2021 IEEE International Conference on Robotics and Automation (ICRA 2021)
May 31 - June 4, 2021, Xi'an, China

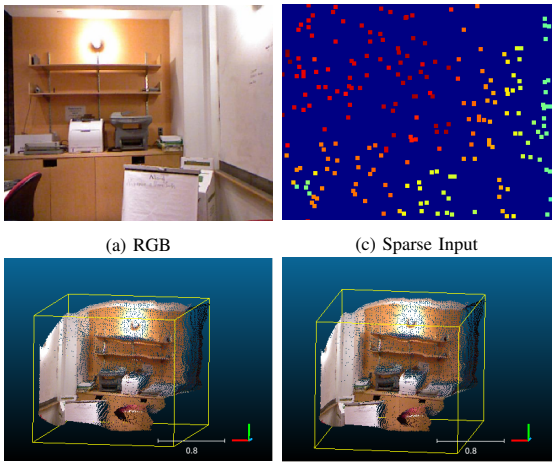
Linear Inverse Problem for Depth Completion with RGB Image and Sparse LIDAR Fusion

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

***Abstract*— Comprehensive depth information from surrounding scenes is important for perception in autonomous driving and robots. Sparse LIDAR sensors give a low-density point cloud of the environment, but are more affordable than their high-density counterparts. In this paper, we propose a novel sensor fusion architecture for sparse LIDAR depth completion. Instead of the traditional end-to-end neural network-based algorithm, we formulate depth completion as a Linear Inverse Problem (LIP) with a multi-modal proximal operator. This sensor fusion architecture allows a better signal prior and finds the unique optimal solution to the LIP. Instead of learning a unified network for the sparse input which treats pixels evenly, the proposed architecture guarantees both the data consistency and smoothness of the predicted depth map. To demonstrate the performance of our algorithm, we benchmark on the simulation dataset TartanAir, and the real indoor NYUdepthv2 and real outdoor KITTI datasets. Our proposed method outperforms previous methods and uses fewer parameters in both indoor and outdoor datasets.**

I. Introduction

A reliable perception system is crucial for the safety of autonomous vehicles, as it influences the behavior estimation of surrounding vehicles and the decision making of the ego-vehicle. Currently, multiple sensors of various types are mounted on state-of-the-art autonomous driving cars. A camera produces an RGB image with comprehensive semantic information of the environment, but it is quite difficult to directly estimate the shape and state of an object from a single RGB image. A high-definition LIDAR gives depth information about the surrounding environment, which roughly measures the shape and location of the objects. However, the point cloud is quite sparse even for high-definition LIDAR, with only 10% of pixel density compared to an image [1]. It is even worse for a sparse LIDAR, though it is more affordable than the multi-layer high-definition LIDAR. Recently, state-of-the-art methods have begun to explore extending sensor fusion architectures to solve the depth completion problem, which augments the sparse point cloud under the guidance of depth semantic features [2]–[4]. These approaches return detailed dense RGB+D information and further benefit other perception tasks including 3D reconstruction, object detection and HD mapping [2], [5]. However, this family of methods usually treats the pixels evenly, which interpolates the missing depth but blurs the given depth values through multiple convolution layers. In this paper, we formulate the depth completion problem as a Linear Inverse Problem (LIP), which guarantees the data consistency of the given depth values. To solve the underdetermined problem, we propose a sensor fusion-based proximal operator which combines the local semantic information to find the unique feasible solution to the completion task. To further improve the performance of the proximal operator, adversarial training techniques are applied to improve the smoothness of the output. As an iterative optimization approach, the proposed method provides a lightweight solution to the on-board computation limitation with fewer network parameters. The proposed method not only achieves a promising result on the indoor and outdoor benchmarks, but also shows the ability to retain the data consistency and details of the input depth map. A sample result of the proposed method is shown in Fig. 1.



(a) RGB (c) Sparse Input

(b) Result (d) Ground Truth

Fig. 1: Visualization result of the proposed method on NYUdepthv2 dataset. We project the depth map back into the world coordinates. We achieve a smooth dense point cloud of the scene.

The remainder of this paper is organized as follows: we briefly summarize previous work on depth completion and the Linear Inverse Problem in Sec. II. Sec. III details the formulation of the depth completion and Linear Inverse problems and proposes an adversarial semantic-enhanced proximal operator. Sec. IV introduces our experiment settings, and details ablation studies and benchmark results. Sec. V summarizes the proposed method and suggests further applications.

Chen Fu is with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA {cfu1}@andrew.cmu.edu

Christoph Mertz is with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA {cmertz}@andrew.cmu.edu

John M. Dolan is with the Department of Electrical and Computer Engineering and Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA {jdolan}@andrew.cmu.edu

II. Related Work

Estimating the depth of a surrounding scene is a regression problem. It can be further divided into depth estimation, depth in-painting and depth completion according to the input modulation of the algorithm. In particular, depth completion aims to complete and interpolate the missing depth pixels within the depth map, given a sparse LIDAR point cloud, stereo images or RADAR, with the guidance of an RGB image.

Depth Completion: In order to achieve a high-resolution pixel-wise depth image from a sparse depth map, [2] proposed network architectures with encoder-decoder networks. These augmentation algorithms provide simple end-to-end early fusion solutions to the depth completion problem. However, these approaches can only predict a rough depth map with fewer features and depth contours. In order to further improve the encoder-decoder architecture, [6] applies the U-net architecture, which passes the feature maps from the encoder to the decoder directly, improving the quality of the depth map. In contrast, [3] refines the depth completion accuracy from the standpoint of the training pipeline, as the ground truth is sparse for the autonomous driving scenario. The proposed self-supervised training method improves the end-to-end network [6] by introducing a photometric loss, taking advantage of the sequential information of the RGB image. However, these methods cannot focus or concentrate on the details in local features of the depth image. In order to solve this problem, [7] proposes a novel spatial refinement convolution network which completes the depth value depending on neighbor pixels. To recover features from both local pixels and global context, [8] proposes a middle-level fusion architecture which allows the local and global context features to guide the depth completion task. As an improvement on previous work, [9] uses 2.5D surface normal information to aid the depth completion of the sparse input. However, these methods assume that the scene is constituted by planes, which is not a reasonable assumption in reality, especially for complex autonomous driving scenarios. To better understand the underlying probability model of the depth completion problem, [10] gives a Gaussian process formulation of the problem and solves it by an inductive fusion architecture. Even though this method gives a smoother depth estimation, it does not guarantee the data consistency with the input depth values.

Linear Inverse Problem (LIP): How to accurately and efficiently solve a LIP is an important research topic, as many image-based problems can be formulated as LIP, such as image denoising and deblurring [11]. Traditional signal priors are applied to regularize the LIP, which ensures an unique solution according to the specific context features from the RGB image [12]. However, these types of priors are usually hand-crafted, and cannot be generalized for complex problems such as image super-pixels or image in-painting. Such a technique does not fit the high-performance requirement of the depth completion problem. Recently, deep learning was applied to directly learn a mapping between the measurements and the true images. This family of methods shows a promising result in the area of image in-painting and image super-resolution, which produces a more natural and realistic image [13], [14]. Unlike the previous learning-based methods which only rely on one modulation, we propose a sensor fusion-based architecture to deal with the under-determined problem in LIP. Instead of merely using depth image as a regularization, context features are also considered in achieving a possible depth map. As a result, our method achieves not only data consistency for the known depth values, but also a smooth and reasonable depth interpolation.

III. Methodology

A. Problem Formulation

First, we define the noise measurement of the depth from sparse LIDAR as $\mathbf{y} \in \mathbf{R}^{H\times W}$ and the ground truth dense depth map as $\mathbf{x} \in \mathbf{R}^{H\times W}$, where H and W are the height and width of the image plane. The corresponding linear transform of the system is defined as A. In the depth completion task, A refers to a linear sampling mask with the same dimension as the input depth map. In the depth completion task, we are solving the under-determined linear problem so that $\mathbf{y} \approx A\mathbf{x}$. Instead of merely using the sparse depth as a constraint on the completion, the semantic information from the RGB image is also included in our formulation as R. In this case, the depth completion problem is formulated as the Linear Inverse Problem:

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \lambda\phi(\mathbf{x}, R) \tag{1}$$

where $\phi(\mathbf{x}, R)$ is the sensor fusion-based signal prior, learned from a large-scale depth completion dataset, and λ is the weighting parameter. One typical solution to solve this linear problem is applying the Alternating Direction Method of Multipliers (ADMM). In order to ease the problem, we rewrite (1) as follows:

$$\begin{aligned} \min_{\mathbf{x},\mathbf{z}} \quad & \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \lambda\phi(\mathbf{z}, R) \\ \text{s.t.} \quad & \mathbf{x} = \mathbf{z} \end{aligned} \tag{2}$$

As shown in (2), the original problem is simplified by splitting the optimization variable into two parts $\mathbf{x}$ and $\mathbf{z}$, which are restricted by the equivalent constraint. As a result, the Augmented Lagrangian defined by ADMM can be applied to allow decomposition:

$$L_\rho(\mathbf{x}, \mathbf{z}, \mathbf{y}) = \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \lambda\phi(\mathbf{z}, R) + \mathbf{y}^T(\mathbf{x} - \mathbf{z}) + \frac{\rho}{2}\|\mathbf{x} - \mathbf{z}\|_2^2 \tag{3}$$

where $\mathbf{y}$ is the dual variable of (2) and ρ is the penalty term of the equivalent constraint. By replacing $\mathbf{u} = \mathbf{y}/\rho$, we achieve a scaled-form Augmented Lagrangian as:

$$L_\rho(\mathbf{x}, \mathbf{z}, \mathbf{u}) = \frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \lambda\phi(\mathbf{z}, R) + \frac{\rho}{2}\|\mathbf{x} - \mathbf{z} + \mathbf{u}\|_2^2 - \frac{\rho}{2}\|\mathbf{u}\|_2^2 \tag{4}$$

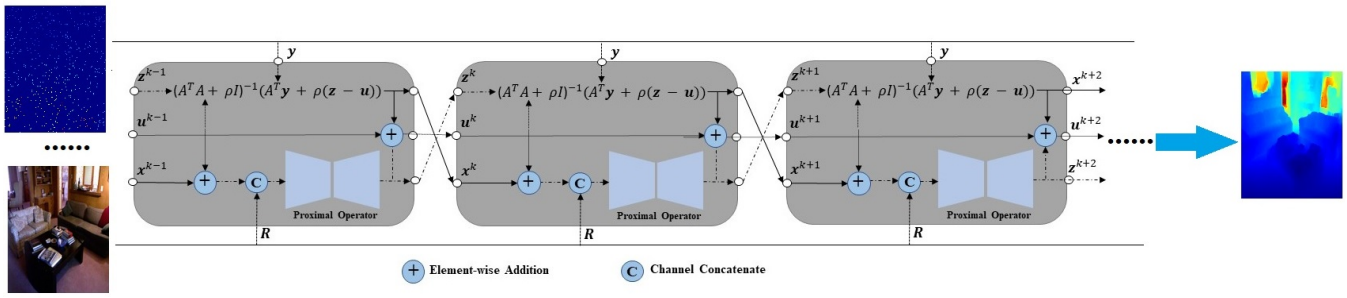


Fig. 2: In this figure, we show the detailed architecture of the proposed method. The RGB image and a sparse depth map are the input to the algorithm with a complete depth map as output. For each iteration, the proximal operator trained by adversarial learning takes the updates from the previous iteration and smooths out the depth map with the context feature from the RGB image.

Instead of jointly optimizing over $(\mathbf{x}, \mathbf{z})$, the original minimization problem is split into two parts and optimized separately through iteration:

$$\begin{aligned} \mathbf{x}^{k+1} &= \underset{\mathbf{x}}{\operatorname{argmin}} L_\rho(\mathbf{x}, \mathbf{z}^k, \mathbf{u}^k) \\ \mathbf{z}^{k+1} &= \underset{\mathbf{z}}{\operatorname{argmin}} L_\rho(\mathbf{x}^{k+1}, \mathbf{z}, \mathbf{u}^k) \\ \mathbf{u}^{k+1} &= \mathbf{u}^k + (\mathbf{x}^{k+1} - \mathbf{z}^{k+1}) \end{aligned} \tag{5}$$

We can further derive the $\mathbf{x}$-update and $\mathbf{z}$-update equations for the three-step iteration as follows:

$$\mathbf{x}^{k+1} = \underset{\mathbf{x}}{\operatorname{argmin}}(\frac{1}{2}\|\mathbf{y} - A\mathbf{x}\|_2^2 + \frac{\rho}{2}\|\mathbf{x} - \mathbf{z}^k + \mathbf{u}^k\|_2^2) \tag{6}$$

$$\mathbf{z}^{k+1} = \underset{\mathbf{z}}{\operatorname{argmin}}(\frac{\rho}{2}\|\mathbf{x}^{k+1} - \mathbf{z} + \mathbf{u}^k\|_2^2 + \lambda\phi(\mathbf{z}, R)) \tag{7}$$

As (6) is in convex form with respect to $\mathbf{x}$, a closed-form solution can be obtained by taking the gradient of the equation and setting the result to zero. In this step, the proposed method is restricting the prediction value on the pixels with depth values from the input to retain consistency. To update (7), we define the proximal operator as:

$$\mathbf{prox}_{\phi, \frac{\lambda}{\rho}}(\mathbf{x}^{k+1} + \mathbf{u}^k,\ R) \tag{8}$$

Taking advantage of the fact that ADMM separates the signal prior by introducing the additional variable $\mathbf{z}$, a standalone proximal operator can be learned through a large-scale dataset, which is independent of the linear operator A. The common proximal operator of the z-update in (5) is soft-thresholding if we consider the L-1 norm as the signal prior. But in the proposed method, we replace the signal prior with a derivable neural network. This enables the proximal operator to combine multiple sensor modulations and smoothly interpolate the depth value.

B. Sensor Fusion based Proximal Operator

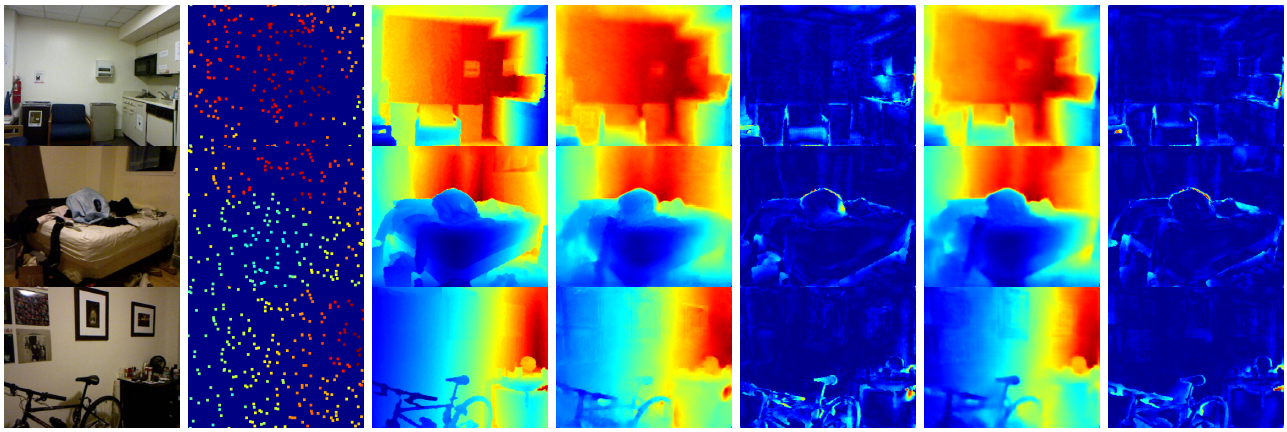
In our approach, the signal prior in the original problem (1) is reformulated as the proximal operator (8) in the ADMM. As shown in (5), $\mathbf{u}$ accumulates the residual between $\mathbf{x}$ and $\mathbf{z}$, which is forced to be equivalent according to the condition in (2). As a result, we can apply a neural network-based proximal operator which updates $\mathbf{z}$ through mapping a noisy depth input to a real depth image, guided by the semantic priors from the RGB image:

$$\mathbf{z}^{k+1} \leftarrow G(\mathbf{x}^{k+1} + \mathbf{u}^k,\ R) \tag{9}$$

As a lightweight solution, we implement a tiny encoder-decoder network architecture for each iteration of the ADMM updates. Different from the traditional proximal operators which only consider the RGB image or single data modulation as prior knowledge, our implementation introduces an early-fusion architecture. This is achieved by concatenating the noisy or sparse depth input and corresponding RGB image. The encoder of the network consists of three "bottleneck" blocks as the Resnet model [15] and three up-Projection blocks detailed in [6]. According to previous experimental experience, skip connections usually pass the gradients better and improve training efficiency of the neural network, so we add two skip connections between the peer-to-peer bottleneck block and up-Projection block. In order to further improve the quality of the neural network-based proximal operator G, we applied adversarial training techniques. Compared to traditional supervised training, adversarial training results in a more natural and accurate depth map, given a noisy depth input. In this case, a classifier or discriminator D is pretrained to distinguish a ground truth depth map from the dataset or a fake depth map generated by the proximal operator, applying the cross-entropy loss. The proximal operator G is also pretrained using the L-2 norm as the loss function. During the adversarial process, the generator, which is also the proximal operator, tries to trick the discriminator D by generating a depth map similar to the ground truth, while the discriminator tries its best to classify the true depth map and generated depth map. We jointly train both the discriminator D and the proximal operator G to ensure they compete with each other. In both outdoor and indoor scenarios, the ground truth depth map contains vacant pixels due to sensor failure or sparsity of the ground truth. To avoid the discriminator's learning greedy features from these data, we masked out the invalid regions in the noisy depth map according to the ground truth depth map before feeding into the discriminator.

IV. EXPERIMENTAL RESULTS

In this section, we conduct a detailed study of the performance of the proposed algorithm. We test the proposed method on the depth completion benchmark, indoor NYUdepthv2 dataset [16] and on-road KITTI depth completion dataset [1]. We also test the performance of the proposed method on the simulation dataset TartanAir [17], which has



(a) RGB image (b) Input (c) Ground truth (d) CSPN [7] (e) Error: CSPN (f) LIP+AD (Our) (g) Error: Our

Fig. 3: Visualization result on NYUdepthV2 dataset using official validation sets. The images from left to right are: RGB images, input sparse depth maps (visually enhanced by bilinear interpolation), ground truth, prediction results for CSPN [7], error maps for CSPN, prediction results for the proposed method (LIP+AD) and error maps for LIP + AD. We are using 200 depth points in this experiment setting. In the error map, the value of the error increases from dark blue to dark red.

exact ground truth. Ablation studies as well as the benchmark comparisons are also provided.

A. Experiment Setup and Evaluation Metric

NYUdepthv2 Dataset: This dataset contains 47584 images in the training set and 654 images in the testing dataset, according to the official dataset splitting. In order to compare the proposed method with previous work, we center-crop and reduce the resolution of the original RGB image and depth map to 256 * 320 [2], [6]. To compare with the benchmark result, we also apply a random sub-sampling strategy to simulate the sparse depth map input.

KITTI Dataset: The KITTI depth completion dataset contains 86,898 RGB image-LIDAR pairs for training, 1k for validation and 1k in the test set. In order to achieve a denser ground truth of the depth completion problem, [1] temporally concatenates LIDAR point clouds and applies stereo to remove the noise. We get rid of the pixels at the top of the image, since there are no LIDAR projections in this area. In this case, we achieve an image size of 1216 * 256 for the input, ground truth and RGB image, which enables a fair comparison to previous work. This dataset is more focused on completing the depth map from a relevantly dense point cloud captured by a high-definition Velodyne 64-layer LIDAR. To simulate the point cloud from sparse LIDAR, we also sub-sampled the KITTI Odometry dataset the same as [2] and reported the depth completion performance using different sub-sample rates.

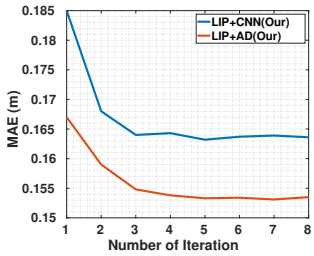
TartanAir Dataset: TartanAir is a large-scale dataset for robot navigation in different simulated environments. It provides multi-modal sensor data with accurate ground truth [17]. It has 30 photo-realistic simulation environments, covering both indoor and outdoor scenes under various weather conditions. In order to achieve the sparse LIDAR input, we down-sample the original depth image to be the same size as in the NYUdepthv2 dataset. We divide the whole dataset and use 90% of the frames for training and 5% each for validation and testing.

Evaluation Metric: Similar to previous state-of-the-art methods, we apply the benchmark evaluation metrics to analyze the performance of the proposed method on different datasets. We provide the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Relative Error (REL), which gets rid of scale. We also apply the δ_j metric to count the percentage of pixels with error within the thresholds 1.25, 1.25^2 and 1.25^3. The root mean squared error of the inverse depth (iRMSE) and mean absolute error of the inverse depth (iMAE) are also provided in KITTI datset.

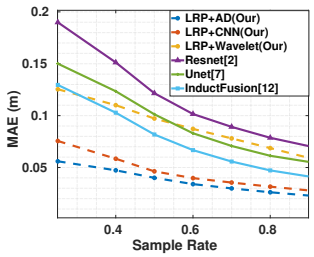
B. Ablation Study

Convergence study: In order to prove the convergence capability of the proposed method, we conduct an analysis of the MAE against different numbers of iterations of the ADMM update. We show the MAE vs. iterations for both LIP+CNN and LIP+AD in Fig. 4a. Our algorithm converges fast, with a small number of iterations, which is boosted by the learning-based proximal operator. For the first iteration, the proximal operator makes a rough estimate of the depth value on vacant pixels considering the observations and RGB features. For later iterations, the proximal operator smooths out the depth map with the guidance of the RGB image and noisy depth estimation from previous iterations. As shown in Fig. 4a, applying the adversarial training further reduces the MAE of the proposed approach by learning a more realistic signal prior. To balance performance and efficiency, we can run fewer iterations with a small number of parameters, instead of using the traditional method of hand-crafting the signal prior.

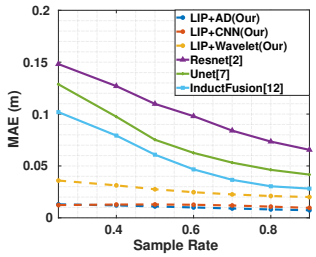
Data Consistency and Depth Completion Accuracy: In this section, we mainly focus on the comparison of the proposed method with previous methods in terms of data consistency and accuracy of depth completion. Detailed comparison is shown in Fig. 4b and 4c. In these figures, we compare the LIP with different proximal operators, including the proposed adversarial-trained encoder-decoder network (LIP+AD), a supervised encoder-decoder network



(a) Convergence study of the proposed method with different types of training strategies. We conduct the test on the NYUdepthv2 dataset with 200 depth samples as input.

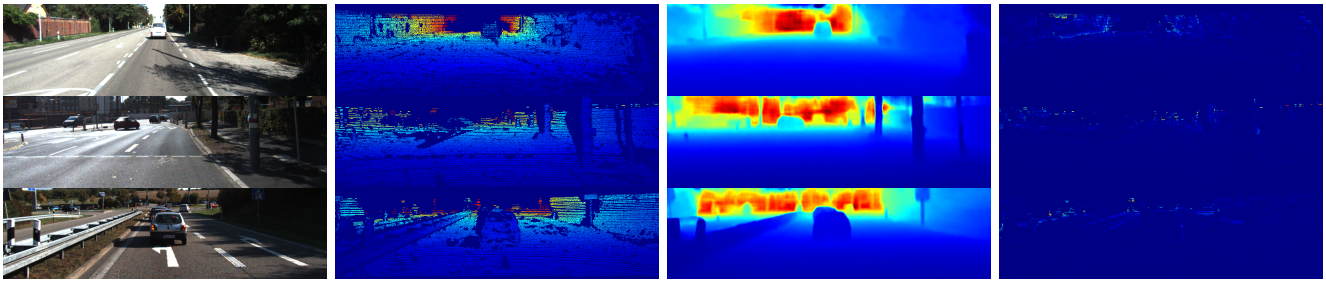


(b) Total MAE over all pixels, which indicates the performance of the prediction. A smaller MAE shows a better overall depth completion accuracy.



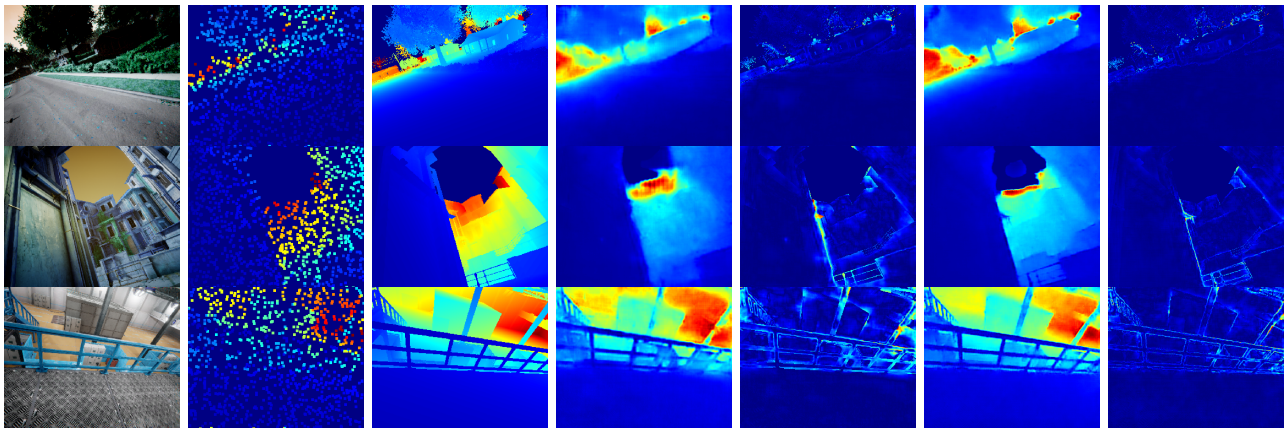
(c) MAE on pixels with depth sample input. A smaller MAE represents a better data consistency performance.

Fig. 4: Ablation Studies. In this figure, we show the convergence study of the proposed method as well as the comparison of the proposed method and previous works [2], [6], [7] in terms of overall performance and data consistency. LIP+AD is the proposed architecture with adversarial training. LIP+CNN is the proposed architecture without adversarial training. LIP+Wavelet applies the traditional wavelet-based proximal operator. We compare the MAE of these methods for different sample rates.



(a) RGB (b) Ground Truth (c) Our (d) Error

Fig. 5: Visualization result on KITTI dataset. The images from left to right are: RGB images, ground truth, prediction results for the proposed method (LIP+AD) and prediction error of the proposed method. In the error map, the value of the error increases from dark blue to dark red.



(a) RGB image (b) Input (c) Ground Truth (d) Ma et al. [2] (e) Error: Ma. [2] (f) LIP+AD (Our) (g) Error: Our

Fig. 6: Visualization result on Airsim dataset. The images from left to right are: RGB images, input sparse depth maps (visually enhanced by bilinear interpolation), ground truth, prediction results for Ma et al. [2], error for [2] and prediction results for the proposed method (LIP+AD) and error for the proposed method. We are using 200 depth points in this experiment setting. In the error map, the value of the error increases from dark blue to dark red.

TABLE I: Comparison of proposed network with state-of-the-art methods on NYUdepthv2 Dataset. In this table, we report the depth completion performance with 200 depth samples.

Method	RMSE	REL	δ^1	δ^2	δ^3
Ma et al. [2]	0.230	0.044	97.1	99.4	99.8
Unet [6]	0.203	0.040	97.6	99.5	**99.9**
Abdelrahman et al. [18]	0.192	0.030	97.9	99.5	99.8
SPN [19]	0.172	0.031	98.3	99.7	**99.9**
InductFusion [10]	0.169	0.028	98.4	**99.9**	**99.9**
CSPN [7]	0.162	0.028	98.6	99.7	**99.9**
LIP (Ours)	0.167	0.030	98.5	99.7	**99.9**
LIP+AD (Ours)	**0.153**	**0.026**	**98.9**	**99.9**	**99.9**

(LIP+CNN) and Wavelet transform (LIP + wavelet). We report the MAE on different subsets of pixels against different sub-sample rates. In Fig. 4b, we show the overall depth completion accuracy for different algorithms. As shown in the figure, LIP+Wavelet (yellow dotted line) does not perform better than the Unet-based [6] or InductFusion [10] methods at a high sample rate, due to the limited representation ability in the frequency domain. This family of domain transfer methods also suffers bottlenecks at extremely low sample rates. The learning-based methods [2], [6], [10] learn to interpolate by the guidance of RGB and depth image

data. By introducing a learning-based proximal operator into LIP, LIP+CNN achieves better performance. The adversarial training (LIP+AD) further improves the overall performance by learning more robust prior knowledge from the dataset. Fig. 4c indicates the data consistency of the depth estimation on given sparse pixel locations. For this figure, we notice that the LIP-based methods perform better than other deep learning-based methods. This is achieved by the fact that the $\mathbf{x}$-update function updates the depth completion result to minimize the objective $\|\mathbf{y} - A\mathbf{x}\|_2^2$, and by the restriction $\mathbf{x} = \mathbf{z}$. During this process, $\mathbf{x}$ is updated by the observation $\mathbf{y}$ and denoised by the signal prior. However, for the pure deep network-based method, the given depth pixel locations are blurred by neighbor values. In general, the learning-based LIP performance is better than the traditional hand-crafting-based LIP. This is achieved by the generalization characteristic of the signal prior trained from the large-scale dataset. The adversarial training only slightly benefits the accuracy, compared to supervised training.

TABLE II: Performance comparison of proposed network with state-of-the-art methods on KITTI validation dataset. We consider the previous RGB image and LIDAR fusion methods for comparison. RMSE and MAE are in meters, iRMSE and iMAE are in (1/km), and PARAM is in millions.

Methods	RMSE	MAE	iRMSE	iMAE	PARAM
CSPN [7]	1019.64	279.46	2.93	1.15	17.41
S2De [3]	814.73	249.95	2.80	1.21	26.10
PwP [20]	777.05	235.17	2.42	1.13	28.99
DLiDAR [9]	758.38	226.05	2.56	1.15	53.44
LIP (Ours)	779.61	238.37	2.48	1.12	7.88
LIP+AD (Ours)	**742.19**	**209.49**	**2.08**	**1.02**	**7.88**

C. Performance Comparison on Benchmark Datasets

In this section, we mainly conduct analysis on benchmark datasets and compare the performance with previous methods. We present the performance of the proposed method and state-of-the-art methods qualitatively and quantitatively.

In Table I, we compare the proposed method with previous methods on the indoor dataset (NYUdepthv2). The LIP with conventional CNN as signal prior does not perform better than previous method [7]. This is due to the fact that the signal prior trained by supervised learning is not as smooth as the CSPN network which better considers local features and details in the RGB image and depth map. LIP+AD performs slightly better than CSPN, with around 5% improvement in terms of RMSE. And we use only 45% of the parameters of CSPN, as shown in Table II. In this case, the proposed method can recover better details in the depth predictions with fewer parameters.

We also visualize the depth completion results for the proposed method LIP+AD and previous method CSPN on the NYUdepthv2 dataset. As shown in Fig. 3, our method makes a better depth prediction on the sofa in case 1 and also recovers more details of the bicycle in case 3. We also achieve a better prediction on the edges of the bed and clothes in case 2, compared to the CSPN method.

In Table II, we compare the proposed method with previous methods on the outdoor Dataset (KITTI). Compared with other methods, we achieve a slightly better performance in terms of RMSE and MAE. Though the proposed method does not have the best iMAE, it uses fewer parameters compared to previous works. Our model is also effective after being trained on KITTI alone, whereas previous works [9] require additional simulation datasets for training, or a pretrained model from Cityscapes [21]. Visualization results are shown in Fig. 5.

TABLE III: Performance comparison of proposed network with state-of-the-art methods on KITTI Odometry dataset and TartanAir dataset. We report the performance of the proposed method and previous methods with 200 and 500 depth samples.

Dateset	Method	#Depth	RMSE	REL	δ^1
KITTI	Unet. [6]	200	3.67	0.072	92.3
	InductFusion [10]	200	3.11	0.058	93.9
	LRP (Ours)	200	2.67	0.055	96.1
	LRP + AD (Ours)	200	**2.46**	**0.047**	**97.3**
KITTI	Ma et al. [2]	500	3.38	0.073	93.5
	Cheng et al. [7]	500	2.98	0.044	95.7
	InductFusion [10]	500	2.84	0.045	95.3
	LRP + AD (Ours)	500	**2.40**	**0.043**	**97.8**
TartanAir	Ma et al. [2]	200	2.49	0.139	91.4
	Unet. [6]	200	2.41	0.135	92.7
	InductFusion [10]	200	2.35	0.154	93.1
	LIP (Ours)	200	2.05	0.130	93.7
	LIP + AD (Ours)	200	**1.75**	**0.125**	**94.3**

In Table III, we show more results on the KITTI Odometry dataset and the simulation dataset TartanAir. As shown in the first two blocks of the table, our method can achieve a better performance with quite sparse input data, which simulates the point cloud captured by the sparse LIDAR. The third block of Table III indicates that our method can also generalize on the large-scale dataset, which mixes both the indoor and outdoor scenarios. We show some visualizations of the proposed method and a previous method [2] in Fig. 6.

V. Conclusion

In this paper, we formulate the sensor fusion-based depth completion problem as a Linear Inverse Problem (LIP). Instead of applying a traditional hand-crafted signal prior, we propose a multiple-modulation proximal operator with an encoder-decoder architecture. To further improve the smoothness of the proximal operator, we apply an adversarial training technique. The proposed method both guarantees data consistency and achieves smooth and realistic depth prediction, which outperforms previous methods. As a fast-converging iterative approach, our method allows a lightweight solution which has fewer parameters compared to other state-of-the-art methods. Future work will further reduce the number of parameters by training a single proximal operator to fit all iterations of the LIP. We will also extend the depth completion results to benefit other perception tasks, including object detection, shape completion and 3D tracking.

VI. Acknowledgments

This work was supported by the CMU Argo AI Center for Autonomous Vehicle Research.

REFERENCES

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

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

[3] 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.

[4] Y. Chen, B. Yang, M. Liang, and R. Urtasun, "Learning joint 2d-3d representations for depth completion," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[5] W. Wang, M. R. U. Saputra, P. Zhao, P. Gusmao, B. Yang, C. Chen, A. Markham, and N. Trigoni, "Deeppco: End-to-end point cloud odometry through deep parallel neural network," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 3248–3254.

[6] C. Fu, C. Mertz, and J. M. Dolan, "Lidar and monocular camera fusion: On-road depth completion for autonomous driving," in *IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 273–278.

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

[8] S. S. Shivakumar, T. Nguyen, S. W. Chen, and C. J. Taylor, "Dfusenet: Deep fusion of rgb and sparse depth information for image guided dense depth completion," *CoRR*, vol. abs/1902.00761, 2019.

[9] J. Qiu, Z. Cui, Y. Zhang, X. Zhang, S. Liu, B. Zeng, and M. Pollefeys, "Deeplidar: Deep surface normal guided depth prediction for outdoor scene from sparse lidar data and single color image," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[10] C. Fu, C. Dong, C. Mertz, and J. M. Dolan, "Depth completion via inductive fusion of planar lidar and monocular camera," *arXiv preprint arXiv:2009.01875*, 2020.

[11] A. Mousavi, A. B. Patel, and R. G. Baraniuk, "A deep learning approach to structured signal recovery," *CoRR*, vol. abs/1508.04065, 2015.

[12] W. Dong, L. Zhang, G. Shi, and X. Wu, "Image deblurring and super-resolution by adaptive sparse domain selection and adaptive regularization," *IEEE Transactions on Image Processing*, vol. 20, no. 7, p. 1838–1857, Jul 2011. [Online]. Available: http://dx.doi.org/10.1109/TIP.2011.2108306

[13] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," *arXiv preprint arXiv:1801.07892*, 2018.

[14] N. Chodosh and S. Lucey, "When to use convolutional neural networks for inverse problems," *arXiv preprint arXiv:2003.13820*, 2020.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *CoRR*, vol. abs/1512.03385, 2015.

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

[17] W. Wang, D. Zhu, X. Wang, Y. Hu, Y. Qiu, C. Wang, Y. Hu, A. Kapoor, and S. Scherer, "Tartanair: A dataset to push the limits of visual slam," *arXiv preprint arXiv:2003.14338*, 2020.

[18] A. Eldesokey, M. Felsberg, and F. S. Khan, "Propagating confidences through cnns for sparse data regression," *arXiv preprint arXiv:1805.11913*, 2018.

[19] R. Liu, G. Zhong, j. Cao, and Z. Lin, "Learning to diffuse: A new perspective to design pdes for visual analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016.

[20] Y. Xu, X. Zhu, J. Shi, G. Zhang, H. Bao, and H. Li, "Depth completion from sparse lidar data with depth-normal constraints," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[21] M. Cordts, M. Omran, S. Ramos, T. Scharwächter, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset," in *CVPR Workshop on The Future of Datasets in Vision*, 2015.