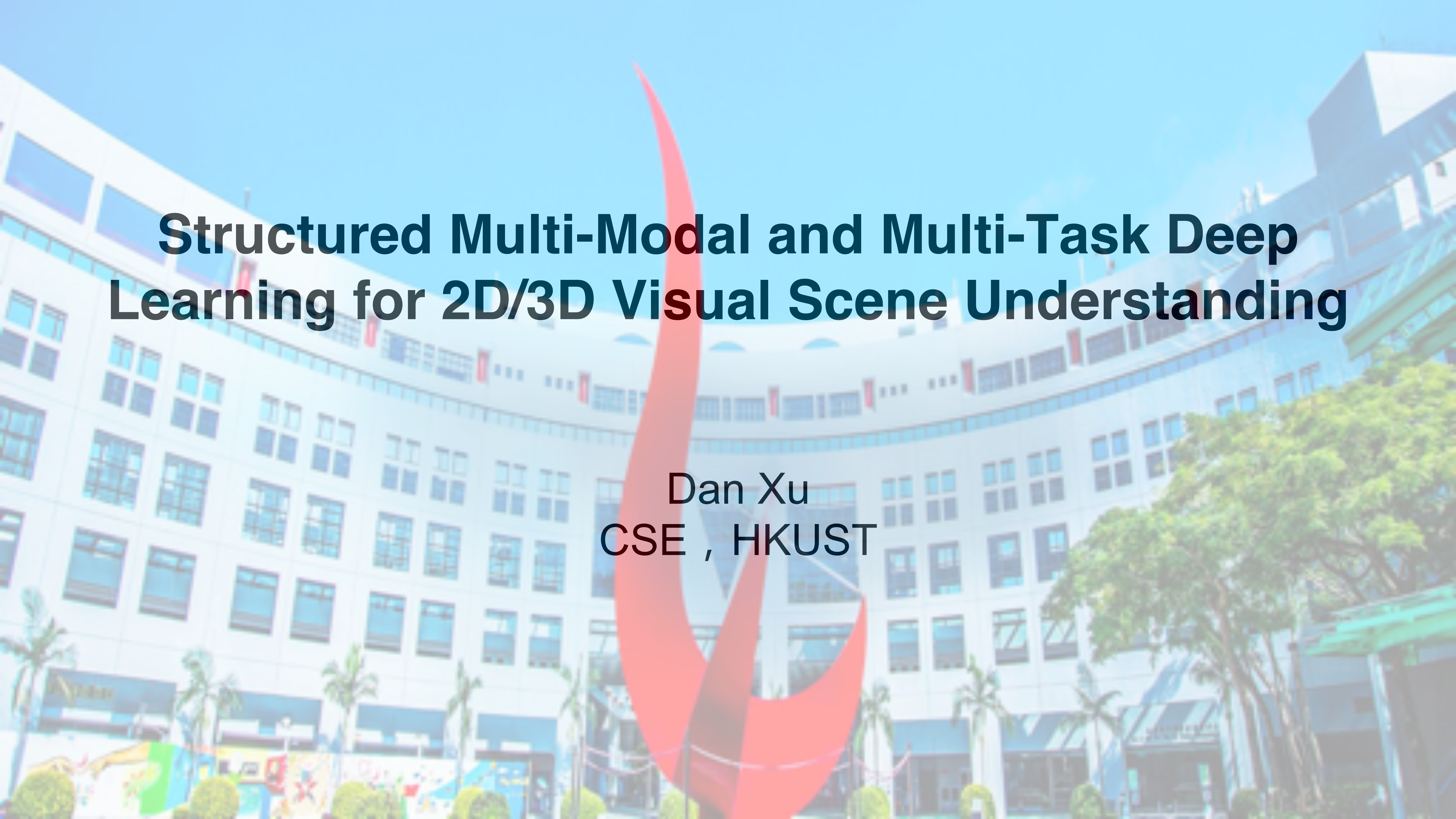
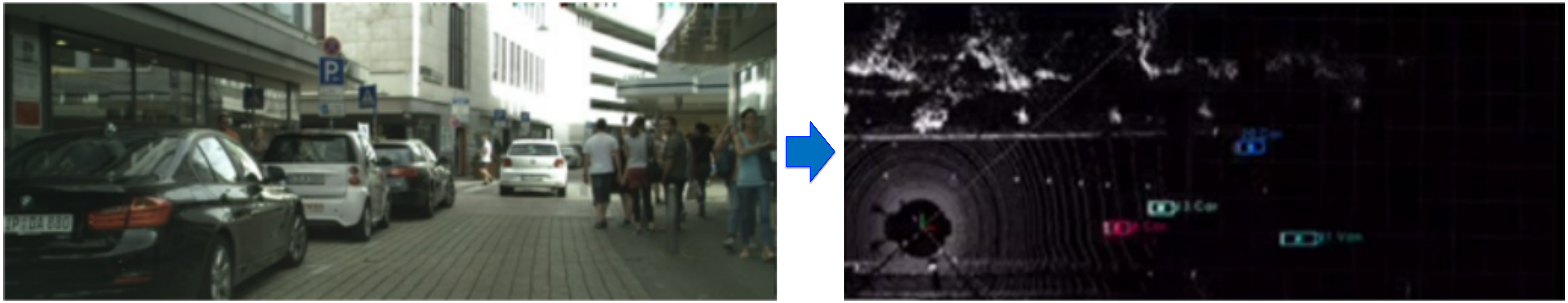


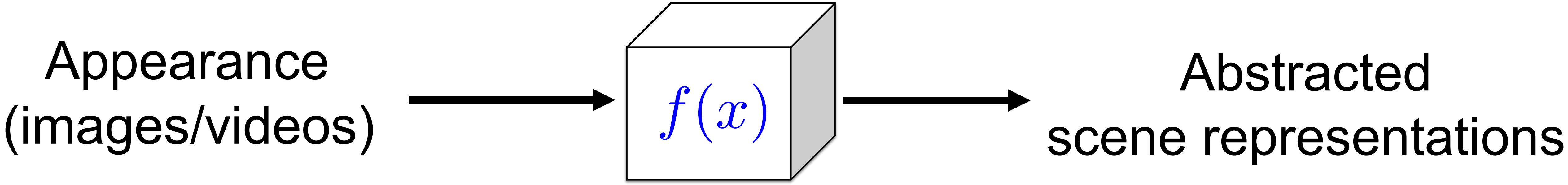
Structured Multi-Modal and Multi-Task Deep Learning for 2D/3D Visual Scene Understanding

Dan Xu
CSE , HKUST

Visual Scene Understanding

- Fundamental research domain in computer vision
- Complex inference: objects, parts, context, interaction and location





Appearance
(images/videos)

$f(x)$

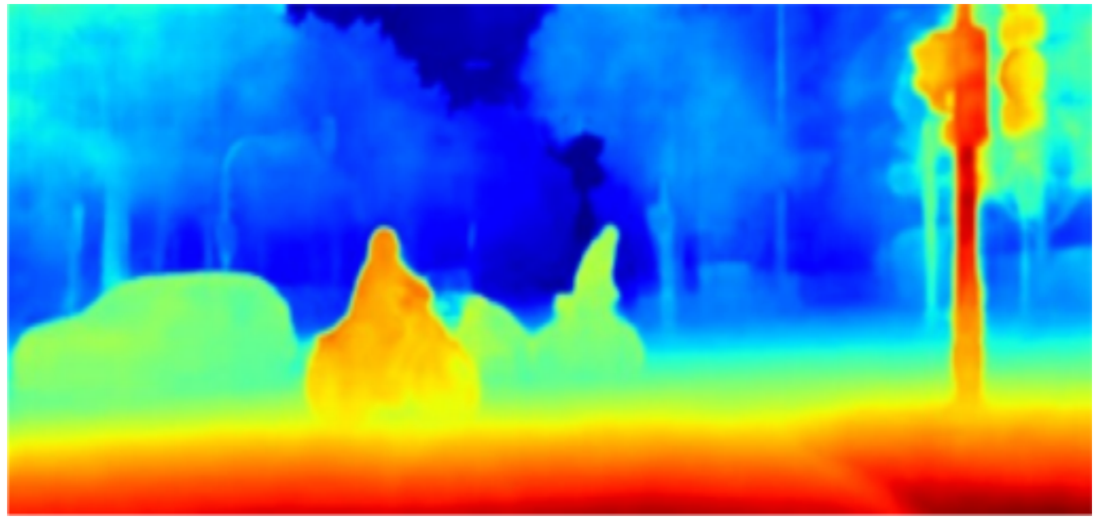
Abstracted
scene representations

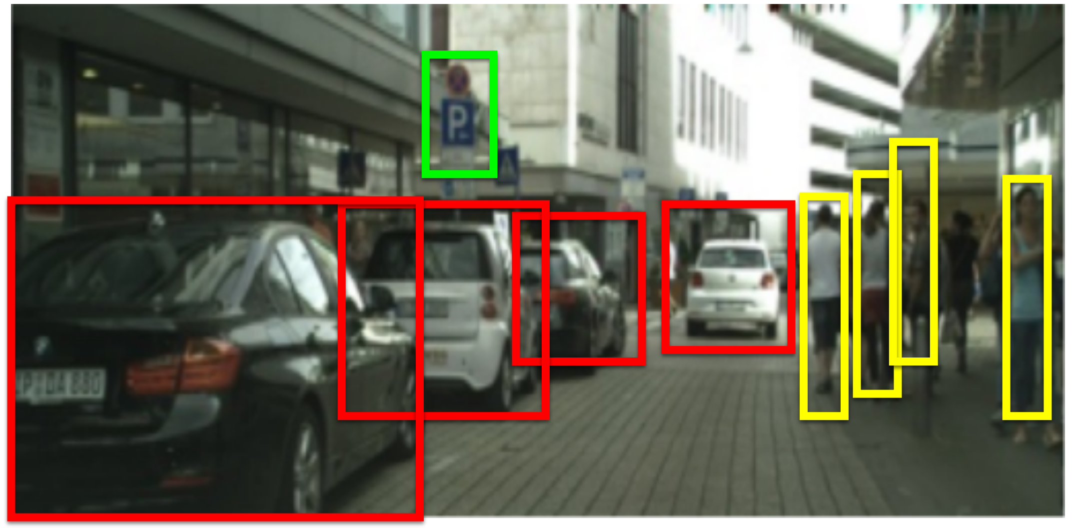
Visual Scene Understanding

- Important tasks: scene parsing, depth estimation, object detection, visual odometry

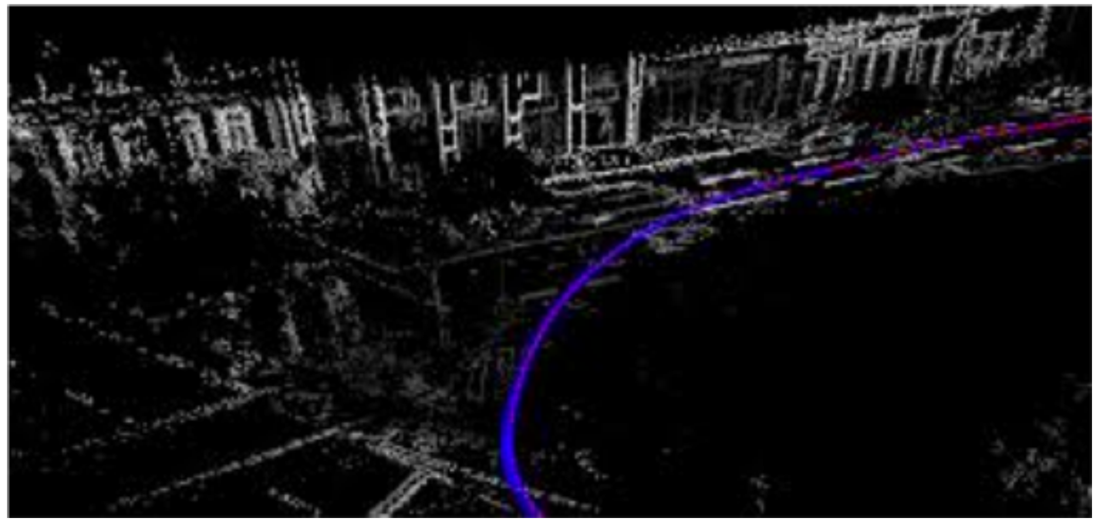

Input RGB


Scene Parsing


Depth Estimation

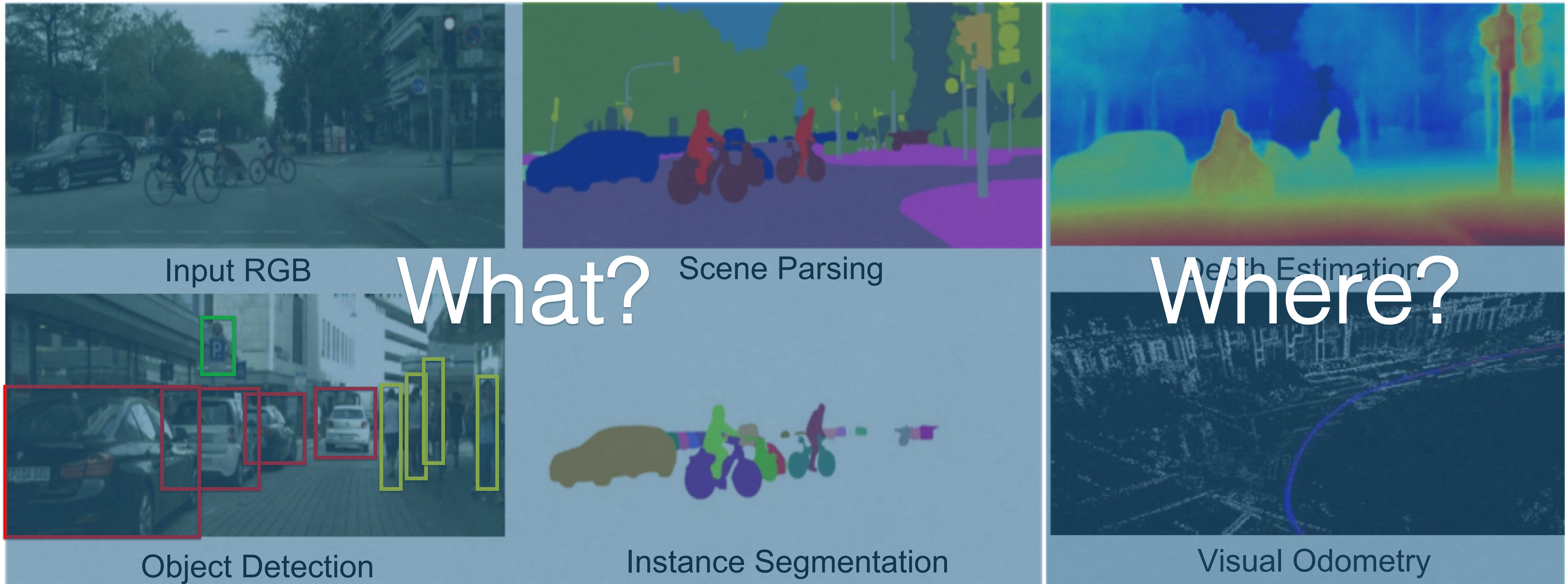

Object Detection


Instance Segmentation


Visual Odometry

Visual Scene Understanding

- Important tasks: scene parsing, depth estimation, object detection, visual odometry



Application

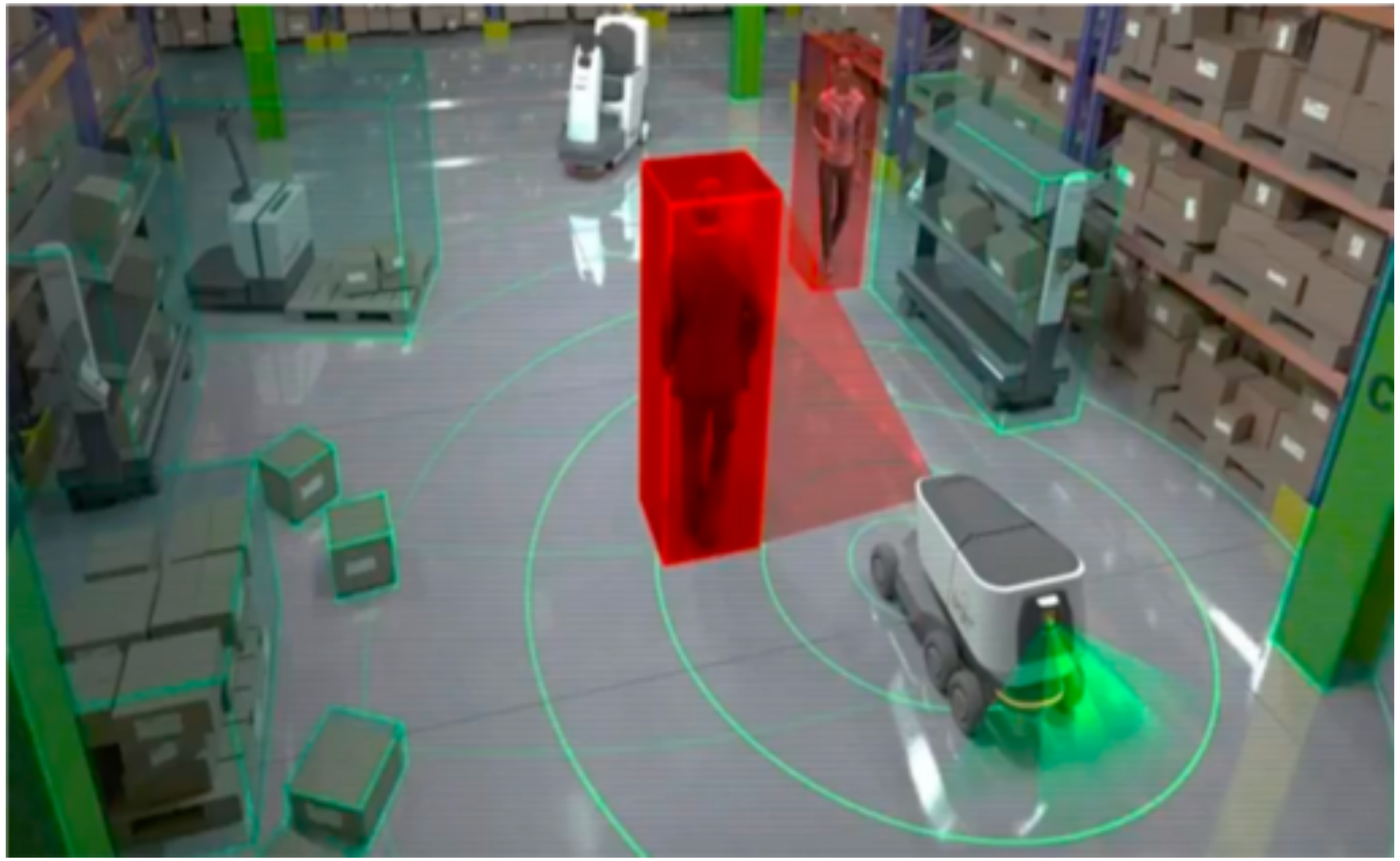
- Self-driving scenarios: automotive driving safety, path planning

Driving safety systems

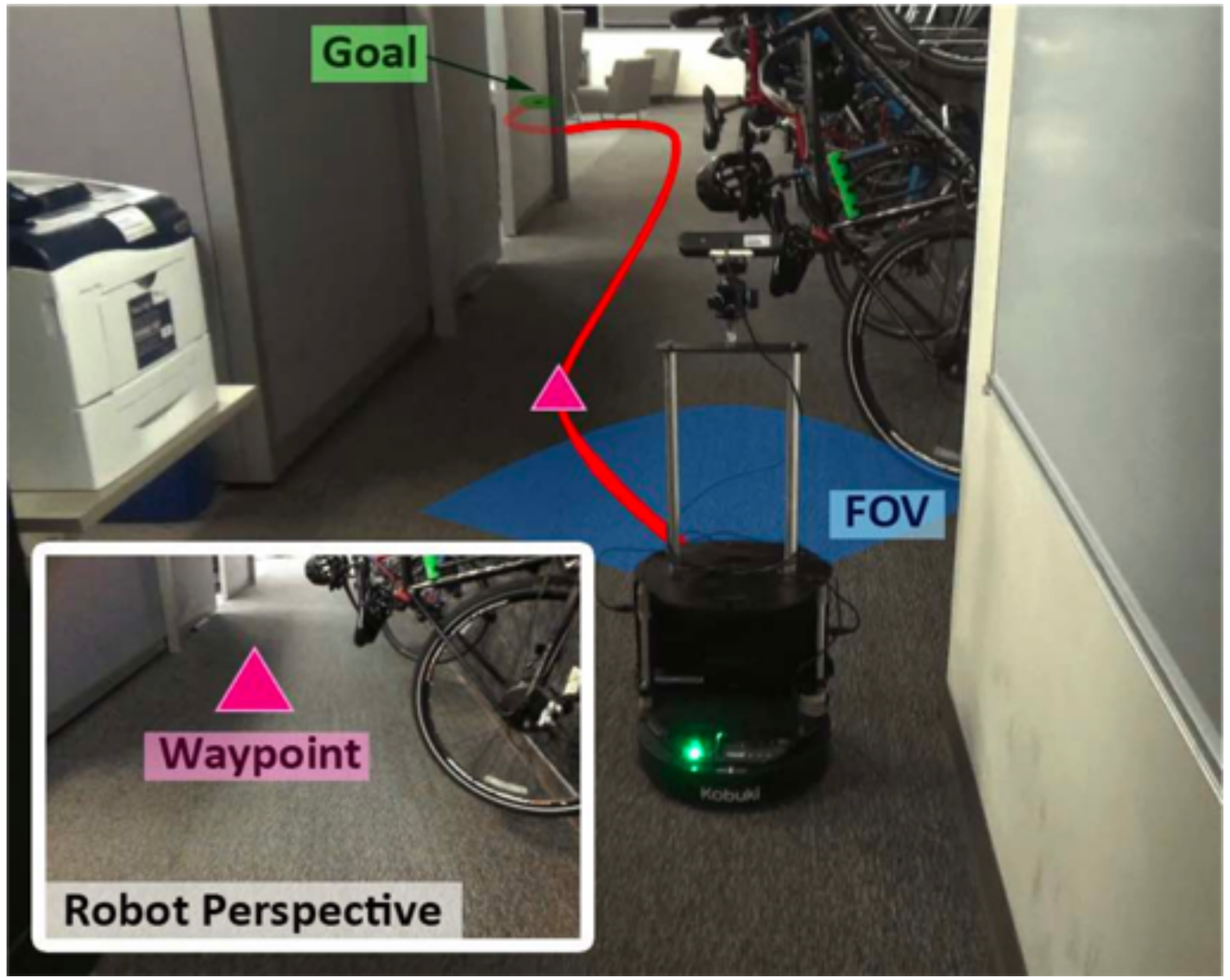
Path planning

Application

- Robotic navigation scenarios: perception and localization



Robot perception



Robot localization

Application

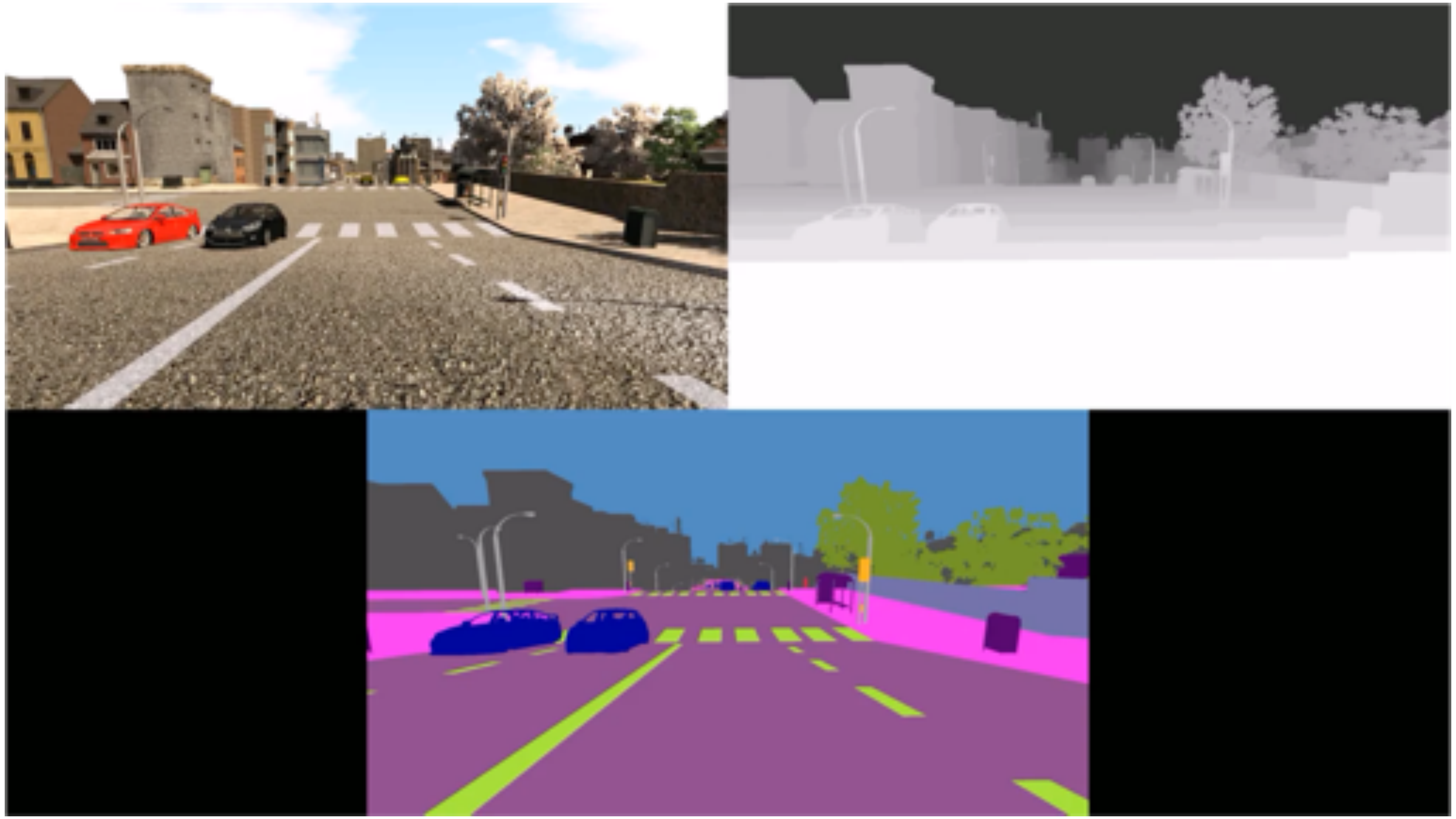
- Public safety and smart cities: transportation monitoring, anomaly detection

Transportation monitoring

Anomaly detection

Research Objectives

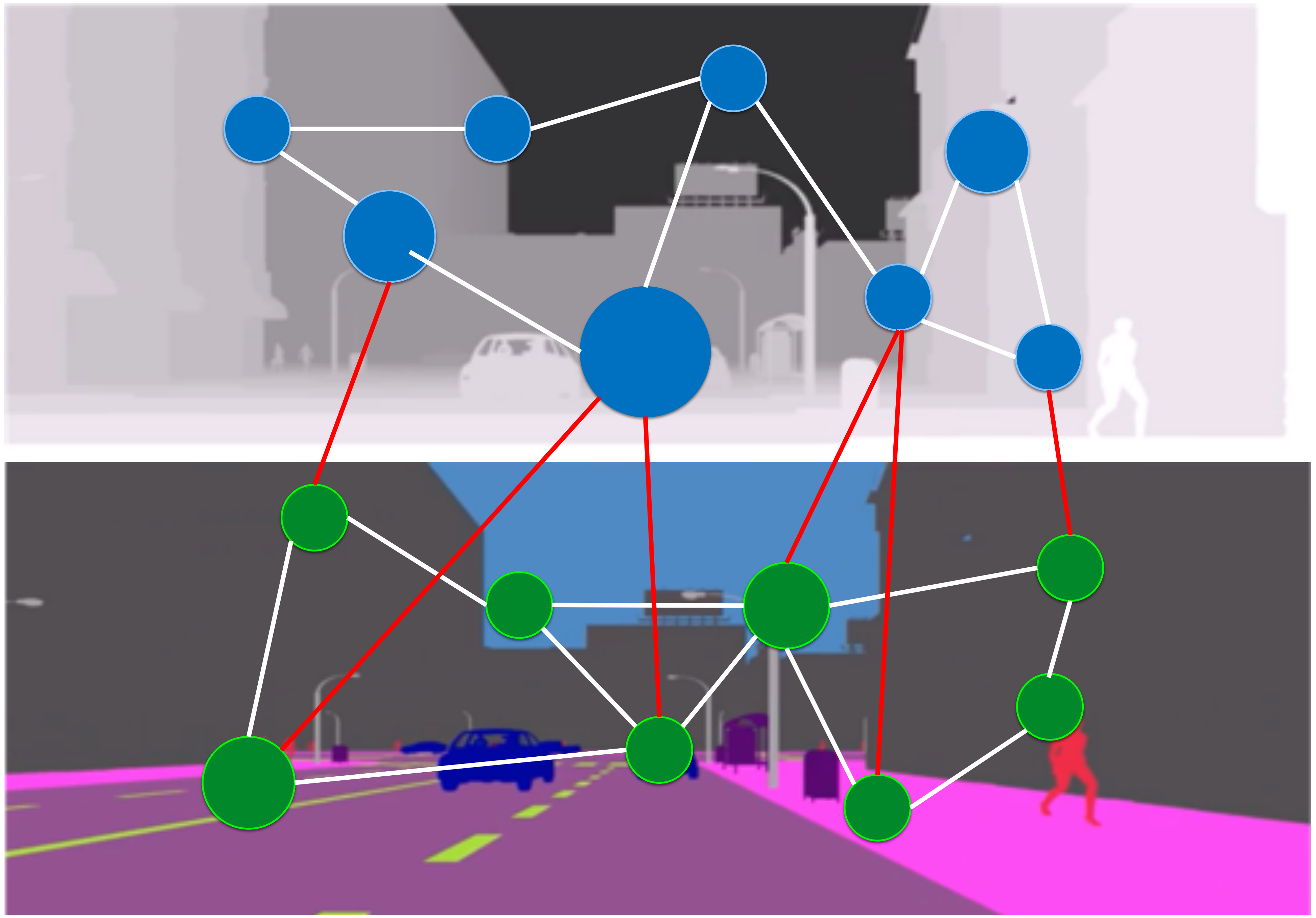
- Effective representations from rich multi-modal and structured data



- Multi-modal data: RGB, depth, thermal, semantics
- Multi-modal deep learning
  - Input is one modality, output is another
  - Multi-modalities are jointly learned
  - One modality assists in the learning of another

Research Objectives

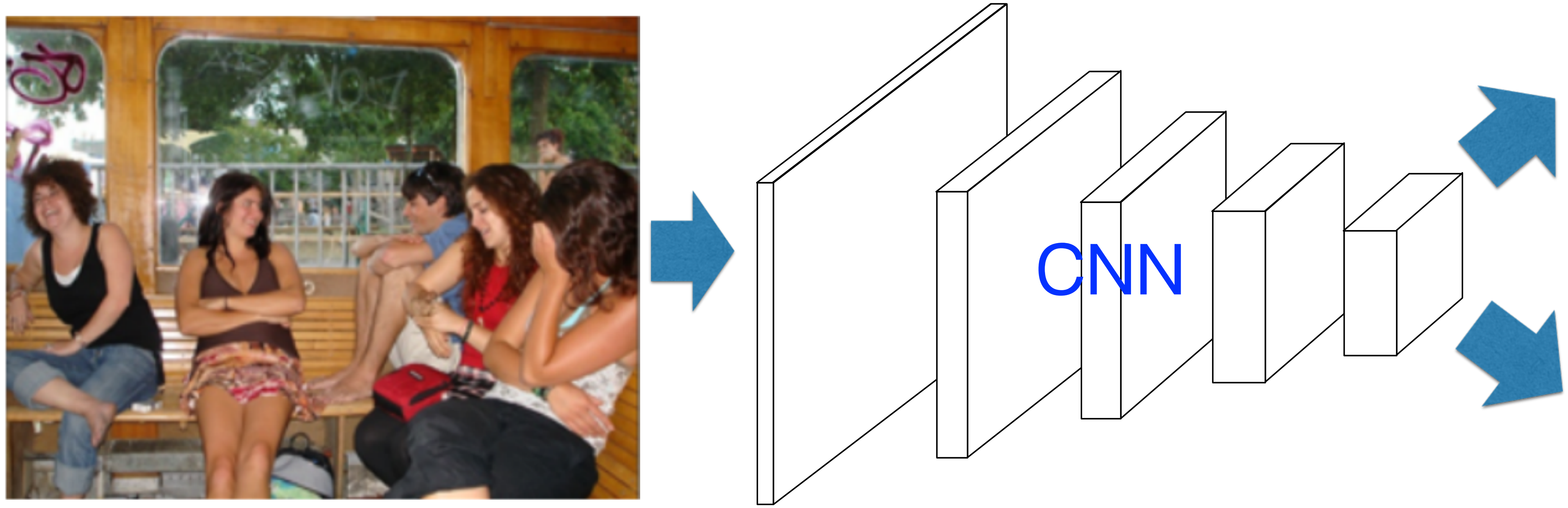
- Effective representations from rich multi-modal and structured data

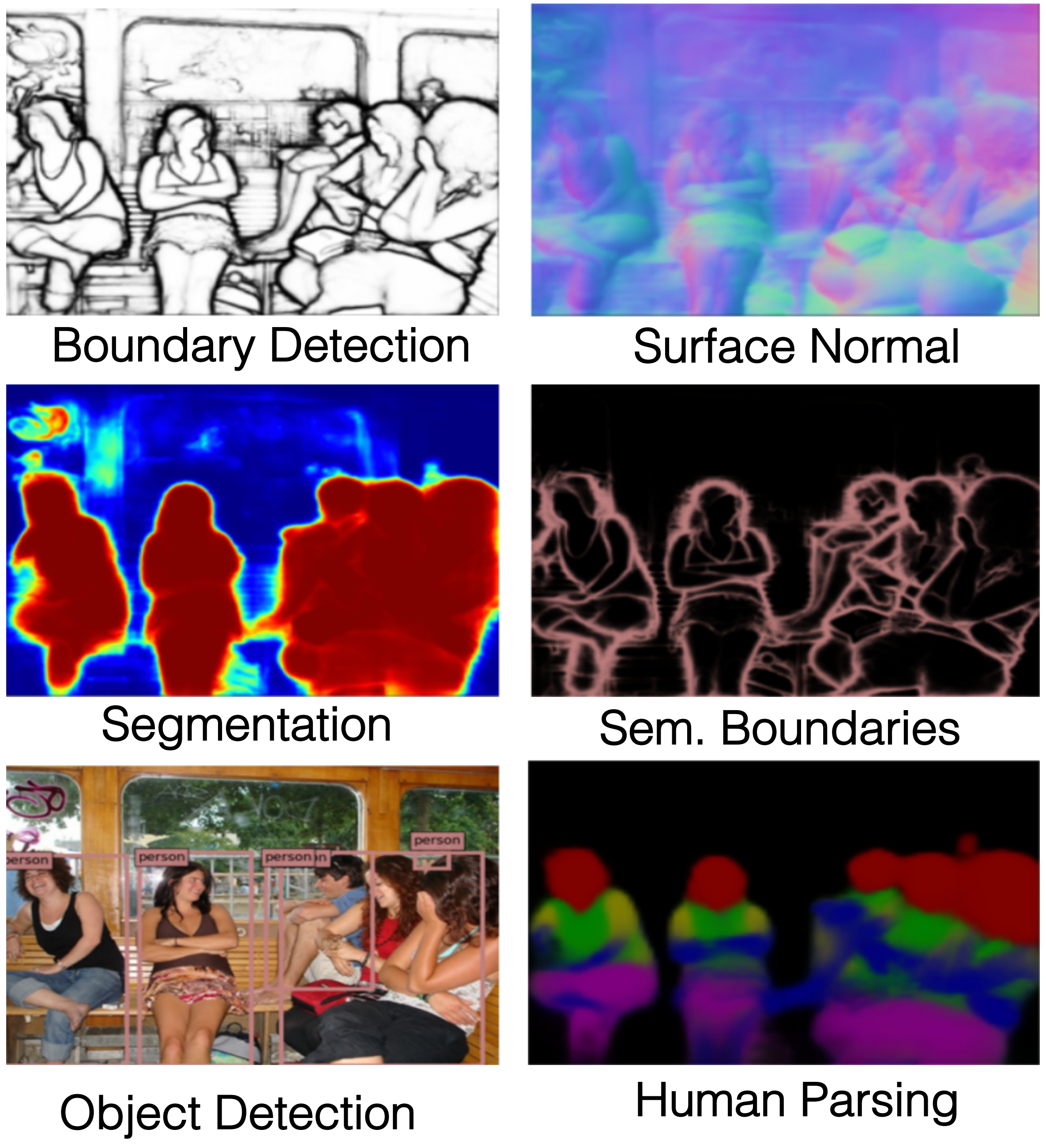


- Highly structured and correlated
- Graph-based modelling and deep network design
- Effective structured representations and predictions

Research Objectives

- Modelling complex task via joint learning of multiple sub-tasks

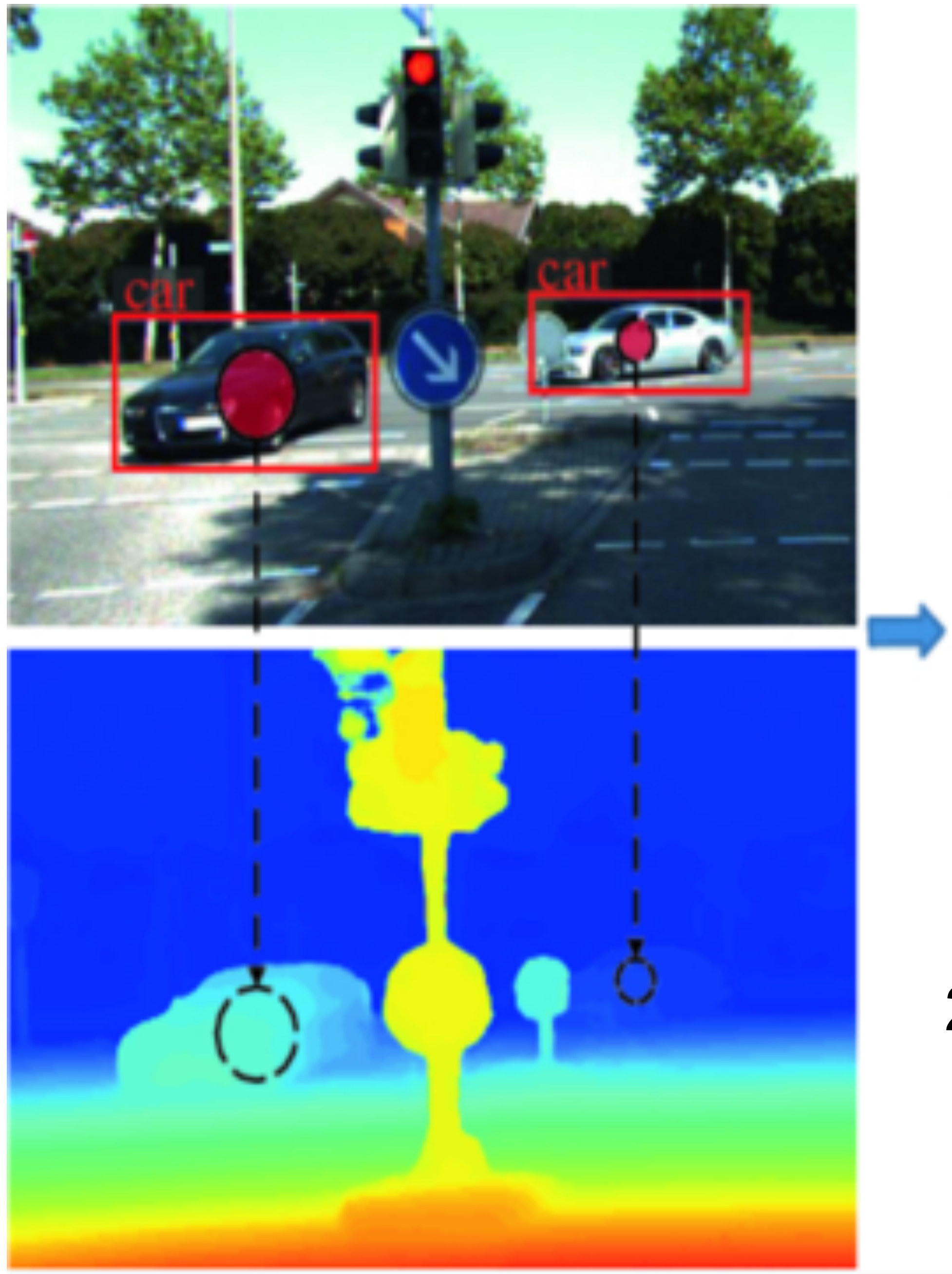




❑ How to design a structure to handle the multiple tasks while using a single effective and efficient deep network?

Research Objectives

- Modelling 2D and 3D for high-level scene understanding



2D location, category and depth

- 2D and 3D data and tasks are beneficial to each other
  - 2D semantics (object categories, appearance and spatial relationships) boost the 3D estimation
  - 3D information (e.g. scene geometry) facilitates the prediction of 2D tasks
- Interaction between 2D and 3D tasks and data in a single deep model

Overview

- Scene depth estimation with structured probabilistic modeling
- A joint multi-modal and multi-task deep learning framework
- Modelling the interaction between 2D and 3D data and tasks
- Hot research & development fields along the direction
- Summary

Overview

- Scene depth estimation with structured probabilistic modeling

- A joint multi-modal and multi-task deep learning framework

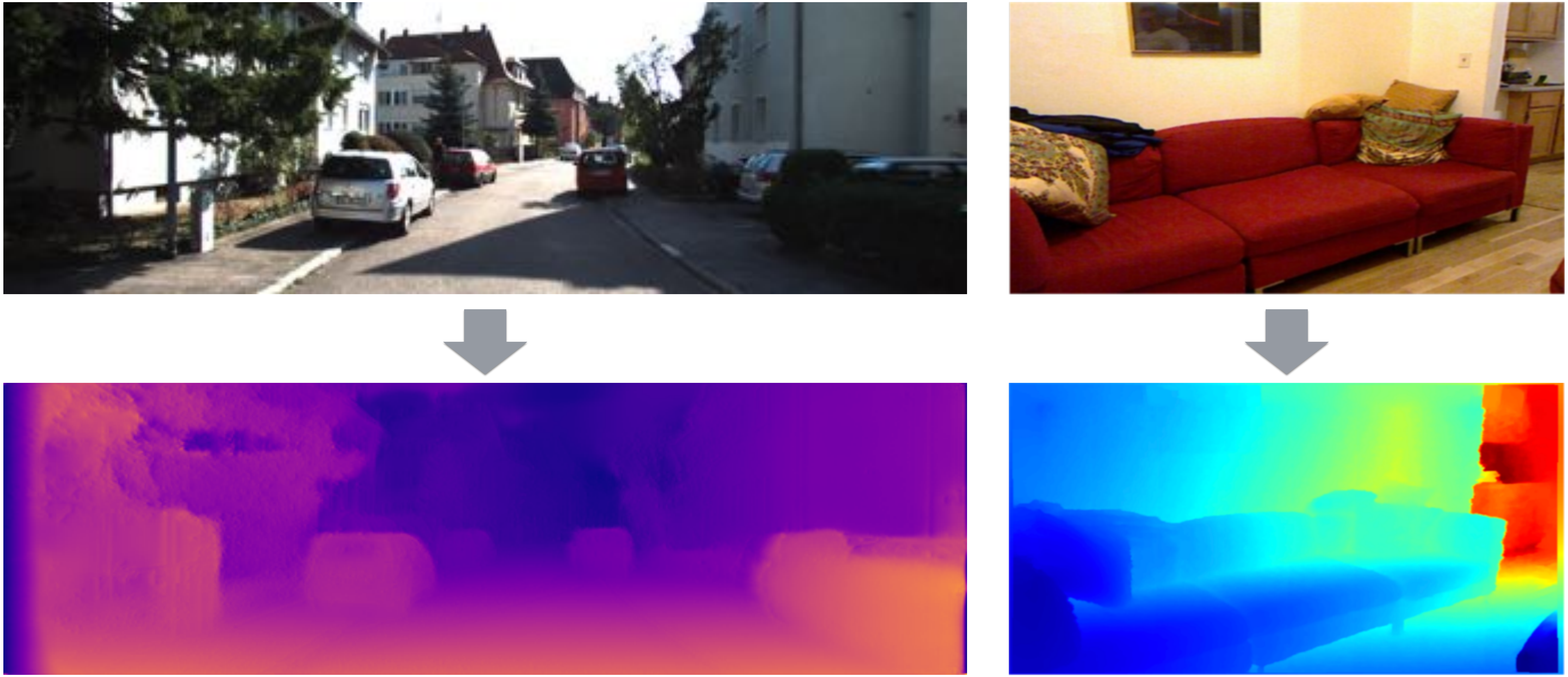
- Modelling the interaction between 2D and 3D data and tasks

- Hot research & development fields along the direction

- Summary

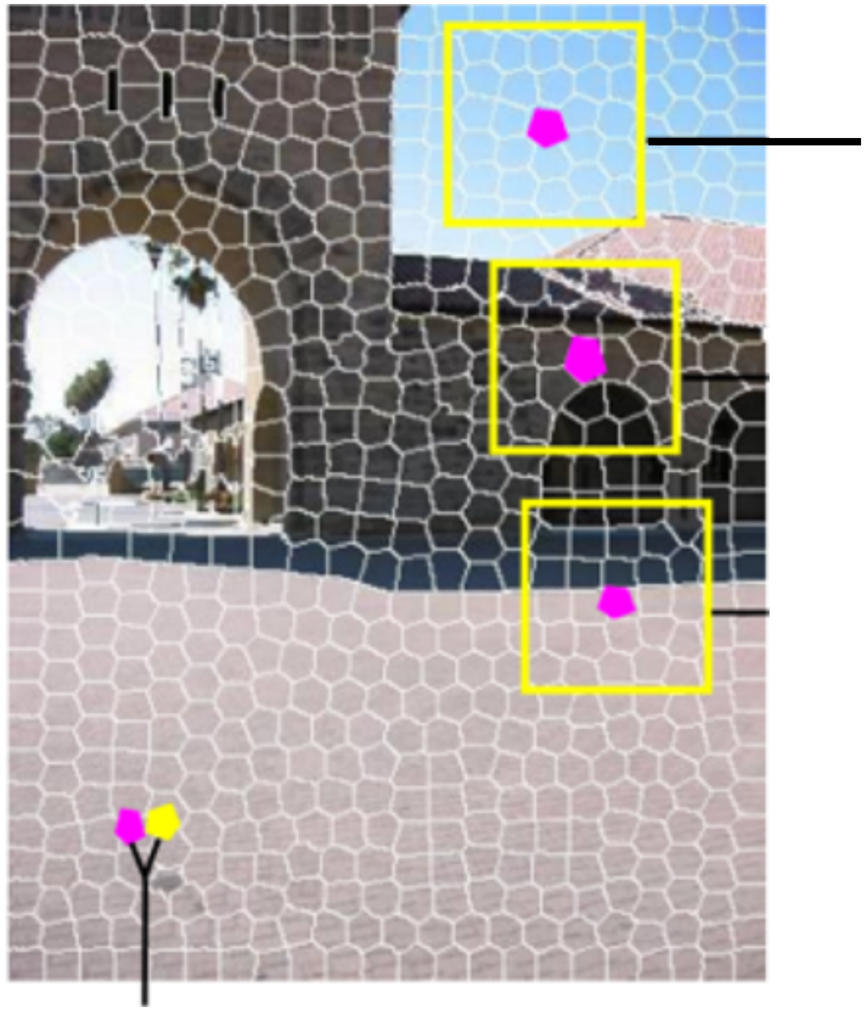
Monocular Depth Estimation

- Regression from RGB → Depth



Structured Modelling on Deep Predictions

- Deep predictions: local kernels, structured information lost

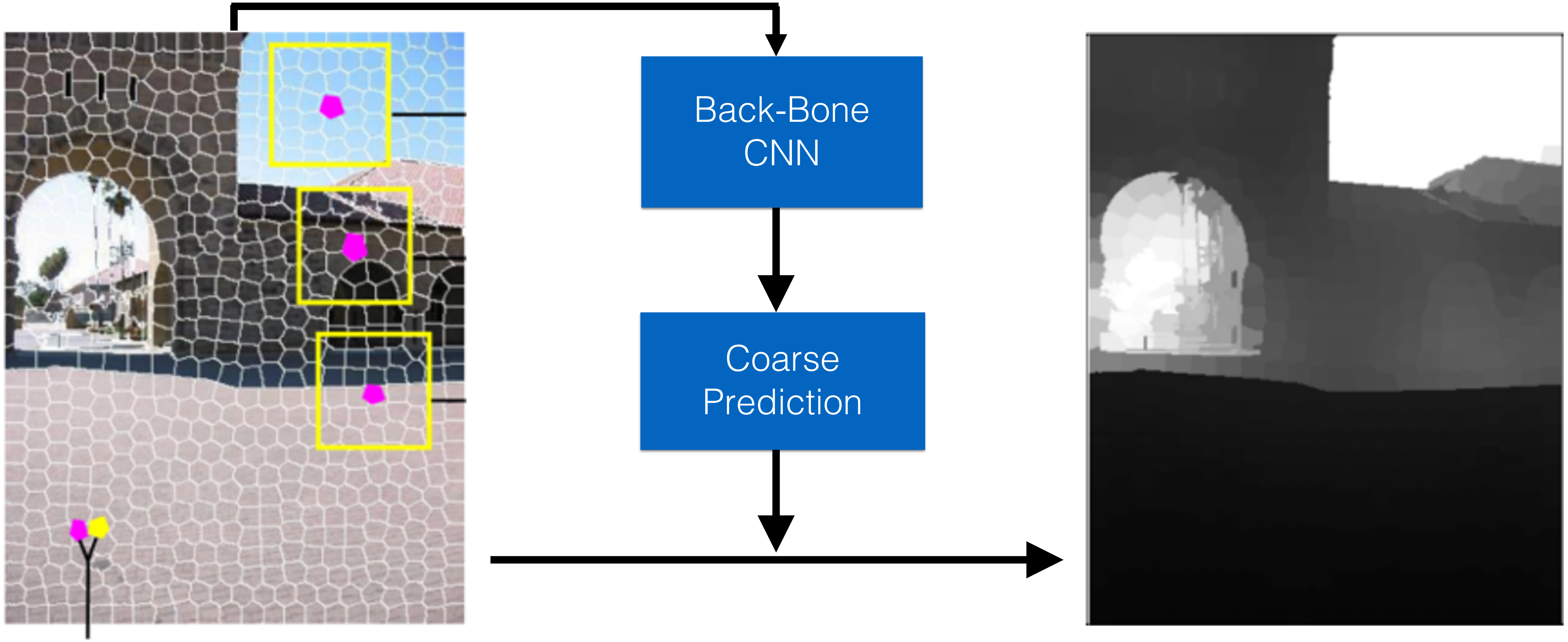


Appearance relationship

Spatial relationship

Structured Modelling on Deep Predictions

- Structured modelling with CRFs for depth regression



Back-Bone CNN

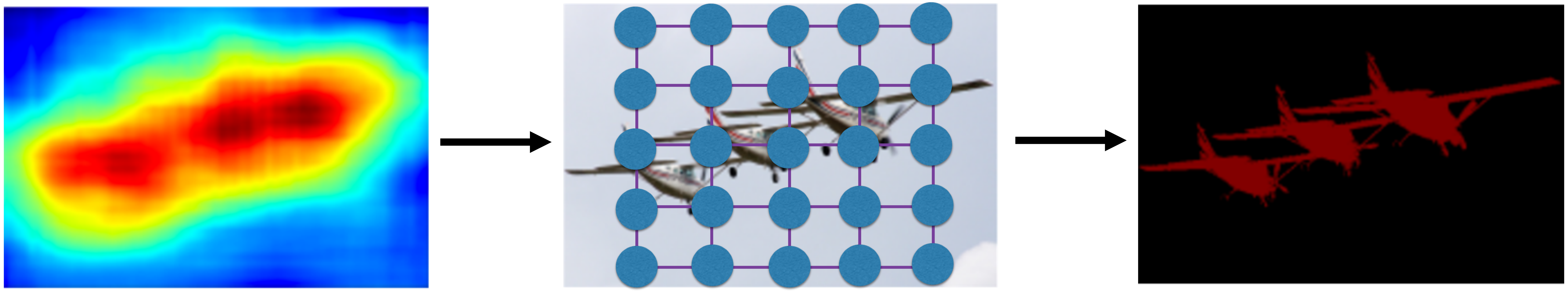
Coarse Prediction

CRF modeling

Refined depth output

Structured Modelling on Deep Predictions

- Deep structured discrete prediction (e.g. semantic segmentation)



CNN coarse output

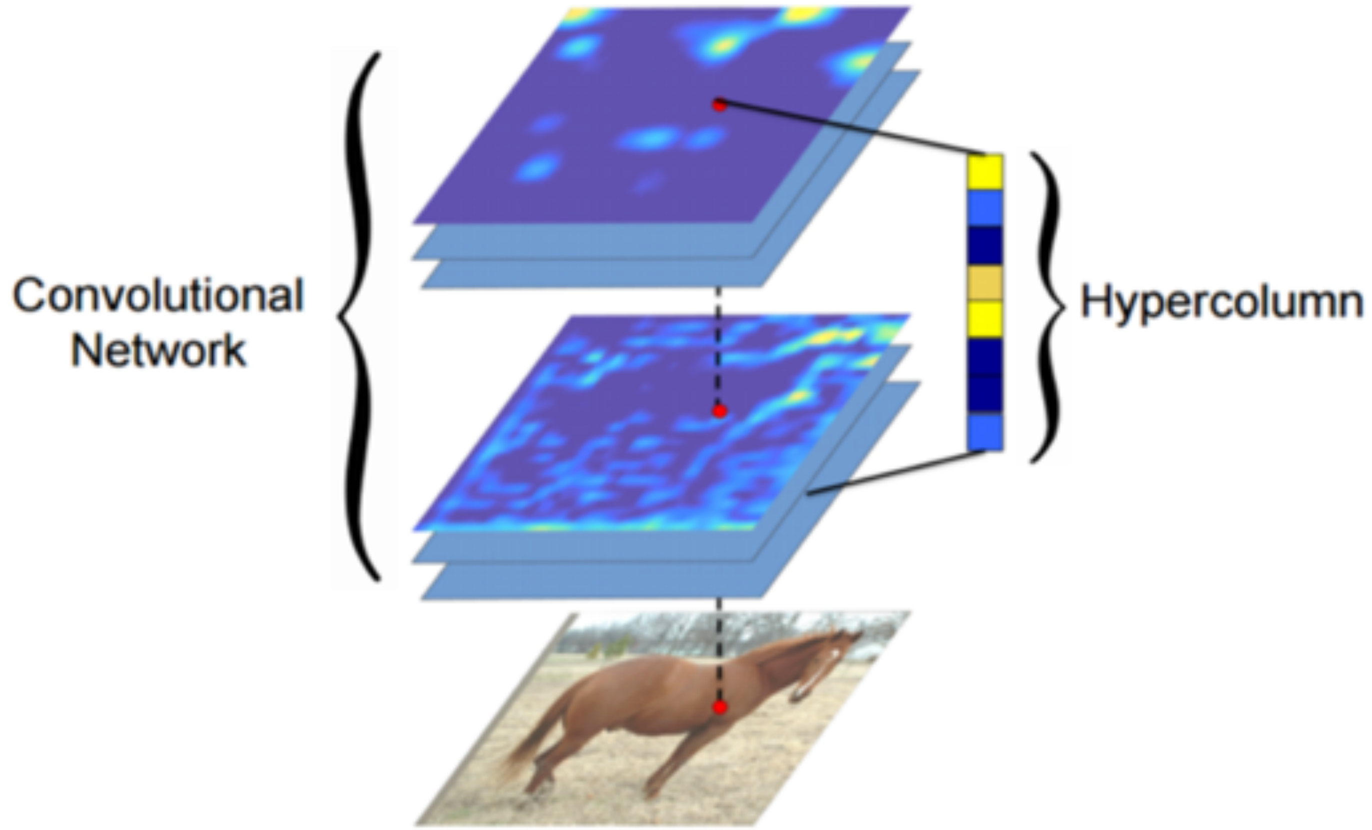
CRF-modeling

Inference

- Representative works:
  - CRF-RNN:
    - Zheng and Torr et al., Conditional random fields as recurrent neural networks. In *ICCV*, 2015.
  - Deep convolutional neural field:
    - Liu and Reid et al., Learning depth from single monocular images using deep convolutional neural fields. *IEEE TPAMI*, 38(10):2024–2039, 2016.
- Applicable in **discrete domain** or in **single scale**

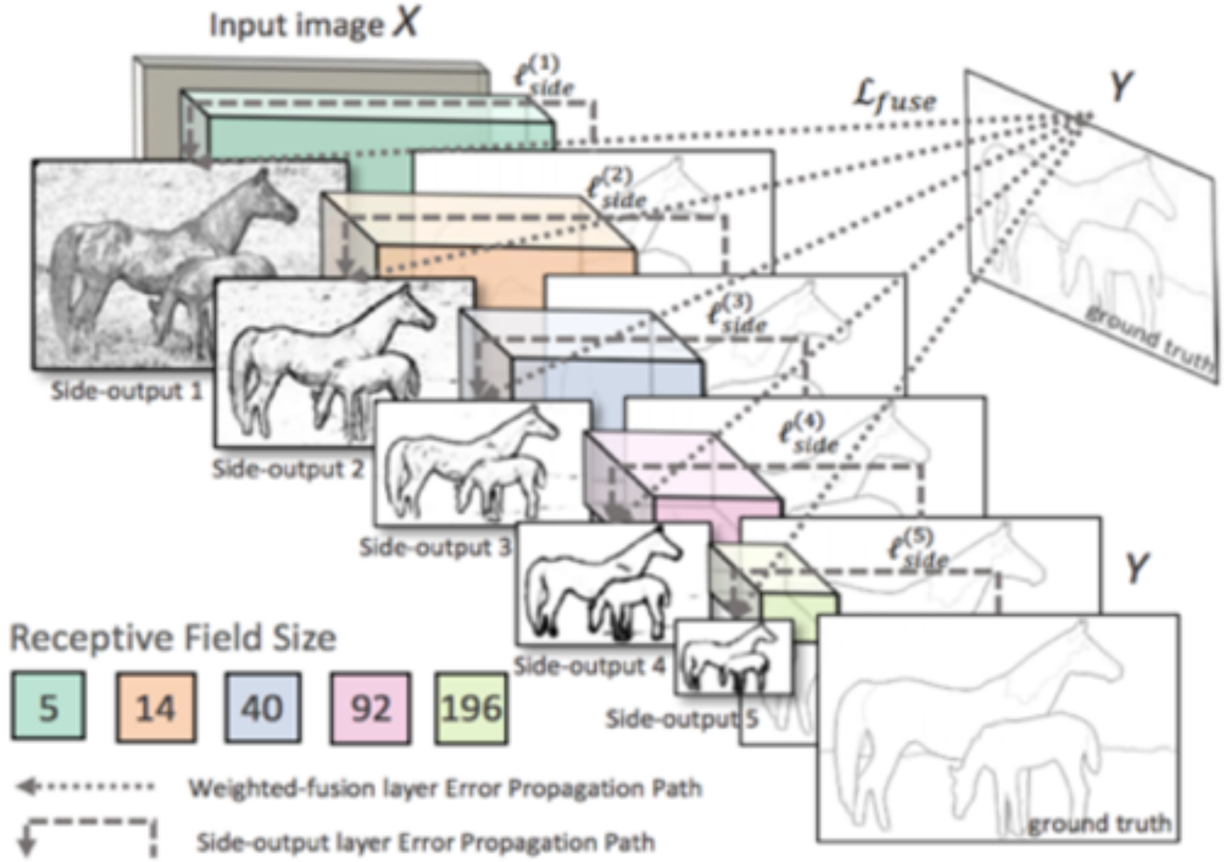
Structured Modelling on Deep Predictions

- Multi-scale information in deep CNN



Hypercolumn

B. Hariharan, P. Arbela ´ez, R. Girshick and J. Malik. Hypercolumns for object segmentation and fine-grained localization. In CVPR, 2015.



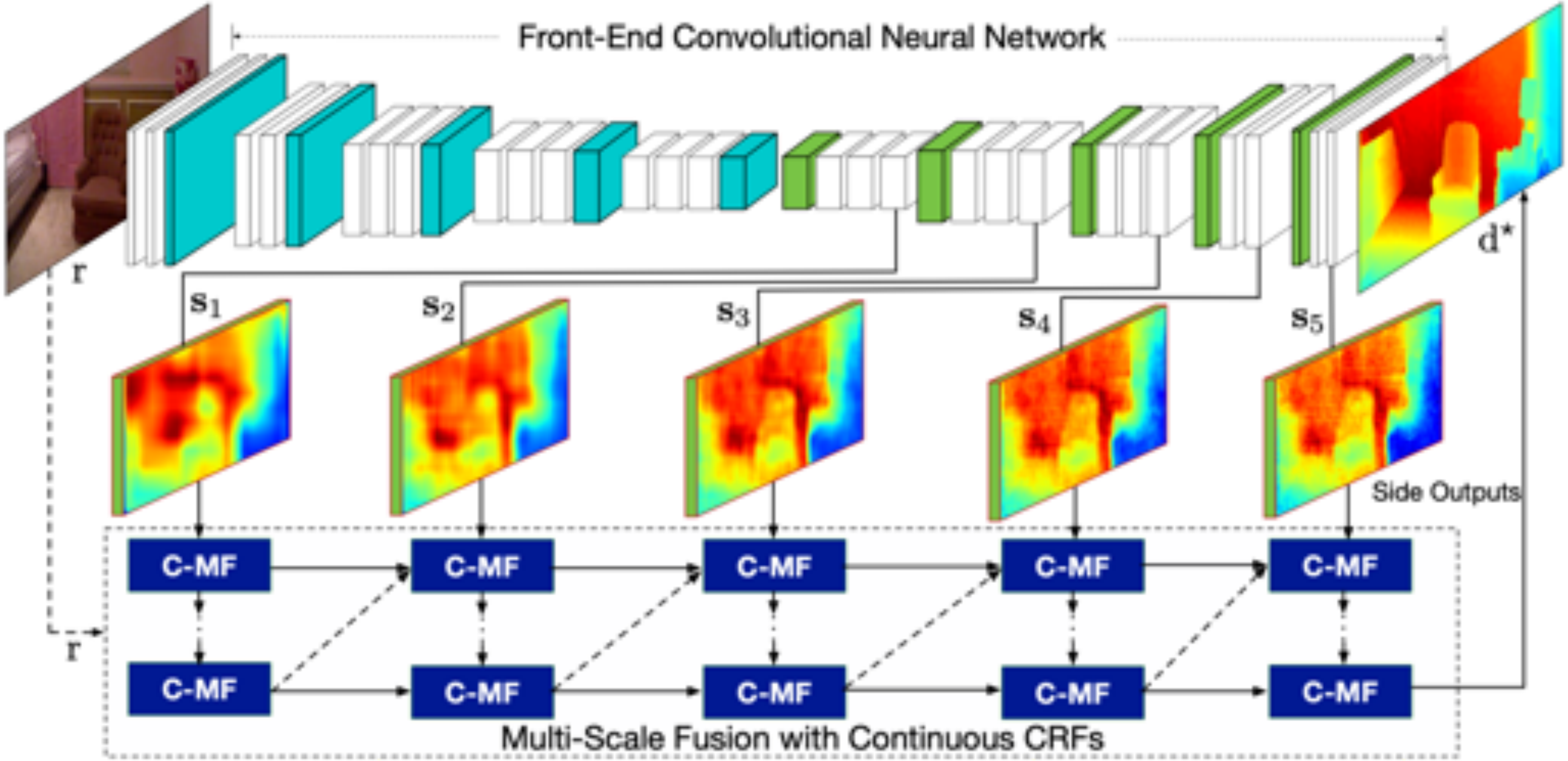
HED

S. Xie and Z. Tu. Holistically-nested edge detection. In *ICCV*, 2015.

- Fusion schemes: **concatenation** or **weighted averaging**

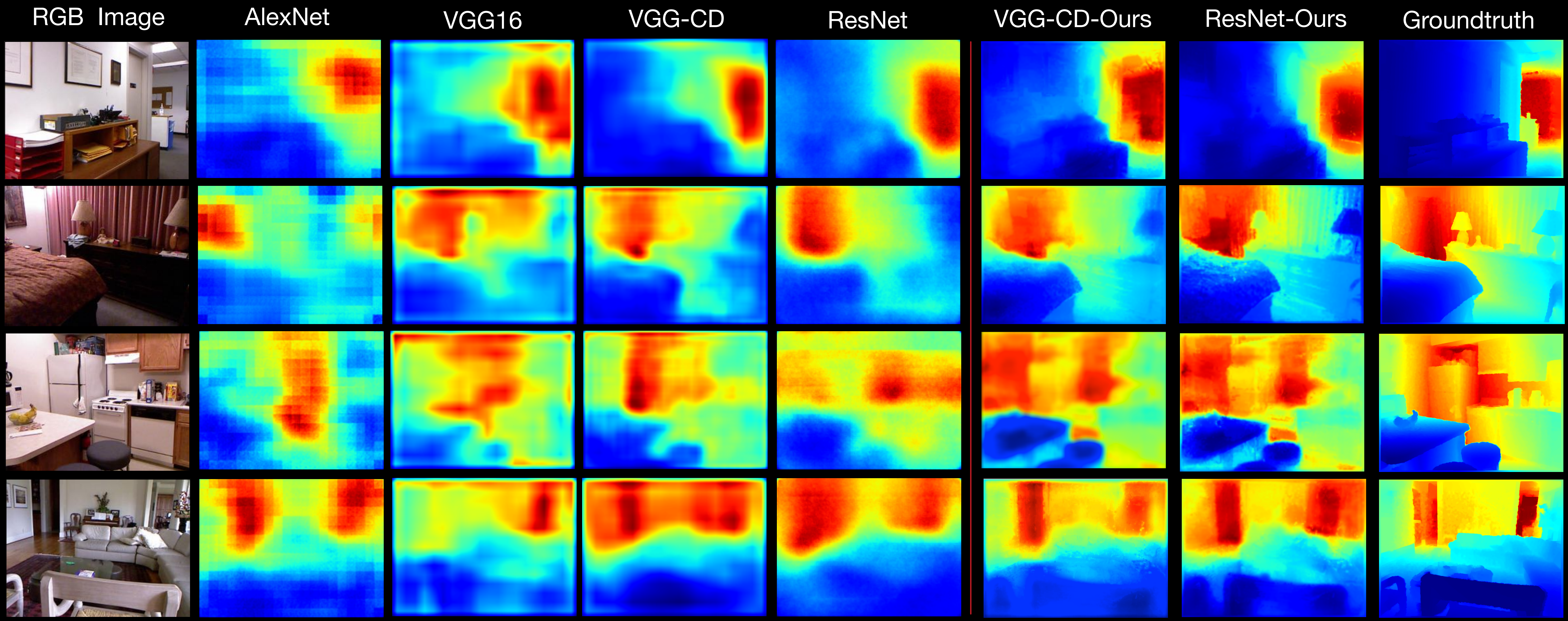
Multi-scale Structured Modelling

- Joint multi-scale CNN-CRF deep framework

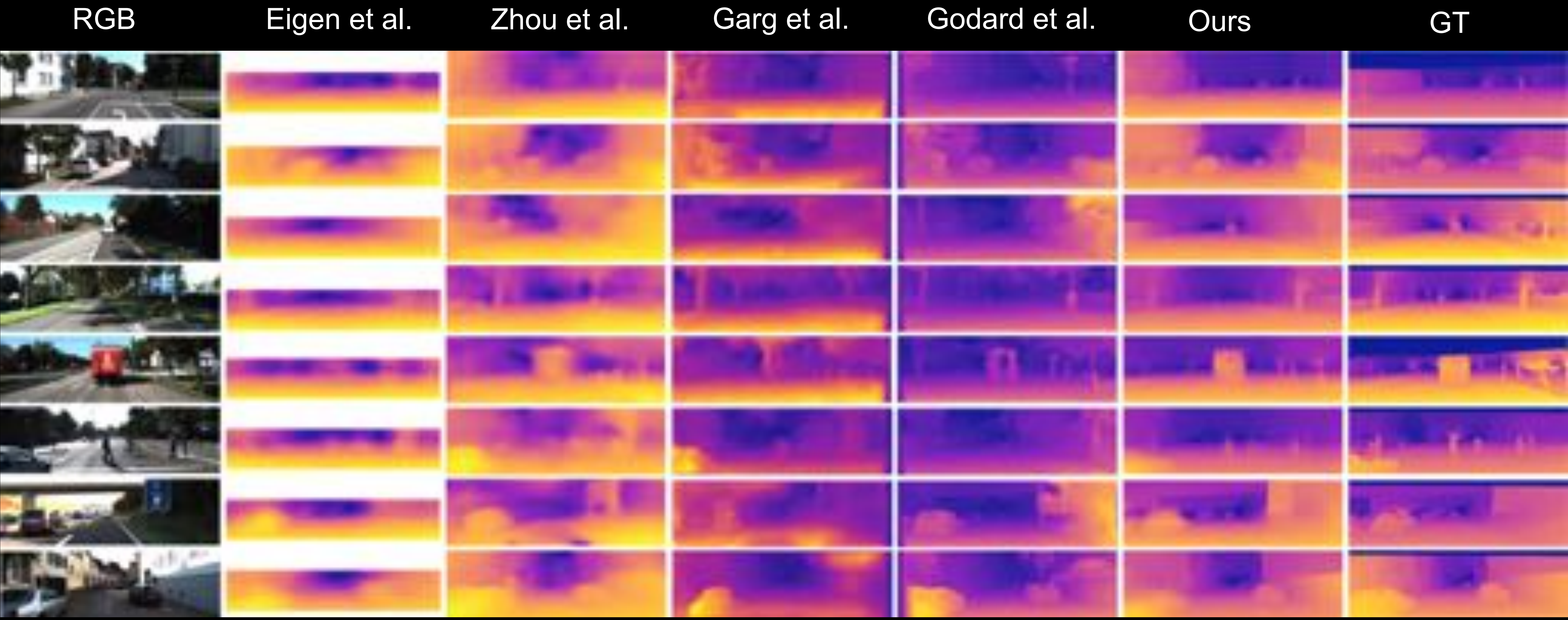


First work for multi-scale deep structured fusion & prediction in continuous domain

Results on NYUD-V2 Benchmark

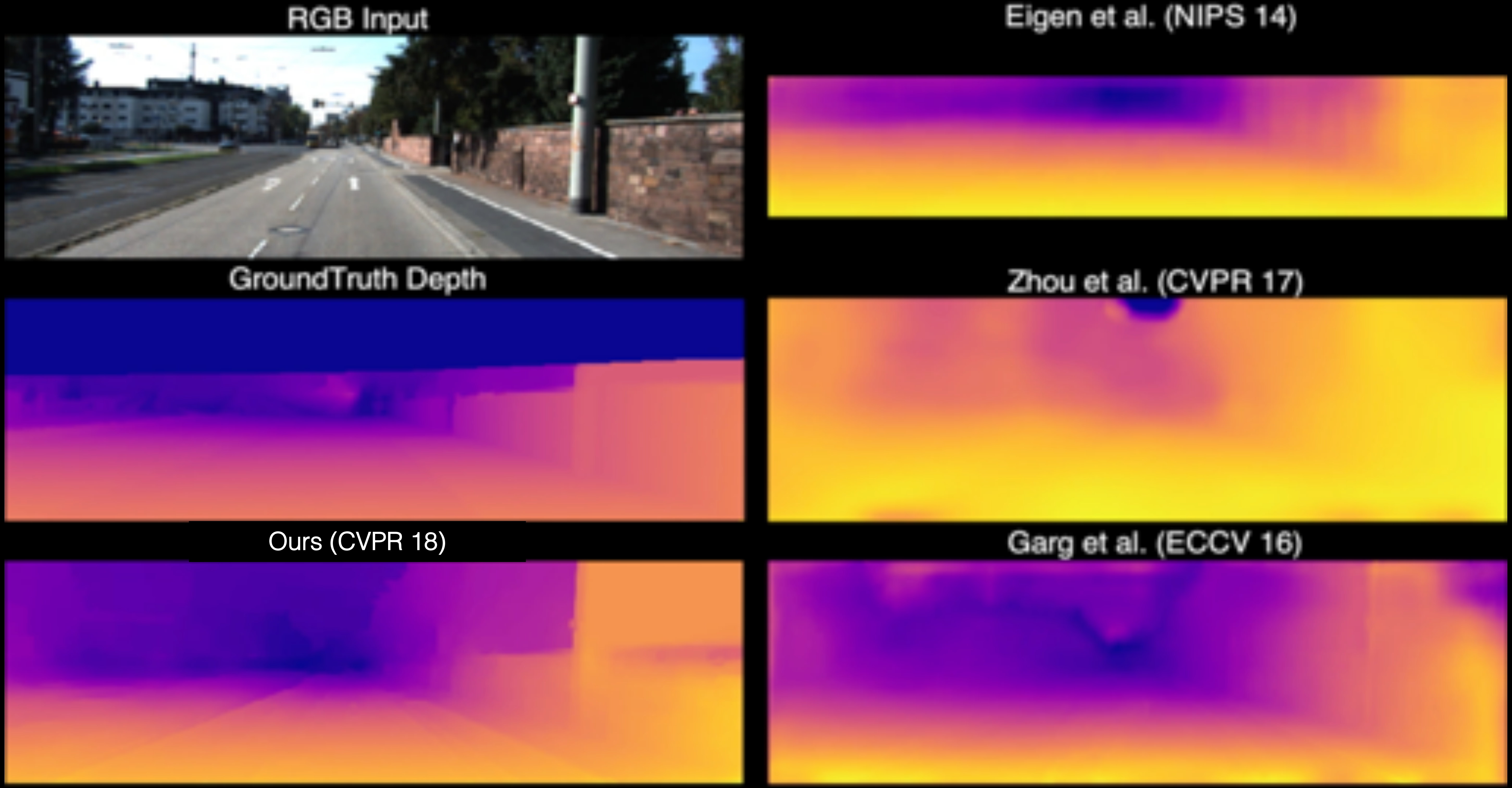


Results on KITTI Benchmark



Better qualitative results with more clear scene structure and details

Results on KITTI Benchmark

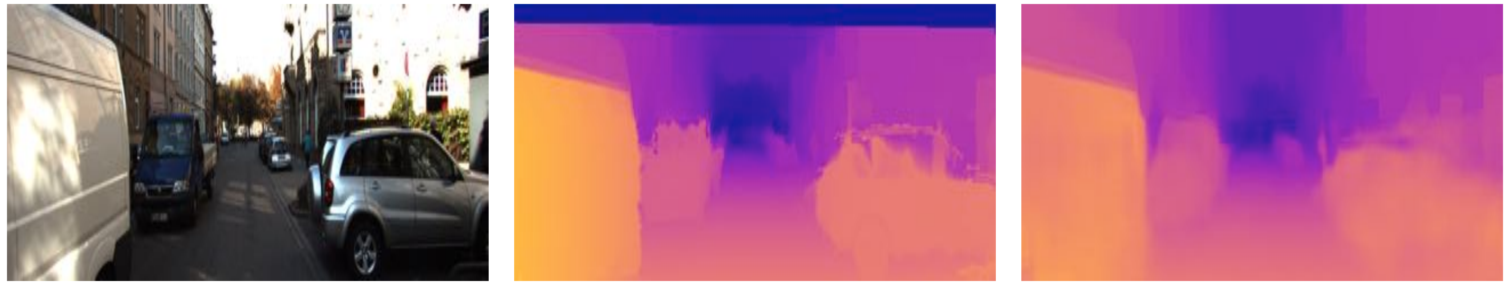


Structured Modelling on Deep Features

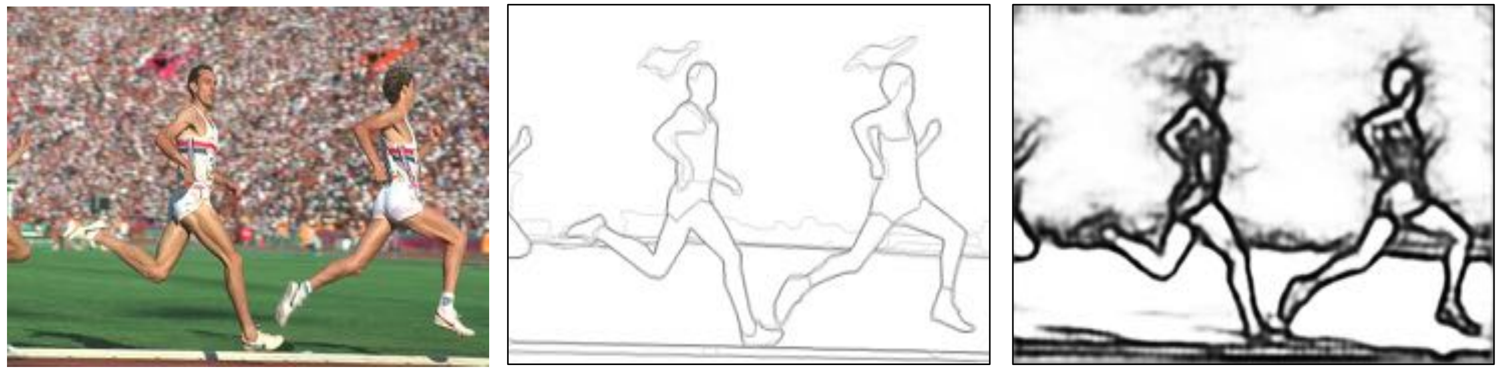
- Limitations in modelling on deep predictions
  - Less flexibility (continuous or discrete tasks)
  - Lose more scene structure information while the network goes deep

Continuous regression tasks

Discrete classification tasks



(a) Monocular Depth Estimation



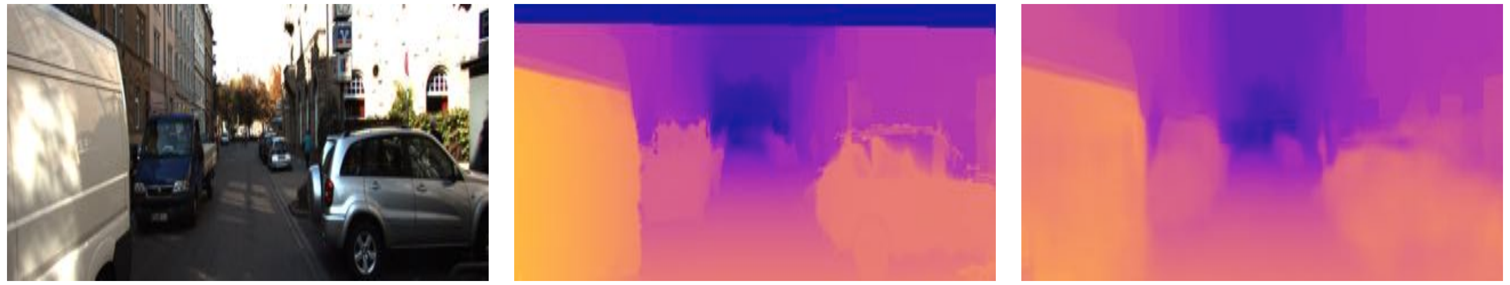
(b) Object Contour Detection

(c) Semantic Segmentation

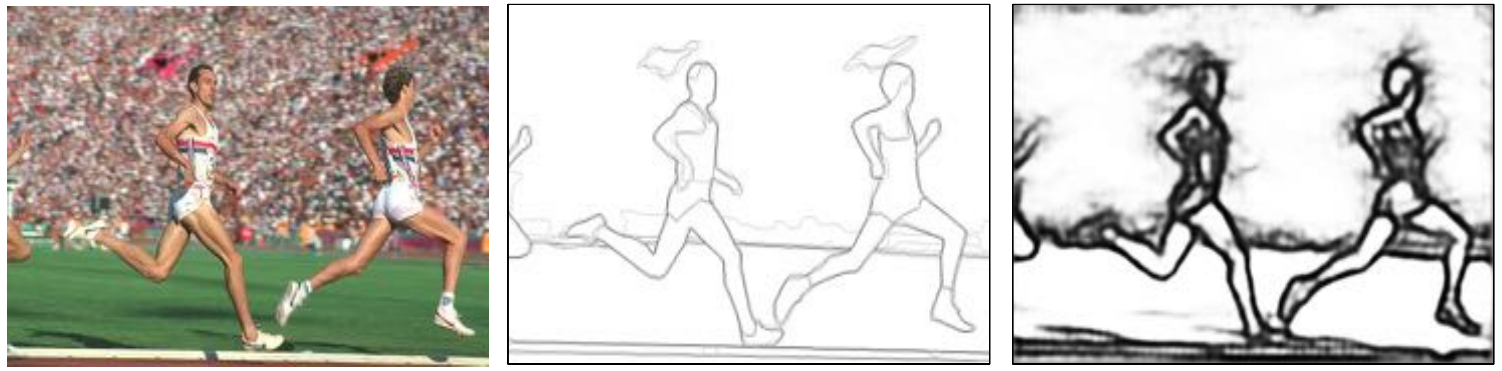
Structured Modelling on Deep Features

- Limitations in modelling on deep predictions
  - Less flexibility (continuous or discrete tasks)
  - Lose more scene structure information while the network goes deep

Design a model working on the intermediate feature level?



(a) Monocular Depth Estimation

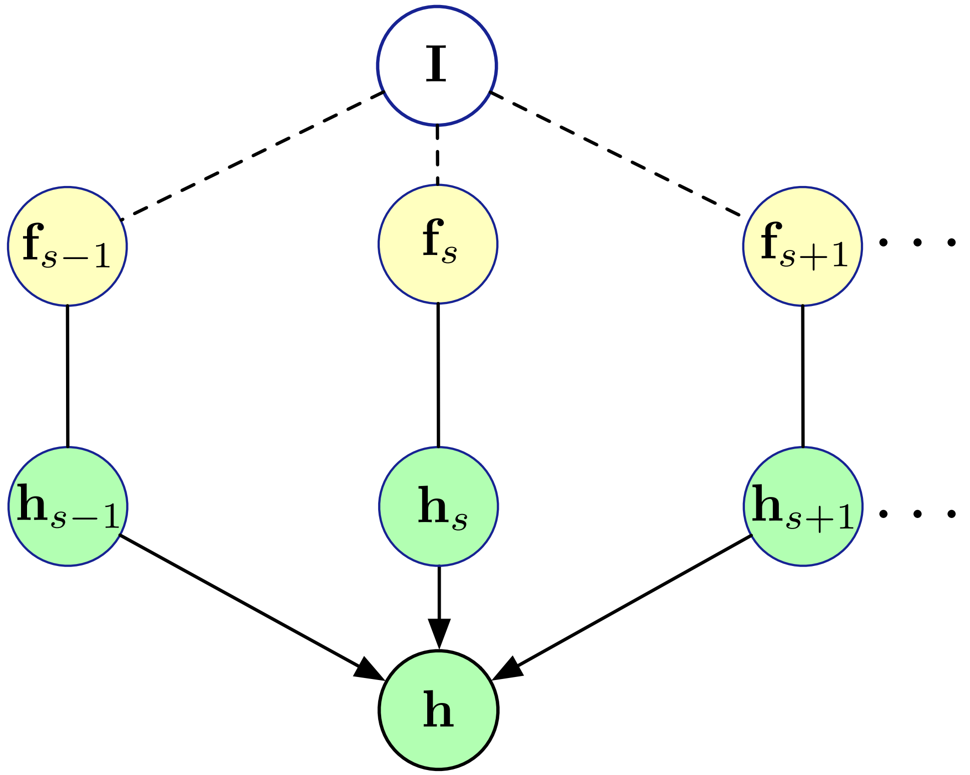


(b) Object Contour Detection

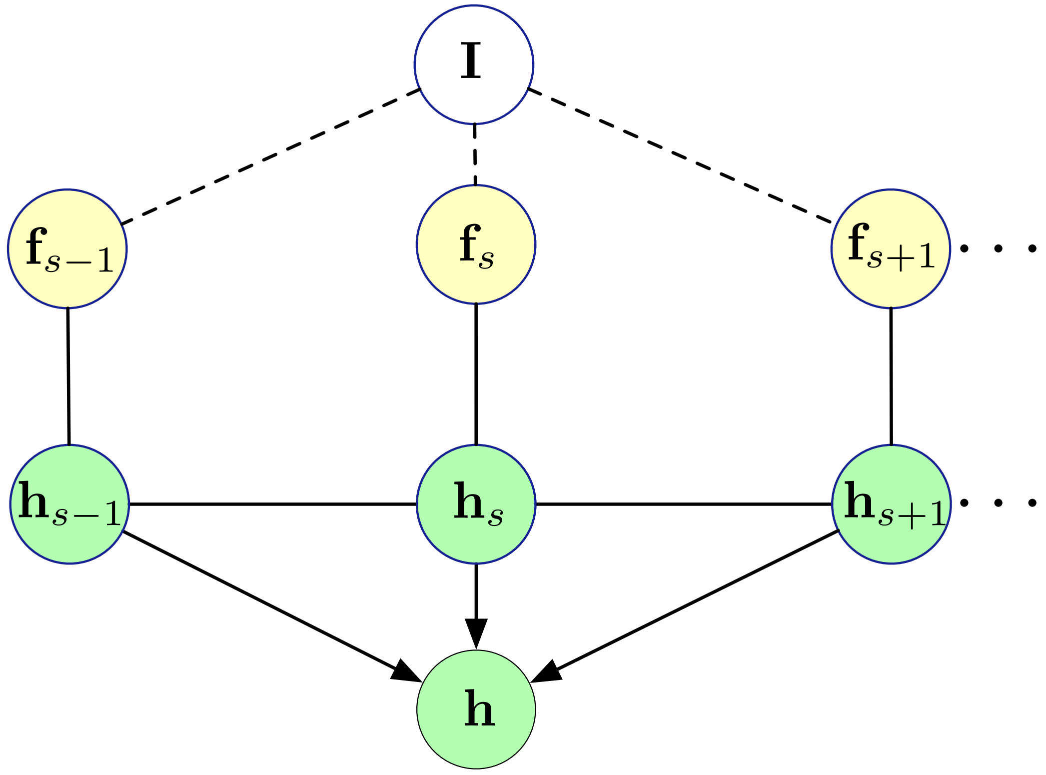
(c) Semantic Segmentation

Structured Modelling on Deep Features

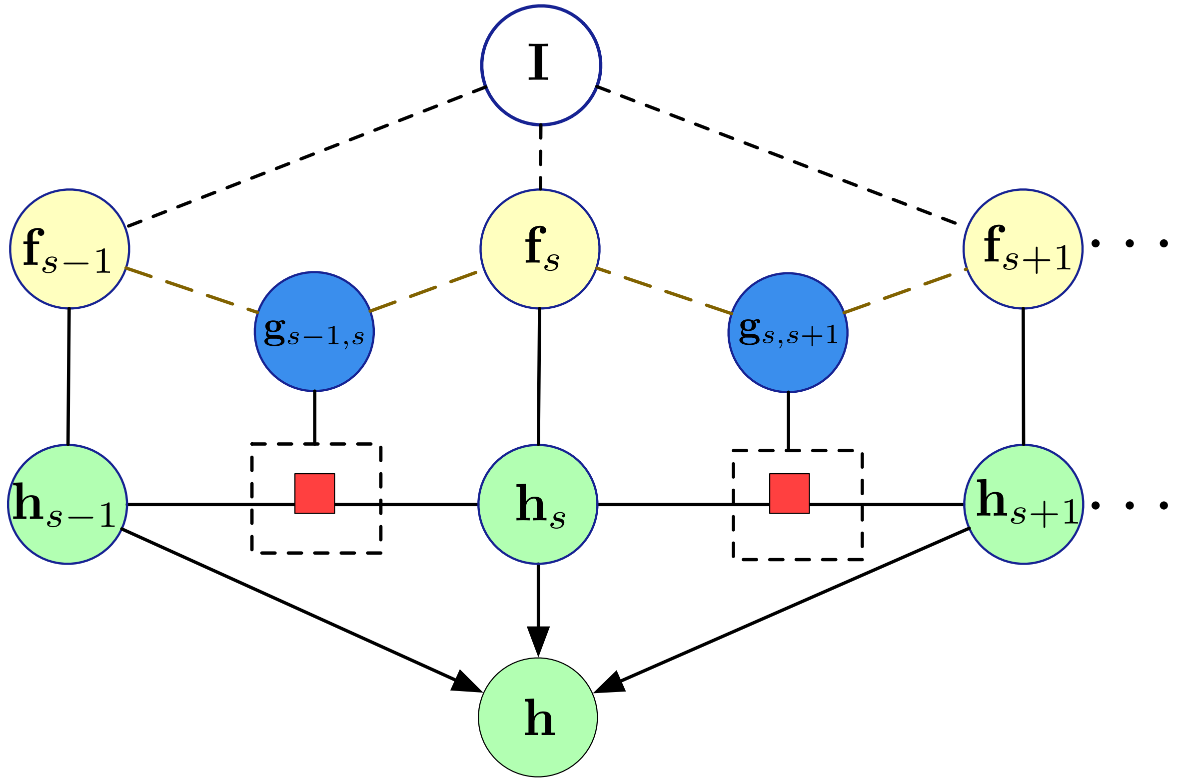
- Probabilistic graph attention network on deep features



(a) Multi-Scale Neural Network



(b) Multi-Scale CRFs

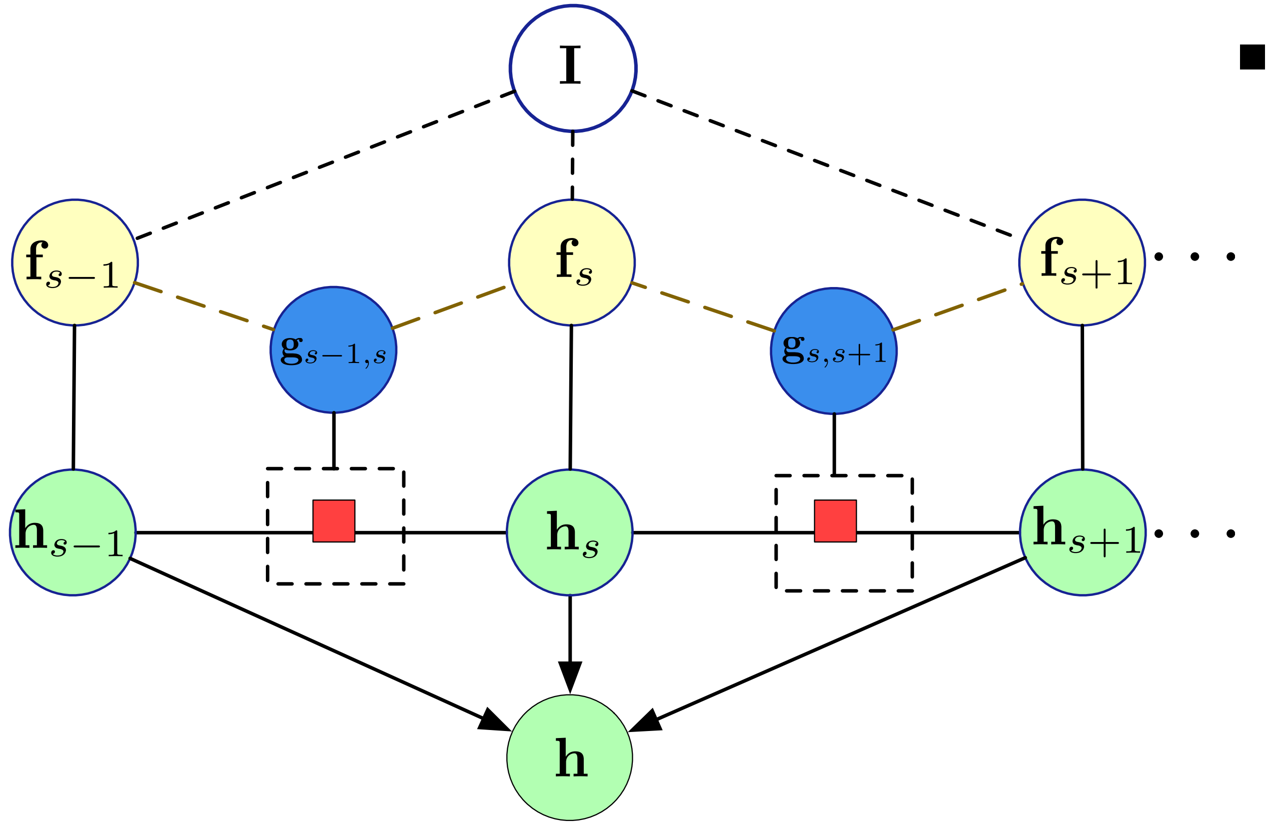


(c) Attention-Gated CRFs

- Attention as gating for controlling message passing between features

Structured Modelling on Deep Features

- Probabilistic graph attention network on deep features



(c) Attention-Gated CRFs

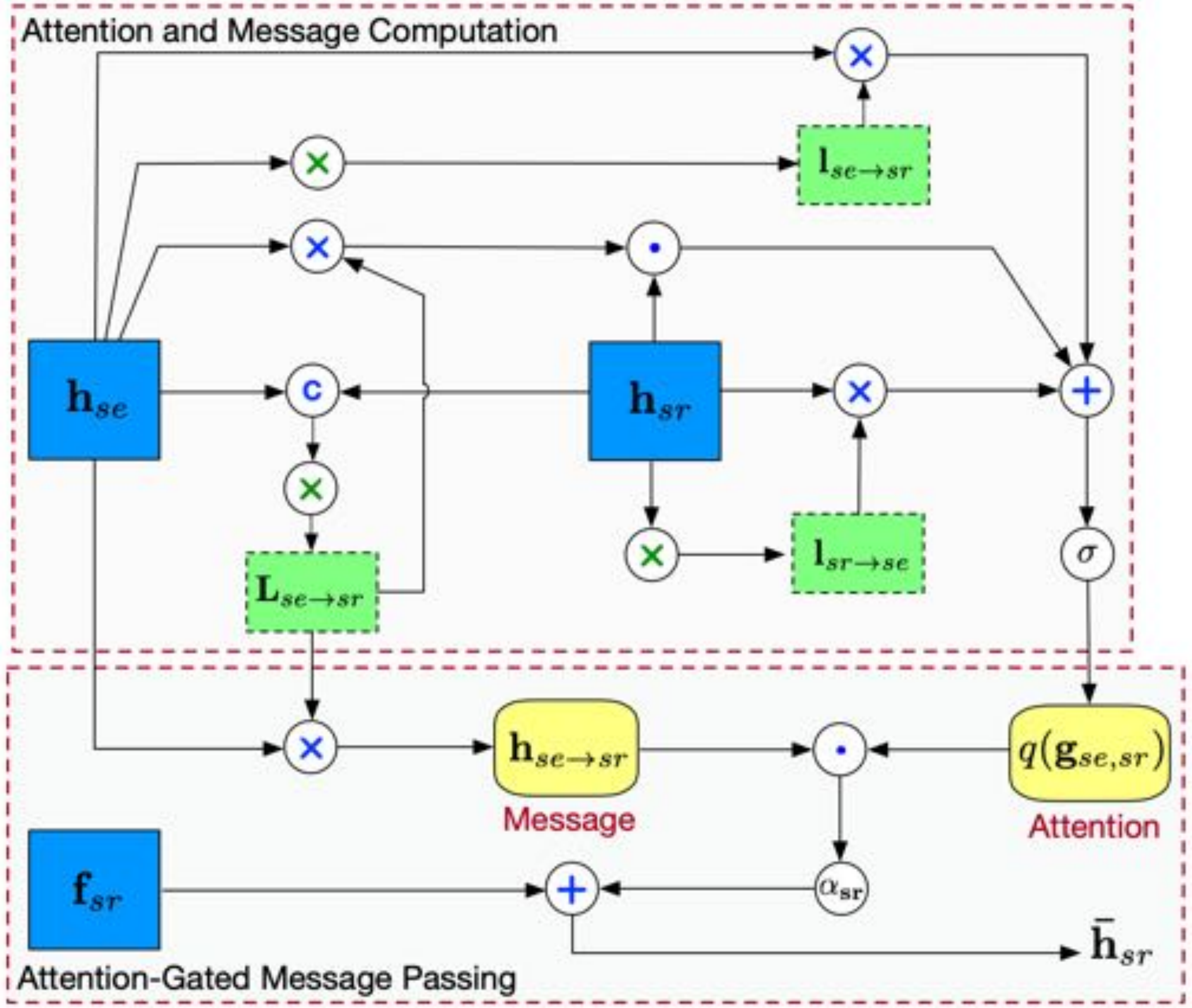
- Model formulation

$$E(\mathbf{H}, \mathbf{G}, \mathbf{I}, \Theta) = \underbrace{\sum_{s} \sum_{i} \phi_h(\mathbf{h}_s^i, \mathbf{f}_s^i)}_{\text{Unary potential}} + \underbrace{\sum_{s_e, s_r} \sum_{i,j} g^i_{s_e, s_r} \psi_h(\mathbf{h}^i_{s_r}, \mathbf{h}^j_{s_e})}_{\text{Gated pairwise potential}} .$$

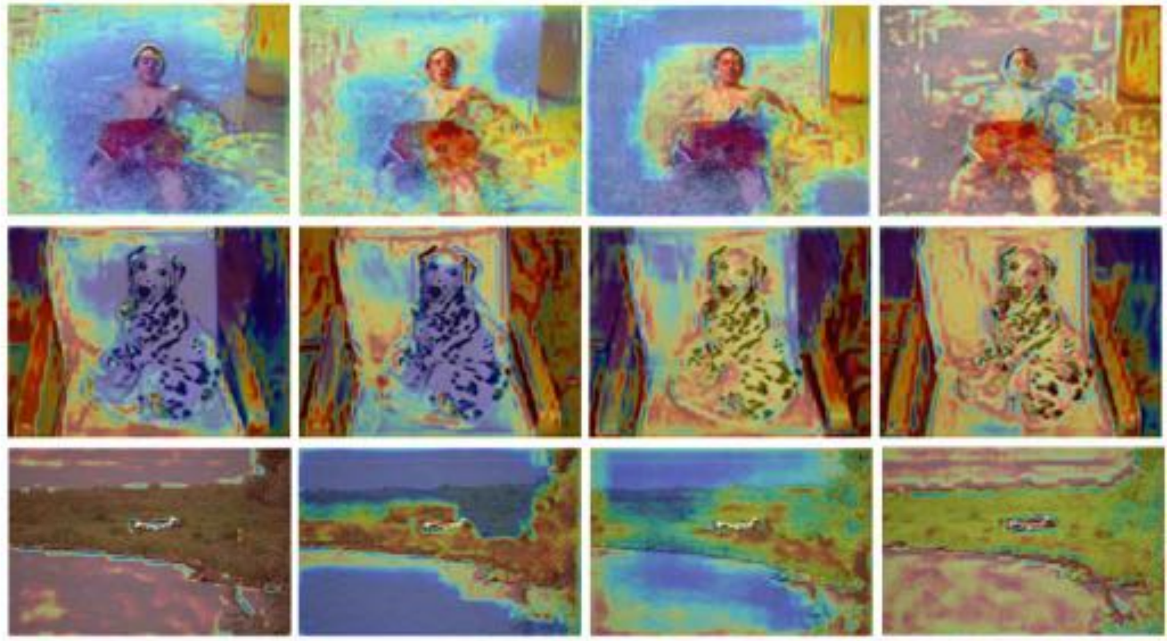
$$= -\sum_s \sum_i \frac{a_s^i}{2} \left\| \mathbf{h}_s^i - \mathbf{f}_s^i \right\|^2 + \sum_{s \in s_k} \sum_{i,j} g^i_{s_s, s_n} \tilde{\mathbf{h}}^i_{s_r} \mathbf{K}^{i,j}_{s_r, s_c} \tilde{\mathbf{h}}^j_{s_c}$$

Structured Modelling on Deep Features

- Probabilistic graph attention network on deep features
  - Neural network implementation



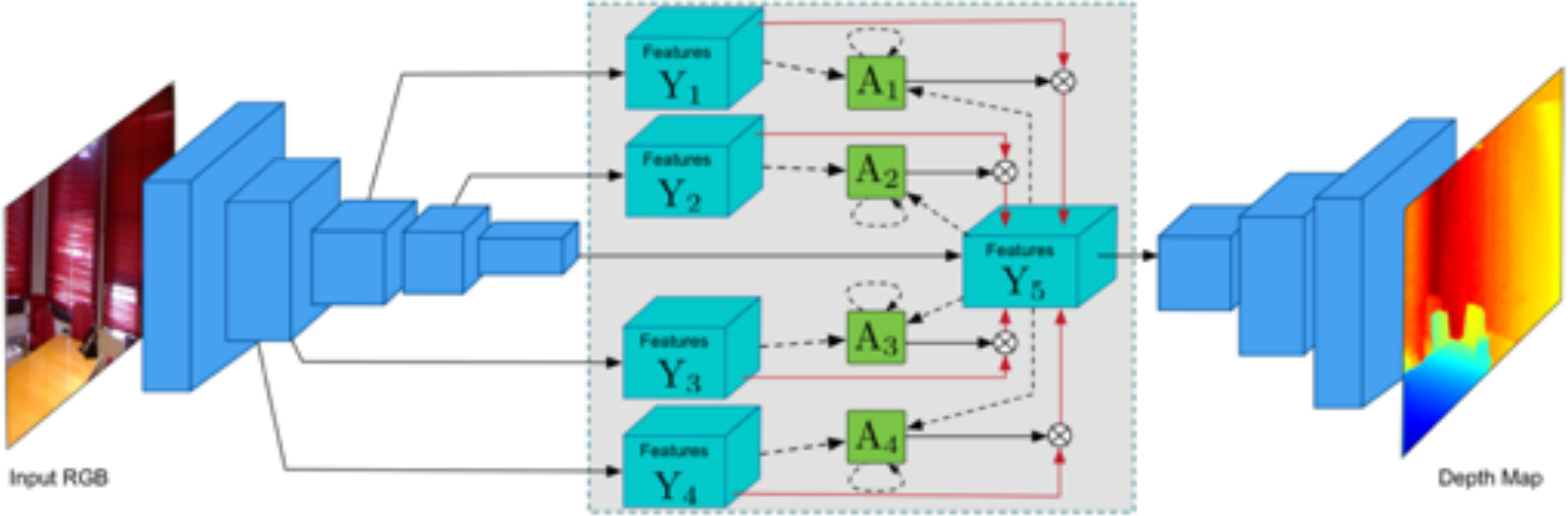
Learned attention on KITTI depth estimation



Learned attention on Pascal-Context segmentation

Structured Modelling on Deep Features

- Probabilistic graph attention network on deep features
  - Applicable in the middle of a CNN for deep structured feature refinement



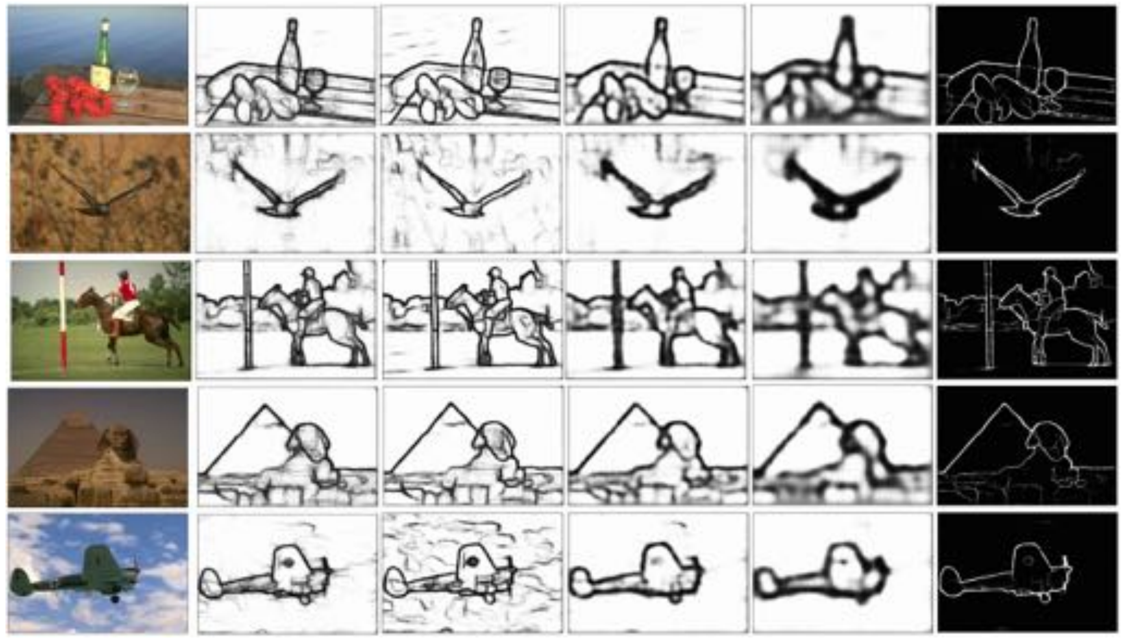
Encoder

Structured graph attention network module

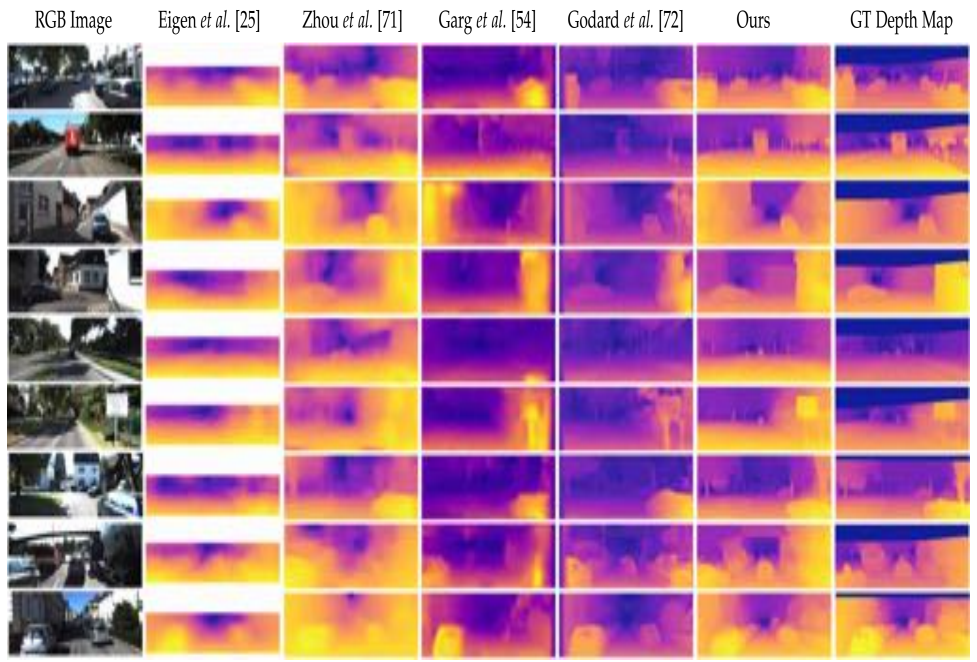
Decoder

Structured Modelling on Deep Features

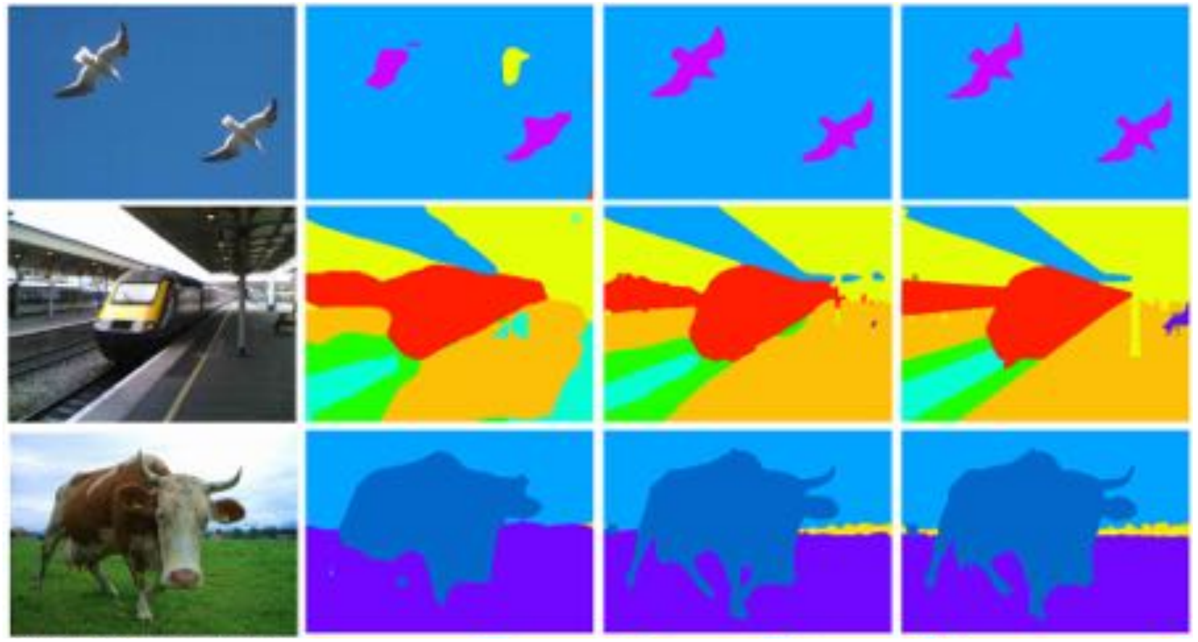
- Significant improvement on different continuous or discrete tasks



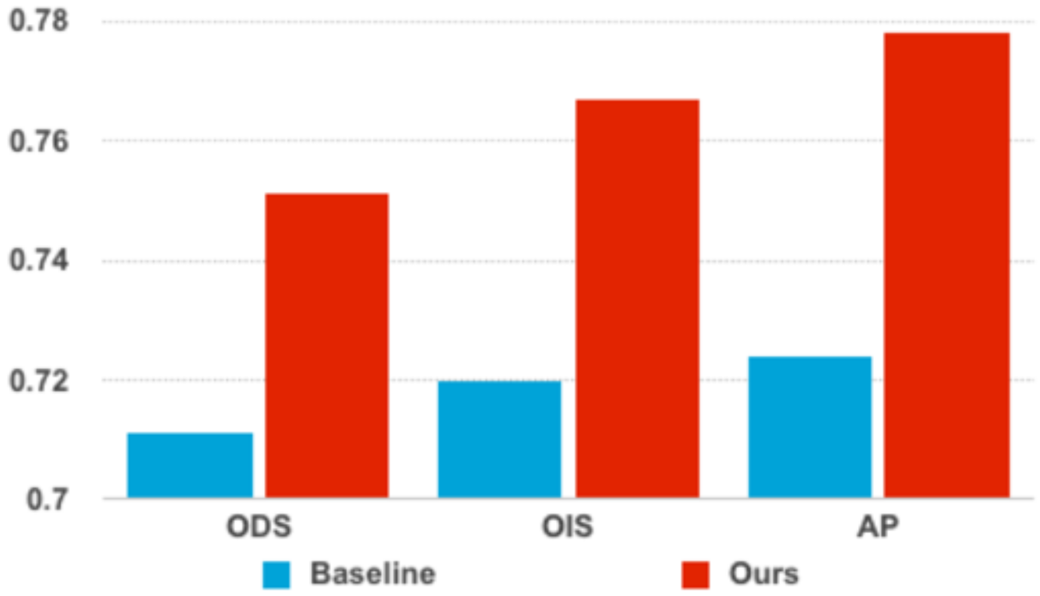
Contour detection on BSDS500

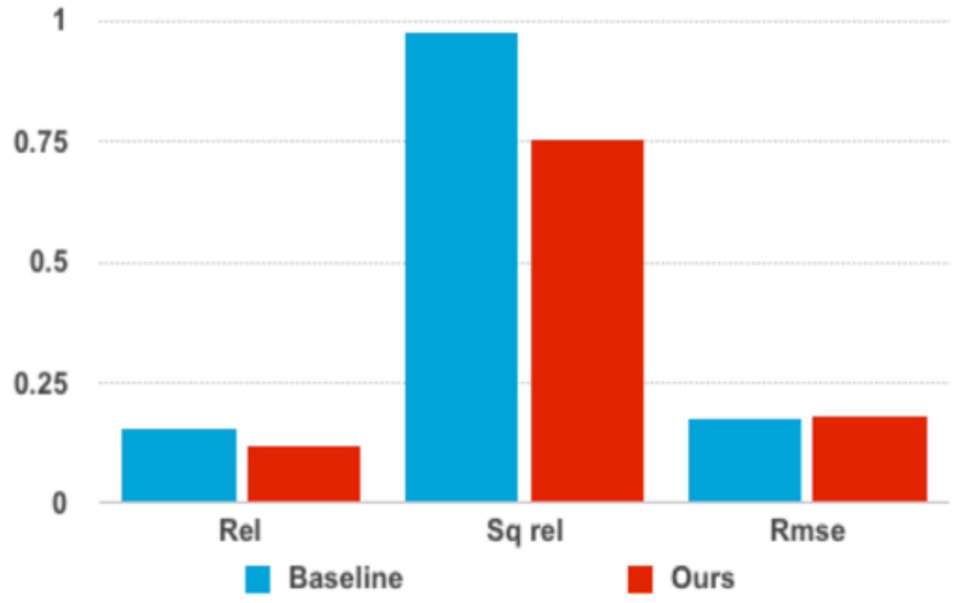


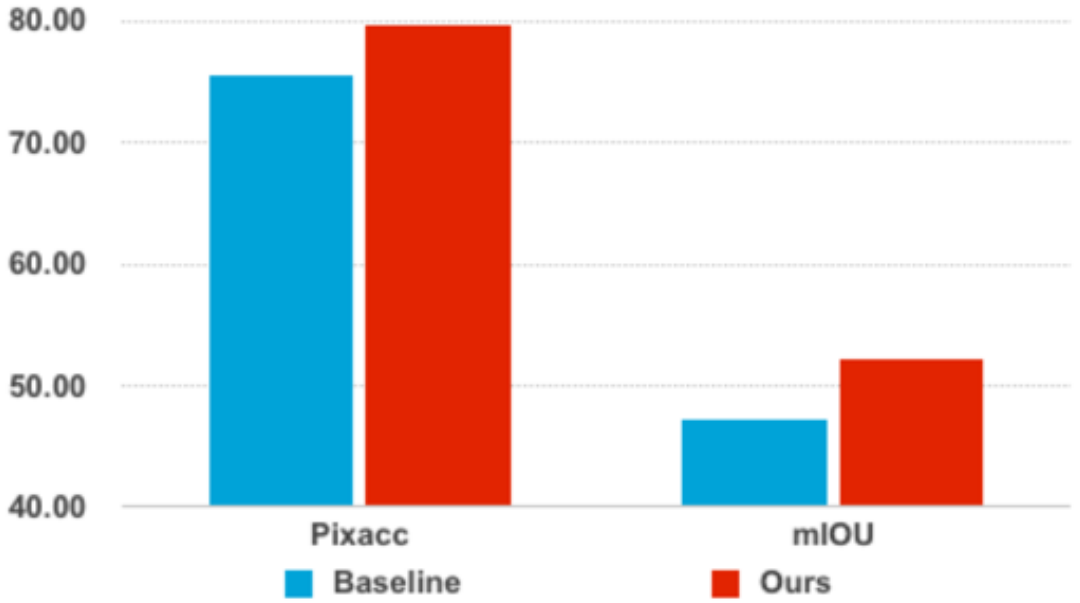
Depth estimation on KITTI



Segmentation on Pascal-Context







Overview

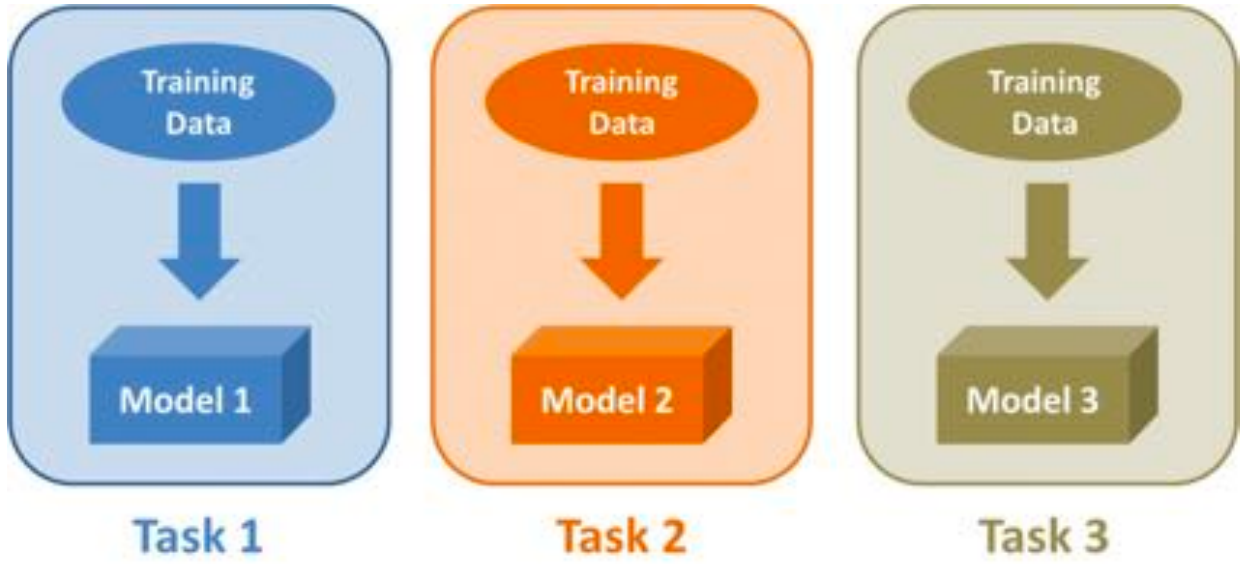
- Scene depth estimation with structured probabilistic modeling
- A joint multi-modal and multi-task deep learning framework
- Mo between 2D
- Ho ent fields alo
- Su

Single-Task Deep Learning

Multi-Modal Multi-Task Deep Learning

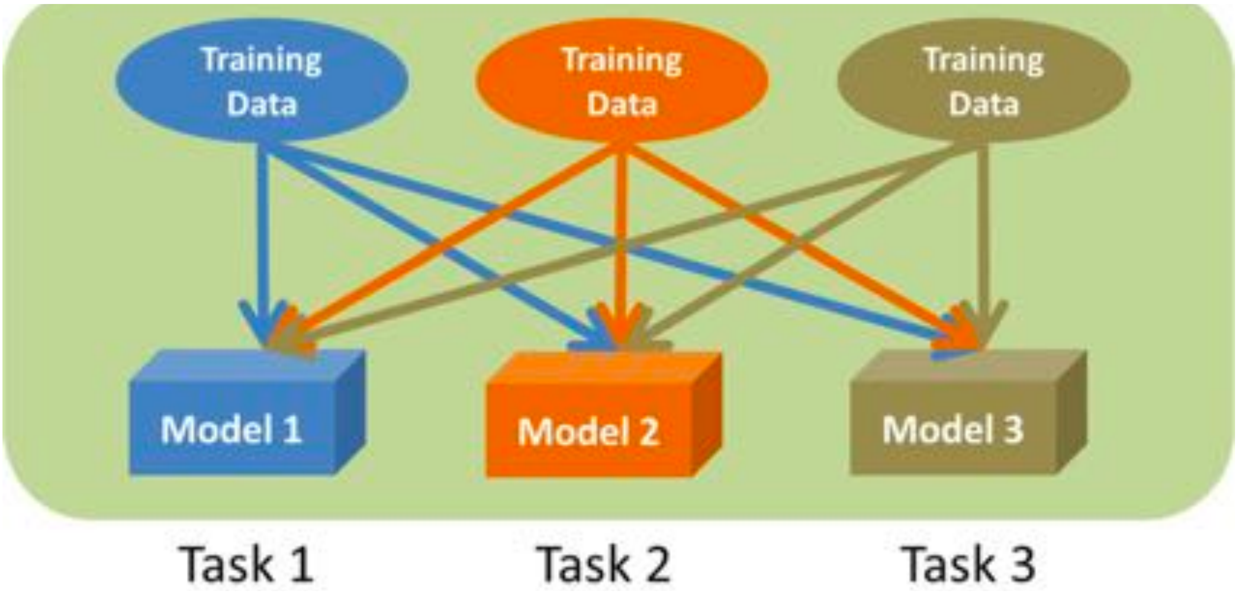
Joint Multi-Modal/Task Deep Learning

- Single task learning vs. multi-task learning



Single task learning

- Independently train each task
- No training data or parameter sharing

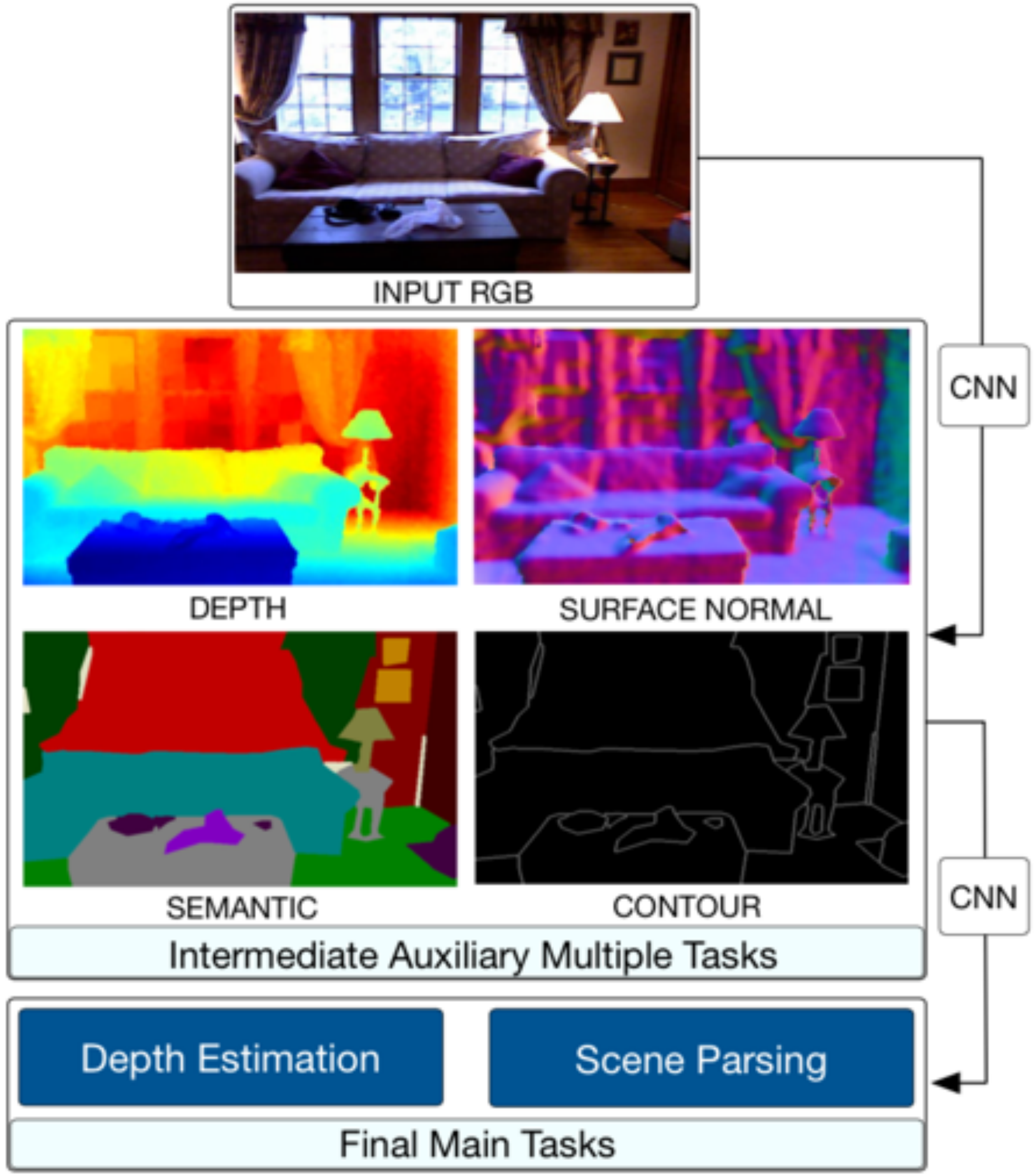


Multi task learning

- Train multi-tasks with shared multi-modal data
- Tasks dependent to each other
- Convenience in deployment

Joint Multi-Modal/Task Deep Learning

- Problems and motivation in multi-task deep learning

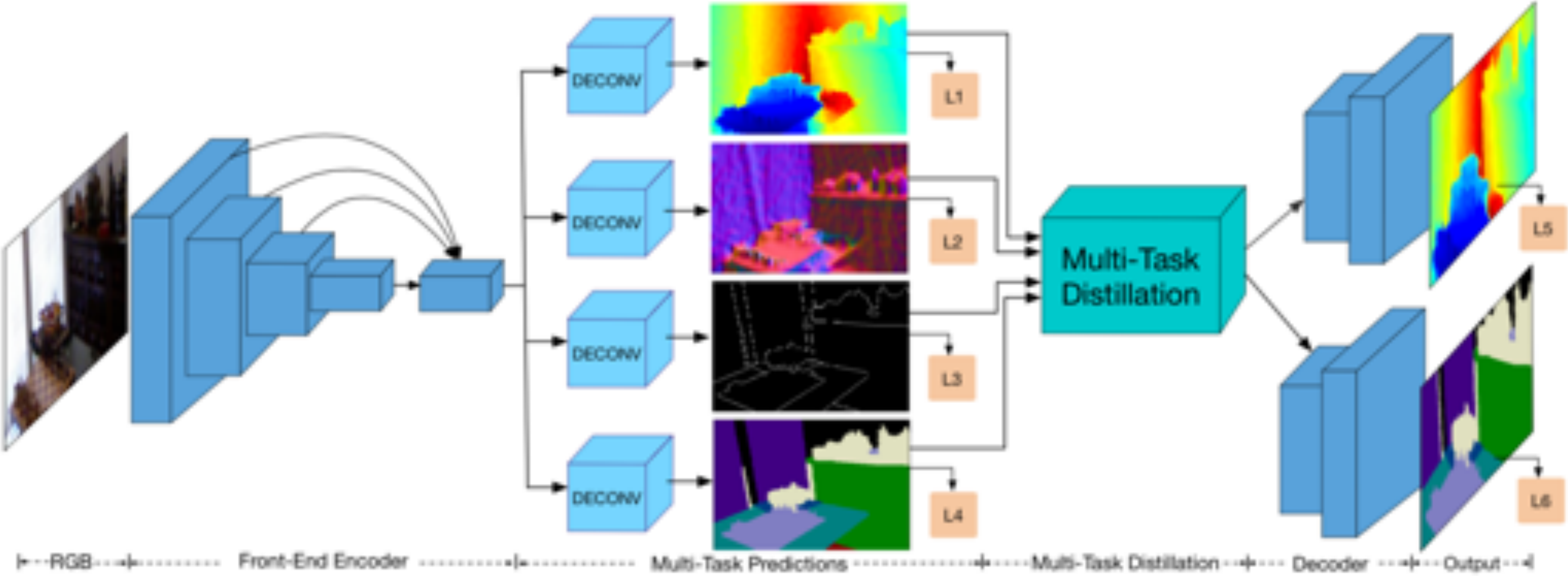


- Difficulty: Directly optimizing multiple tasks given input training data not guarantees consistent gain on all the tasks
- Observation: Multi-modal input data improves training the model

- Could we facilitate final tasks via leveraging intermediate multiple predictions?
- Only one single modal data required?

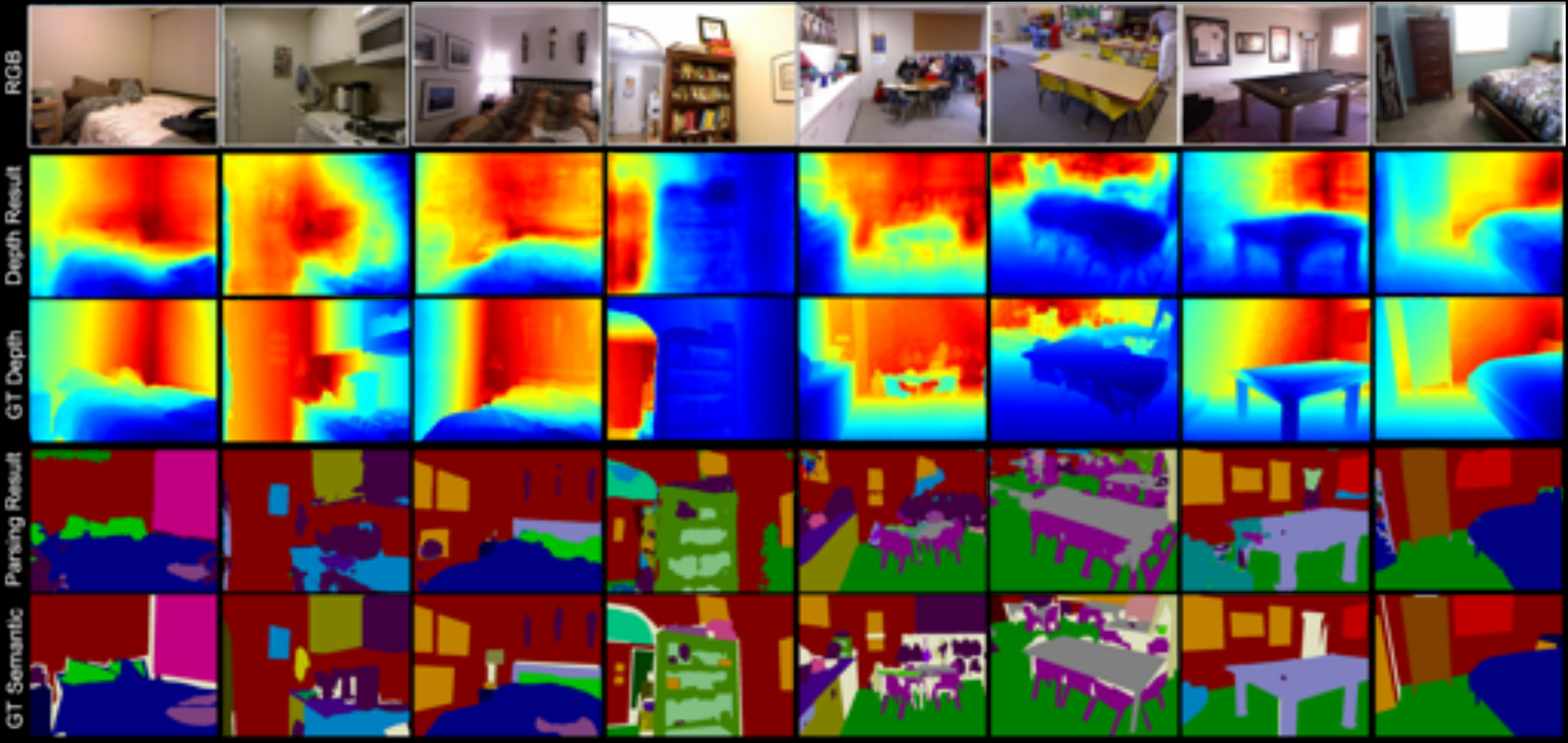
PAD-Net: Prediction and Distillation Network

- Network structure

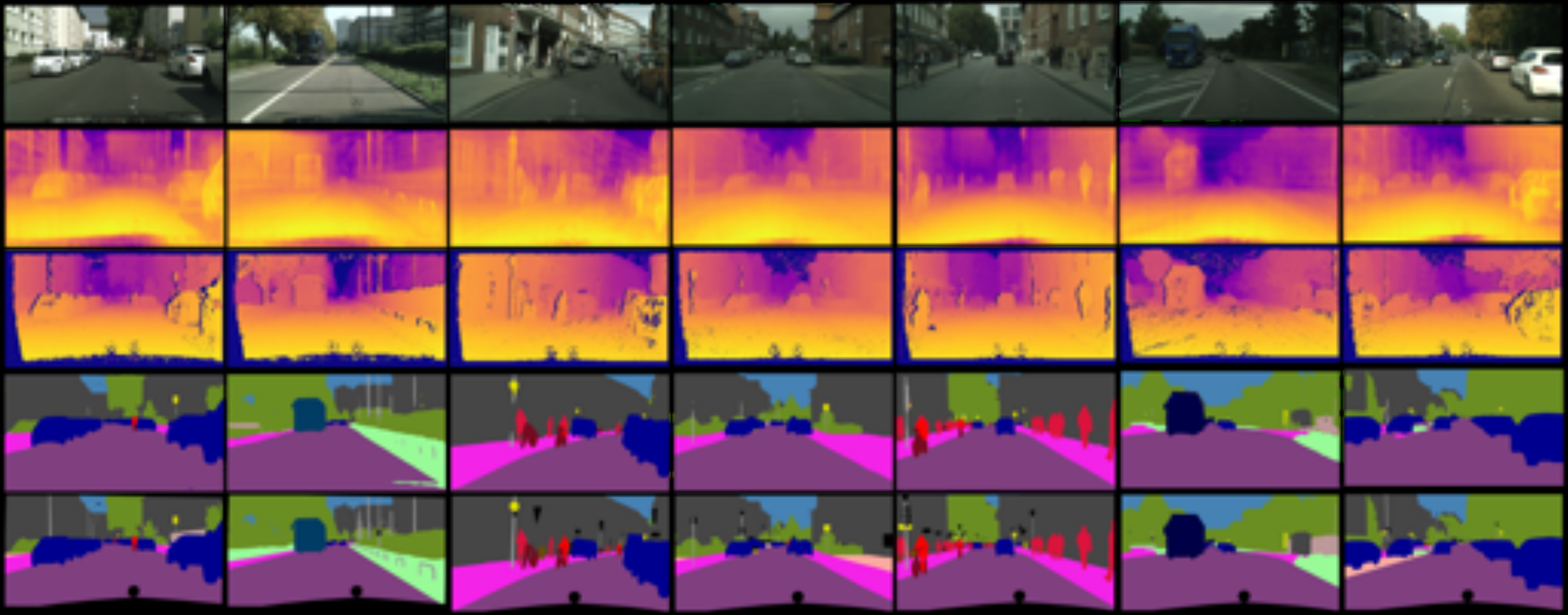


Multi-task distillation network for simultaneous depth estimation and scene parsing.

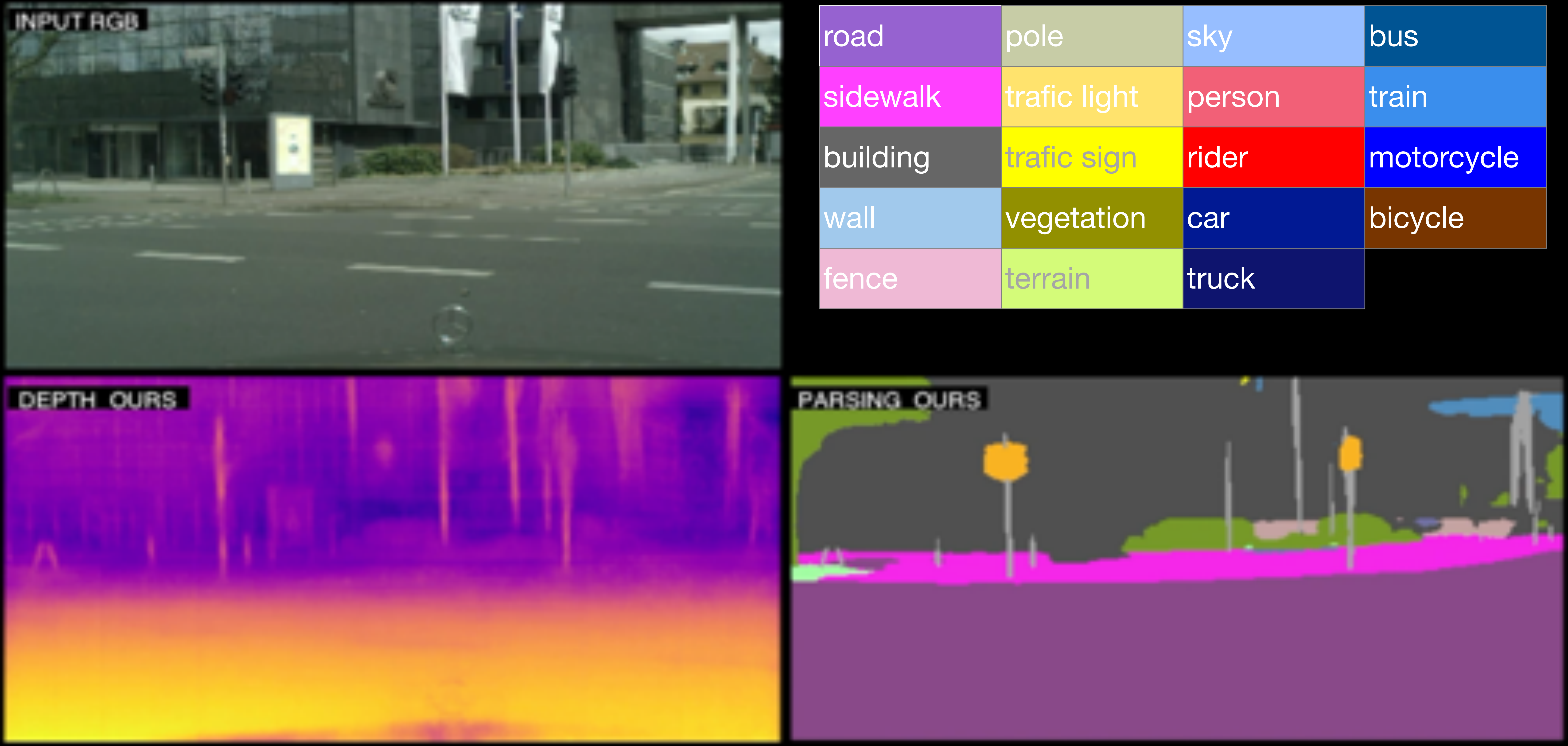
Results on Indoor NYUD-V2



Results on outdoor Cityscapes



Demo on Outdoor Cityscapes Dataset



road	pole	sky	bus
sidewalk	trafic light	person	train
building	trafic sign	rider	motorcycle
wall	vegetation	car	bicycle
fence	terrain	truck	

Overview

- Scene depth estimation with structured probabilistic modeling
- A joint multi-modal and multi-task deep learning framework
- Modelling the interaction between 2D and 3D data and tasks
- H ment fields al
- S

Deep Learning in 2D

→

Deep Learning in the interaction of 2D & 3D

Perception of 3D from 2D

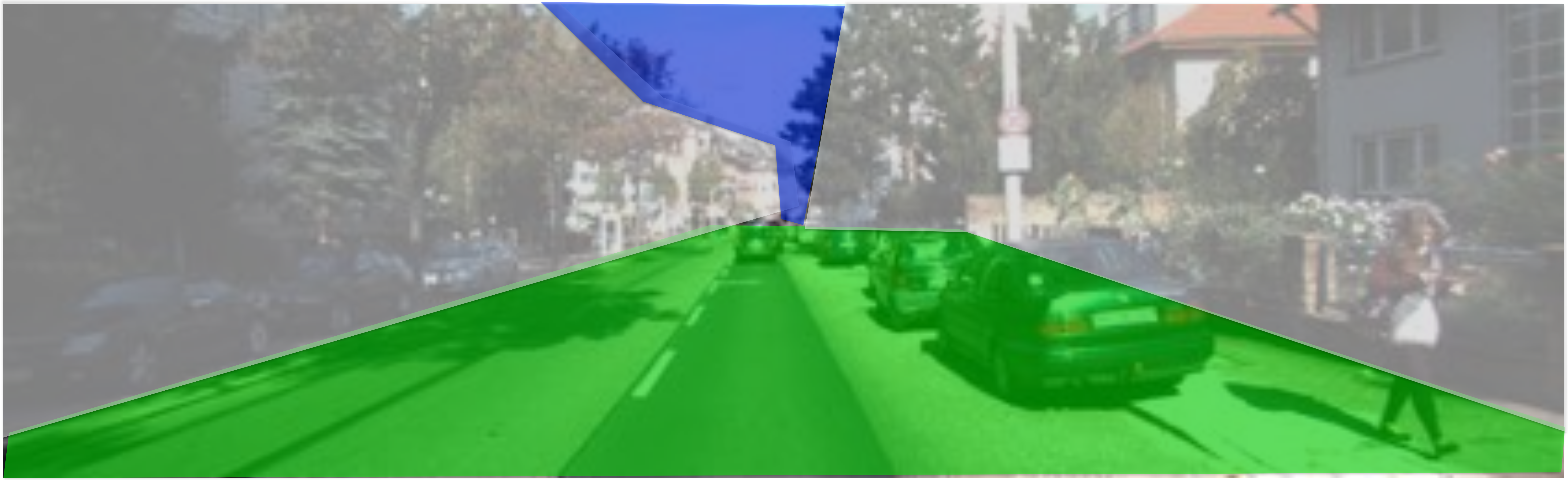
- 2D RGB Image

Perception of 3D from 2D

- 2D RGB Image -> Depth

Perception of 3D from 2D

- 2D RGB Image -> 3D Layout



Learning 3D from 2D

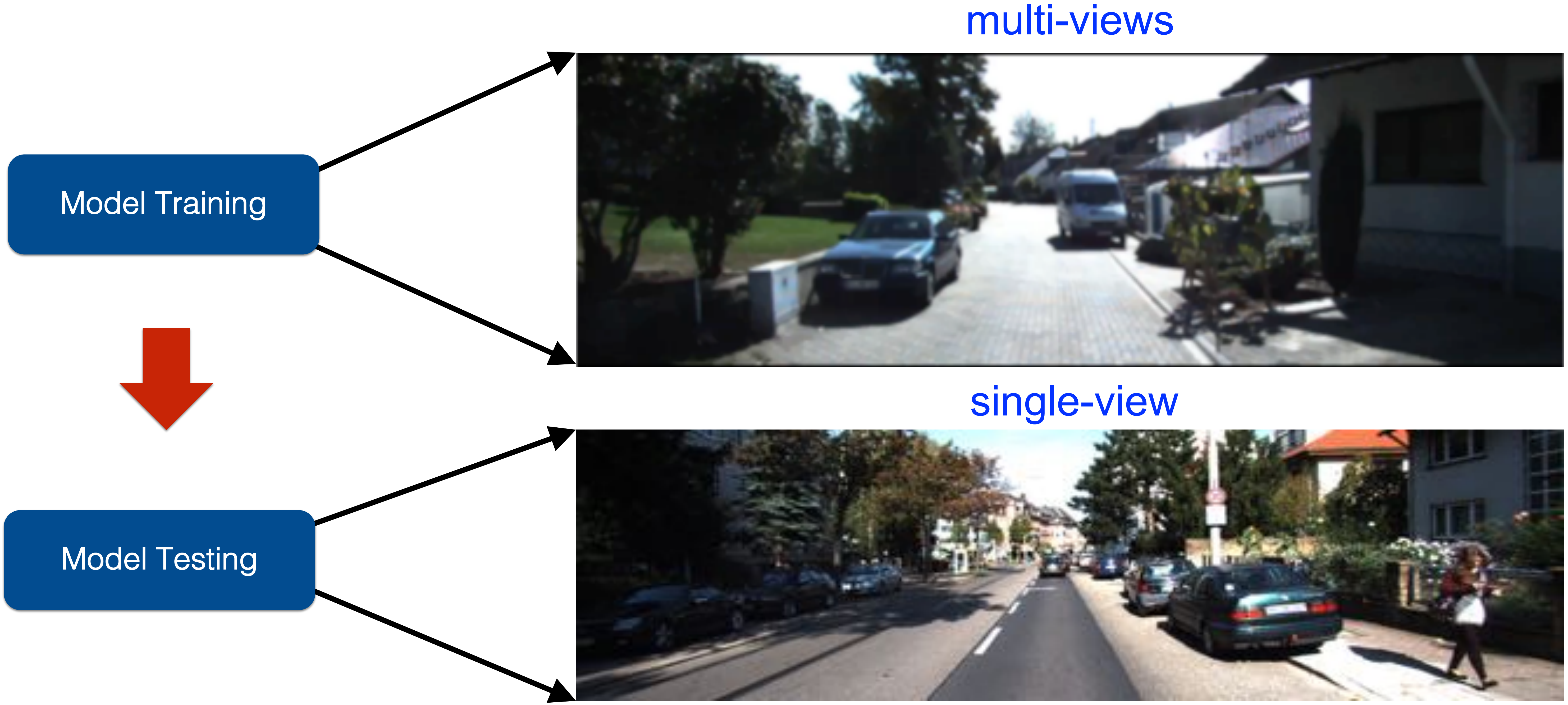
- Ambiguity in 2D: depth lost during projection
- Supervised learning using ground-truth 3D data

RGB

GT Depth

Learning 3D from 2D

- Self-supervised learning from multi-views



multi-views

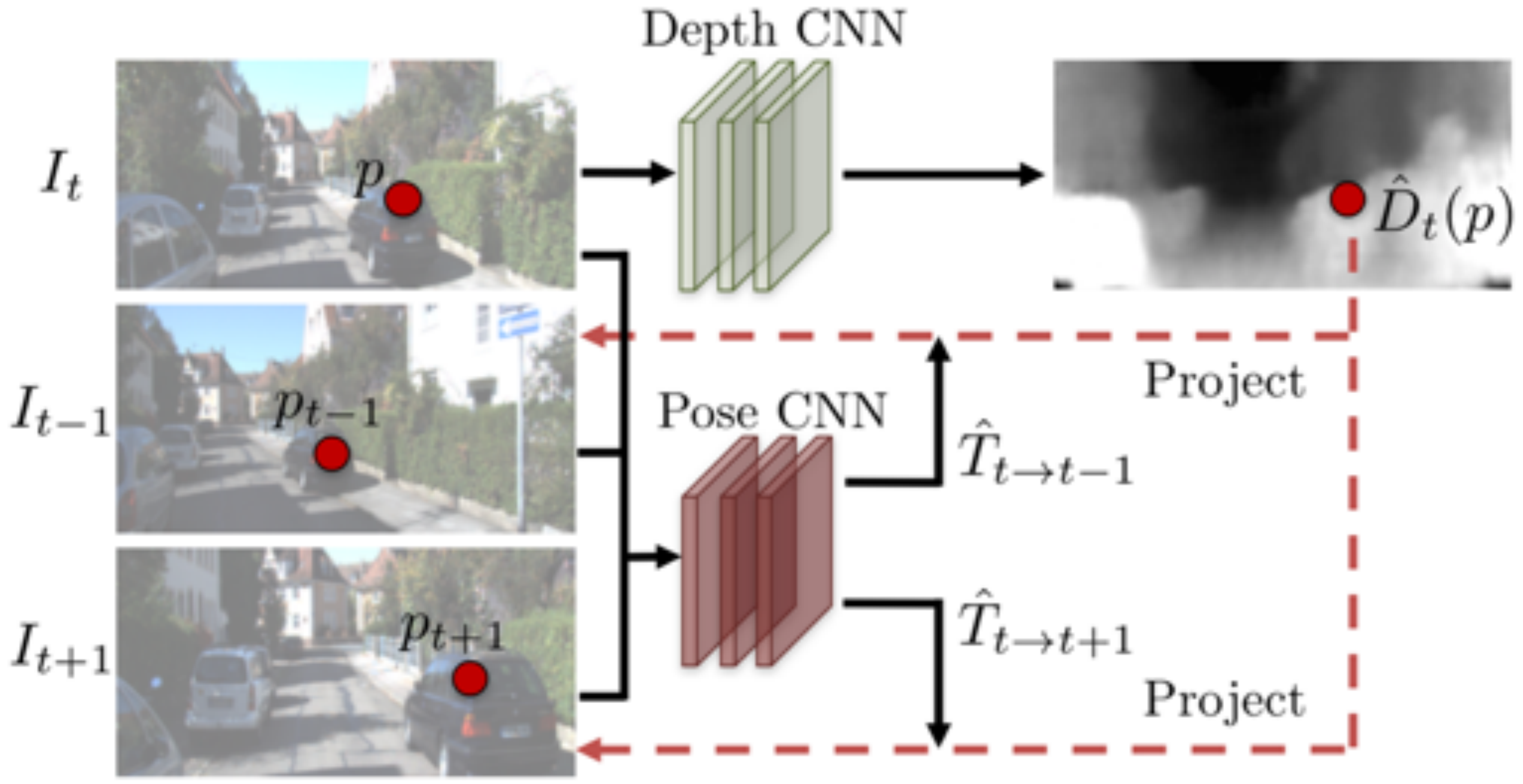
Model Training

single-view

Model Testing

SfM(Structure from Motion)-Learner

- Self-supervised framework for joint learning of depth and pose



- Differentiable image warping

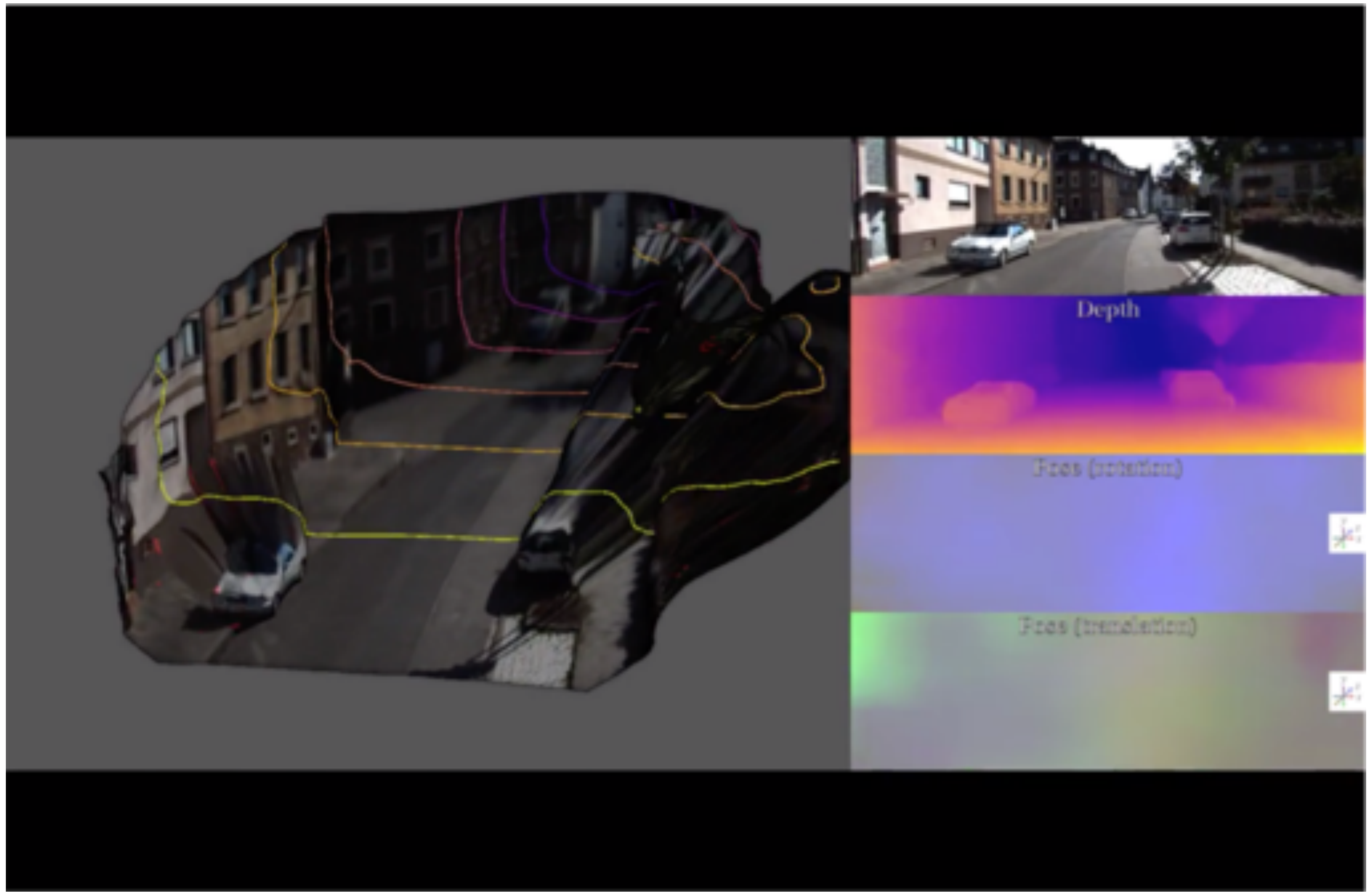
$$p_s \sim K\hat{T}_{t\to s}\hat{D}_t\left(p_t\right)K^{-1}p_t$$

- Photometric consistency

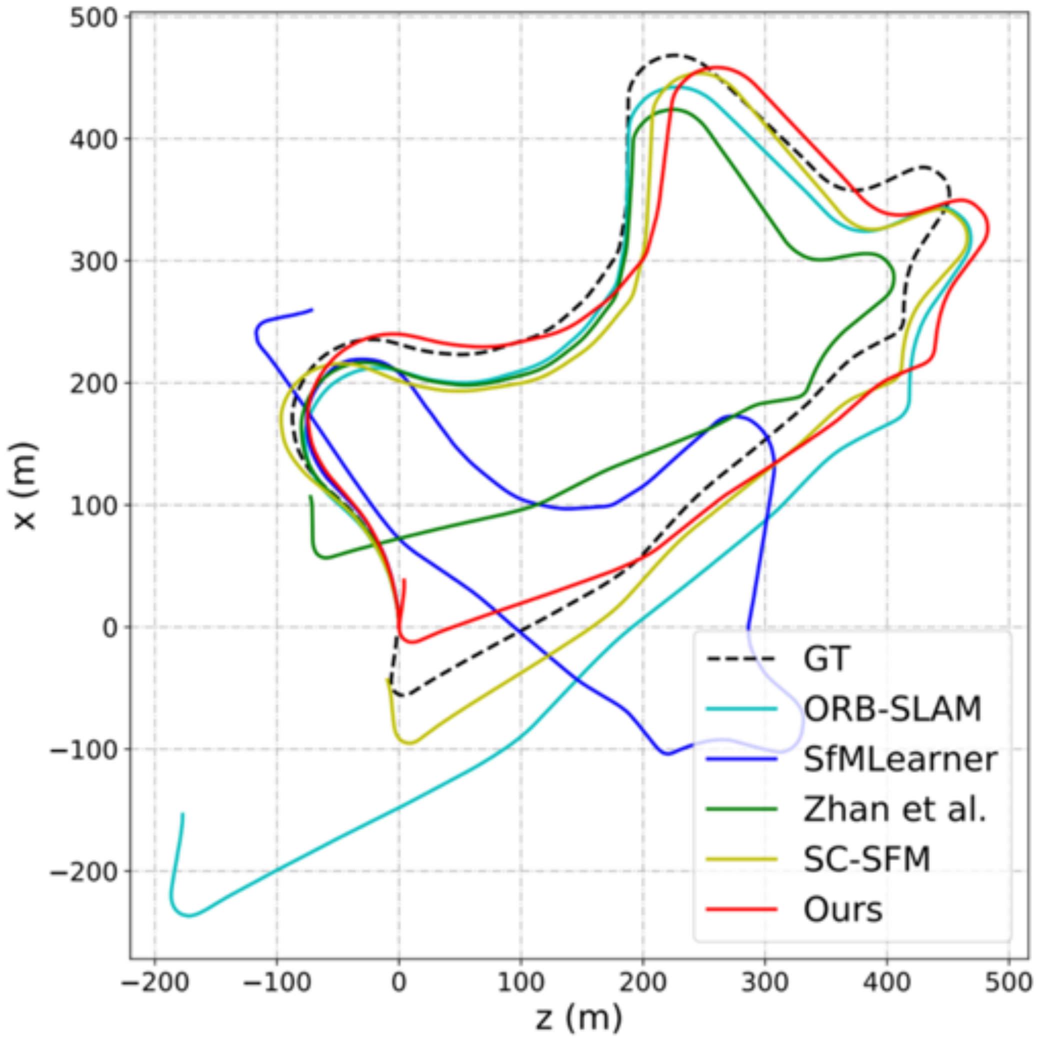
$$\mathcal{L}_{vs} = \sum_s \sum_p \left| I_t(p) - \hat{I}_s(p) \right|$$

Experimental Results

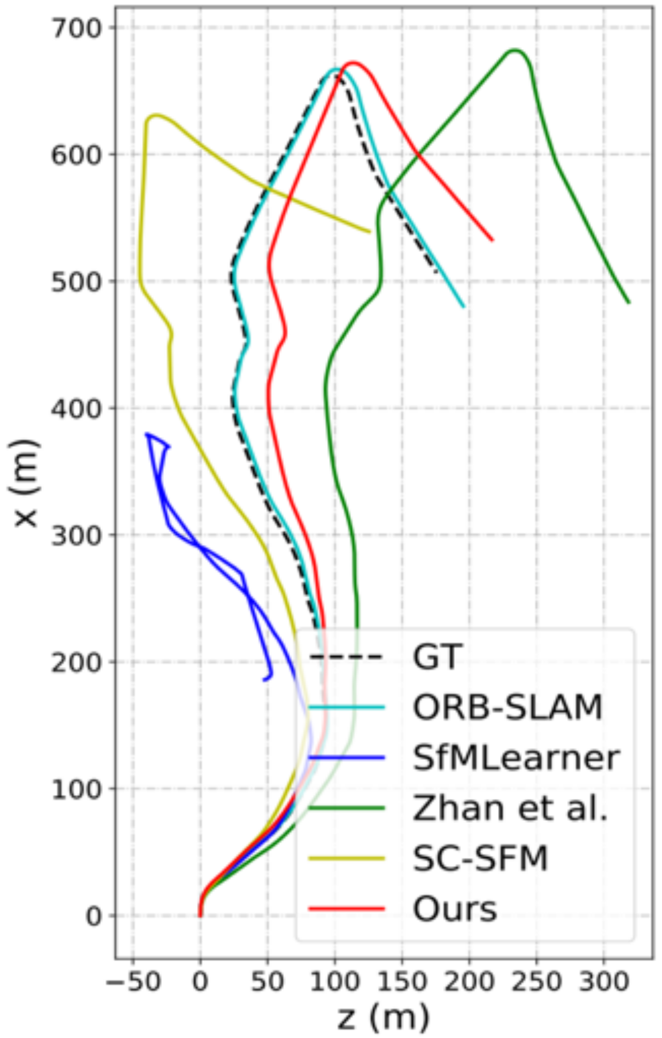
- Kitti visual odometry



Video Demo



Sequence 09



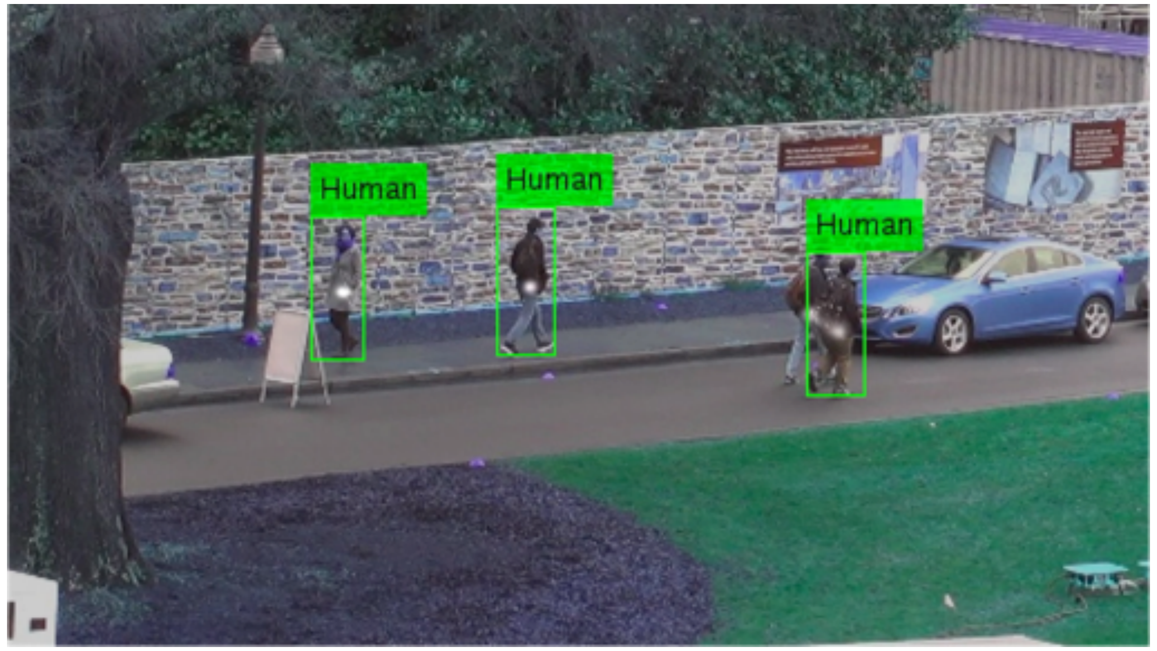
Sequence 10

Utilization of 3D for 2D tasks

- Estimating 3D Scene Geometry for 2D Video Object Detection

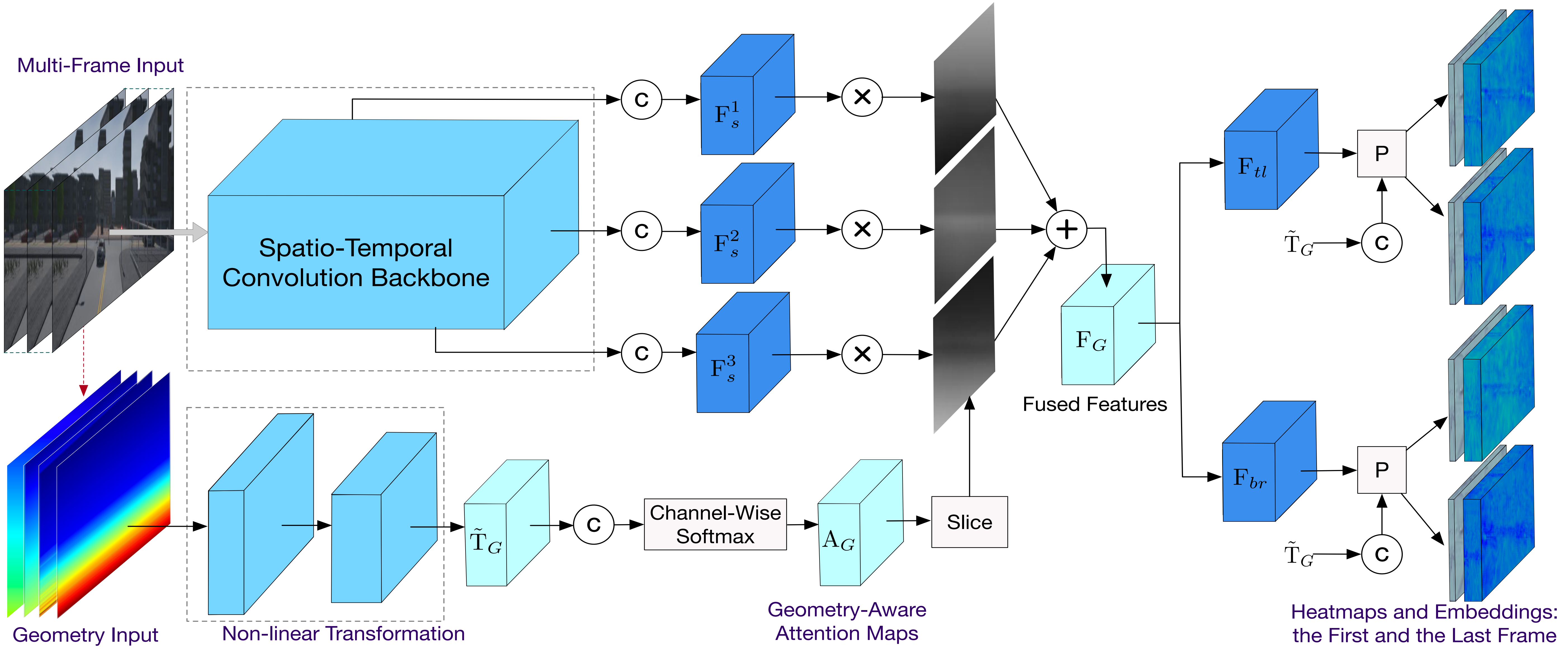
(a) A False Positive Detection Case (b) Height in Pixels of Objects (c) Pseudo Depth Map of Humans

- Geometry (*e.g.* depth) useful for scale ambiguity and occlusion
- Design geometry correlated kernels
- scene geometry can be estimated directly from static cameras, learned from training data
- Geometry-aware feature learning and prediction



Utilization of 3D for 2D tasks

- Estimating 3D Scene Geometry for 2D Video Object Detection



- Achieved significant improvement over one-stage and two-stage video object detectors (Faster RCNN, SSD)

Overview

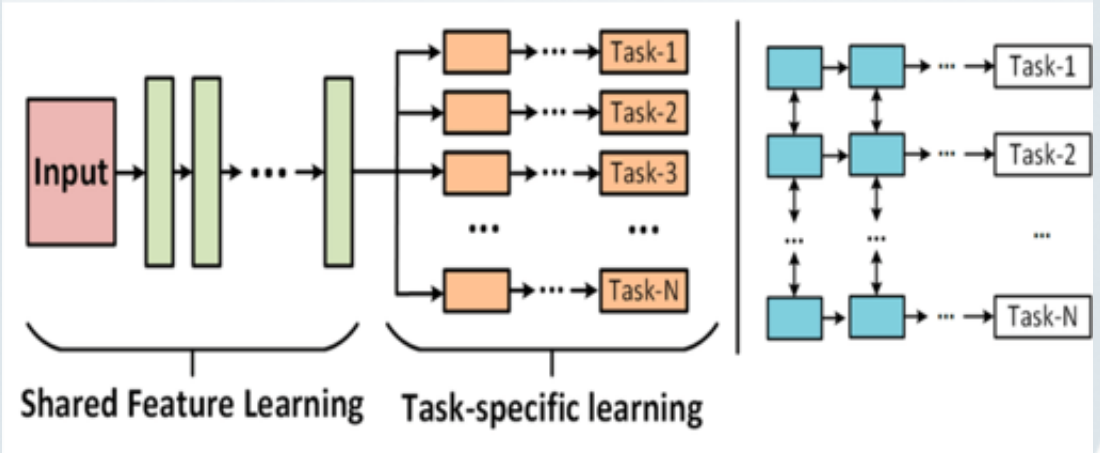
- Scene depth estimation with structured probabilistic modeling
- A joint multi-modal and multi-task deep learning framework
- Modelling the interaction between 2D and 3D data and tasks
- Hot research & development fields along the direction
- Summary

Research Hotspots

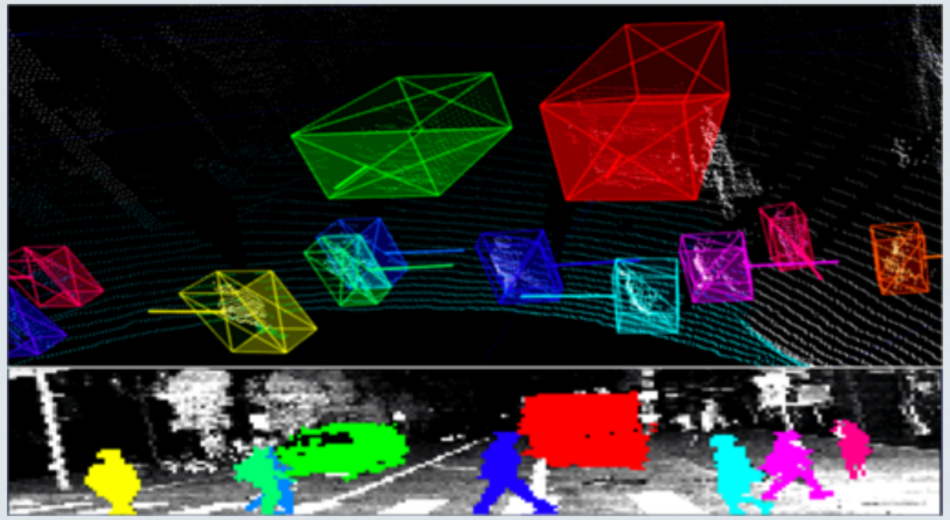
- End-to-End Deep Learning Frameworks and Systems towards Real AI

▪ Statistical Graph Theoretic Framework for Deep Model Design and Explanation

▪ Effective architecture design and learning strategies for deep multi-task learning



▪ High-level scene modelling via complex interaction from 2D & 3D data and tasks



- Develop big application-level systems for realistic large-scale visual scene understanding applications.

Dynamic Graph Network

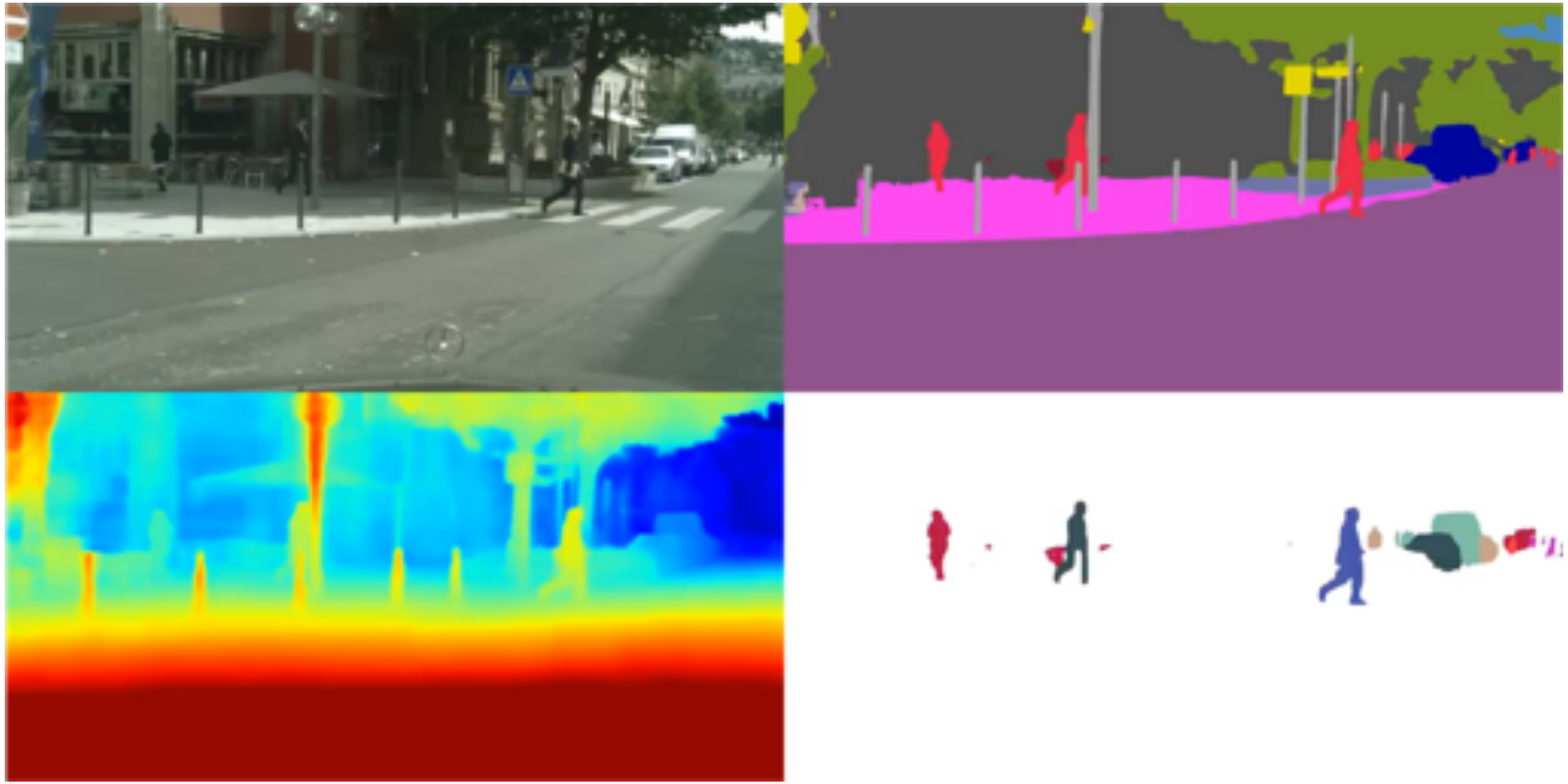
- Statistical Graph Theory Framework Deep Model Design and Explanation

- High efficiency graph deep learning

  - Leaning dynamic graph instead of fully/partially connected static graph
  - Dynamic sampling, dynamic kernels and dynamic affinities

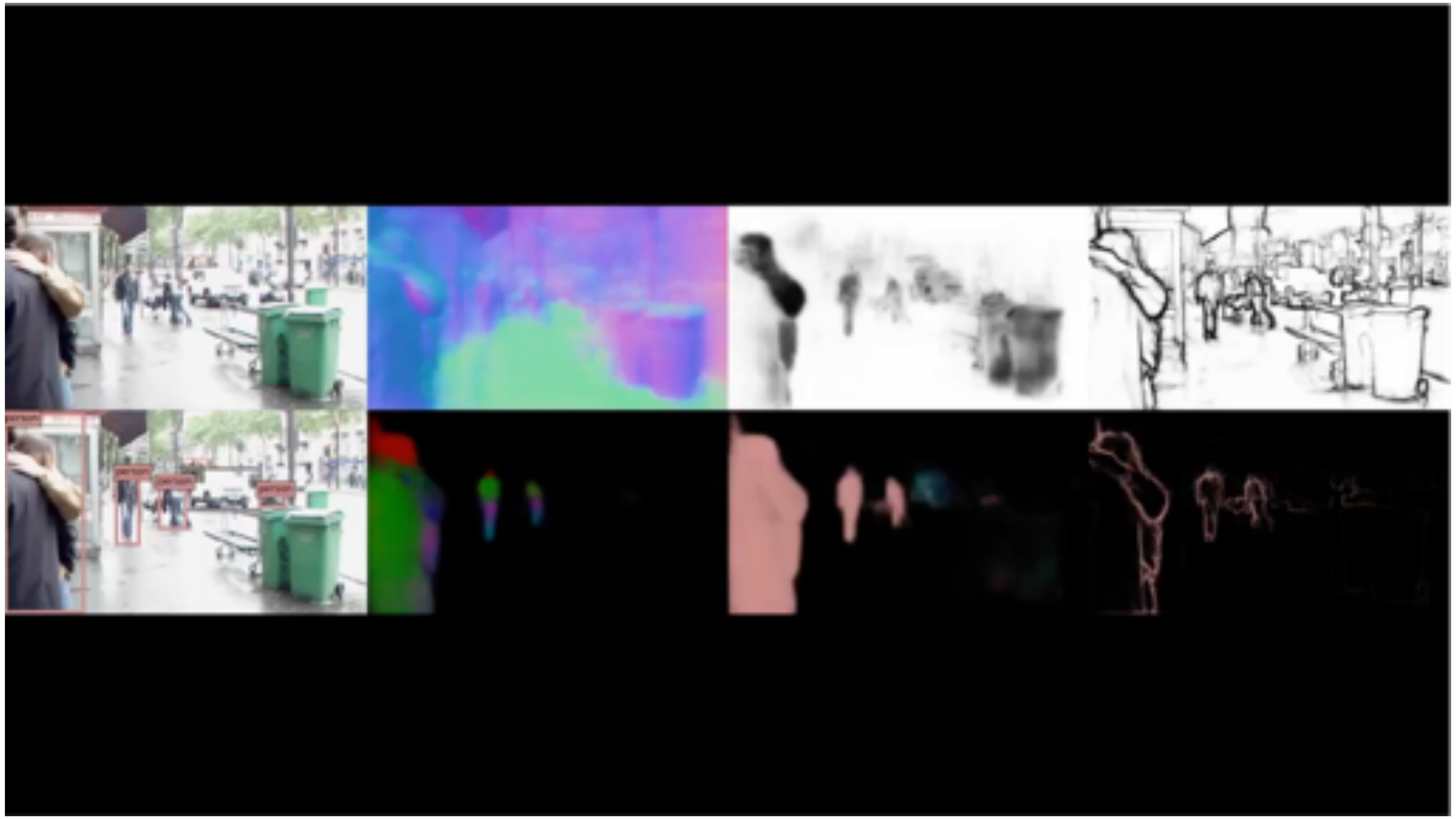
Deep Multi-Task Learning Framework

- Effective Architecture Design and Learning Strategies for Deep Multi-Task Learning
  - Network architecture search for shared and task-specific structure design
  - Exploration of gradient balance and clipping strategies in optimization



Deep Multi-Task Learning Framework

- Effective Architecture Design and Learning Strategies for Deep Multi-Task Learning
- Network architecture search for shared and task-specific structure design
- Exploration of gradient balance and clipping strategies in optimization



End-to-end Deep Visual SLAM

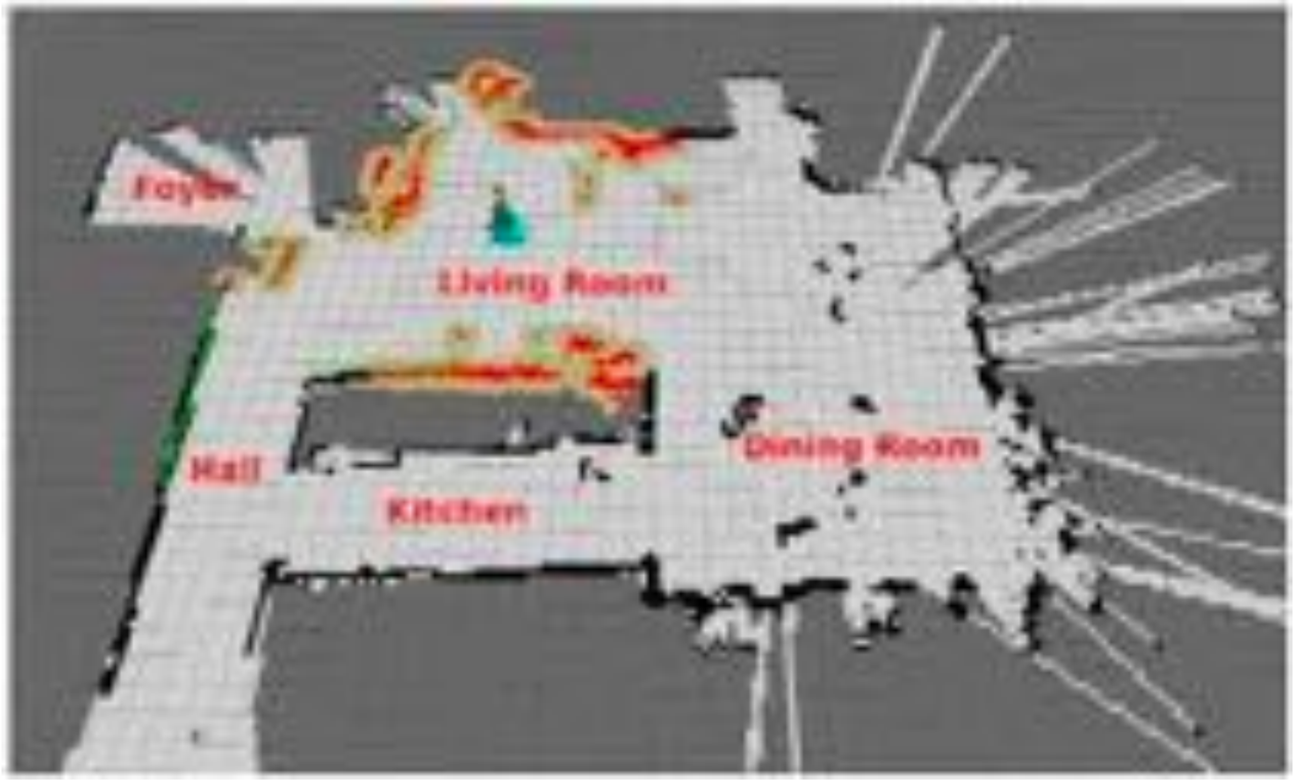
- What is SLAM?

  - Compute the pose of the robot and create a map at the same time

- **Localization:** estimating the robot’s localization

- **Mapping:** building a map

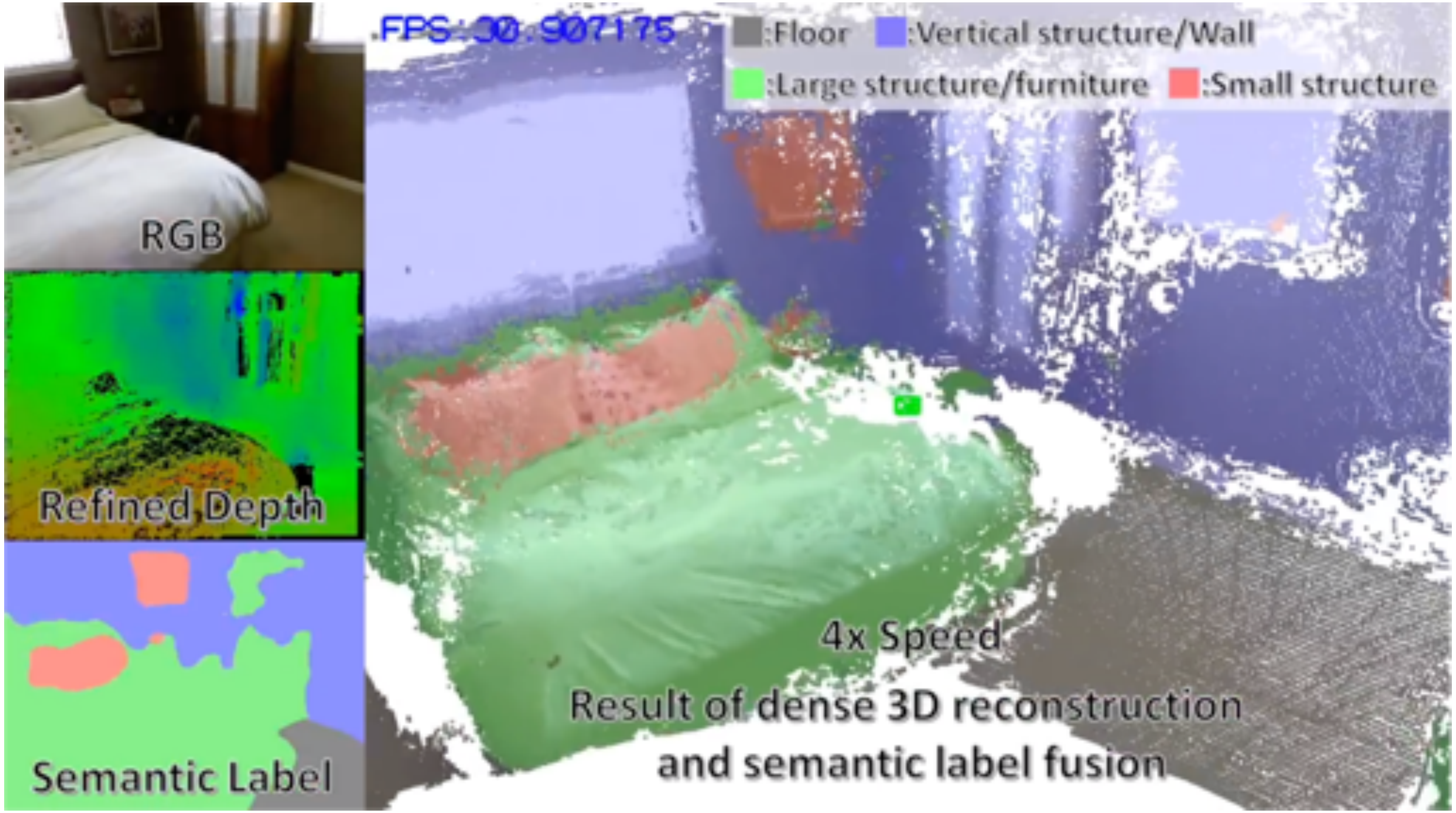
- **SLAM:** simultaneously localizing the robot and building a map



End-to-end Visual SLAM

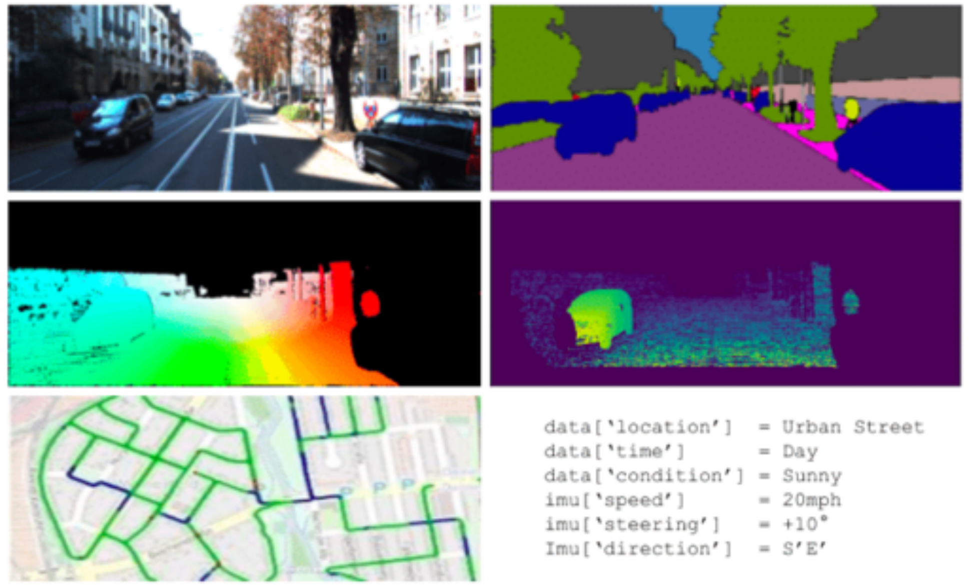
- End-to-end deep learning based visual slam systems

  - Challenges in Key-frame detection, global pose optimization, 3D reconstruction

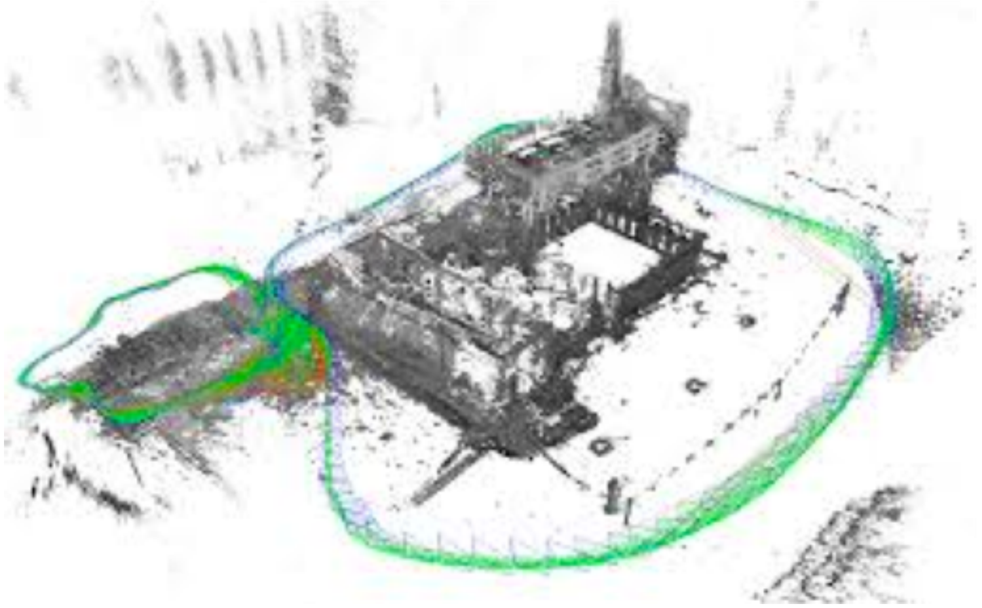


Summary

- Introduced the importance and applications of visual scene understanding
- Introduced an advanced scene depth estimation framework with structured probabilistic modeling
- Described a joint multi-modal deep learning pipeline for simultaneous multi-task inference for complex scene understanding
- End-to-end learning the interaction between 2D and 3D data and tasks
- Hot research trends: graph models for deep learning, effective multi-task deep learning network design, end-to-end visual SLAM system for self-driving and robotics



```
data['location']  = Urban Street
data['time']      = Day
data['condition'] = Sunny
imu['speed']      = 20mph
imu['steering']   = +10°
Imu['direction']  = S'E'
```



Thank you!
Questions?