Connecting the Dots: Learning Representations for Active Monocular Depth Estimation

Gernot Riegler[1,*], Yiyi Liao[2,*], Simon Donne[2], Vladlen Koltun[1], Andreas Geiger[2]
[1]Intel Intelligent Systems Lab, [2]Autonomous Vision Group, MPI-IS / University of Tübingen

CVPR LONG BEACH CALIFORNIA June 16-20, 2019

Introduction

Passive vs. Active Stereo?
Active Stereo can reconstruct textureless regions by projecting into the scene.

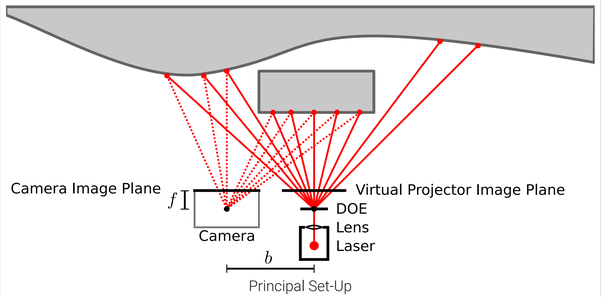
Learning representations?
We find that a conv. network is estimating disparities effectively, despite no absolute location information is encoded.

Where does the supervision come from?
Self-supervision. We use robust photometric and geometric losses.

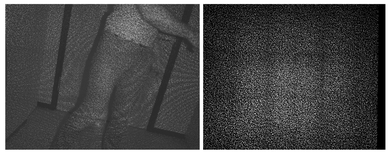
That's all?
We further utilize ambient edge information that can be disentangled from the input.

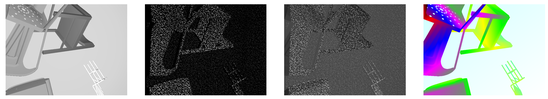
Okay, you got me. Where can I get it?
`https://github.com/autonomousvision/connecting_the_dots`

Set-Up

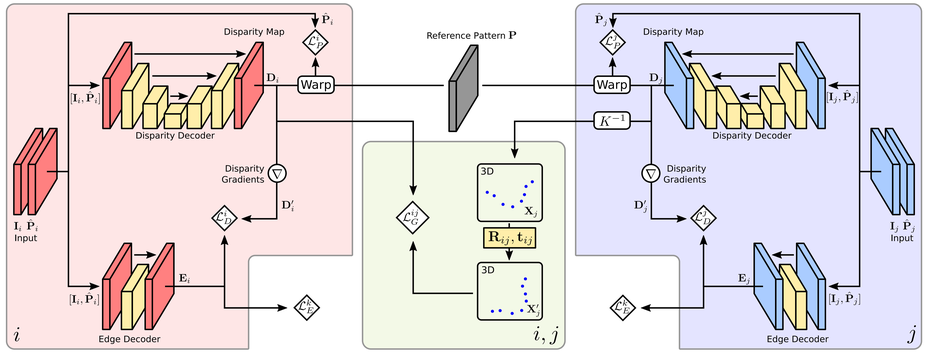


Principal Set-Up


Raw Input


Ambient | Pattern | IR Input | Disparity
Simulator to test in a controlled environment

Method Overview



Quantitative Results

	$o(0.5)$	$o(1)$	$o(2)$	$o(5)$
$\mathbf{I} \rightarrow \mathcal{D}$	**6.22**	**3.00**	**1.63**	**0.85**
$\mathbf{I} \rightarrow \mathcal{D}_{row}$	8.19	4.35	2.40	1.07
$\mathbf{I} \rightarrow \mathcal{P}$	11.83	5.08	2.46	1.11
$\mathbf{I} \rightarrow \mathcal{P}_{row}$	58.12	28.81	8.02	2.11
$\mathbf{A} \rightarrow \mathcal{D}$	90.71	81.41	63.53	32.00

Architectural Choices

	$o(0.5)$	$o(1)$	$o(2)$	$o(5)$
Supervised	6.22	3.00	1.63	0.85
$\mathcal{L}_P$	10.92	6.00	4.10	2.72
$\mathcal{L}_P + \mathcal{L}_D$	8.67	4.23	**2.56**	**1.52**
$\mathcal{L}_P + \mathcal{L}_D + \mathcal{L}_G$	**6.77**	**3.88**	2.57	1.63

Influcence of Loss Function

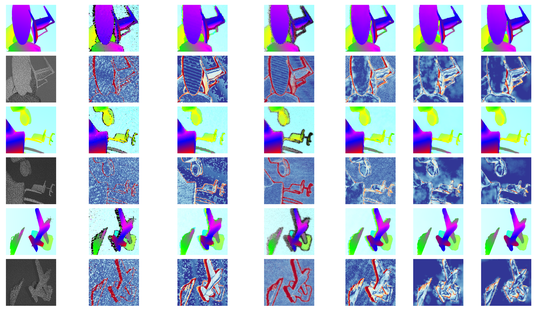
	$o(0.5)$	$o(1)$	$o(2)$	$o(5)$	$o_u(1)$	$o_u(5)$
Block Matching	7.84	7.20	7.06	6.83	4.44	4.23
FastMRF [1]	12.07	8.36	6.71	5.14	5.25	3.57
HyperDepth [2]	15.01	12.63	11.83	11.49	7.39	6.73
Ours	**6.77**	**3.88**	**2.57**	**1.63**	**1.75**	**0.70**

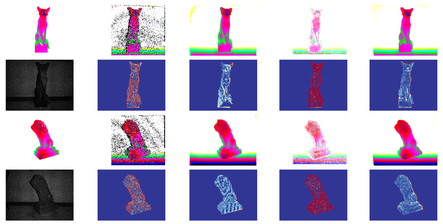
Comparison to baselines

	acc	comp	h. mean
Block Matching	551.082	3.883	7.712
FastMRF [1]	12.690	6.971	8.999
HyperDepth [2]	**8.759**	5.263	6.575
Ours	11.042	**3.147**	**4.898**

Evaluation on dataset of [1]

Qualitative Results


GT, Input | Block Matching | FastMRF [1] | HyperDepth [2] | Ours | Ours | Ours
Results on synthetic data


GT, Input | Block Matching | FastMRF [1] | HyperDepth [2] | Ours
Results on dataset of [1]

Method Details

Photometric Loss - Minimize smooth census loss [3] to maximize photometric consistency:

$$\mathcal{L}_P^i(\hat{\mathbf{P}}_i, \mathbf{P}, \mathbf{D}_i) = \sum_{x,y} \|\hat{\mathbf{p}}_i(x,y) - \mathbf{p}(x - D_i(x,y), y)\|_C \,. \quad (1)$$

Disparity Loss - Encourage disparity discontinuities at edges of the ambient image:

$$\mathcal{L}_D^i(\mathbf{D}'_i, \mathbf{E}_i) = \sum_{x,y} -\log\left(p(D'_i(x,y), E_i(x,y))\right) \,. \quad (2)$$

Edge Loss - Avoid trivial solution to disparity loss:

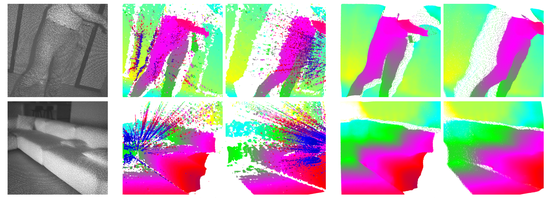
$$\mathcal{L}_E^k(\mathbf{E}_k, \mathbf{A}_k) = -\sum_{x,y} A'_k(x,y) \log E_k(x,y) + w\,(1 - A'_k(x,y)) \log(1 - E_k(x,y)) \,. \quad (3)$$

Geometric Loss - Utilize multiple views of the same static scene:

$$\mathcal{L}_G^{ij}(\mathbf{D}_i, \mathbf{X}'_j) = \sum_{\mathbf{x} \in \mathbf{X}'_j} \min\left(\left|\mathbf{x}_z - bf D_i^{-1}(\mathbf{K}\,\mathbf{x})\right|, \tau\right) \,. \quad (4)$$

$\mathbf{I}$ . . . Input image
$\mathbf{D}$ . . . Predicted disparity map
$\hat{\mathbf{P}}$ . . . LCN input image
$\mathbf{A}$ . . . Ambient image
$\mathbf{P}$ . . . LCN reference pattern
$\mathbf{E}$ . . . Predicted edge map
$\mathbf{D}'$ . . . Disparity magnitude $= |\nabla \mathbf{D}|$
$\mathbf{X}$ . . . Point cloud from predicted disparity map

More Qualitative Results


Input | Block Matching | Ours
Results on own Kinect recordings

Conclusion

- It is possible to train a network for disparity estimation in a self-supervised fashion
- Obtains similar results to training on ground-truth disparities
- Modeling joint distribution of depth and image edges improves depth boundaries
- Constraints on discontinuities from multiple views improves performance

References

[1] Q. Chen and V. Koltun. "Fast MRF Optimization with Application to Depth Reconstruction". In: *CVPR*. 2014.
[2] S. R. Fanello et al. "HyperDepth: Learning Depth from Structured Light without Matching". In: *CVPR*. 2016.
[3] D. Hafner et al. "Why Is the Census Transform Good for Robust Optic Flow Computation?" In: *SSVM*. 2013.