methods are also tested in the untrained situations.

In future works, our loss function will be investigated to refine depth better, by introducing consistencies as suggested in [16]. Furthermore, this approach could be applied to path planning of mobile robot platforms. Additionally, since the mechanism of our system is similar to the structured light range finder systems [34], [35], we will check whether our system is applicable to the structured light systems or not.

## References


[1] H. Durrant-Whyte and T. Bailey, "Simultaneous Localization and Mapping: Part I," *IEEE Robotics & Automation Magazine*, vol. 13, no. 2, pp. 99–110, 2006.

[2] T. Bailey and H. Durrant-Whyte, "Simultaneous Localization and Mapping (SLAM): Part II," *IEEE Robotics & Automation Magazine*, vol. 13, no. 3, pp. 108–117, 2006.

[3] X. Chen, A. Milioto, E. Palazzolo, P. Giguère, J. Behley, and C. Stachniss, "SuMa++: Efficient LiDAR-based semantic SLAM," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 4530–4537.

[4] B. Li, T. Zhang, and T. Xia, "Vehicle detection from 3D LiDAR using fully convolutional network," *arXiv preprint arXiv:1608.07916*, 2016.

[5] S. Khattak, C. Papachristos, and K. Alexis, "Vision-depth landmarks and inertial fusion for navigation in degraded visual environments," in *Proc. International Symposium on Visual Computing*, 2018, pp. 529–540.

[6] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 681–687.

[7] D. Lee, H. Kim, S. Jung, and H. Myung, "Depth-hybrid speeded-up robust features (DH-SURF) for real-time RGB-D SLAM," *Advances in Robotics Research*, vol. 2, no. 1, p. 33, 2018.

[8] H. Kim, S. Song, J. Hyun, S. H. Hong, and H. Myung, "RGB-D and magnetic sequence-based graph SLAM with kidnap recovery," in *Proc. International Conference on Control, Automation and Systems (ICCAS)*, 2018, pp. 1440–1443.

[9] H. Kim, S. Song, and H. Myung, "GP-ICP: Ground plane ICP for mobile robots," *IEEE Access*, vol. 7, pp. 76 599–76 610, 2019.

[10] A. Cherubini, F. Spindler, and F. Chaumette, "Autonomous visual navigation and laser-based moving obstacle avoidance," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 5, pp. 2101–2110, 2014.

[11] A. Saxena, S. H. Chung, and A. Y. Ng, "3-D depth reconstruction from a single still image," *International Journal of Computer Vision*, vol. 76, no. 1, pp. 53–69, 2008.

[12] R. Garg, V. K. BG, G. Carneiro, and I. Reid, "Unsupervised CNN for single view depth estimation: Geometry to the rescue," in *Proc. European Conference on Computer Vision (ECCV)*, 2016, pp. 740–756.

[13] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 270–279.

[14] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in Neural Information Processing Systems*, pp. 2366–2374, 2014.

[15] Y. Liao, L. Huang, Y. Wang, S. Kodagoda, Y. Yu, and Y. Liu, "Parse geometry from a line: Monocular depth estimation with partial laser observation," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 5059–5066.

[16] B.-U. Lee, H.-G. Jeon, S. Im, and I.-S. Kweon, "Depth completion with deep geometry and context guidance," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2019, pp. 3281–3287.

[17] F. Ma and S. Karaman, "Sparse-to-dense: Depth prediction from sparse depth samples and a single image," in *Proc. IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 1–8.

[18] X. Cheng, P. Wang, and R. Yang, "Depth estimation via affinity learned with convolutional spatial propagation network," in *Proc. European Conference on Computer Vision (ECCV)*, 2018, pp. 103–119.

[19] Y. Zhang, S. Khamis, C. Rhemann, J. Valentin, A. Kowdle, V. Tankovich, M. Schoenberg, S. Izadi, T. Funkhouser, and S. Fanello, "ActiveStereoNet: End-to-end self-supervised learning for active stereo systems," in *Proc. European Conference on Computer Vision (ECCV)*, 2018, pp. 784–801.

[20] P. Yin, J. Qian, Y. Cao, D. Held, and H. Choset, "FusionMapping: Learning Depth Prediction with Monocular Images and 2D Laser Scans," *arXiv preprint arXiv:1912.00096*, 2019.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[22] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. IEEE International Conference on 3D Vision (3DV)*, 2016, pp. 239–248.

[23] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *Proc. European Conference on Computer Vision (ECCV)*, 2016, pp. 483–499.

[24] W. Li, Z. Wang, B. Yin, Q. Peng, Y. Du, T. Xiao, G. Yu, H. Lu, Y. Wei, and J. Sun, "Rethinking on multi-stage networks for human pose estimation," *arXiv preprint arXiv:1901.00148*, 2019.

[25] Y. Cao, Z. Wu, and C. Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 11, pp. 3174–3182, 2017.

[26] "Intel RealSense D435i," https://www.intelrealsense.com/depth-camera-d435i/, 2018, [Online; accessed 28 Feb. 2020].

[27] "HOKUYO UST-20LX," https://www.hokuyo-aut.jp/search/single.php?serial=167, 2009, [Online; accessed 28 Feb. 2020].

[28] "Intel Realsense D435 vs Microsoft Kinect V2," https://youtu.be/O0hSzay3fns, 2009, [Online; accessed 28 Feb. 2020].

[29] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 3354–3361.

[30] N. Carlevaris-Bianco, A. K. Ushani, and R. M. Eustice, "University of Michigan North Campus long-term vision and LiDAR dataset," *The International Journal of Robotics Research*, vol. 35, no. 9, pp. 1023–1035, 2016.

[31] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[32] "360 Laser Distance Sensor LDS-01," http://www.robotis.us/360-laser-distance-sensor-lds-01-lidar, 2009, [Online; accessed 28 Feb. 2020].

[33] M. Mancini, G. Costante, P. Valigi, and T. A. Ciarfuglia, "Fast robust monocular depth estimation for obstacle detection with fully convolutional networks," in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 4296–4303.

[34] H. Jeon, Y. Kim, W. Myeong, and H. Myung, "One-way ViSP (Visually Servoed Paired structured light system) for structural displacement monitoring," *Smart Materials and Structures*, vol. 26, no. 8, p. 085044, 2017.

[35] R. B. Fisher, A. Ashbrook, C. Robertson, and N. Werghi, "A low-cost range finder using a visually located, structured light source," in *Proc. IEEE International Conference on 3-D Digital Imaging and Modeling*, 1999, pp. 24–33.