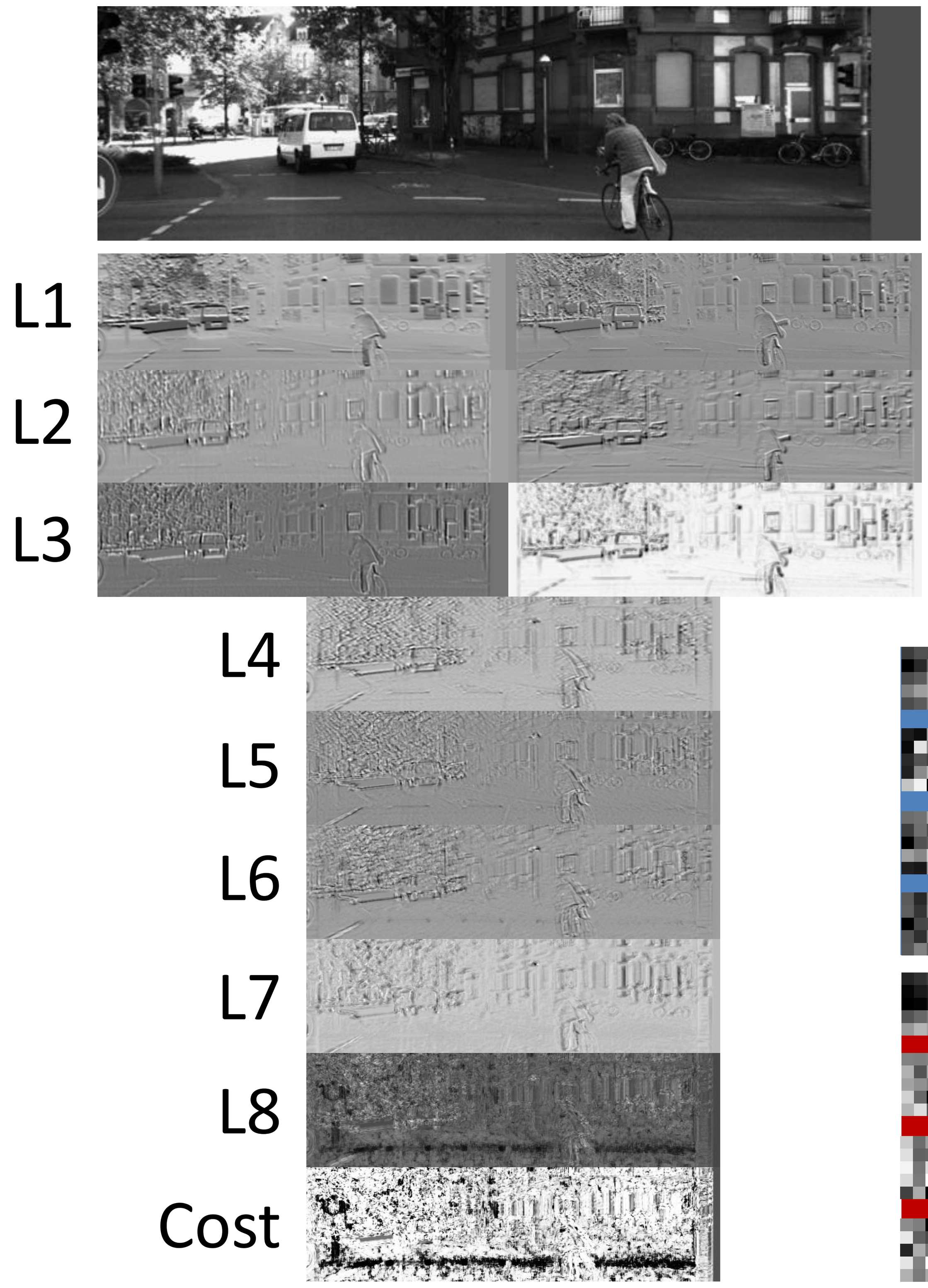
Using CNNs to Estimate Depth from Stereo Imagery

Tyler Jordan, Skanda Shridhar, Jayant Thatte
Department of Electrical Engineering, Stanford University

Motivation

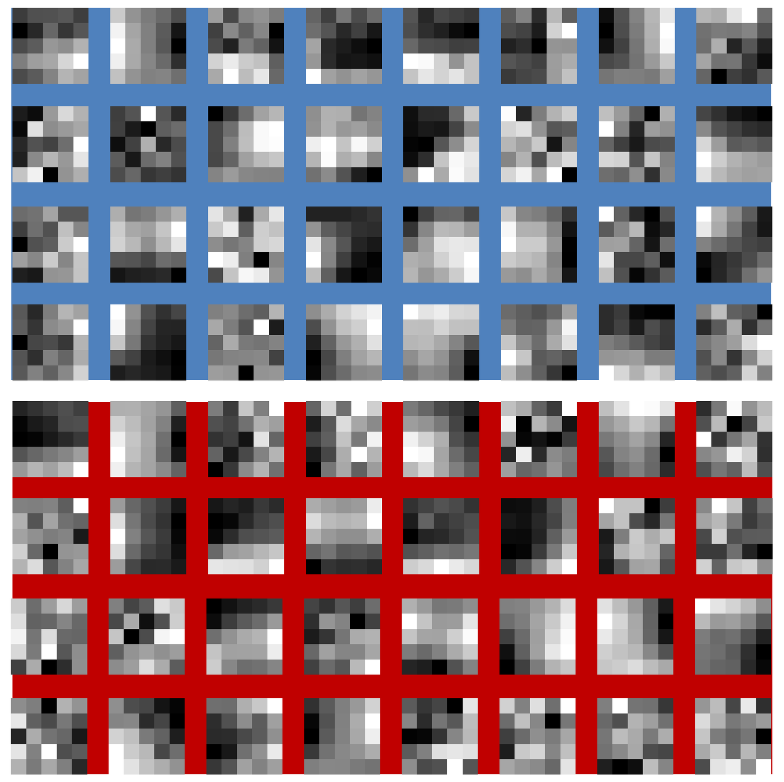
- 3D TV / Free Viewpoint TV
- Virtual Reality / Head-mounted displays
- Augmented Reality
- Computer Vision
- Autonomous Vehicles

Convolutional Neural Network

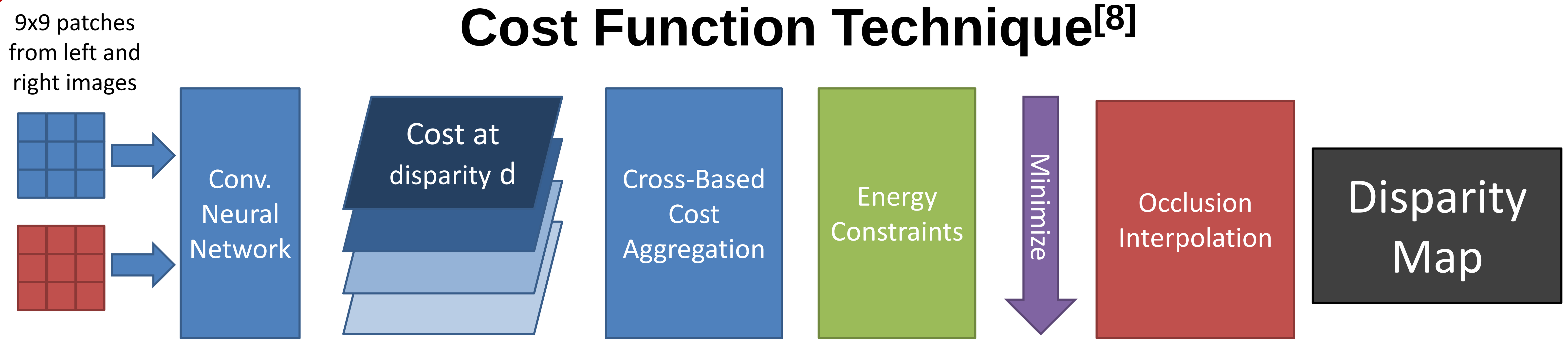


Convolutional Neural Networks are interconnected layers of artificial neurons (perceptrons) that are trained to create a model for image classification. Each layer corresponds to a set of filters which are applied to the output of the previous layer ultimately resulting in a classification label. Our CNN is trained to calculate the similarity of pixels in the stereo imagery at various disparities

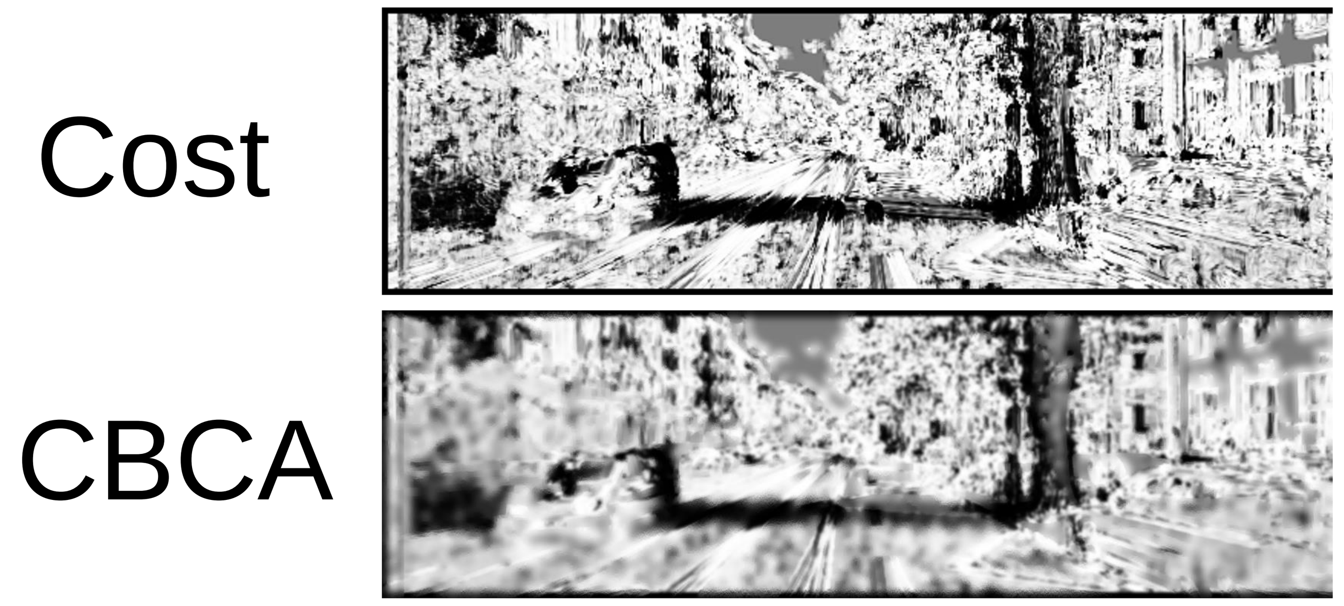
L1 Filters

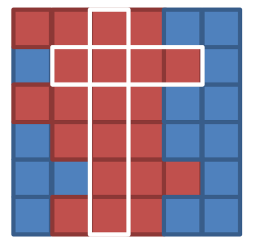


Cost Function Technique[8]



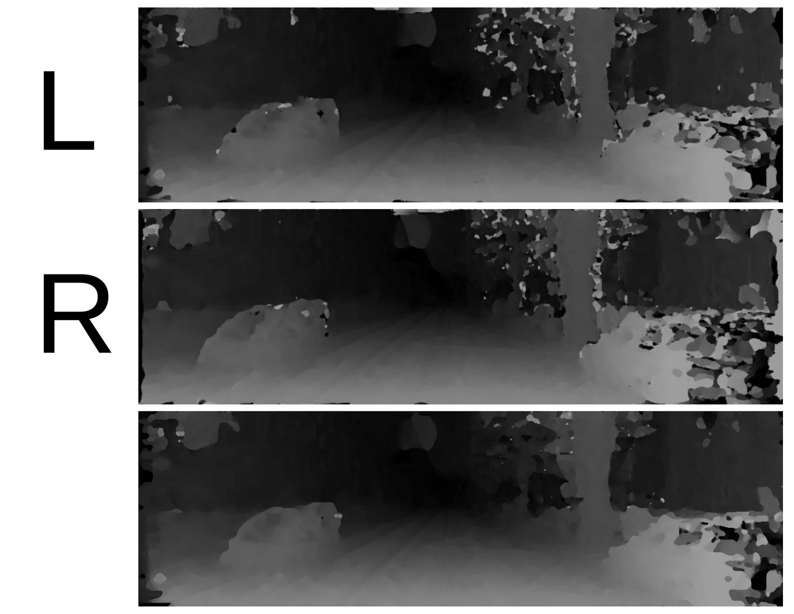
Cross-Based Cost Aggregation



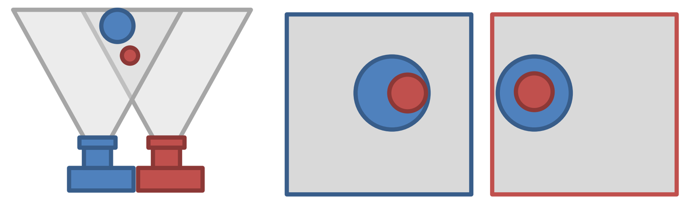


Support region (red) created by union of horizontal crosses along the vertical cross. The cross length are determined by intensity difference and length constraints. This allows for context-based blurring

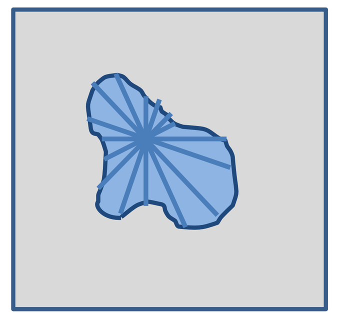
Occlusion Interpolation



Interpolation uses the depth information from the right image corresponding to the disparity in the left to fill in holes.
Regions where the right and left depth map don't agree after occlusion interpolation are filled by the median of the closest good pixels in 16 directions



Regions occluded in the left image (blue) are filled in with data from the right (red)



Experimental Results

Major objects in the scenes like the road, signs, and cars are accurate in the disparity maps. The right and left edges are not as clean as the center of the image due to the lack of redundant data. The CNN approach performs far better than the naïve plane-sweep approach.

References

1. Zhang, Ke, Jiangbo Lu, and Gauthier Lafruit. "Cross-based local stereo matching using orthogonal integral images." Circuits and Systems for Video Technology, IEEE Transactions on 19.7 (2009): 1073-1079.
2. Thomas, Graham, and Oliver Grau. "3D image sequence acquisition for TV & film production." null. IEEE, 2002.
3. Flack, Julien, Philip V. Harman, and Simon Fox. "Low-bandwidth stereoscopic image encoding and transmission." Electronic Imaging 2003. International Society for Optics and Photonics, 2003.
4. Fehn, Christoph. "Depth-image-based rendering (DIBR), compression, and transmission for a new approach on 3D-TV." Electronic Imaging 2004. International Society for Optics and Photonics, 2004.
5. Kim, Hansung, and Adrian Hilton. "3D scene reconstruction from multiple spherical stereo pairs." *International journal of computer vision* 104.1 (2013): 94-116.
6. Kim, Hansung, et al. "Dynamic 3d scene reconstruction in outdoor environments." *In Proc. IEEE Symp. on 3D Data Processing and Visualization*. 2010.
7. Schonbein, Miriam, and Andreas Geiger. "Omnidirectional 3d reconstruction in augmented manhattan worlds." *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*. IEEE, 2014.
8. Žbontar, Jure, and Yann LeCun. "Computing the stereo matching cost with a convolutional neural network." *arXiv preprint arXiv:1409.4326* (2014).