AI70601/CSE61101

3D Vision and Machine Perception

Prof. Kyungdon Joo

3D Vision & Robotics Lab.

AI Graduate School (AIGS) & Computer Science and Engineering (CSE)

Some materials, figures, and slides (used for this course) are from textbooks, published papers, and other open lectures

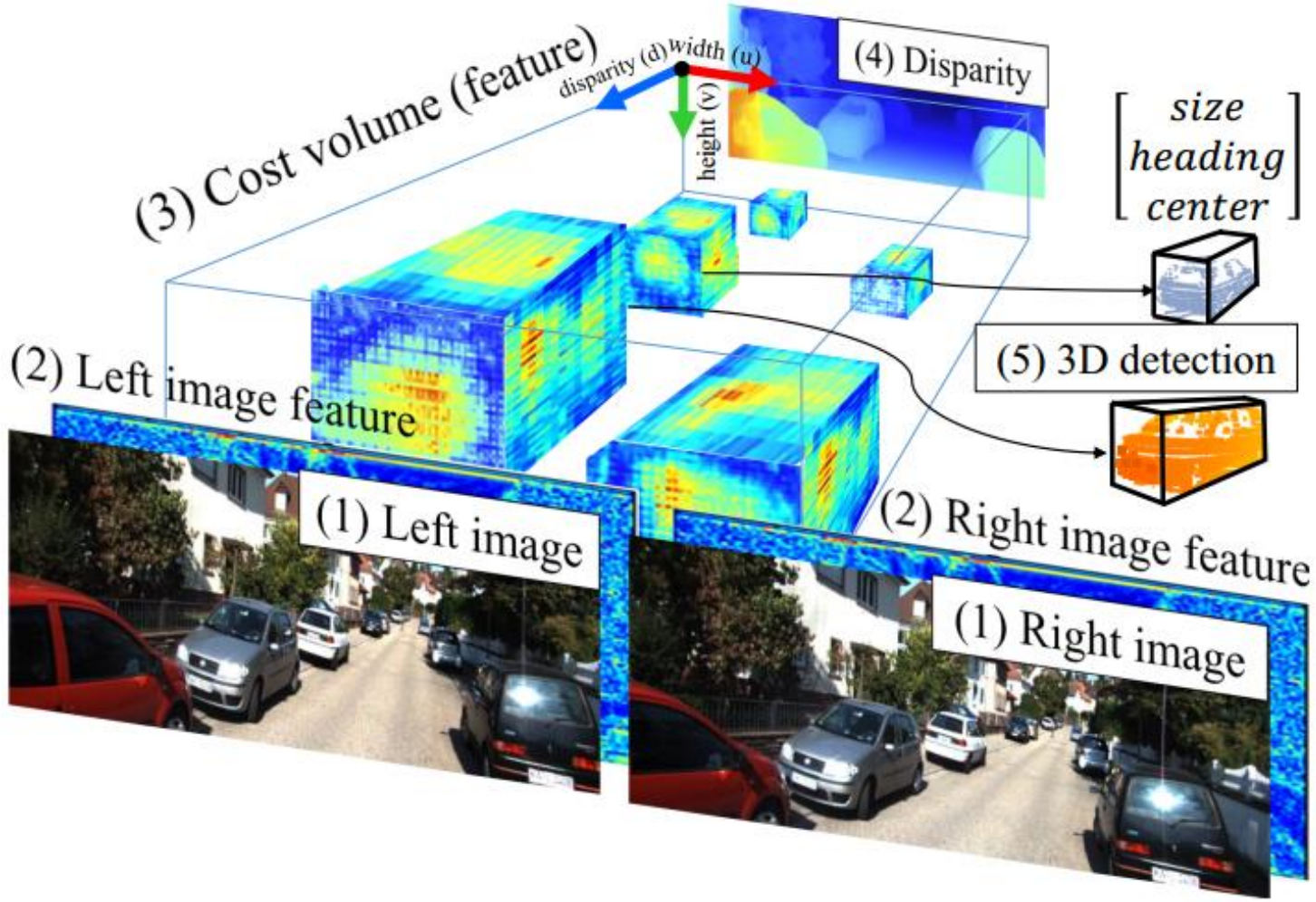
Term Project Evaluation Criteria

- **Completeness of Submission (5%)**
  - Points will be deducted if any of the following items are missing or if the PPT or report lacks effort:
  - Code / PPT / Report

- **Proposed Method (15%)**
  - Contributions: Assess the novelty and suitability for 3D vision.
  - Assess the quality and completeness of the results in relation to the difficulty of implementing the proposed method.

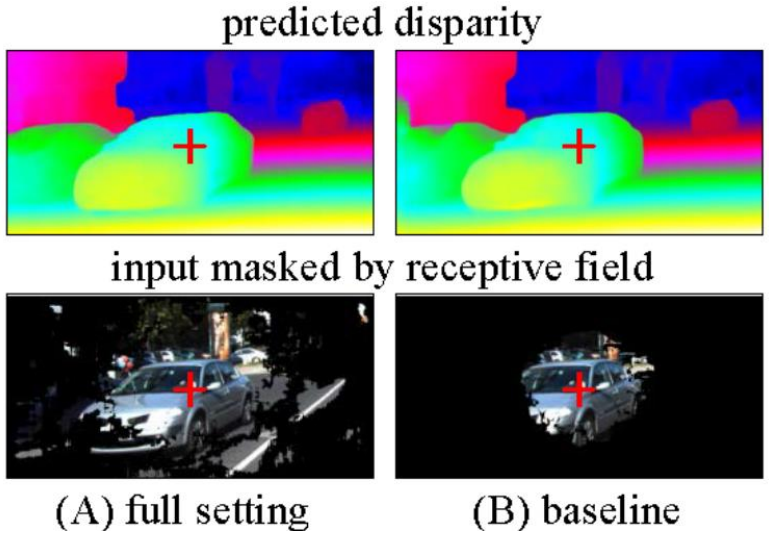
- **Experimental Design and Validation (15%)**
  - Evaluate the impact of the added modules on the overall performance of the system.
  - Assess how well the proposed method is objectively validated through experiments.

Stereo Object Matching Network (ICRA 2021)

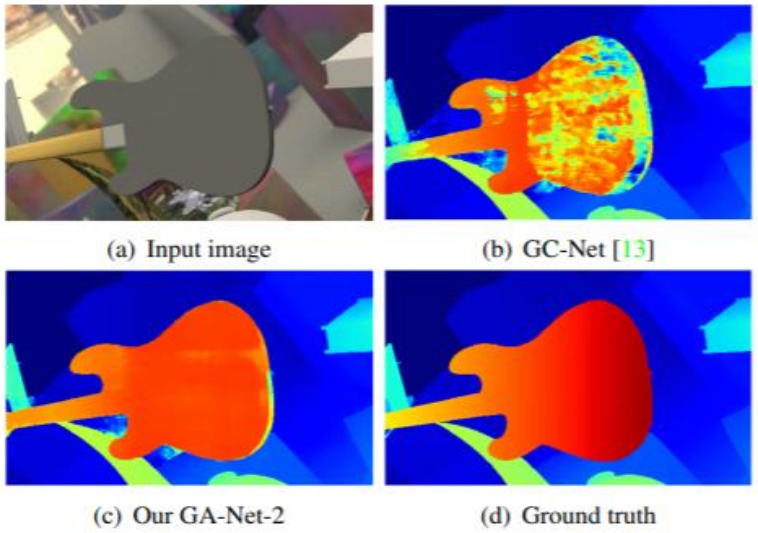


Intro. problem definition

- Limitation / Previous work
  - **Uncertainty** nearby objects' **boundary**.



Chang and Chen [1]



Zhang et al. [2]

→ *depend on only 2D information*

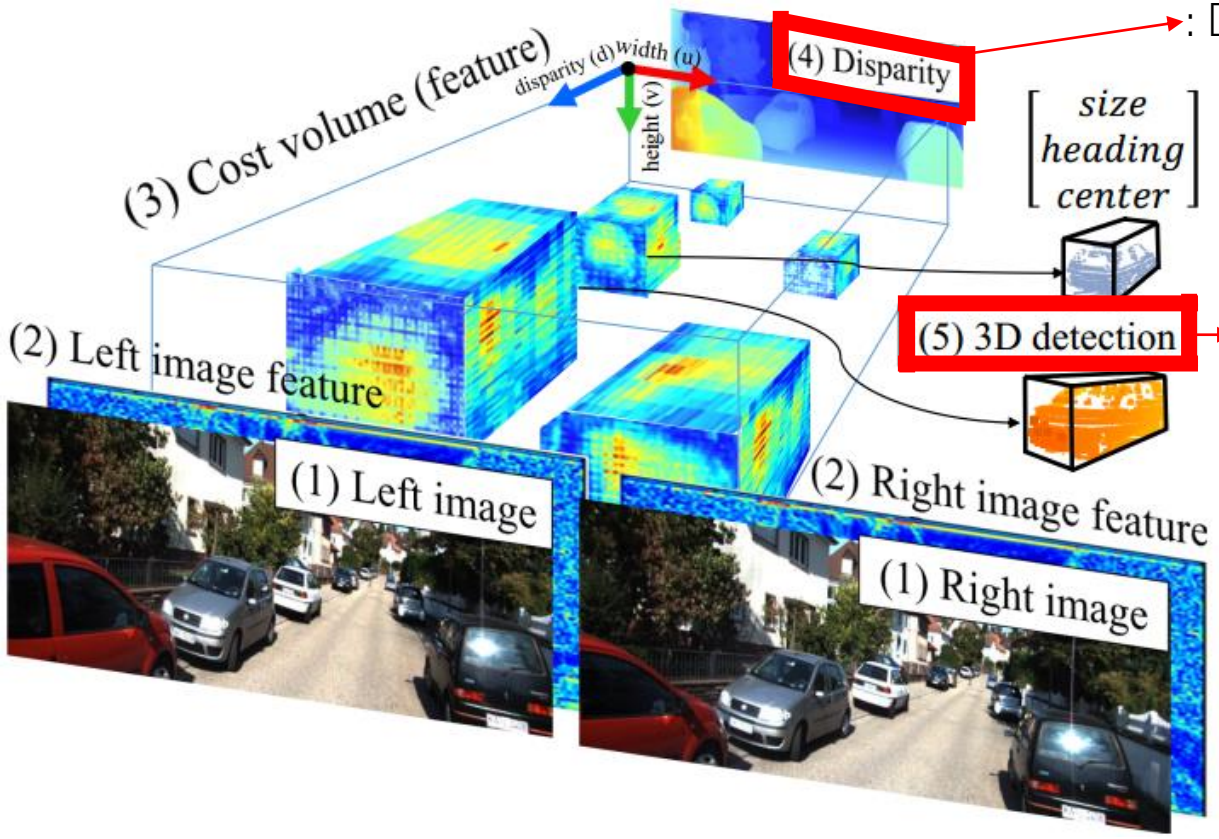
→ *Texture-less, repetitive patterns*

[1] Chang and Cheng., "Pyramid stereo matching network." *CVPR*. 2018.
[2] Zhang et al. "Ga-net: Guided aggregation net for end-to-end stereo matching." *CVPR*. 2019.

Intro. Stereo Object Matching Network

- What is stereo object matching?

Stereo matching + 3D object detection → Instance(object)-aware stereo matching.

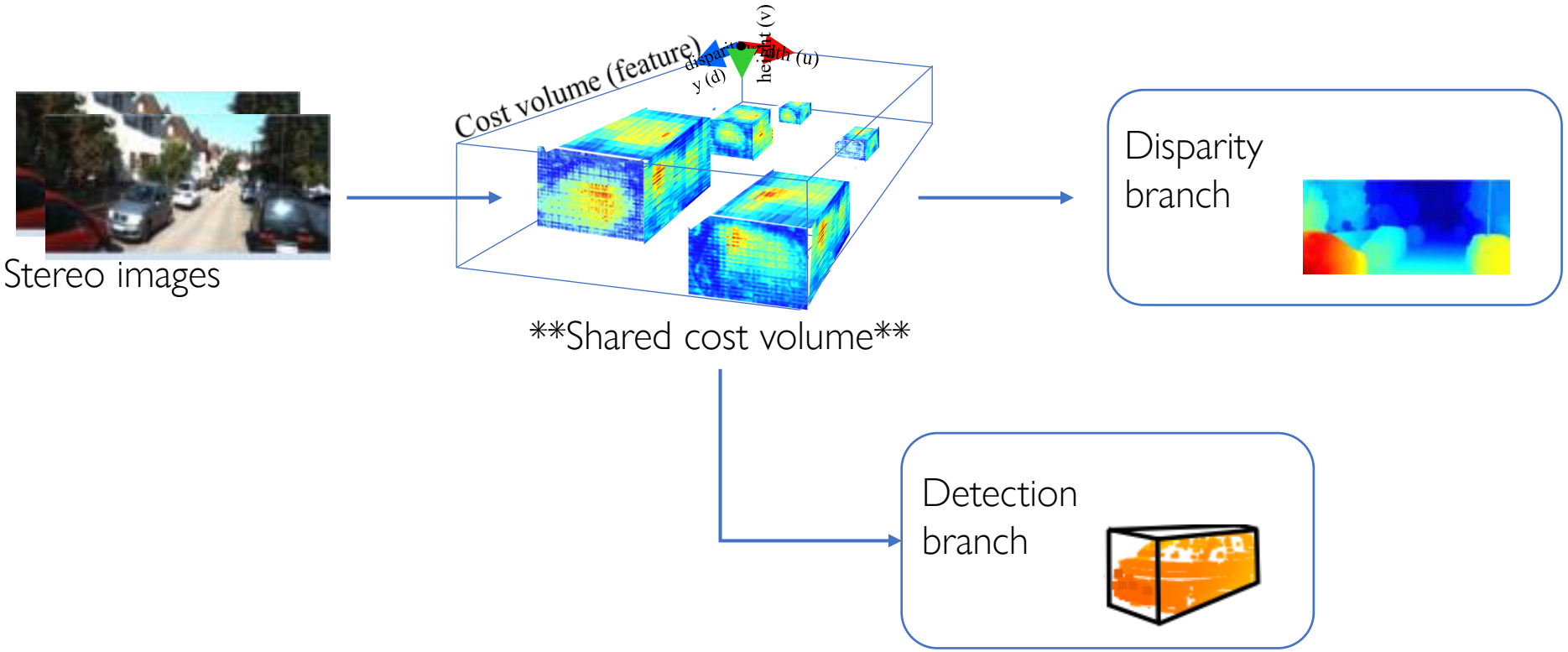


: Disparity estimation via cost volume.

: Cost volume based 3d object detection.

Main. Overview

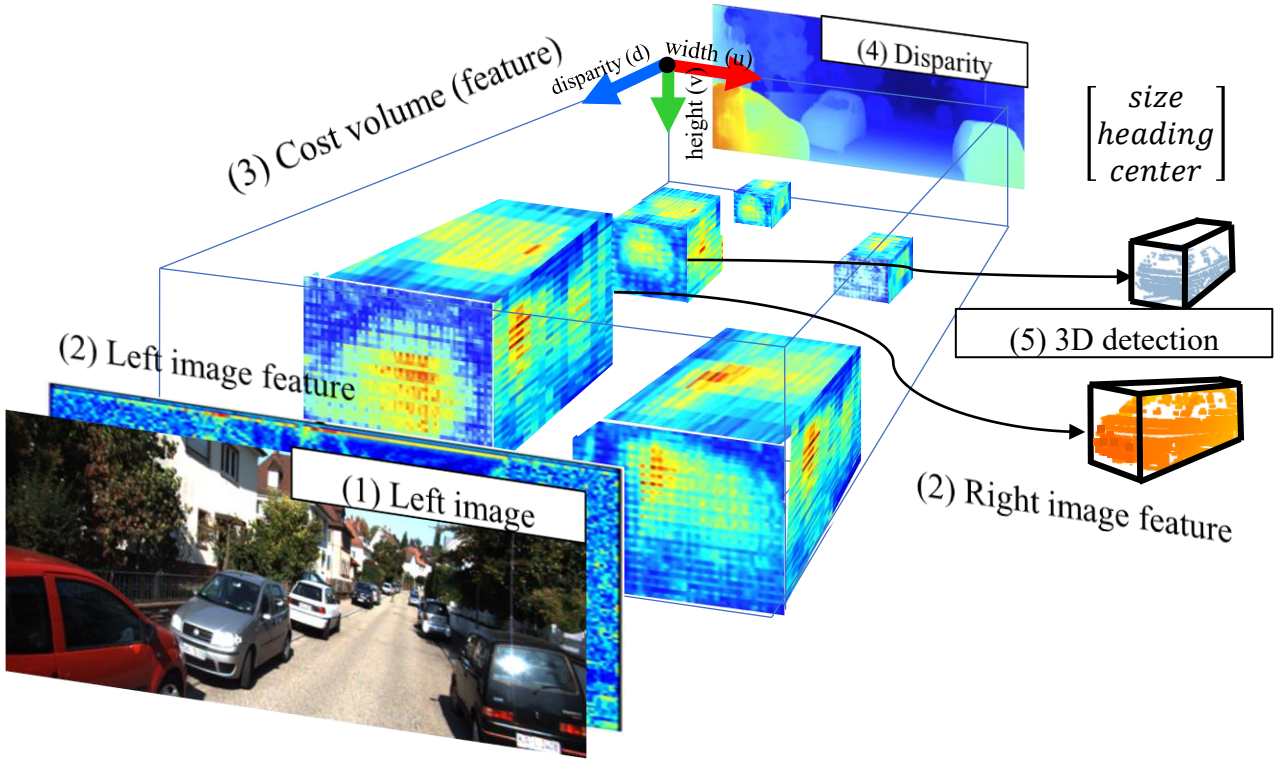
- How to combine stereo matching and object detection?



Disparity branch and detection branch share cost volume as backbone features.

Main. Stereo Object Matching Network

- Revisited differentiable cost volume [4].

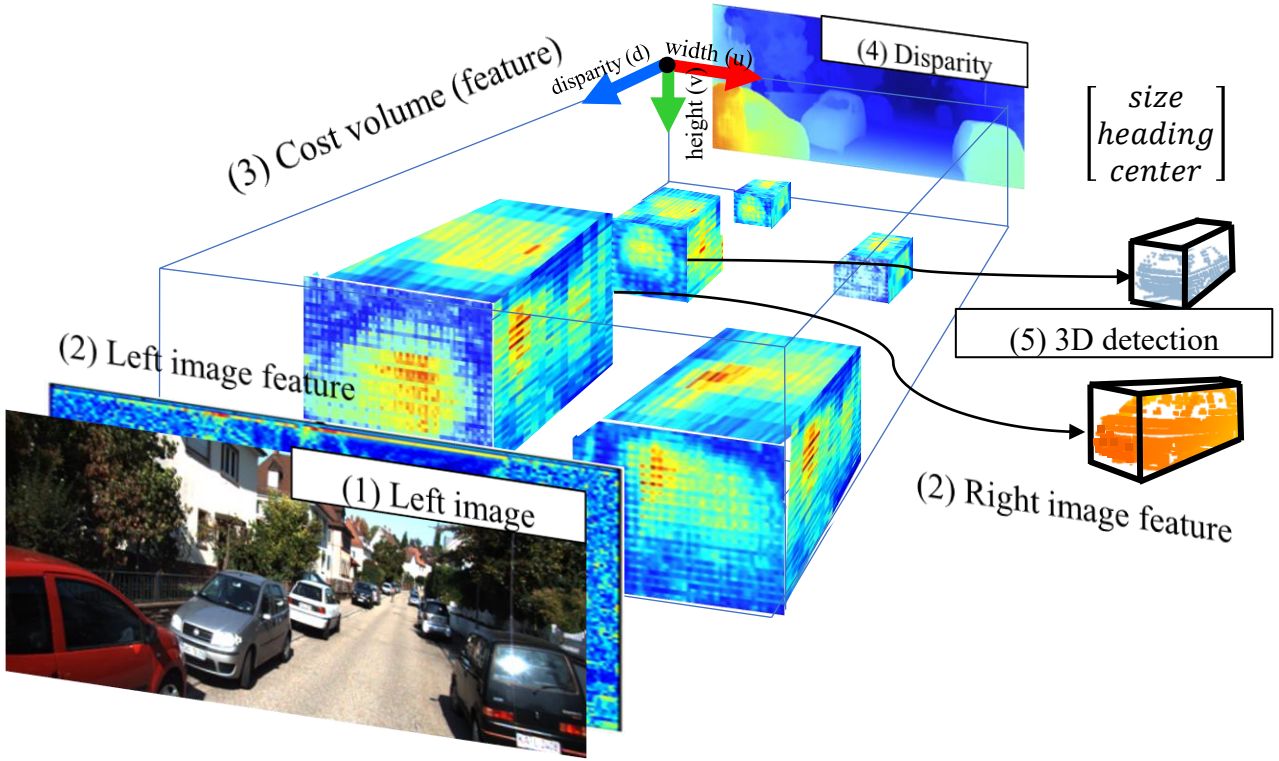


- Cost volume is the volumetric representations of camera rays.
  - Original usage is for disparity estimation [4].

[4] Kendall et al. "End-to-end learning of geometry and context for deep stereo regression." *CVPR*. 2017.

Main. Stereo Object Matching Network

- Revisited differentiable cost volume [4].



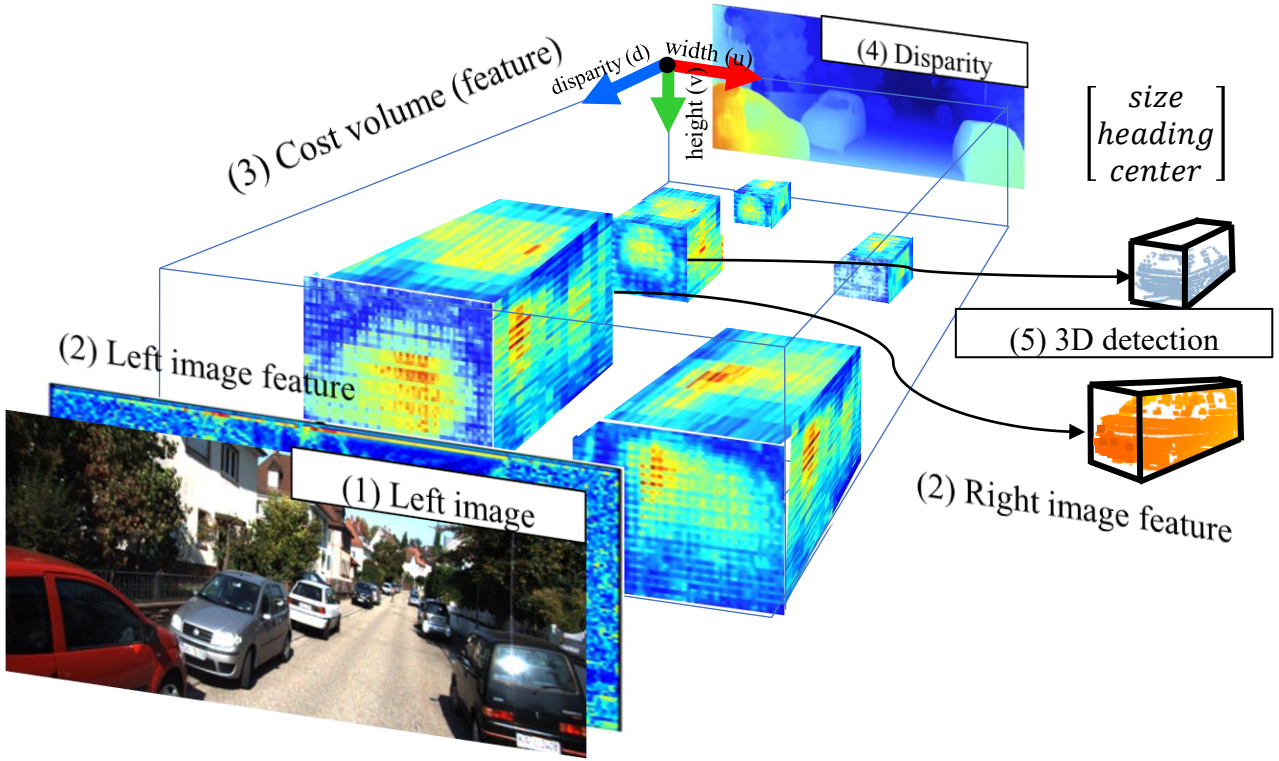
- Cost volume is the volumetric representations of camera rays.
  - Original usage is for disparity estimation [4].

- Accordingly, parts of the cost volume correspond to 3D vehicles' information, such as heading or positions.

[4] Kendall et al. "End-to-end learning of geometry and context for deep stereo regression." *CVPR*. 2017.

Main. Stereo Object Matching Network

- Revisited differentiable cost volume [4].

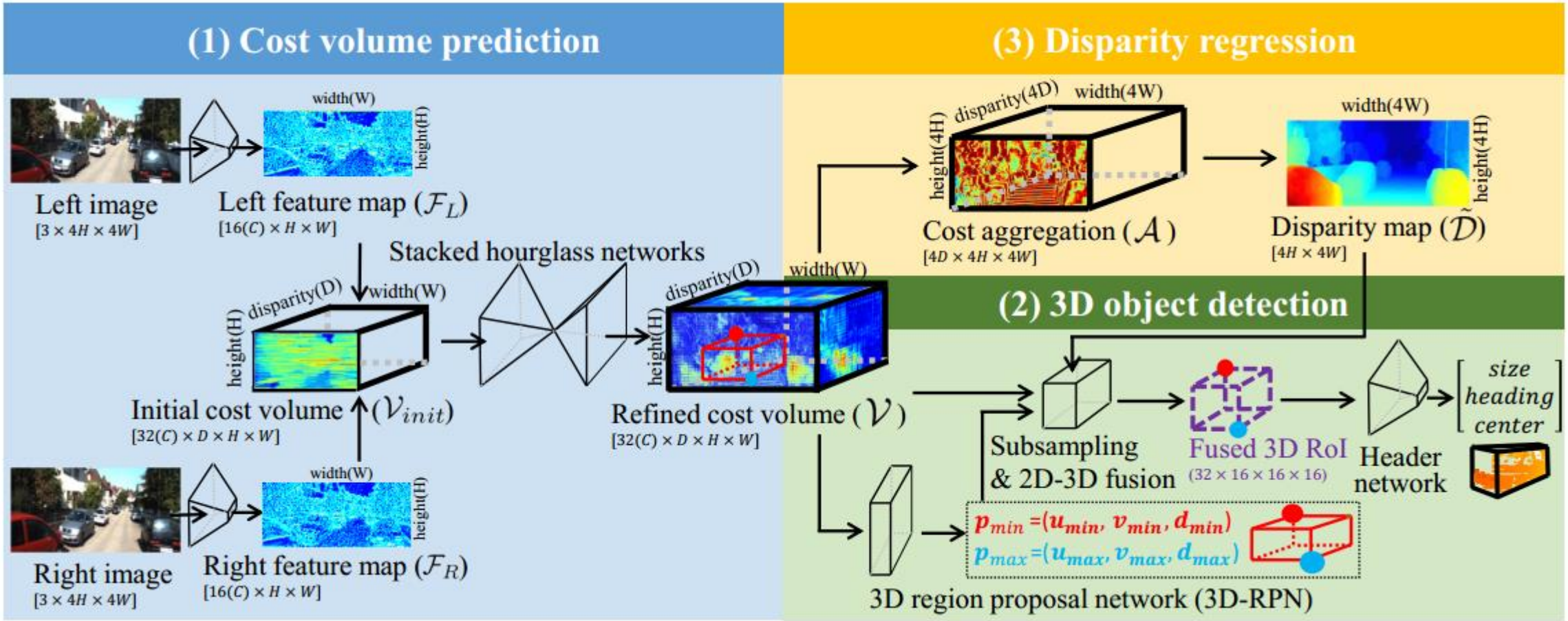


- Cost volume is the volumetric representations of camera rays.
  - Original usage is for disparity estimation [4].
- Accordingly, parts of the cost volume correspond to 3D vehicles' information, such as heading or positions.
- However, 3D attention in cost volume is only valid for the surface of vehicles, not invisible voxels.
  - **Requires additional processes** to properly link cost volume to detection branch.

[4] Kendall et al. "End-to-end learning of geometry and context for deep stereo regression." *CVPR*. 2017.

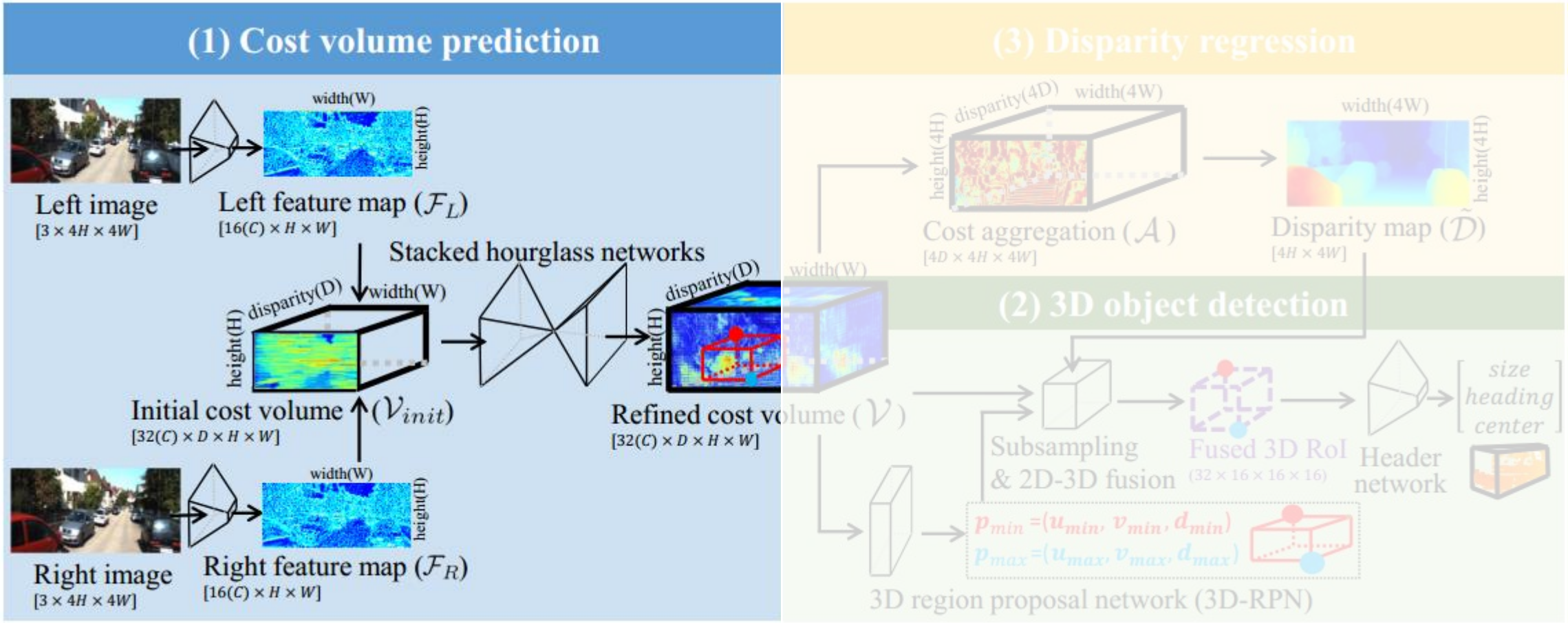
Main. Stereo Object Matching Network

- Architecture



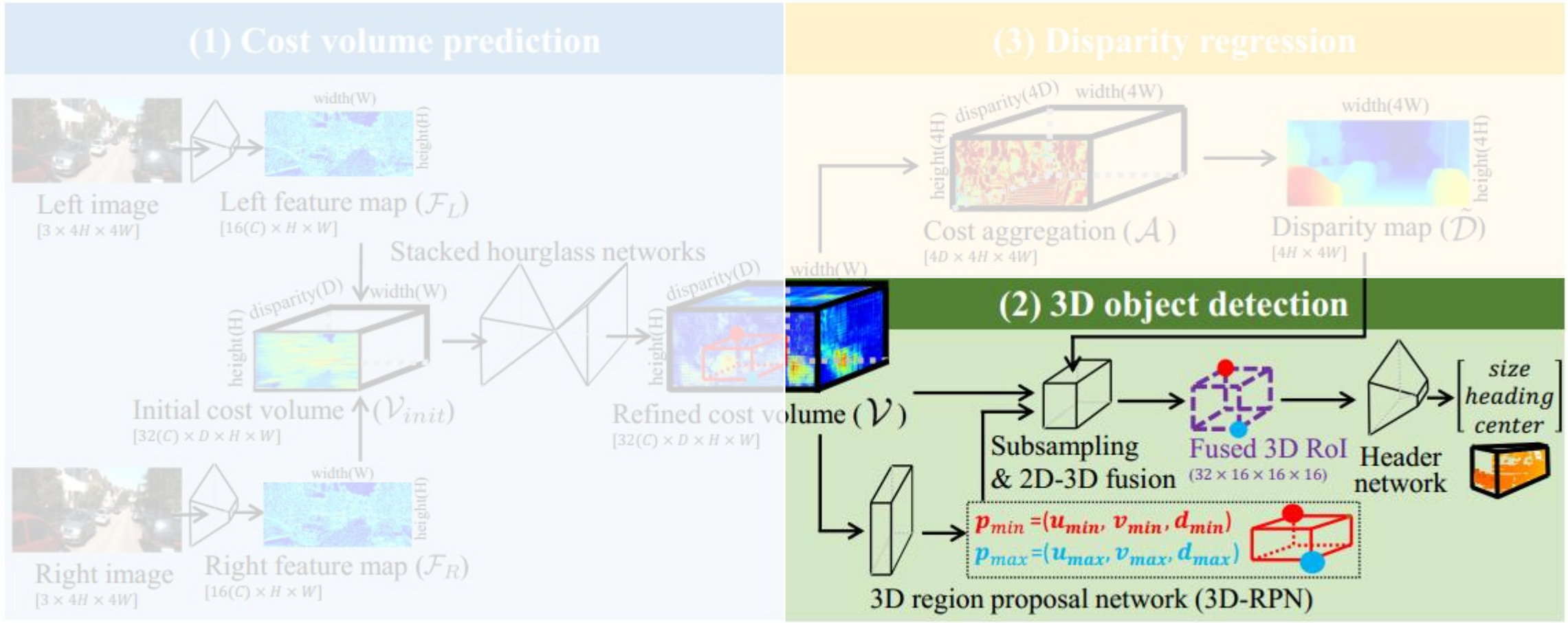
Main. Stereo Object Matching Network

- Architecture



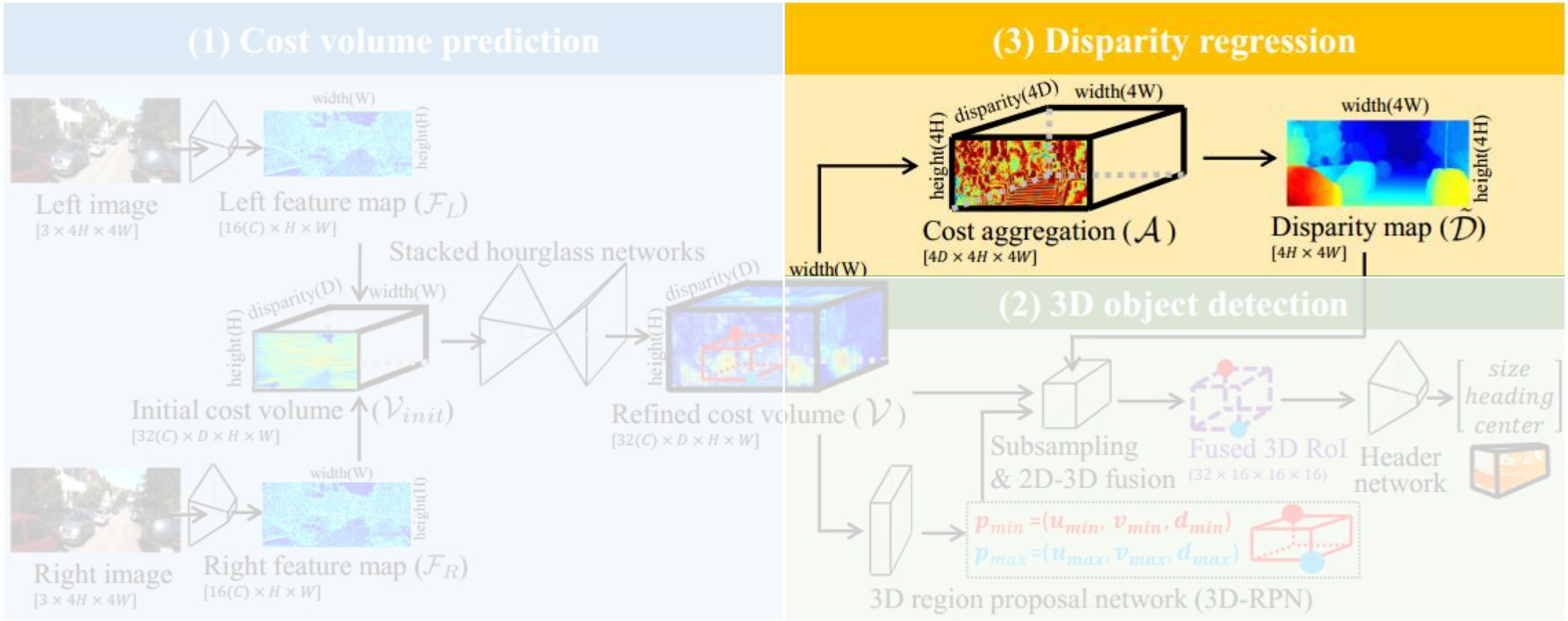
Main. Stereo Object Matching Network

- Architecture



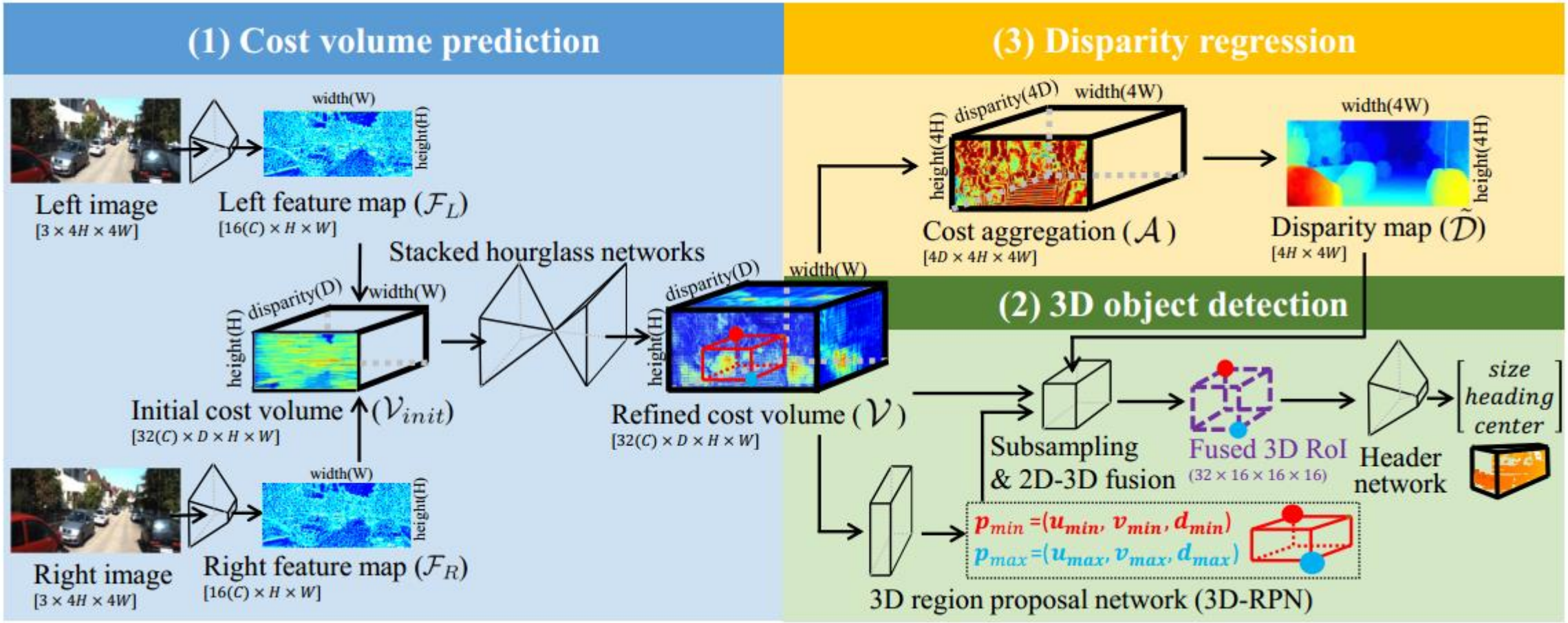
Main. Stereo Object Matching Network

- Architecture



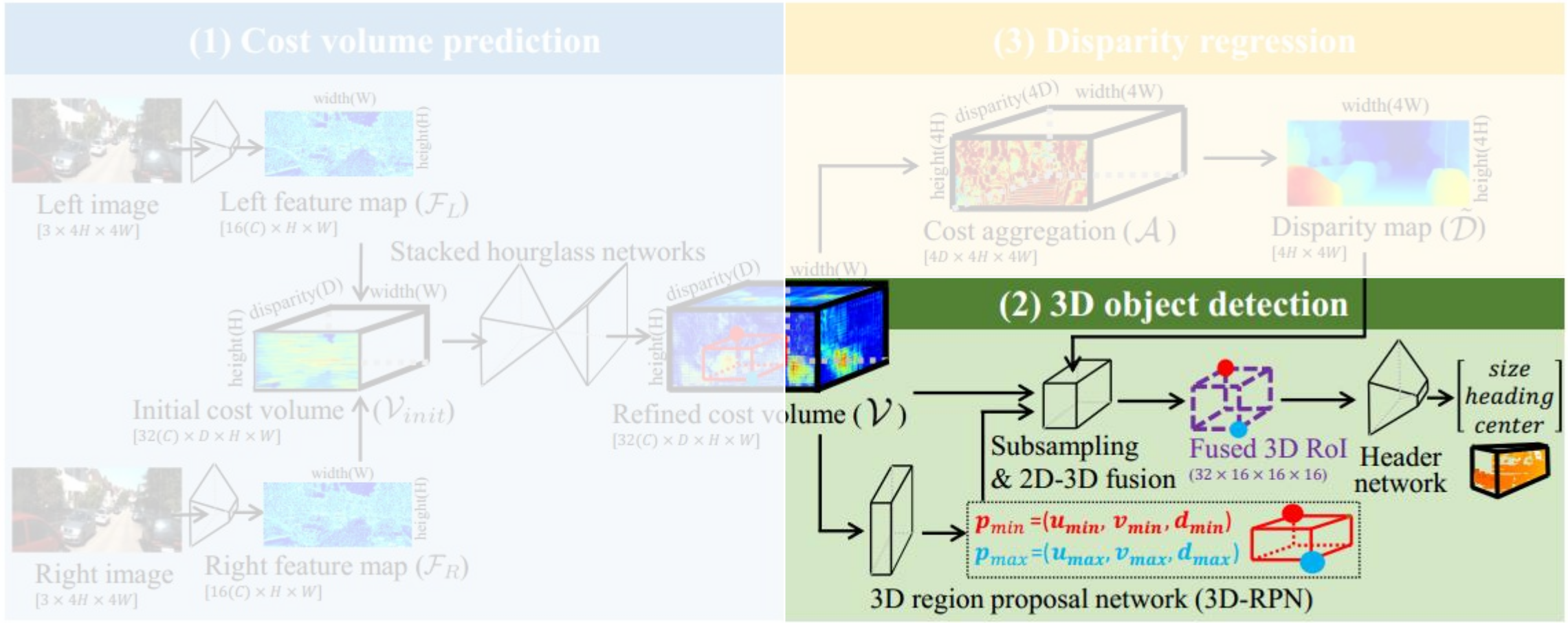
Main. Stereo Object Matching Network

- Architecture



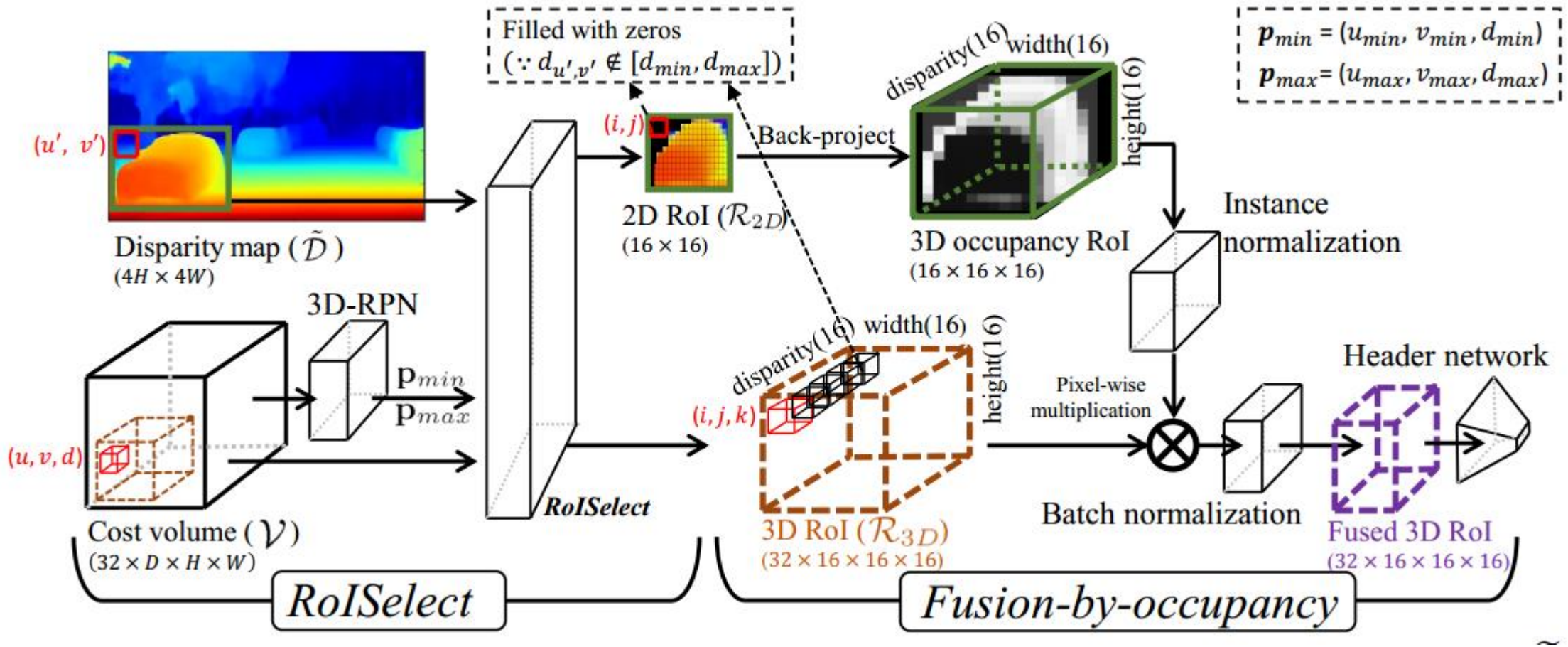
Main. Stereo Object Matching Network

- Architecture



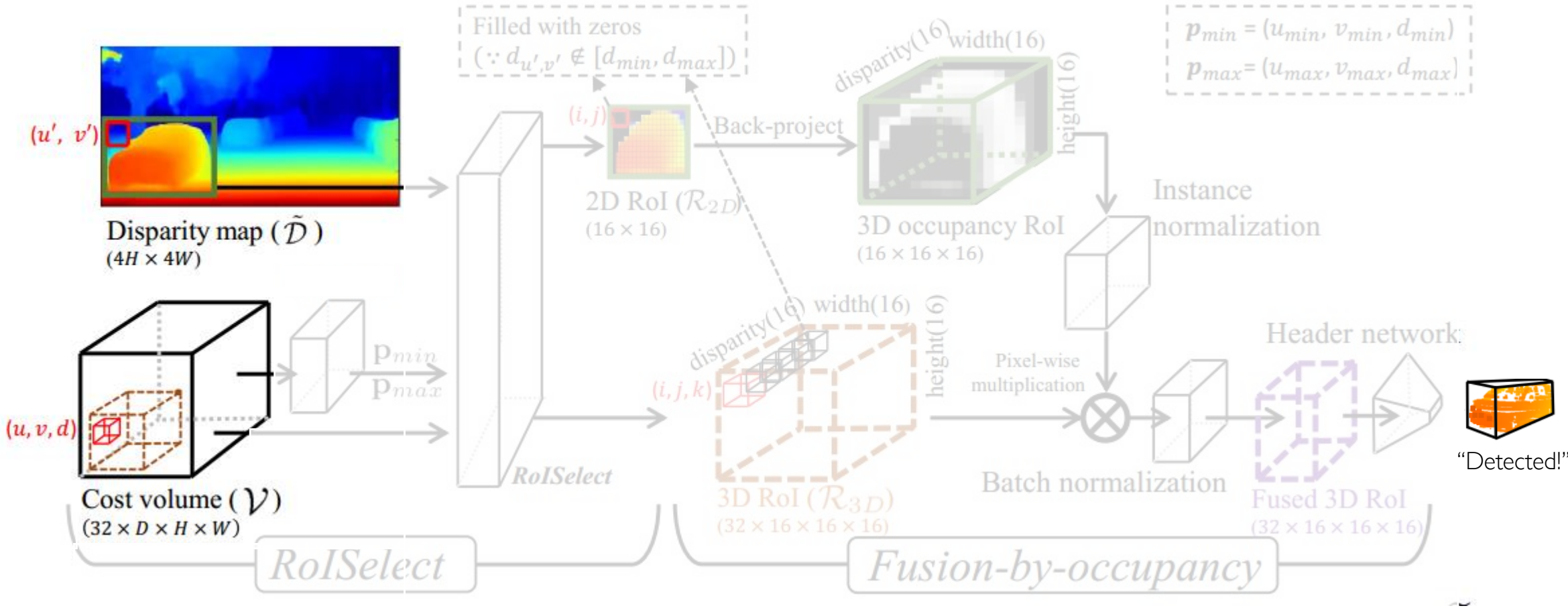
Main. Stereo Object Matching Network

- How to link cost volume and detection branch?



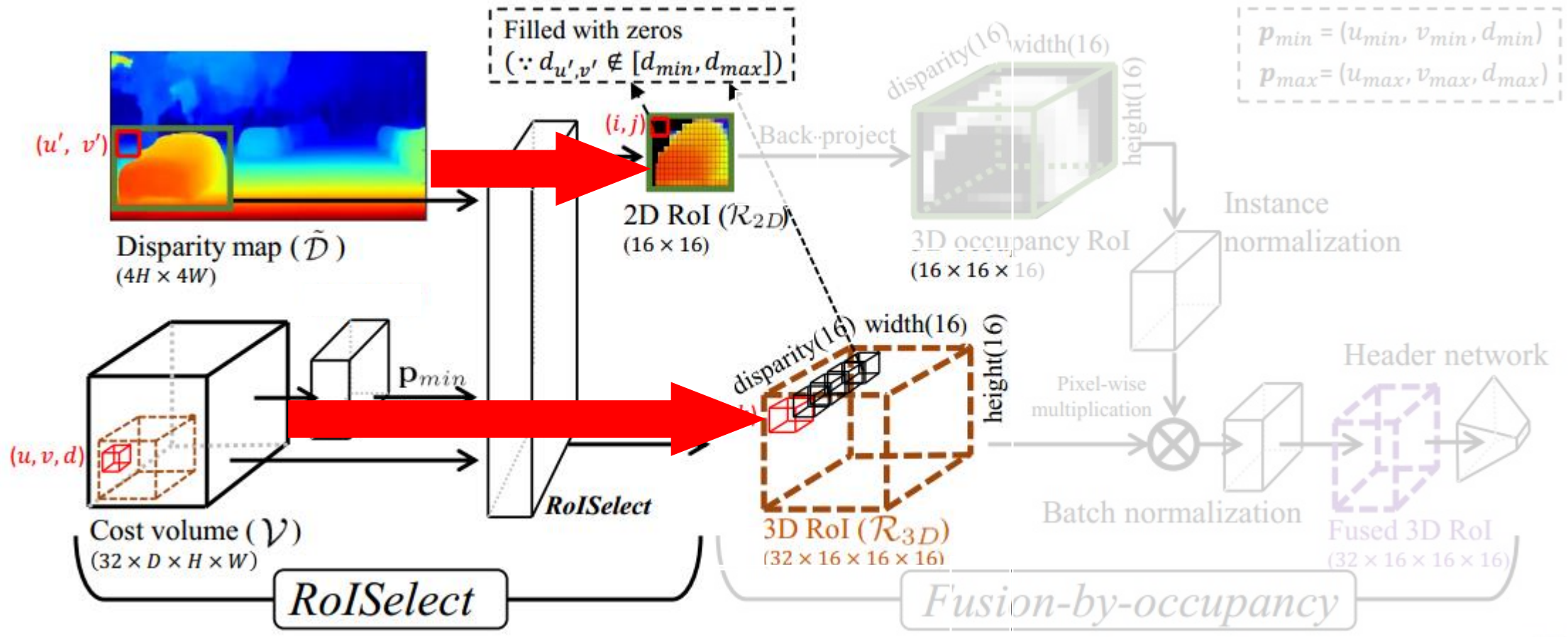
Main. Stereo Object Matching Network

- How to link cost volume and detection branch?



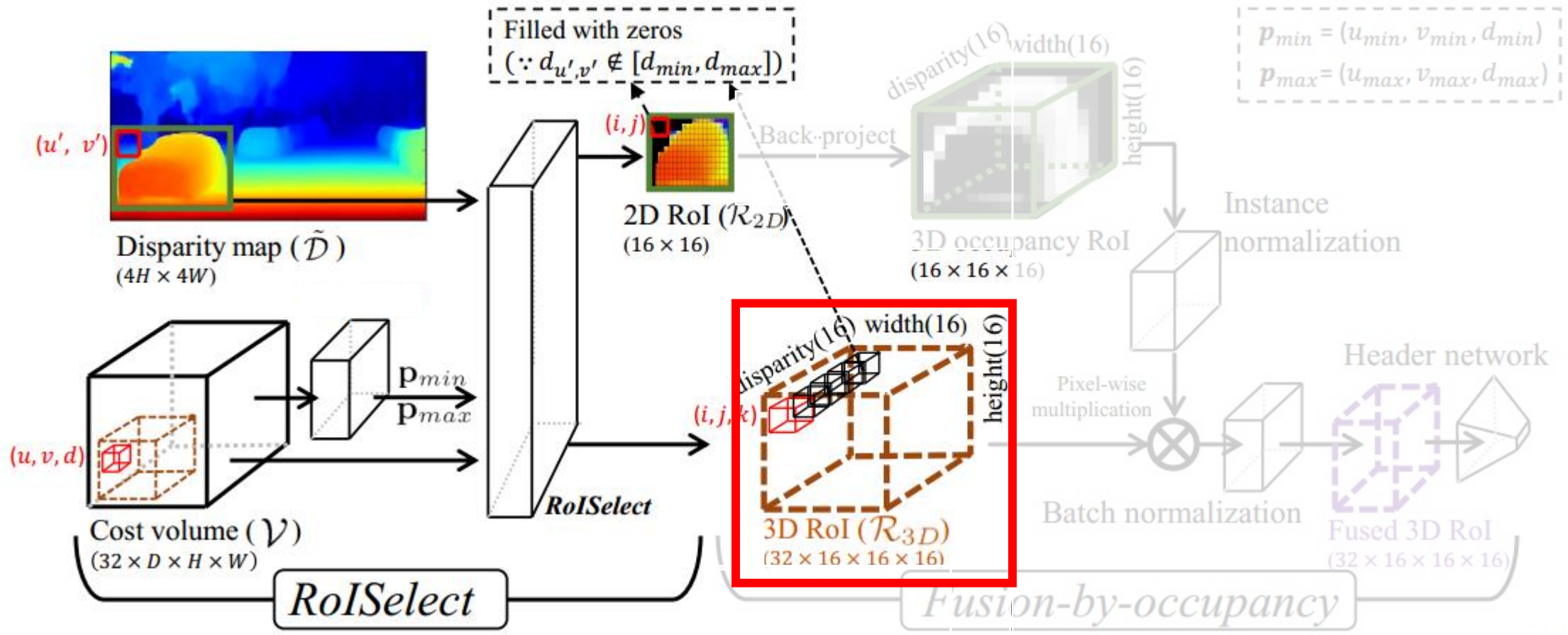
Main. Stereo Object Matching Network

- How to link cost volume and detection branch?



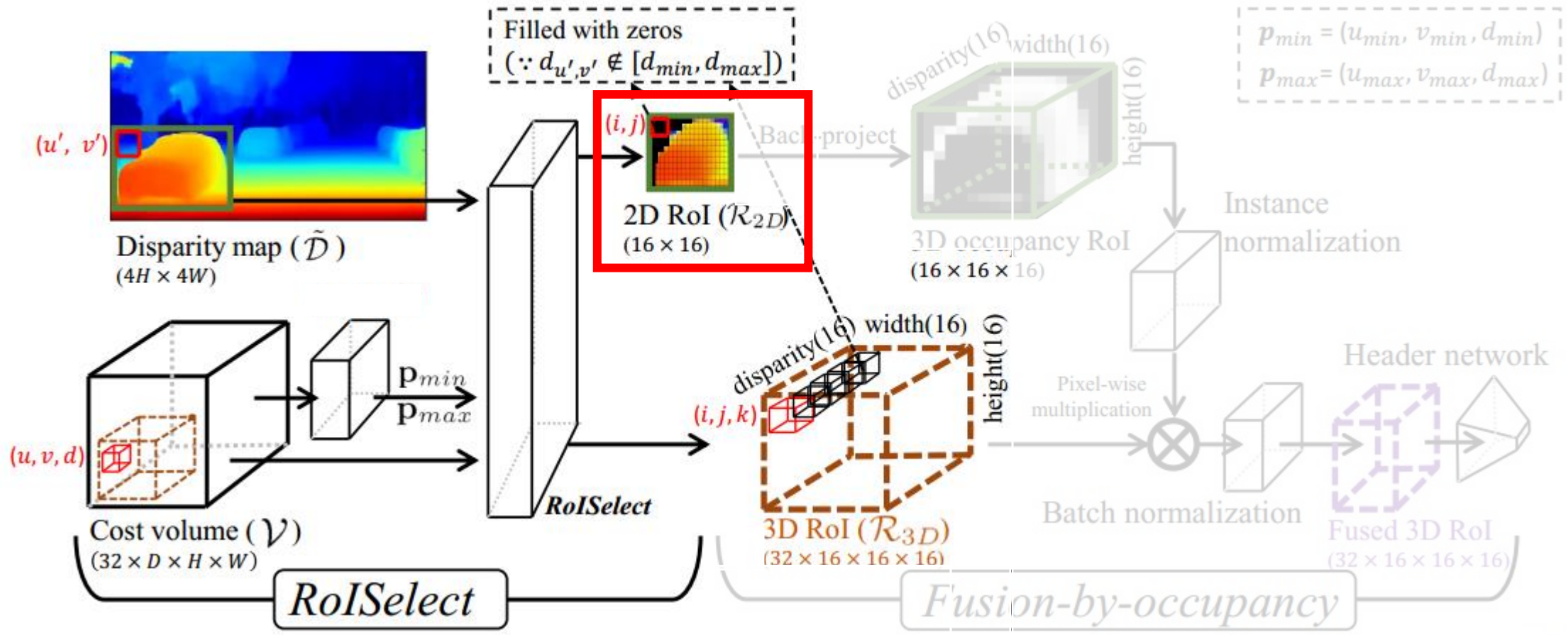
Main. Stereo Object Matching Network

- How to link cost volume and detection branch?



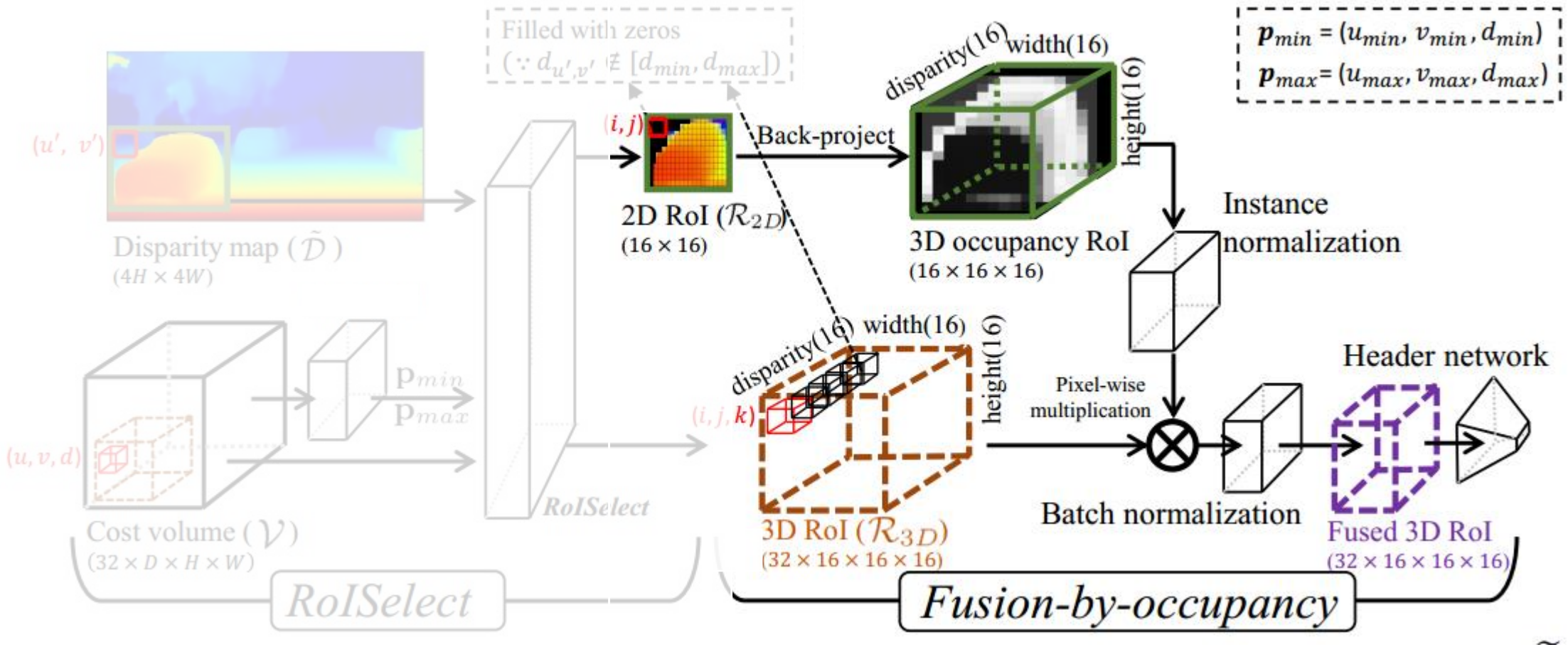
Main. Stereo Object Matching Network

- How to link cost volume and detection branch?



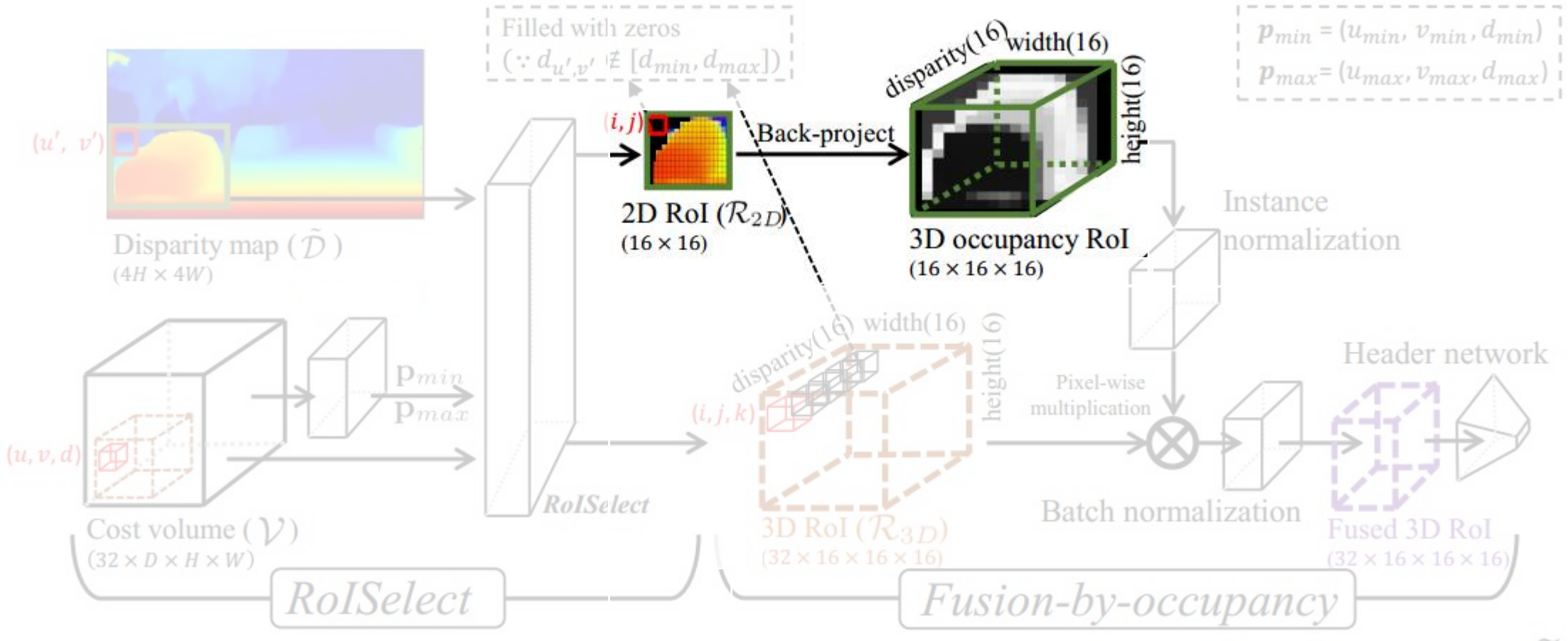
Main. Stereo Object Matching Network

- How to link cost volume and detection branch?



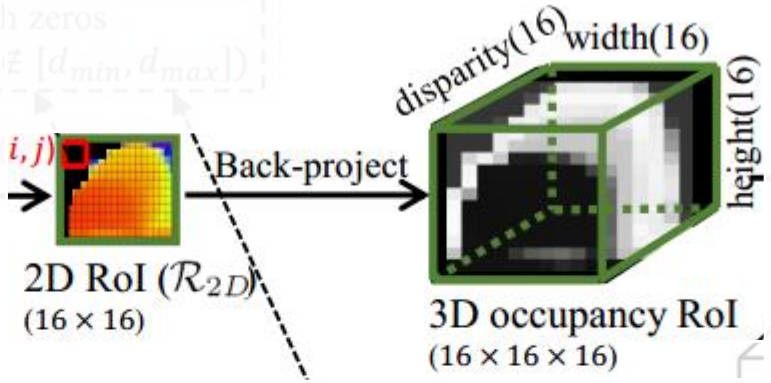
Main. Stereo Object Matching Network

- How to link cost volume and detection branch?



Main. Stereo Object Matching Network

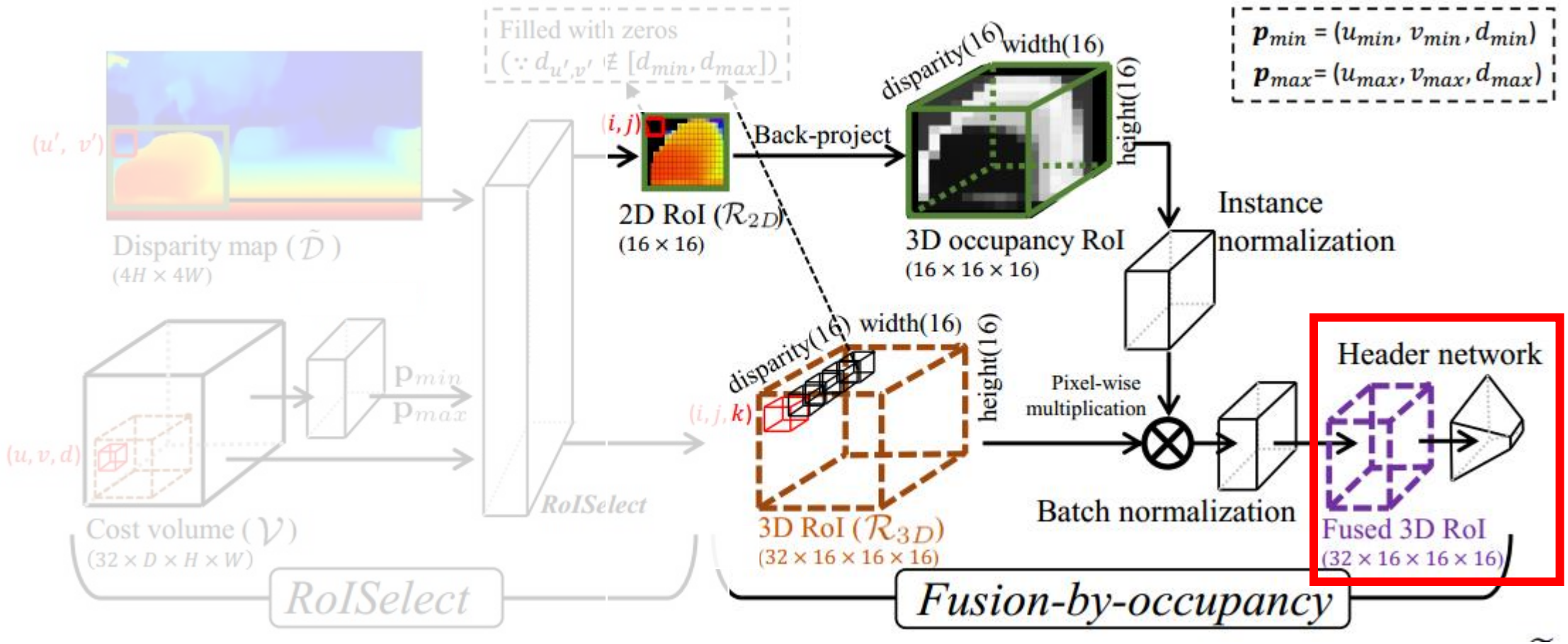
- How to link cost volume and detection branch?



The purpose of 3D occupancy RoI is to **mask voxels inside of vehicles.**
Thus, detection branch only affects surface of vehicles
and **refines 3D surface attention within cost volume while training.**

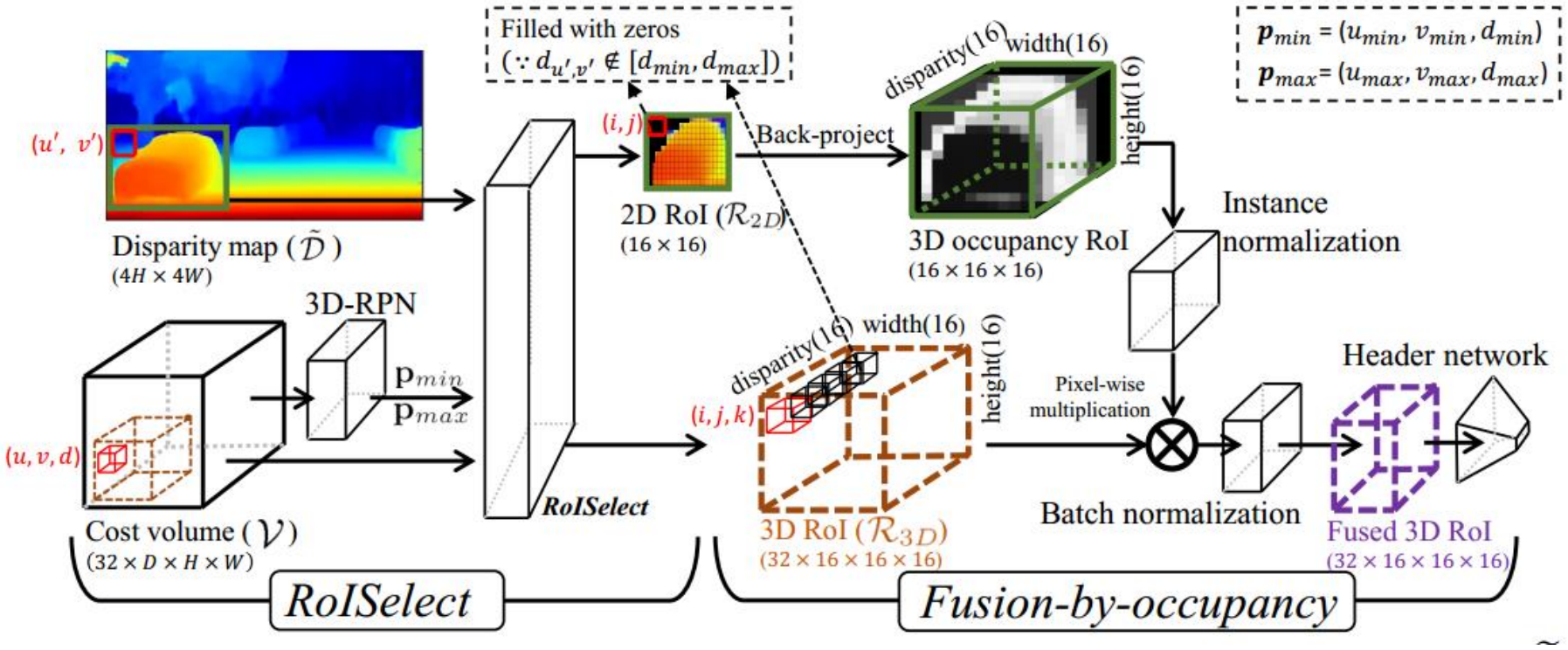
Main. Stereo Object Matching Network

- How to link cost volume and detection branch?



Main. Stereo Object Matching Network

- How to link cost volume and detection branch?



Message: extract vehicles' surface attention from cost volume
for attention refinement of cost volume during training session.

Results. Stereo Object Matching Network

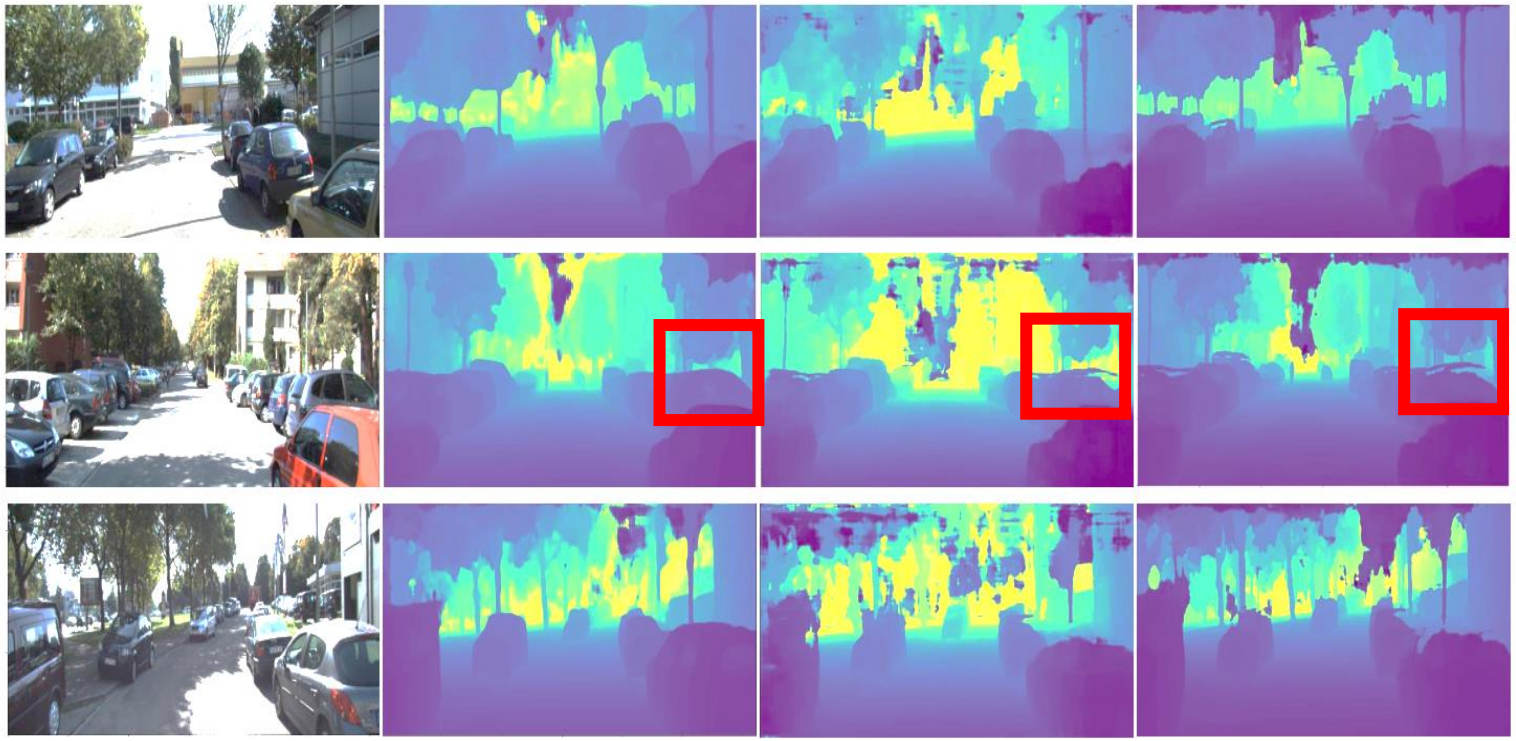
- Depth comparison

Method	Modality	Depth Evaluation (Lower the better)		
		Abs Rel	Sq Rel	RMSE
Eigen [35]	M	0.203	1.548	6.307
DORN [36]	M	0.072	0.307	2.727
PSMnet [16]*	S	0.053	0.234	2.847
PseudoLiDAR [7]	S	0.052	0.281	3.027
DSGN [6]	S	0.064	0.184	2.942
Ours	S	**0.027**	**0.111**	**1.842**

Table 1. Depth accuracy in KITTI Eigen split benchmark.
Note that M and S represent monocular cameras and stereo camera, respectively.

Method	Modality	Depth Evaluation (Lower the better)		
		Abs Rel	Sq Rel	RMSE
PSMnet [16]*	S	0.050	0.517	4.909
PseudoLiDAR [7]	S	0.080	1.044	5.137
DSGN [6]	S	0.059	**0.503**	4.748
Ours	S	**0.044**	0.557	**4.619**

Table 2. Depth accuracy in Virtual KITTI benchmark.



[6] Chen et al. "Dsgn: Deep stereo geometry network for 3d object detection." CVPR. 2020.
[7] Wang et al. "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving." CVPR. 2019.

Results. Stereo Object Matching Network

- Detection comparison.

Method	Modality	AP_{2D} (Moderate)	AP_{BEV} (Moderate)	AP_{3D} (Moderate)
Multi-Fusion [6]	M	-	13.63	5.69
M3D-RPN [38]	M	83.67	21.18	17.06
3DOP [39]	S	88.07	9.49	5.07
TLnet [40]	S	-	21.88	14.26
StereoRCNN [27]	S	**88.27**	48.30	36.69
PseudoLiDAR [7]	S	-	56.80	45.30
DSGN [6]	S	83.59	**63.91**	**54.27**
Ours	S	76.71	30.92	19.88

Ours: Cost volume (pixel unit)
PseudoLiDAR: pointclouds (metric unit)
DGSN: Geometry volume (metric unit)

Table3. Detection results in KITTI detection validation benchmark.

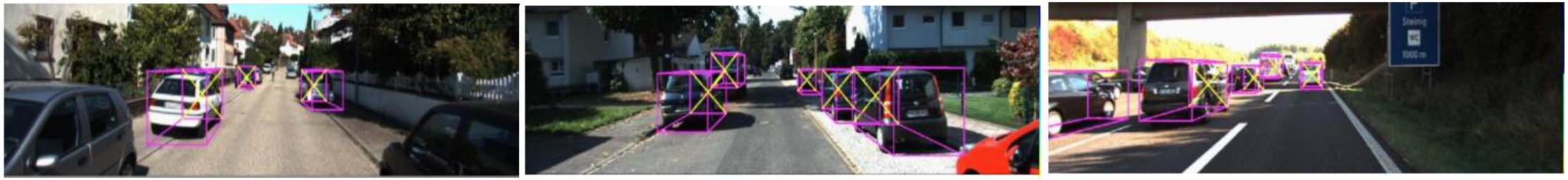


Fig3. Qualitative results from our stereo object matching network.

[6] Chen et al. "Dsgn: Deep stereo geometry network for 3d object detection." CVPR. 2020.
[7] Wang et al. "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving." CVPR. 2019.

Results. Stereo Object Matching Network

- Detection comparison.

Method	Modality	AP_{2D} (Moderate)	AP_{BEV} (Moderate)	AP_{3D} (Moderate)
Multi-Fusion [6]	M	-	13.63	5.69
M3D-RPN [38]	M	83.67	21.18	17.06
3DOP [39]	S	88.07	9.49	5.07
TLnet [40]	S	-	21.88	14.26
StereoRCNN [27]	S	**88.27**	48.30	36.69
PseudoLiDAR [7]	S	-	56.80	45.30
DSGN [6]	S	83.59	**63.91**	**54.27**
Ours	S	76.71	30.92	19.88

Ours: Cost volume (pixel unit)
PseudoLiDAR: pointclouds (metric unit)
DGSN: Geometry volume (metric unit)

Table3. Detection results in KITTI detection validation benchmark.

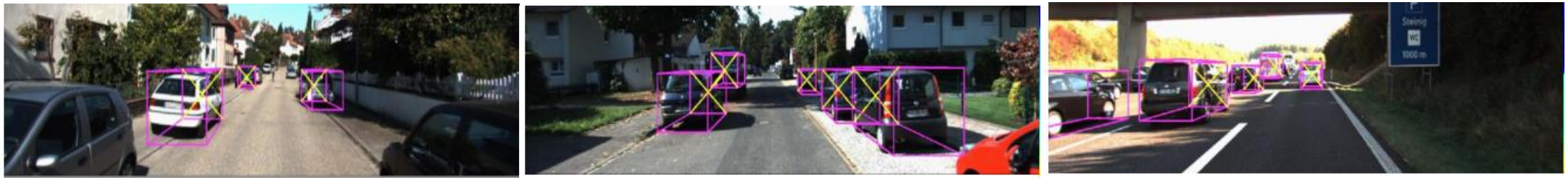


Fig3. Qualitative results from our stereo object matching network.

[6] Chen et al. "Dsgn: Deep stereo geometry network for 3d object detection." CVPR. 2020.
[7] Wang et al. "Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving." CVPR. 2019.

Volumetric Propagation Network: Stereo-LIDAR Fusion for Long-Range Depth Estimation
(RA-L 2021 with ICRA presentation)

Intro. problem definition

- What is stereo-LiDAR fusion?

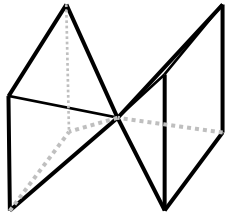
Given stereo images and pointclouds,
this task aims to estimate a dense depth map in left camera viewpoint.

Right image ($\mathcal{I}_R$)

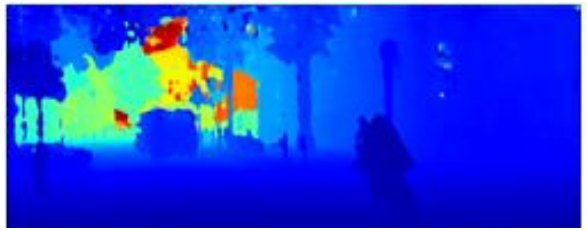
Left image ($\mathcal{I}_L$)

$$\begin{bmatrix} x_1 & x_2 & \cdots & x_i & \cdots & x_k & \cdots & x_N \\ y_1 & y_2 & \cdots & y_i & \cdots & y_k & \cdots & y_N \\ z_1 & z_2 & \cdots & z_i & \cdots & z_k & \cdots & z_N \end{bmatrix}$$

Raw point clouds ($\mathcal{P}$)



Deep network



Stereo-LiDAR fusion

Depth map ($\tilde{\mathcal{Z}}$)

Intro. problem definition

- What is stereo-LiDAR fusion?

Stereo-LiDAR fusion can be viewed as combination of stereo matching and depth completion.

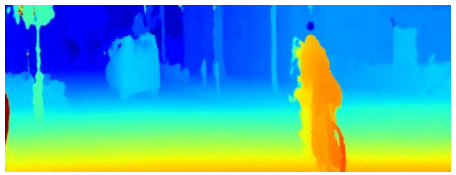
Right image ($\mathcal{I}_R$)

Left image ($\mathcal{I}_L$)

$$\begin{bmatrix} x_1 & x_2 & \cdots & x_i & \cdots & x_k & \cdots & x_N \\ y_1 & y_2 & \cdots & y_i & \cdots & y_k & \cdots & y_N \\ z_1 & z_2 & \cdots & z_i & \cdots & z_k & \cdots & z_N \end{bmatrix}$$

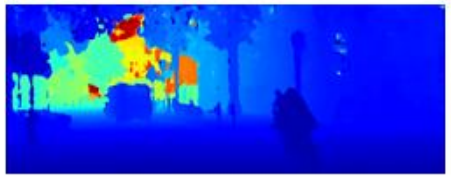
Raw point clouds ($\mathcal{P}$)

Stereo matching



Disparity map

Depth completion



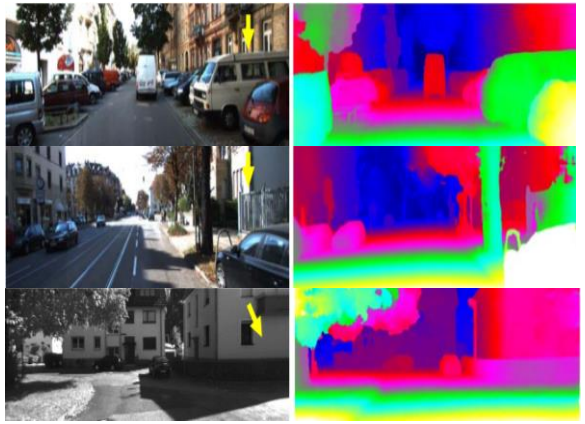
Depth map

Intro. problem definition

- Why stereo-LiDAR fusion?

For driving safety, it is required to estimate long-range & high-quality depth.

Stereo Matching



RMSE=910mm

(Chang and Chen [1])

<

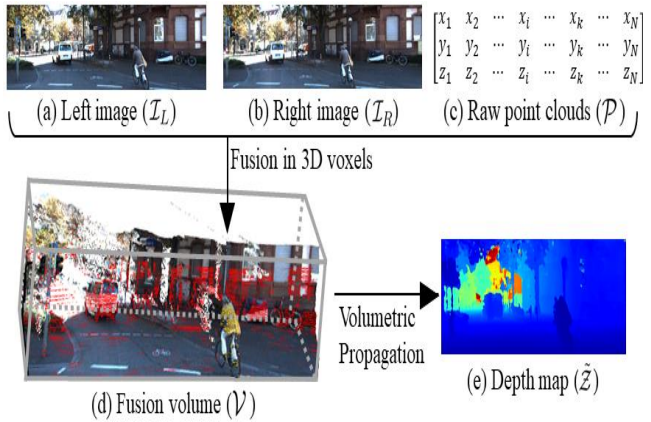
Depth Completion

RMSE=781mm

(Chen et al. [2])

<

Stereo-LiDAR fusion



RMSE=636mm

(Ours)

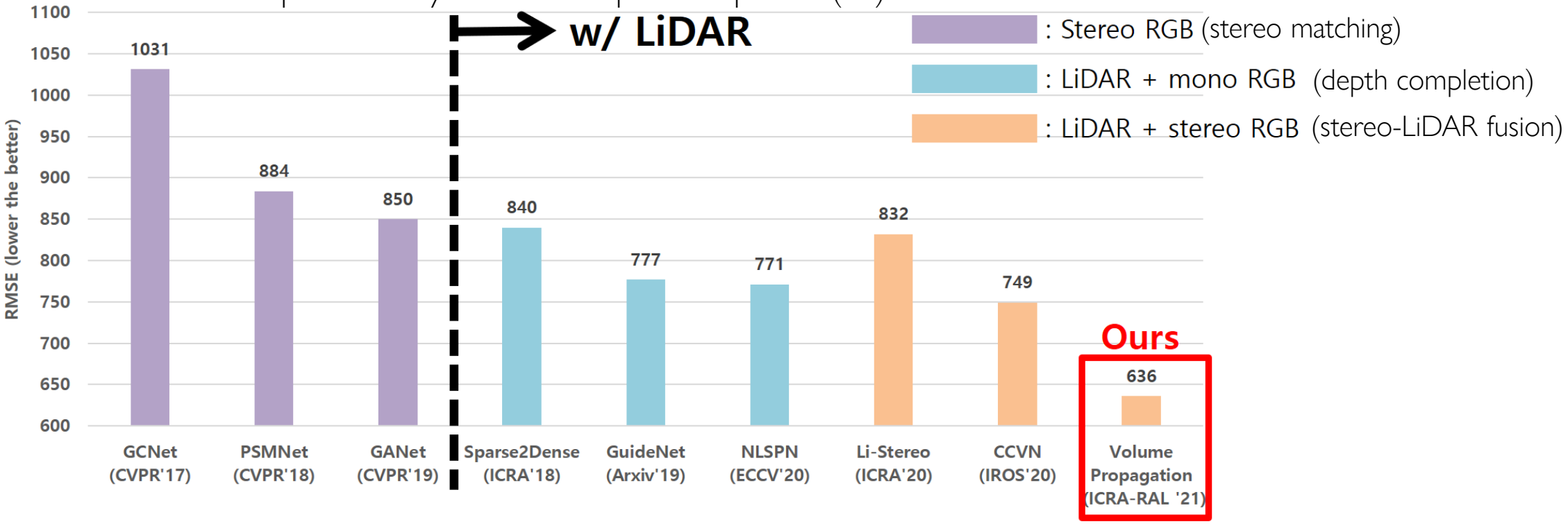
[1] Chang and Cheng., "Pyramid stereo matching network." CVPR. 2018.
[2] Chen et al. "Learning joint 2d-3d representations for depth completion." ICCV. 2019.

Intro. problem definition

- Why stereo-LiDAR fusion?

Theoretically, recent depth estimation methods can infer long-range depth, they suffer from low-quality depth estimation.

< Depth accuracy on KITTI Depth Completion (val) Benchmark >



Intro. challenges

- What is the issue in stereo-LiDAR fusion?

  How to fuse two different modalities, stereo images (cameras) and point clouds (LiDAR)?

Intro. previous works

- How to fuse stereo images and point clouds?

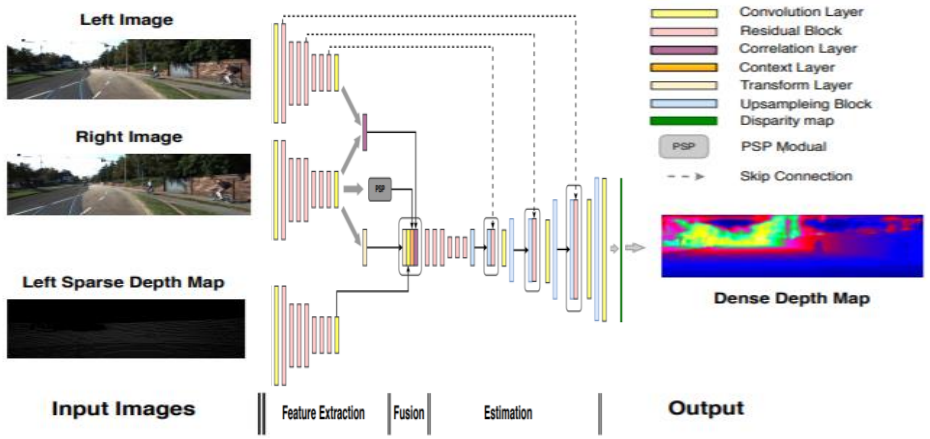


Fig. Architecture of LiStereo [3].

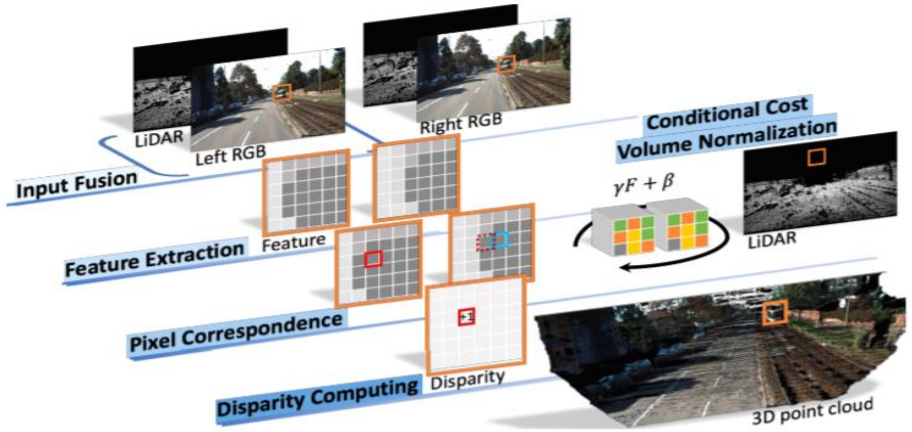


Fig. Overview of CCVN [4].

Common strategy:
pointclouds -> (project) -> sparse depth maps (pre-processing)
-> RGB-D input (pixel-wise fusion)

[3] Zhang et al. "LiStereo: Generate Dense Depth Maps from LIDAR and Stereo Imagery." ICRA, 2020.
[4] Wang et al. "3d lidar and stereo fusion using stereo matching network with conditional cost volume normalization." IROS. 2020.

Intro. challenges

- How to fuse stereo images and point clouds?

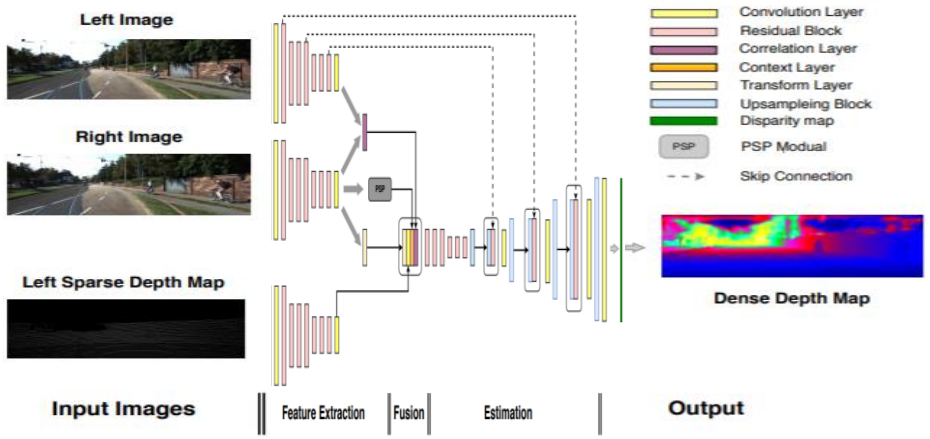


Fig. Architecture of LiStereo [3].

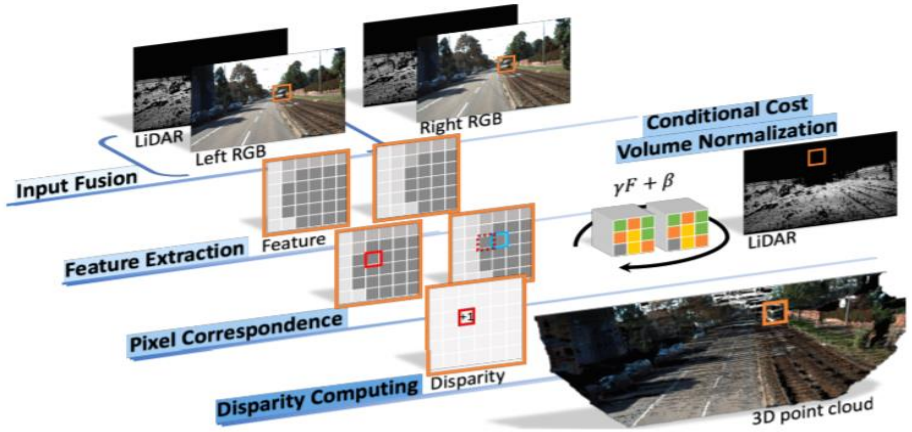
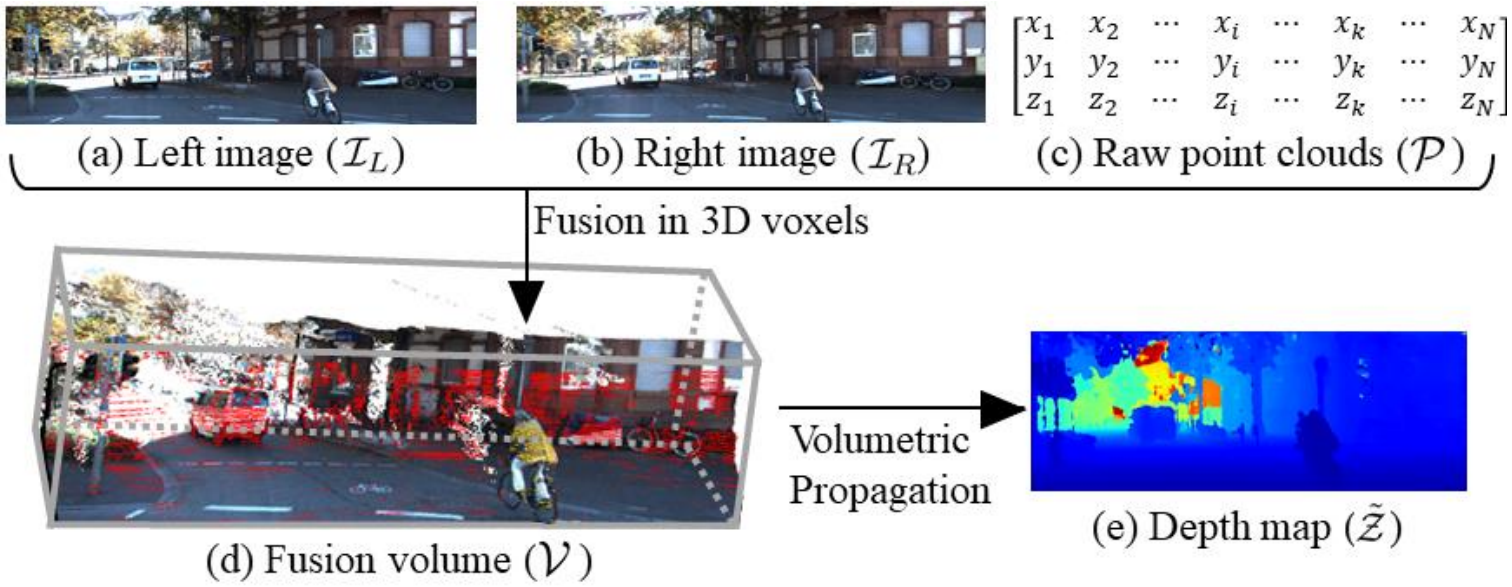


Fig. Overview of CCVN [4].

However, neighboring pixels in a 2D image domain are not necessarily adjacent in a 3D space.
2D fusion can loose depth-wise spatial connectivity,
which can not be an appropriate fusion scheme.

Intro. solutions

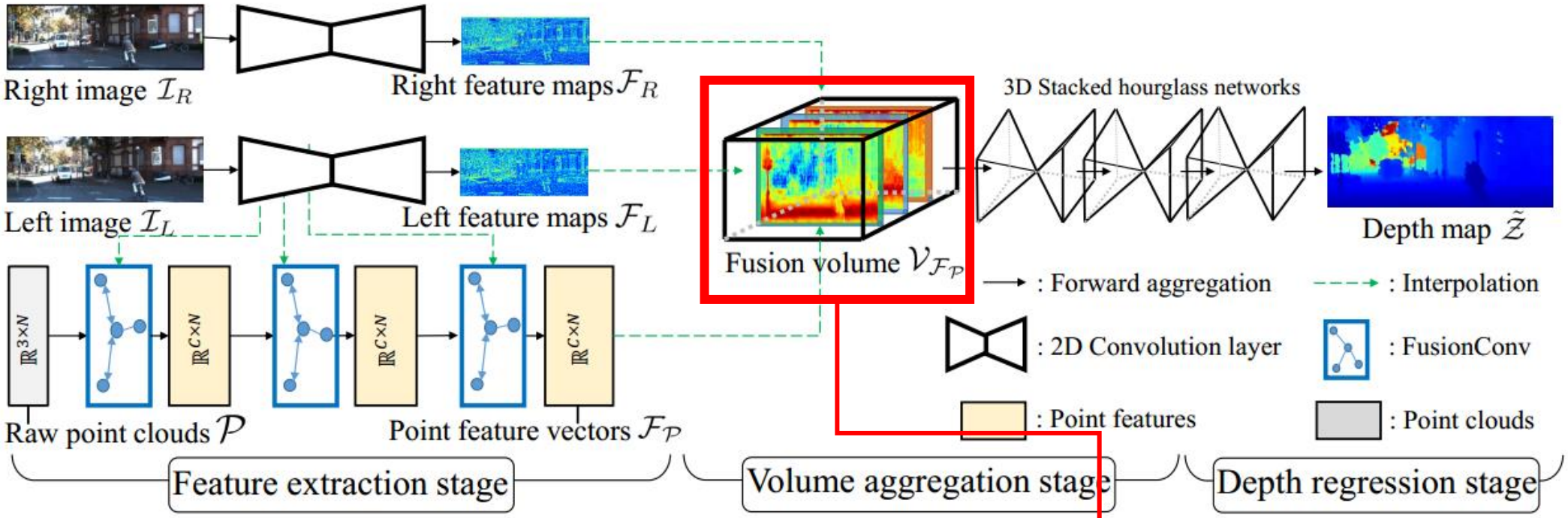
- How to fuse stereo images and point clouds?



In this paper, we propose a **geometry-aware** stereo-LiDAR fusion network for long-range depth estimation, called *volumetric propagation network*.

Main. Volumetric propagation network

- Geometry-aware Fusion



The key idea of our fusion is related to **fusion volume** that
align stereo images' feature maps and point feature vectors directly in 3D space.

Main. Volumetric propagation network

- Geometry-aware Fusion ... (fusion volume)

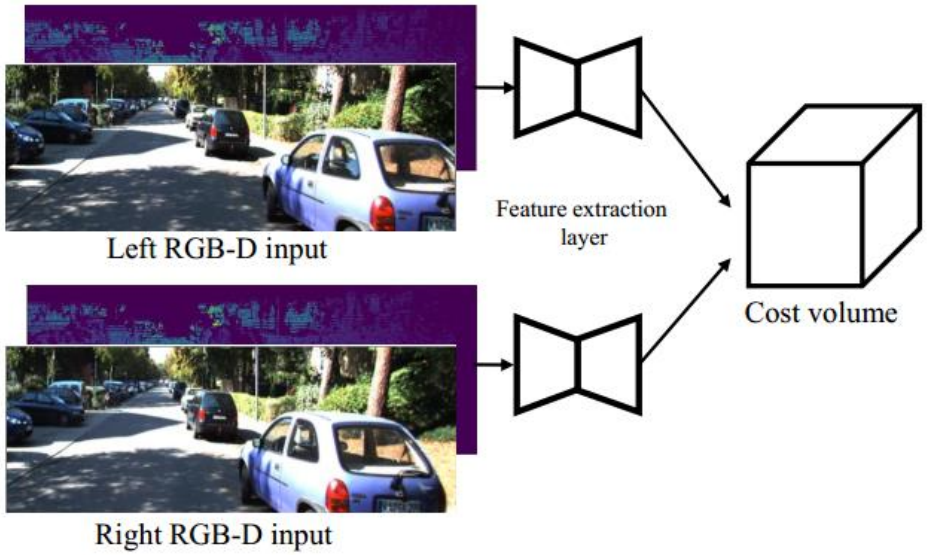


Fig. Simple visualization of **2D fusion (RGB-D)** networks.
→ Can **loss** 3D geometric relation.
→ Compute pixel-wise fusion.

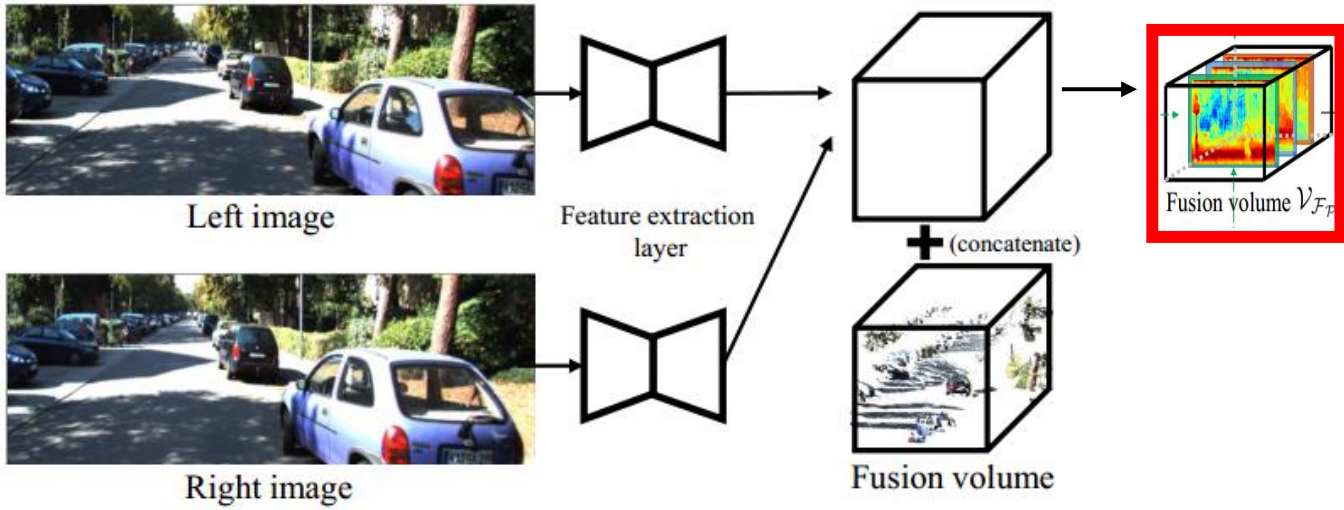
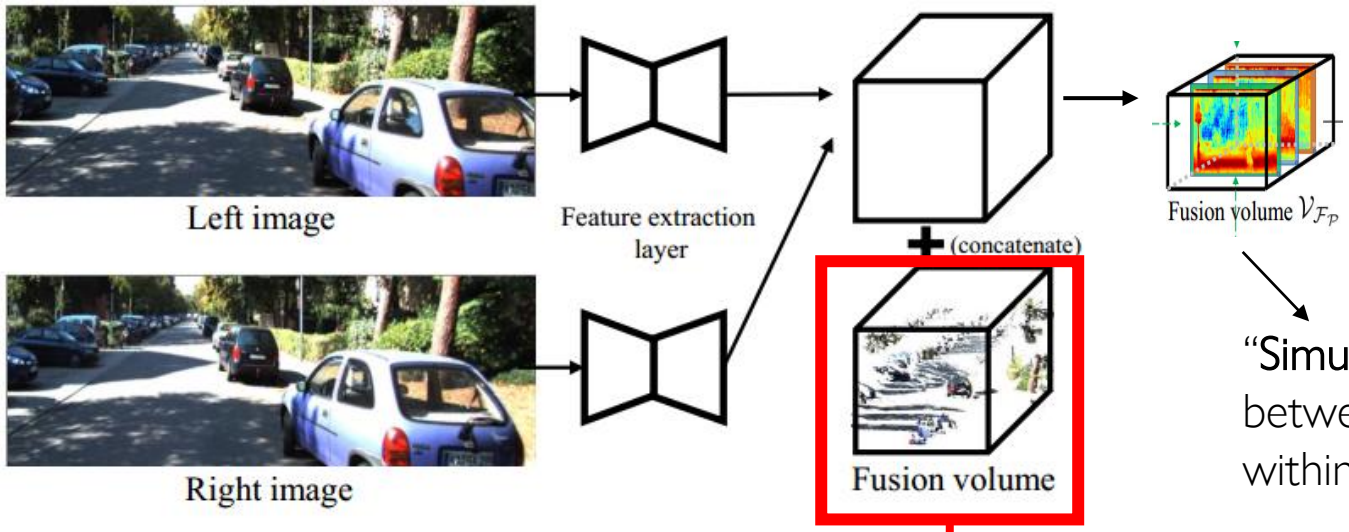


Fig. Simple visualization of **3D fusion** networks (ours).
→ **Maintain** 3D geometric relation.
→ Simultaneously compute matching cost of all input data.

Our fusion volume (right) **maintains the geometric relation** of stereo images and point clouds within 3D volumetric space.

Main. Volumetric propagation network

- Geometry-aware Fusion ... (fusion volume)

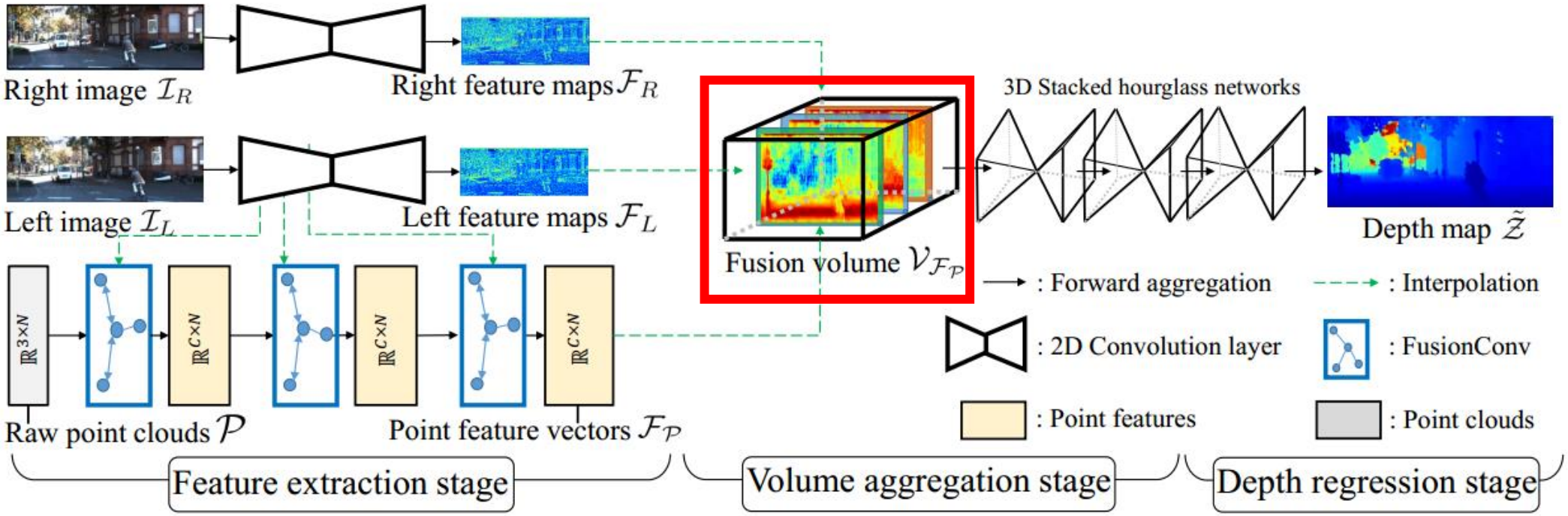


"Simultaneously compute matching cost between stereo images and pointclouds within the 3D volumetric space."

By directly embedding point features into 3D volume ,
we maintain metric-accuracy from pointclouds.
To do so, we formulate stereo-LiDAR fusion as **volumetric propagation**.

Main. Volumetric propagation network

- Geometry-aware Fusion ... (fusion volume)



Thanks to our fusion volume, we can improve the quality of long-range depth.

Results.

- State-of-the-art performance in depth estimation.

TABLE I
QUANTITATIVE RESULTS OF DEPTH ESTIMATION NETWORKS IN KITTI COMPLETION VALIDATION BENCHMARK.
* REPRESENTS THE REPRODUCED RESULTS.

Method	Modality	Depth Evaluation (Lower the better)			
		RMSE (mm)	MAE (mm)	iRMSE (1/km)	iMAE (1/km)
GCnet [5]	Stereo	1031.4	405.4	1.6814	1.0356
PSMnet* [15]	Stereo	884	332	1.649	0.999
Sparse2Dense* [6]	Mono + LiDAR	840.0	-	-	-
Guidenet [7]	Mono + LiDAR	777.78	221.59	2.39	1.00
NLSPN [18]	Mono + LiDAR	771.8	197.3	2.0	0.8
CSPN++ [30]	Mono + LiDAR	725.43	207.88	-	-
Park *et al.* [11]	Stereo + LiDAR	2021.2	500.5	3.39	1.38
LiStereo [12]	Stereo + LiDAR	832.16	283.91	2.19	1.10
CCVN [10]	Stereo + LiDAR	749.3	252.5	**1.3968**	**0.8069**
Ours	Stereo + LiDAR	**636.2**	**205.1**	1.8721	0.9870

Results.

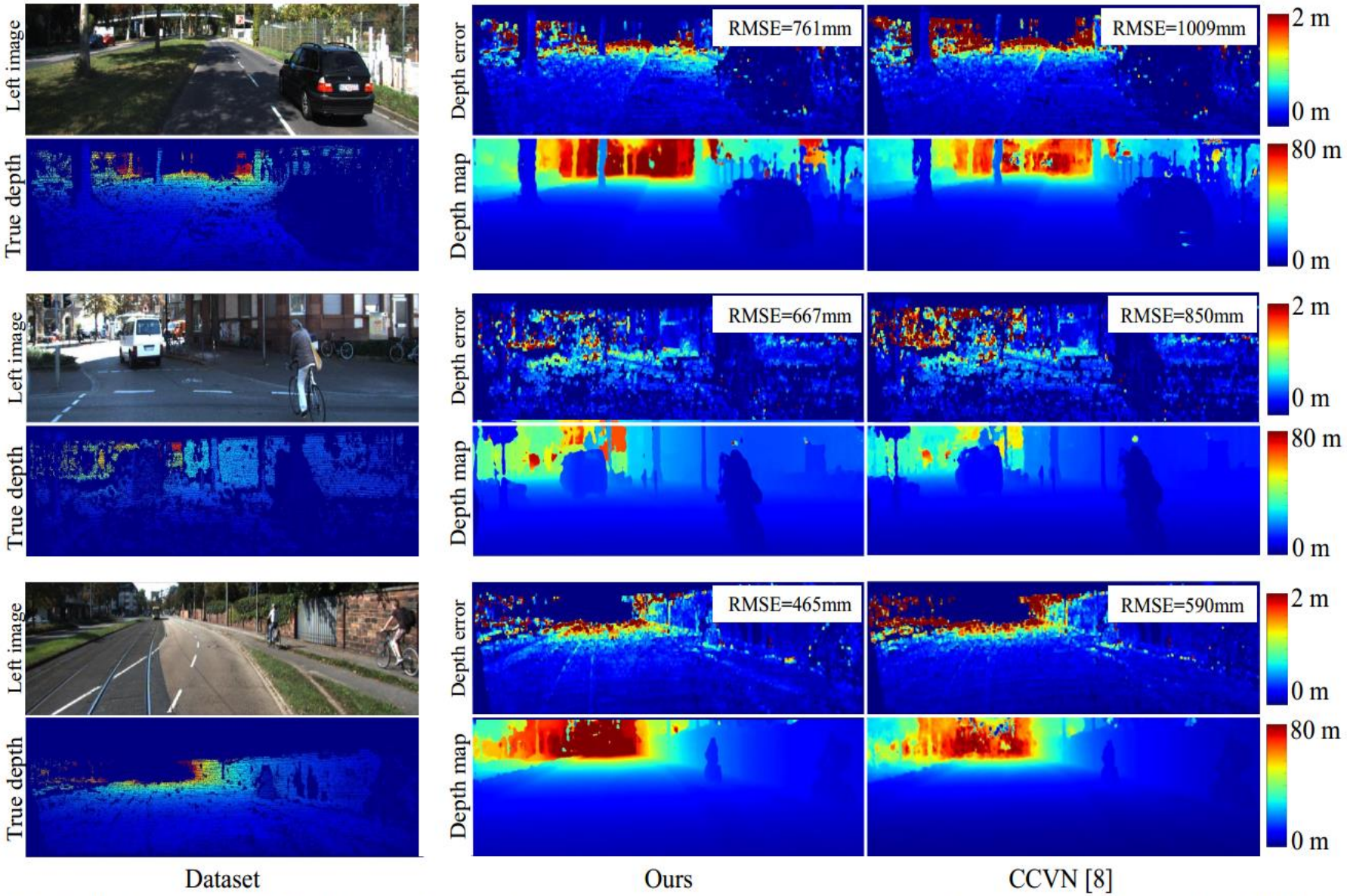


Fig. 4. **Qualitative results on KITTI dataset.** We visualize depth maps and depth errors from ours and the recent stereo-Lidar method (CCVN [10]) in three different cases. We also include the depth metric RMSE (lower the better).

Results.

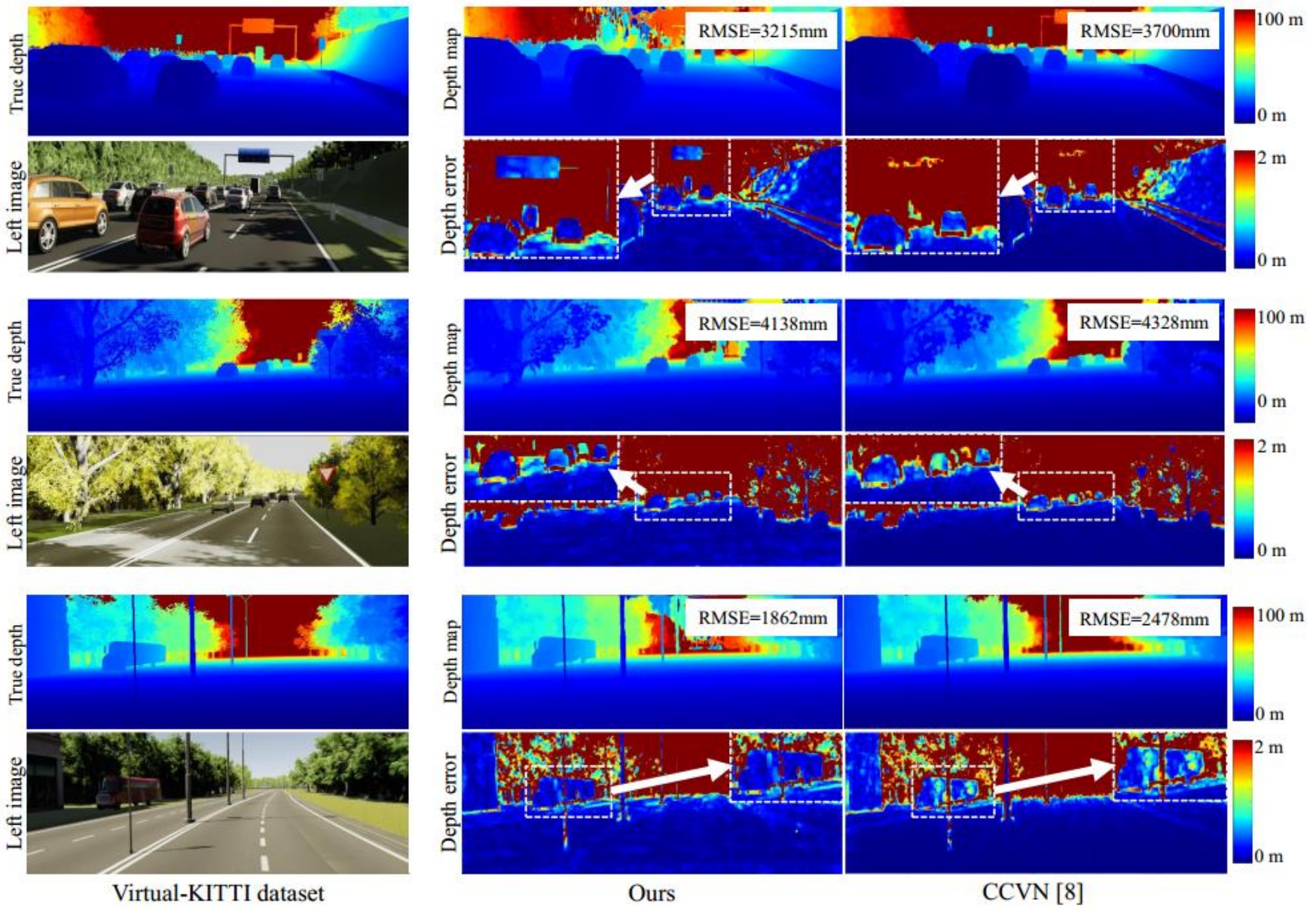


Fig. 5. **Qualitative results on the Virtual-KITTI 2.0 dataset.** We evaluate the estimated depth from both our network and the recent Stereo-LiDAR fusion network by Wang *et al*. [10]. This synthetic data covers the wide range of depth upto $655m$, but we clamp true depth maps and estimated depth maps upto $100m$ during the evaluation, as in Table II. For the detailed visualization, we crop and enlarge the part of depth error maps in each frame. Mainly, the cropped images correspond to the farther area to validate our long-range depth estimation.

Thank you ☺