PAPER ANALYSIS

Presented by Yannis He

-

Paper: **Depth Completion from Sparse LiDAR Data with Depth-Normal Constraints**

Conference: ICCV 2019

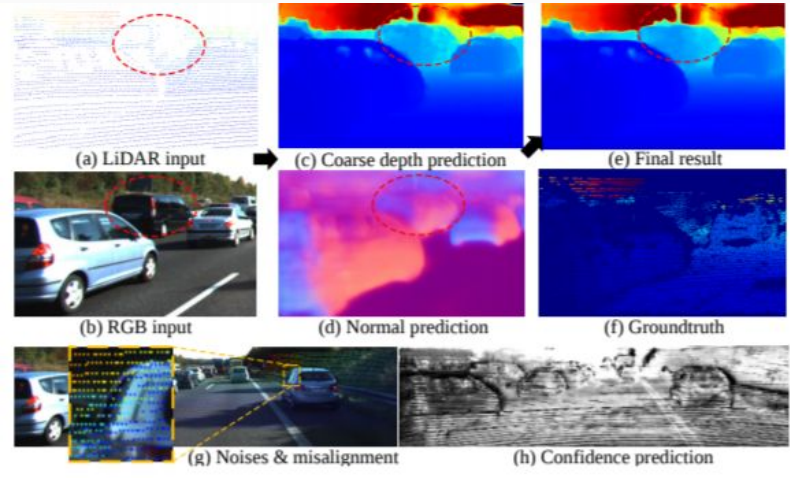
Authors: Yan Xu, Xinge Zhu, Jianping Shi, Guofeng Zhang, Hujun Bao, Hongsheng Li

- SenseTime Research
- The Chinese University of Hong Kong
- State Key Lab of CAD&CG, Zhejiang University

https://arxiv.org/abs/1910.06727

INTRODUCTION

- Depth Completion:
  - Aims to recover dense depth maps from sparse depth measurements
- Existing methods:
  - Directly train a network to learn a mapping from sparse depth input to dense depth maps output
    - Has difficulties in utilizing the 3D geometric constraints and handling the practical sensor noises.
  - Mainstream: input the sparse depth maps into an encoder-decoder network and predict dense depth maps
    - These black-box methods force the CNN to learn a mapping from sparse depth measurement to dense maps
      - Generally a challenging task and lead to unsatisfactory completion results as shown in the image
- Idea: proper geometric constraints should be incorporated into the end-to-end framework to regularize the completion process and make it more interpretable.
  - Depth and surface normal are two strongly correlated factors in the 3D world
  - The locally linear orthogonality between them can be utilized in depth completion



Presented by Yannis He

INTRODUCTION (CONT')

- Contribution:
  - To regularize the depth completion and improve robustness against noise
  - Propose a unified CNN framework that
    - Models the geometric constraints between depth and surface normal in a diffusion module
      - Model the locally linear orthogonality between depth and normal by associating them in the plane-origin distance space
    - Predicts the confidence of sparse LiDAR measurements to mitigate the impact of noise
  - The encoder-decoder backbone predicts surface normals, coarse depth, and confidence of LiDAR inputs simultaneously
  - Subsequently inputted into the diffusion refinement module to obtain the final completion results
  - 3 main contributions:
    1. Reposition the focus of depth completion from 2D space to 3D space based on the assumption that a 3D scene is constituted by piecewise planes
    2. Propose a unified two-stage CNN framework to achieve depth completion from very sparse input
    3. Framework can be trained in an end-to-end manner, and achieves state-of-the-art performance.
- Process:
  1. Adopt a CNN-based backbone to estimate the surface normal and depth from sparse LiDAR measurements and color images
  2. Transform the predicted depth and normal to the plane-origin distance space
  3. Conduct a refinement process in this space via a diffusion model to enforce the geometric constraints

Presented by Yannis He

RELATED WORK

- Depth Completion
  - Existing approaches: mainly mainly aim to handle the incomplete depth measurements from 2 types of sensors:
    - Structured-light scanners
      - Widely used in 3D reconstruction post-processing
    - LiDAR
      - Usually require real-time responses in the scenarios of robotics
  - Classic methods generally employ hand-crafted features or kernels to complete the missing values
    - Most are task-specific and usually confronted with performance bottleneck due to generalization ability
  - Learning-based methods:
    - Sparsity-invariant convolution layer to enhance the depth measurements from LiDAR
      - Also model the confidence propagation through layers and reduce the quantity of model parameters
    - Sparse depth and color images as inputs of an off-the-shelf network with self-supervised LiDAR completion
    - Encoder-decoder framework tends to predict the depth maps comprehensively
      - But fail to concentrate on the local areas
    - Convolutional spatial propagation refinement network to post process the depth completion results with neighboring depth values
      - Refinement in 2D depth space based on assuming the depth values are locally constant
        - This assumption is sub-optimal and outdoor scene performance is barely satisfactory
  - Current approaches ignore the noises in LiDAR measurements, which are inevitable in practices

RELATED WORK

- Depth and Normal
  - The Relations between depth and surface normal has been exploited to improve the depth accuracy
  - Monocular depth estimation:
    - Compute normal from depth and then recover the depth from normal inversely
  - Can benefit from geometric constraints & solve by Cholesky factorization
    - Optimize for linear system is hard to be employed in end-to-end framework and achieve joint optimization
  - Suitable for post processing the RGB-D camera data
  - Can hardly achieve real-time processing
- Anisotropic Diffusion
  - Anisotropic diffusion originally models the physical process that equilibrates concentration differences without creating or destroying mass, e.g. heat diffusion.
  - Has been extensively used in image denoising, depth compilation, segmentation.
  - Previous methods define diffusion conductance only based on the similarity in diffusion space or in the guidance map (e.g. a color image)
    - This limits the performance
    - In this work, we take advantage of feature extraction capability of CNN and use the high-dimension features to calculate the conductance.

METHOD

- Assumption
  - 3D scene is constituted by piecewise planes
  - Distances between these planes and the origin are piecewise constant
- Proposal:
  - A two-stage end-to-end deep learning framework: 1. Prediction network, 2. Refinement network
    - regularizes the depth completion process using the constraints between depth and surface normal

1. Prediction Network:
   a. Estimates the surface normal map, the coarse depth map, and confidences of sparse depth inputs with a shared-weight encoder and independent decoders
   b. The sparse input and coarse depth maps are transformed to the plane-origin distance subspace with normal estimation
2. Refinement Network:
   a. A diffusion model recurrently refines plain-origin distance.
   b. This enforces the piecewise plain constraints and regularizes the depth completion.

  - This framework utilizes the depth & surface normal geometric constraint

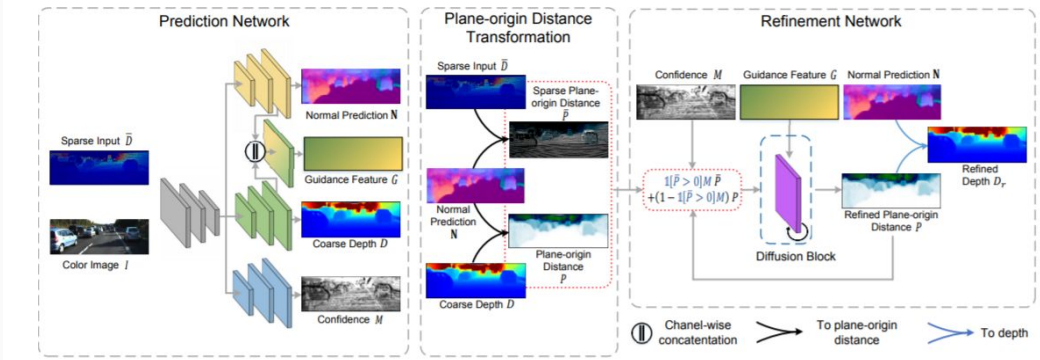


Figure 2: Overview of our proposed framework. The prediction network first predicts maps of surface normal $\mathbf{N}$, coarse depth D and confidence M of sparse depth input with a shared-weight encoder and independent decoders. Then, the sparse depth inputs $\bar{D}$ and coarse depth D are transformed to the plane-origin distance space as $\bar{P}$ and P, using Eq. (5). Next, the refinement network, an anisotropic diffusion module, refines the coarse depth map D in the plane-origin distance subspace to enforce the constraints between depth and normal and to incorporate information from the confident sparse depth inputs. During the refinement, the diffusion conductance depends on the similarity in guidance feature map G (See Eq. (7)). Finally, the refined P is inversely transformed back to obtain the refined depth map D_r when the diffusion is finished.

METHOD (CONT')

- Prediction Network
  - Takes sparse depth and the corresponding color image as input
    - Predicts surface normal map, coarse depth completion, and confidence map via separate decoders
    - U-Net architecture for prediction network:
      - ResNet-34 variant as encoder & cascaded upsampling layers as decoders
- Recurrent Refinement Network:
  - Issue:
    - The encoder-decoder architecture doesn't exploit the geometric constraints between depth and surface normal to regularize the estimated depth
    - It has difficulties of taking full advantages of the sparse inputs.
  - Proposed solution:
    - We need to further refine the completion results in a novel plane-origin distance subspace via an anisotropic diffusion module
      - Based on the assumption that the 3D surface of the scene is constituted by piecewise planes and the plane-origin distance is piecewise constant.

Algorithm 1 The refinement procedure

1: **for all** $\mathbf{x}$ **do**
2: $\quad \bar{P}(\mathbf{x}) \leftarrow \bar{D}(\mathbf{x})\mathbf{N}(\mathbf{x})\mathbf{C}^{-1}\mathbf{x}$
3: $\quad P(\mathbf{x}) \leftarrow D(\mathbf{x})\mathbf{N}(\mathbf{x})\mathbf{C}^{-1}\mathbf{x}$
4: **end for**
5: $i \leftarrow 0$
6: **while** $i < max\_iteration$ **do**
7: $\quad$ **for all** $\mathbf{x}$ **do**
8: $\quad\quad P(\mathbf{x}) \leftarrow \mathbb{1}[\bar{P}(\mathbf{x}) > 0]M(\mathbf{x})\bar{P}(\mathbf{x}) + (1 - \mathbb{1}[\bar{P}(\mathbf{x}) > 0]M(\mathbf{x}))P(\mathbf{x})$
9: $\quad$ **end for**
10: $\quad$ **for all** $\mathbf{x}$ **do**
11: $\quad\quad$ Conduct the refinment using Eq. (6)
12: $\quad$ **end for**
13: $\quad i \leftarrow i + 1$
14: **end while**
15: **for all** $\mathbf{x}$ **do**
16: $\quad D(\mathbf{x}) \leftarrow P(\mathbf{x})/(\mathbf{N}(\mathbf{x})\mathbf{C}^{-1}\mathbf{x})$
17: **end for**

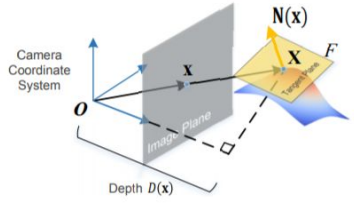
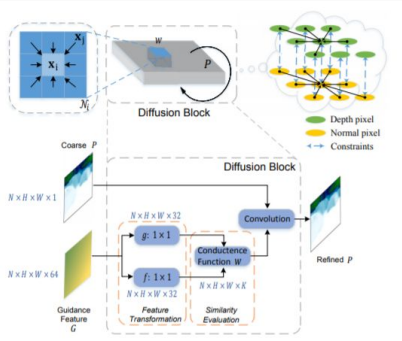


Figure 3: In camera coordinate system, the relation between depth and normal can be established via the tangent plane equation.

Figure 4: The proposed differentiable diffusion block. In each refinement iteration, high-dimensional feature vectors (*e.g.*, of dimension 64) in guidance feature map G are independently transformed via two different functions f and g (modeled as two convolution layers followed by normalization). Then, the conductances from each location $\mathbf{x}_i$ (in plane-origin distance map P) to its neighboring K pixels ($\mathbf{x}_j \in \mathcal{N}_i$) are calculated using Eq. (7). Finally, the diffusion is performed through a convolution operation with the kernels defined by the previous computed conductances. Through such diffusion, depth completion results are regularized by the constraint between depth and normal.

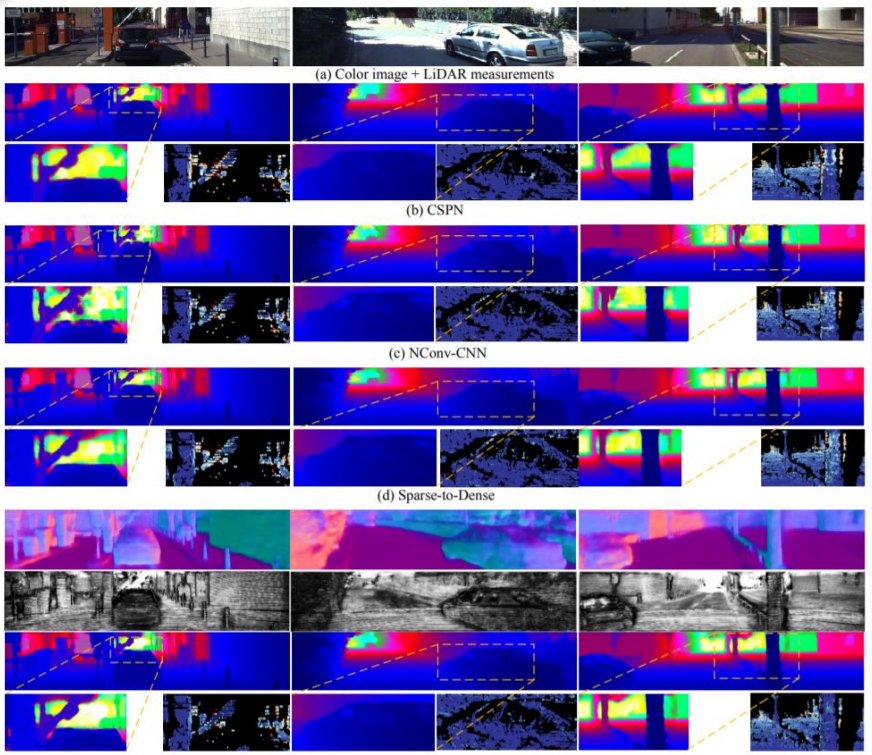


RESULTS

- 1st on the test set of KITTI depth completion benchmark with RMSE metric
- Quantitative comparison with competing approaches as shown on the right

Table 1: The evaluation results on the test set of KITTI depth completion benchmark. The root mean square error (RMSE) and mean absolute error (MAE) are in millimeters, while inverse RMSE and inverse MAE are in 1/kilometer.

Method	RMSE	MAE	iRMSE	iMAE
Ours	777.05	235.17	2.42	1.13
Sparse-to-Dense [23]	814.73	249.95	2.80	1.21
NConv-CNN [10]	829.98	233.26	2.60	1.03
Spade-RGBsD [16]	917.64	234.81	2.17	0.95
HMS-Net [14]	937.48	258.48	2.93	1.14
CSPN [3]	1019.64	279.46	2.93	1.15
Morph-Net [7]	1045.45	310.49	3.84	1.57
DFuseNet [33]	1206.66	429.93	3.62	1.79



Presented by Yannis He