COMPUTER VISION University of Freiburg

CVPR Lehrstuhl für Computer Vision & Pattern Recognition

DATASETS + DISPNET CODE

A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation

Nikolaus Mayer — mayern@cs.uni-freiburg.de
Eddy Ilg — ilg@cs.uni-freiburg.de
Philip Häusser — haeusser@cs.tum.edu
Philipp Fischer — fischer@cs.uni-freiburg.de
Daniel Cremers — cremers@tum.de
Alexey Dosovitskiy — dosovits@cs.uni-freiburg.de
Thomas Brox — brox@cs.uni-freiburg.de

PAPER + POSTER

CVPR2016

lmb.informatik.uni-freiburg.de
vision.in.tum.de

Datasets

Task and Hallmarks

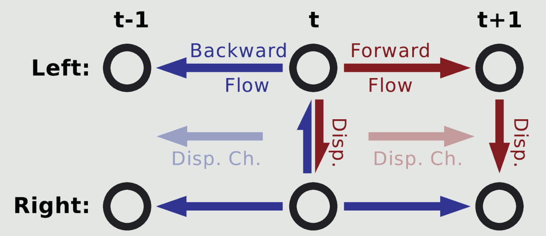
Create a dataset to train deep CNNs for scene flow

- Synthetic data creation → **accurate dense groundtruth**
- Procedurally generated scenes → **huge amounts of data**
- Many data modalities included → **enables a multitude of tasks**

Data Generation

What is Scene Flow?

Scene flow describes the **3D motion of scene points**, just like optical flow describes the 2D motion. Instead of metric values, we can use a pixel-space formulation to decouple the task from camera intrinsics.



The task thus splits into
- optical flow,
- disparities and
- disparity change.

The last component fills gaps left by occlusions.

Augmentation

- Due do epipolar constraints, disparity and scene flow cannot use all available augmentation types (less than optical flow)
- We need more source data to get the same amount of effective data

	Optical flow			Disparity			Scene flow		
Augmentation	all	time	stereo	all	time	stereo	all	time	stereo
Color changes	✓	✓	–	✓	–	✓	✓	✓	✓
Rotation	✓	✓	–	✗	–	✗	✗	✗	✗
Translation	✓	✓	–	(✓)	–	(✓)	(✓)	(✓)	(✓)
Crop / Zoom	✓	✓	–	(✓)	–	✗	(✓)	(✓)	✗
General warp	✓	✓	–	(✓)	–	✗	(✓)	(✓)	✗

- (✓) is technically OK, but does not conserve projective effects which might otherwise be exploited by a learning method
- – "not applicable", e.g. there is no time component in disparity

Available Data

Stereo RGB
- "Clean" and "Final" pass images
- Lossless PNG and efficient WebP

Optical flow
- Left and right view
- Forwards and backwards in time

Disparity
- Left and right view
- Full `float32` precision

Disparity change
- Left and right view
- Forwards and backwards in time
- No occluded regions

Segmentation
- Left and right view
- Object and Material level

Motion boundaries
- Left and right view
- Forwards and backwards in time

+ Full camera intrinsics / extrinsics

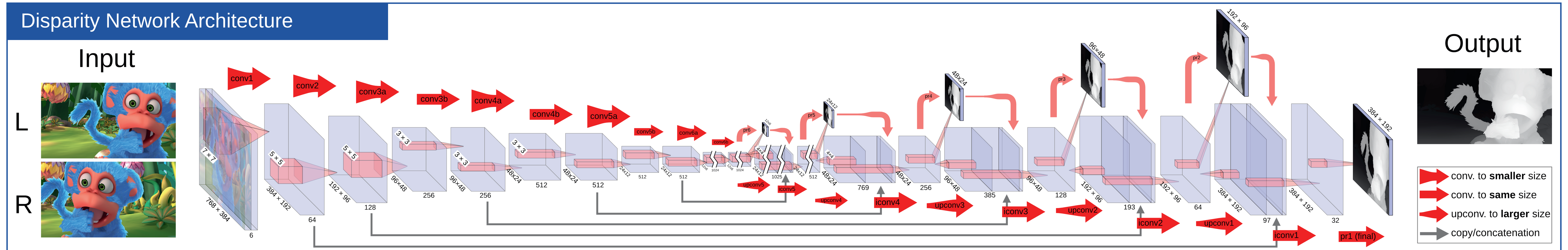
39049 frames

Train your networks for:

Optical flow
Disparity/depth from stereo
Scene flow
Depth from single image
Segmentation
Motion de-blurring
.
.
.

Get the Datasets:
http://lmb.informatik.uni-freiburg.de/resources/datasets/SceneFlowDatasets.en.html

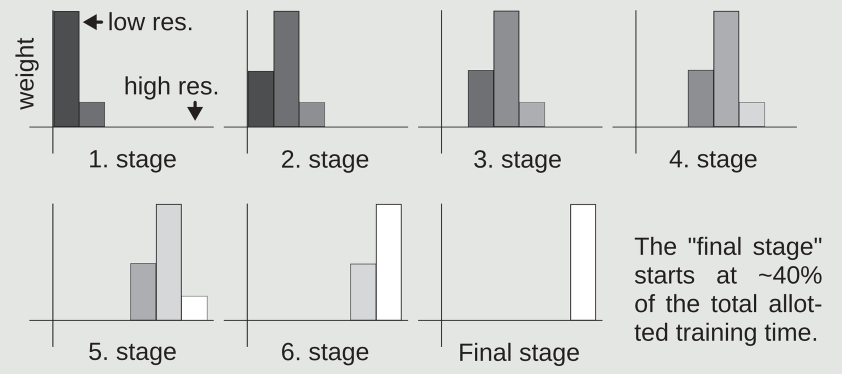
Disparity Network Architecture



Disparity Estimation

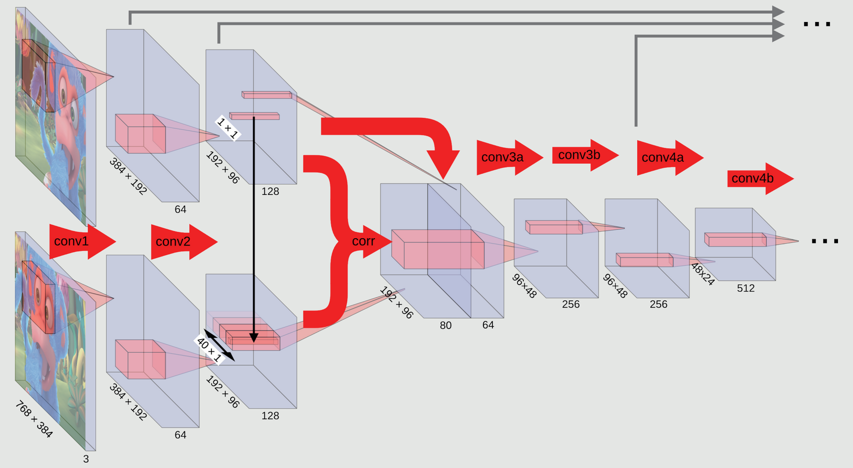
Training Schedules

- The network is very large and takes long to train
- In the beginning, we only train early **low-resolution losses**
- Later, we **enable higher resolutions** and **phase out low-res** losses
- The network can then **repurpose** the deeper layers when they are no longer constrained by directly attached losses



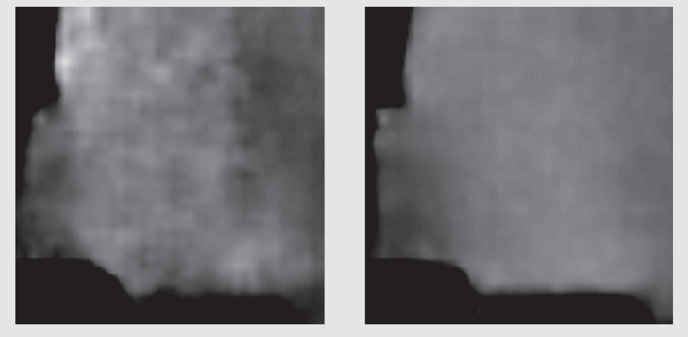
The "final stage" starts at ~40% of the total allotted training time.

1D-Correlation for Disparity



- Disparity is a **1D problem**, with known structure on rectified images
- Images are **processed separately** to extract features before joining
- Explicit **feature correlation on scanlines** enables efficient matching

Convolutions Between Upsamplings



Close-up of a predicted disparity map without (left) and with (right) convolutions between up-convolutions.

The additional steps in the synthesis/decoder result in much **smoother predictions**.

Scores

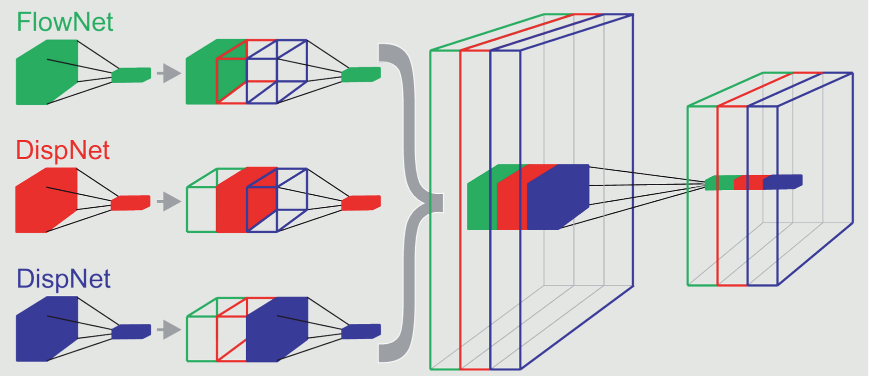
Method	KITTI 2012 train	KITTI 2012 test	KITTI 2015 train	KITTI 2015 test (D1)	Driving clean	FlyingThings3D clean test	Monkaa clean	Sintel clean train	Time
DispNet	2.38	—	2.19	—	15.62	2.02	5.99	5.38	0.06s
DispNetCorr1D	1.75	—	1.59	—	16.12	1.68	5.78	5.66	0.06s
DispNet-K	1.77	—	(0.77)	—	19.67	7.14	14.09	21.29	0.06s
DispNetCorr1D-K	1.48	1.0†	(0.68)	4.34%	20.40	7.46	14.93	21.88	0.06s
SGM	10.06	—	7.21	10.86%	40.19	8.70	20.16	19.62	1.1s
MC-CNN-fst	—	—	—	4.62%	19.58	4.09	6.71	11.94	0.8s
MC-CNN-acrt	—	0.9	—	3.89%	—	—	—	—	67s

14.5 fps on GTX Titan X

Scene Flow Estimation

Building the SceneFlowNet

Training from scratch is not feasible (9x the size of a FlowNet/DispNet!)
1. Interleaving 3 pretrained networks (1x FlowNet [2] and 2x DispNets)
2. Joint retraining on optical flow, 2x disparity, **and disparity change**



Benefits of Joint Retraining

Conjecture: A joint network for scene flow should be able to predict optical flow and disparity better than two separate specialist networks.

Rationale: The all-in-one network has access to more scene data (e.g. it could infer object structure from disparity and use it for optical flow).

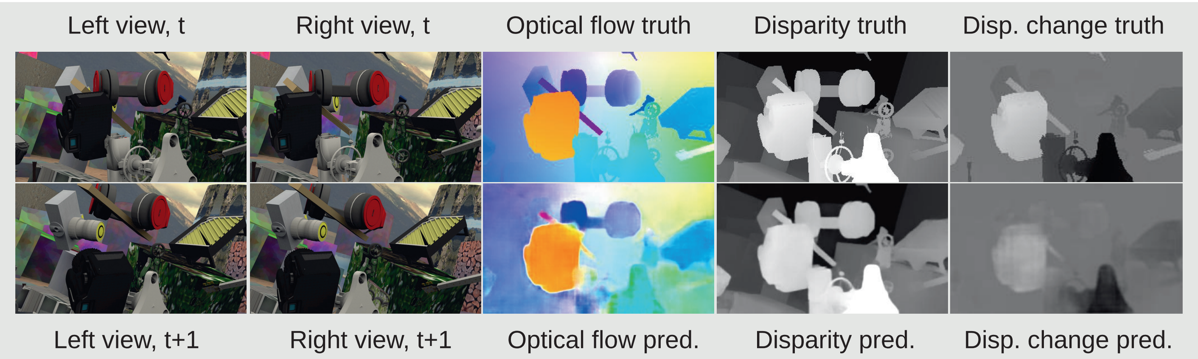
	Flow	Disparity	Disp. Ch
FlowNet	13.78		
DispNet		2.41	
FlowNet +500k	12.18		
DispNet +500k		2.37	
SceneFlowNet +500k	10.99	2.21	0.79

Our experiments show that **all tasks** are performed better after joint retraining, compared to individual extra training for the same time.

Results

Example scene flow prediction using a jointly retrained all-in-one network.

(Note that this network does not yet enforce any explicit consistencies between the scene flow components.)



DATASETS + DISPNET CODE

PAPER + POSTER

Acknowledgements

Partially funded by the ERC Starting Grant VideoLearn, the ERC Consolidator Grant 3D Reloaded, and by the DFG Grants BR 3815/7-1 and CR 250/13-1.

References

[1] M. Savva, A. X. Chang, and P. Hanrahan. Semantically-Enriched 3D Models for Common-sense Knowledge. CVPR 2015 Workshop on Functionality, Physics, Intentionality and Causality, 2015.

[2] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. van der Smagt, D. Cremers, and T. Brox. FlowNet: Learning optical flow with convolutional networks. In ICCV, 2015.

[3] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In ECCV, Part IV, LNCS 7577, pages 611–625, Oct. 2012.

[4] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

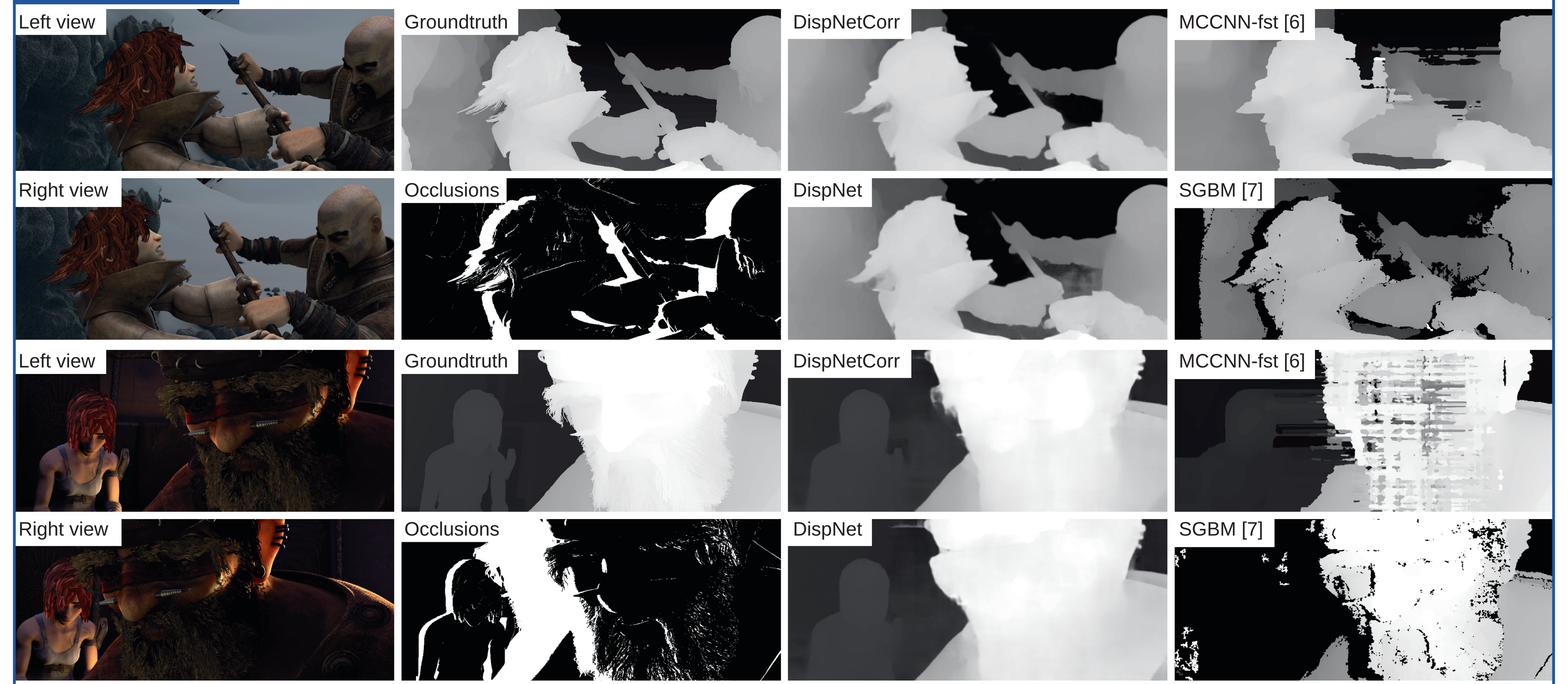
[5] https://store.blender.org/product/monkaa-open-movie

[6] J. Zbontar and Y. LeCun. Stereo matching by training a convolutional neural network to compare image patches. arXiv preprint arXiv:1510.05970, 2015.

[7] http://docs.opencv.org/2.4/modules/calib3d/doc/camera_calibration_and_3d_reconstruction.html

Disparity Results

Sintel [3]



KITTI 2015 [4]

Our Datasets