LIDAR and Monocular Camera Fusion: On-road Depth Completion for Autonomous Driving

Chen Fu[1], Christoph Mertz[2] and John M. Dolan[1,2]

***Abstract*— LIDAR and RGB cameras are commonly used sensors in autonomous vehicles. However, both of them have limitations: LIDAR provides accurate depth but is sparse in vertical and horizontal resolution; RGB images provide dense texture but lack depth information. In this paper, we fuse LIDAR and RGB images by a deep neural network, which completes a denser pixel-wise depth map. The proposed architecture reconstructs the pixel-wise depth map, taking advantage of both the dense color features and sparse 3D spatial features. We applied the early fusion technique and fine-tuned the ResNet model as the encoder. The designed Residual Up-Projection block recovers the spatial resolution of the feature map and captures context within the depth map. We introduced a depth feature tensor which propagates context information from encoder blocks to decoder blocks. Our proposed method is evaluated on the large-scale indoor NYUdepthV2 and KITTI odometry datasets which outperforms the state-of-the-art single RGB image and depth fusion method. The proposed method is also evaluated on a reduced-resolution KITTI dataset which synthesizes the planar LIDAR and RGB image fusion.**

I. Introduction

To provide an autonomous vehicle with a sufficient level of autonomy and safety, the perception system needs a robust object detection unit. Unlike object detection in computer vision, simply providing a bounding box in the 2D image plane or 3D real world coordinates is not enough for autonomous driving. Additional information including heading angle of the vehicle, 3D location and distance of the obstacle, as well as rough shape are all important for the decision-making and trajectory planning of an autonomous vehicle. Different sensors have different capabilities and properties. Cameras provide dense texture and semantic information about the scene, but have difficulty directly measuring shape and location of a detected object. LIDAR provides accurate distance measurement of the object using time of flight (TOF). In order to estimate the coarse shape and location of the object, the LIDAR point cloud should be segmented. However, precise point cloud segmentation is difficult due to the sparsity in horizontal and vertical resolution of the scanning points. The RADAR provides object-level speed and location relative to the ego-vehicle via range and range-rate, but does not give accurate shape of the objects.

The point cloud of a LIDAR scan is usually sparse even for high-definition LIDAR, especially for faraway objects, compared to the density of an RGB image. Object detection and segmentation merely using low-cost planar LIDAR is even more difficult. As a result, how to fuse dense RGB semantic information with sparse depth information from sparse LIDAR measurements to achieve a better perception capability is an important topic for both academia and industry [1]. The main contributions of this paper are: 1) We apply the early-fusion technique and fine-tune the famous ResNet-50 model as the feature encoder. 2) A Residual Up-Projection block (RUB) is designed to recover the spatial resolution of the depth map. 3) Context information is propagated from encoder blocks to RUB by skip connections.

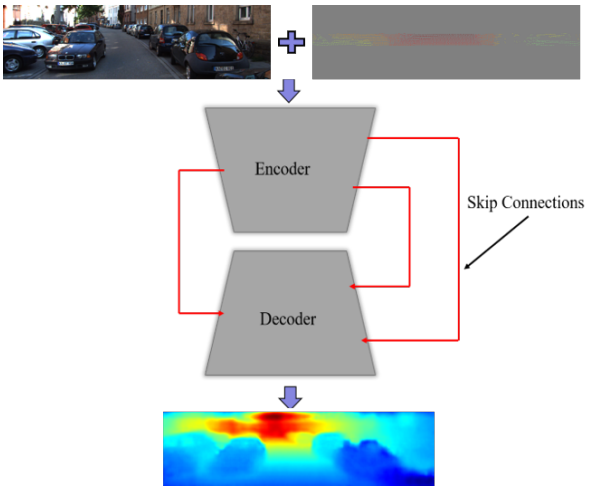


Fig. 1: We fuse the planar LIDAR and RGB camera at an early stage which takes the RGB image and sparse depth map as input. Skip connections and a decoder network improve dense depth prediction.

The proposed method is shown in Figure 1. This paper is organized as follows: Section II briefly reviews the prior work in depth sensing, using various types of sensors. Section III-A introduces the proposed Residual Up-Projection block. Sections III-C and III-D detail the proposed network architecture and Berhu Loss function. The experimental setup and dataset are respectively explained in Sections III-E. Finally, Sections IV-B and V discuss the experimental results of our method and give conclusions.

II. Related Work

Depth sensing is one of the difficult regression problems in robotics perception and autonomous driving. This problem can be quite different depending on the type of sensors mounted on the robot. It includes depth completion from sparse LIDAR scanning points and RGB-D image, and prediction from a single RGB image [1].

Depth Prediction: Depth prediction from a single RGB

[1]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA cfu1@andrew.cmu.edu
[2]The Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA {mertz, jmd}@cs.cmu.edu

image is a challenging task, as an RGB image has limited depth information. However, the problem is not unsolvable, considering the heuristic in human vision: nearby objects look bigger, distant objects smaller, and object texture is also helpful. Some early solutions modeled the relationship between visual cues by a Markov Random Field (MRF), associating depth with superpixels [2], [3]. To further improve the performance of pixel-based MRF, [4] combined semantic labels to inform depth prediction. However, this family of algorithms uses predefined models and hand-crafted features, which is difficult to apply to the variety of on-road scenes in autonomous driving. Various deep learning techniques have also been applied to this problem. In [5], a multi-scale network is proposed to estimate depth in indoor scenes. Taking advantage of high-quality features extracted from ResNet, [6] predicts the depth map of the scene. However, these methods lose details of the scene through convolution and pooling, which leads to a severe decrease in depth map resolution. To solve this problem, [7] concatenates a multi-scale feature learner and ordinal regression optimizer with a feature encoder and scene understanding module to preserve the quality of the high-resolution depth map. In this paper, we concatenate an Up-Projection block chain with the encoder, which achieves a high-resolution depth map.

Depth Completion: The depth completion problem can be broken into two categories: depth inpainting and dense completion. Distance, transparency and bright surfaces are the most common failure cases for commodity-grade RGB-D cameras. As a result, large missing areas and holes within the depth channel need to be filled. To fill holes within a Kinect depth map, [8], [9] propose algorithms that take spatial and temporal information from neighbor pixels. However, these methods only consider the local depth instead of local geometry. To address this limitation, [10] provides a local tangent plane estimation algorithm that enhances the depth image. In [11], the deep network predicts surface normal and occlusion boundaries from RGB images and solves a linear optimization problem taking depth information as regularization. Recently, a Generative Adversarial Network (GAN) has been applied to depth completion as well [12]. Even though the RGB-D image contains more depth information, it is difficult for the depth camera to sense large-scale on-road scenarios, as it can only capture the depth of objects within a few meters. As a result, a RGB-D camera with the state-of-the-art depth inpainting algorithm is not a solution to depth estimation for autonomous driving.

In the dense completion problem, a low-quality depth map is completed or super-resolved into a pixel-wise depth map. In [13], a novel data term formulation is applied to the MRF, which qualitatively and quantitatively improves the high-resolution depth maps. The relationship between image segmentation boundaries and depth boundaries is fully used in [14] to predict depth from aggressively sub-sampled images and videos. To recover the dense depth information from a cropped depth map, [15] proposed a Sparsity Invariant Convolution network which is invariant to sparsity level. However, a binary observation mask which served as prior knowledge needs to be passed to the sparse convolution layer.

The perception system of autonomous vehicles usually contains multiple types of sensors. Sensor fusion provides a robust object detection and scene understanding result. In [1], a self-supervised training framework has been applied to RGB images and sparse depth images to generate a dense point cloud. However, the performance of this method highly relies on sequential information of the RGB image and accurate pose translation of the ego-vehicle. Fusing LIDAR with RGB camera through CNN, [16] accomplished depth completion or semantic segmentation with or even without a dense RGB image. This method can deal with potential sensor failure in real autonomous driving cases. As a result, the fusion of multiple sensors can improve the robustness of the perception system.

In this work, we propose an early fusion technique which outperforms single image-based methods as the sparse 3D spatial information from LIDAR enhanced the texture information. No further prior information is needed, such as binary masks in [15]. In addition, the Residual Up-Projection blocks and proposed skip connections reconstruct a more detailed depth map, compared to previous methods.

III. Methodology

A. Residual Up-Projection Block

For the sake of clarity, we introduce the Residual Up-Projection blocks in the decoder network. The most common Up-Pooling layer increases the spatial resolution by considering the features from nearby patches. Bilinear interpolation or a nearest neighbor interpolation mechanism can be applied, which is widely used in fully convolutional neural networks for the semantic segmentation task [17]. However, this Up-Sampling layer is not sufficient for depth completion, as it does not consider the geometry and semantic information. As a result, it increases the prediction error on boundaries of objects. The de-convolution layer defined in [18] up-samples the feature map by Up-Pooling kernels which recover details in the image. Our Residual Up-Projection block (RUB) further optimizes the up-sampling process by introducing a residual into the block [19]. The detailed structure of the RUB is shown in Fig. 3. The $5*5$ convolution layers recover the local details of the feature map. The projection shortcut helps the training step by escaping the $3*3$ convolution layer.

B. Skip connections

Compared with the traditional network architecture proposed in [20], three skip connections are added to pass feature tensors from encoding residual blocks to decoding blocks. By concatenating feature tensors from the encoding blocks and previous decoding blocks, the RUB receives a larger number of feature channels. This can improve the depth prediction, as more context information is propagated to higher-resolution layers. In most deep networks, it is difficult to recover detailed textures and context information. The skip connections also forward missing detailed features such as object boundaries to higher-resolution decoding blocks,

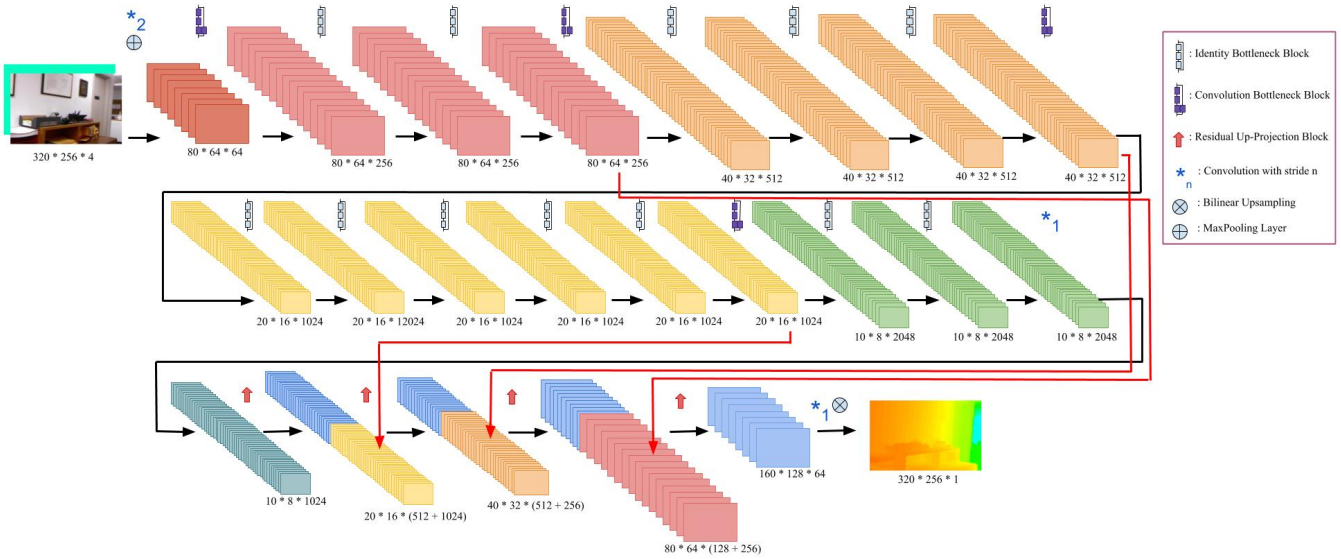


Fig. 2: The architecture of the proposed network is shown above. The encoder network takes a pretrained Resnet model with RGB image and sparse depth map as input. The decoder network concatenates four Residual Up-Projection blocks (RUB). Three skip connections pass the feature map from the residual encoding block to the RUB.

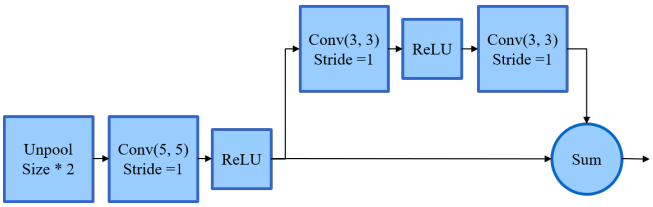


Fig. 3: The proposed Residual Up-Projection block.

which mirror the input. The proposed skip connections are shown as the red arrows in Fig. 2.

C. Network Architecture

In total, there are 4 channels in the input layer of the network after early data fusion, shown as the input in Fig. 2. Instead of predicting the pixel-wise depth from a single RGB image, the sparse 3D features from LIDAR provide a heuristic for depth regression. In reverse, the texture features from the RGB image encode the semantic information, which completes the sparse LIDAR point cloud. In all, the network learns the local geometry of each pixel considering the prior from the RGB image and LIDAR scanning points. Unlike the RGB image, the LIDAR projection image does not have obvious texture, and it is difficult to find patterns in the projected image. As a result, we cannot easily find the associations between LIDAR point clusters in the projected image and features in the RGB image. This problem can be solved by using the proposed early fusion network. We applied Resnet as the encoding network, which means feature tensors are propagated through successive residual blocks. We chained four RUB with skip connections as the decoder to achieve a high-resolution dense depth map.

D. Berhu Loss Function

The most common and popular choice of loss function for regression problems is mean squared error (MSE). However, MSE is not adequate for the depth completion task, as it tends to penalize more heavily for larger errors. It learns to smooth and blur edges on object boundaries, which is even worse in outdoor scenarios in autonomous driving cases [1]. To avoid these problems, we applied the Berhu loss as the loss function for training. The Berhu Loss is defined as follows:

$$B(e) = \begin{cases} |e|, & \text{if } |e| \leq c \\ \frac{e^2+c^2}{2c}, & otherwise \end{cases} \quad (1)$$

The term c is a batch-dependent parameter, which considers the maximum absolute error over all the pixels in the predicted depth map. In this paper, we take c as 20% of the maximum absolute error in a batch. If the element-wise absolute value of the prediction error is smaller than c, Berhu loss behaves as a mean absolute error. Otherwise, it acts approximately as mean square error.

E. Network Training and data Augmentation

We take the Resnet pretrained on the ImageNet dataset as the encoder and fine-tune the Residual blocks with the RGB-D input introduced above. Due to the limitation of our computation resources, we use a smaller batch size of 16 and train the network for 20 epochs. We choose to use the SGD optimizer with a decreasing learning rate, starting from 0.01. We conducted an online data augmentation process, which randomly transforms the original images.

IV. EXPERIMENTAL RESULTS

In this section, we detail our experimental setup and explain the training techniques of the proposed depth completion architecture. To verify the performance of the proposed network, we compare our model with previous depth completion methods through quantitative results on the NYUDepth V2 dataset and KITTI odometry dataset. To evaluate the performance of the proposed network on planar LIDAR

TABLE I: Performance comparison of proposed network with previous single RGB-based and fusion methods on NYUdepthV2 dataset

Input	#Depth Sample	Methods	RMSE (m)	REL	δ_1 (%)	δ_2 (%)	δ_3 (%)
RGB	0	Eigen et al.[21]	0.641	0.158	76.9	95.0	98.8
	0	Laina et al.[22]	0.573	0.127	81.1	95.3	98.8
RGB+D	225	Liao et al.[20]	0.442	0.104	87.8	96.4	98.9
	200	Ma et al.[23]	0.230	0.044	97.1	99.4	99.8
	200	Proposed	**0.203**	**0.040**	**97.6**	**99.5**	**99.9**

and camera fusion, we also test our method on a reduced-resolution KITTI dataset.

A. Experimental Setup and Evaluation Metric

NYUDepthV2 is one of the largest RGB and depth datasets for indoor scene understanding [25]. In total, the training and testing scenes contain 47584 and 654 images, respectively. In order to compare the proposed method with previous algorithms fairly, we down-sampled the original image to half-resolution and center-cropped the image to $320 * 256$ pixels. In the training process, a sparse input depth map is sampled randomly from the ground truth depth image. Instead of using a fixed sparse depth input for every training sample, we randomly generate sparse depth in each training epoch. This can augment the training data and achieve a more robust network. We not only evaluated the proposed network on an indoor depth dataset, we also tested the performance on the on-road autonomous driving KITTI odometry dataset. We picked 46416 data samples with ground truth for our proposed model training, and 3200 images for evaluation. The depth map is constructed by projecting the LIDAR point cloud onto the image plane. Unlike the NYUDepthV2 dataset, in which all image pixels have a depth value, only some of the image pixels contain depth values. As a result, we cropped the bottom part of the image with a size of $928 * 256$, where projected LIDAR points exist.

To compare the proposed network with state-of-the-art methods, we use the following evaluation metrics provided by the benchmark dataset. We compare the root mean square error (RMSE), which directly measures the average error over all pixels. To get rid of the scaling problem we also compare the Mean Absolution Relative Error (REL). In order to count the percentage of pixels within a certain threshold, we also consider the δ_j metric.

B. Results and Discussion on Benchmark Dataset

In order to compare the proposed method with state-of-the-art depth completion and depth prediction methods on the NYUDepthV2 dataset and KITTI odometry dataset, we take the reported accuracy of previous methods from their original papers. In comparing with the method which takes a single RGB image, fusing sparse depth information significantly improves the overall prediction accuracy. In method [20], the architecture simply uses a chain of de-convolution layers to recover the high-resolution depth map. Our method improves the depth prediction by chaining up four RUBs, which captures more detailed textures in the depth map. By applying our proposed architecture with skip connections, we achieve a better performance compared with the state-of-the-art method [23] by 11.3% on the NYUDepth V2 dataset. For the KITTI odometry dataset, we achieve a 40.4% improvement compared with the state-of-the-art single RGB image-based network and a 13% in REL with the state-of-the-art single RGB image and LIDAR fusion method. Detailed comparison is shown in Table I and Table II

C. Test on Reduced Resolution KITTI Odometry Dataset

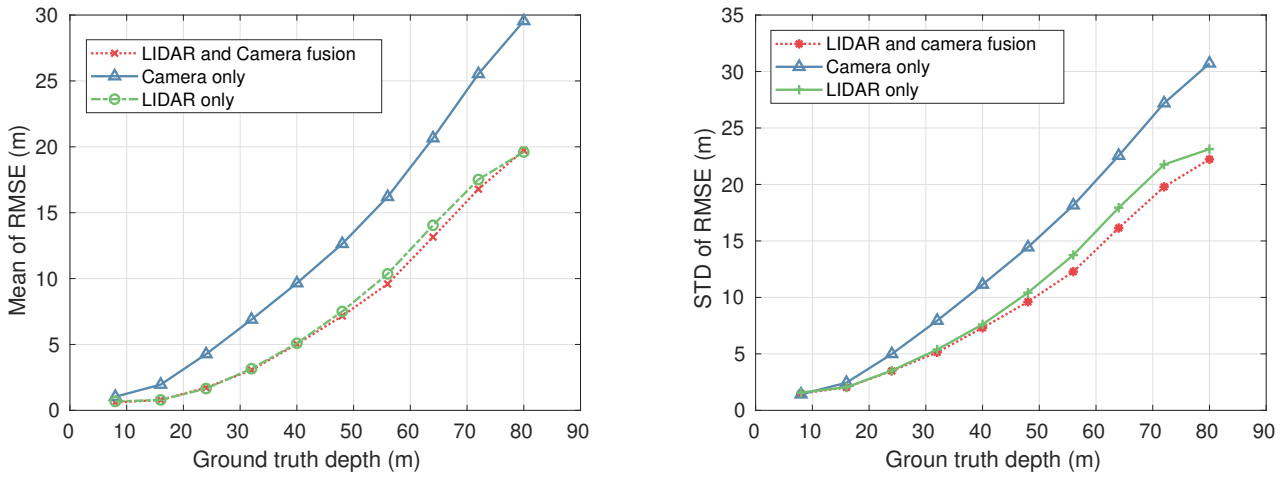
In the KITTI odometry dataset, a high-definite Velodyne HDL-64E is mounted on top of the testing vehicle. It is a a 64 channel and 360° FOV LIDAR sensor with adjustable data update rate. This powerful Velodyne sensor has a horizontal angular resolution of around 0.09° and 26.8° of vertical FOV with approximate 0.4° angular resolution. However, the planar LIDAR sensors are installed at the bumper height of the vehicles, different from the Velodyne LIDAR mounted on top of the vehicle [26]. In order to reduce the resolution of the dense point cloud of KITTI odometry data, we select a band of LIDAR points from the dense point cloud. In this way, the reduced-resolution Velodyne data have similar point cloud features to the planar LIDAR.

In order to test the performance of the proposed architecture we apply different input to the network. As shown in Table III, the early fusion architecture performs better than using a single RGB image and single planar LIDAR depth map as input, which improves the RMSE respectively by around 30% and 10%. As a result, the early fusion technique achieves better pixel-wise depth prediction, as the LIDAR provides depth guides and the RGB images provide semantic information.

We also conducted a statistical study on the RMSE of proposed method on different types of input. In general, the mean and standard deviation of the depth prediction RMSE increase monotonically with the ground truth depth. Based on our analysis, the reason can be summarized as follows. First of all, we have fewer LIDAR points in the faraway regions, compared with nearby regions. Secondly, the textures and features of RGB images at far distances, especially at the vanishing points which are usually tiny and the depth value of these pixels varies sharply. As shown in Fig. 4a and Fig. 4b, the LIDAR and camera fusion technique has better mean RMSE, compared with taking single RGB as input. We have a mean RMSE error less than 5 meters when the depth ground truth is within 40 meters. In comparing with single LIDAR input, we achieve a slightly better mean and STD

TABLE II: Performance comparison of proposed network with previous single RGB-based and fusion methods on KITTI dataset

Input	#LIDAR sample	Methods	RMSE (m)	REL	δ_1 (%)	δ_2 (%)	δ_3 (%)
RGB	0	Mancini et al.[24]	7.51	-	31.8	61.7	81.3
	0	Eigen et al.[21]	6.16	0.190	69.2	89.9	96.7
RGB+D	200	Liao et al.[20]	4.50	0.113	87.4	96.0	98.4
	200	Ma et al. [23]	3.85	0.083	91.9	97.0	**98.9**
	200	Proposed	**3.67**	**0.072**	**92.3**	**97.3**	**98.9**



(a) Mean of the RMSE with different types of inputs.

(b) Standard Deviation of the RMSE with different types of inputs.

Fig. 4: Performance comparison of the proposed method with different types of inputs. We analyze the mean and standard deviation of the RMSE at different ground truth depths.

TABLE III: Performance comparison of proposed network on reduced-resolution KITTI odometry dataset with different types of inputs

Input	RMSE (m)	REL	δ_1 (%)	δ_2 (%)	δ_3 (%)
RGB	5.92	0.193	67.3	91.6	97.6
Depth	4.61	0.095	88.8	95.5	98.2
RGB+D	**4.16**	**0.092**	**89.4**	**96.3**	**98.6**

using LIDAR and camera fusion. Even though it seems that we do not achieve a significant improvement from above figures, it is resulted in the sparsity of ground truth depth map. Actually, the improvement of proposed fusion method is shown in the qualitative results discussed below.

To visually compare the predicted depth map of the proposed method with different types of input, we provide some qualitative result on various road scenes in Fig. 5. Our network generates a more detailed depth map with LIDAR and RGB image fusion, especially for scene 1 shown in Fig. 5a. Comparing with only RGB image and planar LIDAR, the fusion result generates more detailed features on the vehicle boundaries with a clear depth trend. In Fig. 5c, we can notice a more detailed contour of the parking vehicles on the right side of the road, compared to blurred boundaries taking single RGB image or planar LIDAR as input. Even though in Fig. 4a and Fig. 4b the LIDAR-only method has similar mean and STD of RMSE, the visualization result of the LIDAR-only method has a 'stripe-like' feature in its predicted depth maps. This phenomenon is caused by the features of the projected LIDAR depth image where LIDAR points are arranged as lines or arcs in the image plane. With the texture information of RGB image, the fusion method generates a more smooth depth map and more reasonable depth textures of the scenes.

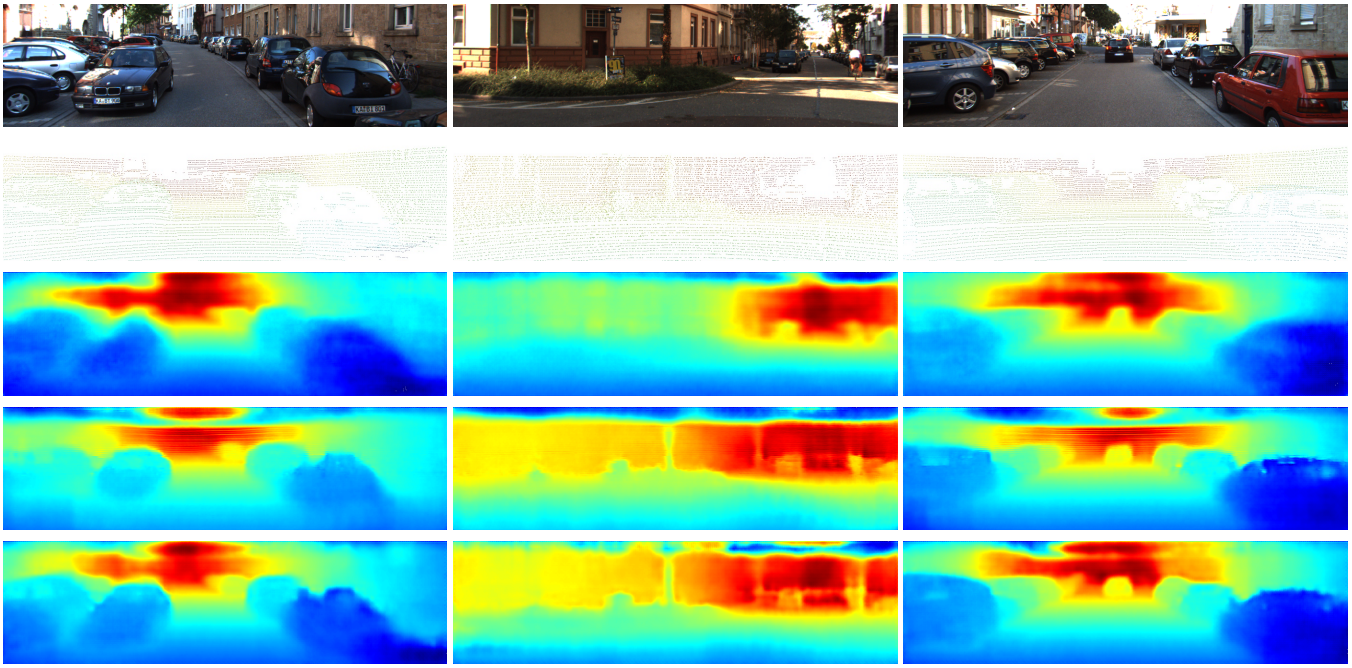
V. Conclusion

In this paper, we propose a deep fusion architecture which fuses LIDAR with RGB images to complete the depth map of the surrounding environment. A Residual Up-Projection block is applied to recover the dense depth map. Skip connections pass the feature map from encoder blocks to decoder blocks, which helps the decoder network capture more context information from feature tensors. Our method outperforms conventional methods on the NYUDepthV2 dataset and KITTI odometry dataset. We also applied the proposed method to the reduced-resolution KITTI odometry dataset to estimate the pixel-wise depth map. Further work will test the proposed method on real planar LIDAR and dense point cloud datasets.

References

[1] F. Ma, G. V. Cavalheiro, and S. Karaman, "Self-supervised sparse-to-dense: Self- supervised depth completion from lidar and monocular camera," *arXiv preprint arXiv:1807.00275*, 2018.

[2] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning Depth from Single Monocular Images," *Advances in Neural Information Processing Systems*, vol. 18, pp. 1161–1168, 2006.



(a) Scene 1 (b) Scene 2 (c) Scene 3

Fig. 5: This group of images shows the qualitative result of our network with and without fusion. The first row of images is the RGB image of the scene. The second row of images is the ground truth depth map. The third and fourth row of images are the prediction depth map of the proposed network using RGB image and synthesized planar LIDAR as input. The fifth row of images is the prediction depth map of the proposed network with planar LIDAR and RGB image fusion. The dark red reflects farther distance and the dark blue reflects closer distance. The range of distance lies between $0m$ and $80m$.

[3] A. Saxena, M. Sun, and A. Y. Ng, "Make3D: learning 3D scene structure from a single still image." *PAMI*, vol. 31, no. 5, pp. 824–40, 2009.

[4] B. Liu, S. Gould, and D. Koller, "Single image depth estimation from predicted semantic labels," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1253–1260.

[5] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Advances in Neural Information Processing Systems 27*, 2014, pp. 2366–2374.

[6] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016*, 2016, pp. 239–248.

[7] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," *CoRR*, vol. abs/1806.02446, 2018.

[8] L. S. Massimo Camplani, "Efficient spatio-temporal hole filling strategy for kinect depth maps," pp. 8290 – 8290 – 10, 2012.

[9] D. Zhang, Y. Yao, D. Zang, and Y. Chen, "A spatio-temporal inpainting method for Kinect depth video," in *IEEE ICSIPA 2013 - IEEE International Conference on Signal and Image Processing Applications*, 2013, pp. 67–70.

[10] K. Matsuo and Y. Aoki, "Depth image enhancement using local tangent plane approximations," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3574–3583.

[11] Y. Zhang and T. Funkhouser, "Deep depth completion of a single rgb-d image," *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[12] I. Goodfellow, J. Pouget-Abadie, and M. Mirza, "Generative Adversarial Networks," *arXiv preprint arXiv: ...*, pp. 1–9, 2014.

[13] J. Lu, D. Min, R. S. Pahwa, and M. N. Do, "A revisit to MRF-based depth map super-resolution and enhancement," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2011, pp. 985–988.

[14] J. Lu and D. Forsyth, "Sparse depth super resolution," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June-2015, 2015, pp. 2245–2253.

[15] J. Uhrig, N. Schneider, L. Schneider, U. Franke, T. Brox, and A. Geiger, "Sparsity invariant cnns," in *International Conference on 3D Vision (3DV)*, 2017.

[16] J. Maximilian, d. C. Raoul, W. Emilie, P. Xavier, and N. Fawzi, " Sparse and Dense Data with CNNs: Depth Completion and Semantic Segmentation," 2018.

[17] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June, no. i, pp. 1–9, 2015.

[18] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2011, pp. 2018–2025.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *Arxiv.Org*, vol. 7, no. 3, pp. 171–180, 2015.

[20] 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," 05 2017, pp. 5059–5066.

[21] E. David and F. Rob, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," 2014.

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

[23] F. Ma and S. Karaman, "Sparse-to-dense: Depth prediction from sparse depth samples and a single image," 2018.

[24] M. Mancini, G. Costante, P. Valigi, and T. Ciarfuglia, "Fast robust monocular depth estimation for obstacle detection with fully convolutional networks," 07 2016.

[25] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[26] C. Fu, P. Hu, C. Dong, C. Mertz, and J. Dolan, "Camera-based semantic enhanced vehicle segmentation for planar lidar," 11 2018, pp. 3805–3810.