Learning 3D representations, disparity estimation, and structure from motion

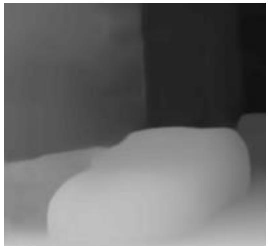
Thomas Brox
University of Freiburg, Germany

Research funded by the ERC Starting Grant VideoLearn, the German Research Foundation, and the Deutsche Telekom Stiftung

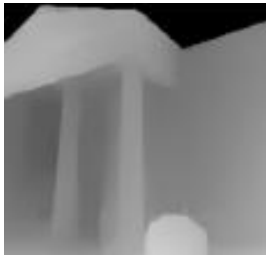
Outline

3D shape and texture from a single image

FlowNet: end-to-end optical flow



DispNet: end-to-end disparities

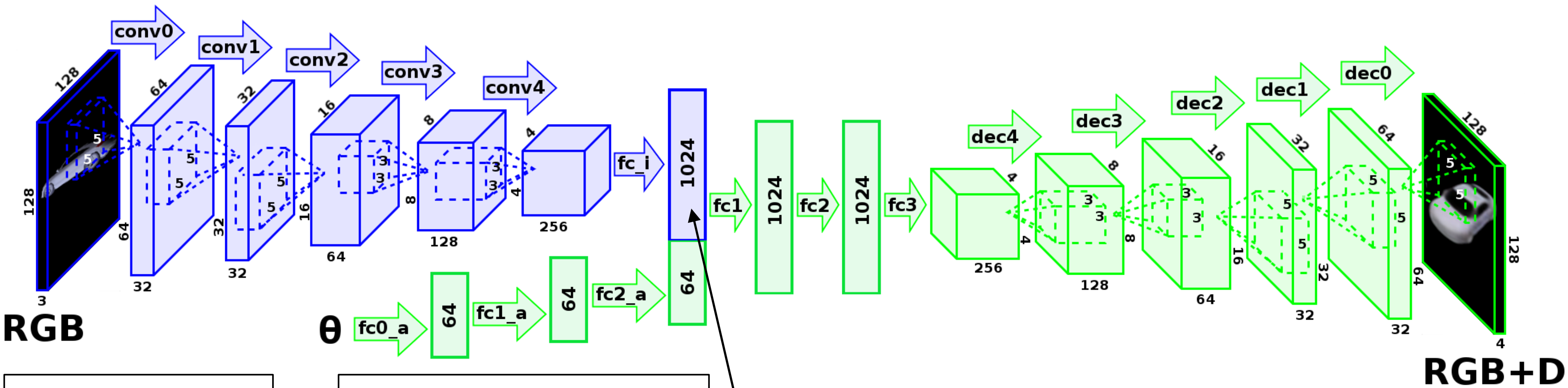


DeMoN: end-to-end structure from motion

Single-view to multi-view

Maxim Tatarchenko
Alexey Dosovitskiy
ECCV 2016

Up-convolutional part



New image from arbitrary view

Choose desired output view

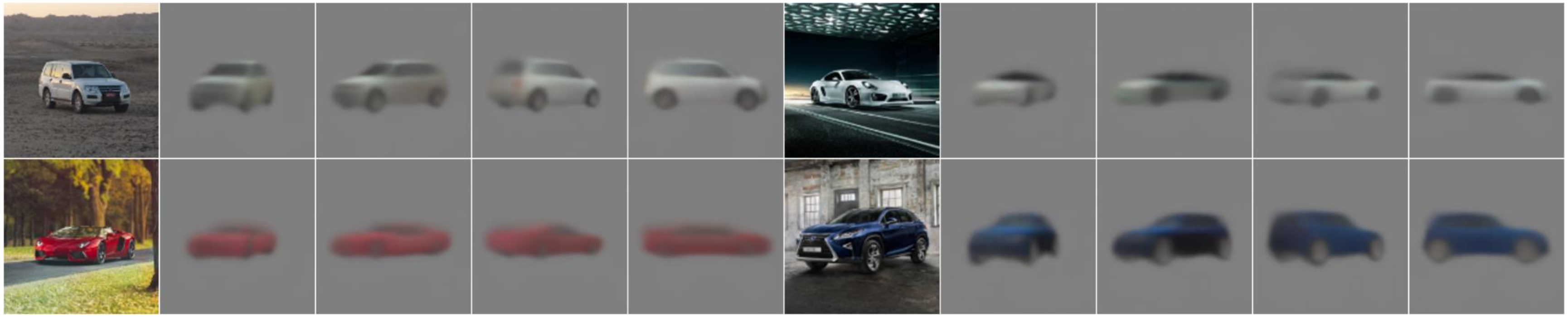
Additional depth map

Analysis part

Canonical 3D representation?

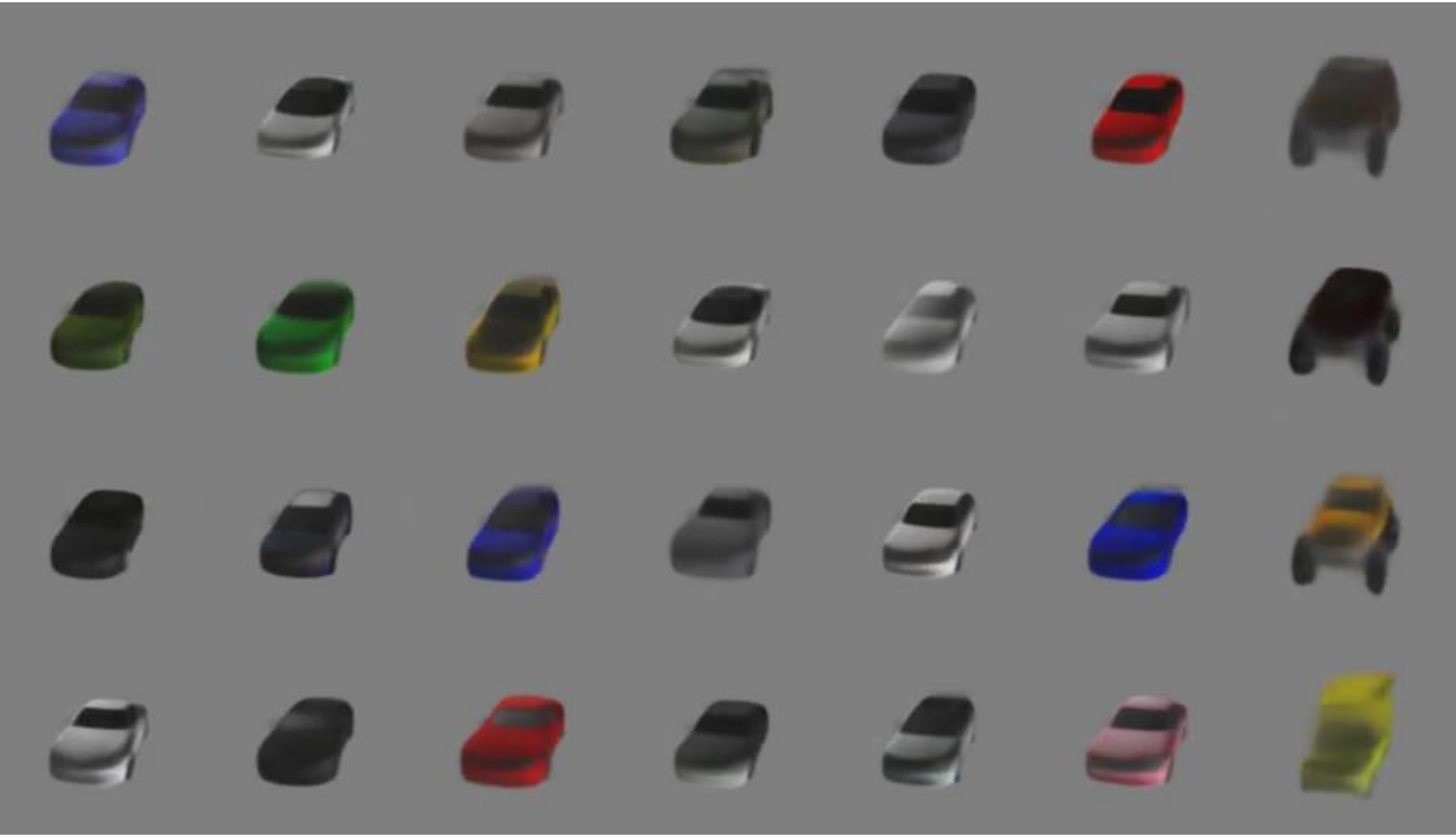
Single-view to multi-view

Synthetic images

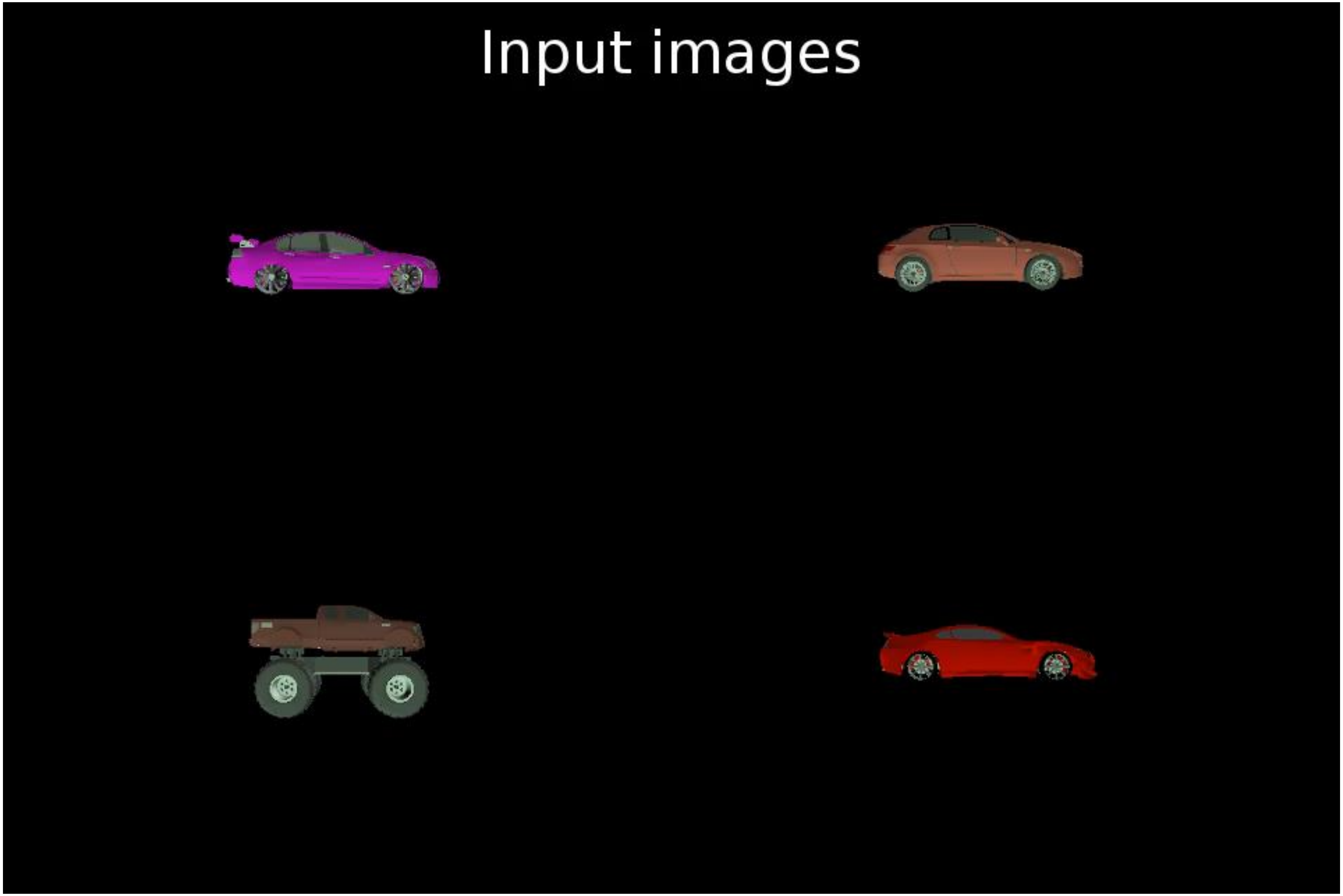


Real images

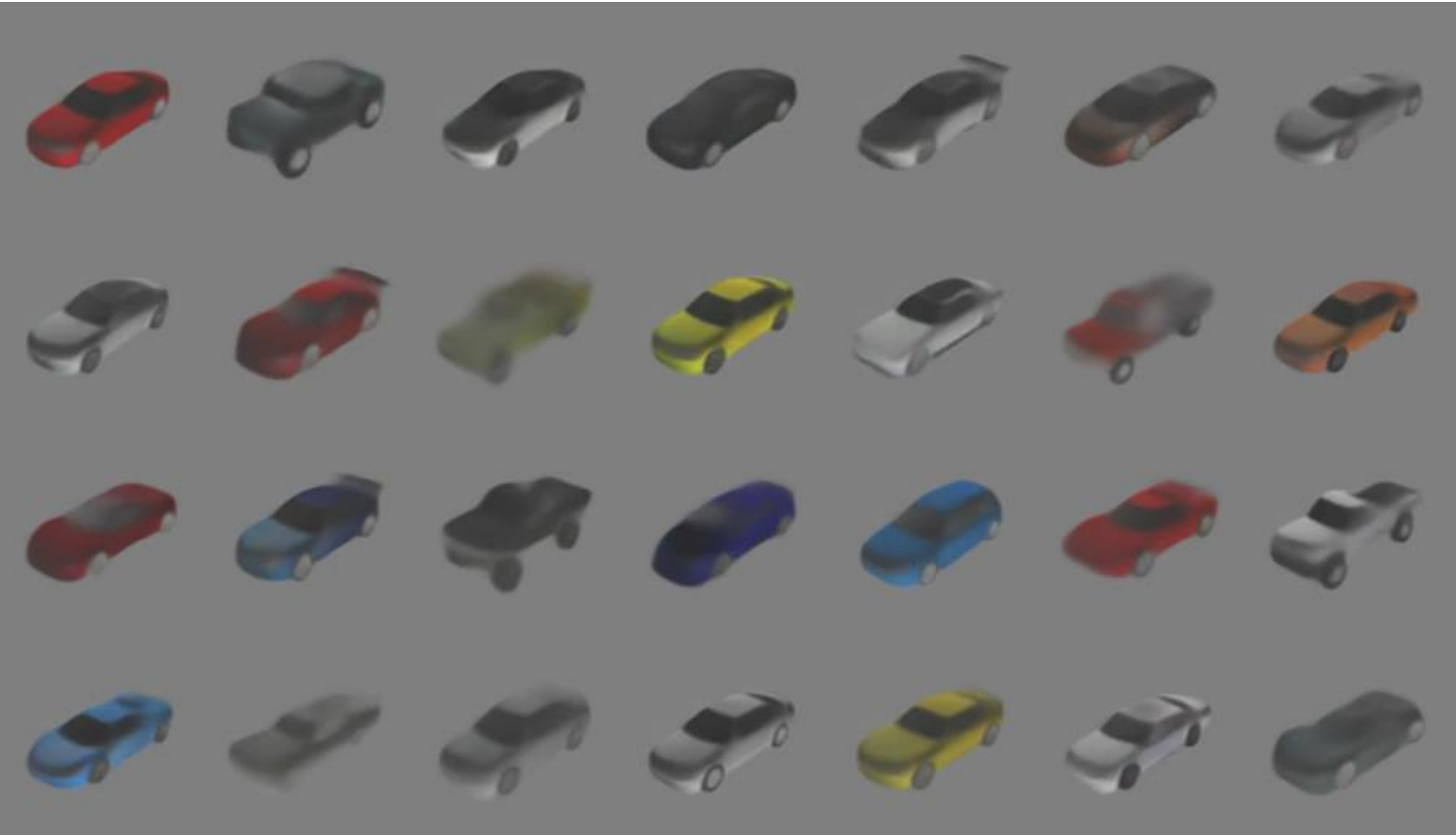
Multi-view looks like 3D



Reconstructing explicit 3D models



Multiview morphing

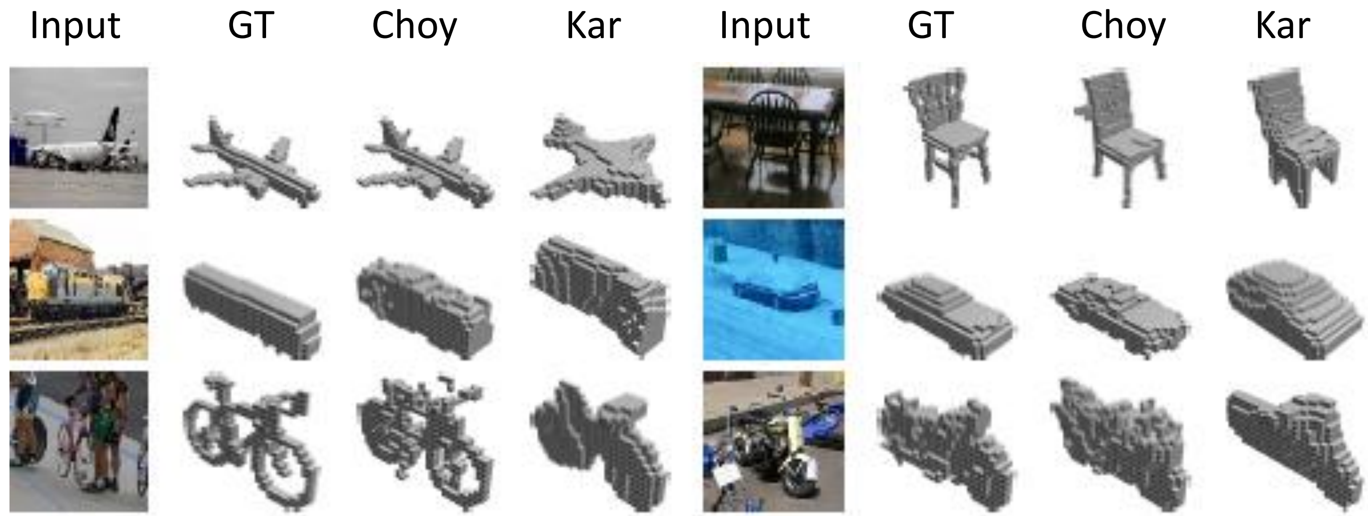


Other interesting work

Yang et al. NIPS 2015
Recurrent network, incrementally rotates the object

Ours for comparison

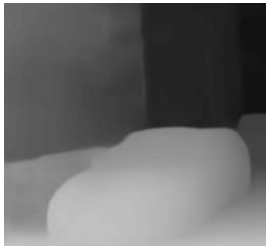
Kar et al. CVPR 2015
Choy et al. 2016



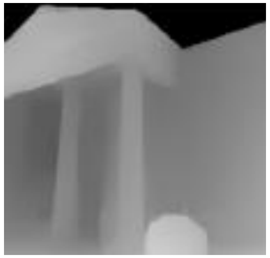
Outline

3D shape and texture from a single image

FlowNet: end-to-end optical flow

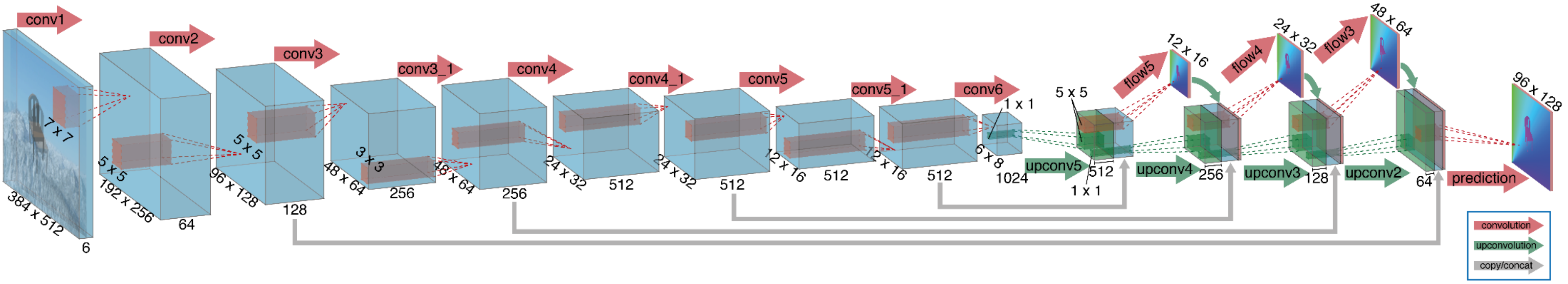


DispNet: end-to-end disparities



DeMoN: end-to-end structure from motion

FlowNet: estimating optical flow with a ConvNet



- Can networks learn to find correspondences?

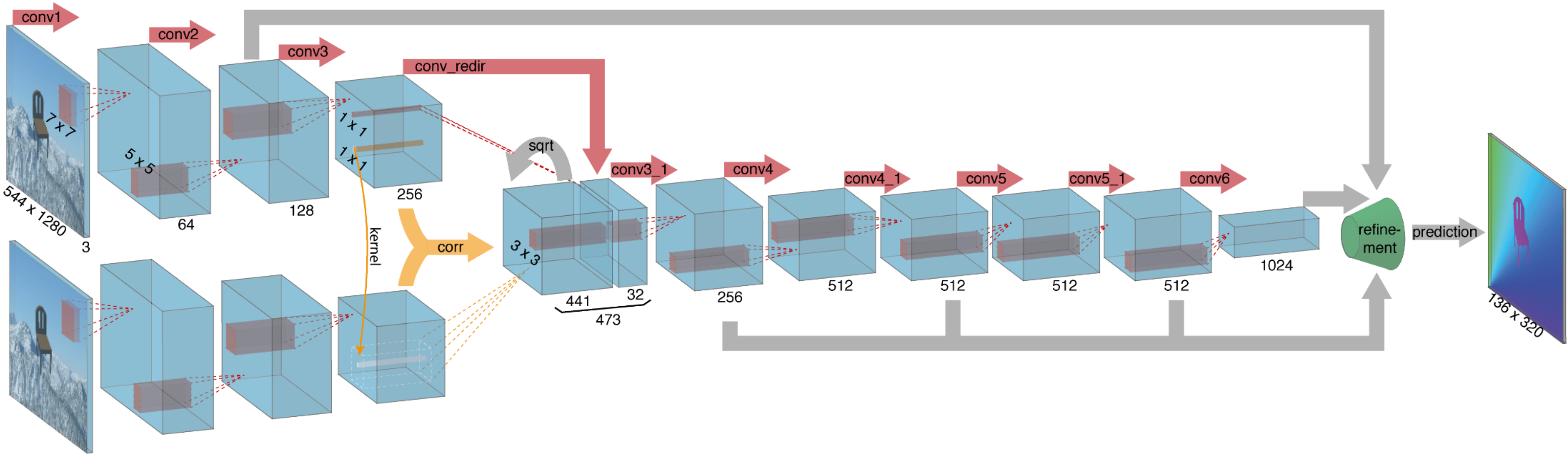
- New learning task!

  (very different from classification, etc.)

Dosovitskiy et al.
ICCV 2015

Can networks learn to find correspondences?

→Help the network with an explicit correlation layer



Dosovitskiy et al.
ICCV 2015

Enough data to train such a network?

- Getting ground truth optical flow for realistic videos is hard

- Existing datasets are small:

	Frames with ground truth
Middlebury	8
KITTI	194
Sintel	1041
Needed	>10000

Realism is overrated: the “flying chairs” dataset

Image pair — Optical flow

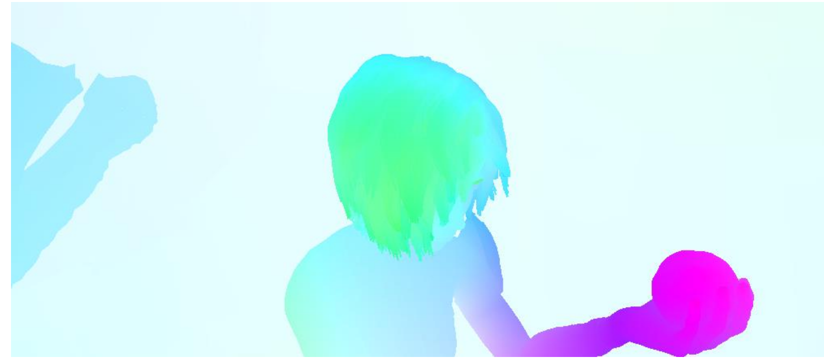
Synthetic 3D datasets

Mayer et al. CVPR 2016

Driving, Monkaa, FlyingThings3D datasets publicly available

Generalization: it works!

Input images



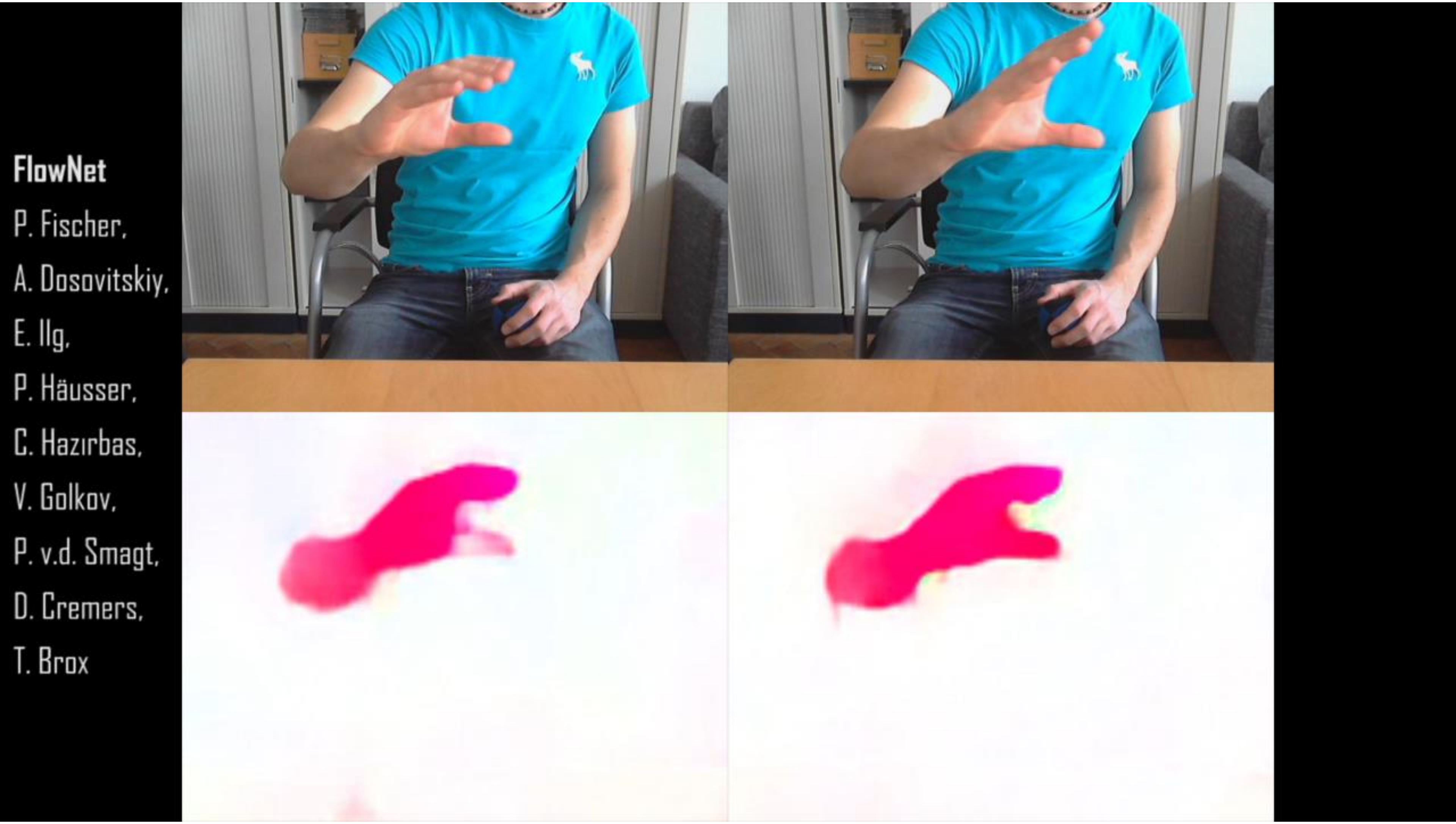
Ground truth

FlowNetSimple

FlowNetCorr

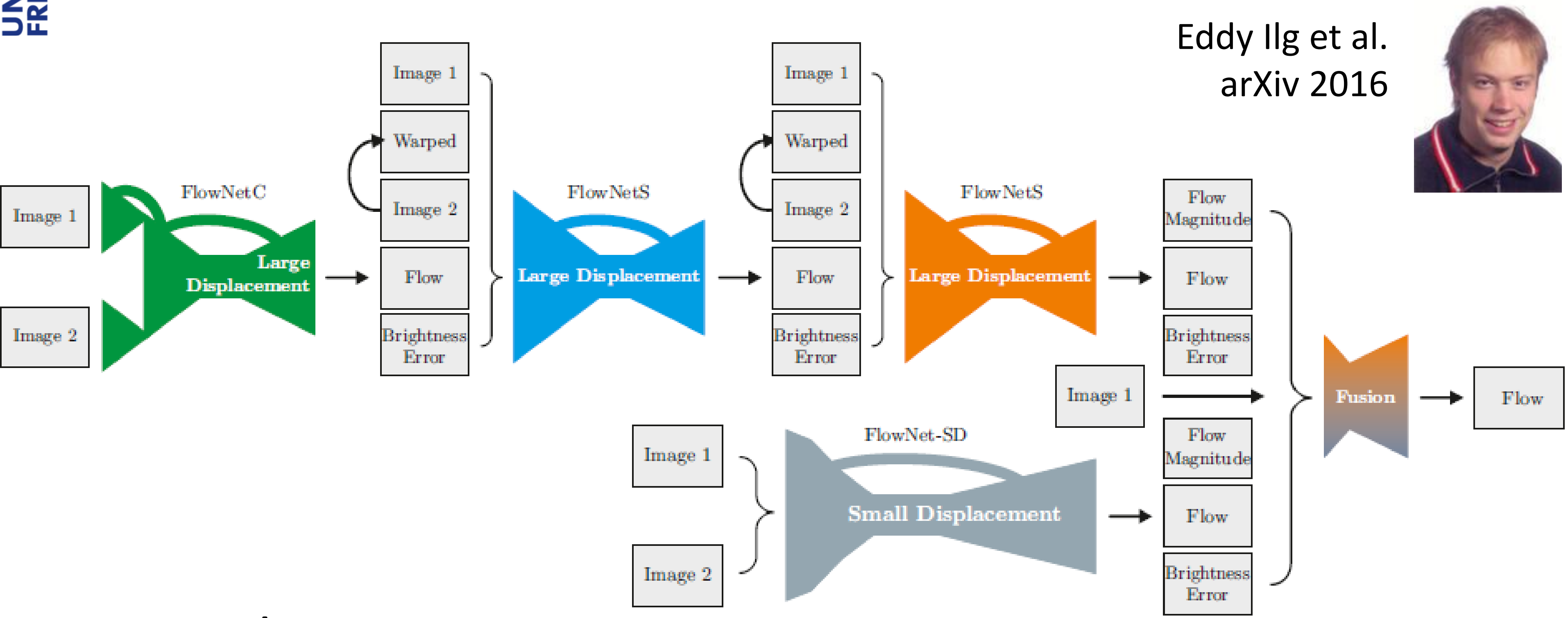
Although the network has only seen flying chairs for training, it predicts good optical flow on other data

Optical flow estimation in 18ms



FlowNet 2.0

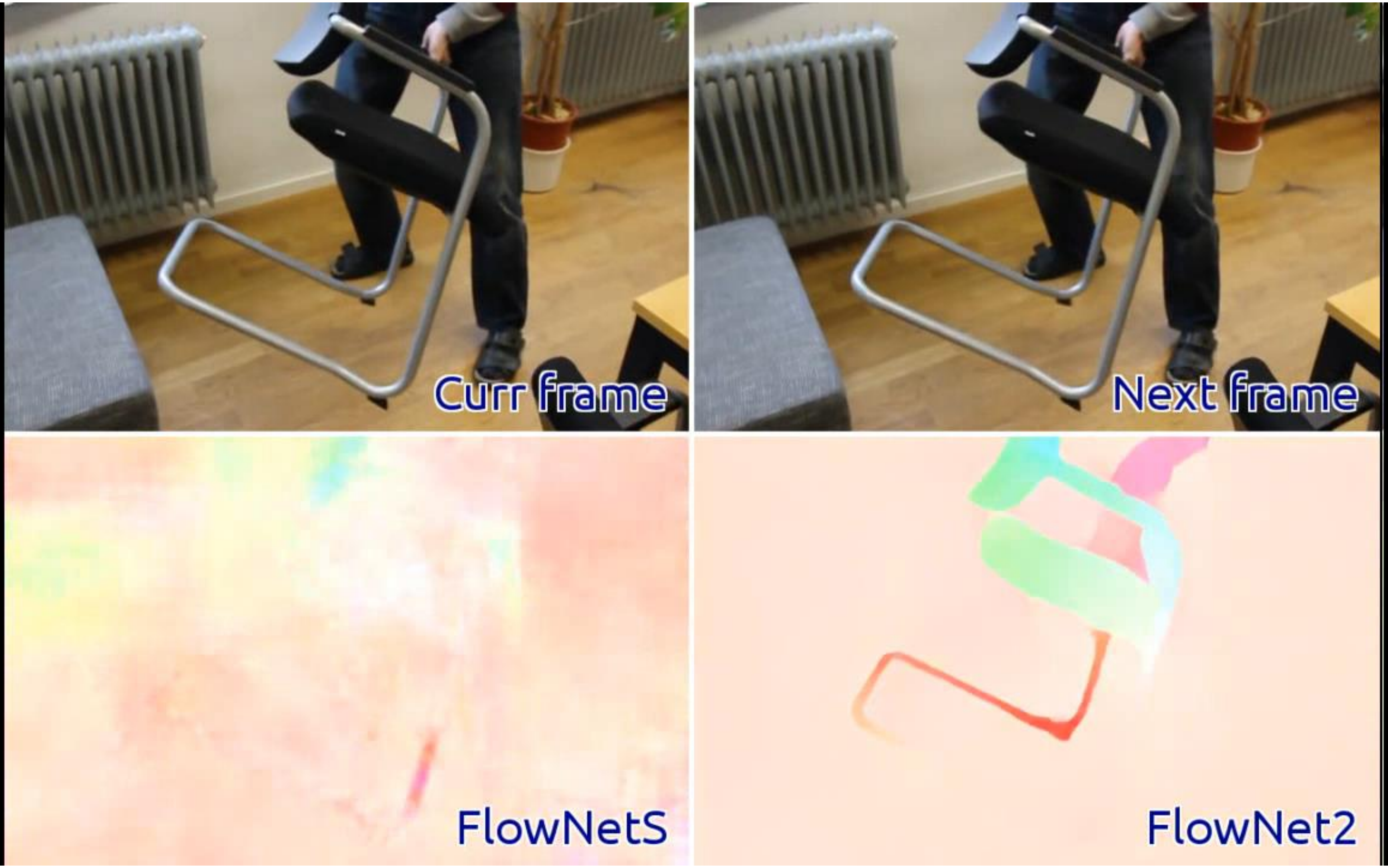
Eddy Ilg et al.
arXiv 2016



Major changes:

- Improved data and training schedules
- Stacking of networks with motion compensation
- Special small displacements and fusion network

FlowNet vs. FlowNet 2.0

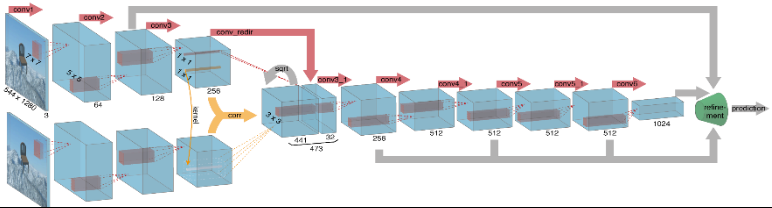


Numbers....

	Sintel	KITTI	runtime
DeepFlow (Weinzaepfel et al. 2013)	7.21	5.8	51940 ms
FlowFields (Bailer et al. 2015)	5.81	3.5	22810 ms
PCA Flow (Wulff & Black 2015)	8.65	6.2	140 ms
FlowNet (Dosovitskiy et al. 2015)	7.52	-	18 ms
FlowNet 2.0	**5.74**	**1.8**	123 ms

DispNet: disparity estimation

Mayer et al.
CVPR 2016

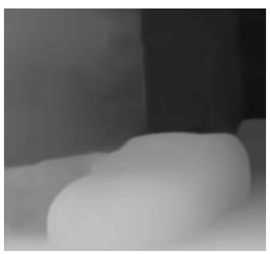


DispNet: disparity estimation

Outline

3D shape and texture from a single image

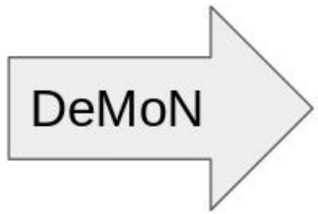
FlowNet: end-to-end optical flow

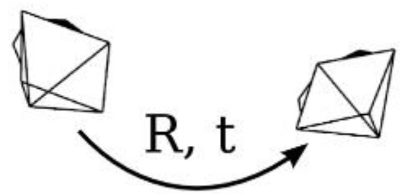


DispNet: end-to-end disparities

DeMoN: end-to-end structure from motion

DeMoN: Structure from motion with a network

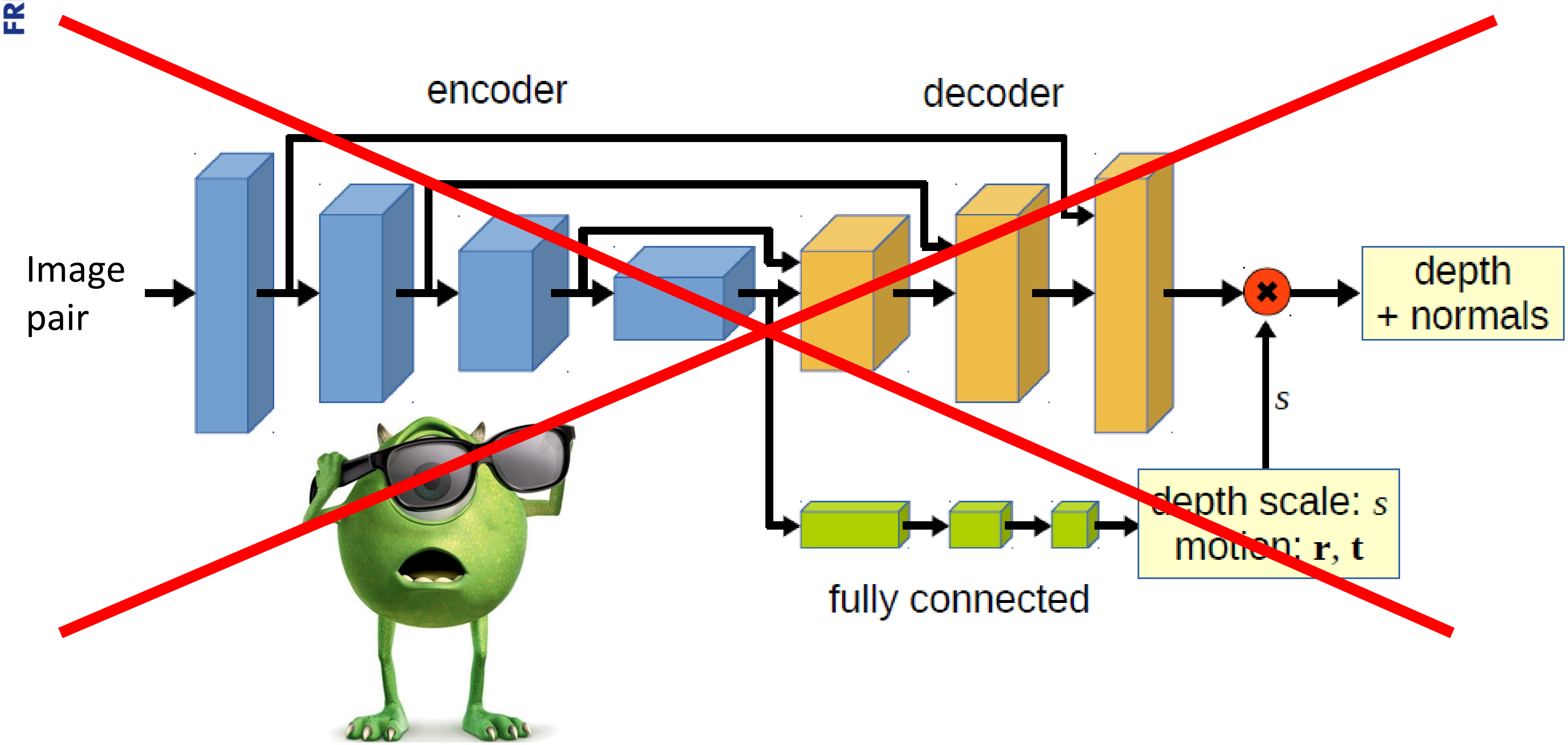




Benjamin Ummenhofer
Huizhong Zhou
arxiv 2016

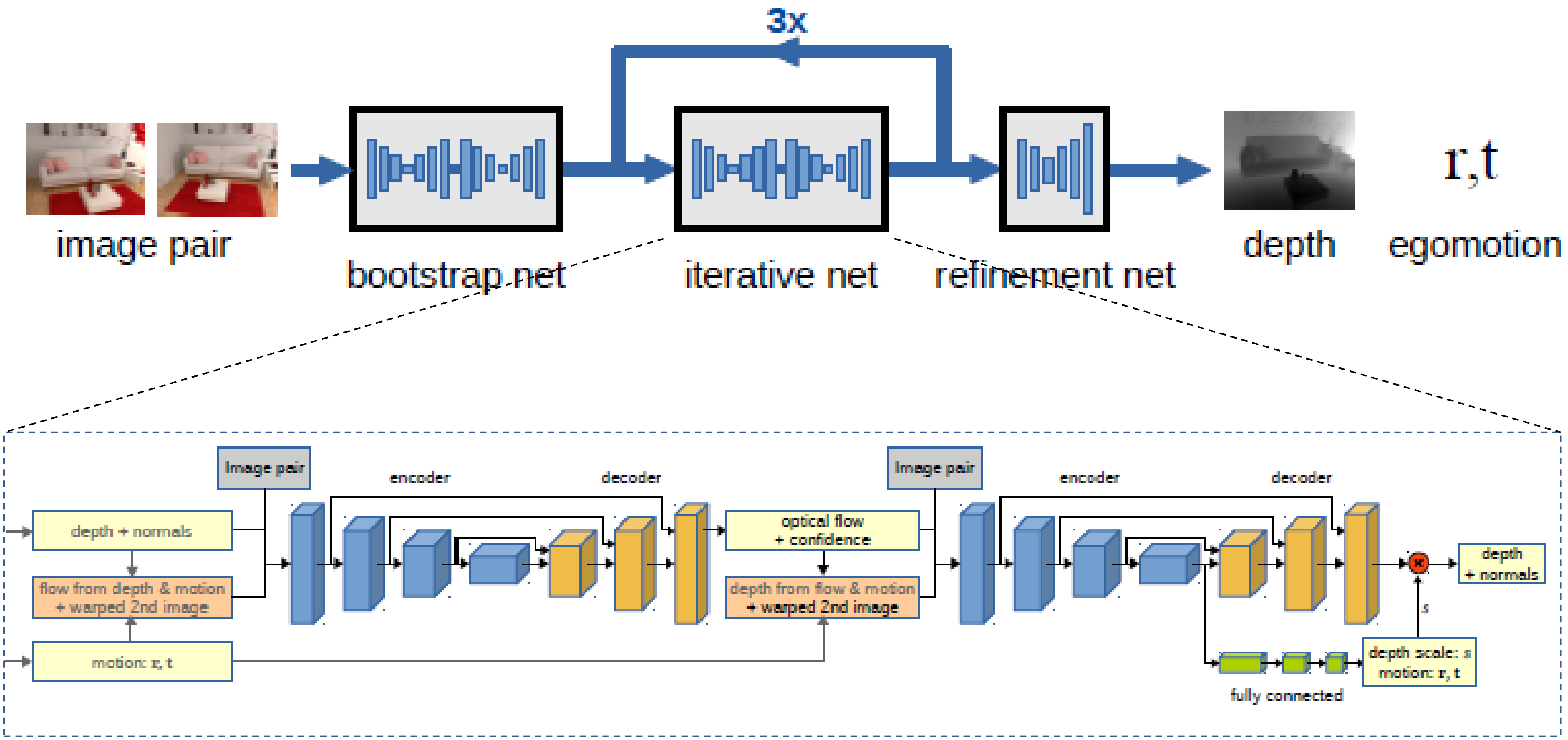
Egomotion estimation and depth estimation are mutually dependent

Straightforward idea



Network ignores the second image
→ motion parallax not learned

DeMoN architecture

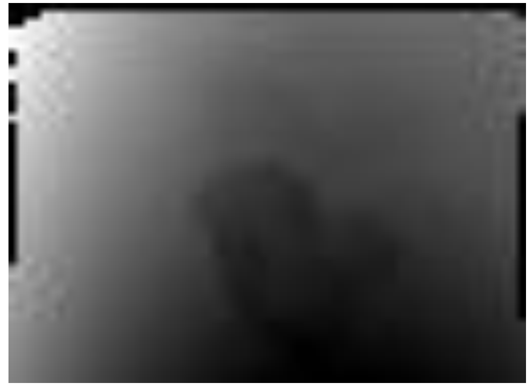


Estimates optical flow

Estimates depth and egomotion

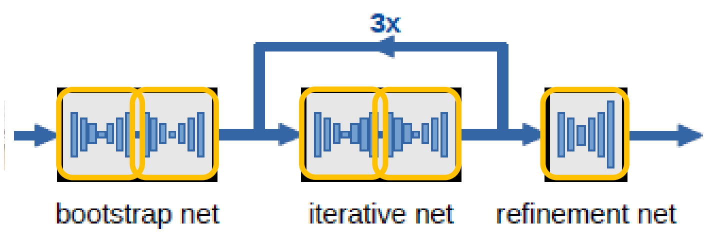
Iterative refinement

Input images

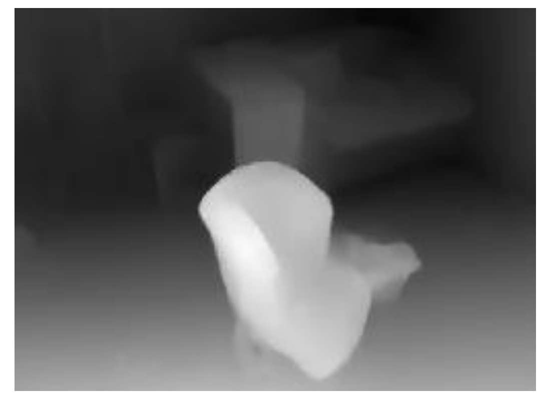


Ground truth
Optical Flow

Estimated optical flow

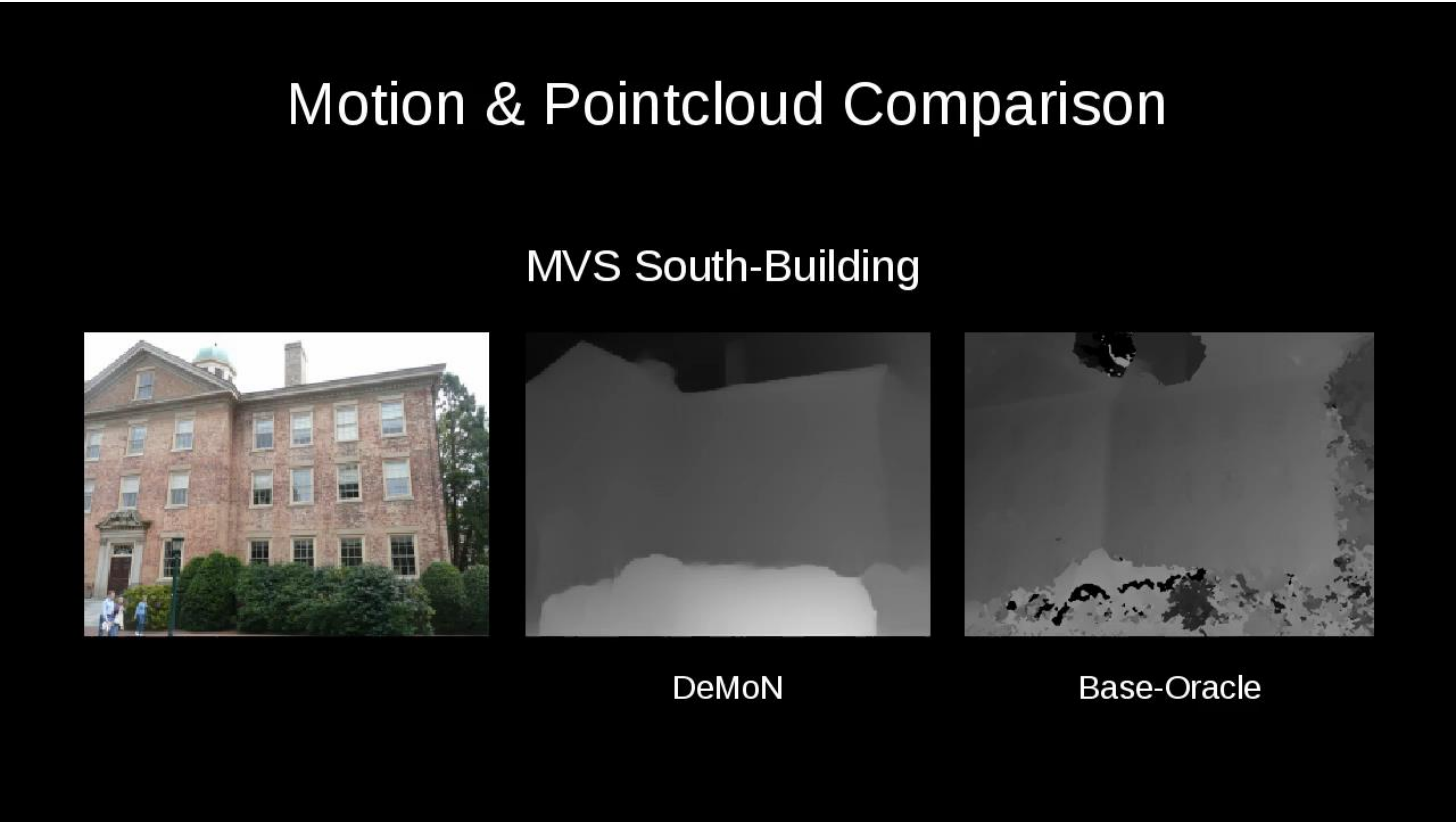


Ground truth
Depth



Estimated depth

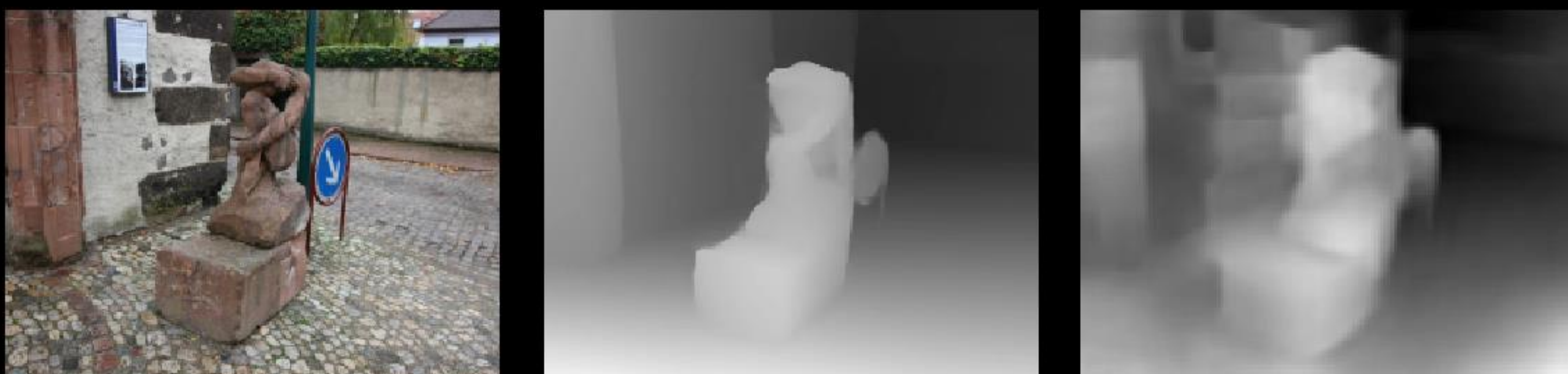
Outperforms two-frame SfM baselines



Two images generalize better than one image

Pointcloud Comparison

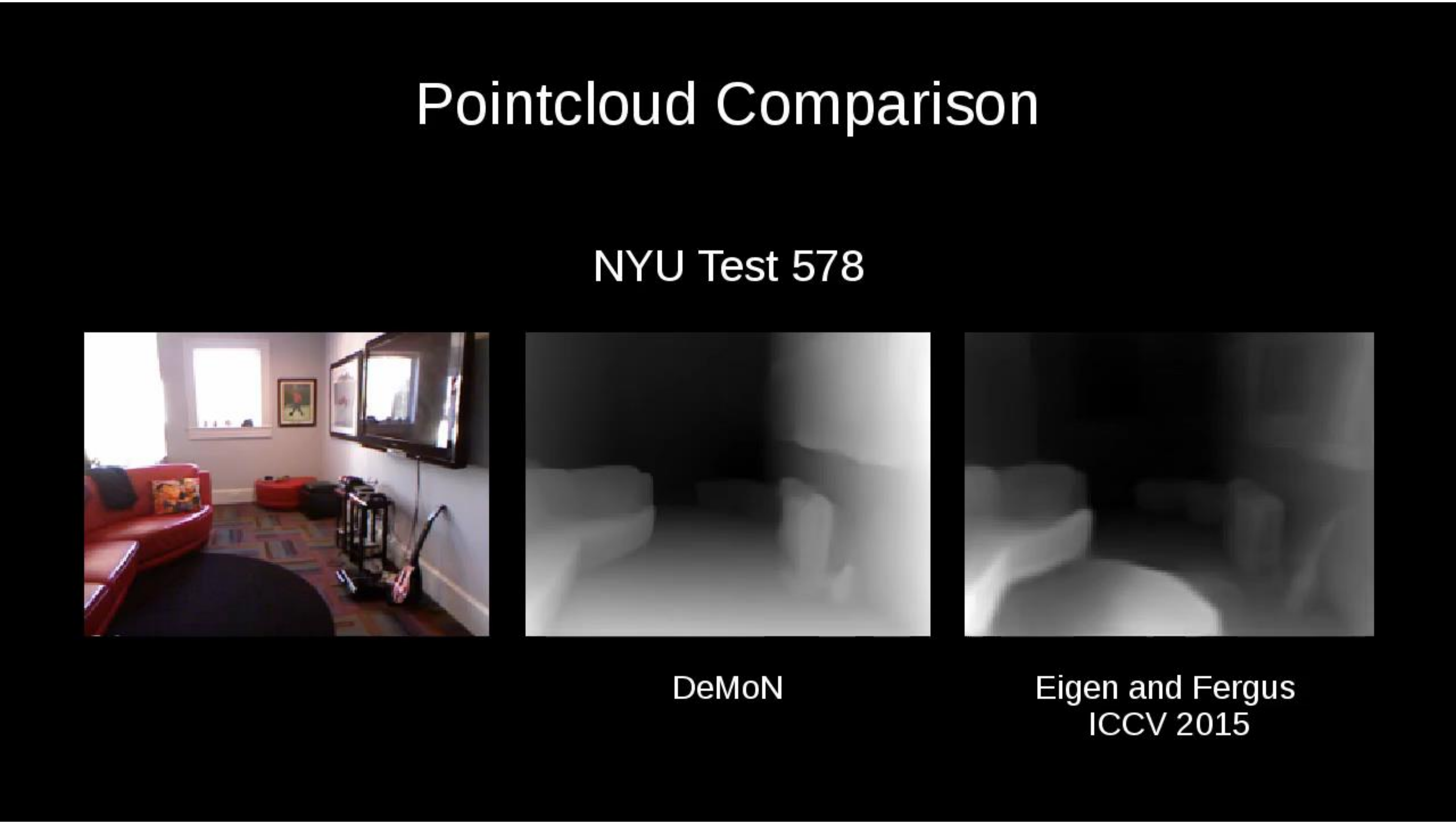
Sculpture



DeMoN

Eigen and Fergus
ICCV 2015

Two images generalize better than one image



Structure from motion at 7fps

Bike Ride

T | T+10 | predicted depth

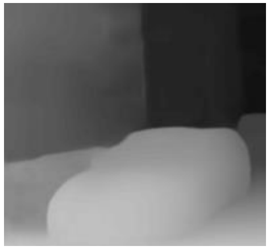
Estimated camera trajectory

Example from RGB-D SLAM dataset (Sturm et al.) Red: DeMoN. Black: Ground truth.

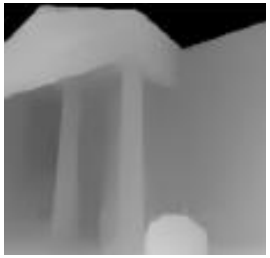
Deep learning for 3D Vision is promising

3D shape and texture from a single image

FlowNet: end-to-end optical flow



DispNet: end-to-end disparities



DeMoN: end-to-end structure from motion