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

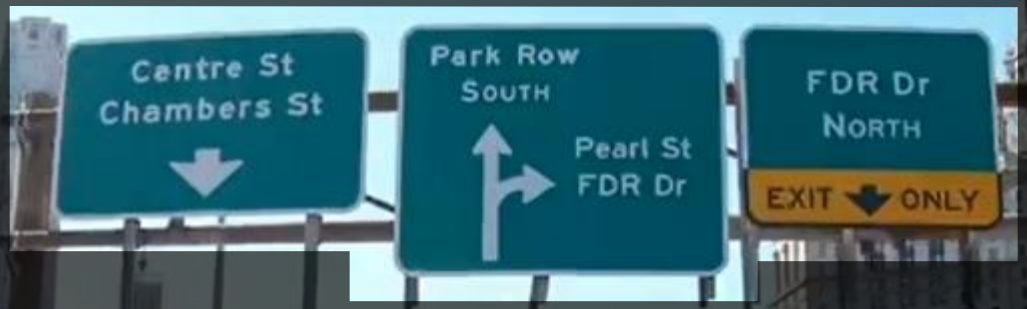
Seokyoung Kim, Chansoo Kim*
Chonnam National University

AIM Lab
Artificial Intelligence & Mobility Laboratory

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!



Traffic sign

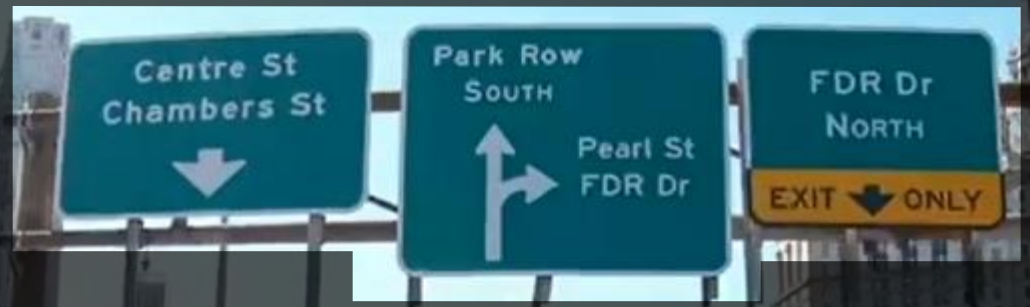
Vehicles

Fence

Fence

Road

Perception of the surrounding environment is crucial!



Traffic sign

Perception tasks require accurate depth values!

Object Detection

6m 9m 3m

Fence

Fence

Obstacle Avoidance

Research Background (I)

❑ Sensors for Depth Estimation

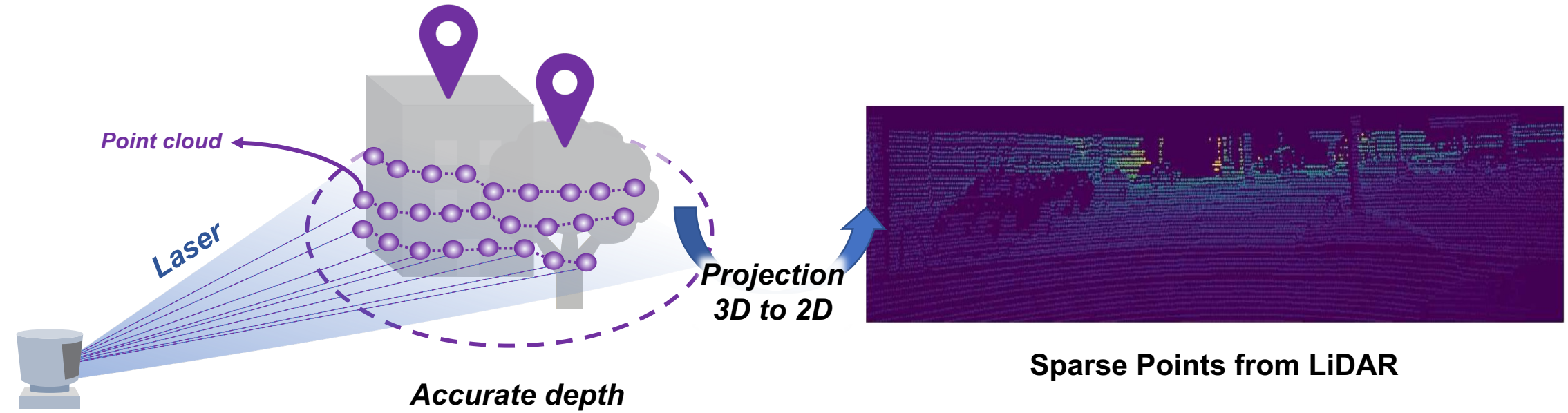
3D LiDAR

Stereo camera

Research Background (II)

❑ Point cloud from LiDAR

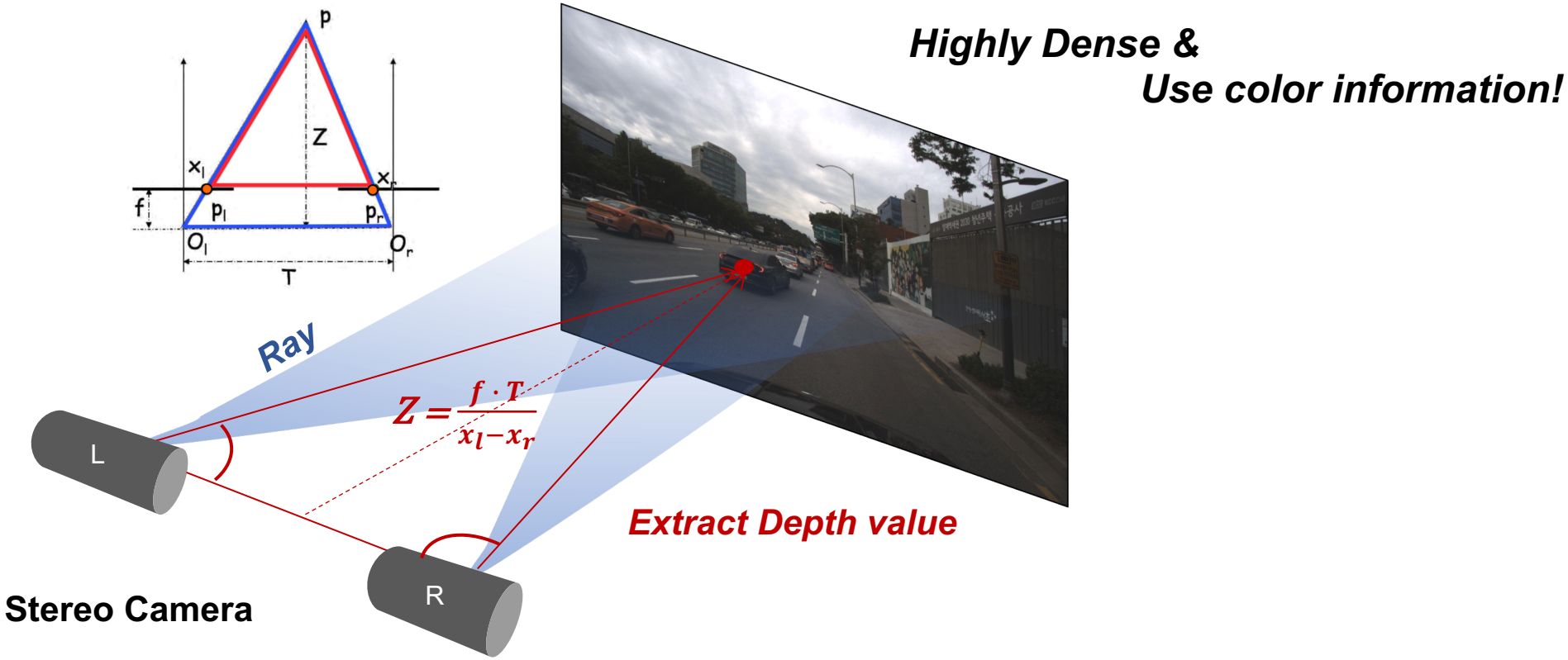
- Laser scanning data from of the driving environment
- Accurate depth of the objects
- Sparsity (~5% density) issues when projecting to the image plane



Research Background (III)

❑ Stereo Camera (RGB Image)

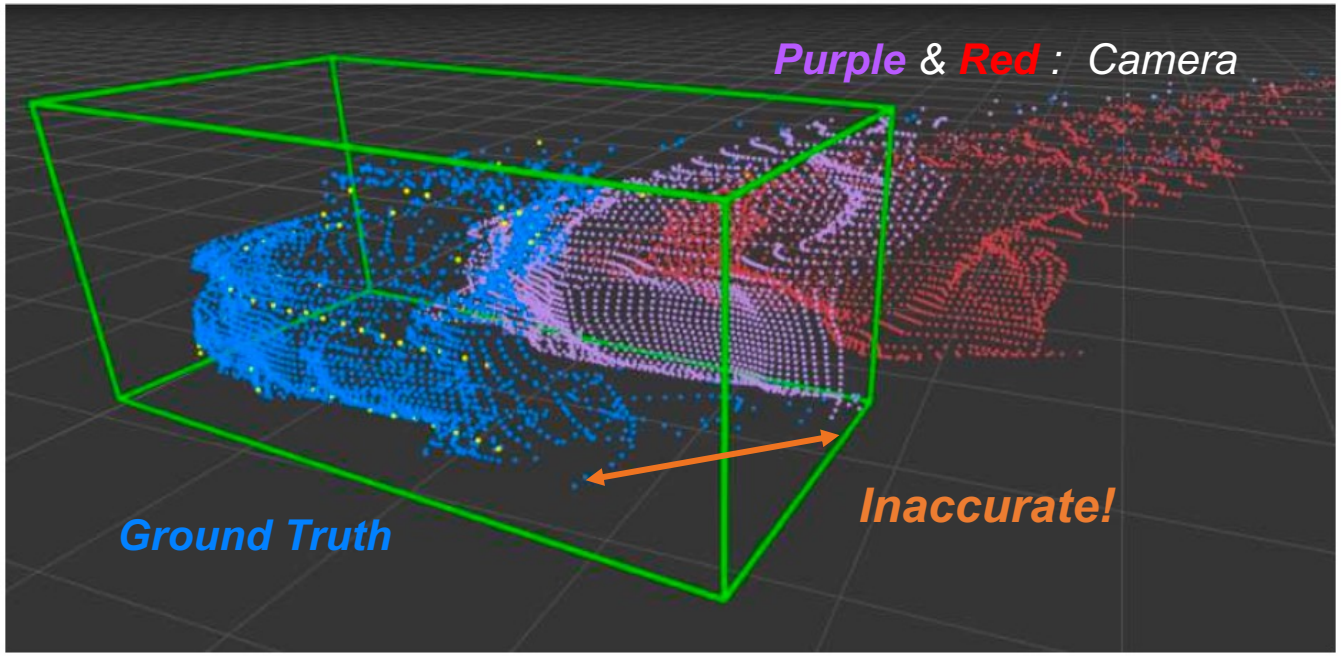
- Depth estimation using triangulation methods



Research Background (IV)

❑ Image-based depth estimation has limitations

- Inaccurate depth estimation especially for faraway objects

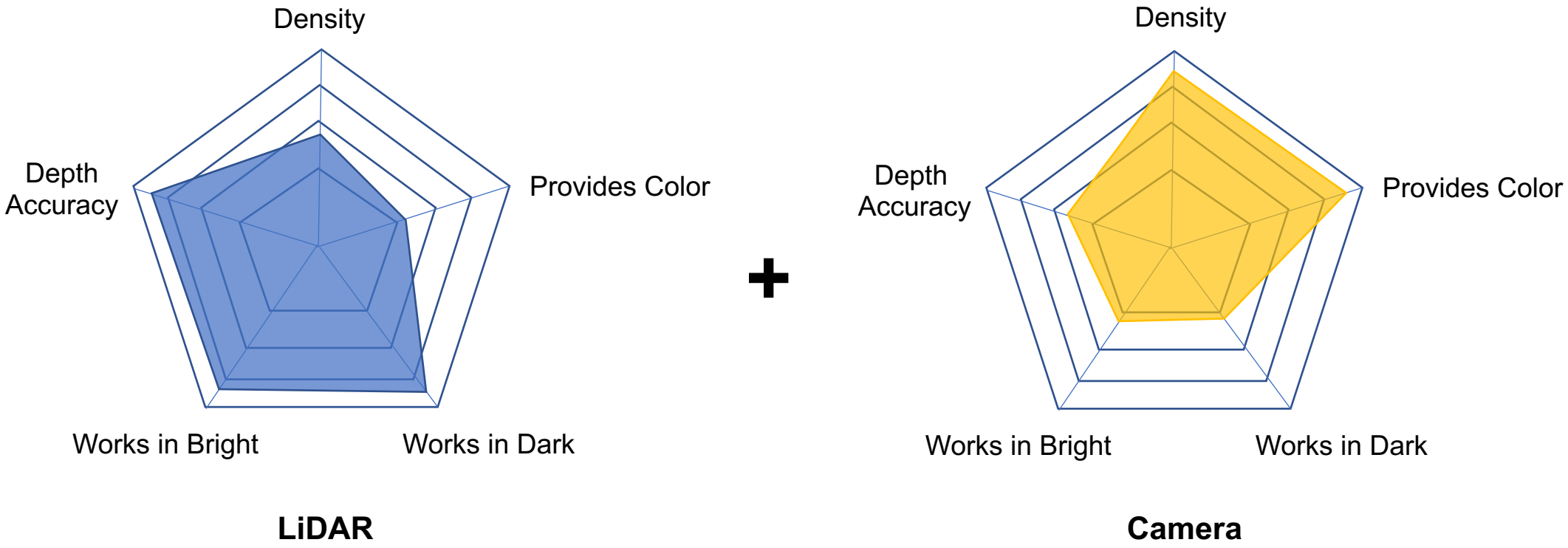


* Yurong You, Yan Wang and Kilian Q.Weinberger. “Pseudo-LiDAR++: Accurate depth for 3D Object Detection in Autonomous Driving” in Proc. of the IEEE/RSJ Intl. ICLR, 2020

How can we improve?

❑ Sensor fusion

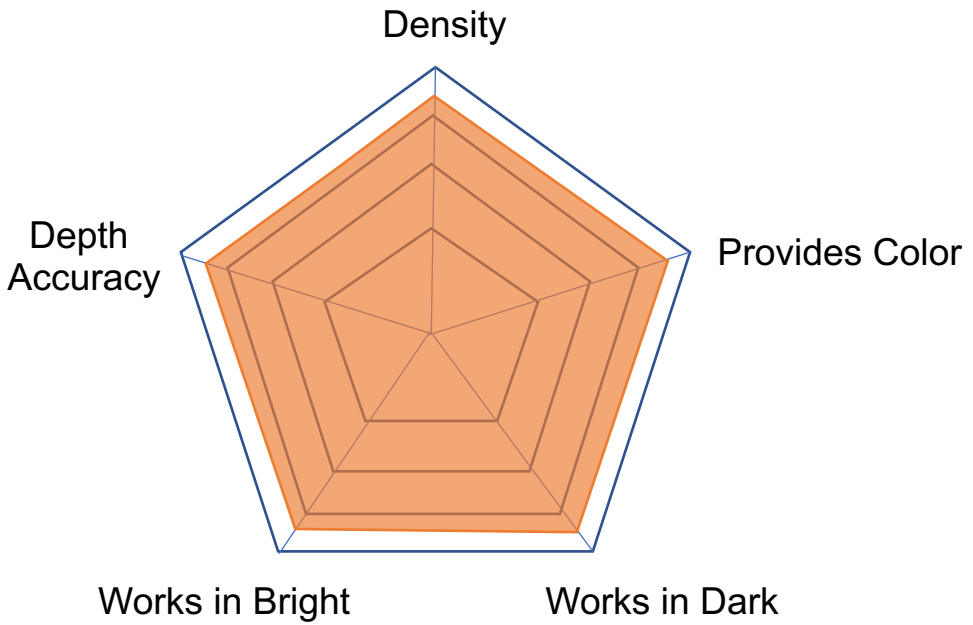
- Leverage complementary characteristics of **LiDAR** and **Camera** sensor



How can we improve?

❑ Sensor fusion

- Leverage complementary characteristics of **LiDAR** and **Camera** sensor

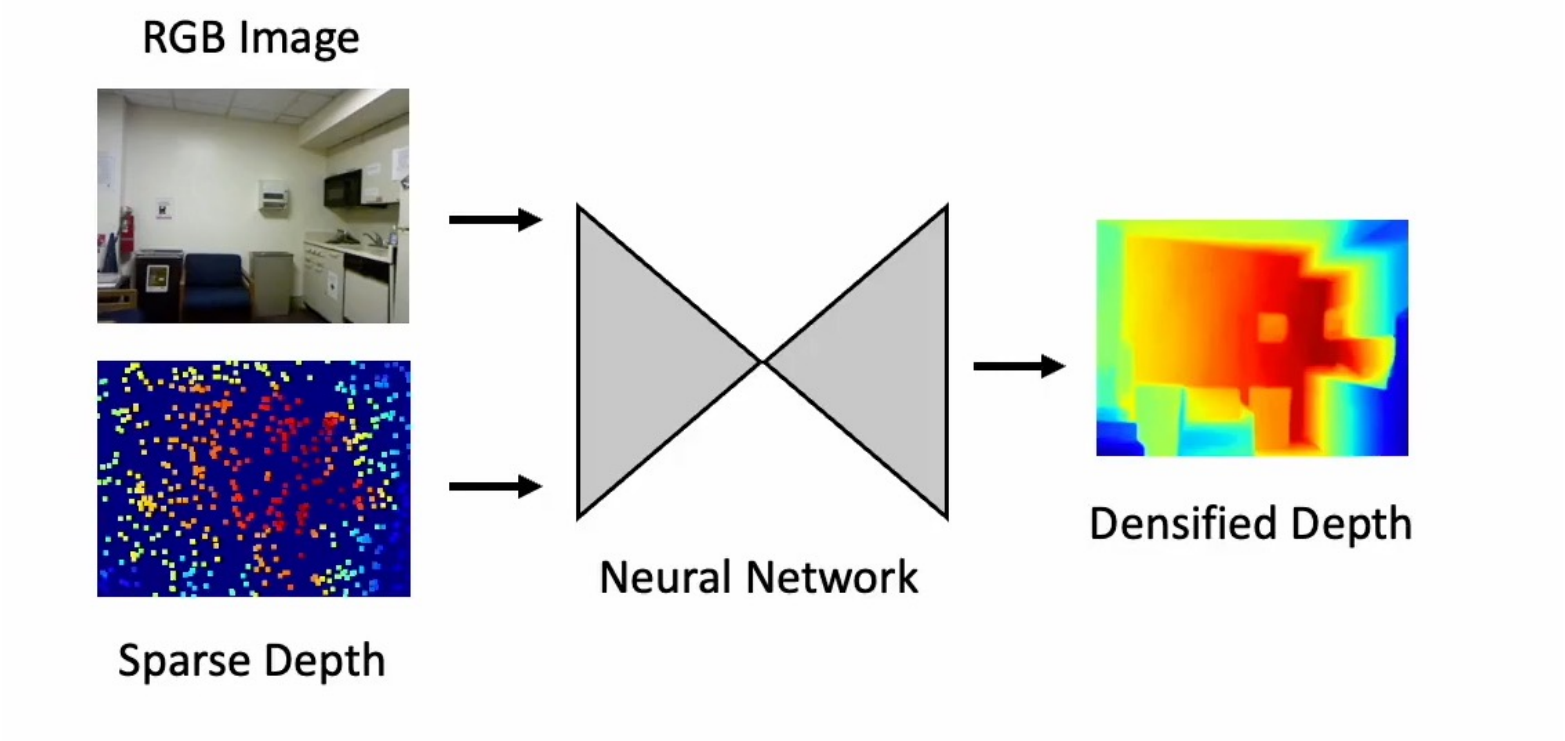


Sensor Fusion

Task of Sensor Fusion

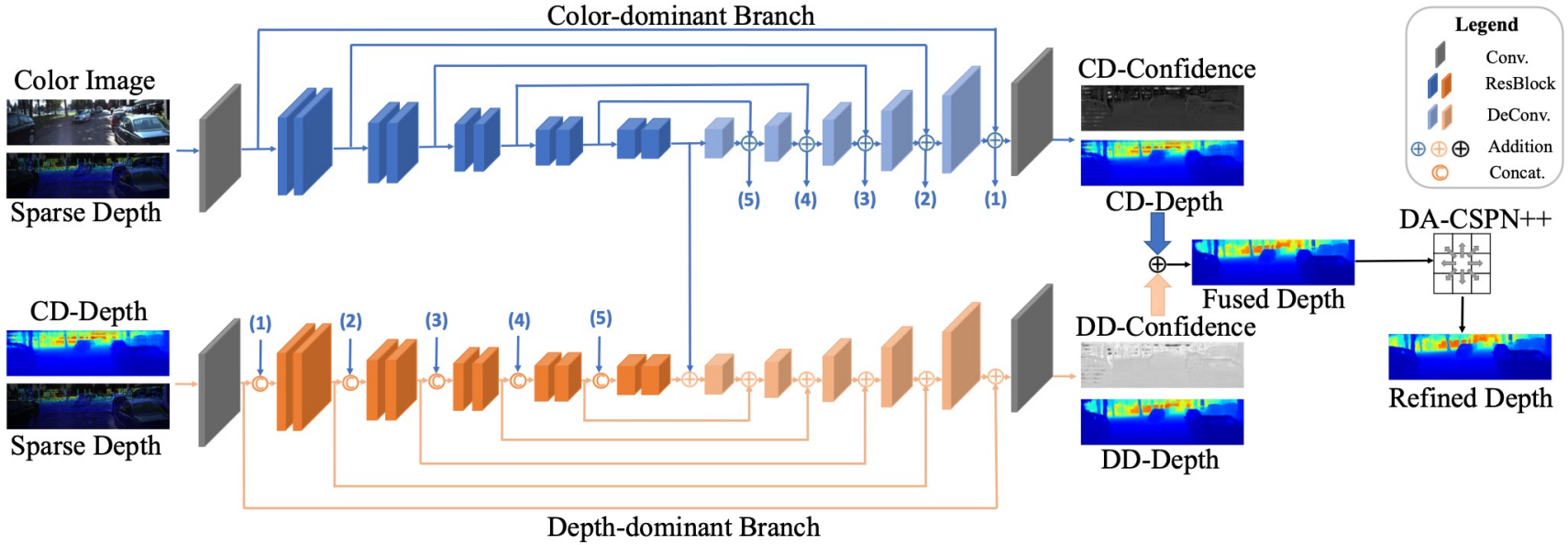
❑ Depth Completion

- 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

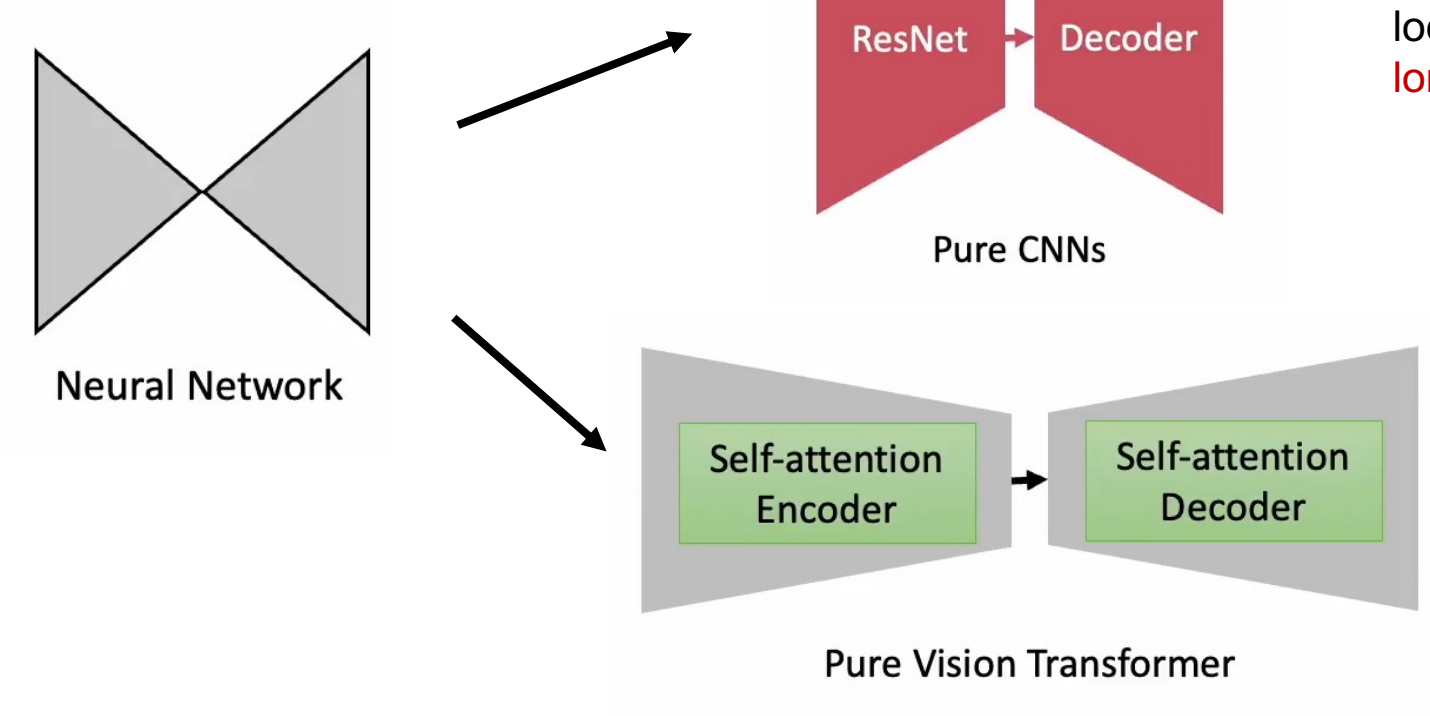


- Two-branch backbones consists of a color-dominant path and a depth-dominant path. It can thoroughly exploit and fuse color and depth modalities

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

Related Works (II)

❑ Problems of Previous Architecture Design

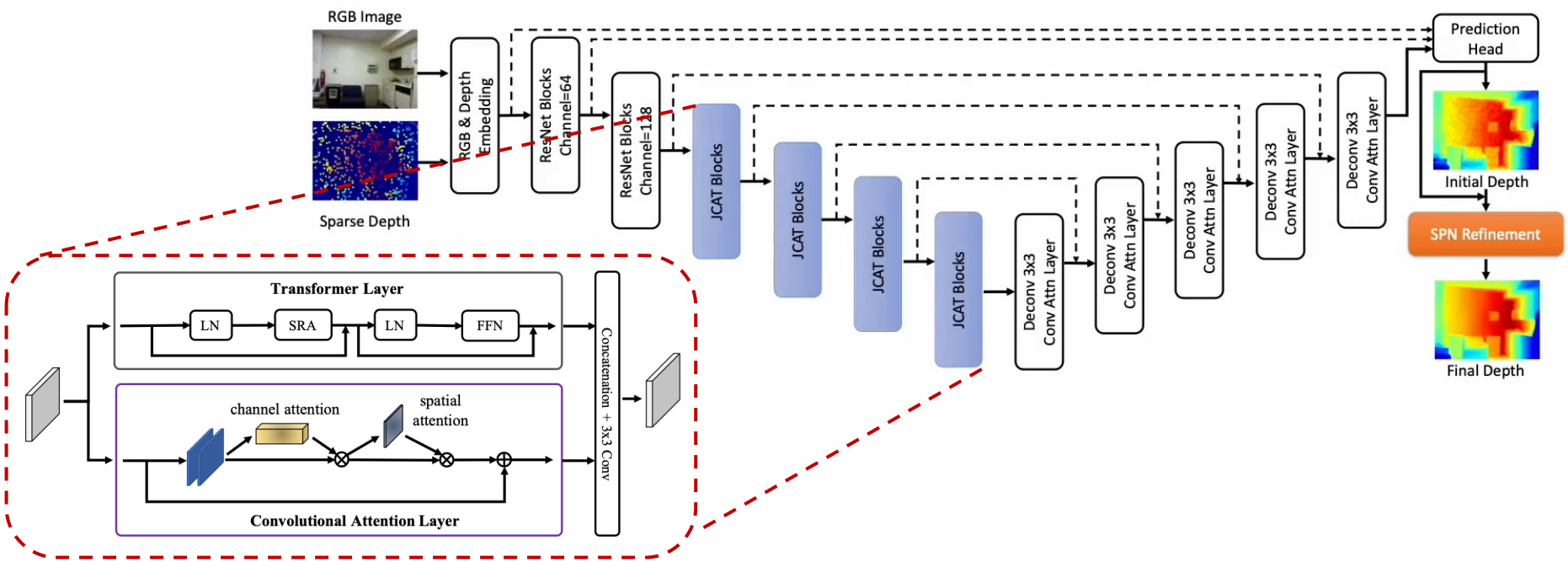


- CNNS can only aggregate within local regions, tough to model global long-range relationship

- Pure Vision Transformer projects image patches into vectors, causing the loss of local details, and high computation cost

Related Works (III)

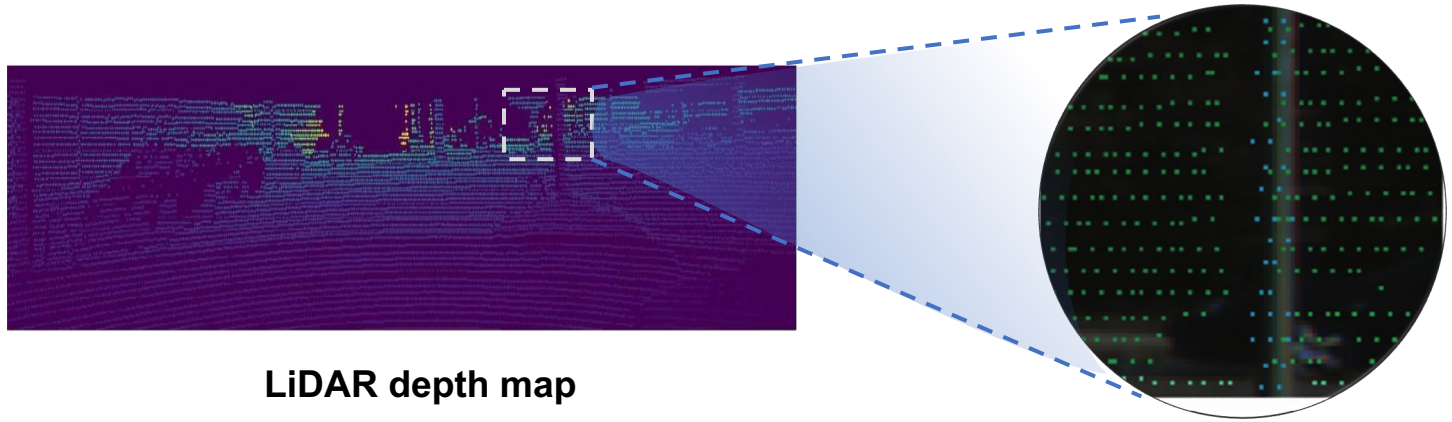
❑ CompletionFormer : Depth Completion with Convolutions and Vision Transformers



- 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

Difference between LiDAR and Camera modality

❑ The data of two modalities are complementary to each other



LiDAR depth map

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

RGB image

- Predicted depth map from RGB image is relatively reliable around object boundaries but may be too sensitive to the change of color or texture

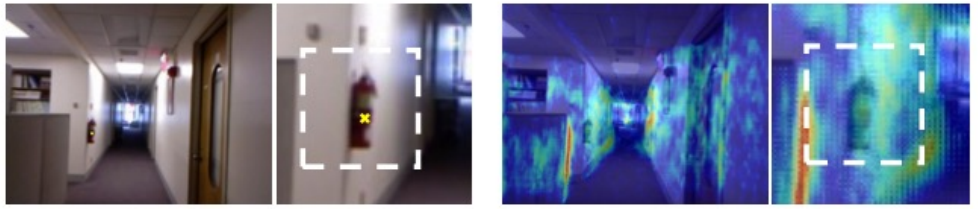
Research Objectives

❑ Proposing a Two-branch depth completion model using the CNN and Vision Transformer

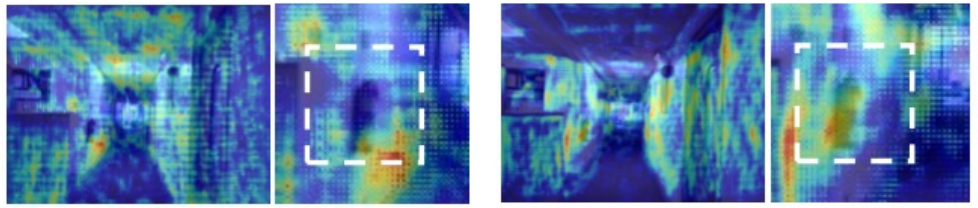
(a) RGB (c) RGB + LiDAR (b) LiDAR

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

(d) RGB (e) Pure CNNs



(f) Pure Transformer (g) CNN + Transformer



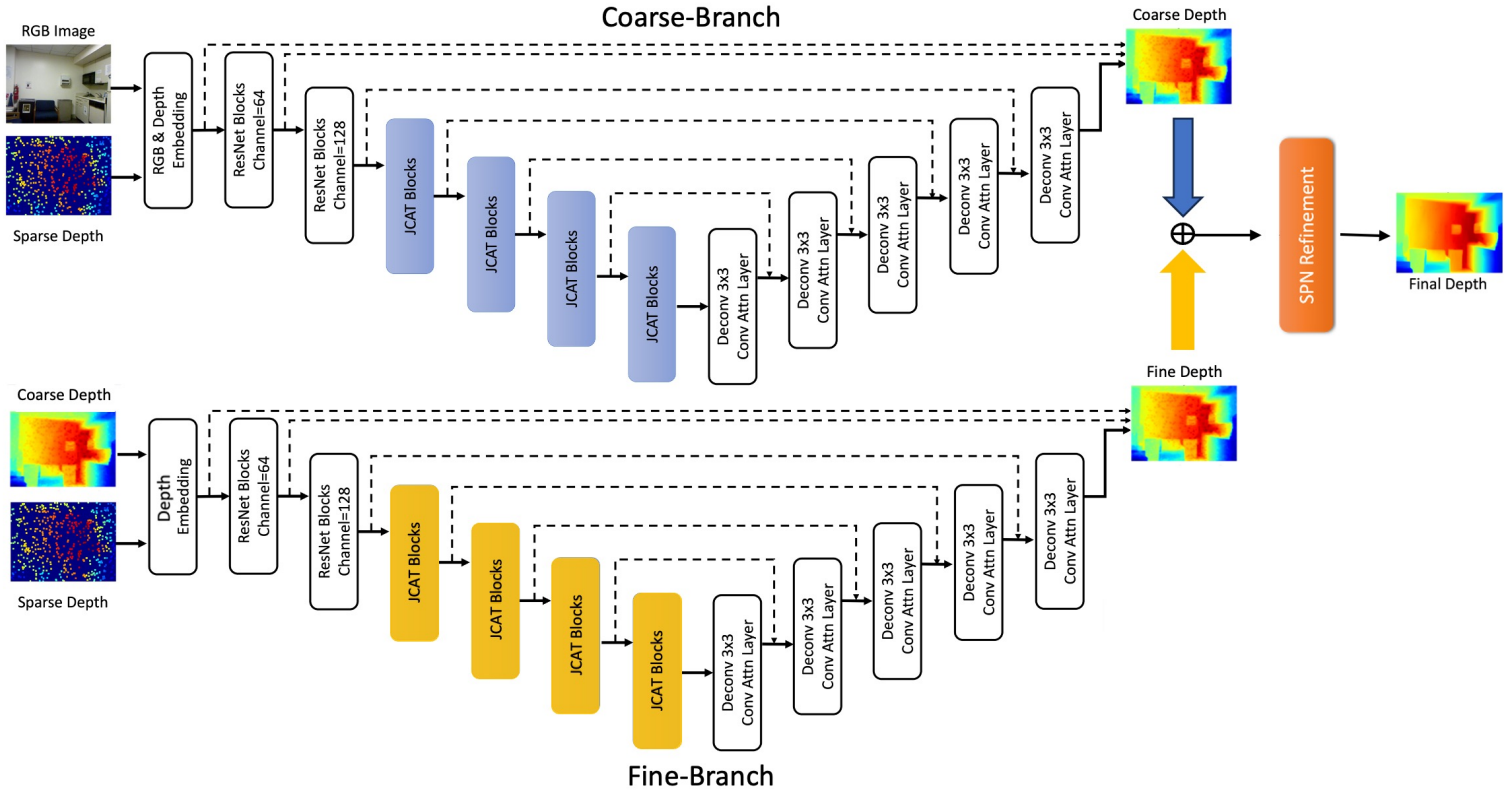
- Proposed model enables the extraction local and global features for accurate depth completion

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

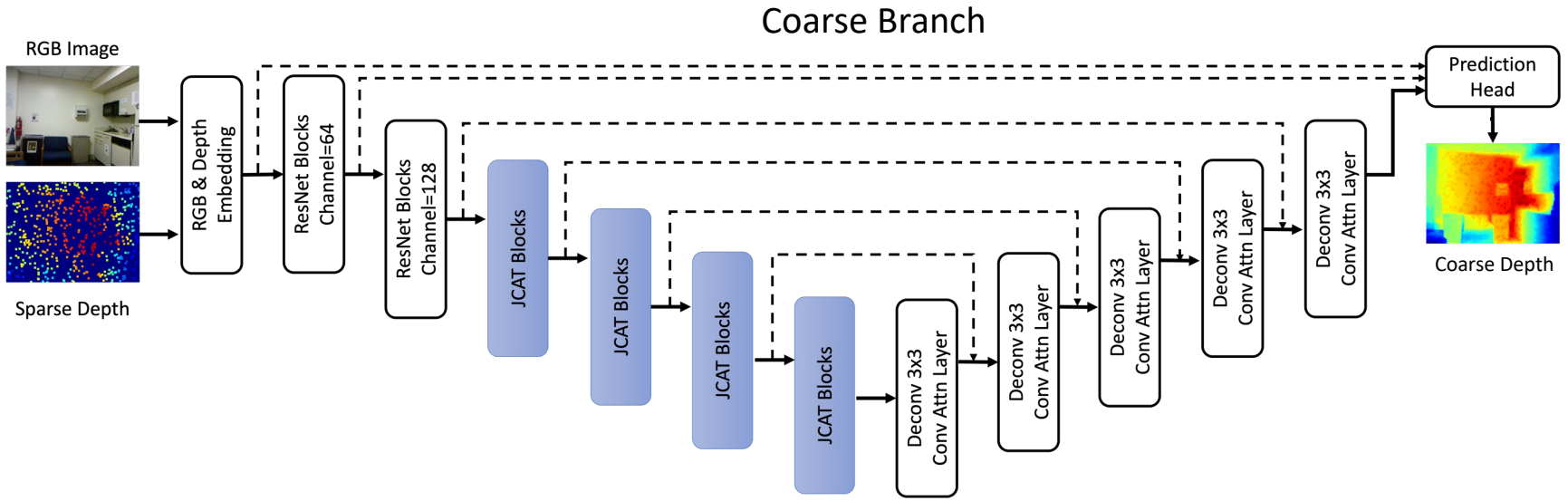
Step 1. Coarse-branch

Step 2. Fine-branch

Step 3. Depth fuse & Refine

Proposed Methods (II)

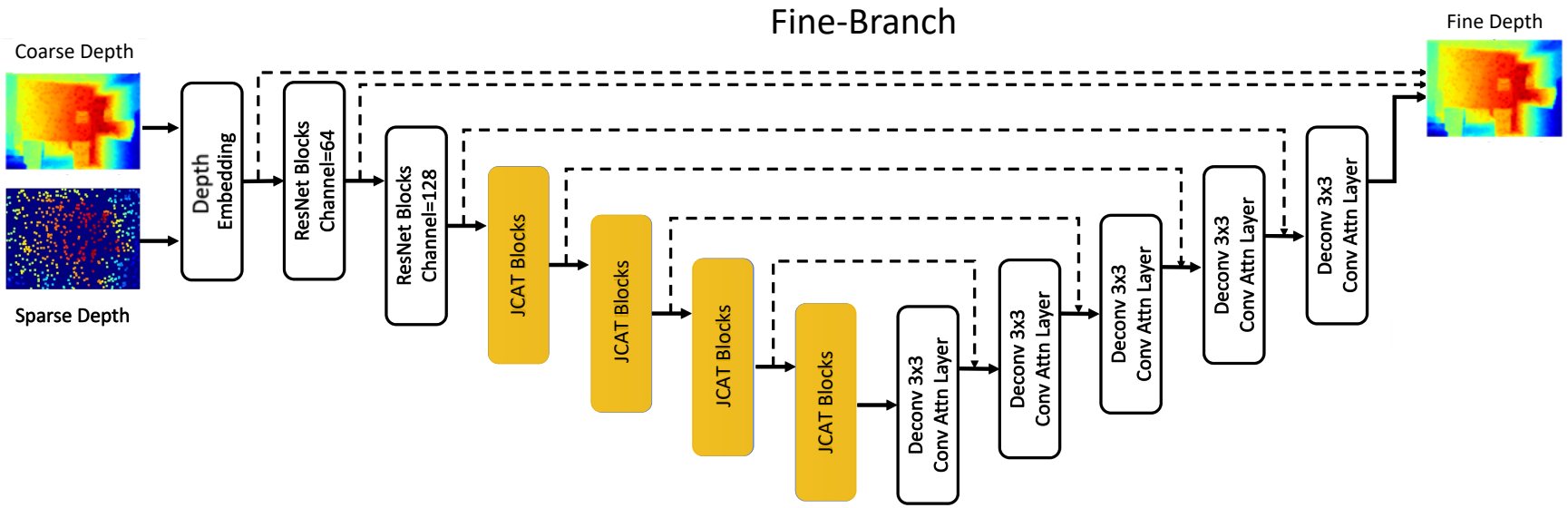
❑ Step 1. Coarse-Branch



- 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

Proposed Methods (III)

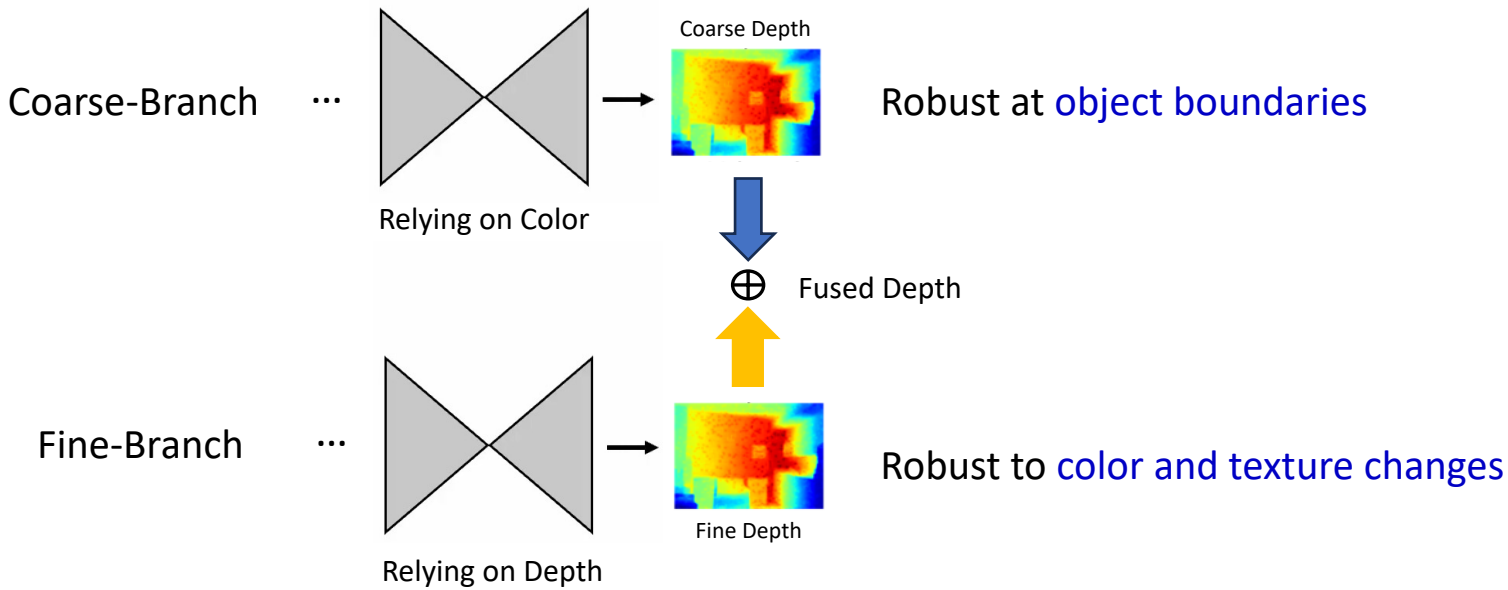
❑ Step 2. Fine-Branch



- 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

Proposed Methods (III)

❑ Step 2. Fine-Branch

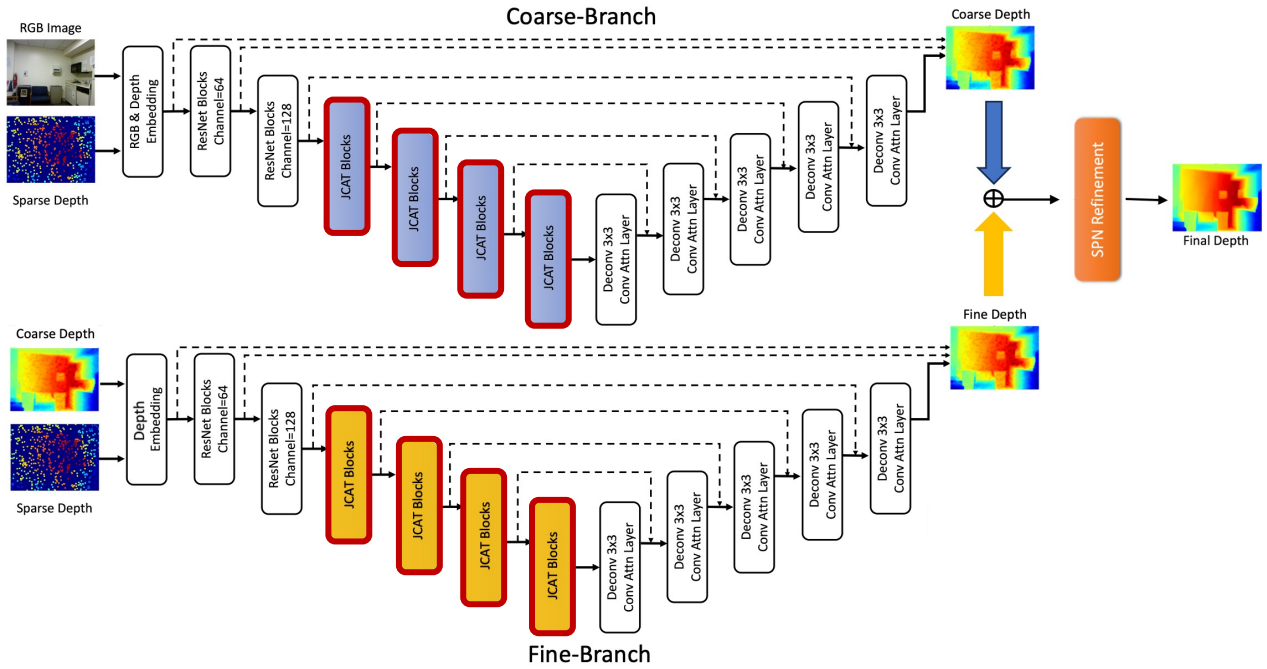


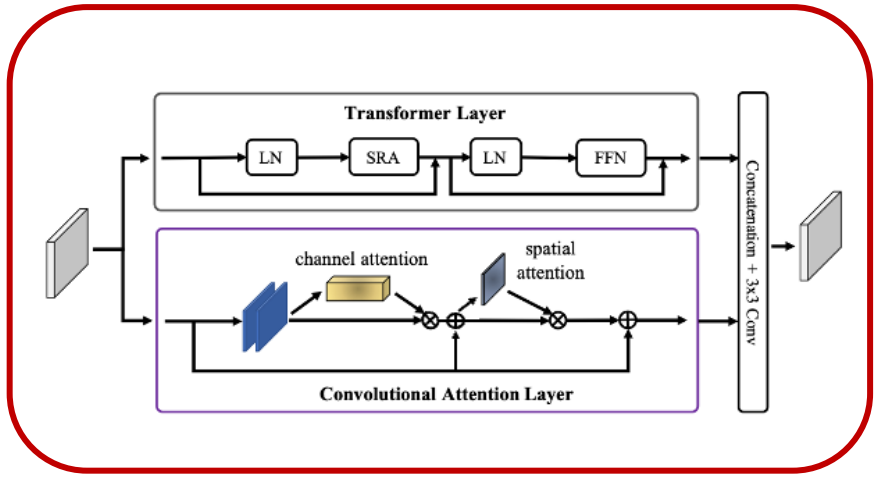
- Fine-Branch also predicts a dense depth map but depending more on depth information
- Depth prediction result obtained from the Coarse-Branch is input to Fine-Branch, and we constructed same encoder-decoder as previous branch. The depth maps predicted from two branches are adaptively fused
- Our methods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

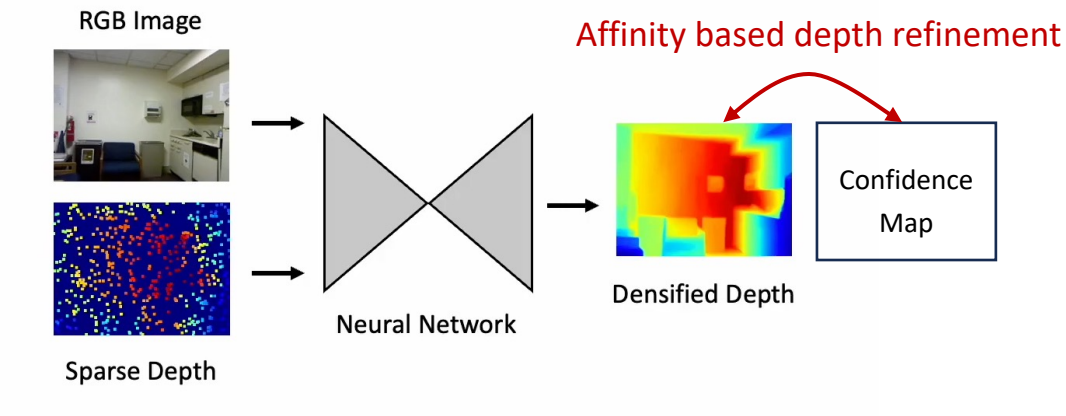




Residual Connection in JCAT Block

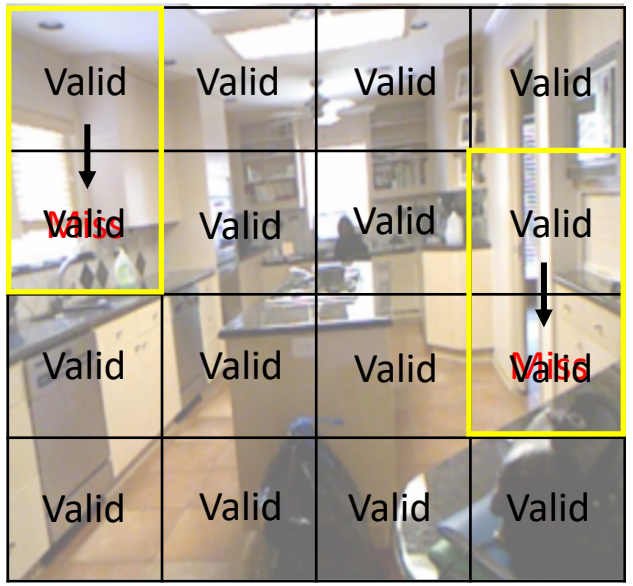
Proposed Methods (V)

❑ 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

Example of SPN process



- Refining missing values in the pixel to obtain the accurate final depth

Loss Functions

Training Loss : L2

$$L(\hat{D}) = \left\| (\hat{D} - D_{gt}) \odot \mathbf{1}(D_{gt} > 0) \right\|^2$$

Predicted depth map — $\hat{D}$

GT for Supervision — D_{gt}

Indicator (Consider only valid piexel) — $\mathbf{1}(D_{gt} > 0)$

Depth prediction Loss

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

Empirical setting hyperparameter — λ_{cb}, λ_{fb}

Experiments

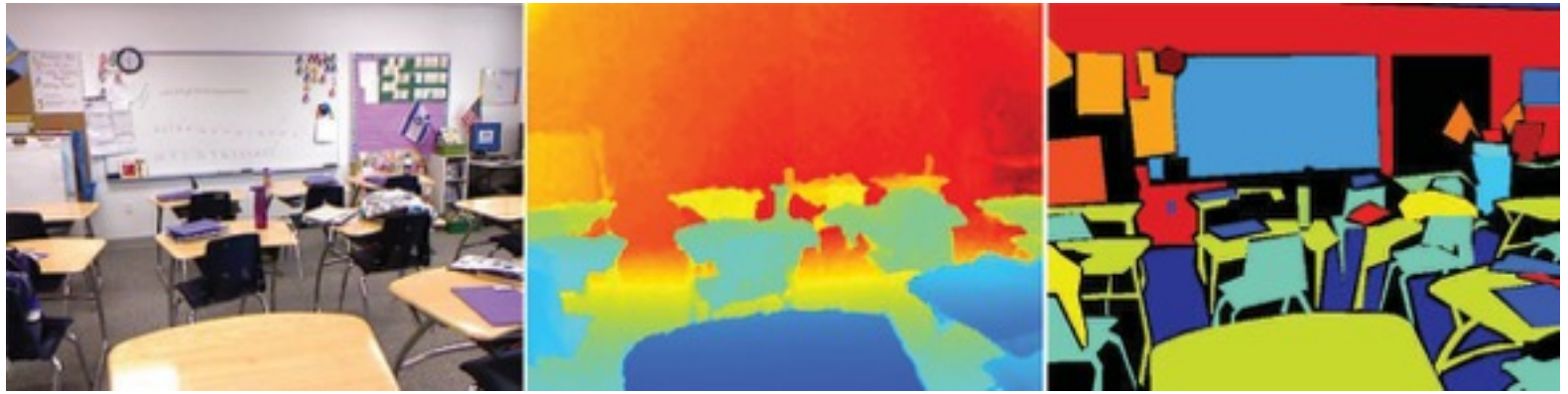
Experimental Environments

❑ Dataset for training and testing

- Open dataset : *NYUv2*
  - 20k indoor color images and sparse depth map

❑ Computing unit

- GPU : NVIDIA RTX A6000 (x2)

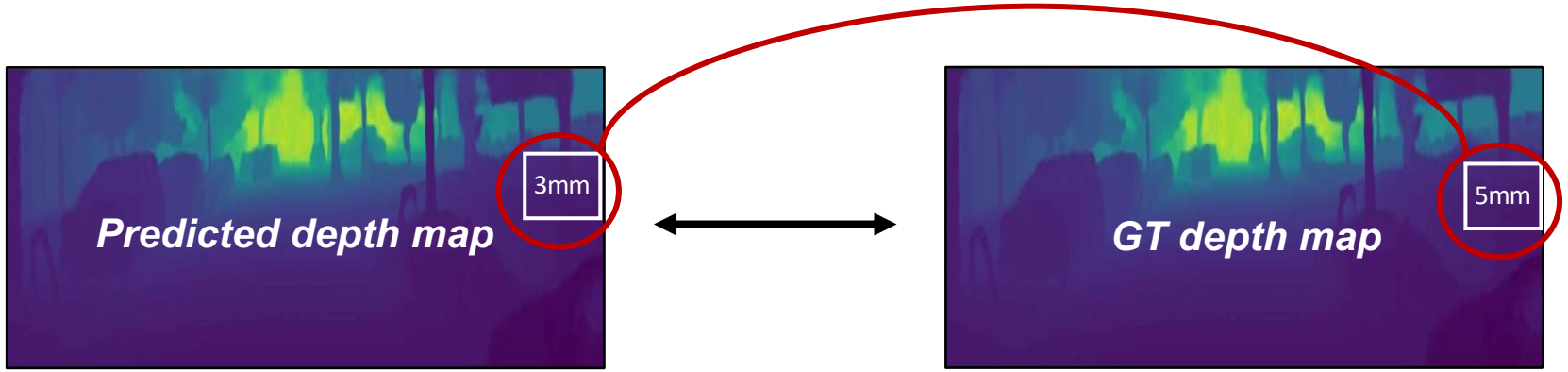


Evaluation Metrics

❑ Evaluation Metric for depth completion

- Root Mean Squared Error (RMSE)
  - Depth difference between predicted depth map and Ground Truth (GT) depth map

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(predicted_i - Ground\ Truth_i)}{N}}$$



Comparison with State-of-the-art Methods

❑ Benchmark on NYUv2 datasets

Method	RMSE (m)	REL (m)
CSPN++	0.117	0.016
DeepLiDAR	0.115	0.022
TWISE	0.097	0.013
NLSPN	0.092	0.012
RigNet	0.090	0.012
CompletionFormer	0.090	0.012
DYSPN	0.090	0.012
BEV@DC	**0.089**	0.012
Ours	**0.089**	**0.011**

	RMSE (mm)	REL (mm)
CompletionFormer	907.1	121.3
Ours	**890.1**	**115.9**

- Our proposed model outperforms the baseline methods on NYUv2 datasets
- It also performs as well as or better than State-of-the-arts models

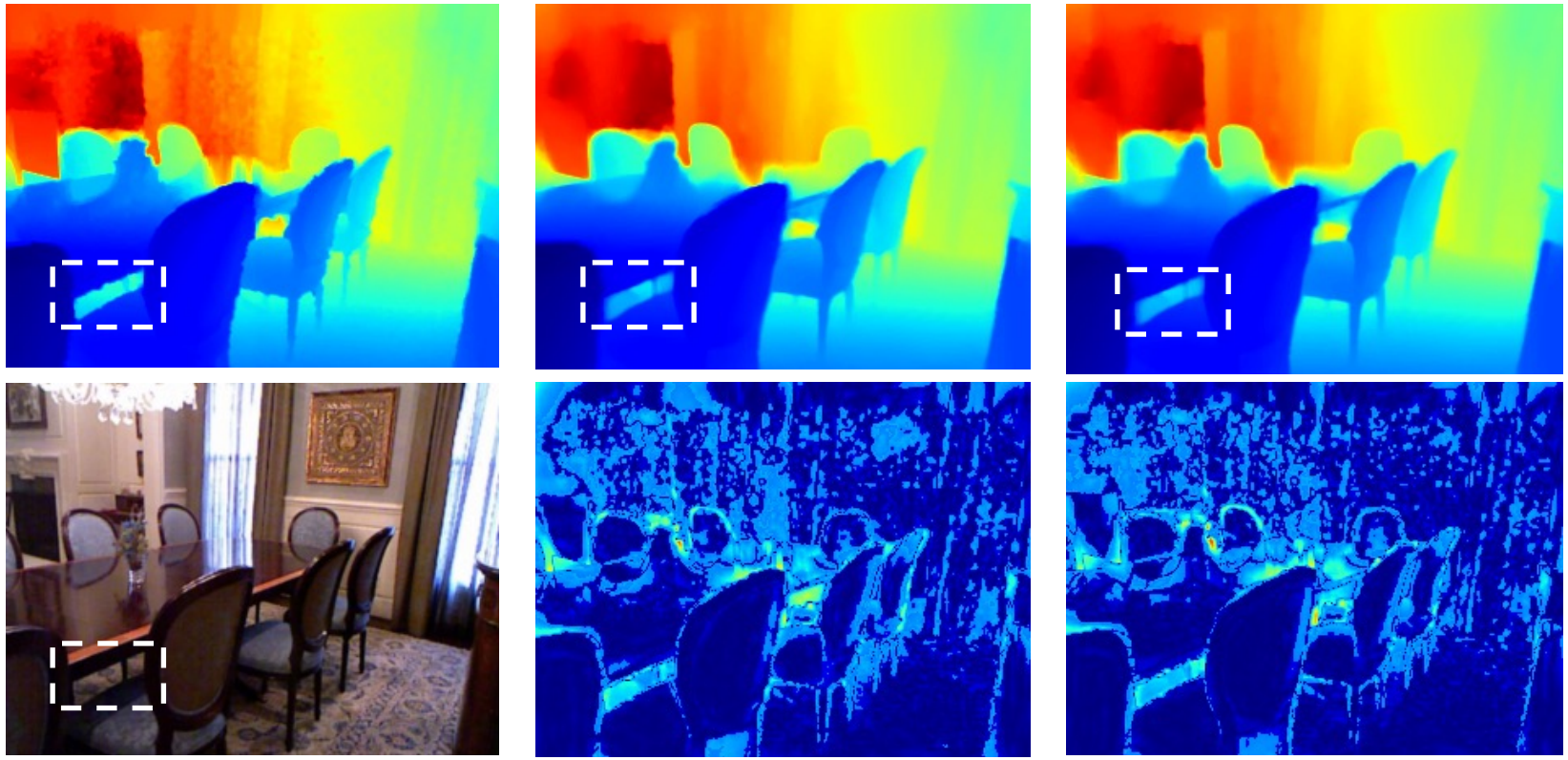
Qualitative Comparison with base Methods

❑ Visualization on NYUv2 Test Set

(a) GT Depth / RGB

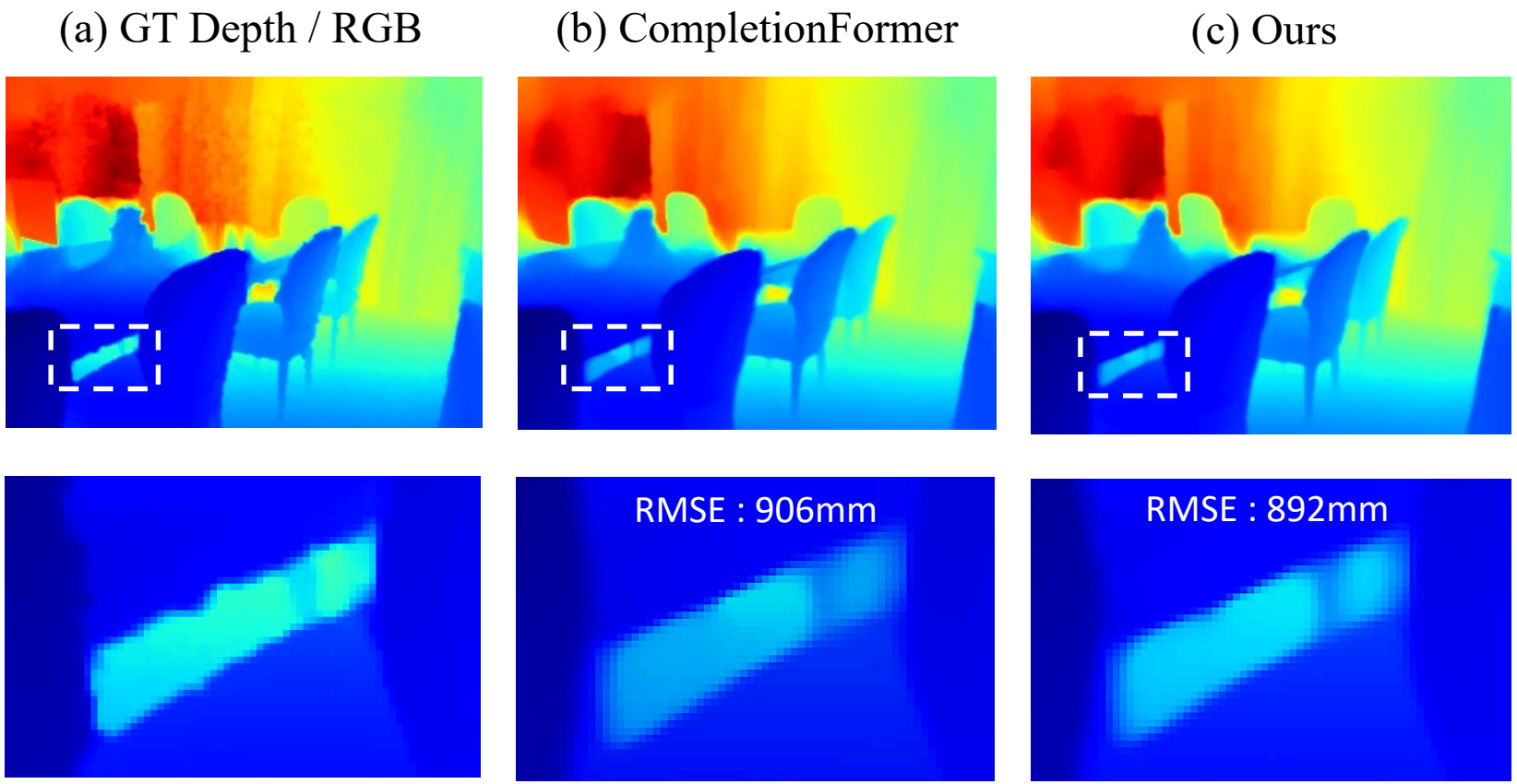
(b) CompletionFormer

(c) Ours



Qualitative Comparison with base Methods

❑ Visualization on NYUv2 Test Set



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