CVPR SEATTLE WASHINGTON
JUNE 16-18 2020

Depth Sensing Beyond LiDAR Range

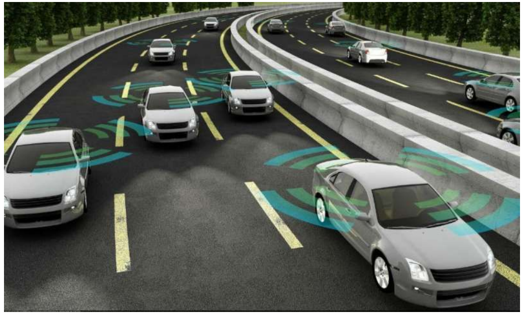
Kai Zhang — *Cornell Tech*
Jiaxin Xie — *HKUST*
Noah Snavely — *Cornell Tech*
Qifeng Chen — *HKUST*

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

Motivation

Self-driving datasets	
Kitti	80 meters
Waymo	80 meters
...	

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



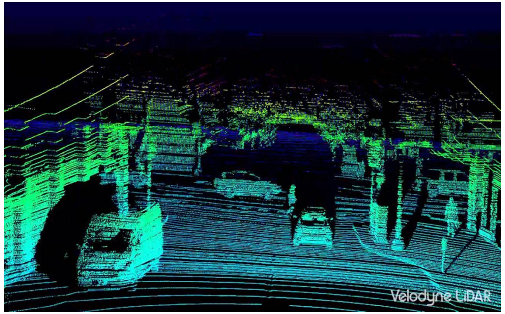
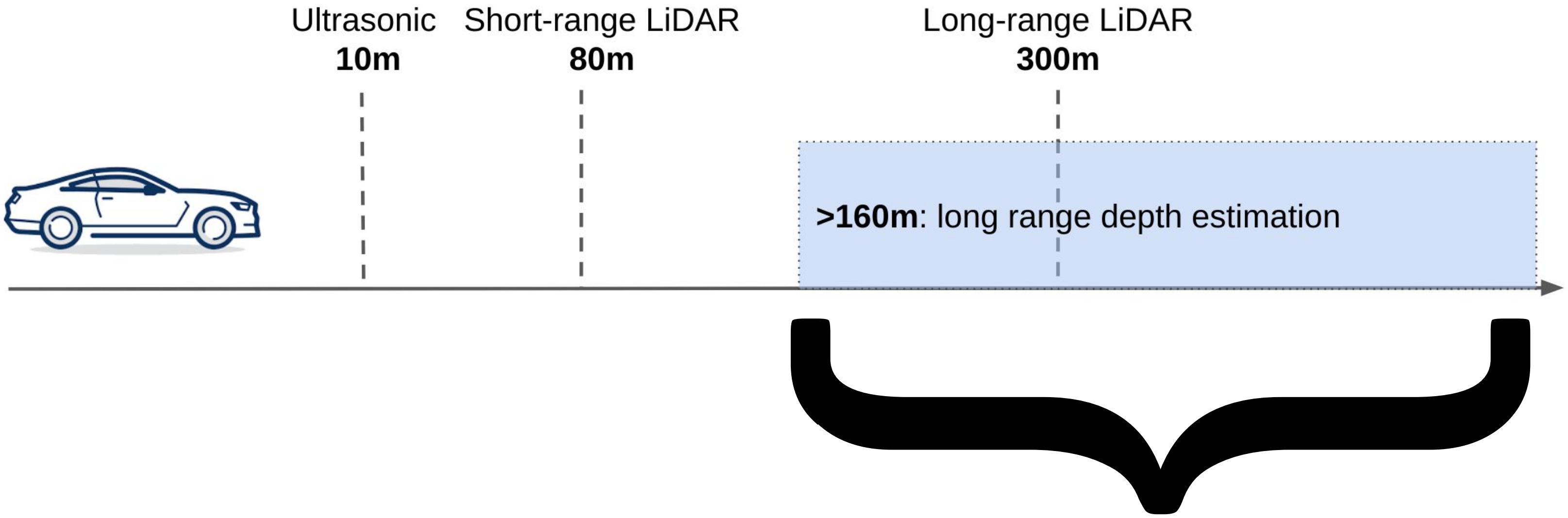


Image sources: velodyne lidar

Question: can we achieve *dense* depth sensing beyond LiDAR range with *low-cost cameras*? (e.g., >300 meters)

Example application:
Autonomous trucks driving on highway

Long-range depth sensing is hard



Long-range LiDAR: sparse and expensive

Long-range depth sensing is hard

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.

Long-range depth sensing is hard

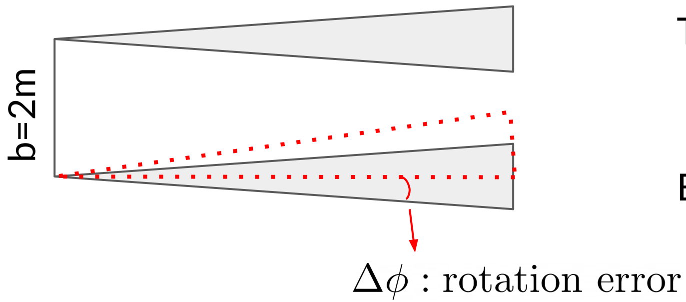
Important camera setup constraint:

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.



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

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

Relative error in estimated depth

Tentative solution - SfM

Bas-relief ambiguity in SfM[1]

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

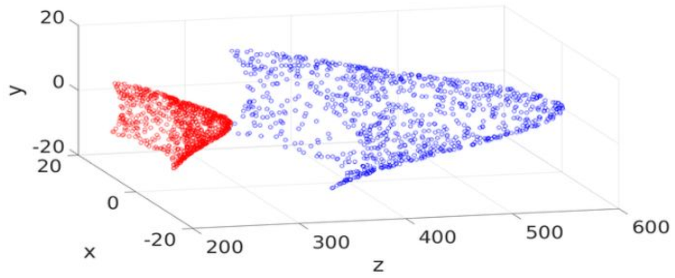


Figure 2: Ground-truth (blue) and the reconstructed (red) scene points. The unit for x, y, z axes is meter.

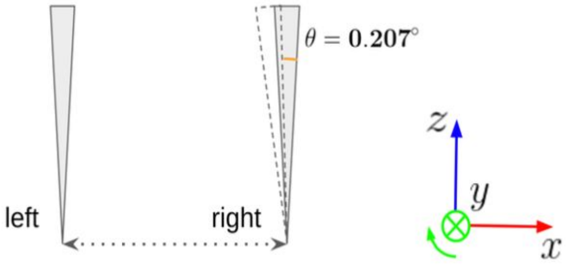
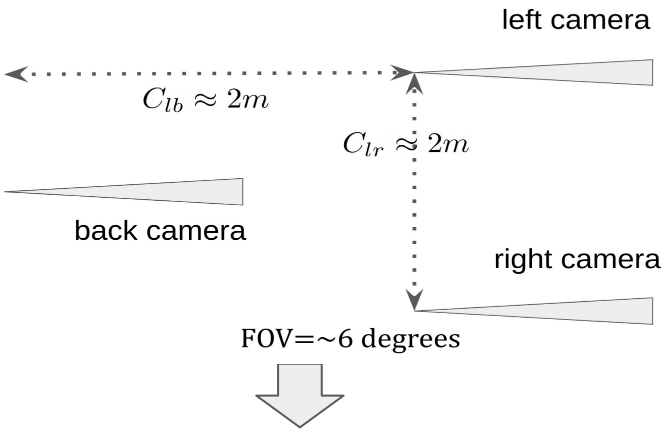
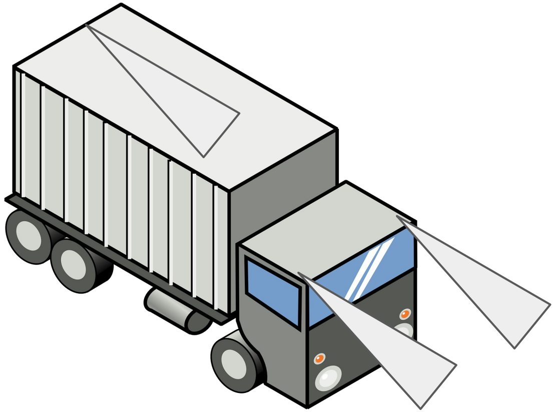


Figure 3: Top-down view of ground-truth relative pose (solid) and the recovered one (dashed). θ is exaggerated for illustration.

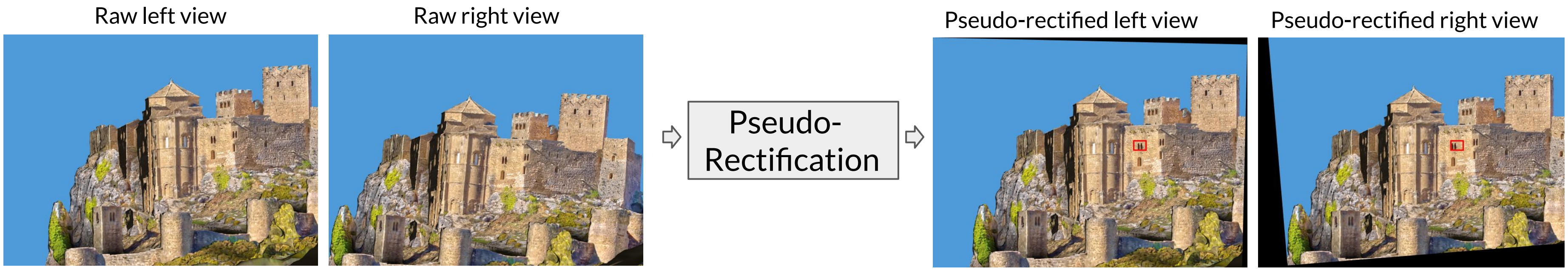
[1] Richard Szeliski and Sing Bing Kang. Shape Ambiguities in Structure From Motion. In *Proc. European Conf. on Computer Vision (ECCV)*, pages 709–721. Springer, 1996.

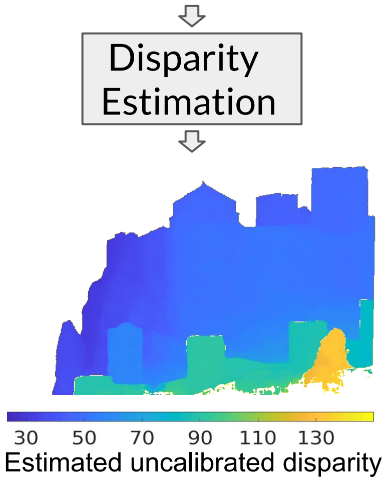
Our approach: a new three-camera vision system





Our approach: novel depth estimation pipeline





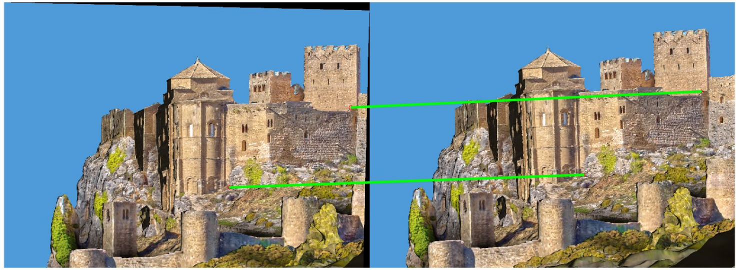
Proposition 1. *When a small-FOV camera is rotated by a small amount, the homography warping the original image to the rotated view is approximately an affine transformation.*

Our approach: novel depth estimation pipeline

Proposition 2. *For two pixels in the left image with the same depth, if they are m_l pixels apart, while their corresponding pixels in the back image are m_b pixels apart, then the depth of these two pixels in the left camera's coordinate frame is*

$$z = \frac{C_{lb}}{\frac{m_l}{m_b} - 1}. \quad (8)$$

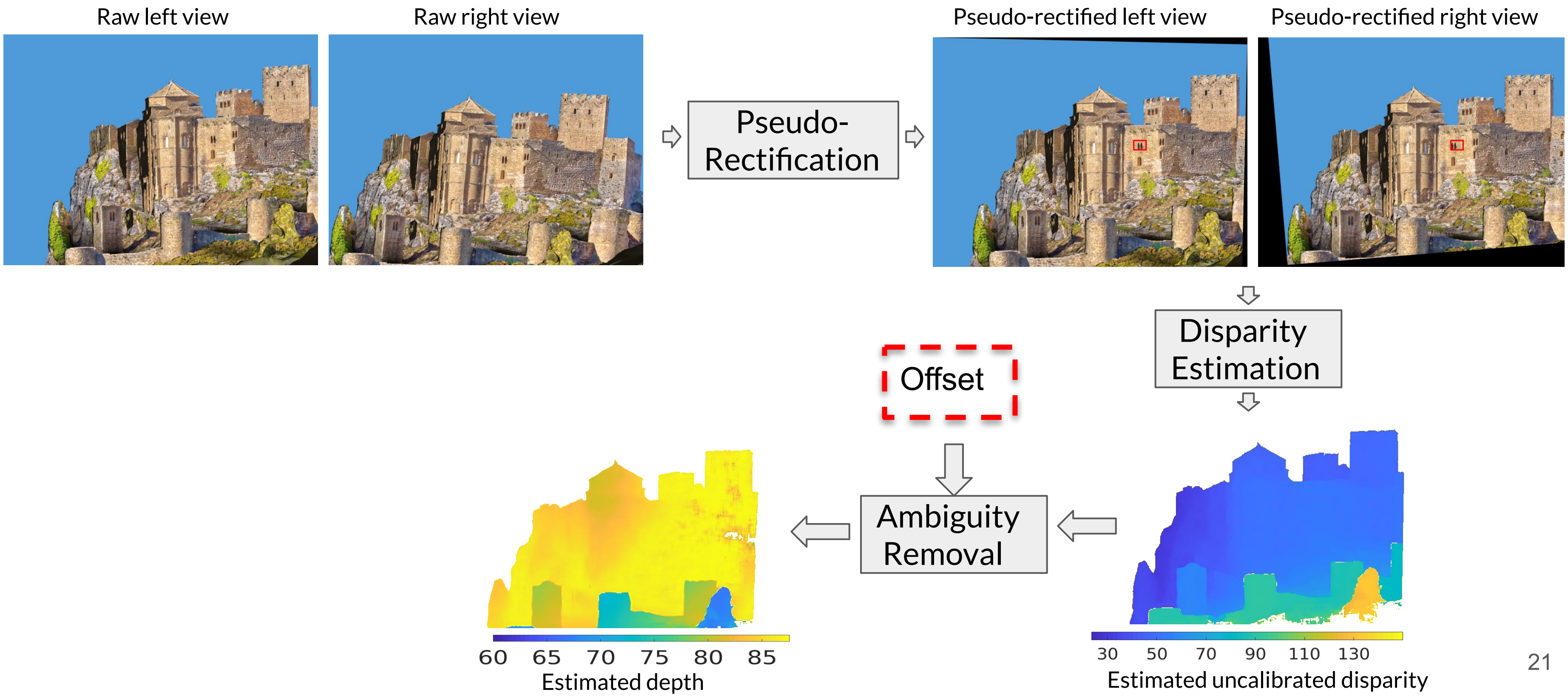
Pseudo-rectified left view   Raw back view



→ Offset

Intuition: to estimate this unknown offset, one essentially needs to know the metric depth of at least one 3D point.

Our approach: novel depth estimation pipeline



Results on synthetic data[1]

	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.

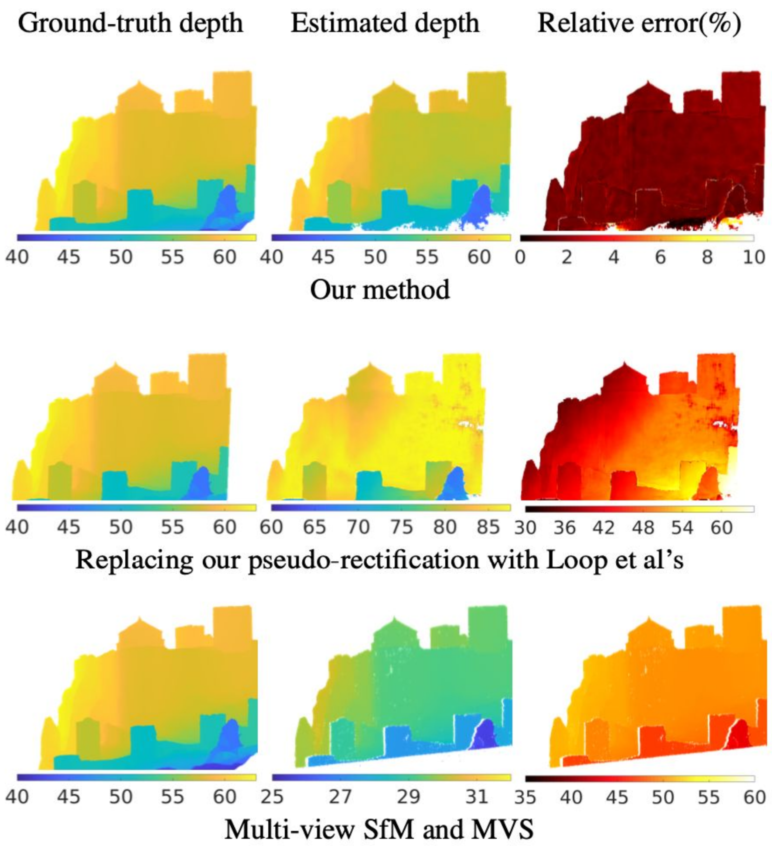
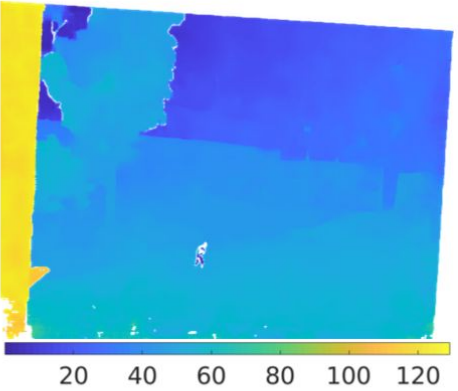


Figure 7: Comparison among different algorithms. For rectification-based methods, the ground-truth depth map has been warped to align with the rectified view. For SfM, we have used the full ground-truth intrinsic matrix.

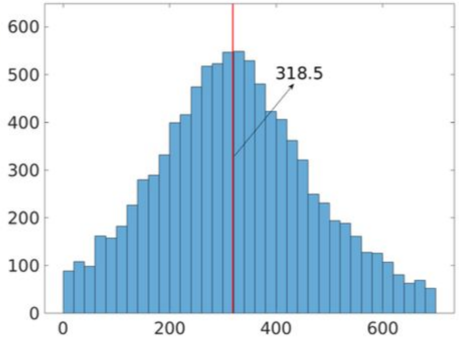
[1] Synthetic scenes might not be in their real-world scale. In experiments, we fix the baseline/depth ratio to be ~1/150.

Results on real-world data

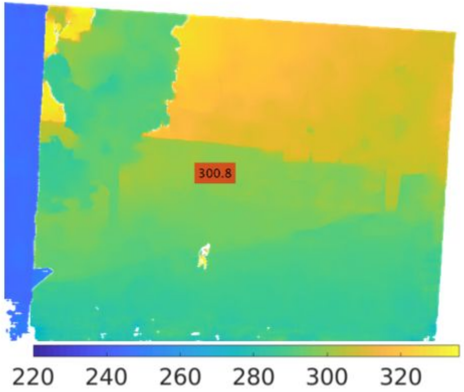
Pseudo-rectified left view



Estimated uncalibrated disparity



Estimated unknown offset



Estimated final depth

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

Advantages

- Low-cost camera-based solution;
- Not require *full* pre-calibration of camera intrinsics and extrinsics;
- Robust to small camera vibrations: important when mounted on moving vehicles.

Limitations

- Due to lack of equipment and facilities, the system has not been built and tested on the road with real-autonomous cars.
- Our method relies on stereo matching as backbone, thus suffering from common issues as stereo matching, e.g., textureless areas.

Thank you!

More technical details can be found in our paper: