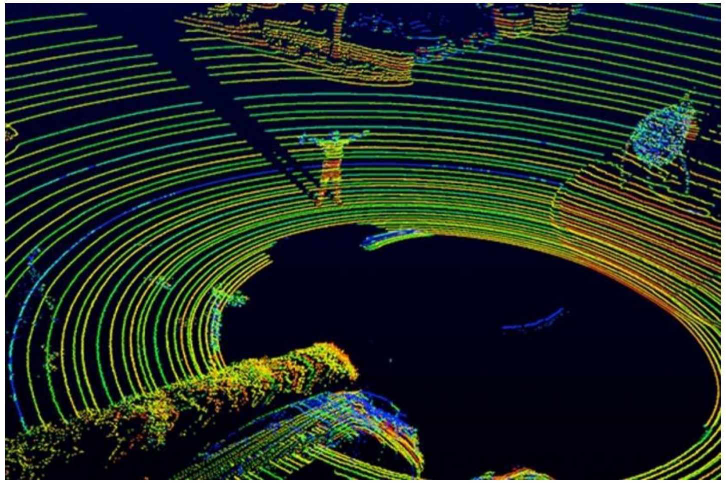
Exploring the New Designs of 3D Sensors

Qifeng Chen
Assistant Professors
CSE & ECE



3D Sensors

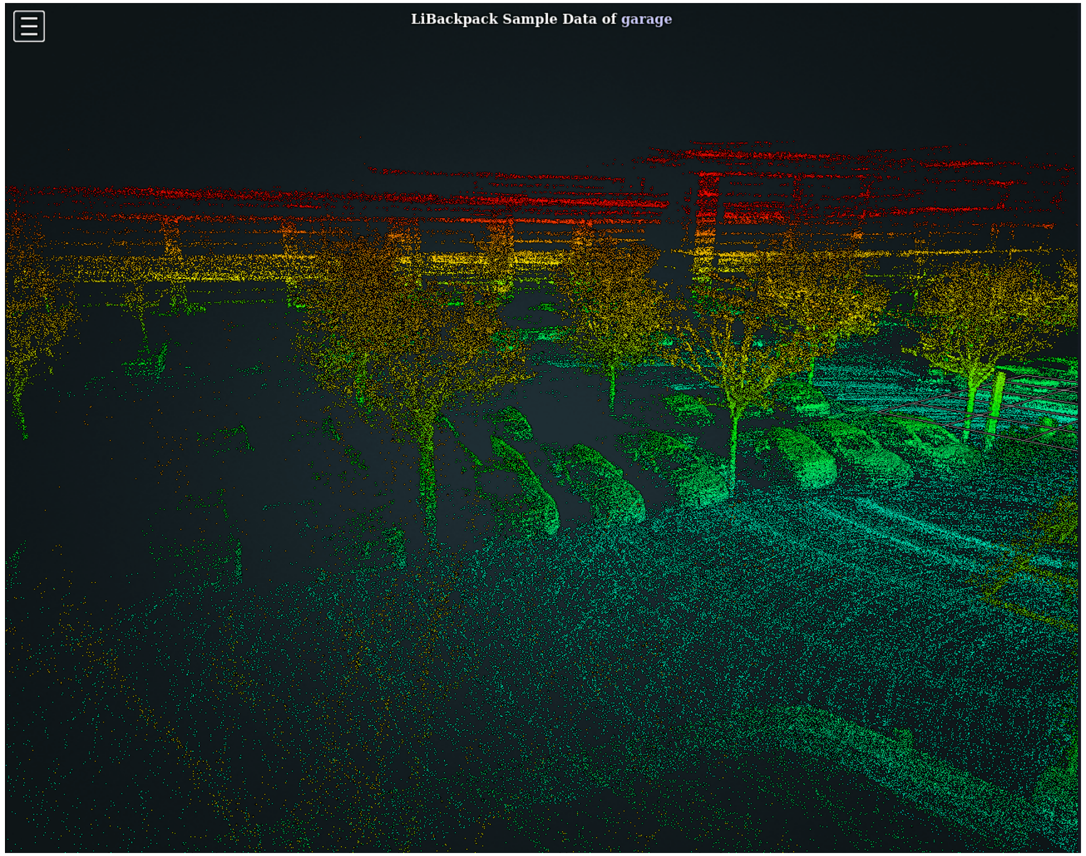
- The real world is 3D
- 3D sensors play a role in autonomous driving and visual perception
  - Reliable distance information
  - Infer semantic information

Sample "Gesture Control" Image Taken by ToF Camera

Time-of-flight camera

Stereo cameras

LiDAR



Kinect

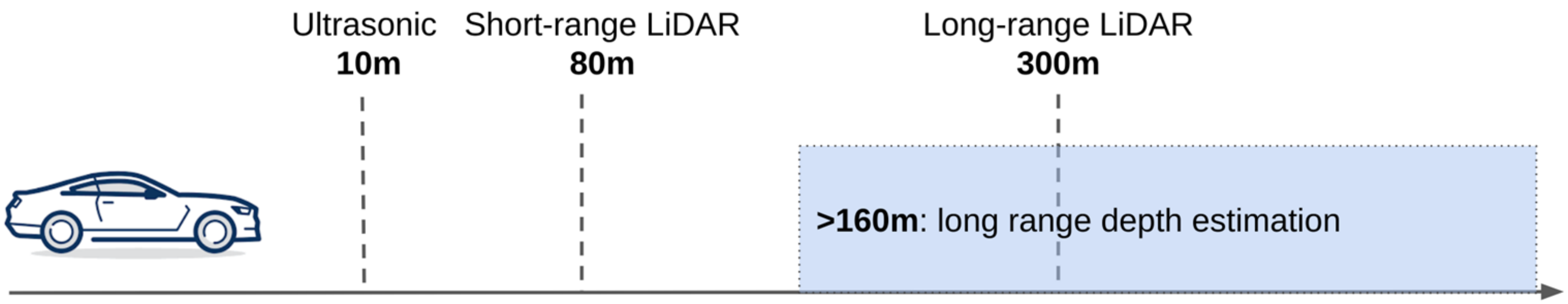
CORNELL TECH

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

CVPR SEATTLE WASHINGTON JUNE 16-18 2020

Depth Sensing Beyond LiDAR Range

Kai Zhang, Jiaxin Xie, Noah Snavely, Qifeng Chen



Depth Sensing Beyond LiDAR Range

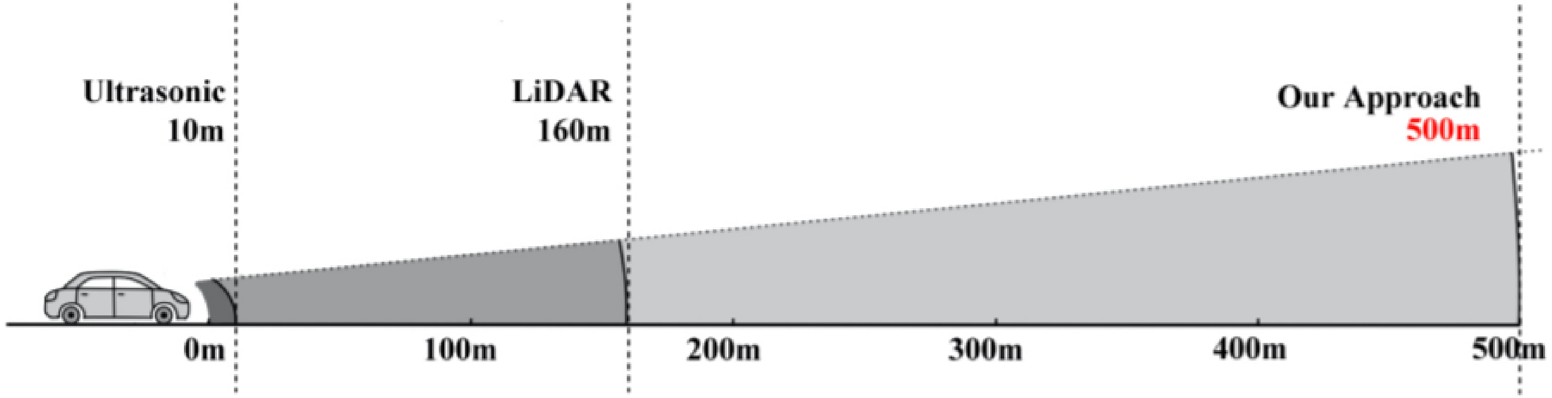
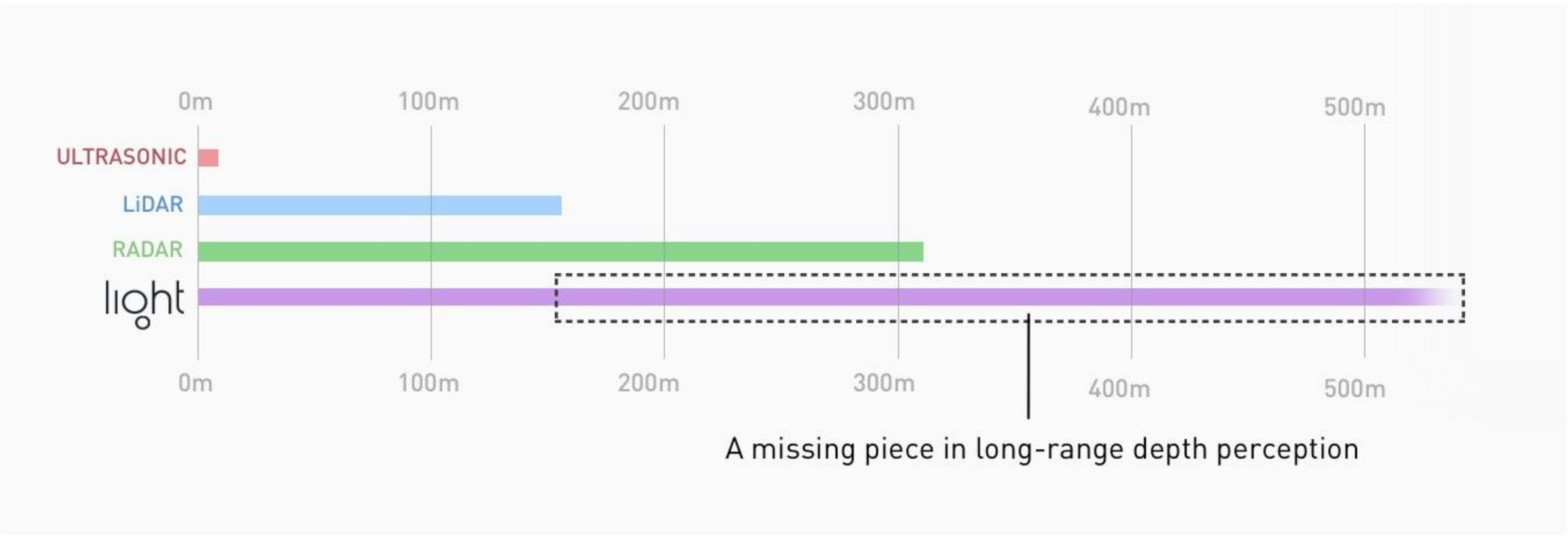


Figure 1: Visualization of existing depth-sensing solutions' maximum range.



Motivation

Self-driving datasets	
Kitti	80 meters
Waymo	80 meters
…	

60 mph = 96 km/h = 27 m/s
80 meters roughly means 3 seconds

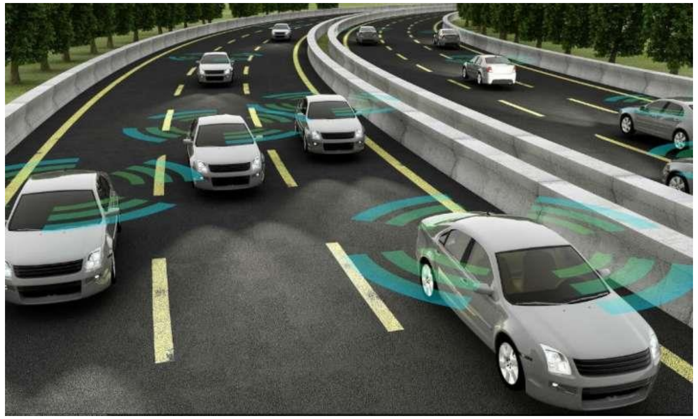
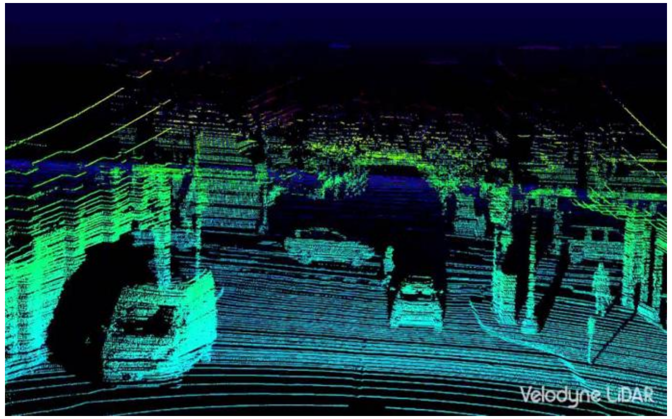



Image sources: velodyne lidar

Question: can we achieve *dense* depth sensing beyond LiDAR range with *low-cost cameras*?

Example application:
Autonomous trucks driving on highway

Problem Setup

Basic idea: use two **cameras with telephoto lens** to capture a stereo pair, then reconstruct a dense depth map.

Nikon P1000

Canon SX70

Industrial cameras[1]

[1] Industrial cameras are usually much cheaper than consumer ones.

Problem Setup

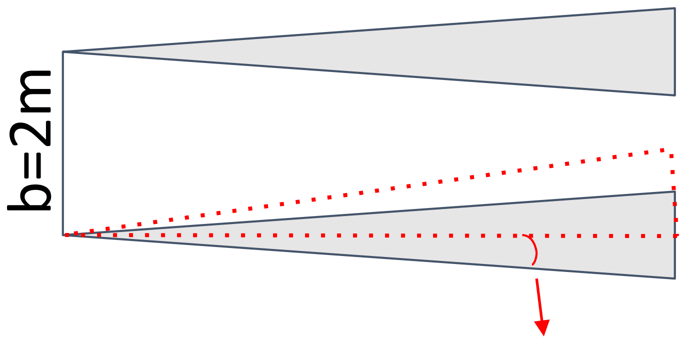
Important camera setup constraint:

Baseline is restricted to ~2 meters because of typical vehicle size.

What does this mean?

Depth estimation is **very sensitive** to pose error, especially rotation error.

It's difficult for hardwares to achieve and maintain this precision.



$\Delta\phi$: rotation error

Triangulation angle: $\theta \approx \frac{b}{z} \approx \frac{2m}{300m} \approx 0.382^\circ$

Estimated depth: $\hat{z} \approx z \cdot (1 - \boxed{\frac{\Delta\phi}{\theta}})$

Relative error in estimated depth

Tentative Solution - SfM

Bas-relief ambiguity in SfM[1]

Big focal length → Near-orthographic camera (Weak perspectivity)

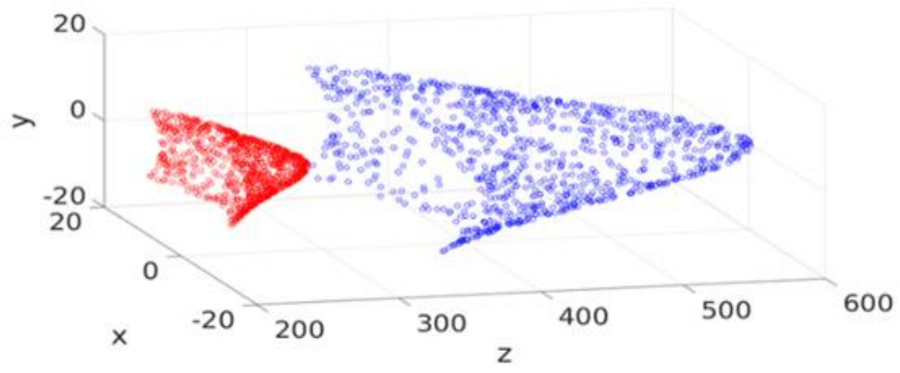


Figure 2: Ground-truth (blue) and the reconstructed (red) scene points. The unit for x, y, z axes is meter.

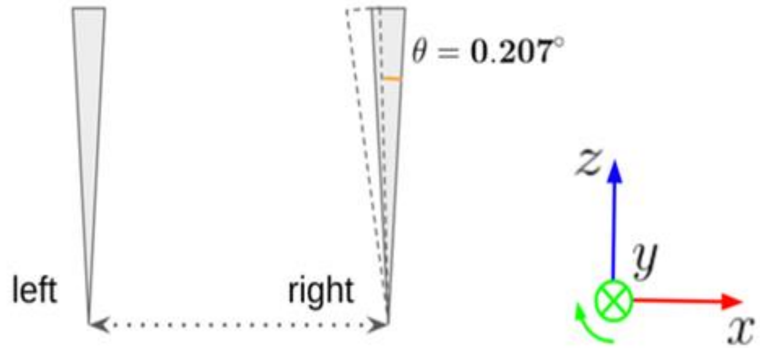
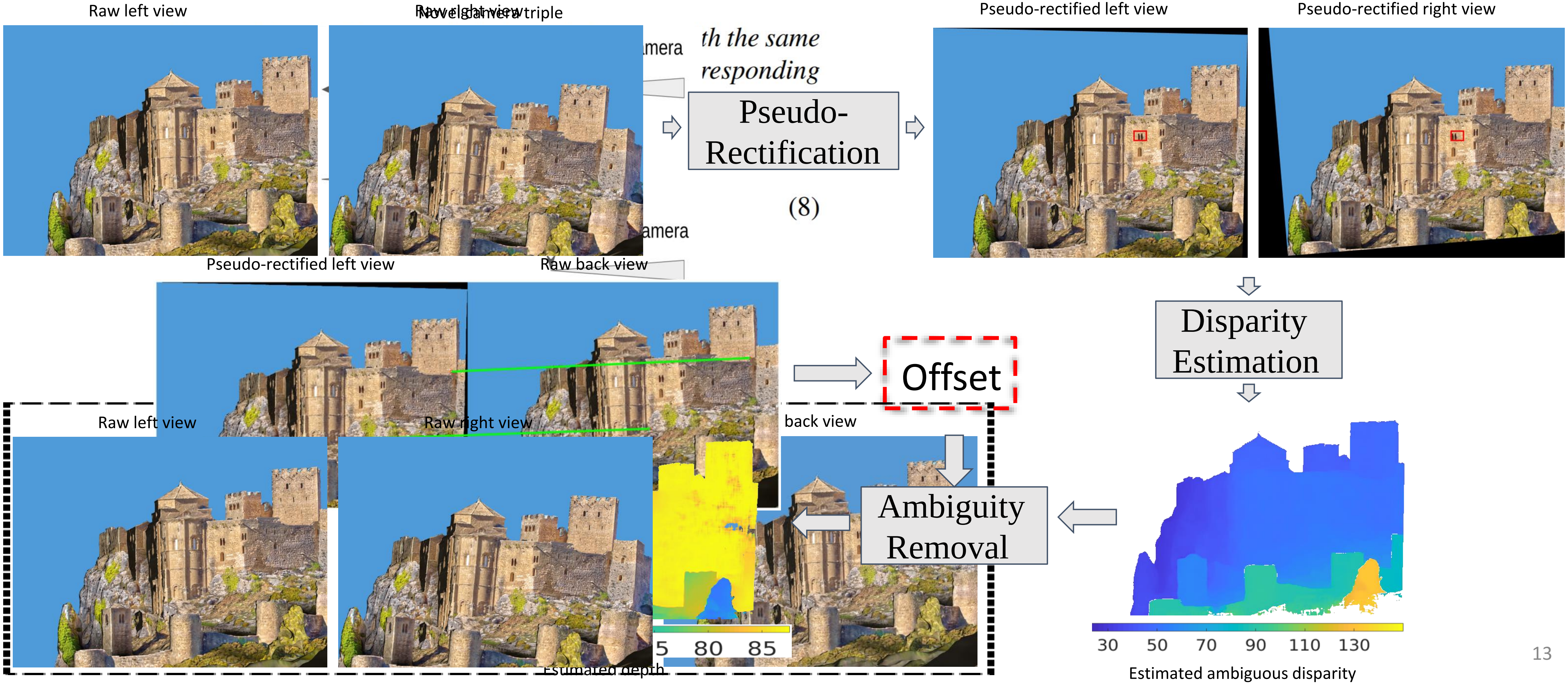


Figure 3: Top-down view of ground-truth relative pose (solid) and the recovered one (dashed). θ is exaggerated for illustration.

[1] Richard Szeliski and Sing Bing Kang. Shape Ambiguities in Structure From Motion. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 709–721. Springer, 1996.

Our Approach



Results on synthetic data

	Failure	<1%	<2%	<3%
Ours	0	**45.3%**	**80.1%**	**96.9%**
Loop and Zhang [18]	0	1.14%	2.73%	5.99%
SfM+MVS [19, 20]	15	6.71%	12.7%	19.1%

Table 1: Quantitative results on 40 synthetic scenes for methods in Fig. 7. "Failure" means the number of scenes for which a method fails to output a depth map. The metric is the portion of pixels with relative depth error below certain threshold, i.e., 1%, 2%, 3%, averaged over the successful scenes.

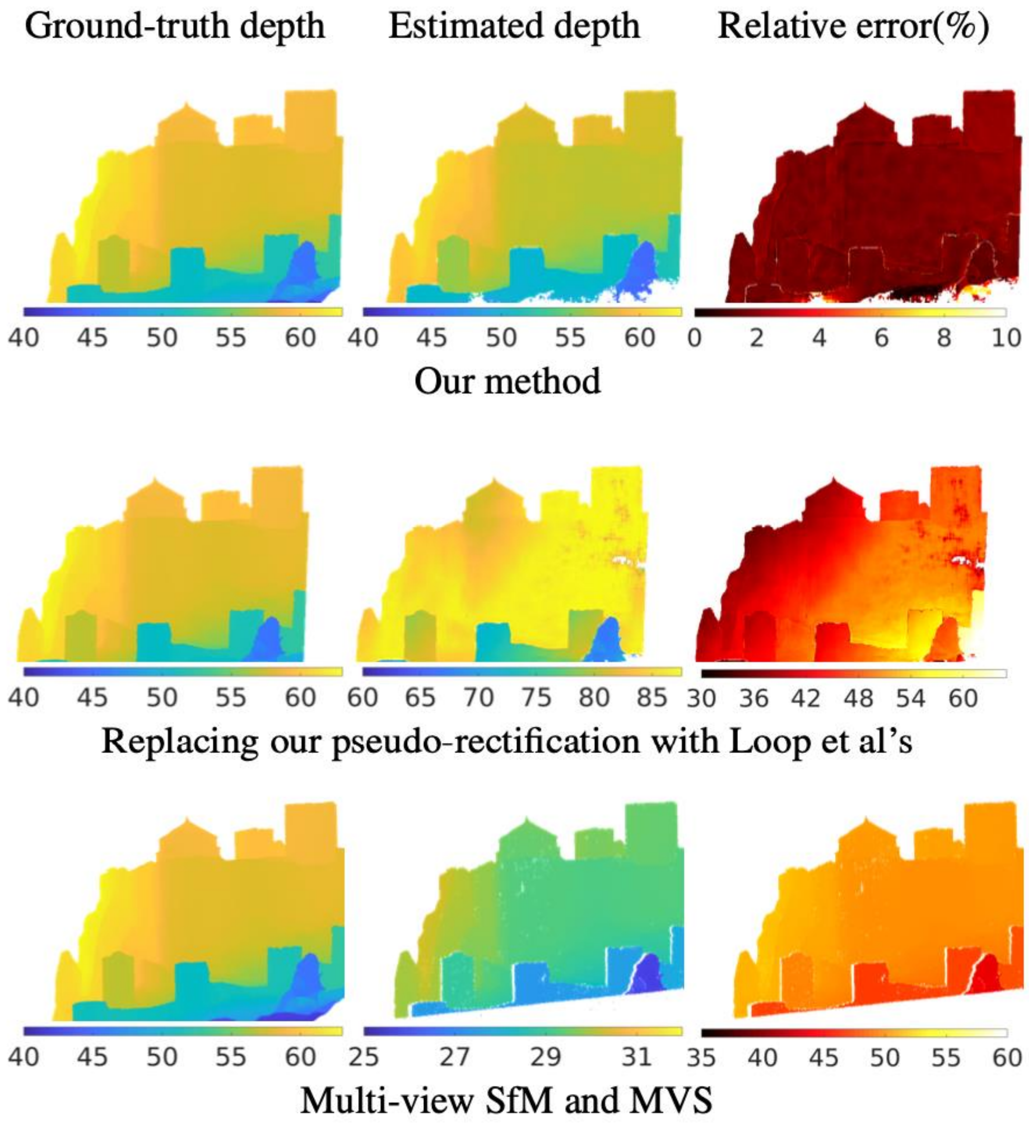
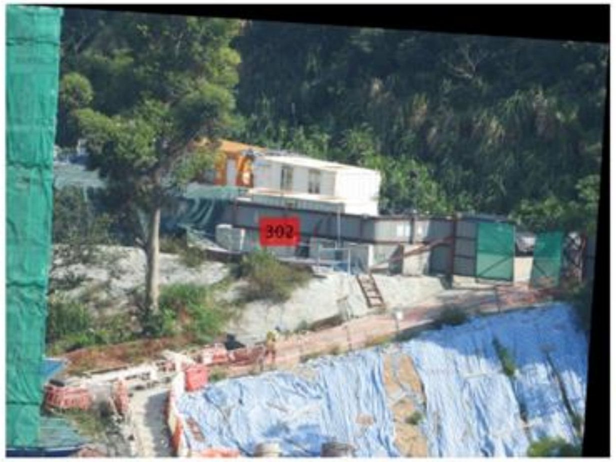
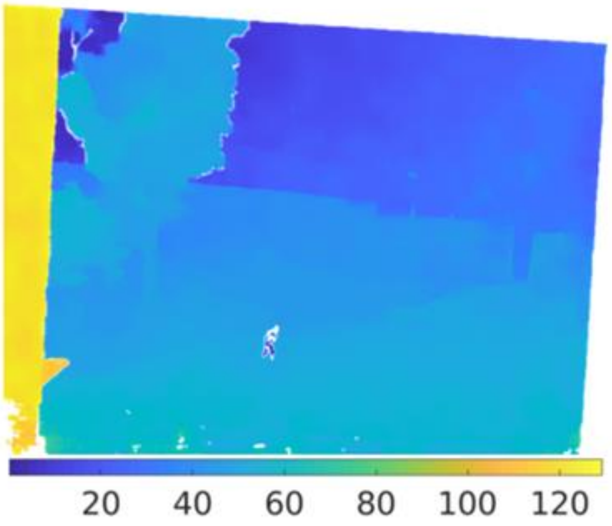


Figure 7: Comparison among different algorithms. For rectification-based methods, the ground-truth depth map has been warped to align with the rectified view. For SfM, we have used the full ground-truth intrinsic matrix.

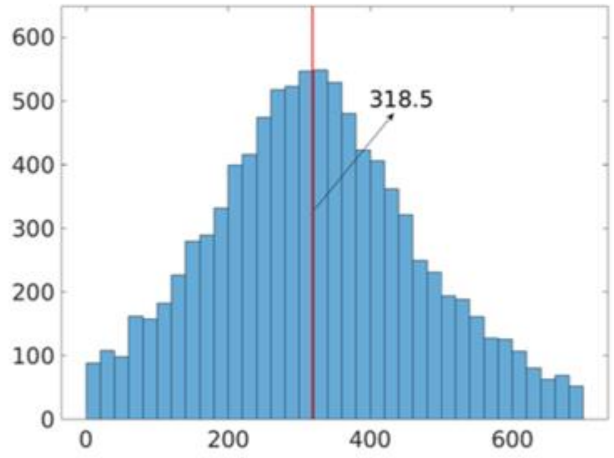
Results on real-world data



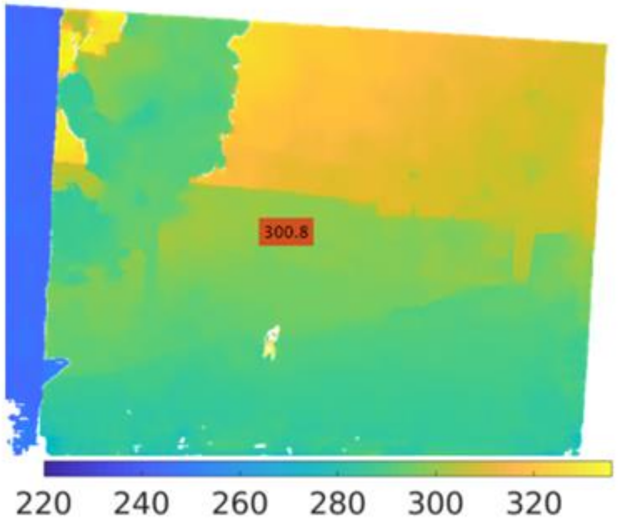
Pseudo-rectified left view



Estimated ambiguous disparity



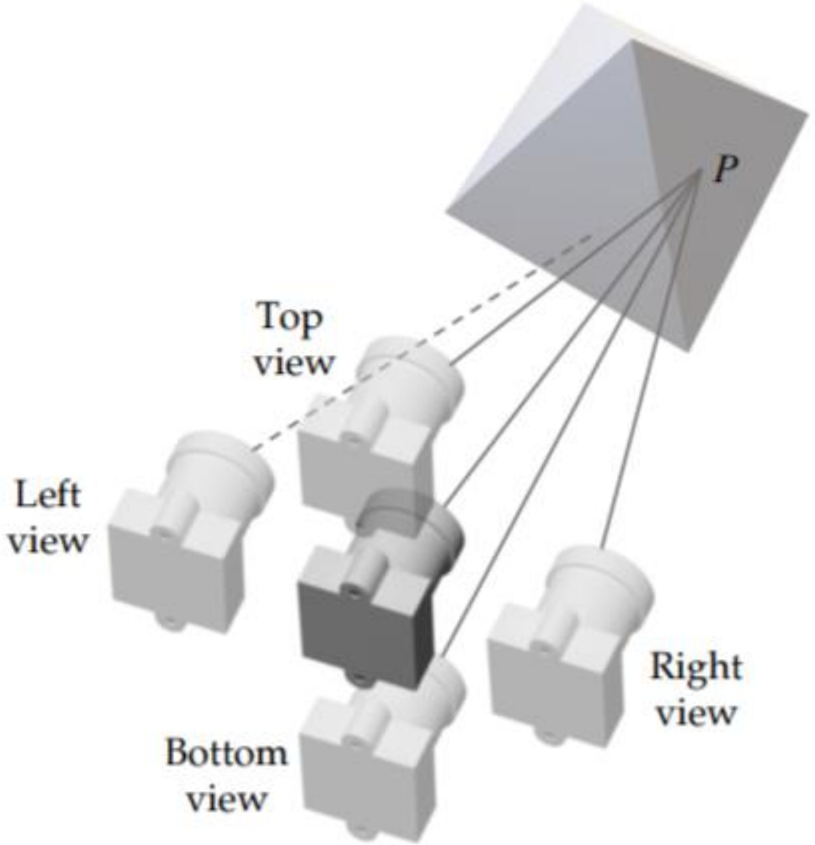
Ambiguity removal histogram



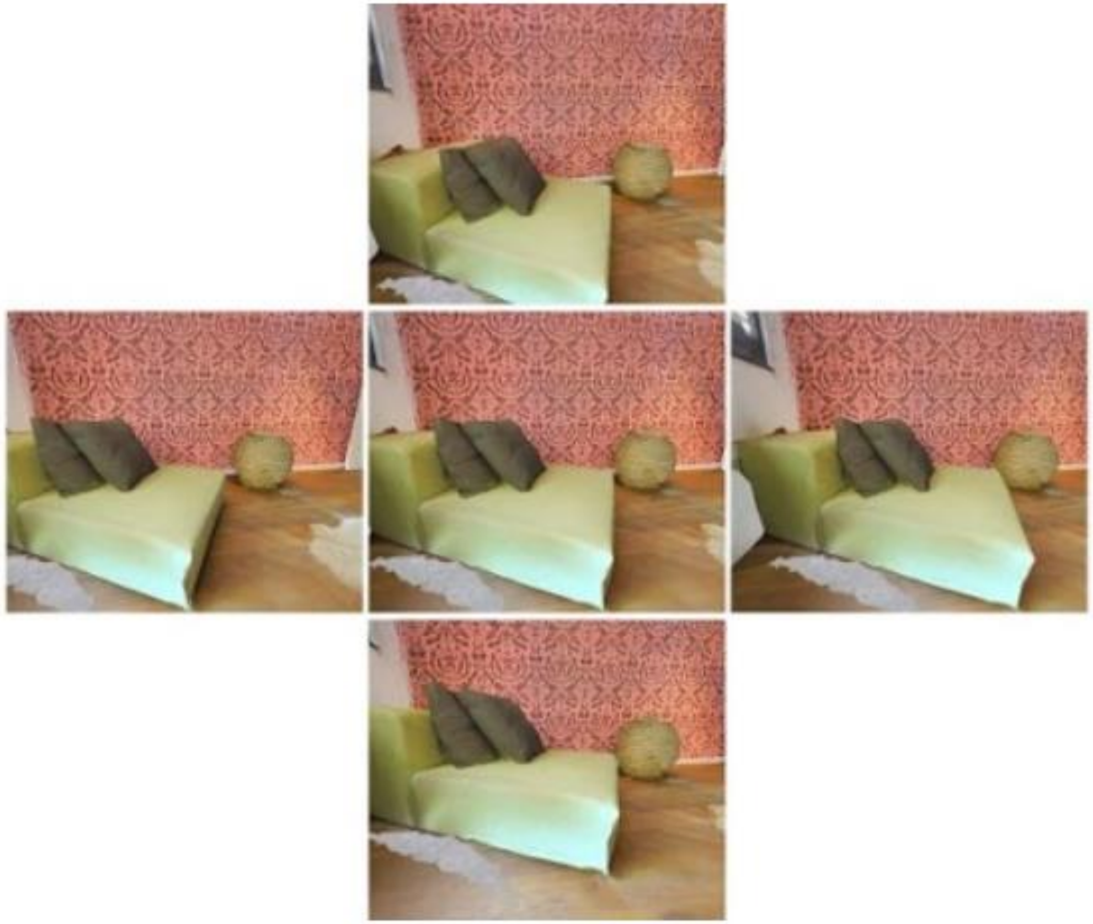
Estimated final depth

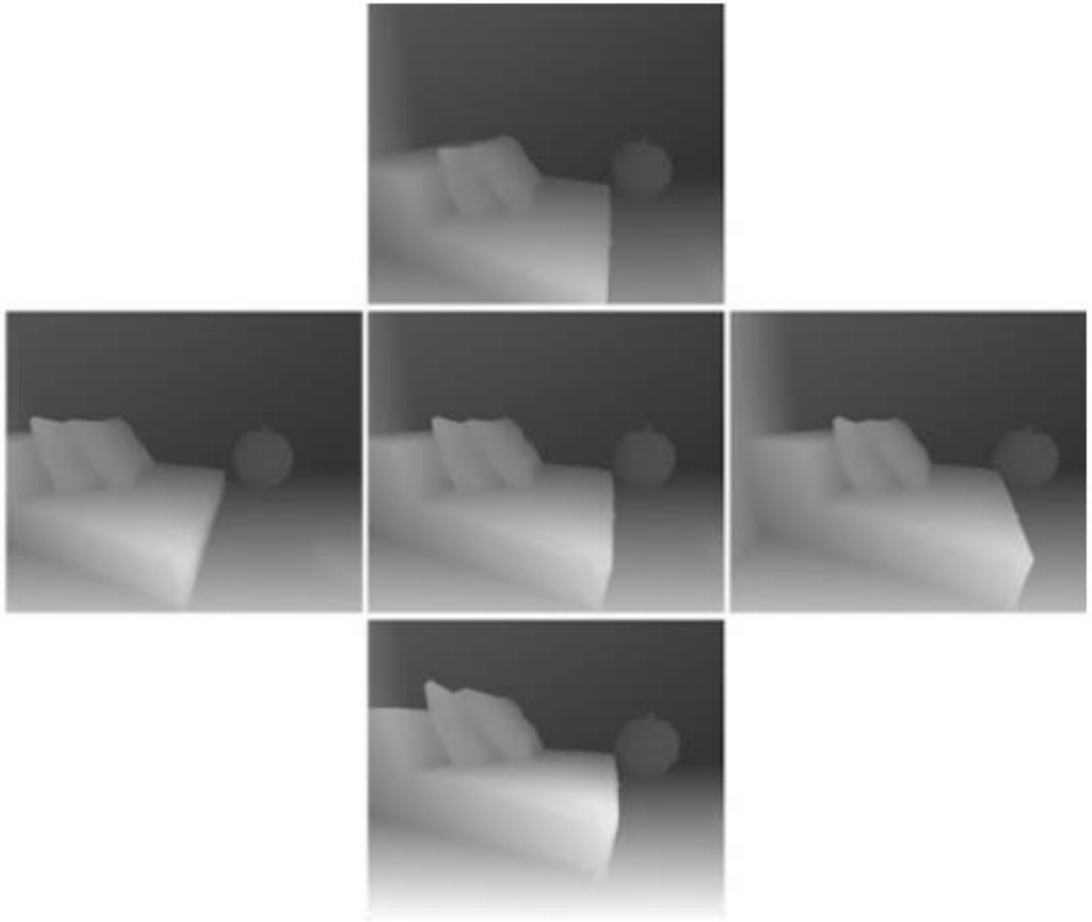
- Ground-truth depth is acquired by the laser rangefinder: only point-wise measurement
- Ground-truth: 302m Estimated: 300.8m

MFuseNet: Robust Depth Estimation with Learned Multiscopic Fusion (ICRA 2020)



Multiscopic Images





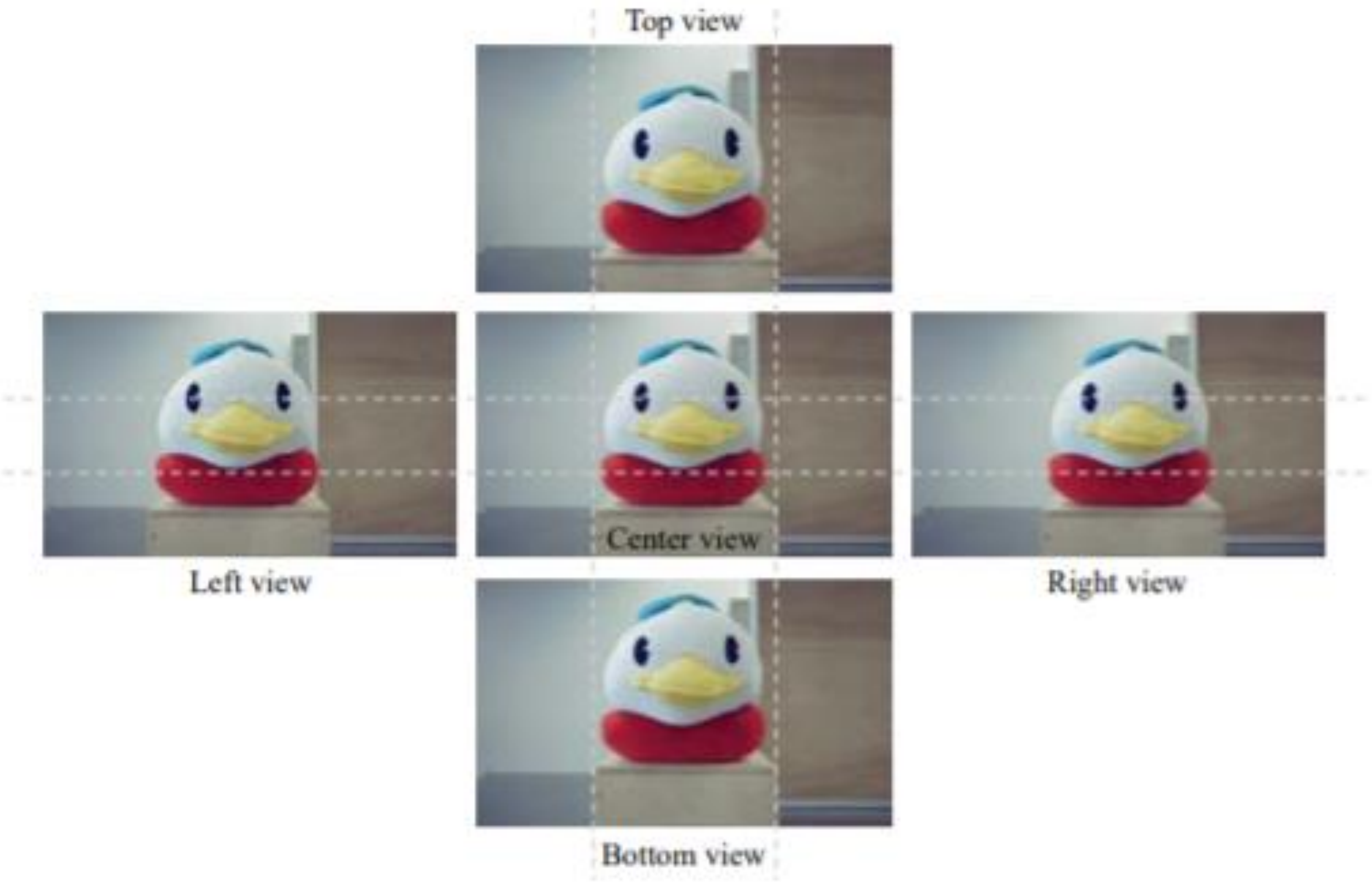


Fig. 2: Five images captured using our multiscopic perception system from different viewpoints. The parallax between the center view and any adjacent view is the same.

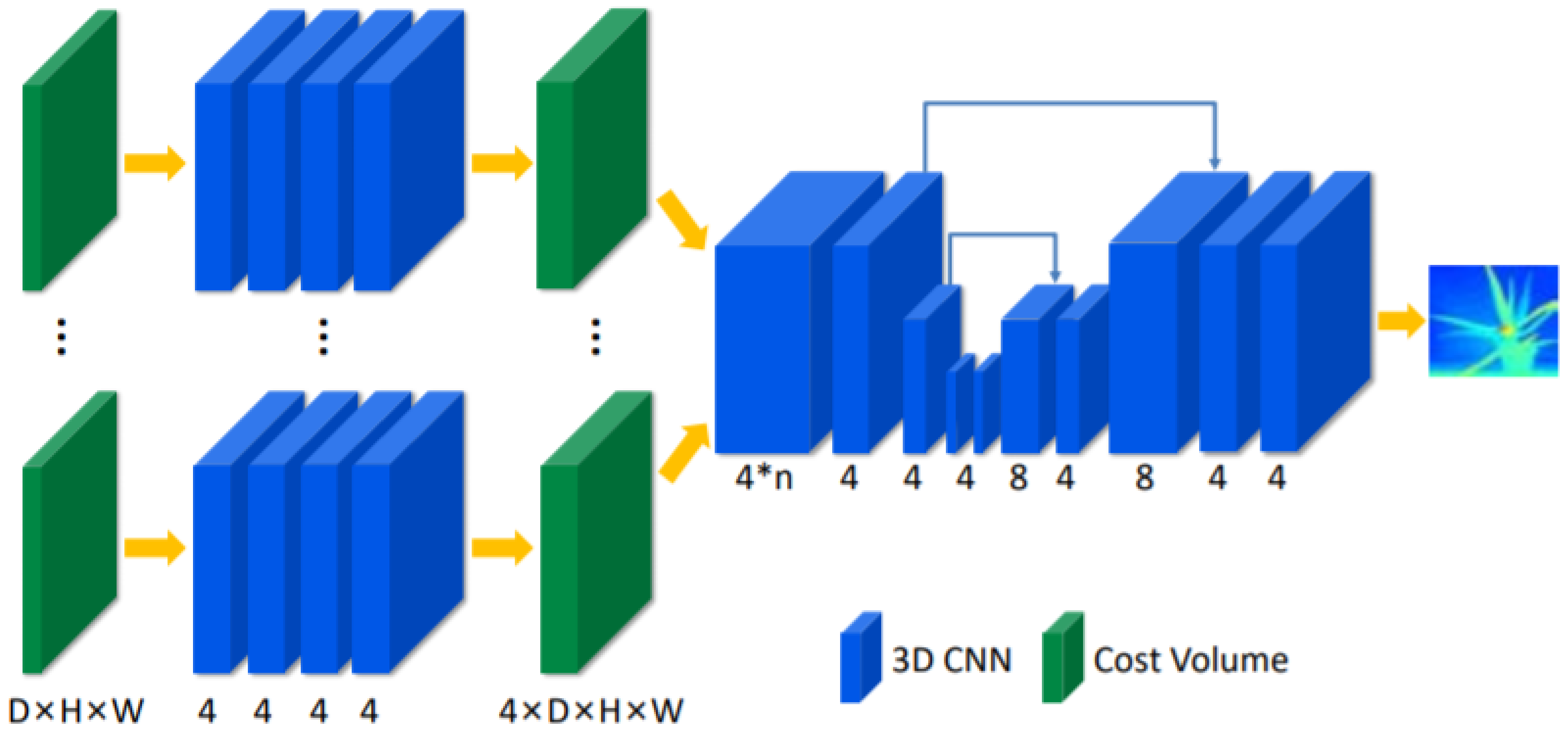
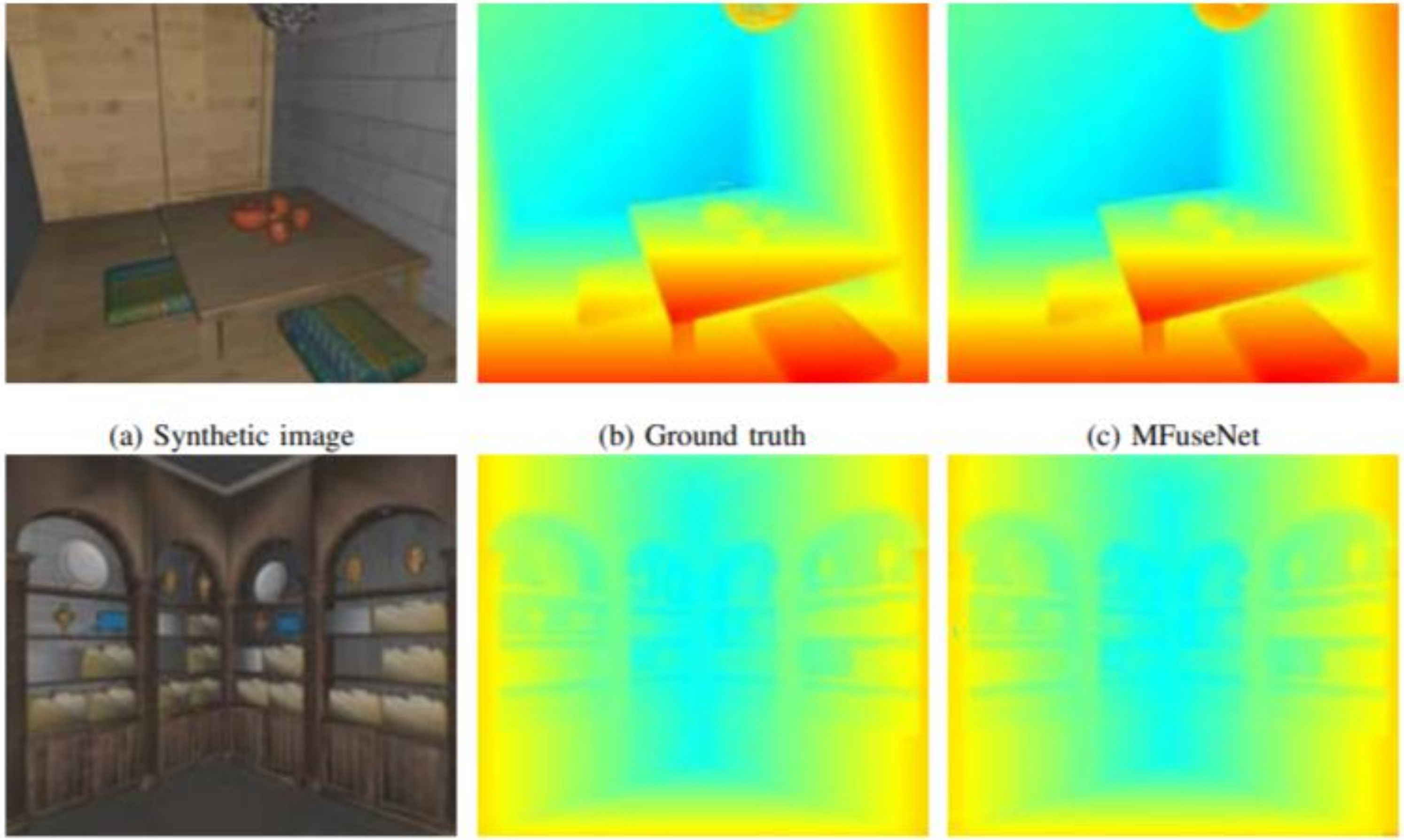


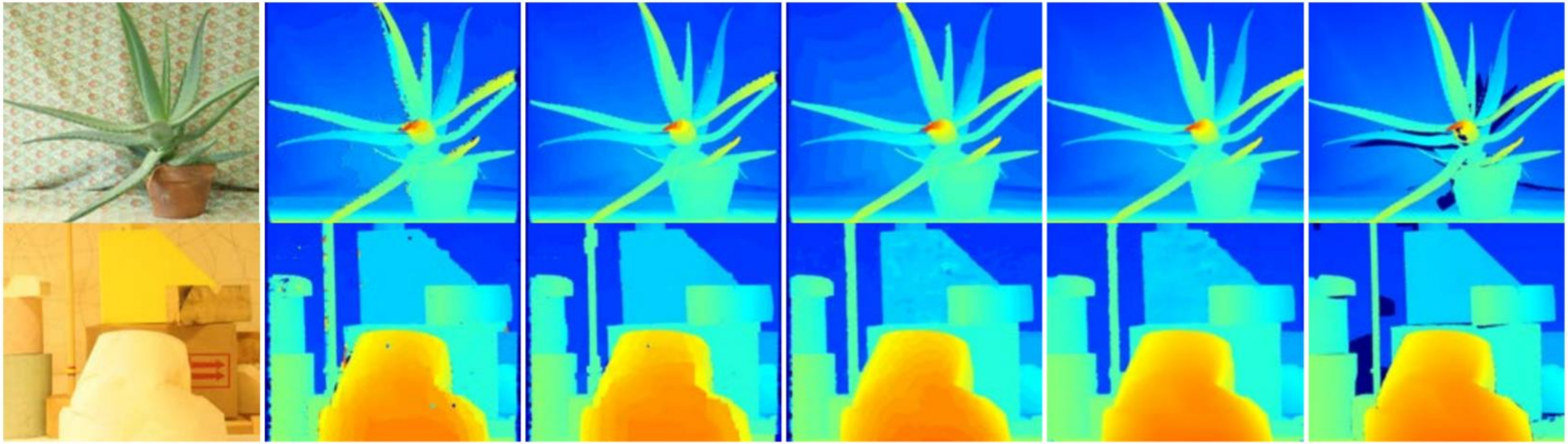
Fig. 6: The network structure of MFuseNet. For n cost volumes with size $D \times H \times W$, they are processed respectively and then fused to get the final disparity. The feature channels of 3D CNN is 4 such that the size of each cost volume before concatenation is $4 \times D \times H \times W$.



(a) Synthetic image (b) Ground truth (c) MFuseNet

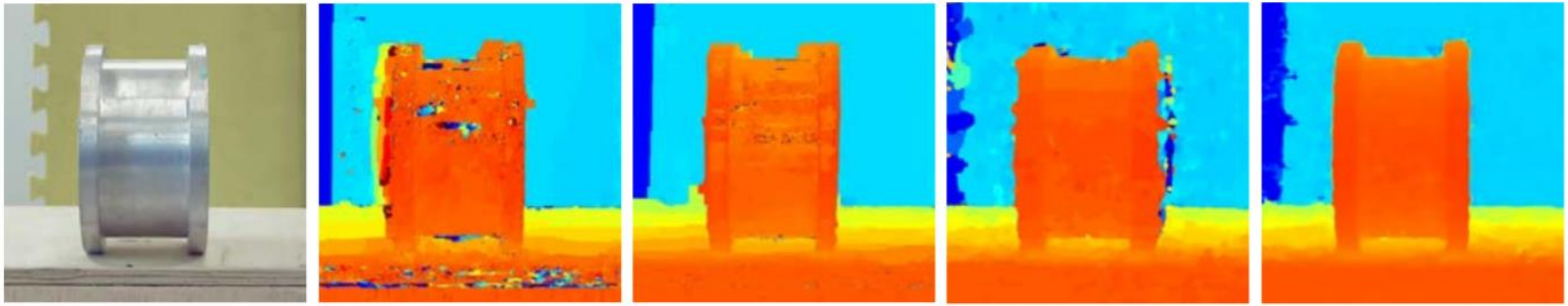
(d) Synthetic image (e) Ground truth (f) MFuseNet

Fig. 7: Color images and ground-truth disparity maps in the synthetic multiscopic dataset, and the disparity maps obtained by MFuseNet.



(a) Reference image (b) Stereo GC (c) Multiscopic GC (d) Stereo MC-CNN (e) MFuseNet Fusion (f) Ground truth

Fig. 9: The disparity estimation results of different algorithms for two sets of images, Aloe and Lampshade in the Middlebury 2006 stereo dataset. The first image is the reference RGB image, i.e., the left image for stereo algorithms and the center image for multiscopic algorithms. Two images are used for stereo algorithms, and three images are used for multiscopic algorithms.



(a) Center image (b) Stereo GC (c) Multiscopic GC (d) Stereo MC-CNN (e) MFuseNet

Fig. 11: The disparity estimation results of different algorithms for a reflective workpiece.