TB-CompletionFormer: Improved Depth Completion based on Two-branch Backbone

Seokyoung Kim, Chansoo Kim*
Chonnam National University
Contents

- Background and Objectives
- Two branch Backbone CompletionFormer
  - Coarse-Branch
  - Fine-Branch
  - Spatial Propagation Network
- Experimental results
- Conclusion
Perception of the surrounding environment is crucial!
Perception of the surrounding environment is crucial!

Perception tasks require accurate depth values!
Research Background (I)

- Sensors for Depth Estimation

3D LiDAR

Stereo camera
Research Background (II)

- **Point cloud from LiDAR**
  - Laser scanning data from of the driving environment
  - Accurate depth of the objects
  - Sparsity (~5% density) issues when projecting to the image plane

Point cloud from LiDAR

LiDAR

Accurate depth

Sparse Points from LiDAR

Projection 3D to 2D
Research Background (III)

- **Stereo Camera (RGB Image)**
  - Depth estimation using triangulation methods

\[
Z = \frac{f \cdot T}{x_l - x_r}
\]

Highly Dense & Use color information!
- Image-based depth estimation has limitations
  - Inaccurate depth estimation especially for faraway objects

Sensor fusion
- Leverage complementary characteristics of LiDAR and Camera sensor

Density
Depth
Accuracy

LiDAR
Works in Bright
Works in Dark

Camera
Provides Color

KSAE 2023 Annual Fall Conference
How can we improve?

Sensor fusion
- Leverage complementary characteristics of LiDAR and Camera sensor

Density
Depth Accuracy
Provides Color
Works in Bright
Works in Dark

Sensor Fusion
Depth Completion

- Given a image and sparse depth map from LiDAR, depth completion aims to obtaining a dense prediction by information propagation.
Related Works (I)

- PE-Net: Towards Precise and Efficient image guided depth completion

- Two-branch backbones consists of a color-dominant path and a depth-dominant path. It can thoroughly exploit and fuse color and depth modalities.

- Designed the backbone using a pure Convolutional Neural Network (CNN)
Related Works (II)

Problems of Previous Architecture Design

- CNNS can only aggregate within local regions, tough to model global long-range relationship.
- Pure Vision Transformer projects image patches into vectors, causing the loss of local details, and high computation cost.
• Propose a Joint Convolutional Attention and Transformer (JCAT) block, by the integration of CNNs and Vision Transformer, enables both local and global propagation for depth completion

• JCAT block consists of a convolutional path and a single transformer path respectively
The data of two modalities are complementary to each other

- Overall depth is reliable but suffered from the heavy noise existing near object boundaries in the sparse input.

- Predicted depth map from RGB image is relatively reliable around object boundaries but may be too sensitive to the change of color or texture.
Proposing a Two-branch depth completion model using the CNN and Vision Transformer

- We designed a two-branch backbone that adaptively fuses color and depth modalities thoroughly

- Proposed model enables the extraction local and global features for accurate depth completion
Proposed Methods
Proposed Methods (I)

- Architecture of TB-CompletionFormer
  - Two-branch backbone to fuse complementary modality
  - Add residual connection within the JCAT block

Steps of Methods

Step 1. Coarse-branch
Step 2. Fine-branch
Step 3. Depth fuse & Refine
Proposed Methods (II)

- **Step 1. Coarse-Branch**

  - Coarse-Branch predicts a dense depth map mainly relying on **color information**
  - Coarse-Branch extracts color-dominant features for depth prediction so that the depth around object boundaries can be learned by taking advantage of structure information in the color image
  - Still, it is hard to predict accurate depth values
Step 2. Fine-Branch

- Fine-Branch predicts a dense depth map but depending more on depth information.
- Depth prediction result obtained from the Coarse-Branch is input to Fine-Branch, and we constructed same encoder-decoder as previous branch.
Step 2. Fine-Branch

- Fine-Branch also predicts a dense depth map but depending more on depth information.
- Depth prediction result obtained from the Coarse-Branch is input to Fine-Branch, and we constructed same encoder-decoder as previous branch. The depth maps predicted from two branches are adaptively fused.
- Our methods can exploit color and depth-dominant information respectively from two branches.
Proposed Methods (IV)

- **Residual Connection in Step 1. & Step 2.**
  - Processes the input as received and adds the residual information
  - The deeper the depth, the higher the accuracy
  - Add a residual block within the JCAT block considering to long learning process
Proposed Methods (V)

Step 3. Depth refinement: Spatial Propagation Network (SPN)

- Failure to fully preserve the initial valid depth value while learning depth information
- The fused map is further fed into the refinement module to enhance the depth quality
- Refining missing values in the pixel to obtain the accurate final depth
Loss Functions

- **Training Loss : L2**

\[
L(\hat{D}) = \| (\hat{D} - D_{gt}) \odot 1(D_{gt} > 0) \|^2
\]

- **Depth prediction Loss**

\[
L = \lambda_{cb} L(\hat{D}_{cb}) + \lambda_{fb} L(\hat{D}_{fb})
\]

Empirical setting hyperparameter
Experiments
Experimental Environments

- Dataset for training and testing
  - Open dataset: NYUv2
    - 20k indoor color images and sparse depth map

- Computing unit
  - GPU: NVIDIA RTX A6000 (x2)
Evaluation Metrics

- **Evaluation Metric for depth completion**
  - Root Mean Squared Error (RMSE)
  - Depth difference between predicted depth map and Ground Truth (GT) depth map

\[
RMSE = \sqrt{\frac{\sum_{i=1}^{N} (predicted_i - Ground Truth_i)^2}{N}}
\]

- **Predicted depth map**
  - Error: 3mm

- **GT depth map**
  - Error: 5mm
## Comparison with State-of-the-art Methods

### Benchmark on NYUv2 datasets

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE (m)</th>
<th>REL (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSPN++</td>
<td>0.117</td>
<td>0.016</td>
</tr>
<tr>
<td>DeepLiDAR</td>
<td>0.115</td>
<td>0.022</td>
</tr>
<tr>
<td>TWISE</td>
<td>0.097</td>
<td>0.013</td>
</tr>
<tr>
<td>NLSPN</td>
<td>0.092</td>
<td>0.012</td>
</tr>
<tr>
<td>RigNet</td>
<td>0.090</td>
<td>0.012</td>
</tr>
<tr>
<td>CompletionFormer</td>
<td>0.090</td>
<td>0.012</td>
</tr>
<tr>
<td>DYSPN</td>
<td>0.090</td>
<td>0.012</td>
</tr>
<tr>
<td>BEV@DC</td>
<td>0.089</td>
<td>0.012</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>0.089</strong></td>
<td><strong>0.011</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>RMSE (mm)</th>
<th>REL (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CompletionFormer</td>
<td>907.1</td>
<td>121.3</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>890.1</strong></td>
<td><strong>115.9</strong></td>
</tr>
</tbody>
</table>

- Our proposed model outperforms the baseline methods on NYUv2 datasets.
- It also performs as well as or better than State-of-the-arts models.
Qualitative Comparison with base Methods

- Visualization on NYUv2 Test Set

(a) GT Depth / RGB
(b) CompletionFormer
(c) Ours
Qualitative Comparison with base Methods

Visualization on NYUv2 Test Set

(a) GT Depth / RGB  (b) CompletionFormer  (c) Ours

RMSE : 906mm  RMSE : 892mm
Conclusion
Conclusion

- **TB-CompletionFormer**
  - We designed two-branch backbone based on CompletionFormer
  - Our method is able to **exploit and fuse complementary modalities thoroughly**
  - It also enables the extraction of **local** and **global** features for accurate depth completion
  - Compared to the base model, it **outperformed results in all metrics**

- **Future works**
  - We need to decrease its runtime further to use real-time
  - It is necessary to sufficiently verify the model based on the KITTI Dataset
Thank you