Self-Supervised Monocular Depth Estimation

Oisin Mac Aodha
University of Edinburgh

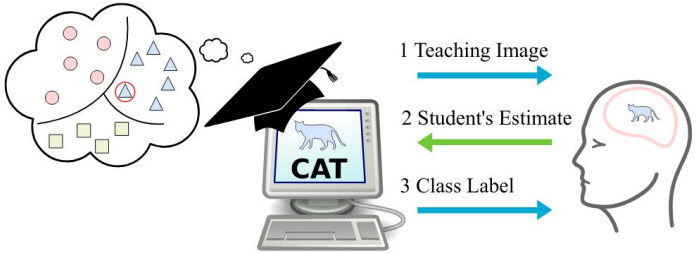
www.oisin.info
@oisinmacaodha

Background


Undergrad
PhD
Postdoc
Canada
Hudson Bay
Labrador Sea
United States
Mexico
Gulf of Mexico
Cuba
Puerto Rico
Guatemala
Caribbean Sea
Nicaragua
Venezuela
North Atlantic Ocean
Norway
Denmark
United Kingdom
Ireland
Poland
Germany
Austria
France
Italy
Spain
Portugal
Tunisia
Mediterranean Sea
Morocco
Algeria
Libya
Western Sahara
Mauritania
Mali
Niger
Chad
Burkina Faso
Guinea
Nigeria

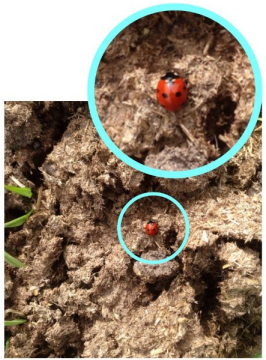
Research Interests

Machine Teaching



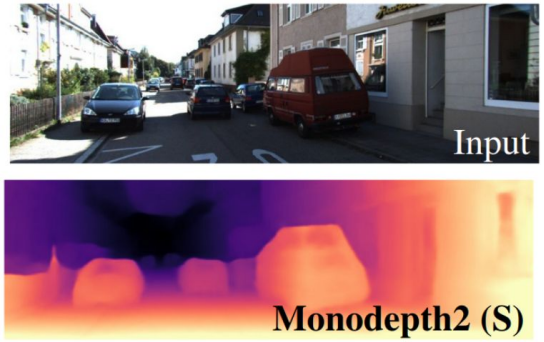
Fine-Grained Classification

Two-spotted ladybug
Adalia bipunctata



Seven-spotted ladybug
Coccinella septempunctata

Self-Supervised Learning



Joint Work With

Clément Godard
Skydio

Michael Firman
Niantic

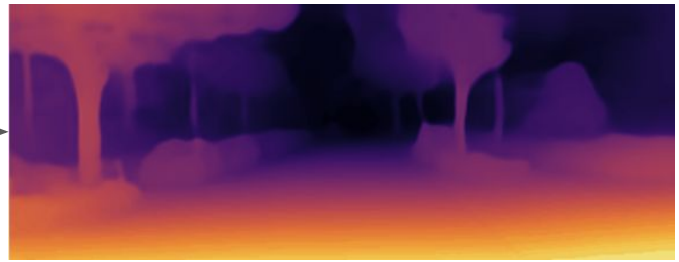
Gabriel J. Brostow
UCL & Niantic

Can we train a deep network to predict **depth** from a **single image**?

Deep Net

Input Image

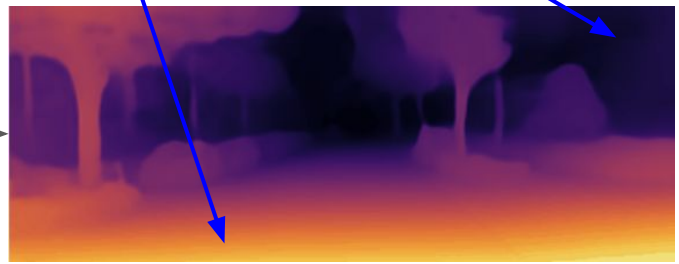
Output Predictions



Input Image

Output Predictions

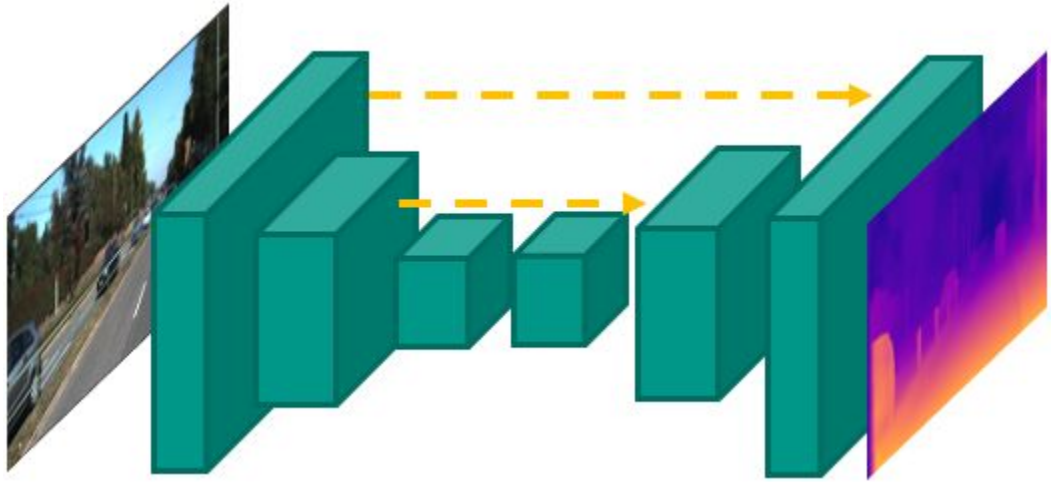
Deep Net



Input Image

Output Predictions

Depth Network



Input Image

Output Depth

Why Depth - Photo Editing

Karsch et al. Rendering synthetic objects into legacy photographs, SIGGRAPH Asia 2011

Why Depth - Augmented Reality

Without depth estimation

Slide Credit: Niantic

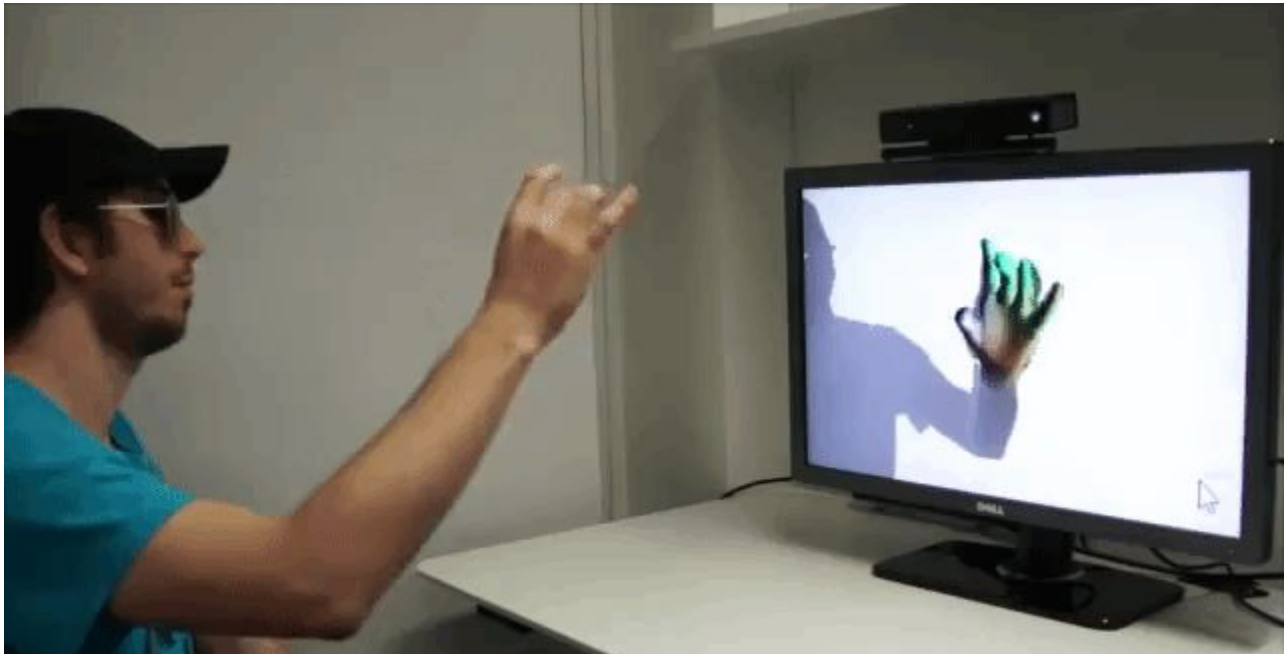
Why Depth - Augmented Reality

Without depth estimation

With depth estimation

Slide Credit: Niantic

Why Depth - Perception

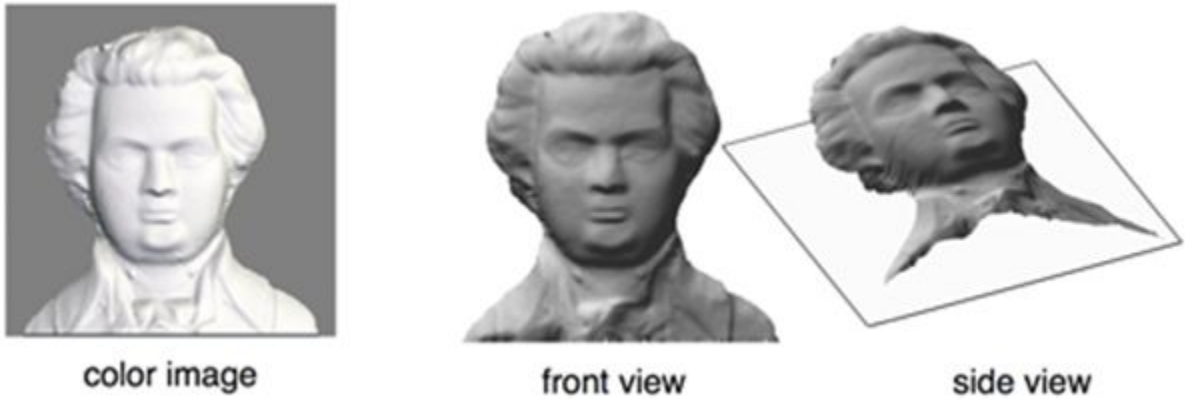


Sharp et al. Accurate, Robust, and Flexible Real-time Hand Tracking, CHI 2015

Related Work - Structure from Motion

http://vision.princeton.edu/courses/SFMedu/

Related Work - Shape from Shading

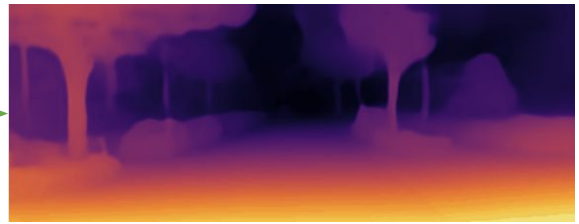


e.g Johnson and Adelson CVPR 2011

Related Work - Learning Approaches

Input frame

Estimated depth

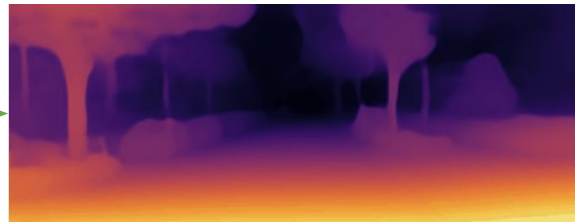


Eigen et al. NeurIPS 2014, Eigen et al. ICCV 2015, Laina et al. 3DV 2016, ...

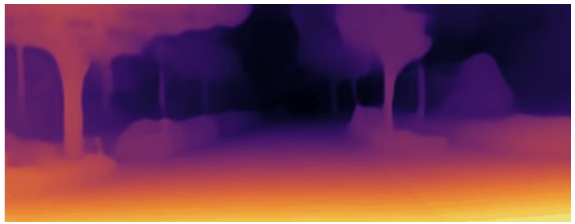
Related Work - Learning Approaches

Input frame

Estimated depth



Supervised GT



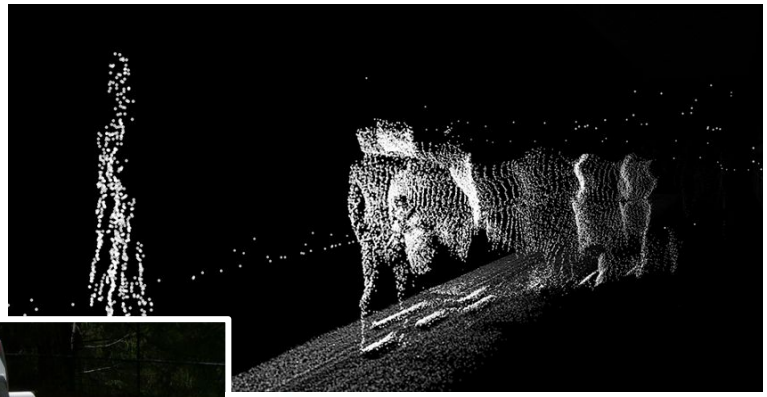
$$L_s = \sum ||f(I^l) - d||^2$$

Eigen et al. NeurIPS 2014, Eigen et al. ICCV 2015, Laina et al. 3DV 2016, ...

Sources of depth supervision

✓ Best results on standard datasets

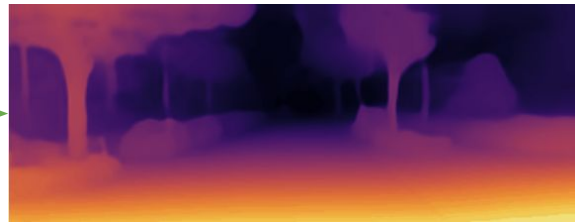
❌ Interference from sun
❌ Sparse points
❌ No shiny or moving surfaces
❌ Heavy and expensive devices
❌ No large online repositories



Related Work - Learning Approaches

Input frame

Estimated depth



Supervised GT

$$L_s = \sum ||f(I^l) - d||^2$$

Eigen et al. NeurIPS 2014, Eigen et al. ICCV 2015, Laina et al. 3DV 2016, ...

How can we train **without** ground truth **depth supervision**?

Stereo data

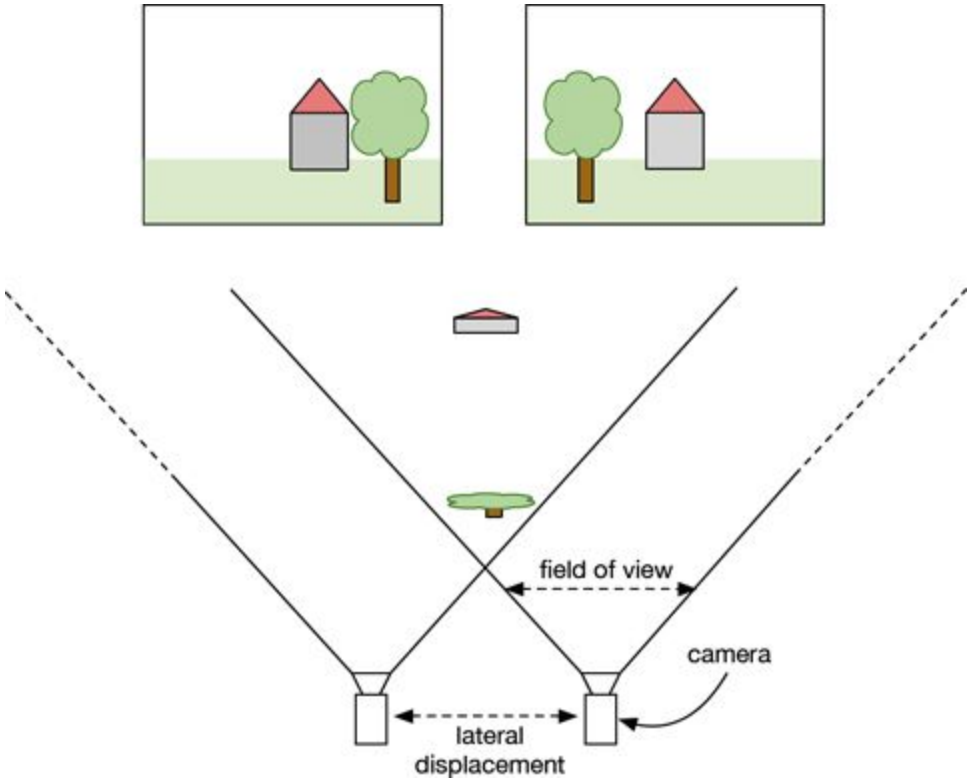


Image Credit https://mint.fh-hagenberg.at/?p=1149

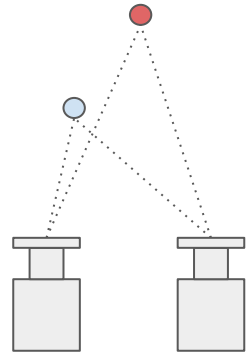
Learning depth from stereo

Input frame L

Source frame R

Learning depth from stereo

Input frame L



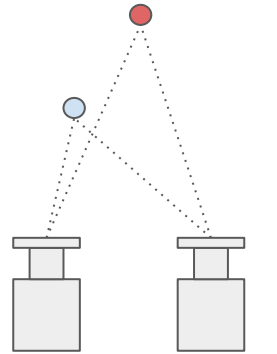
Source frame R

Learning depth from stereo

Input frame L

Source frame R

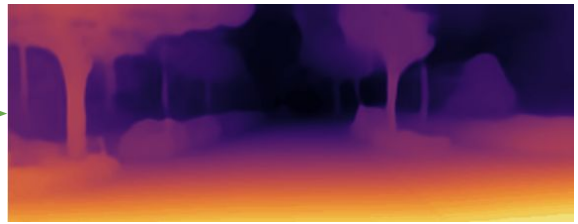
Key Observation
If we know the corresponding location for each pixel in the other view (and the distance between the cameras) we can estimate depth.



Learning depth from stereo

Input frame L

Estimated depth

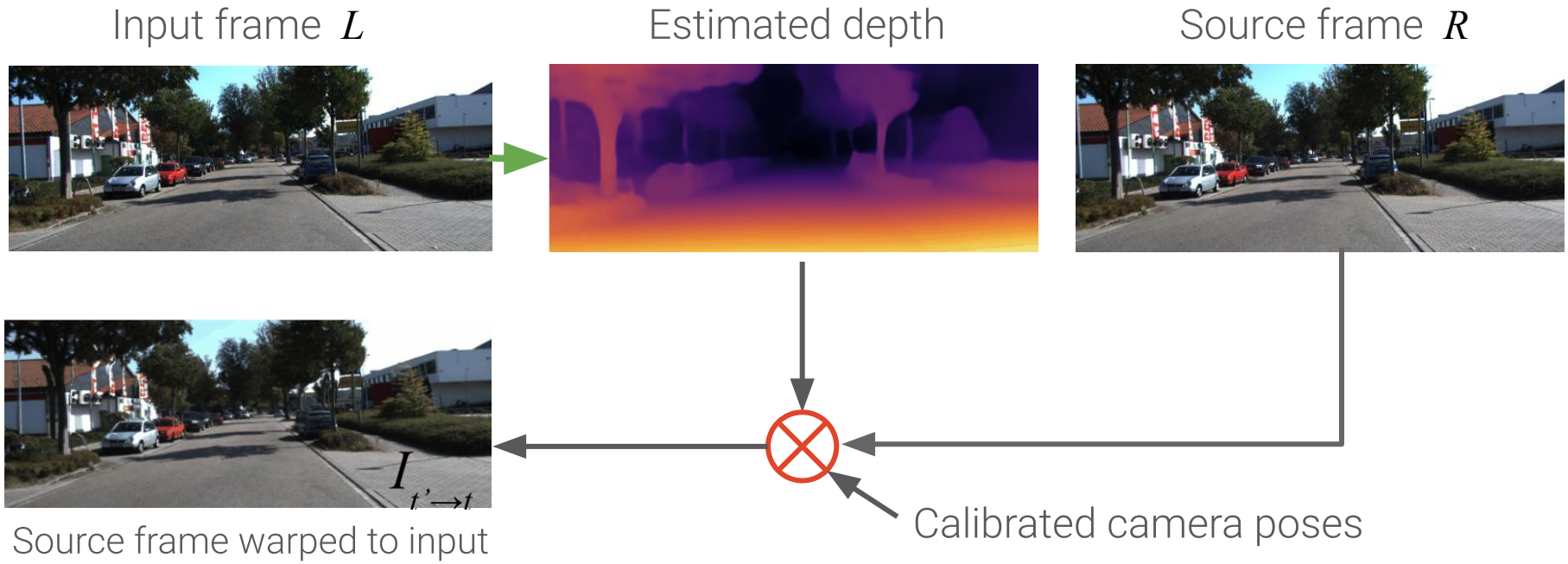


Source frame R

E.g. Garg et al., Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue. ECCV 2016
Godard et al., Unsupervised Monocular Depth Estimation with Left-Right Consistency CVPR 2017

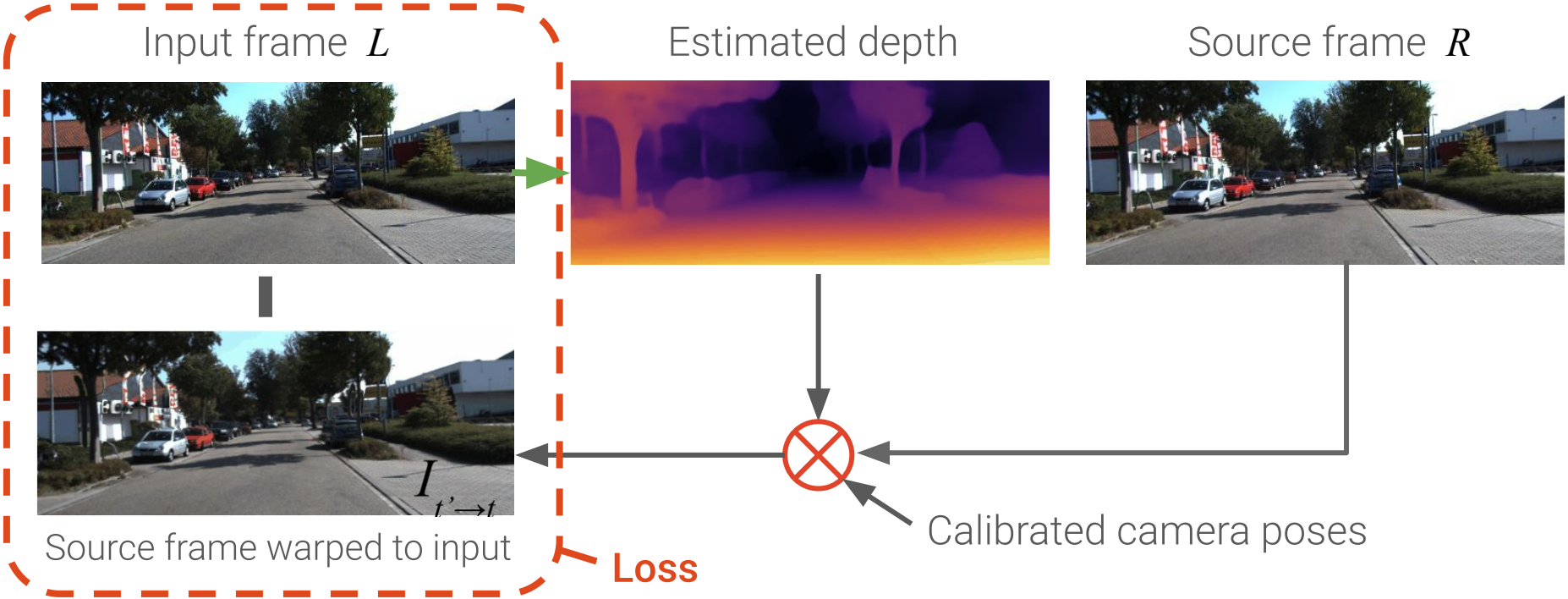
← Monodepth

Learning depth from stereo



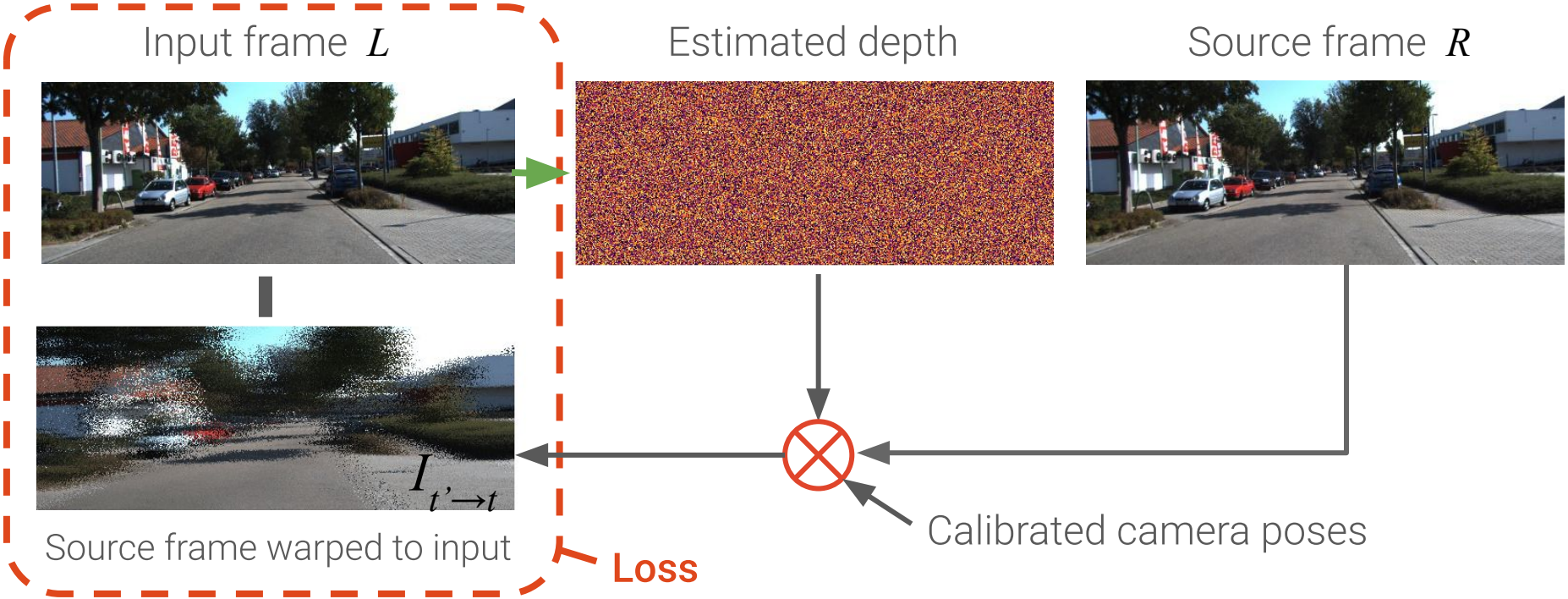
E.g. Garg et al., Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue. ECCV 2016
Godard et al., Unsupervised Monocular Depth Estimation with Left-Right Consistency CVPR 2017 ← Monodepth

Learning depth from stereo



E.g. Garg et al., Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue. ECCV 2016
Godard et al., Unsupervised Monocular Depth Estimation with Left-Right Consistency CVPR 2017 ← Monodepth

Learning depth from stereo



E.g. Garg et al., Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue. ECCV 2016
Godard et al., Unsupervised Monocular Depth Estimation with Left-Right Consistency CVPR 2017 ← Monodepth

Supervised

$$L_s = \sum ||f(I^l) - d^l||$$

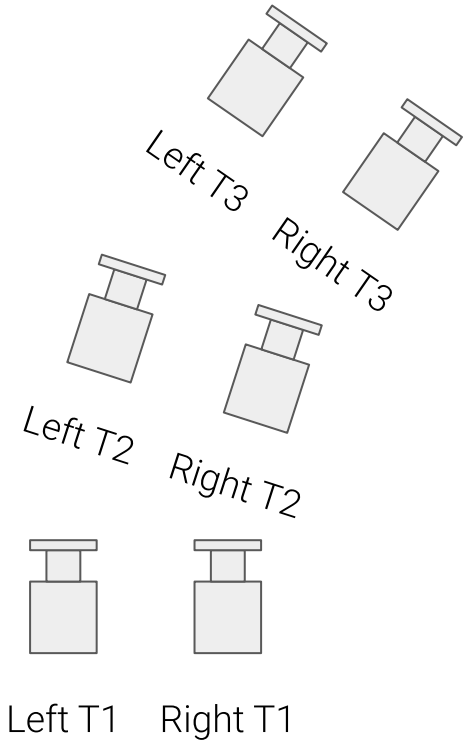
~~Supervised~~ $$L_s = \sum ||f(I^l) - d^l||$$

Self-Supervised

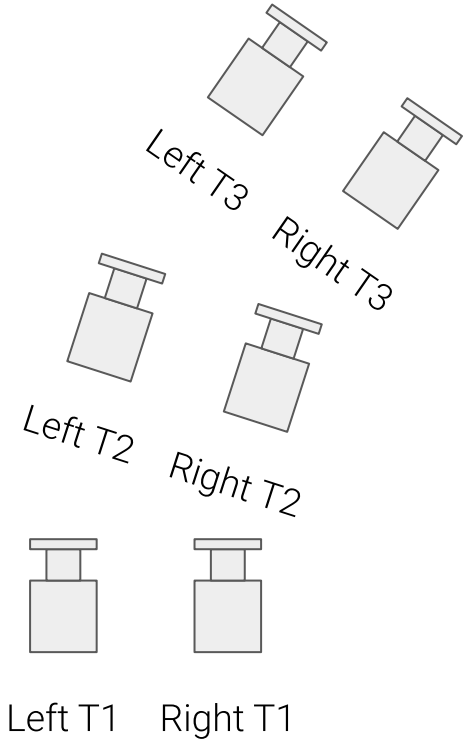
$$\tilde{d}^l = f(I^l)$$
$$\tilde{I}^l = I^r(\tilde{d}^l)$$
$$L_{ss} = \sum ||I^l - \tilde{I}^l||$$

How can we train **without stereo supervision**?

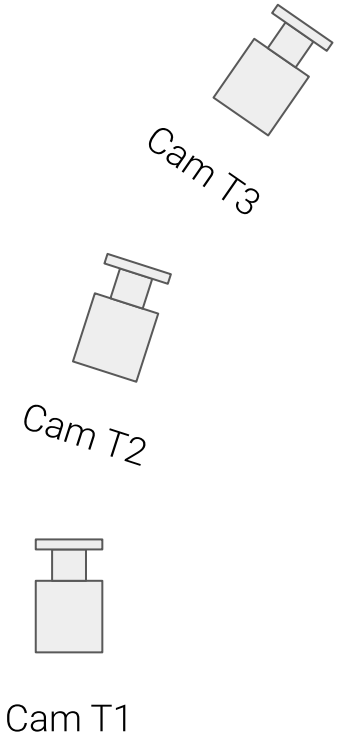
Stereo Data



Stereo Data



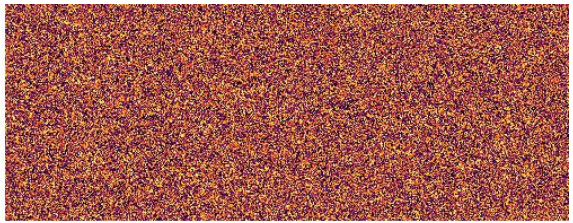
Monocular Video



Learning depth from ~~stereo~~ monocular video

Input frame I_t

Depth estimation



Source frames $I_{t'}$

Source frames warped to input $I_{t'\rightarrow t}$

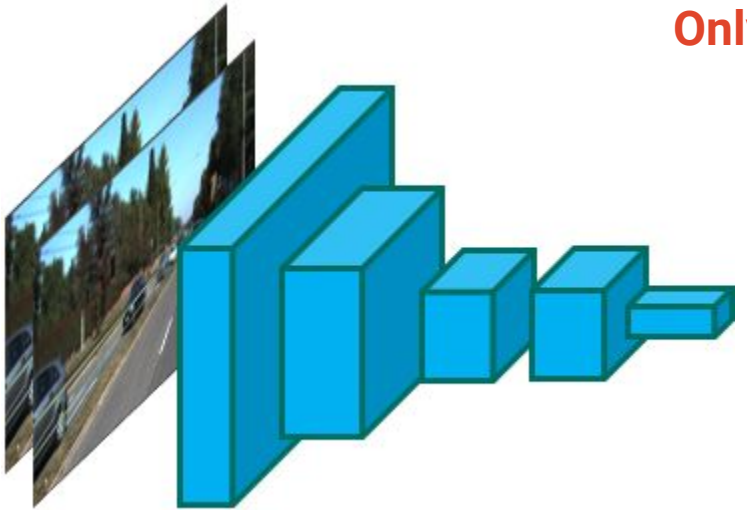
~~Calibrated camera poses~~

Pose prediction network

E.g. Zhou et al., Unsupervised Learning of Depth and Ego-Motion from Video, CVPR 2017

Godard et al., Digging Into Self-Supervised Monocular Depth Estimation ICCV 2019 ← Monodepth2

Pose network



Only needed during training!

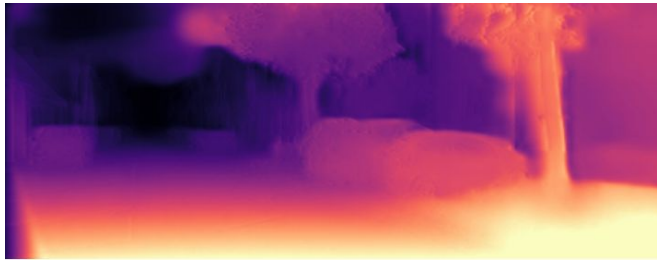
[T, R]

Input Image

Output Translation
and Rotation

Monocular vs. Stereo training

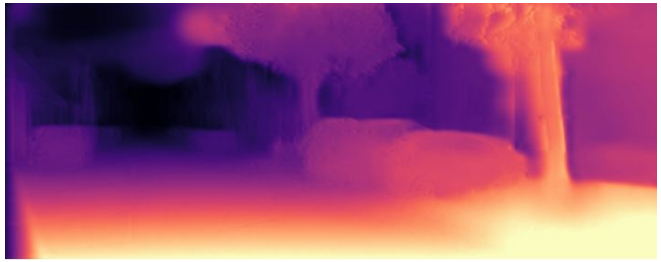
Input



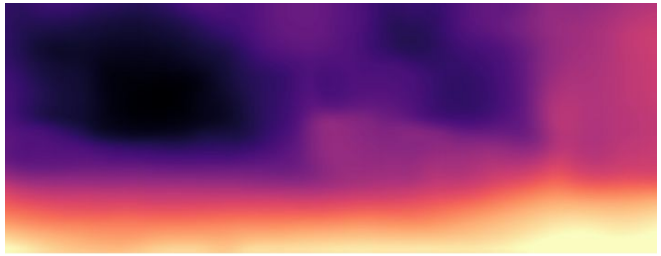
Monodepth
(Ours CVPR 2017)
Stereo training

Monocular vs. Stereo training

Input



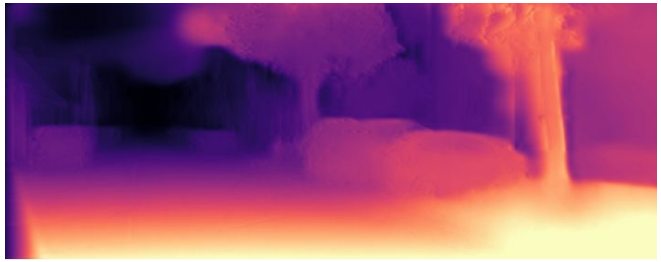
Monodepth
(Ours CVPR 2017)
Stereo training



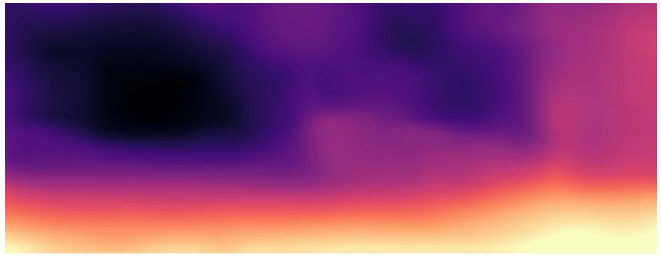
Zhou et al.
(CVPR 2017)
Mono training

Monocular vs. Stereo training

Input



Monodepth
(Ours CVPR 2017)
Stereo training



Zhou et al.
(CVPR 2017)
Mono training

Monodepth2
(Ours ICCV 2019)
Mono training

Monodepth2

1) Occlusion

2) (Some) moving objects

3) Multi-scale depth

Digging Into Self-Supervised Monocular Depth Estimation

Clément Godard[1] Oisin Mac Aodha[2] Michael Firman[3] Gabriel Brostow[3,1]
[1]UCL [2]Caltech [3]Niantic

www.github.com/nianticlabs/monodepth2

Abstract

Per-pixel ground-truth depth data is challenging to acquire at scale. To overcome this limitation, self-supervised learning has emerged as a promising alternative for training models to perform monocular depth estimation. In this paper, we propose a set of improvements, which together result in both quantitatively and qualitatively improved depth maps compared to competing self-supervised methods.

Research on self-supervised monocular training usually explores increasingly complex architectures, loss functions, and image formation models, all of which have recently helped to close the gap with fully-supervised methods. We show that a surprisingly simple model, and associated design choices, lead to superior predictions. In particular, we propose (i) a minimum reprojection loss, designed to robustly handle occlusions, (ii) a full-resolution multi-scale sampling method that reduces visual artifacts, and (iii) an auto-masking loss to ignore training pixels that violate camera motion assumptions. We demonstrate the effectiveness of each component in isolation, and show high quality, state-of-the-art results on the KITTI benchmark.

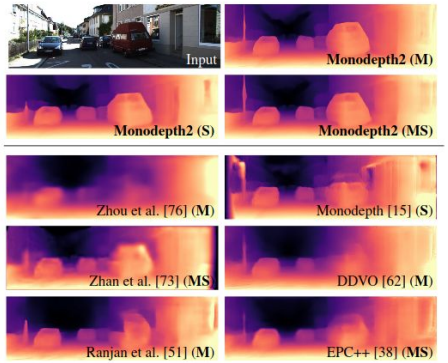


Figure 1. **Depth from a single image.** Our self-supervised model, **Monodepth2**, produces sharp, high quality depth maps, whether trained with monocular (M), stereo (S), or joint (MS) supervision.

approaches have shown that it is instead possible to train monocular depth estimation models using only synchronized *stereo pairs* [12, 15] or *monocular video* [76].

Godard et al. ICCV 2019

Problem: **Occlusions between frames**

Problem: **Occlusions between frames**

Source frame I_{t-1} Target frame I_t Source frame I_{t+1}

Traditional approach: Match every pixel to <u>all</u> source images

Solution: **Minimum reprojection loss**

Source frame I_{t-1}

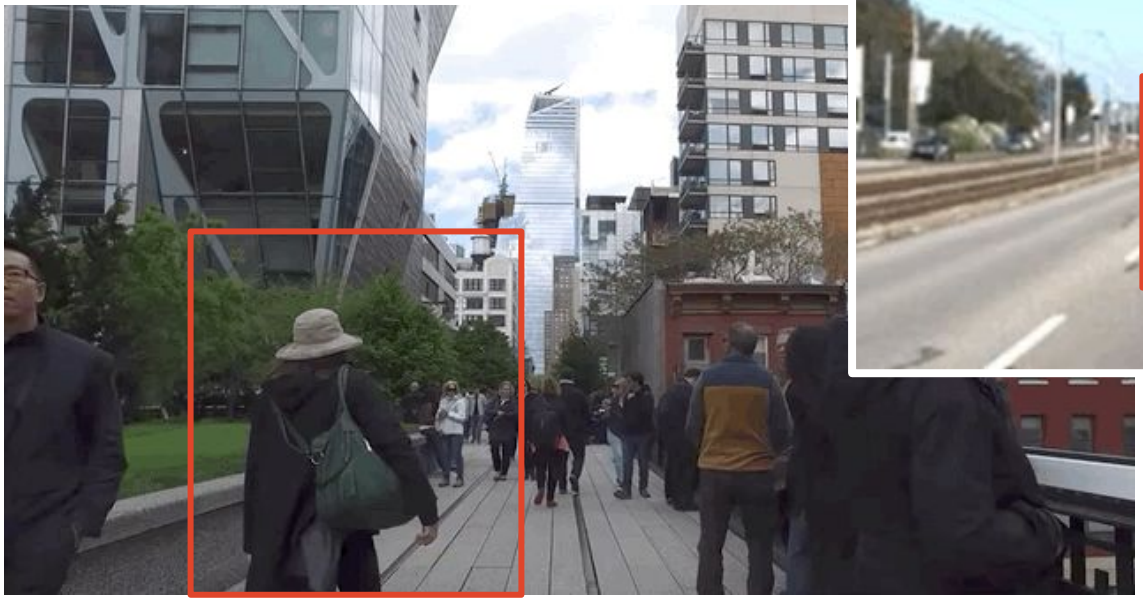
Target frame I_t

Source frame I_{t+1}

Traditional approach: Match every pixel to all source images

Monodepth2: Match every pixel to the best source image

Problem: **Objects moving at ego-velocity**

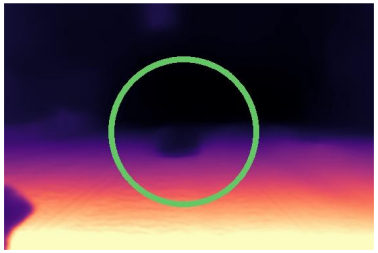


KITTI dataset

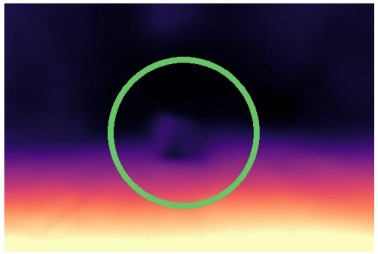
Wind Walk Travel Videos: https://www.youtube.com/watch?v=jv4m41pbOJE

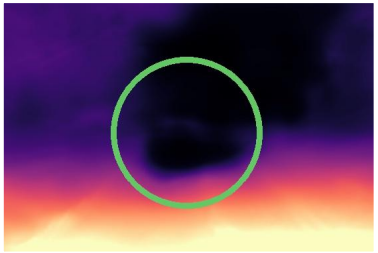
Objects moving at ego-velocity

Ranjan et al.

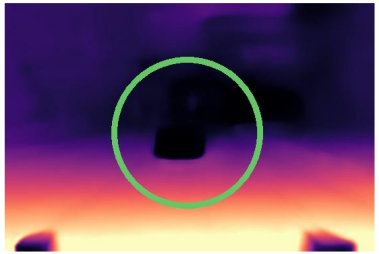


Every Pixel Counts++





Geonet



Our baseline

Ranjan et al., Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. CVPR 2019.
Z. Yin and J. Shi. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. CVPR 2018.
Luo et al. Every pixel counts++: Joint learning of geometry and motion with 3D holistic understanding. arXiv:1810.06125, 2018

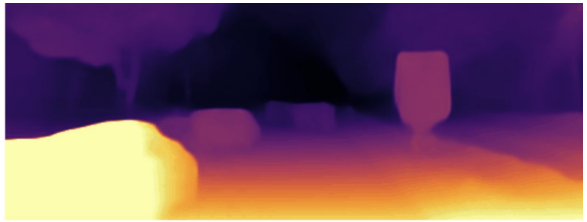
Solution: **Auto-masking**

Aim: Remove training pixels which remain static in image space

Solution: **Auto-masking**

Target frame I_t

Depth estimation



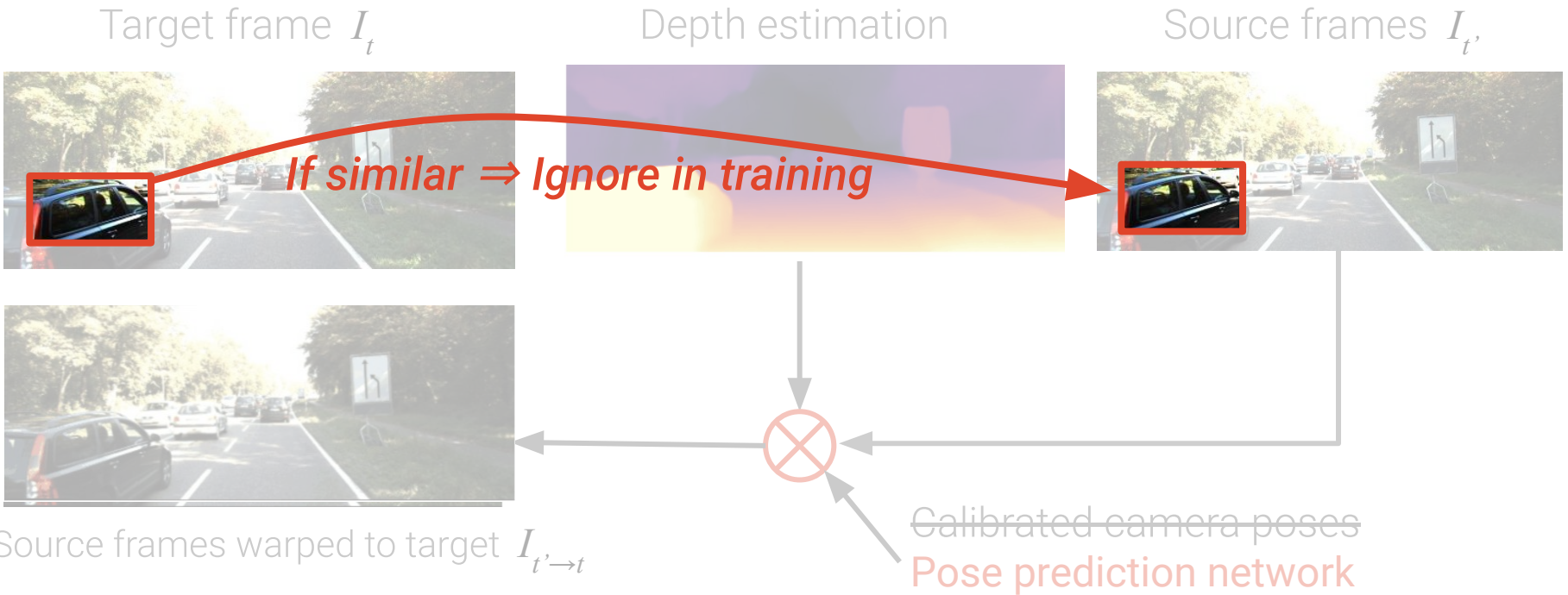
Source frames $I_{t'}$

Source frames warped to target $I_{t' \to t}$

~~Calibrated camera poses~~

Pose prediction network

Solution: **Auto-masking**



Solution: **Auto-masking**

Target frame I_t

Depth estimation

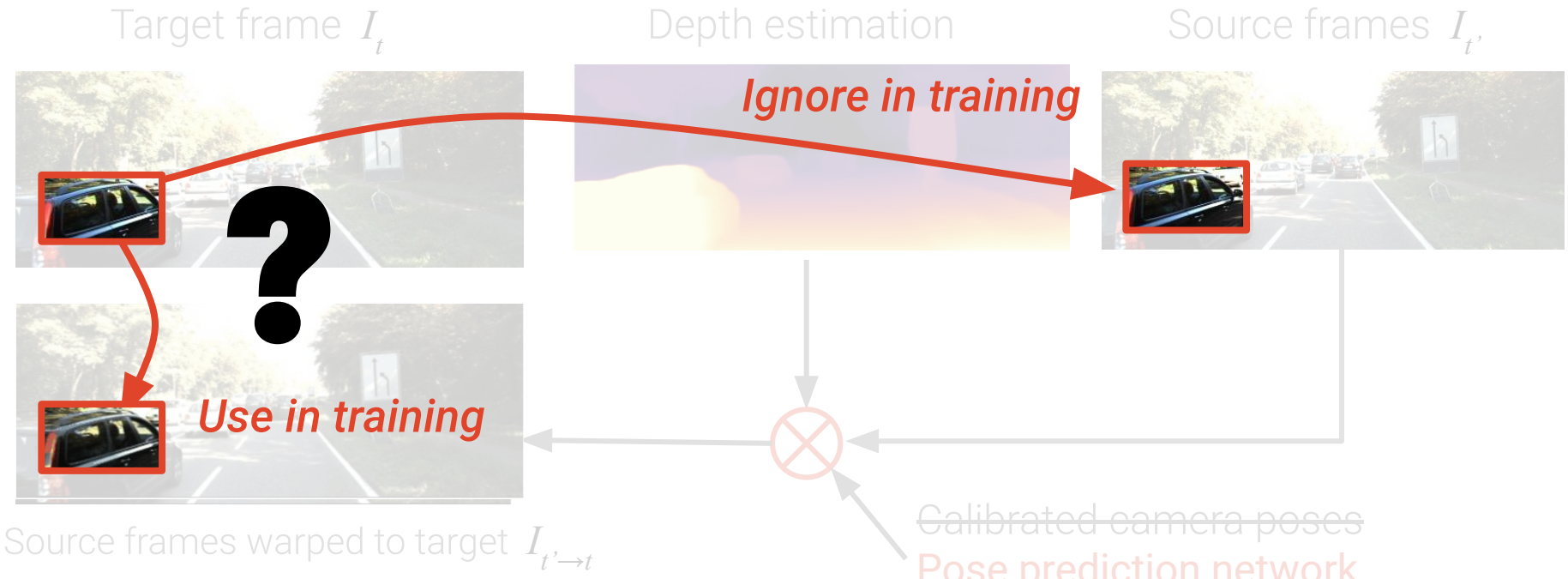
Source frames $I_{t'}$

Source frames warped to target $I_{t' \to t}$

~~Calibrated camera poses~~

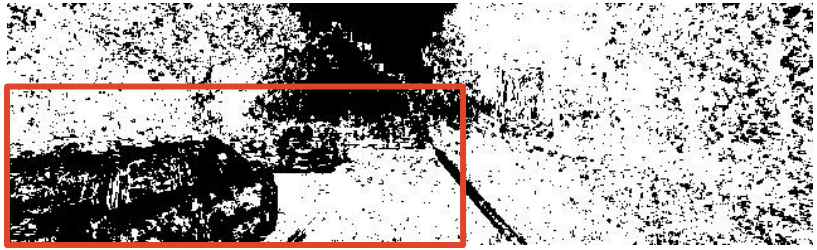
Pose prediction network

Solution: **Auto-masking**



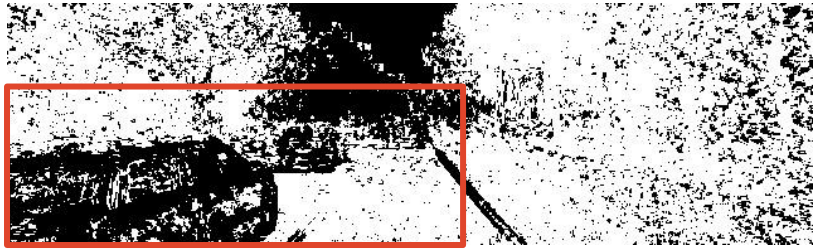
Auto-masking in practice

Ego-velocity objects → Objects removed

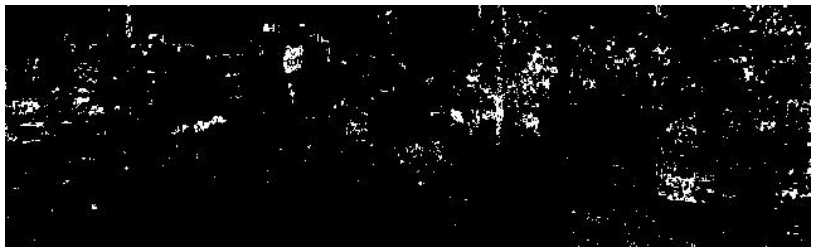


Auto-masking in practice

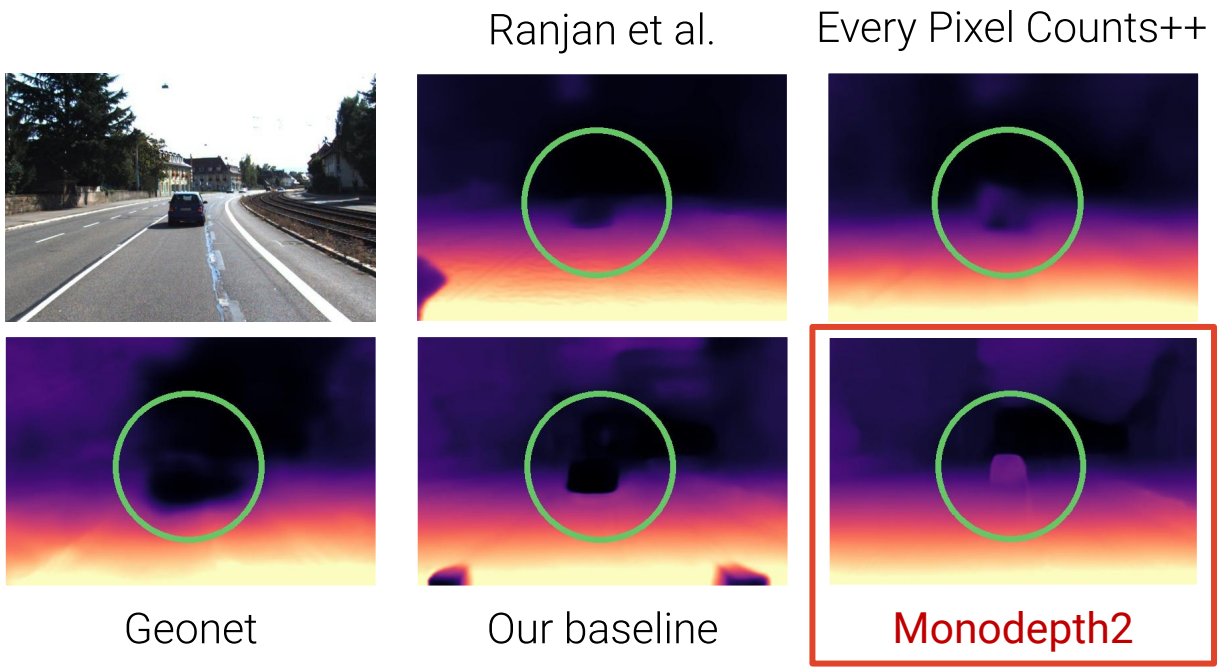
Ego-velocity objects → Objects removed



Static camera → Full frame ignored

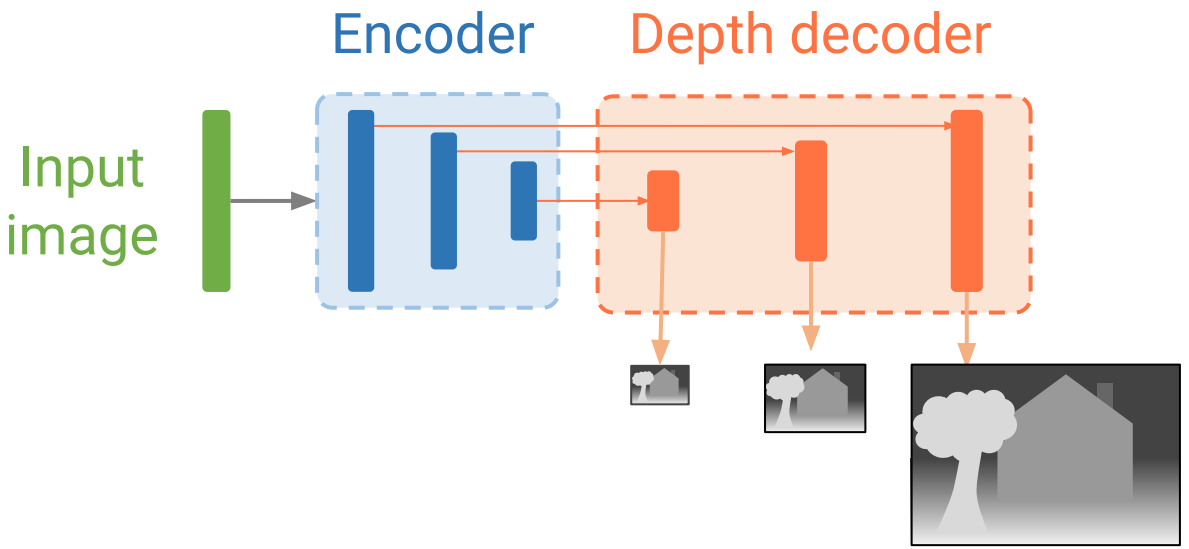


Objects moving at ego-velocity



Ranjan et al., Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. CVPR 2019.
Z. Yin and J. Shi. GeoNet: Unsupervised learning of dense depth, optical flow and camera pose. CVPR 2018.
Luo et al. Every pixel counts++: Joint learning of geometry and motion with 3D holistic understanding. arXiv:1810.06125, 2018

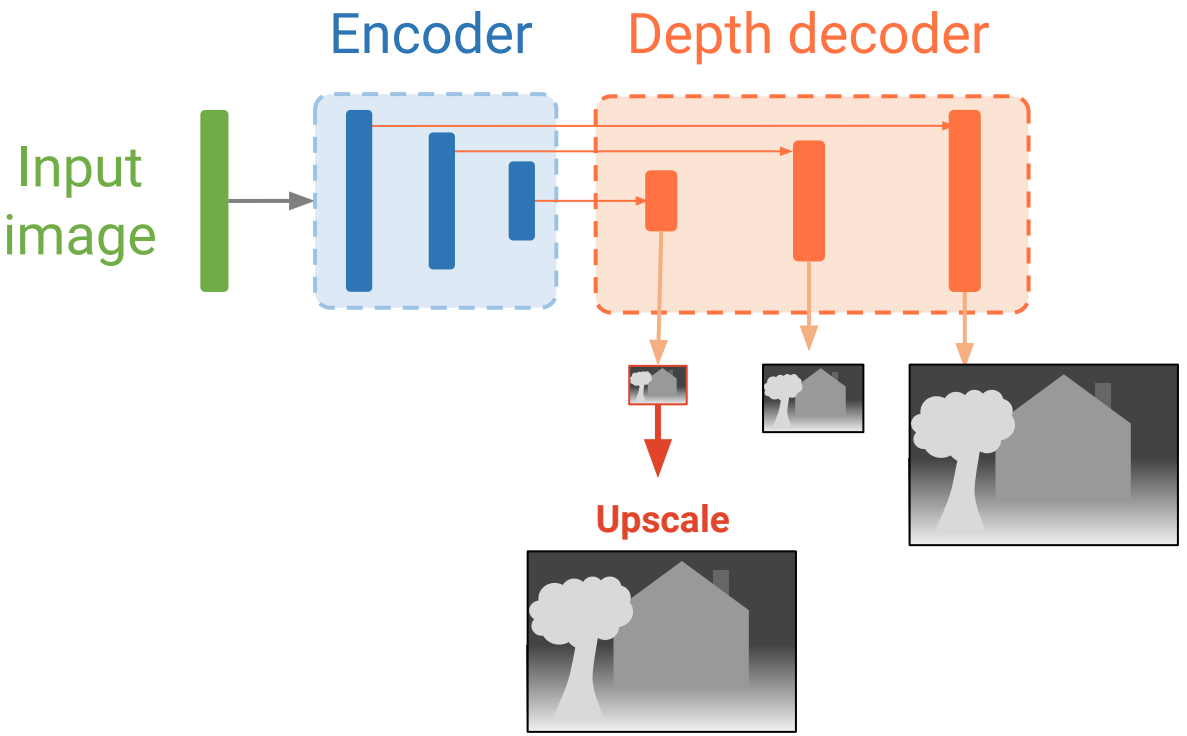
Problem: Multi-scale depth



Problem: Multi-scale depth

Baseline

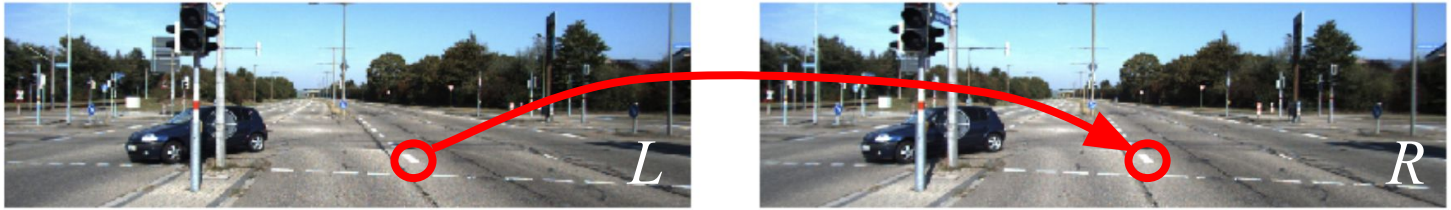
Solution: Full-resolution multi-scale



Solution: Full-resolution multi-scale

Baseline

Ours



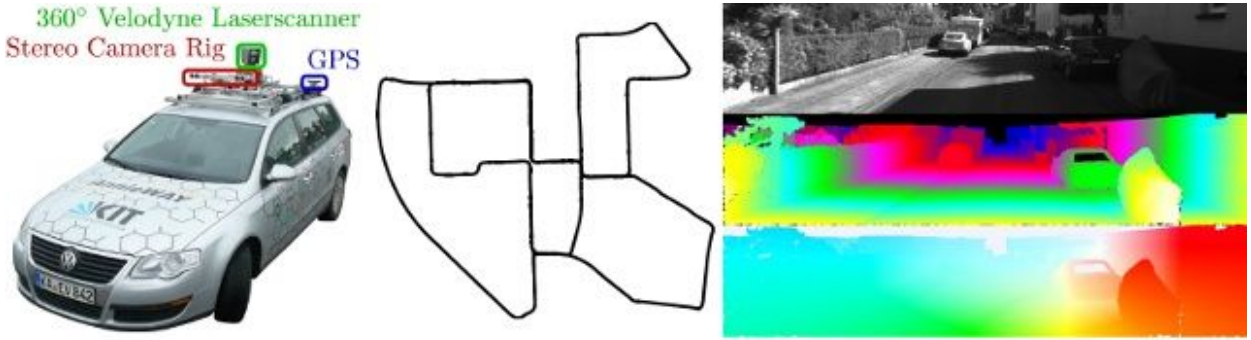
Experiments

KITTI

Driving dataset

Different locations

Stereo and video

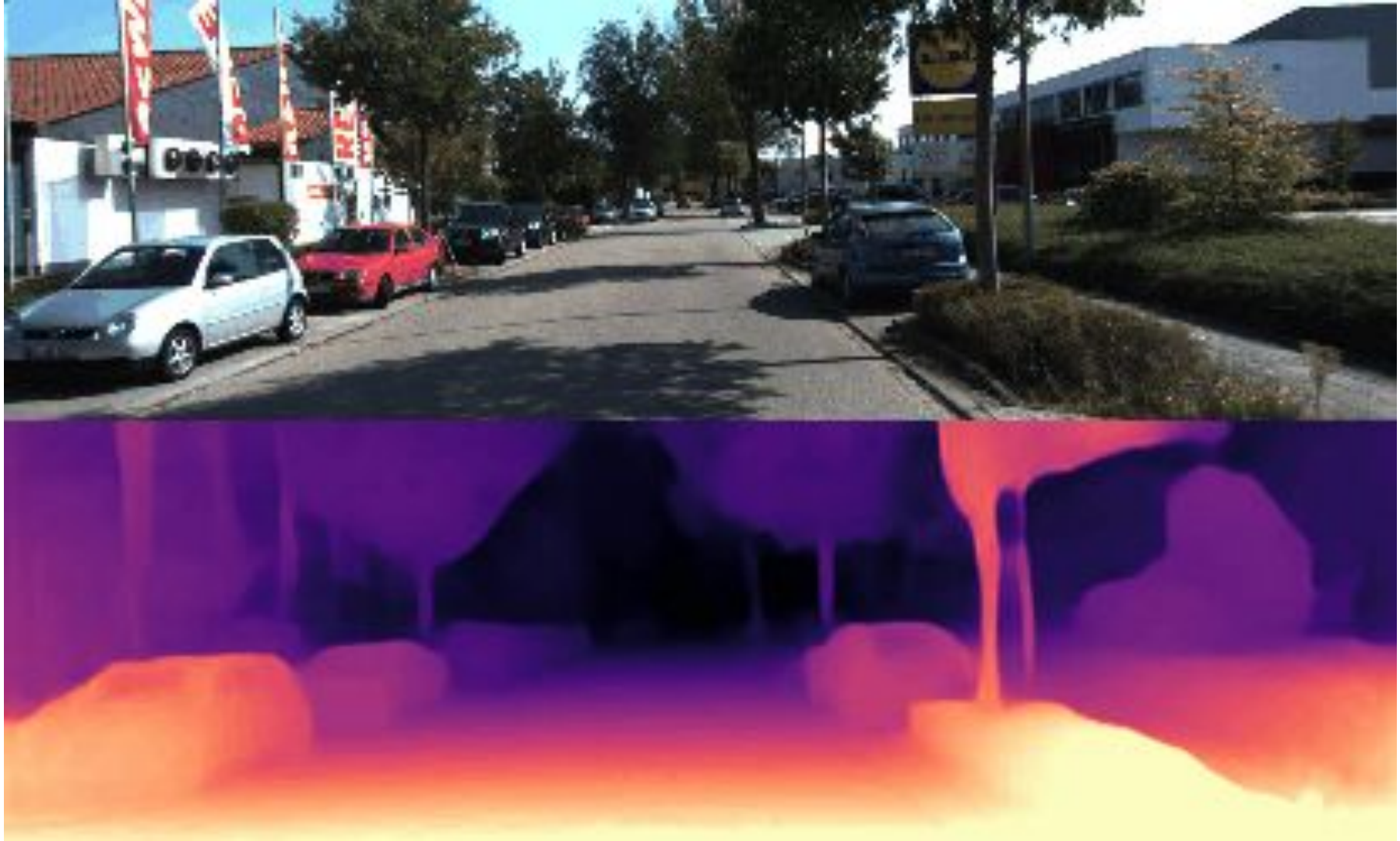


ResNet18 encoder

Train ~40K frames

Test ~700 frames with **lidar ground truth**

Geiger et al. Vision meets robotics: The KITTI dataset, 2013



Baseline comparison

	Absolute relative
Zhou et al.	0.183
Our baseline	0.140

$$\sum |\tilde{d} - d| / d$$

Baseline comparison

	Absolute relative
Zhou et al.	0.183
Our baseline	0.140
Baseline + min. reprojection	0.122
Baseline + auto-masking	0.124
Baseline + full res.	0.124

Baseline comparison

	Absolute relative
Zhou et al.	0.183
Our baseline	0.140
Baseline + min. reprojection	0.122
Baseline + auto-masking	0.124
Baseline + full res.	0.124
Monodepth 2	**0.115**

Baseline comparison

Lower better *Higher better*

	Absolute relative	Squared relative	RMSE	RMSE log	□ < 1.25
Zhou et al.	0.183	1.595	6.709	0.270	0.734
Our baseline	0.140	1.610	5.512	0.223	0.852
Baseline + min. reprojection	0.122	1.081	5.116	0.199	0.866
Baseline + auto-masking	0.124	0.936	5.010	0.206	0.858
Baseline + full res.	0.124	1.170	5.249	0.203	0.865
Monodepth 2	**0.115**	**0.903**	**4.863**	**0.193**	**0.877**

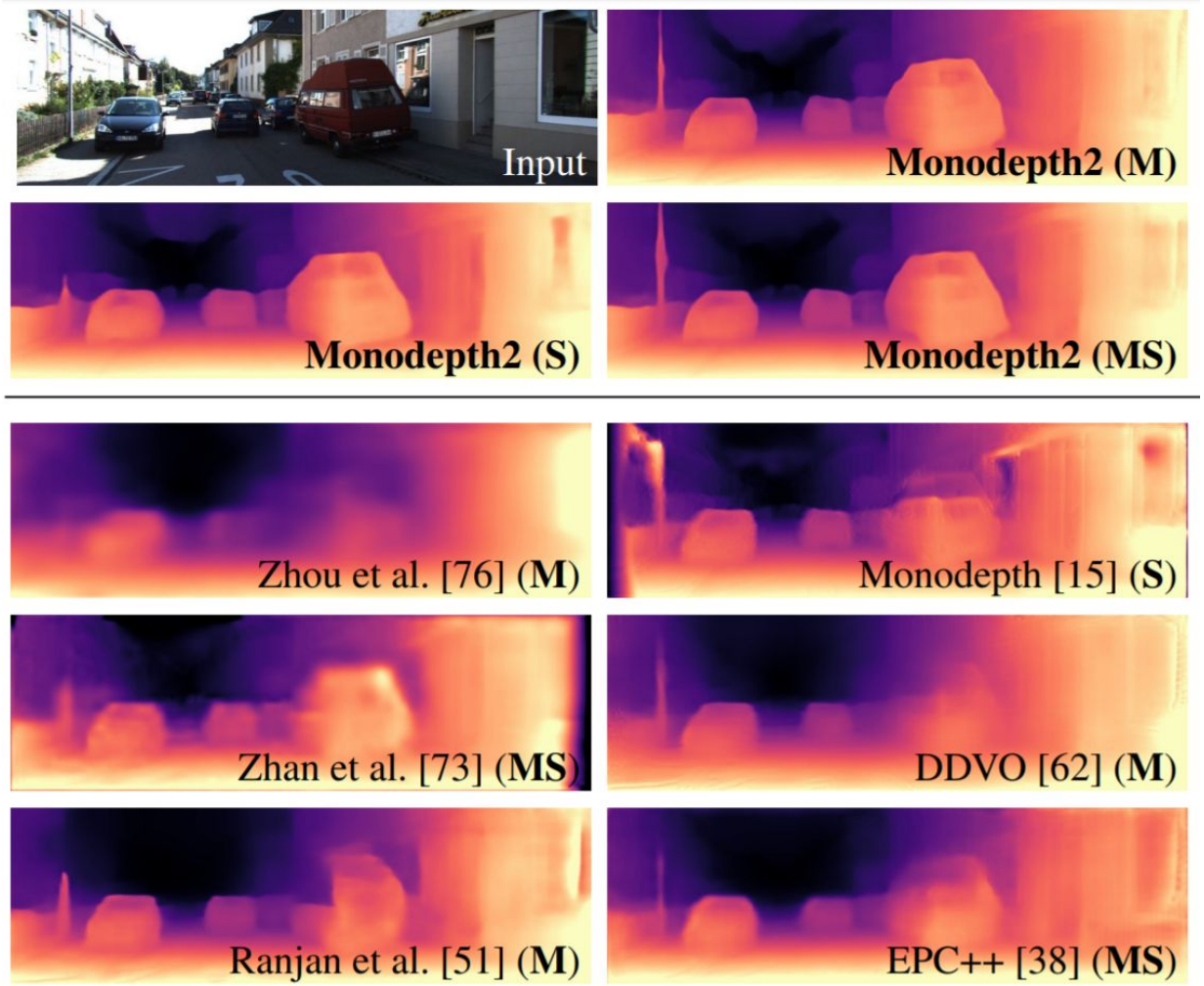
Baseline comparison

Monocular training

Stereo training

Mono + stereo training

Method	Train	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
Zhou [76]†	M	0.183	1.595	6.709	0.270	0.734	0.902	0.959
Yang [70]	M	0.182	1.481	6.501	0.267	0.725	0.906	0.963
Mahjourian [41]	M	0.163	1.240	6.220	0.250	0.762	0.916	0.968
GeoNet [71]†	M	0.149	1.060	5.567	0.226	0.796	0.935	0.975
DDVO [63]	M	0.151	1.257	5.583	0.228	0.810	0.936	0.974
DF-Net [79]	M	0.150	1.124	5.507	0.223	0.806	0.933	0.973
LEGO [69]	M	0.162	1.352	6.276	0.252	-	-	-
Ranjan [52]	M	0.148	1.149	5.464	0.226	0.815	0.935	0.973
EPC++ [39]	M	0.141	1.029	5.350	0.216	0.816	0.941	0.976
Struct2depth '(M)' [5]	M	0.141	1.026	5.291	0.215	0.816	0.945	0.979
Monodepth2 w/o pretraining	M	0.132	1.044	5.142	0.210	0.845	0.948	0.977
Monodepth2	M	**0.115**	**0.903**	**4.863**	**0.193**	**0.877**	**0.959**	**0.981**
Monodepth2 (1024 × 320)	M	**0.115**	**0.882**	**4.701**	**0.190**	**0.879**	**0.961**	**0.982**
Garg [13]†	S	0.152	1.226	5.849	0.246	0.784	0.921	0.967
Monodepth R50 [16]†	S	0.133	1.142	5.533	0.230	0.830	0.936	0.970
StrAT [44]	S	0.128	1.019	5.403	0.227	0.827	0.935	0.971
3Net (R50) [51]	S	0.129	0.996	5.281	0.223	0.831	0.939	0.974
3Net (VGG) [51]	S	0.119	1.201	5.888	0.208	0.844	0.941	**0.978**
SuperDepth [48] (1024 × 382)	S	0.112	0.875	**4.958**	**0.207**	0.852	0.947	0.977
Monodepth2 w/o pretraining	S	0.130	1.144	5.485	0.232	0.831	0.932	0.968
Monodepth2	S	**0.109**	**0.873**	4.960	0.209	**0.864**	**0.948**	0.975
Monodepth2 (1024 × 320)	S	**0.107**	**0.849**	**4.764**	**0.201**	**0.874**	**0.953**	0.977
UnDeepVO [34]	MS	0.183	1.730	6.57	0.268	-	-	-
Zhan FullNYU [73]	D*MS	0.135	1.132	5.585	0.229	0.820	0.933	0.971
EPC++ [39]	MS	0.128	0.935	5.011	0.209	0.831	0.945	**0.979**
Monodepth2 w/o pretraining	MS	0.127	1.031	5.266	0.221	0.836	0.943	0.974
Monodepth2	MS	**0.106**	**0.818**	**4.750**	**0.196**	**0.874**	**0.957**	**0.979**
Monodepth2 (1024 × 320)	MS	**0.106**	**0.806**	**4.630**	**0.193**	**0.876**	**0.958**	**0.980**

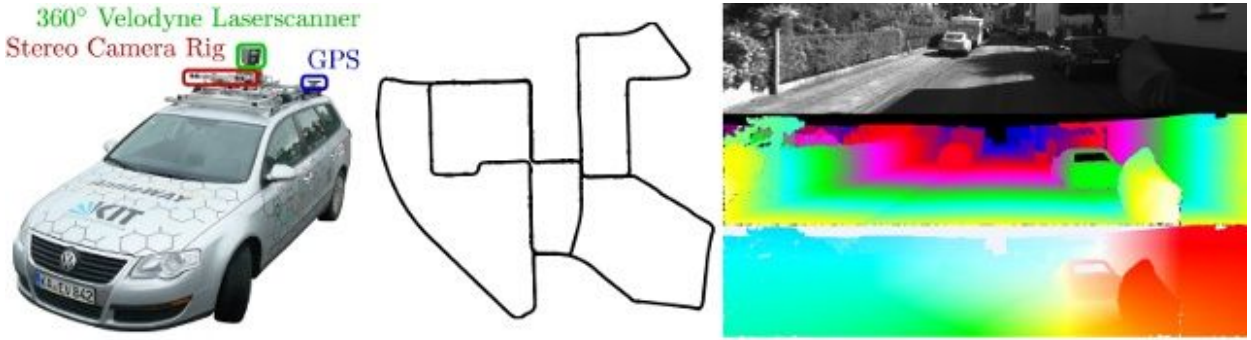

Input
Monodepth2 (M)
Monodepth2 (S)
Monodepth2 (MS)
Zhou et al. [76] (M)
Monodepth [15] (S)
Zhan et al. [73] (MS)
DDVO [62] (M)
Ranjan et al. [51] (M)
EPC++ [38] (MS)

Current Limitations

Limitations - Benchmark Datasets

KITTI

- ✓ Widely adopted
- ✓ Lidar ground truth
- ❌ Limited variation
- ❌ “Simple” scenes

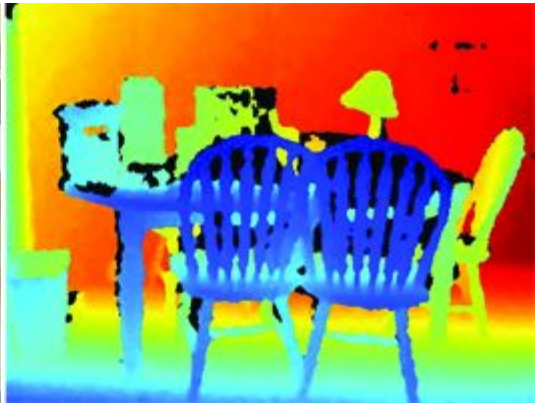


Geiger et al. Vision meets robotics: The KITTI dataset, 2013

Limitations - Benchmark Datasets

NYU Depth v2

- ✓ Widely adopted
- ✓ Indoor scenes
- ❌ Kinect GT
- ❌ Indoors only
- ❌ Static scenes



Silberman et al. Indoor segmentation and support inference from rgbd images, ECCV 2012

Limitations - Benchmark Datasets

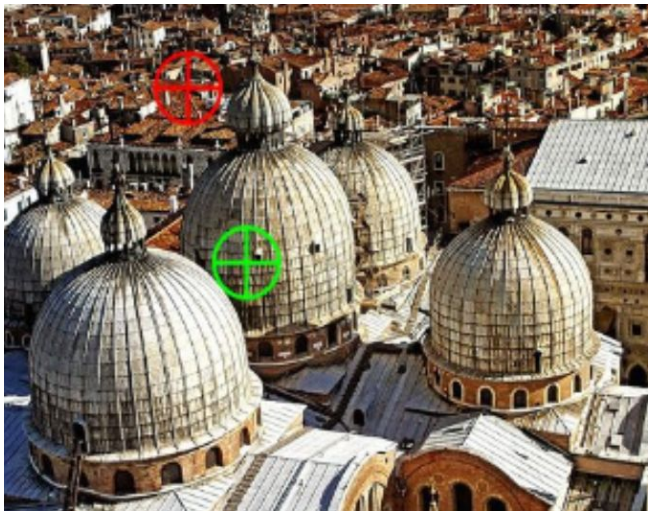
Depth in the Wild

✓ Varied

❌ No video

❌ Manually annotated

❌ Sparse



Chen et al. Single-Image Depth Perception in the Wild, NeurIPS 2016

Limitations - Benchmark Datasets

Other Sources:

Multiview stereo, stereo movies,
stereo photos, ...

✓ Varied

❌ No ground truth

Li et al. MegaDepth 2018, Xian et al. ReDWeb 2018, Ranftl et al. 3D Movies 2019, ...

Limitations - Generalization

MD2 M+S Trained on KITTI tested on Google Street View New York

Limitations - Generalization

MD2 M+S Trained on KITTI

Limitations - Generalization

MD2 M+S Trained on KITTI

Limitations - Moving Objects

Objects that are typically observed as moving during training (e.g. predestrains) are still a problem for monocular training.

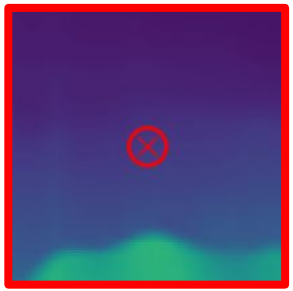
Limitations - Thin Structures

Left Camera Image

Monodepth2 prediction

Example from Watson et al. Self-Supervised Monocular Depth Hints, ICCV 2019

Limitations - Thin Structures



Left Camera Image

Zoom-in on highlighted area

Monodepth2 prediction

Example from Watson et al. Self-Supervised Monocular Depth Hints, ICCV 2019

Limitations - Temporal Stability

Frames are processed independently causing “flickering”.

We can train deep networks to predict **depth** from a single image **without depth supervision**.

Still lots to be done ….

PyTorch code and pretrained models: https://github.com/nianticlabs/monodepth2

Wind Walk Travel Videos: https://www.youtube.com/watch?v=jv4m41pbOJE