Machine-learning for low-level vision problems

Stefano Mattoccia

http://vision.disi.unibo.it/~smatt

Dipartimento di Informatica, University of Bologna

DEEP LEARNING ON-CHIP

September 20-22, 2017 – Politecnico di Torino, Torino (Italy)

Outline

The talk is organized as follows:

1) Machine learning applied to low-level vision problems
    - Depth sensing
    - Confidence estimation
    - Recent trends

2) Mapping of computer vision algorithms with HLS tools
    - Most demanding layers of a CNN (convolutions)

For 1) a small background on stereo vision and confidence estimation is needed

For a more detailed introduction to stereo vision:

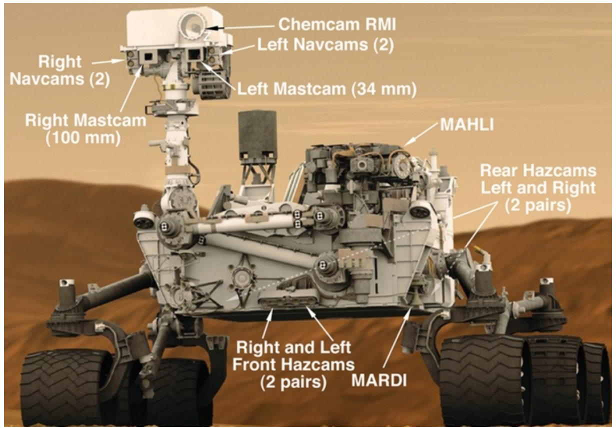
http://vision.deis.unibo.it/~smatt/Seminars/StereoVision.pdf

Most computer vision applications rely on low level features extracted from images:

- **Feature detection and description**
- **Segmentation**
- **...**
- **Depth**

Depth is of paramount importance for several applications:

- **Autonmous driving**
- **Robot picking**
- **Augmented reality**
- **Face recognition**
- **Gaming**
- **. . .**



NASA Mars rover

Google car

DJI drones

Apple iPhone X

Microsoft Xbox and Kinect

Depth sensors

Provide a depth/disparity map D and, in most cases, a conventional 2D (RGB) image

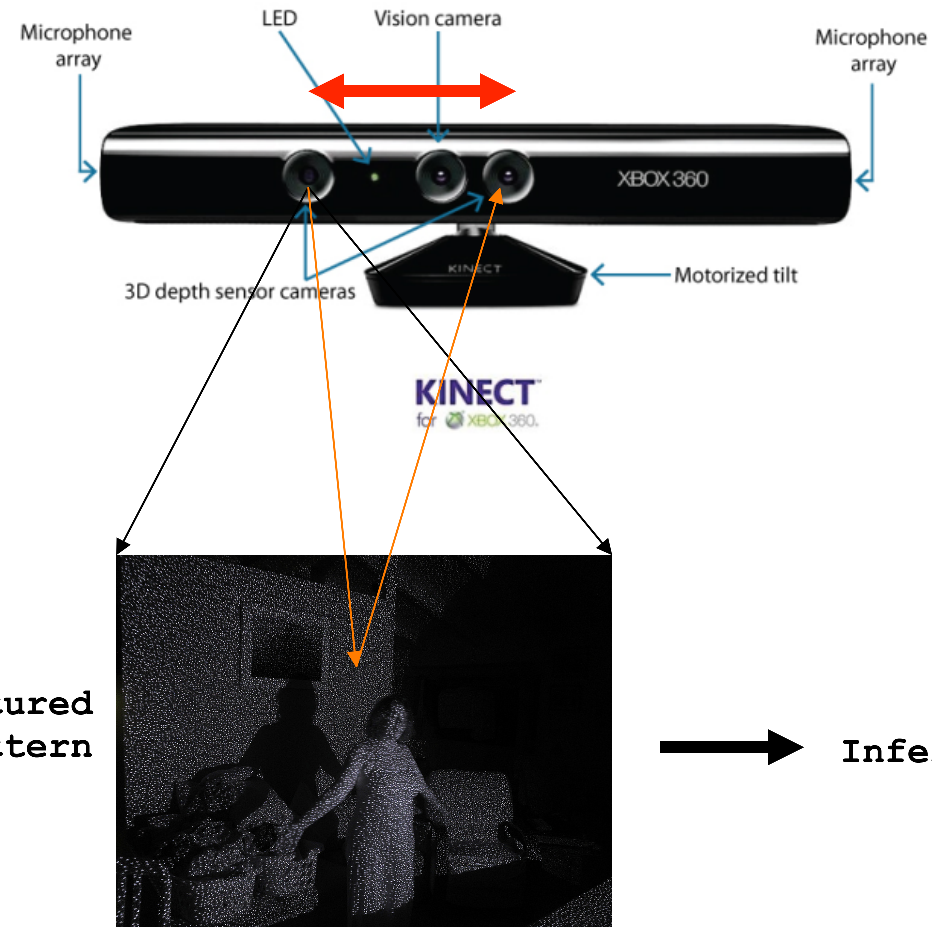
Two main technologies:

- Active
  - Structured light (Kinect 1)
  - ToF - Time of Flight (Kinect 2)
  - LIDAR (Velodyne)
- Passive
  - Stereo vision
  - Monocular depth* sensors based on ML

Active RGBD sensors: structured light

- **Kinect 1**

  PrimeSense/Microsoft now Apple (iPhone X?)

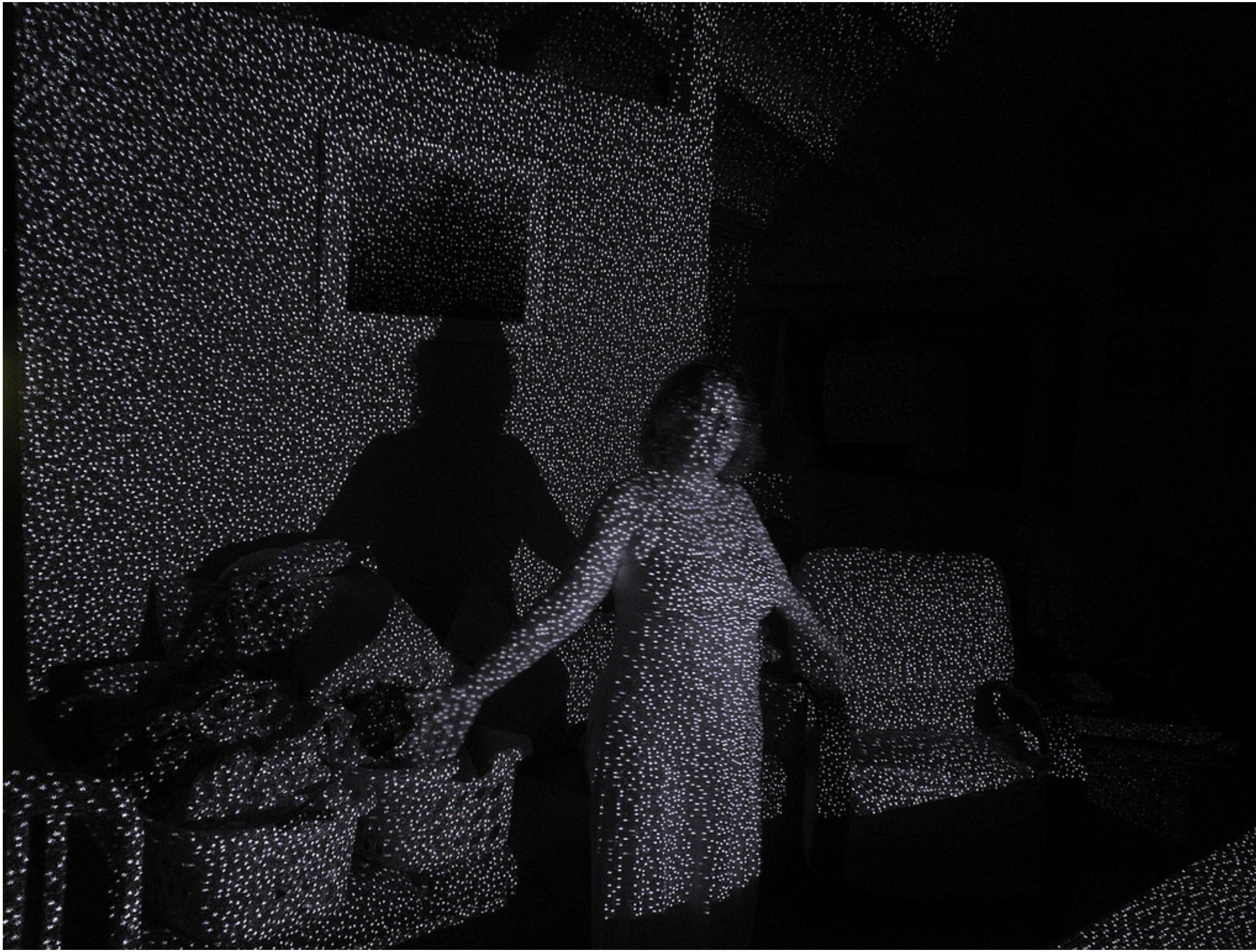


Structured IR pattern

www.flickr.com

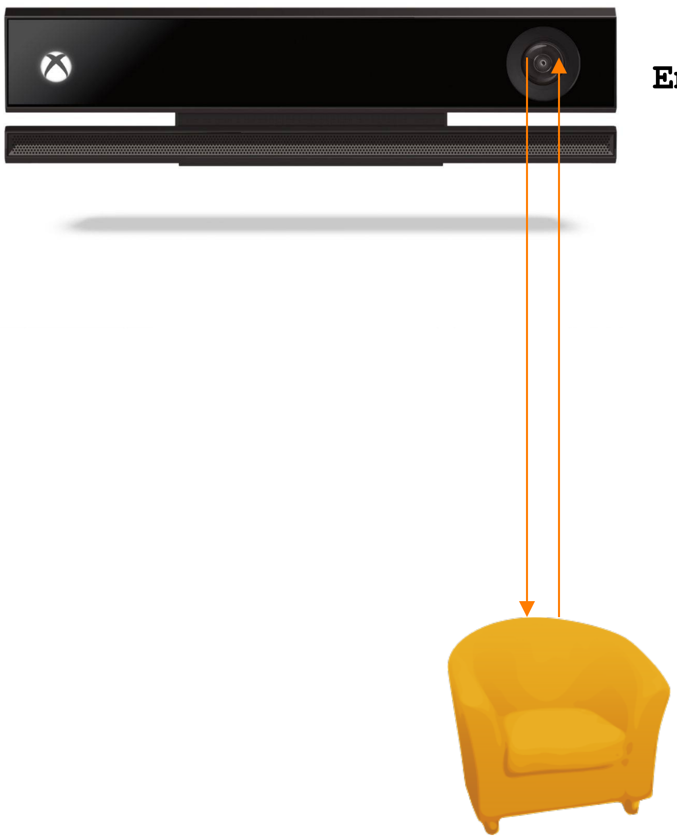
Infers depth by triangulation

Accuracy	
Indoor	
Outdoor	
Wearable	
Long range	
RGB	
Cost	$



Active RGBD sensors: ToF

- Microsoft Kinect 2



Emitter/Receiver

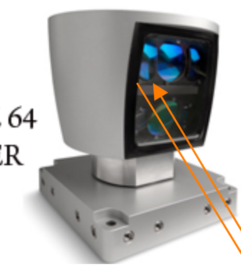
Accuracy	
Indoor	
Outdoor	
Wearable	 
Long range	
RGB	
Cost	$

Infers depth by measuring the bouncing time of signals (path, from emitter to receiver)

Active depth sensors: LIDAR

- LIght Detection And Ranging

VELODYNE 64
BEAM LASER

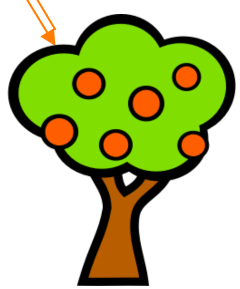


Emitter/Receiver

Mechanical parts

www.velodyne.com

Light/Laser



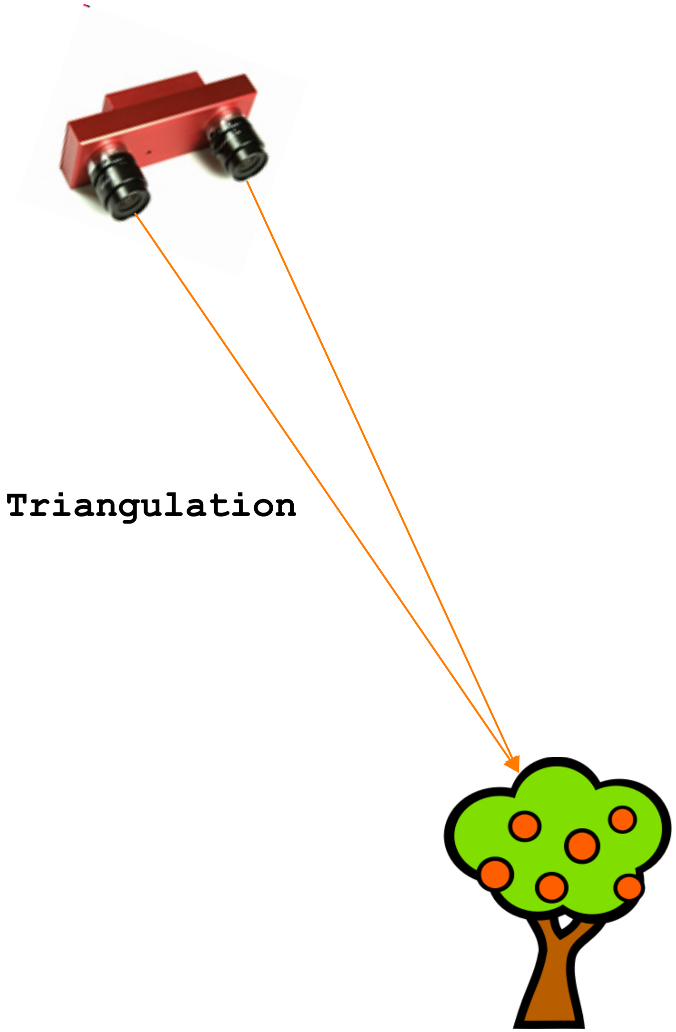
www.google.com

Accuracy	👍👍
Indoor	👍
Outdoor	👍
Wearable	👎
Long range	👍👍
RGB	👎
Cost	$$$$$$$

Infers depth by measuring bouncing time of a laser signal (path, from emitter to receiver)

Passive RGBD sensors: stereo vision

- Passive stereo (binocular)



Accuracy

Indoor

Outdoor

Wearable*

Long range

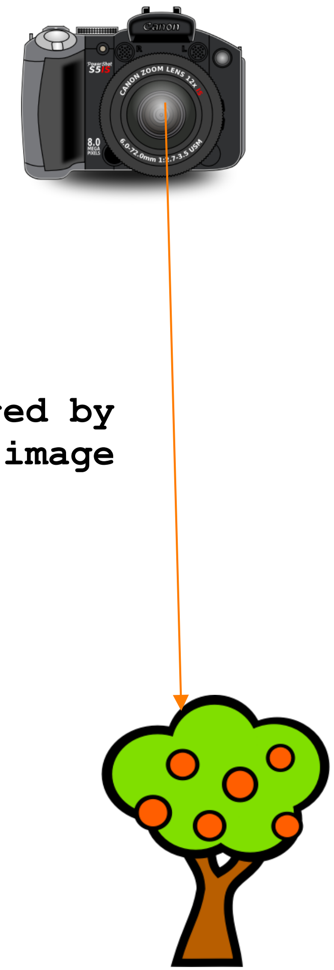
RGB

Cost $

Infers depth by finding
corresponding points in two images

Passive RGBD sensors: monocular depth* camera

- **Monocular camera**



Depth* is inferred by
a single image

Accuracy

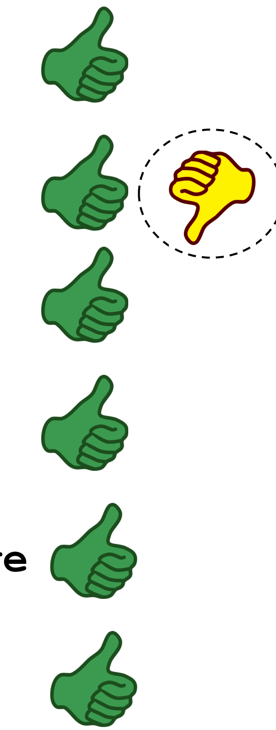
Indoor

Outdoor

Wearable*

Long range

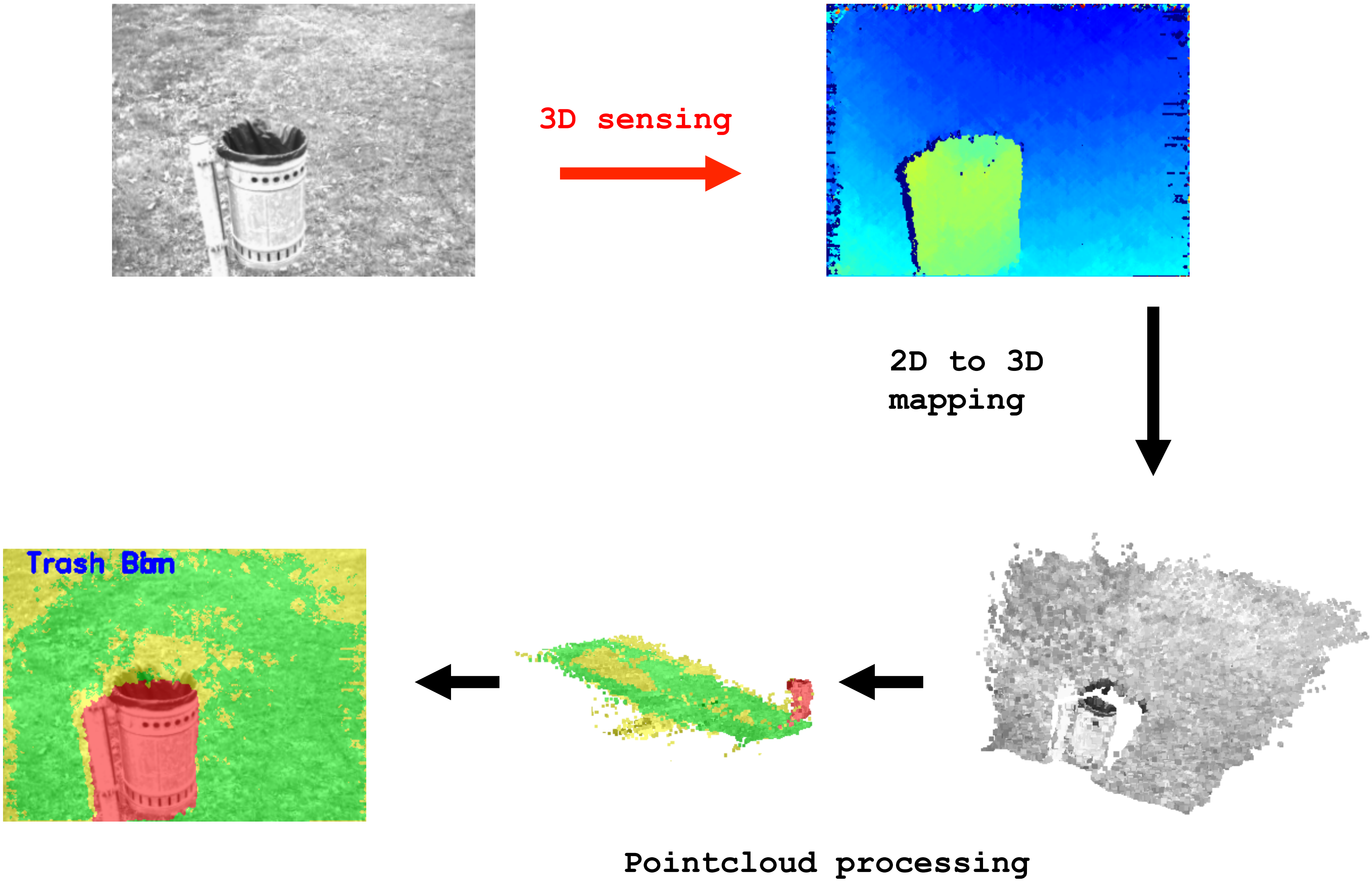
RGB



Cost $

* Not a "true" depth: the absolute distance is unknown

http://visualfunhouse.com/

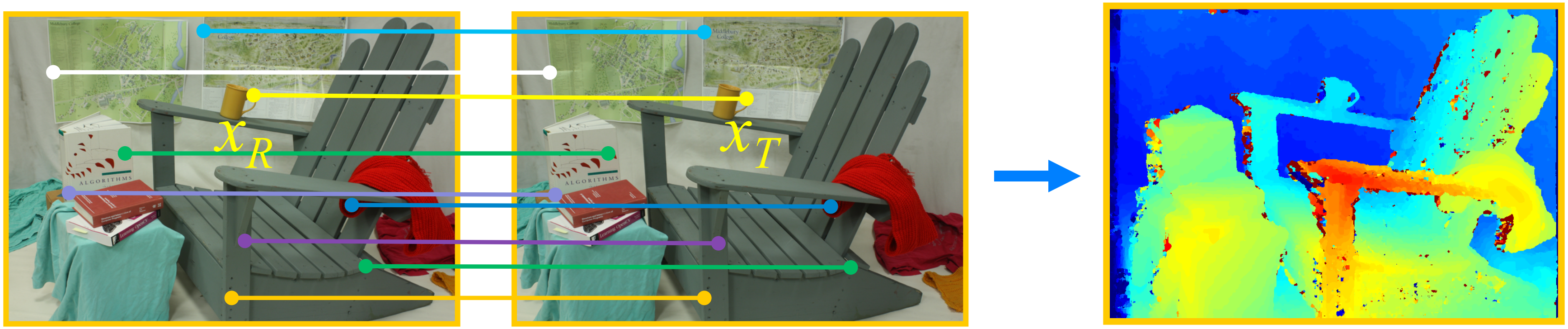

3D sensing
2D to 3D
mapping
Trash Bin
Pointcloud processing

Stereo sensing: problem definition

- Given two (or more) synchronised* images of the same area infer the 3D coordinates of each point in the sensed scene

- Images acquired with standard cameras

1) Find corresponding points (difficult)

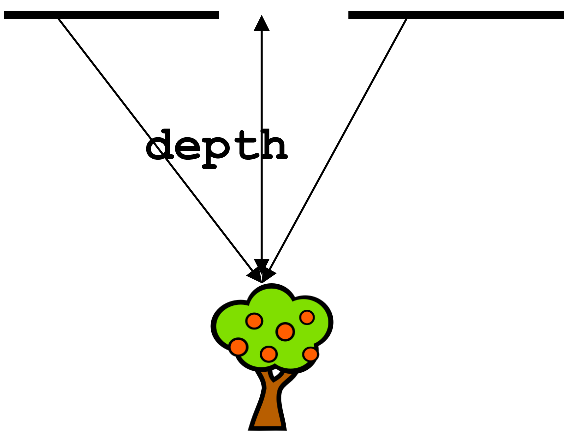


Reference

Target

Disparity map

2) Triangulate to infer depth (straightforward)



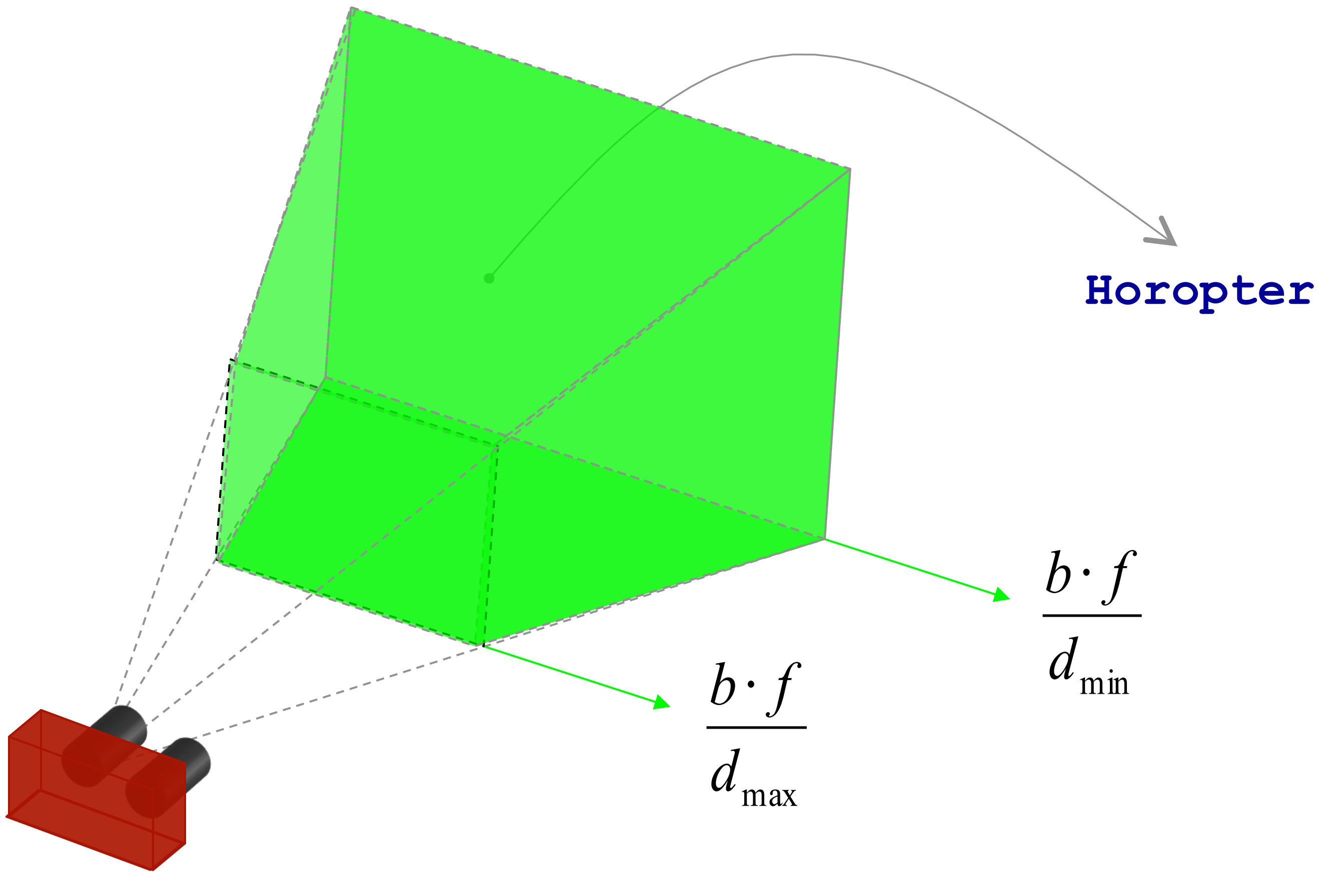
$$Z = \frac{b \cdot f}{x_R - x_T} = \frac{b \cdot f}{d}$$

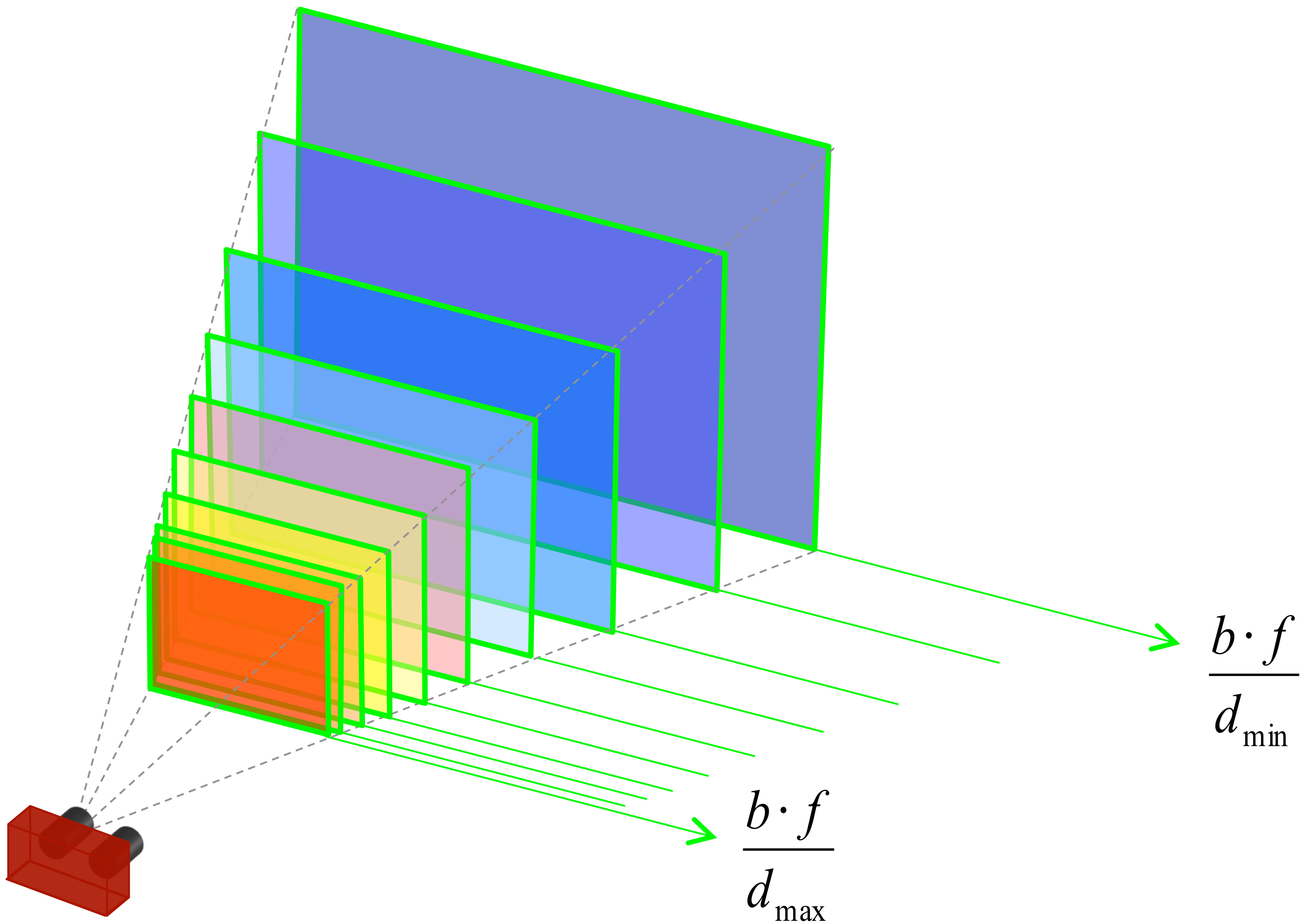
$$X = \frac{Z \cdot (x - xc)}{f}$$

$$Y = \frac{Z \cdot (y - yc)}{f}$$

Range field (Horopter)

Given a stereo rig with baseline b and focal length f, the range field of the system is constrained by the disparity range [d_{min}, d_{max}].



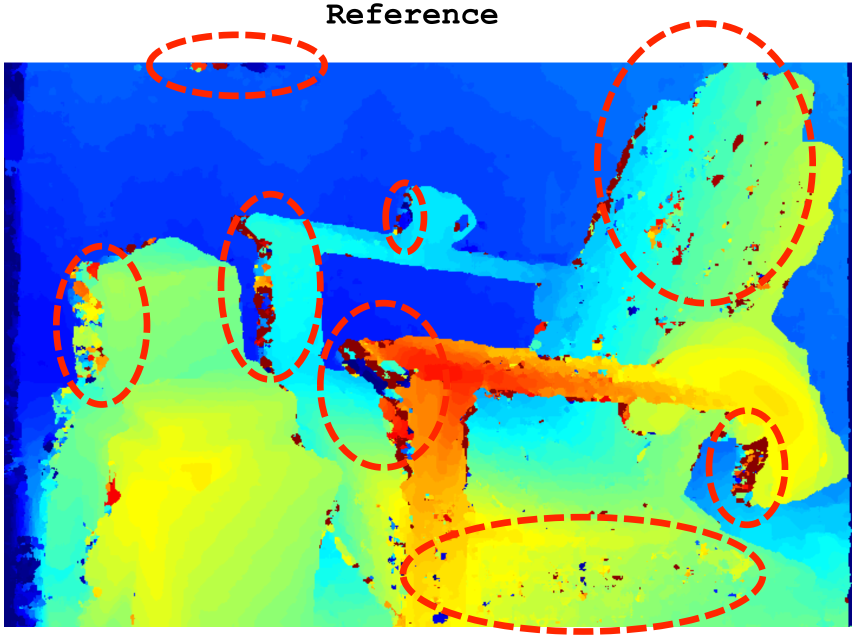


- **Depth measured by a stereo vision system is discretized into parallel planes (one for each disparity value)**
- **A better (virtual) discretization can be achieved with subpixel techniques**

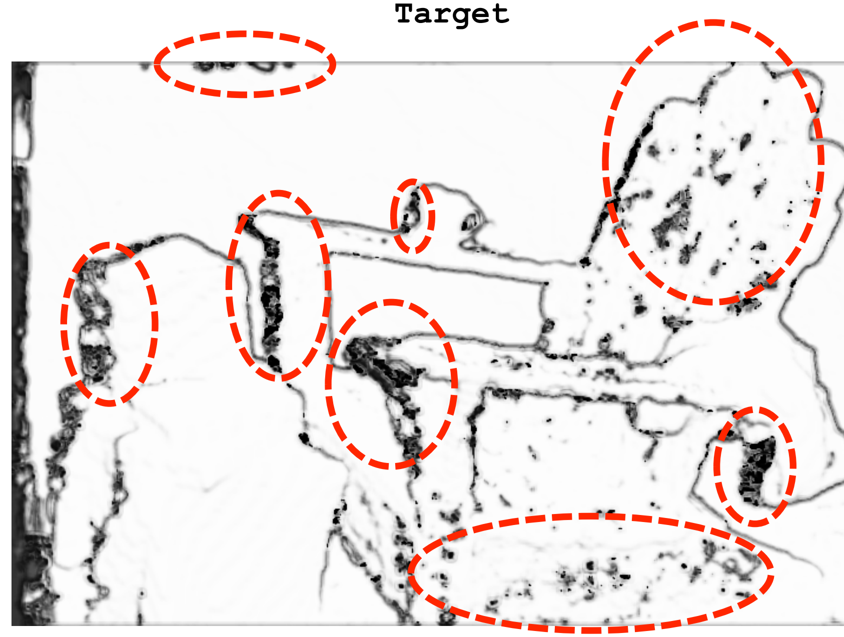
Confidence measure

Reference

Target

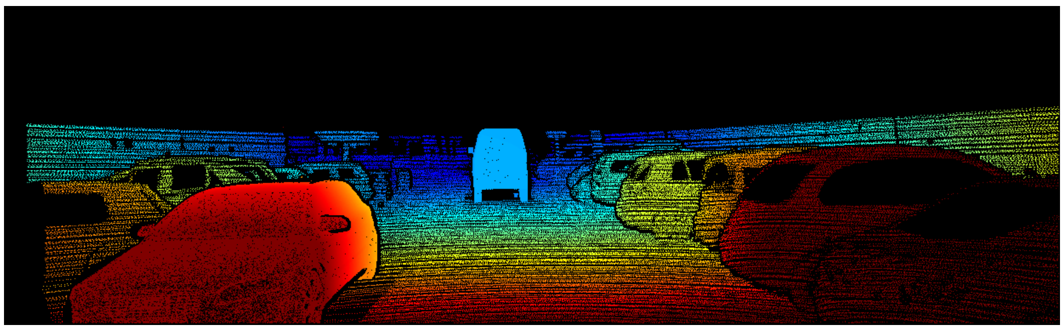


Disparity map



Confidence map

Outdoor dataset: KITTI 2012 (and 2015)

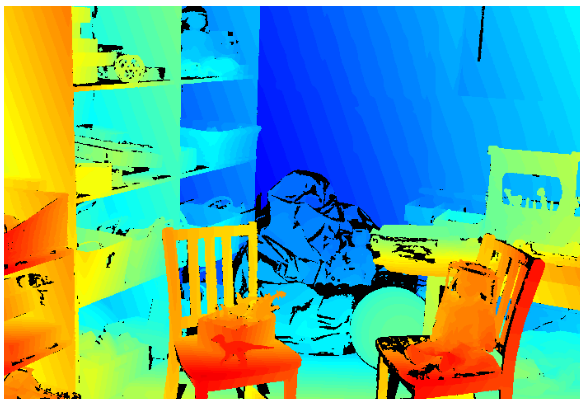


Groundtruth (GT)

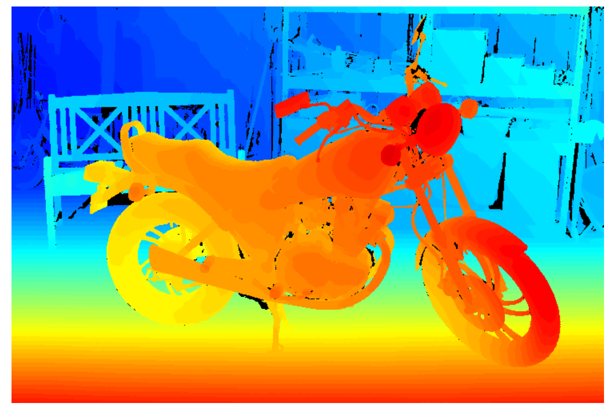
Training	:	194 (200)
Sequences	:	21 images/frame without GT
Testing	:	195 (200)

http://www.cvlibs.net/datasets/kitti/

Indoor dataset: Middlebury 2014



Groundtruth (GT)



Groundtruth (GT)

Training	:	15
Sequences	:	Na
Testing	:	10

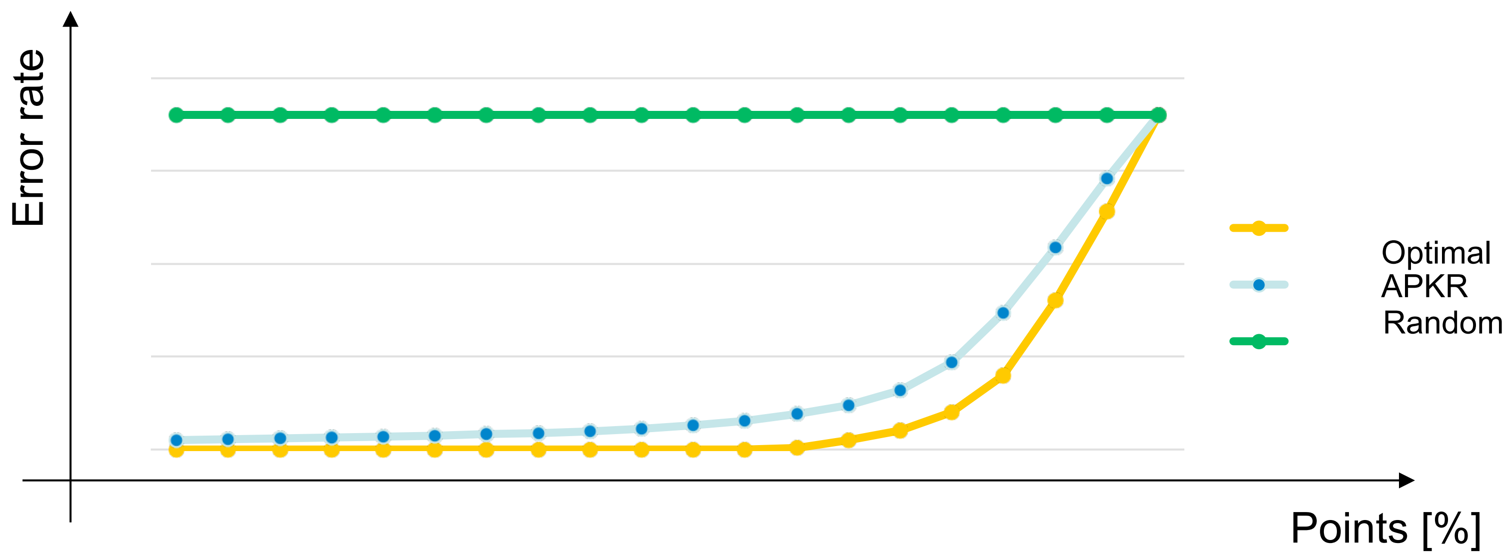
http://vision.middlebury.edu/stereo/eval3/

Stereo algorithms evaluation

- **Error rate or MSE wrt GT data**
- **Often the error bound is set > 1 (not perfect GT)**
- **Testing GT data not available to users**

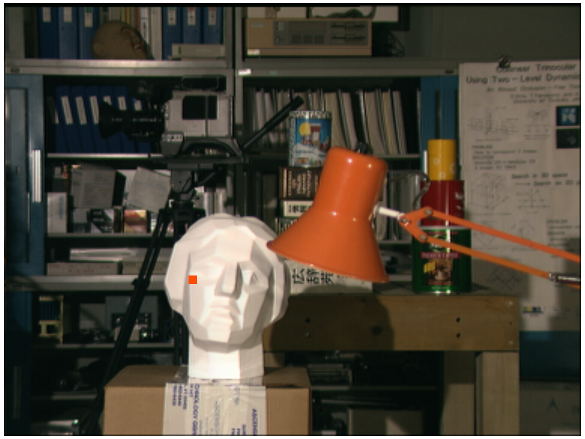
Confidence measures: evaluation

- **The Area Under the ROC Curve (AUC) is the metric to evaluate confidence measures**
- **The lower, the better**



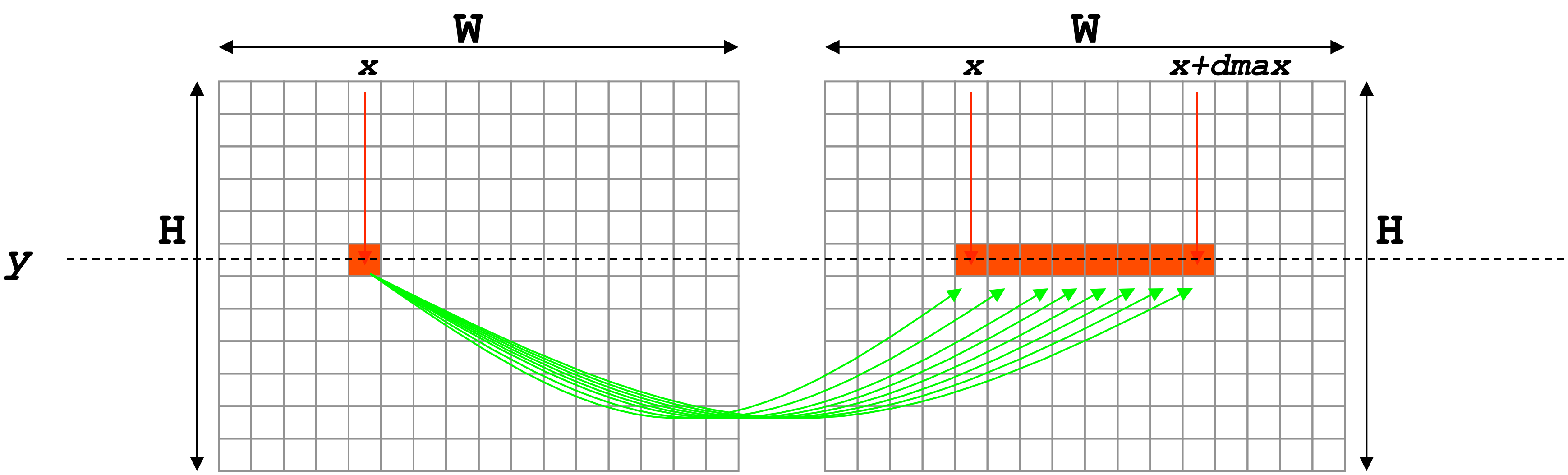
X. Hu and P. Mordohai, "A Quantitative Evaluation of Confidence Measures for Stereo Vision", PAMI 2012

The simplest stereo approach



Reference (R)

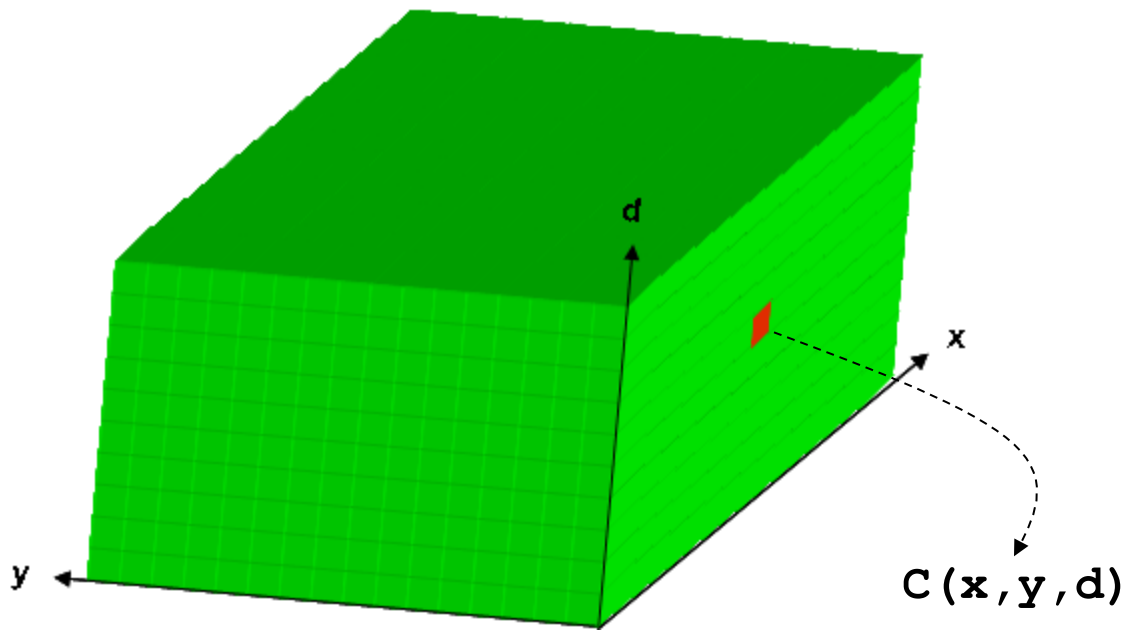
Target (T)



Reference (R)

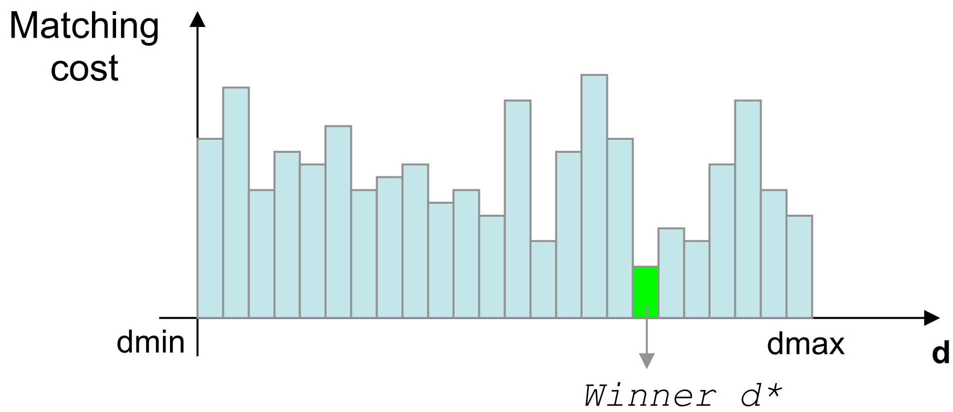
Target (T)

The Disparity Space Image (DSI), aka Cost Volume, is a 3D matrix (WxHx(d_{max}-d_{min}))



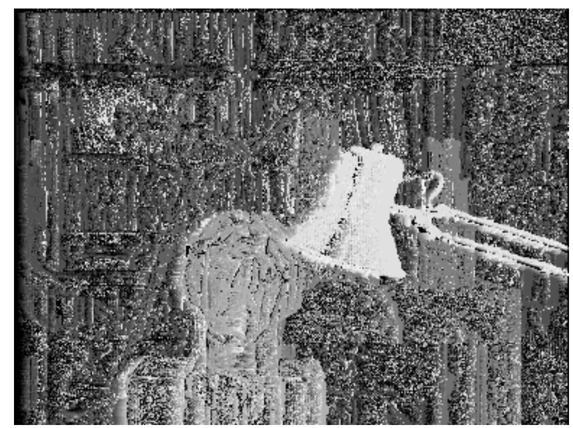
likelihood/confidence of each correspondence

Each element $C(x,y,d)$ of the DSI represents the cost of the correspondence between $I_R(x_R,y)$ and $I_T(x_R+d,y)$ according to the adopted cost function (e.g. Sum of Absolute Differences)



Reference

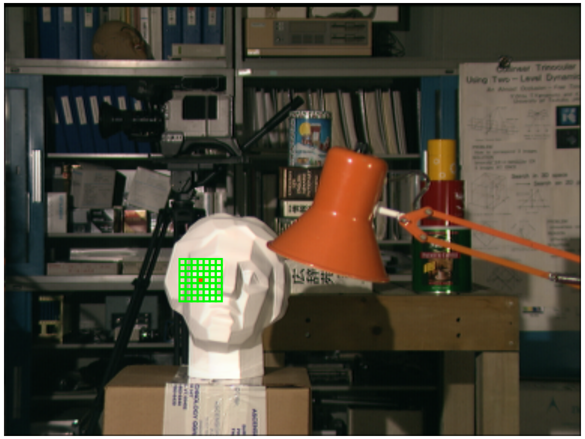
GT



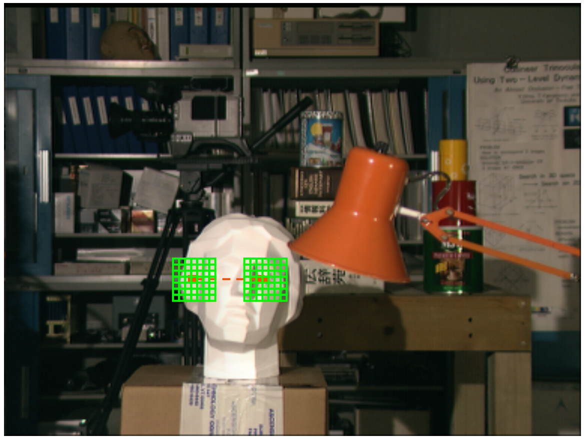
Poor results

Local approaches

To reduce ambiguity costs are aggregated over a patch



Reference (R)



Target (T)

Global (and semi-global*) approaches

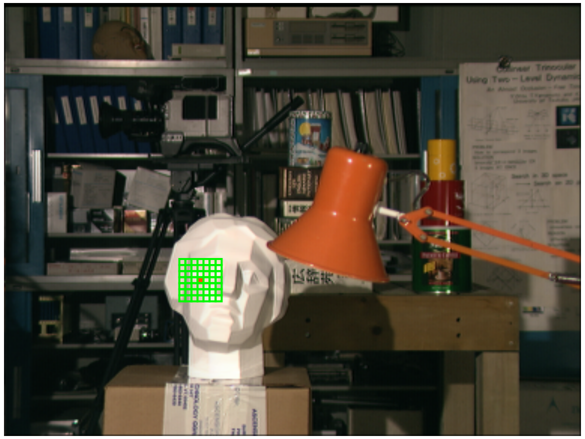
Minimize an energy term over the whole* stereo pair

$$E(d) = E_{data}(d) + E_{smooth}(d)$$

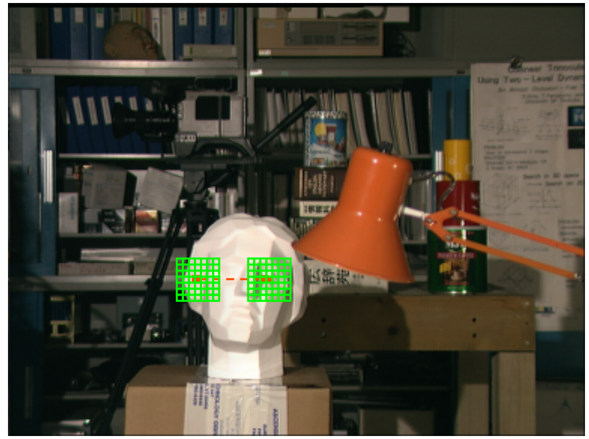
* subset of the stereo pair

Fixed window (aka BM)

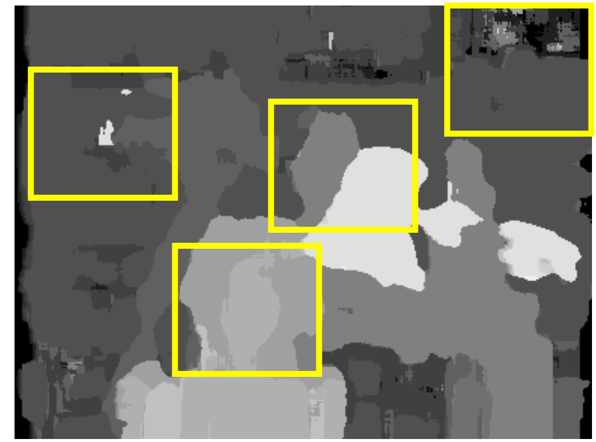
- Simple cost aggregation/mean over a patch



Reference (R)

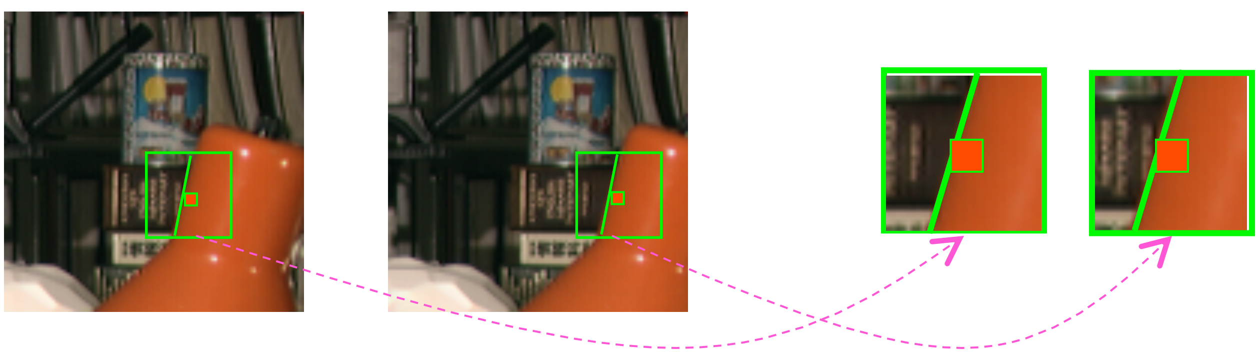


Target (T)



Fixed Window (FW)

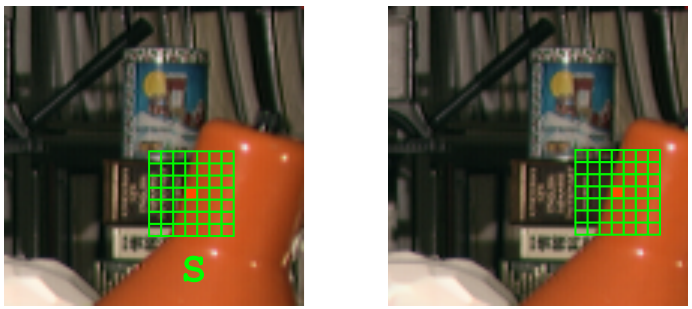
What's wrong with this method?



Background is misaligned !

[ADAPTIVE] K. Yoon and I. Kweon. Adaptive support-weight approach for correspondence search, PAMI, 2006

State-of-the-art cost aggregation strategies aim at shaping the support in order to include only points with the same (unknown) disparity



FW

Ideal

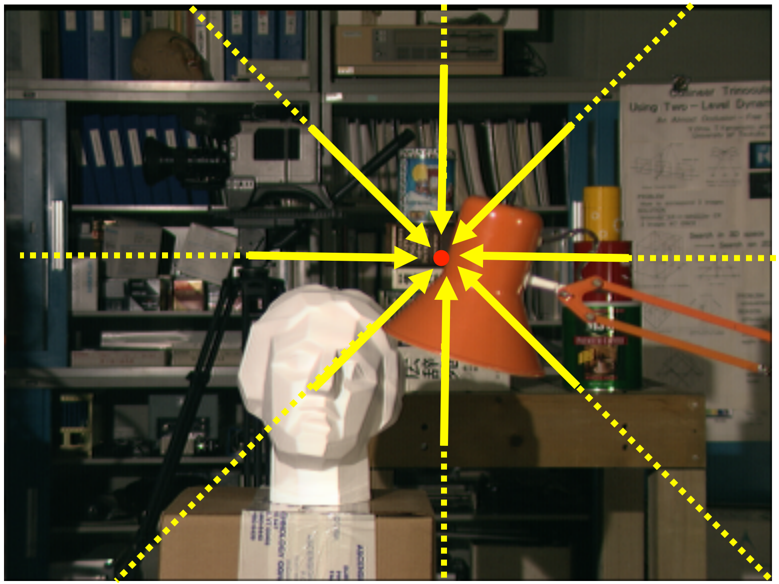
FW: decreasing the size of the support helps in reducing the border localization problem

However, this choice renders the correspondence problem more ambiguous (especially when dealing with uniform regions)

In practice, the choice of the optimal size empirically set

S. Mattoccia, S. Giardino,A. Gambini, Accurate and efficient cost aggregation strategy for stereo correspondence based on approximated joint bilateral filtering, Asian Conference on Computer Vision (ACCV2009)

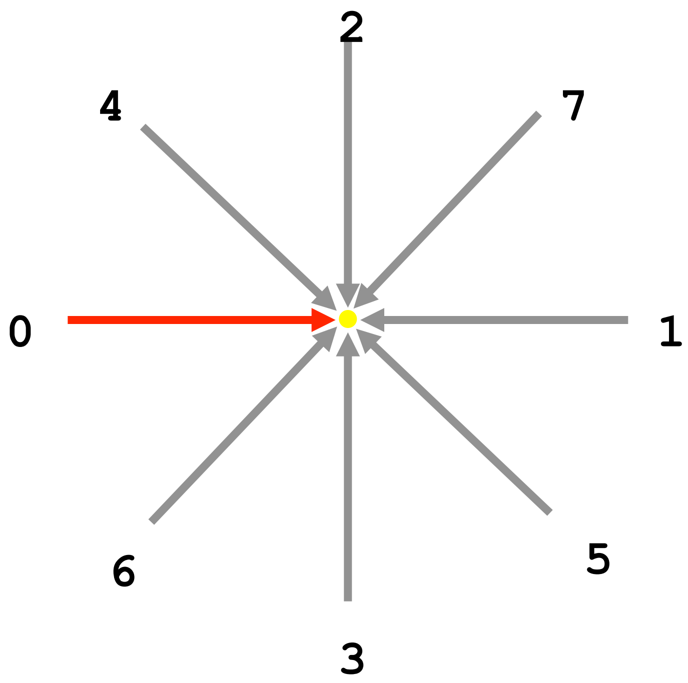
Semi Global Matching (SGM)



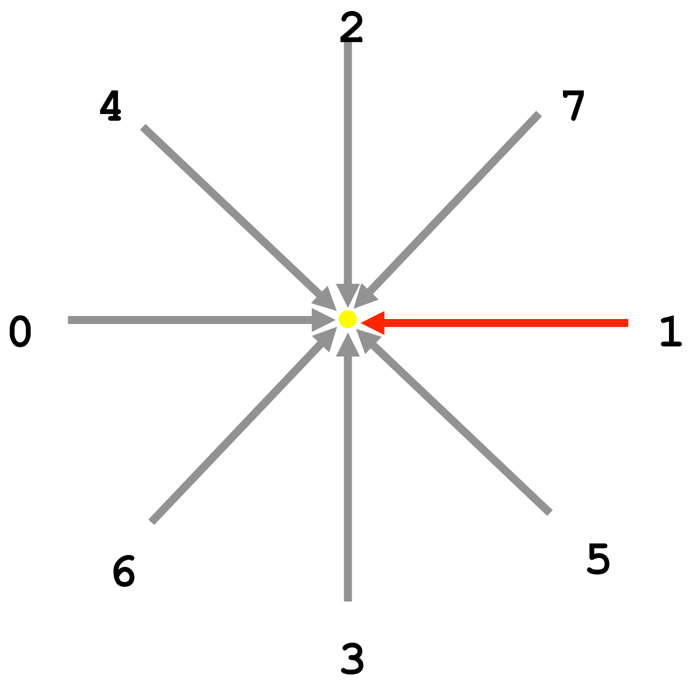
- Fast
- Accurate near discontinuities and in texture-less regions
- Combine/sum of simple disparity optimizatations along multiple scanlines
- High memory footprint (the whole DSI is required)

[SGM] H. Hirschmüller. Stereo vision in structured environments by consistent semi-global matching, PAMI 2008

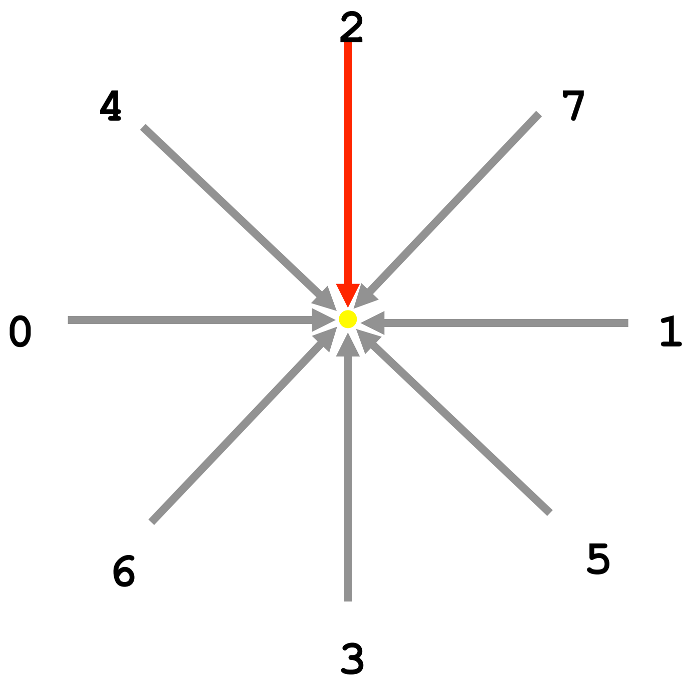
Scanline 0



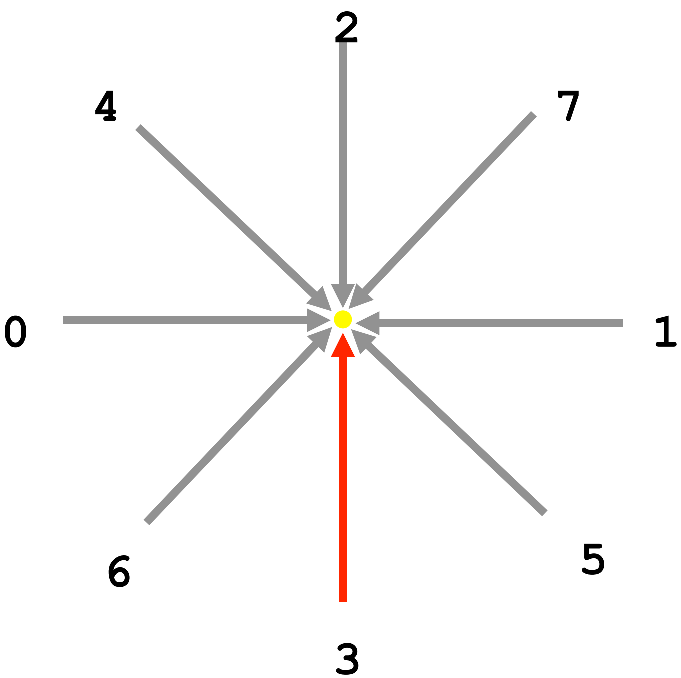
Scanline 1

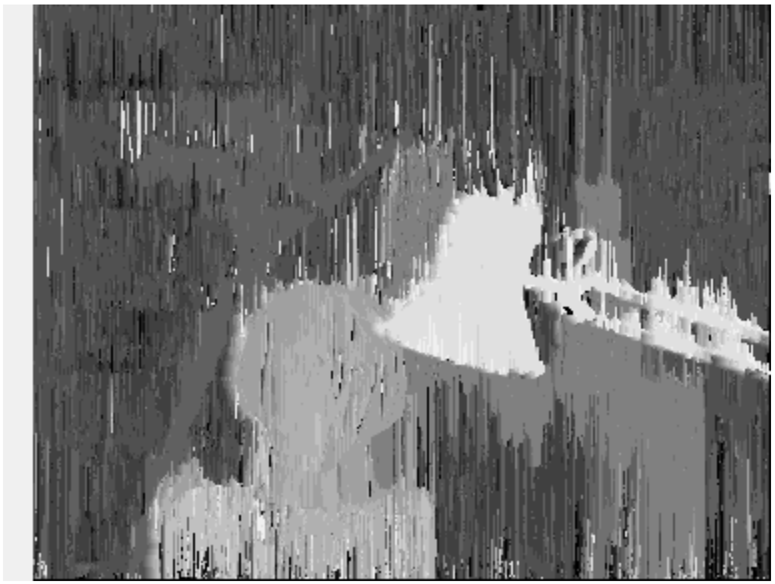


Scanline 2

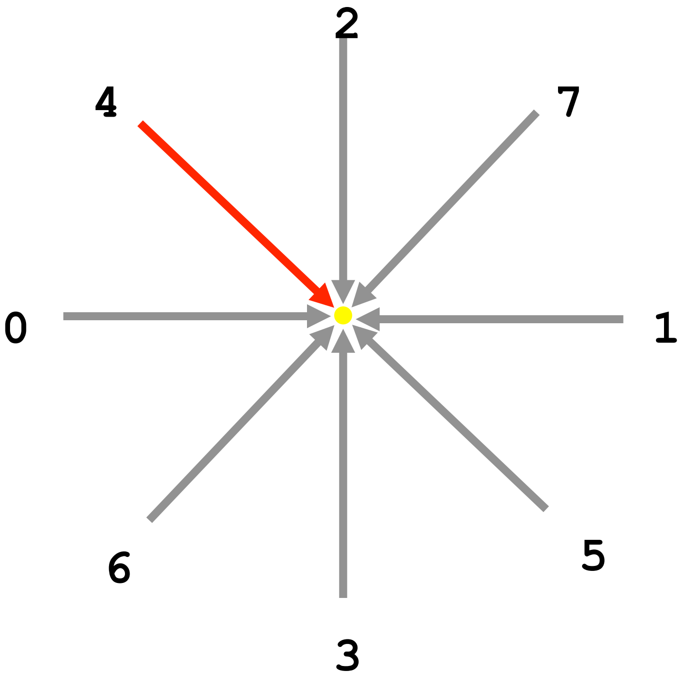


Scanline 3

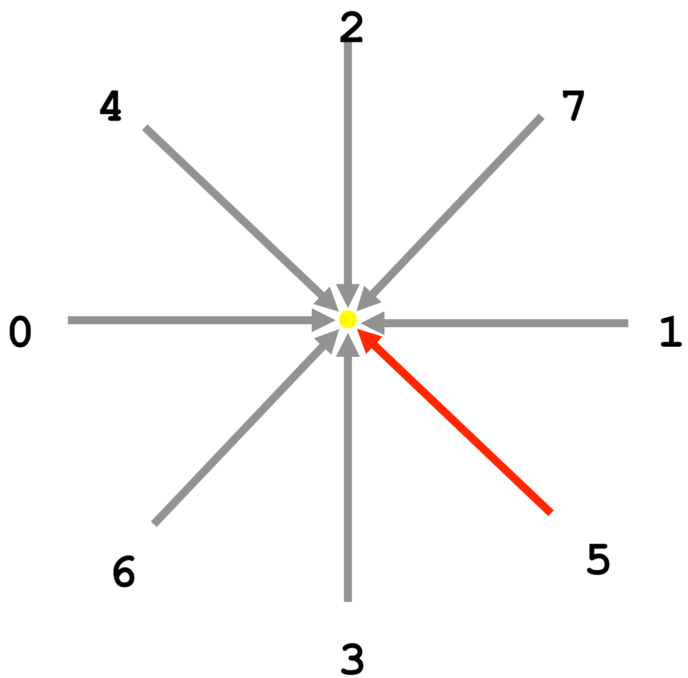


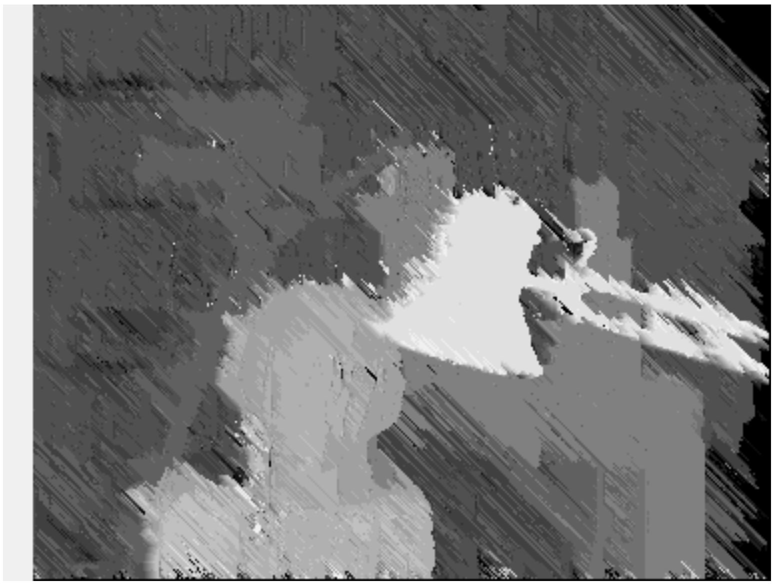


Scanline 4

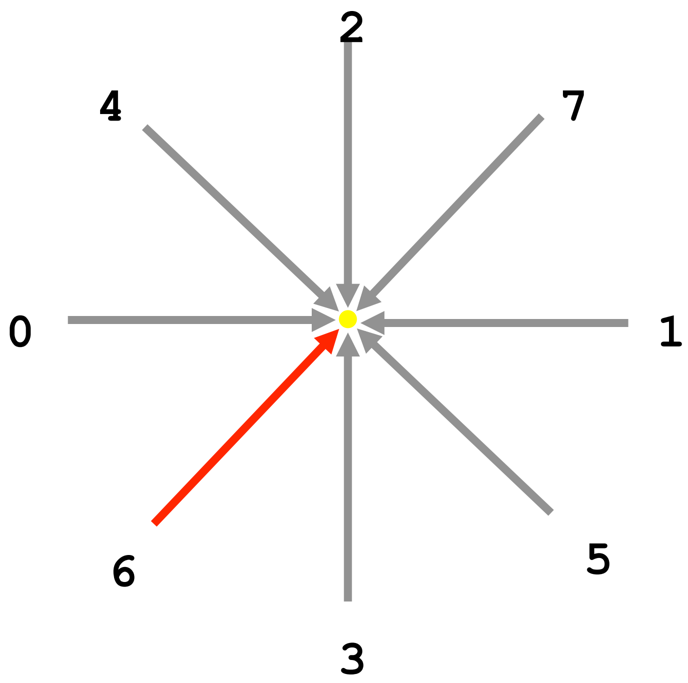


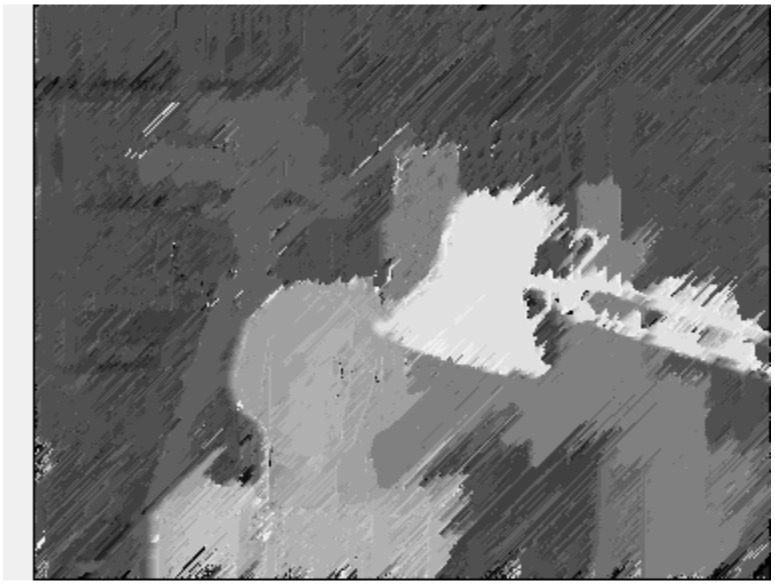
Scanline 5



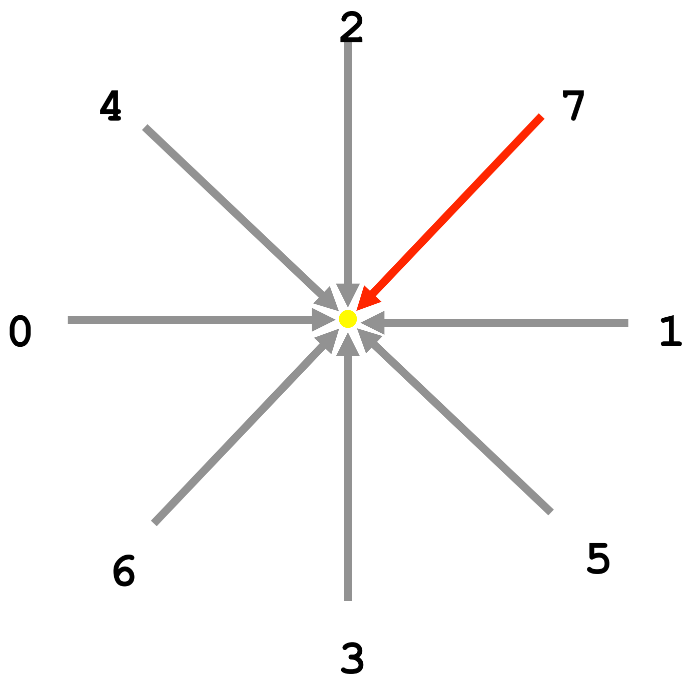


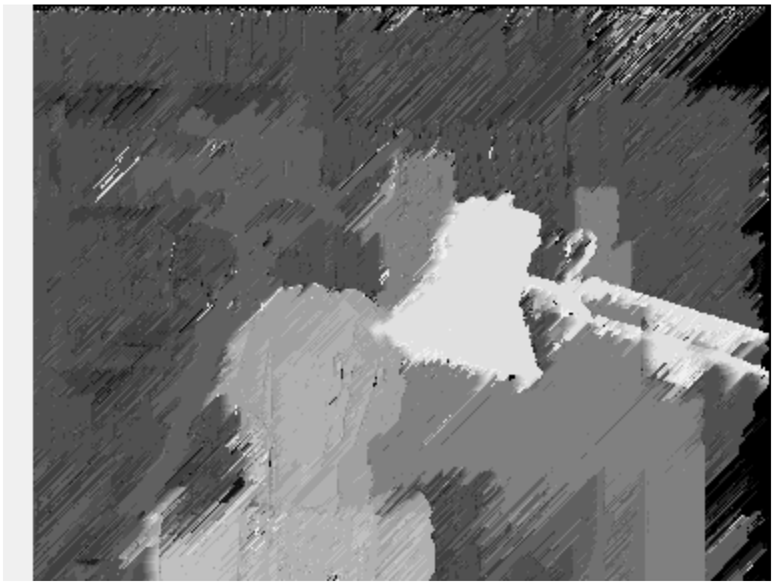
Scanline 6



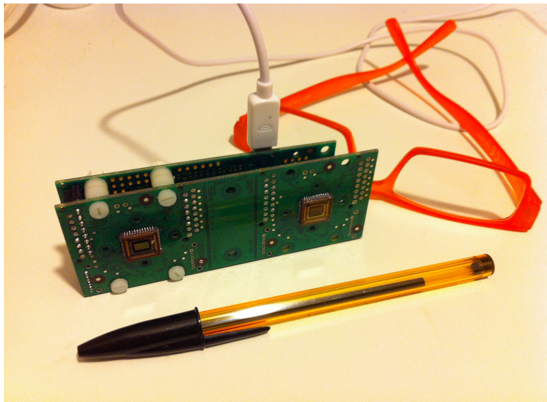


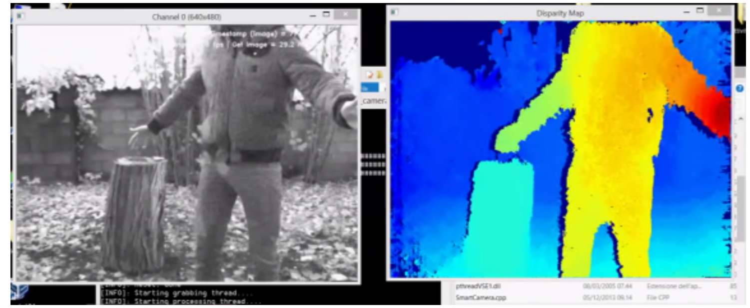
Scanline 7



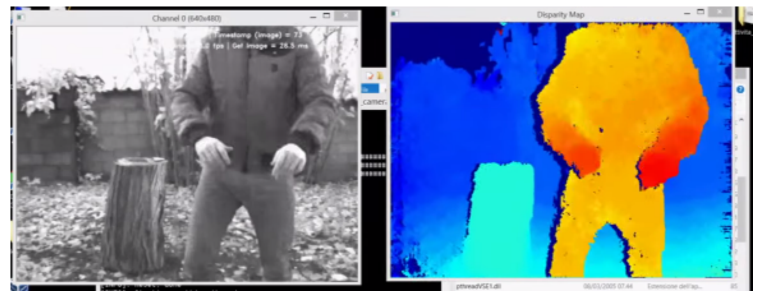


Custom FPGA-based stereo camera: BM and SGM





www.youtube.com/watch?v=KXFWIvrcAYo



- Processing at 30+ fps (640x480)
- Power consumption: < 2.5 Watt
- Self powered via USB cable
- Weight: < 80 g with lens and holder
- Devices: Xilinx Spartan 6 and Zynq

Matching cost: SAD

34	45	44
26	38	45
17	27	31

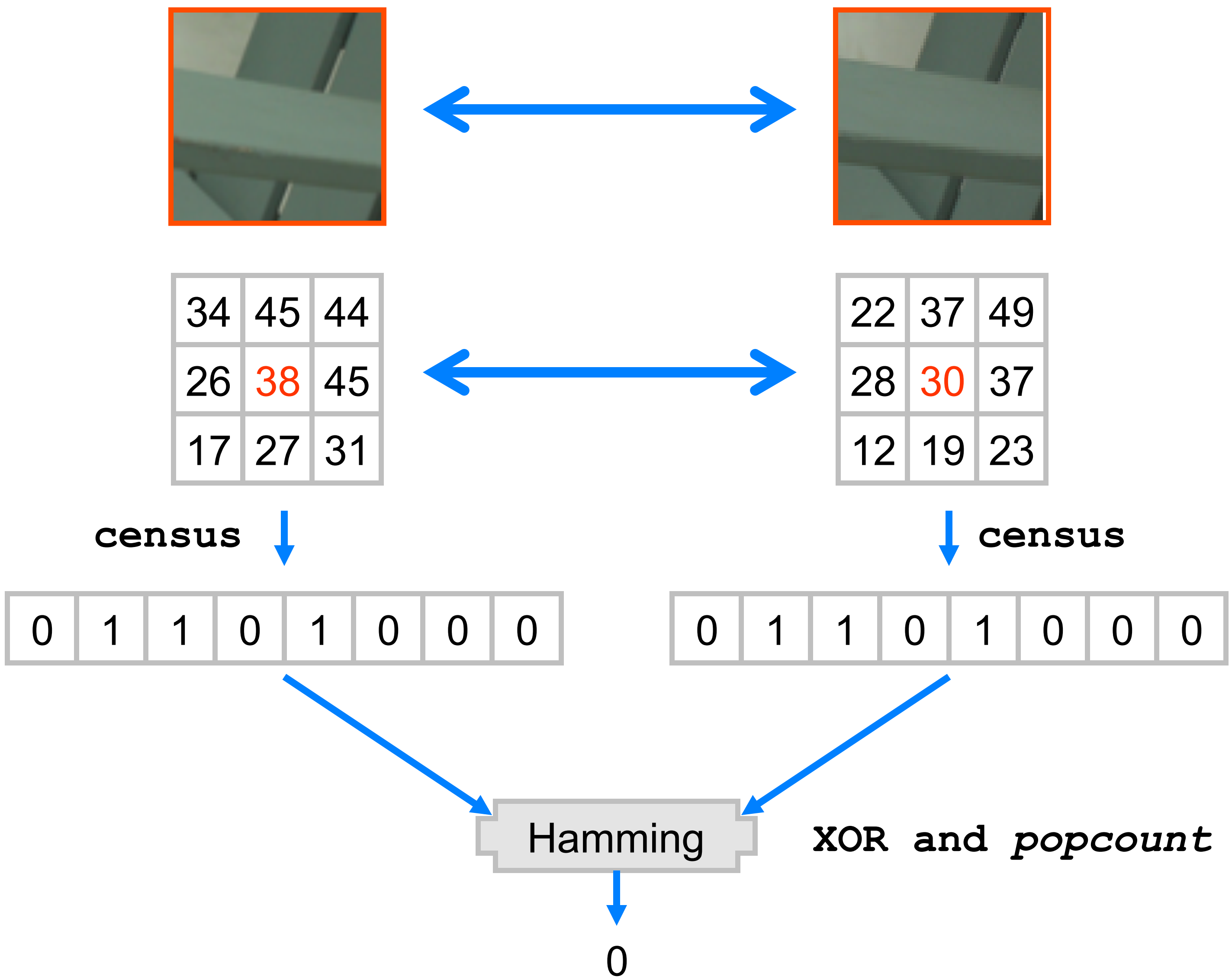
22	37	49
28	30	37
12	19	23

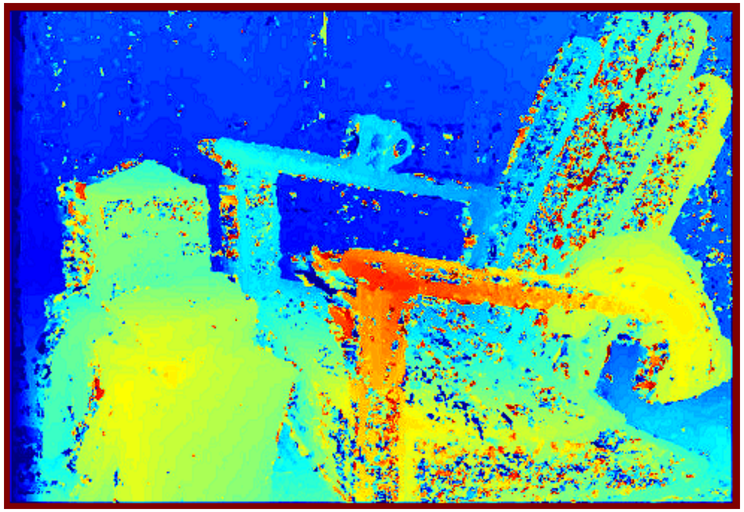
SAD = 64

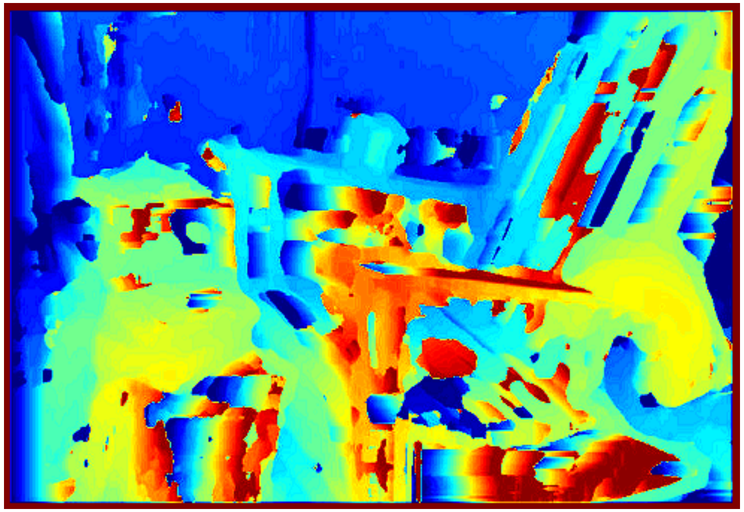
|34-22| + |45-37| + |44-49| + |26-28| + |38-30| + |45-37| + |17-12| + |27-19| + |31-23|

Matching costs: census

A more robust matching function is based on the census transform and the Hamming distance





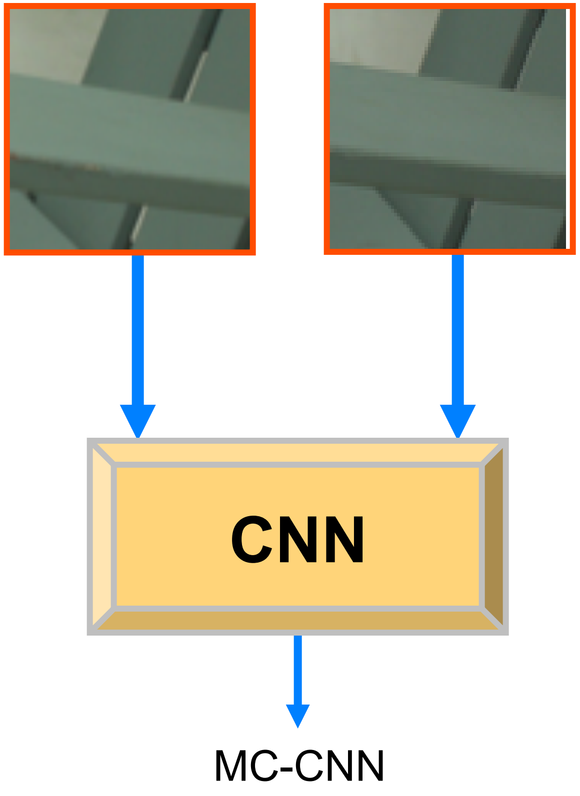


census + Hamming

SAD

Matching cost with deep-learning: MC-CNN

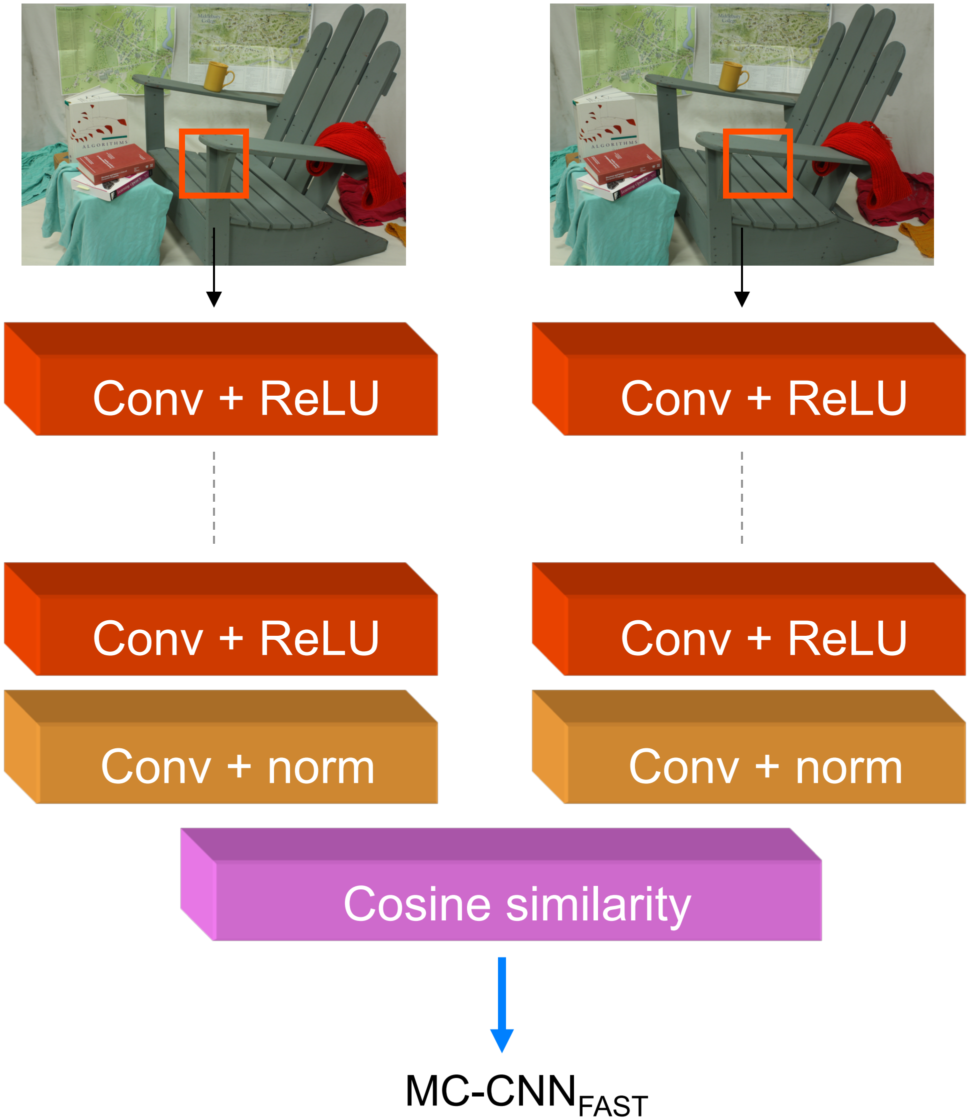
- First end-to-end approach to learn a cost function
- Trained with right and wrong samples (from GT data)
- State-of-the-art method



Training phase:
1 point is 1 sample

J. Zbontar, Y. LeCun, “Computing the stereo matching cost with a convolutional neural network”, CVPR 2015

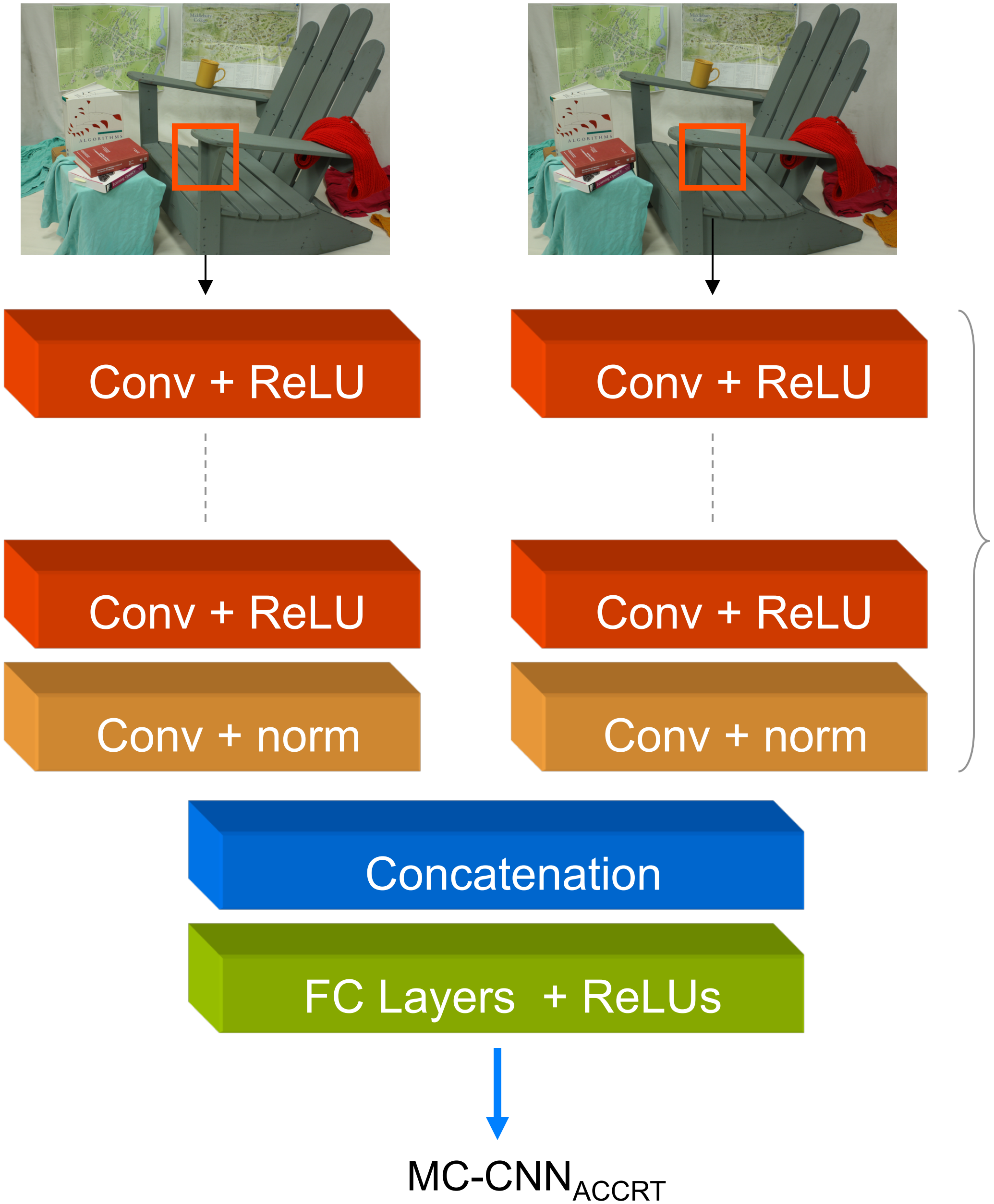
Matching cost with CNN (MC-CNN) fast



Siamese network with shared weights

< 1 sec

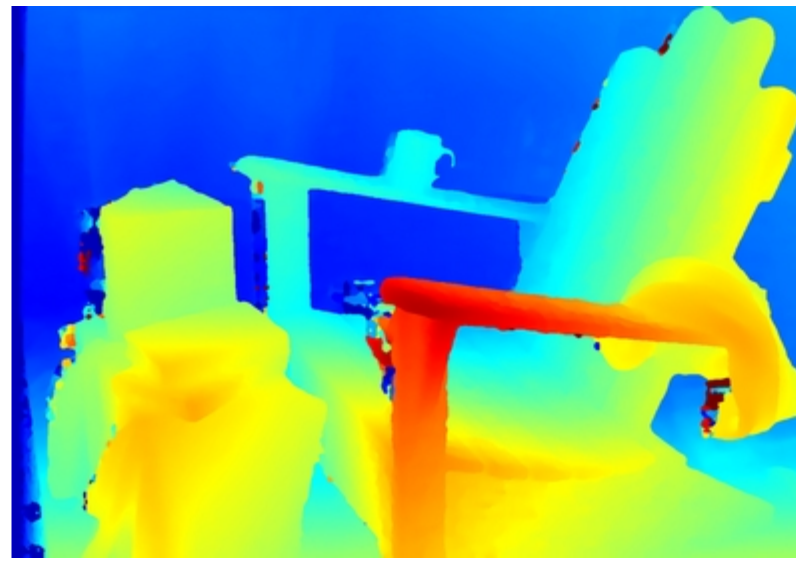
Matching cost with CNN (MC-CNN) accurate



Siamese network with shared weights

≈ 70 sec

- **MC-CNN + adaptive aggregation* + SGM = top performance**
- **The whole system is not *end-to-end* (SGM, cost aggregation)**
- **Most top performing stereo methods now rely on MC-CNN**



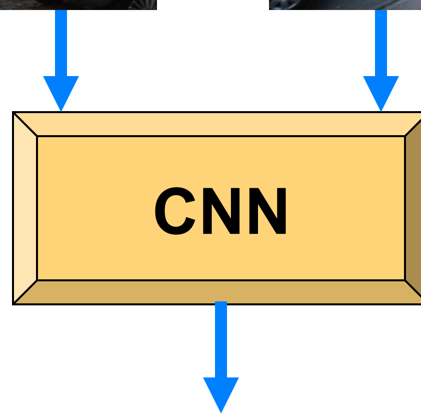
MC-CNN + aggregation + SGM

* K. Zhang, J. Lu, G. Lafruit, "Cross-based Local Stereo Matching Using Orthogonal Integral Images", IEEE Trans. Cir. and Sys. for Video Technol, 2009

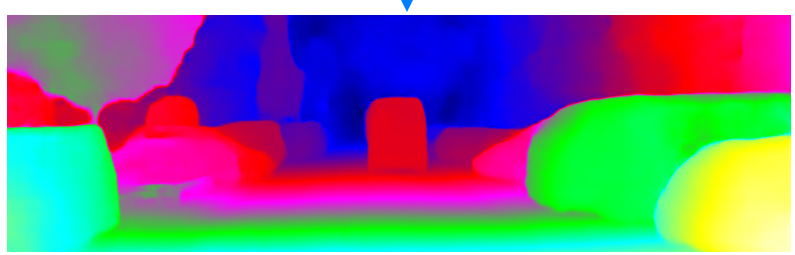
End-to-end stereo: DispNet

A further step forward consists in learning to compute a disparity map from a stereo pair

There isn't a *conventional* stereo algorithm here



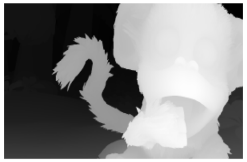
Training phase:
1 image is 1 sample



N. Mayer, E. Ilg P. Häusser and P. Fischer, D. Cremers and A. Dosovitskiy and T. Brox, "A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation", CVPR 2016

DispNet: training and fine tuning

Currently* there aren't datasets with thousands images for training



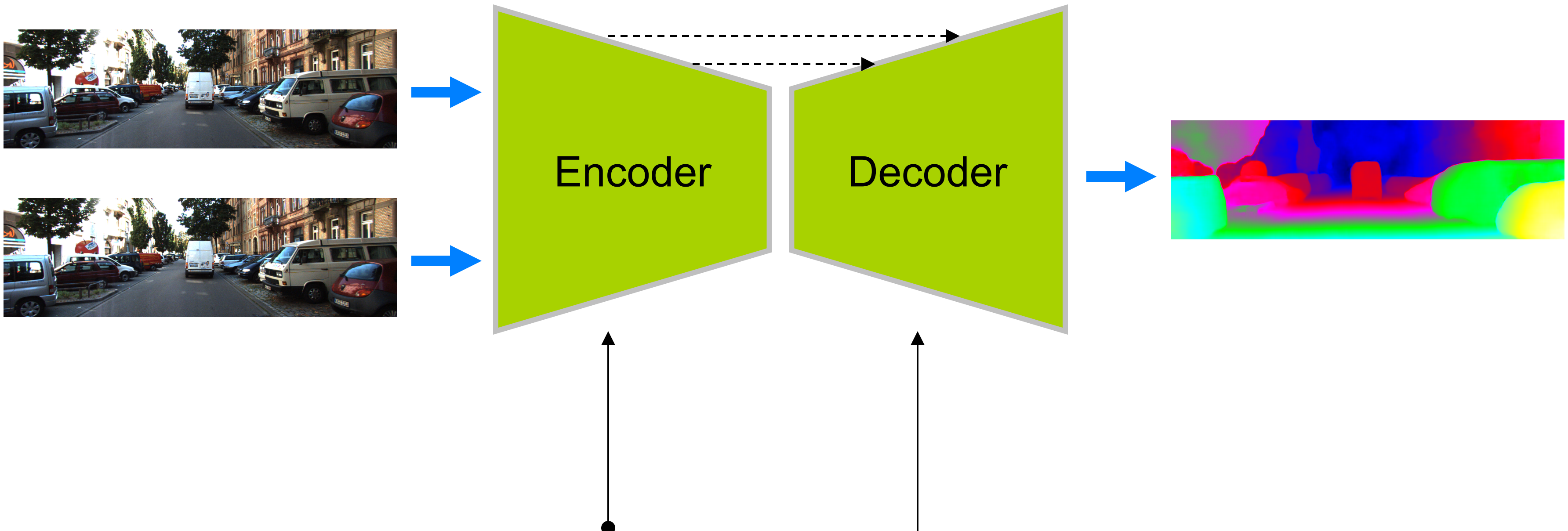
The authors tackled this problem training the deep-network on large synthetic datasets and then fine tuned it on KITTI

Nevertheless, the training dataset is a major issue

Good performance but MCC-CNN + adaptive + SGM performs better

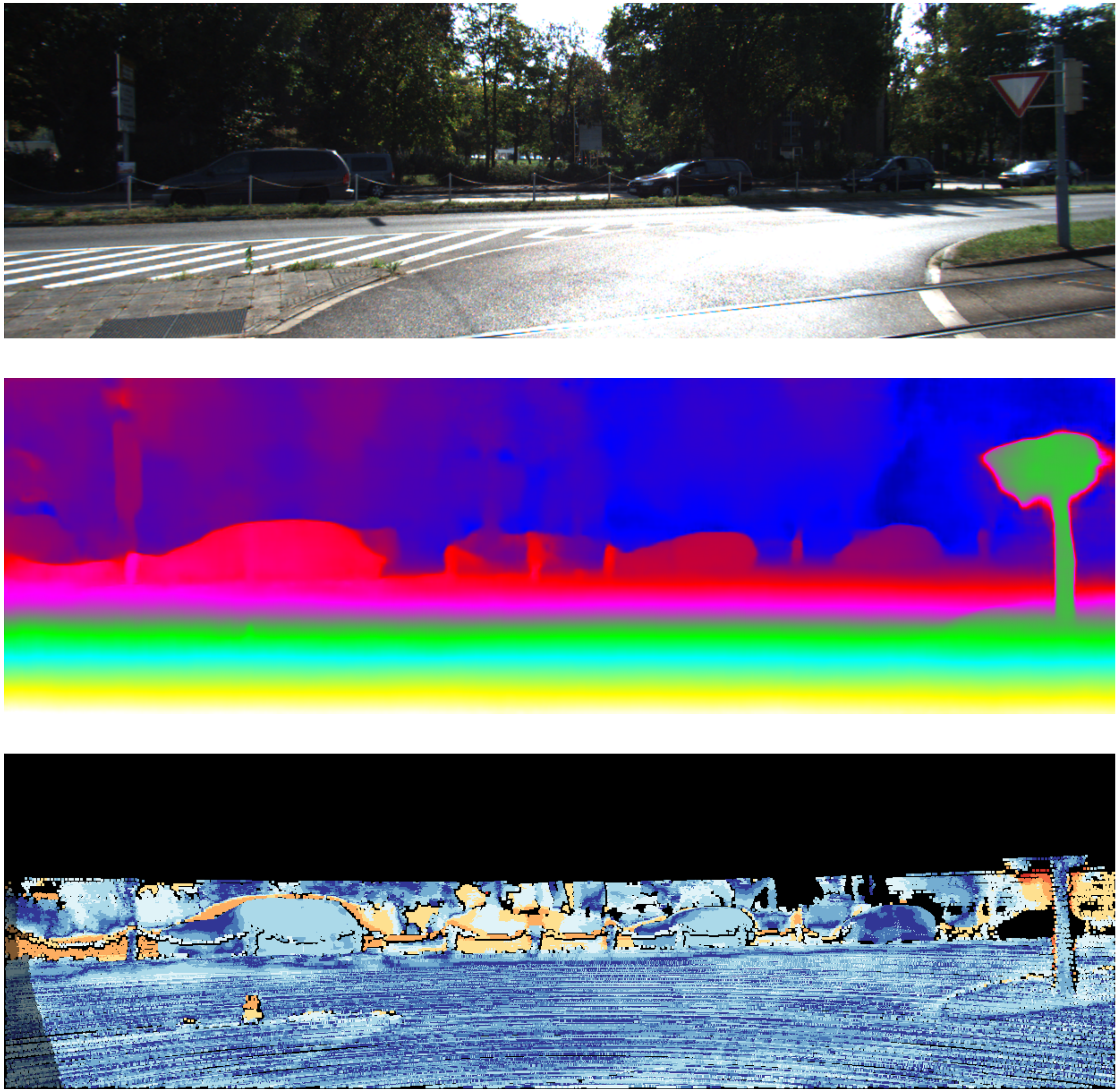
Very fast: 0.06 sec on a Titan X

DispNet: architecture



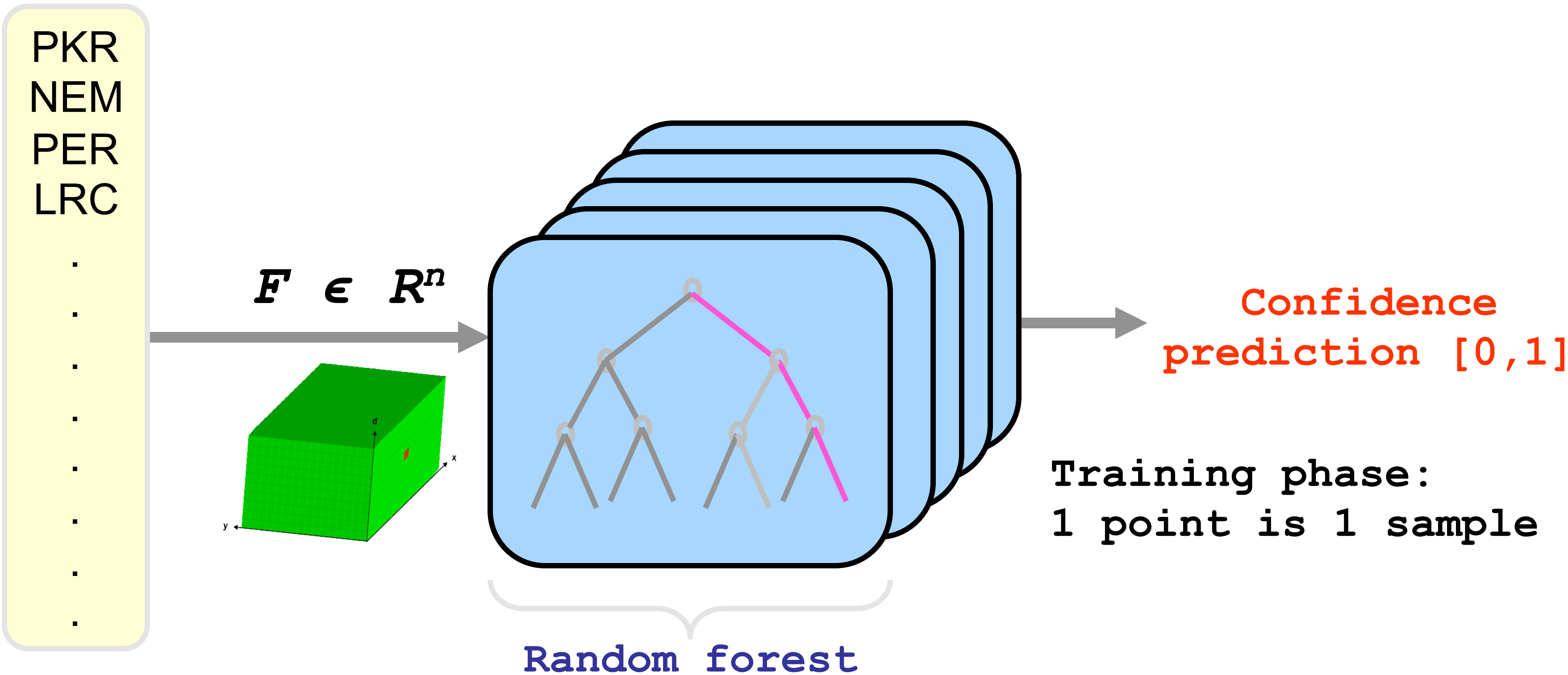
Contracting part:
convolutions

Expanding part: up-convolutions
concatenated with features from
the contracting part



Confidence prediction and machine-learning

- In [Ensemble] a pool of confidence measures is fed to a random forest trained to obtain a *better confidence*



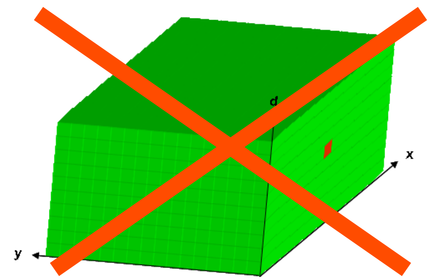
- The features are 23 conventional confidence measures
- The *ensemble* is a much more effective than each confidence measure

[Ensemble] R. Haeusler, R. Nair, and D. Kondermann, “Ensemble learning for confidence measures in stereo vision”, CVPR 2013

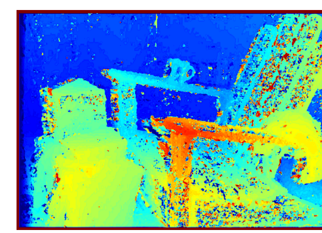
GCP and LEV confidence measures

- **Ensemble was improved by [GCP] using 8 better features and then by [LEV] using 22 even better features**
- **Same strategy for the three methods (Random Forests)**
- **However, they rely on features extracted from the DSI: it is not always available (e.g., RealSense)**

Intel RealSense

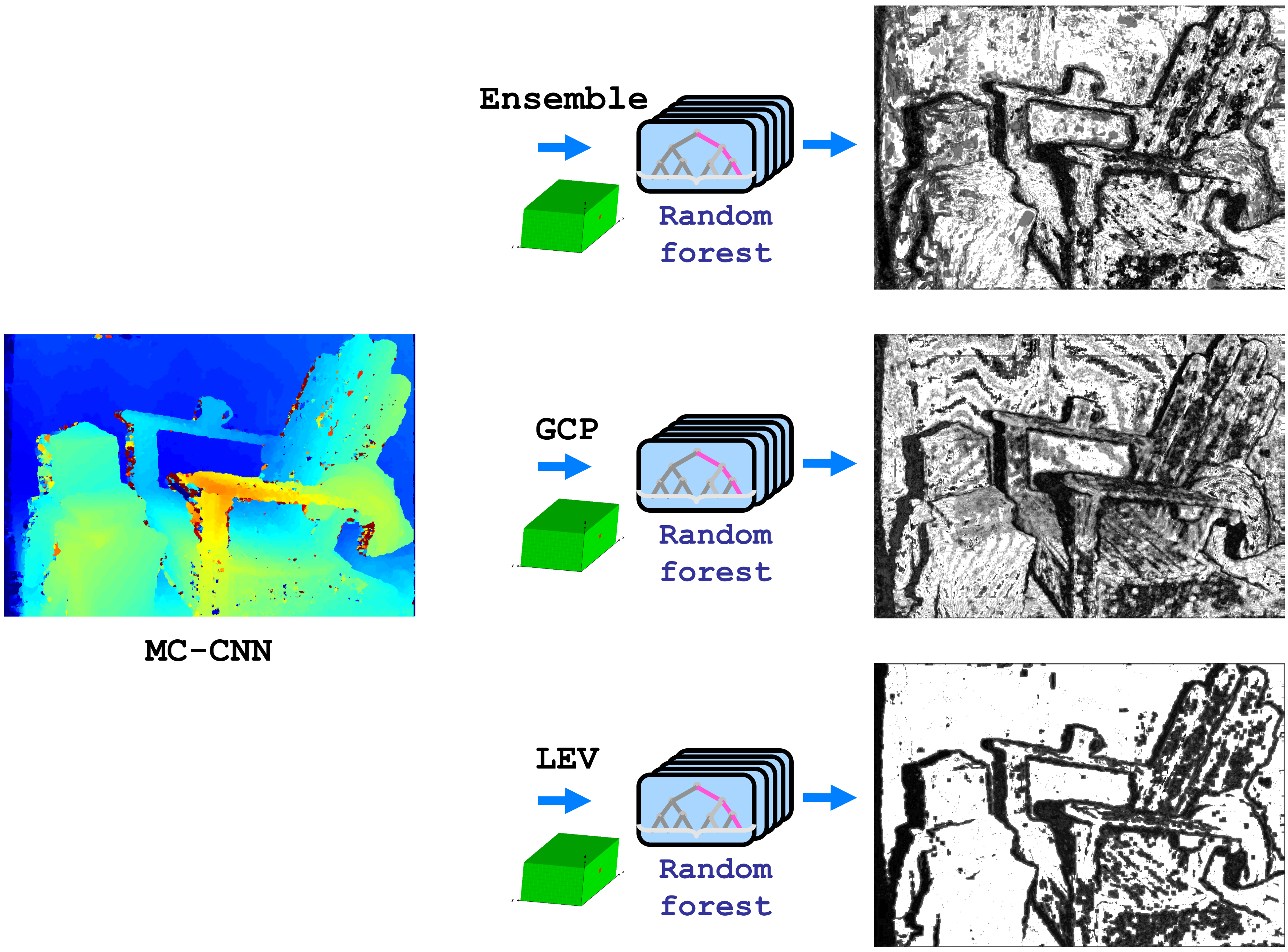


PKR=?
NEM=?
PER=?
LRC=?
?
?
?
?



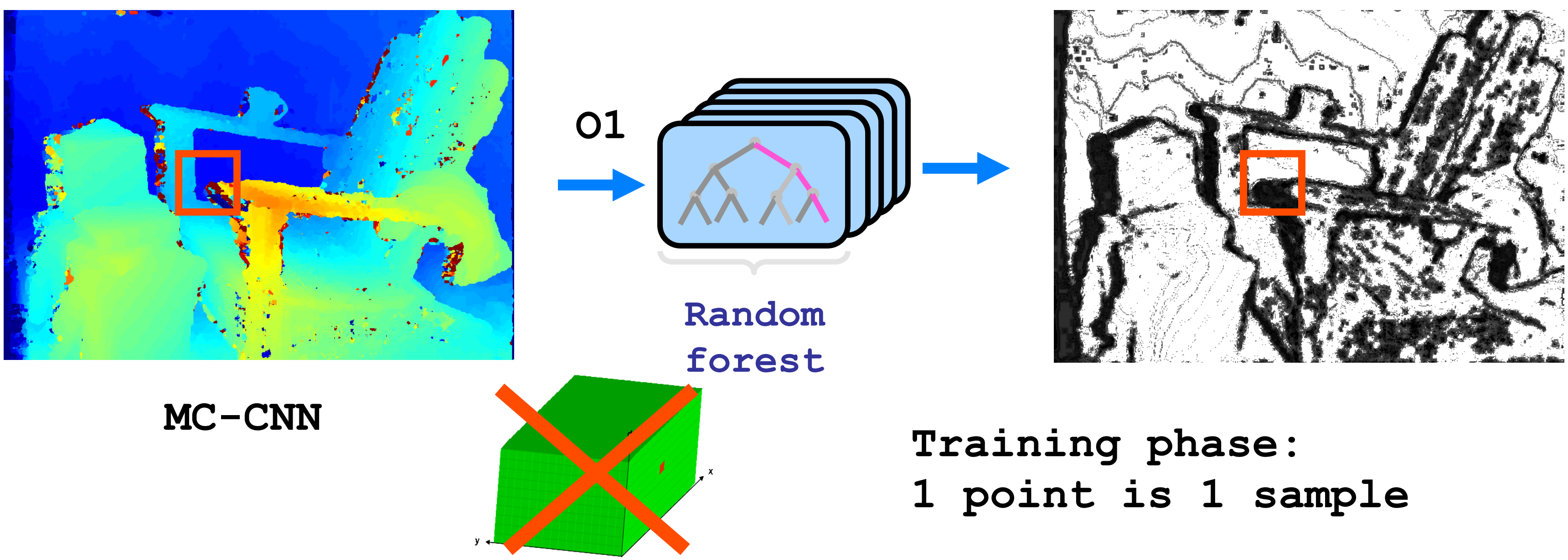
[GCP] Spyropoulos, Komodakis, Mordohai, "Learning to detect ground control points for improving the accuracy of stereo matching", CVPR 2014.

[LEV] Park and Yoon, "Leveraging stereo matching with learning-based confidence measures", CVPR 2015


Ensemble
Random forest
GCP
Random forest
LEV
Random forest
MC-CNN

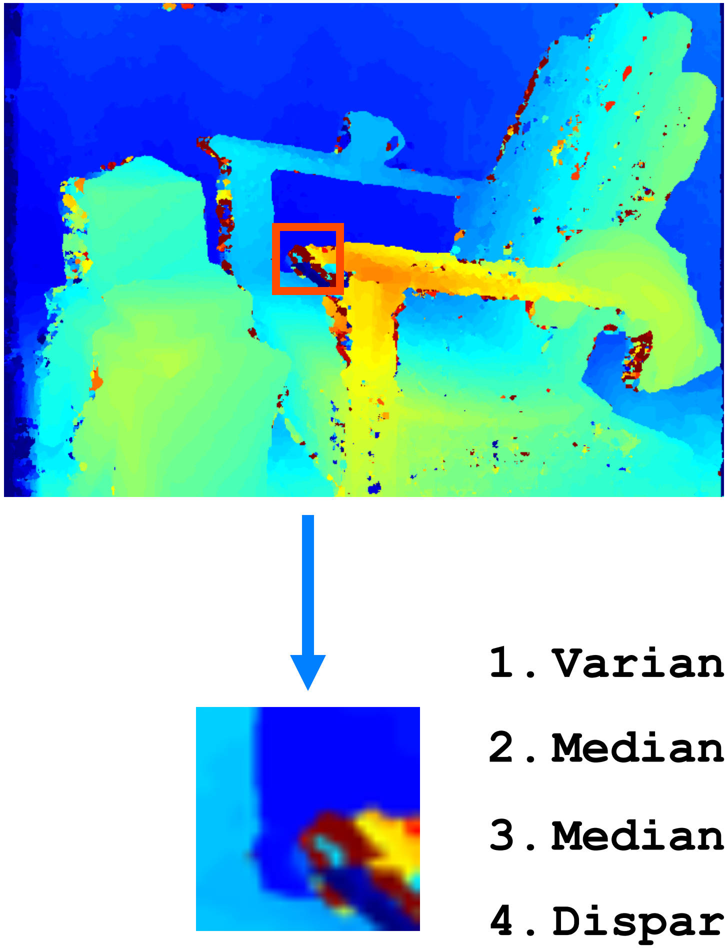
O1 confidence measure

- Aims at removing previous issue concerned with DSI
- 5 features at 4 scales from the disparity map
- Each feature is computed in constant time (O1)
- Same strategy of previous methods (Random-Forest)
- Outperforms state-of-the-art [Ensemble, GCP and LEV]



[O1] Poggi and Mattoccia, “Learning a general-purpose confidence measure based on O(1) features and a smarter aggregation strategy for semi global matching”, 3DV 2016

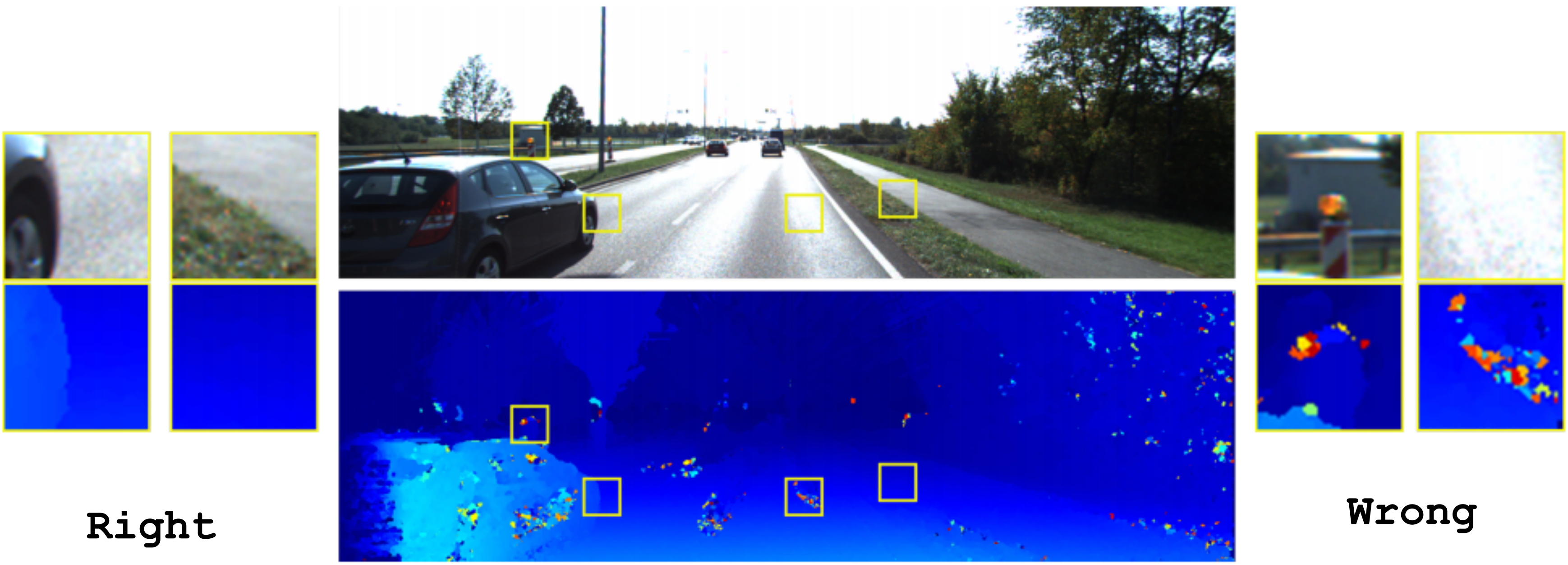
- The features encode the *local behaviour* of the disparity map



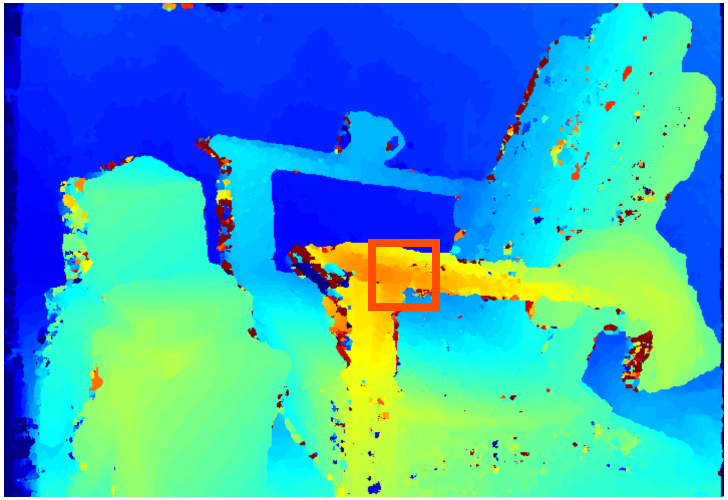
1. Variance (VAR)
2. Median (MED)
3. Median deviation (MDD)
4. Disparity agreement (DA)
5. Disparity scattering (DS)

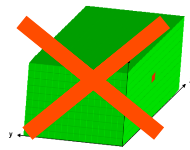
Learning from scratch a confidence measure

- End-to-end learning of a confidence measure [CCNN]
- As for O1, the network is fed only with a disparity map
- The DSI is not required



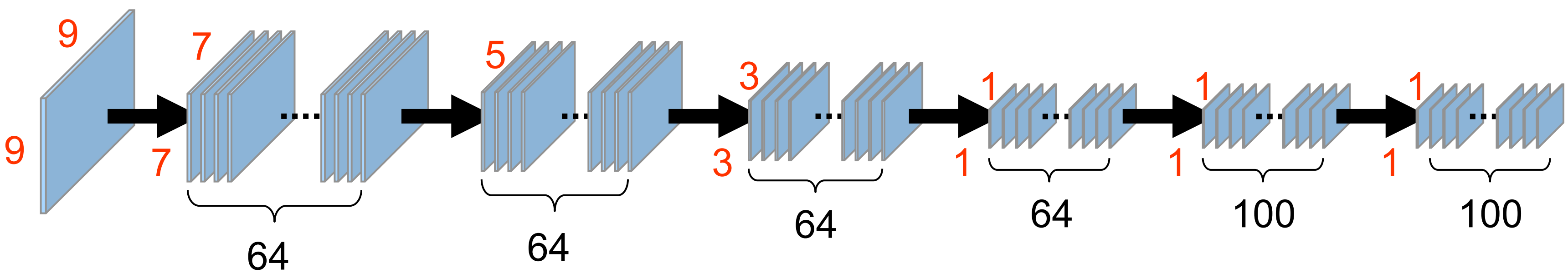
[CCNN] Poggi and Mattoccia, “Learning from scratch a confidence measure”, BMVC 2016





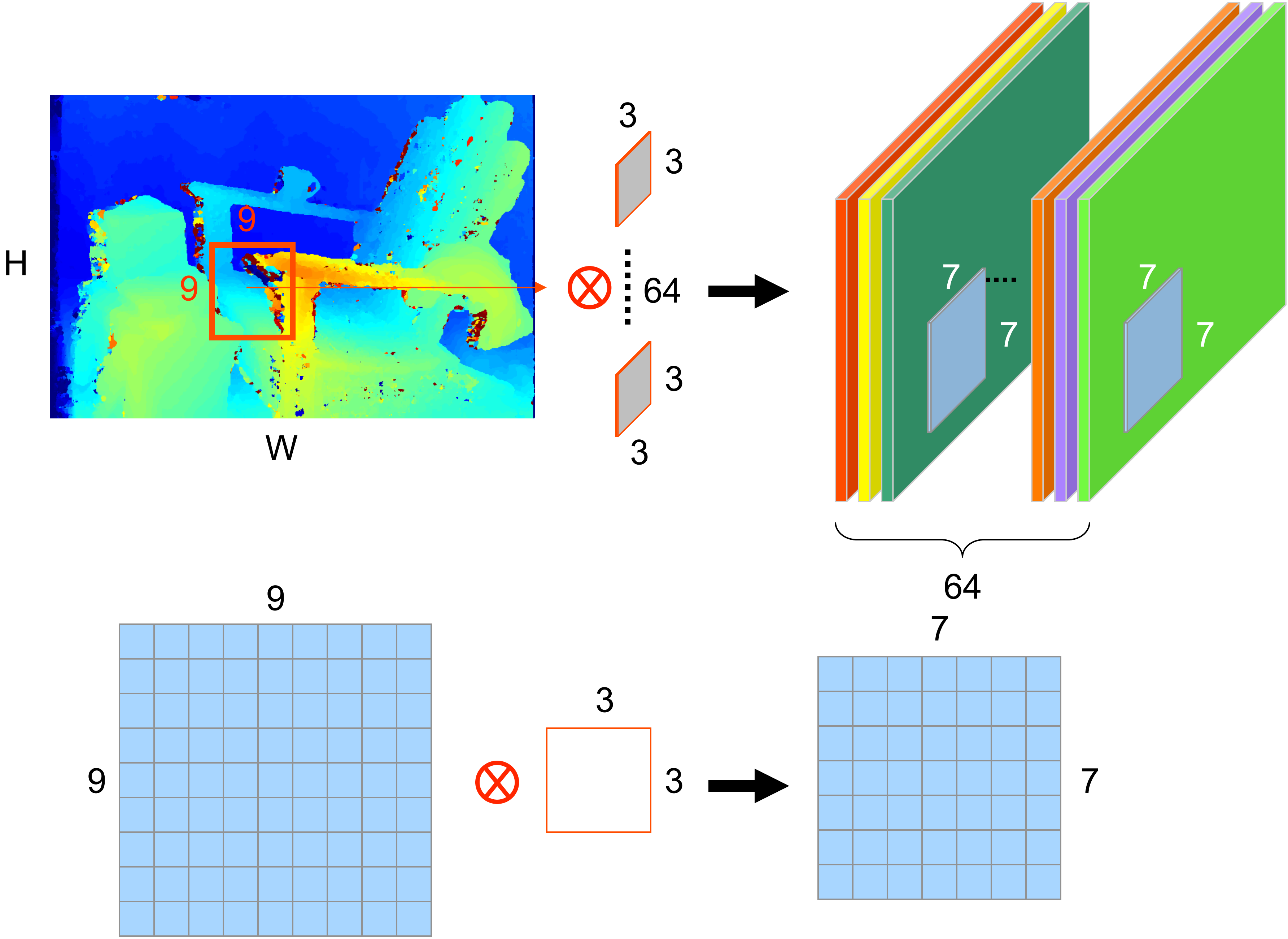
Training phase:
1 point is 1 sample

≈ 0.1 sec (Titan)

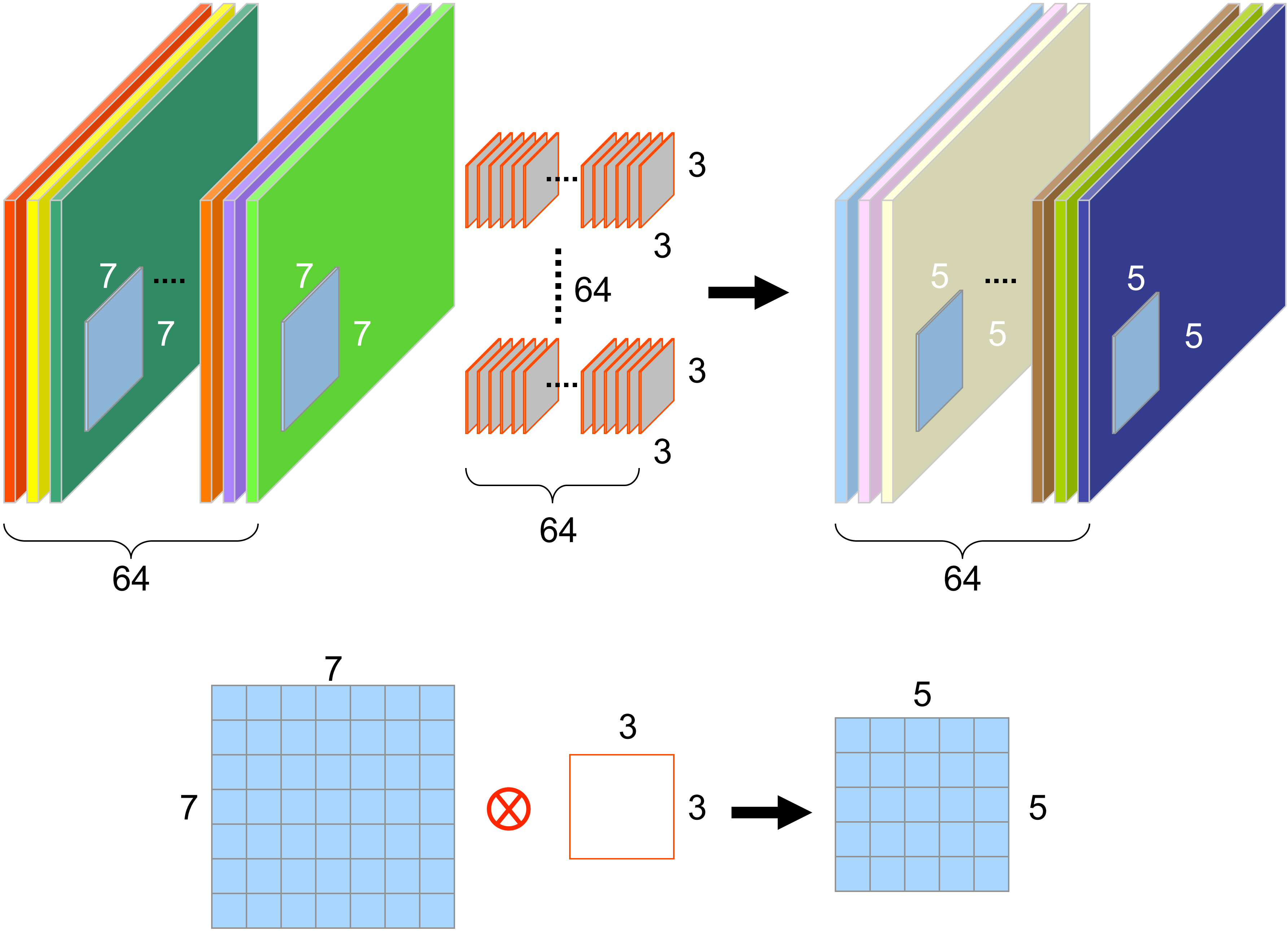


6 layers

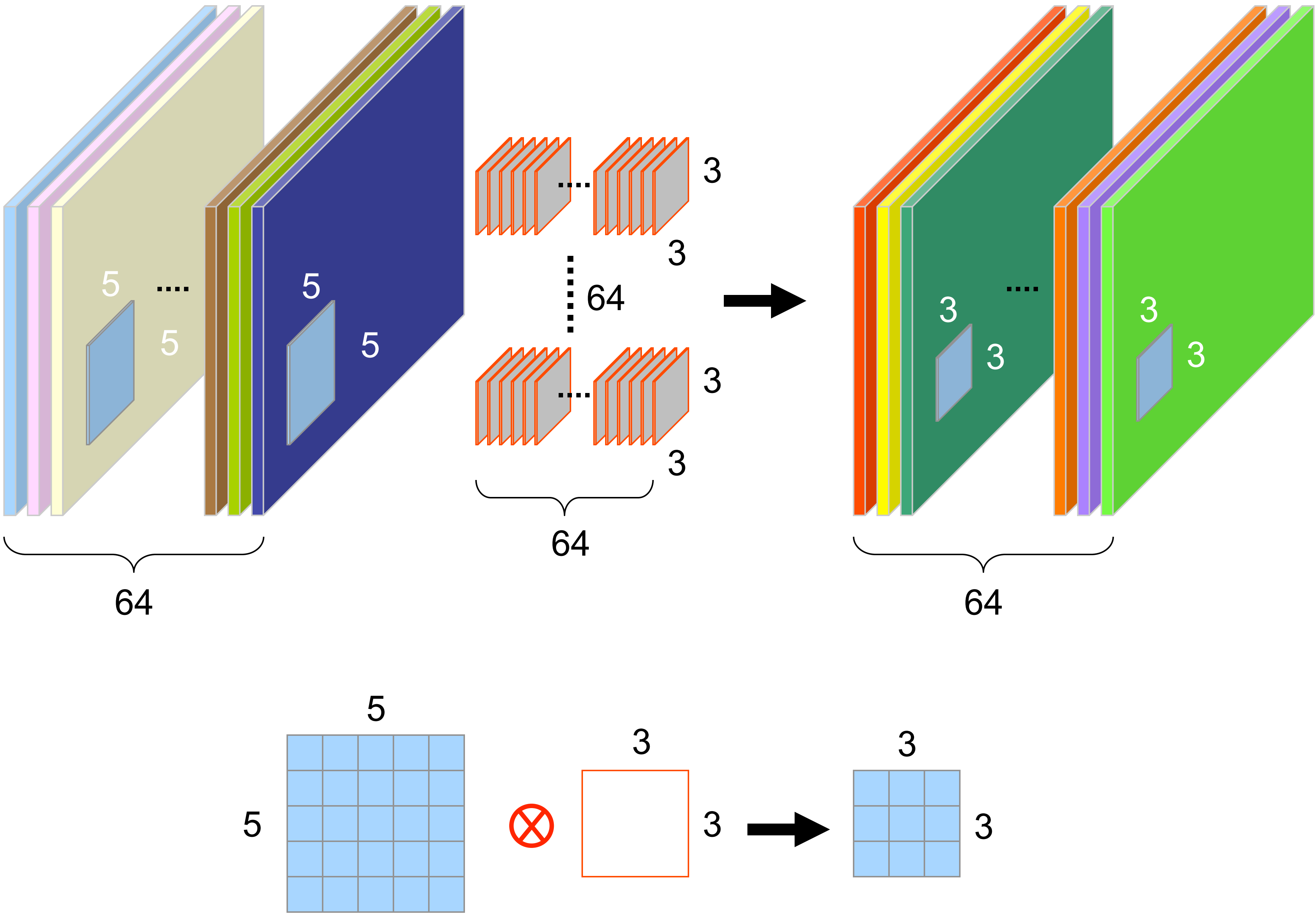
Layer 1: conv + ReLU



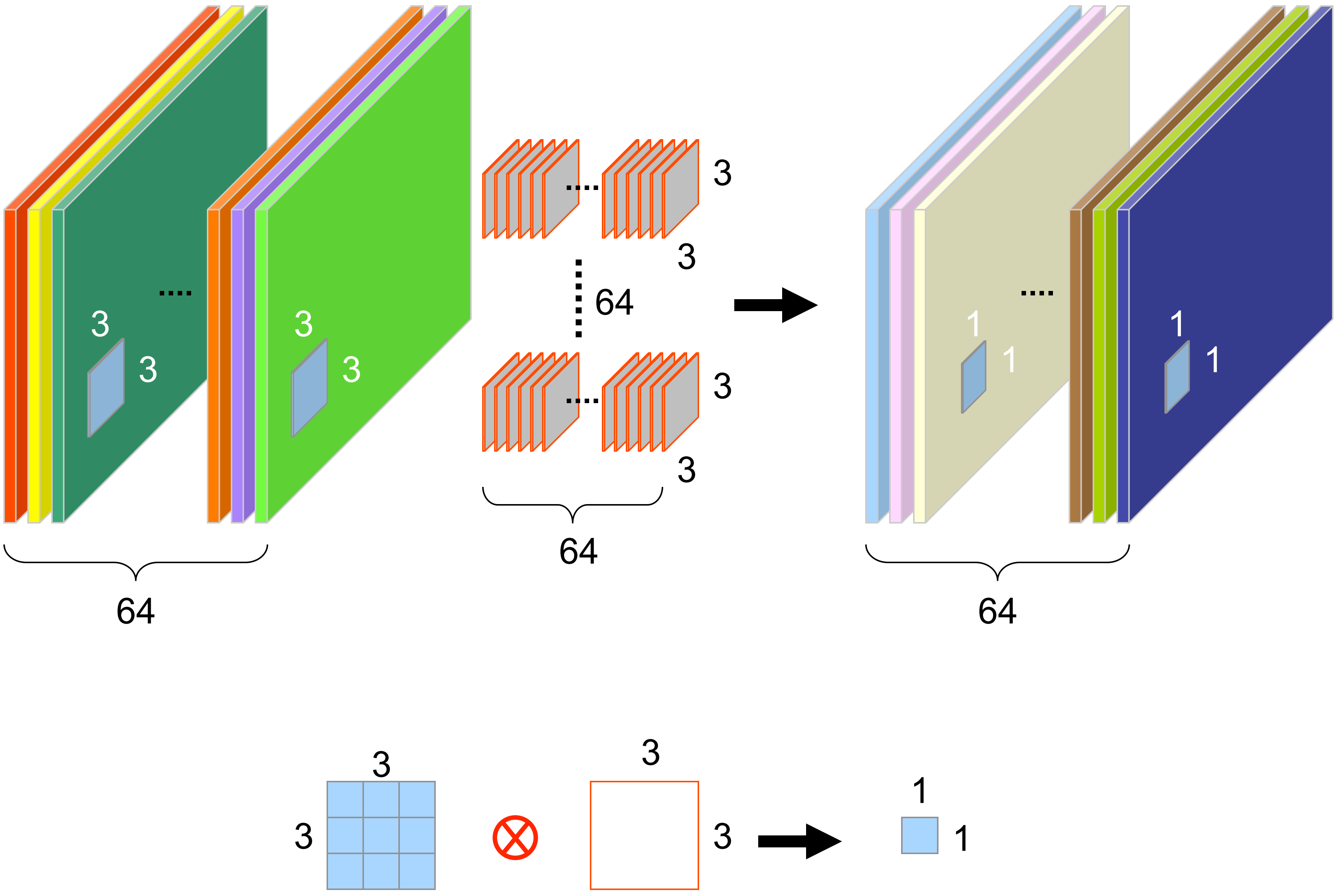
Layer 2: conv + ReLU



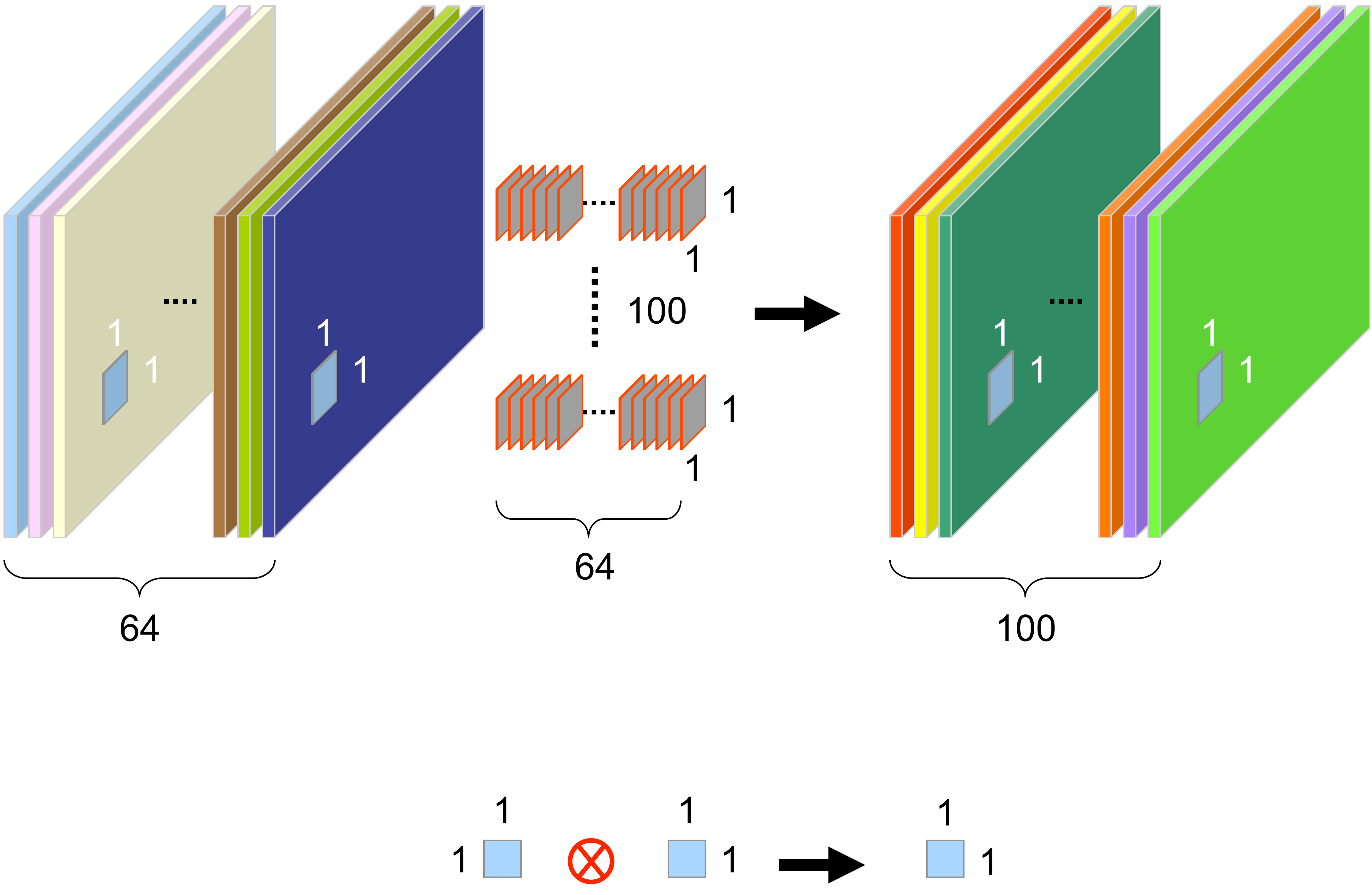
Layer 3: conv + ReLU



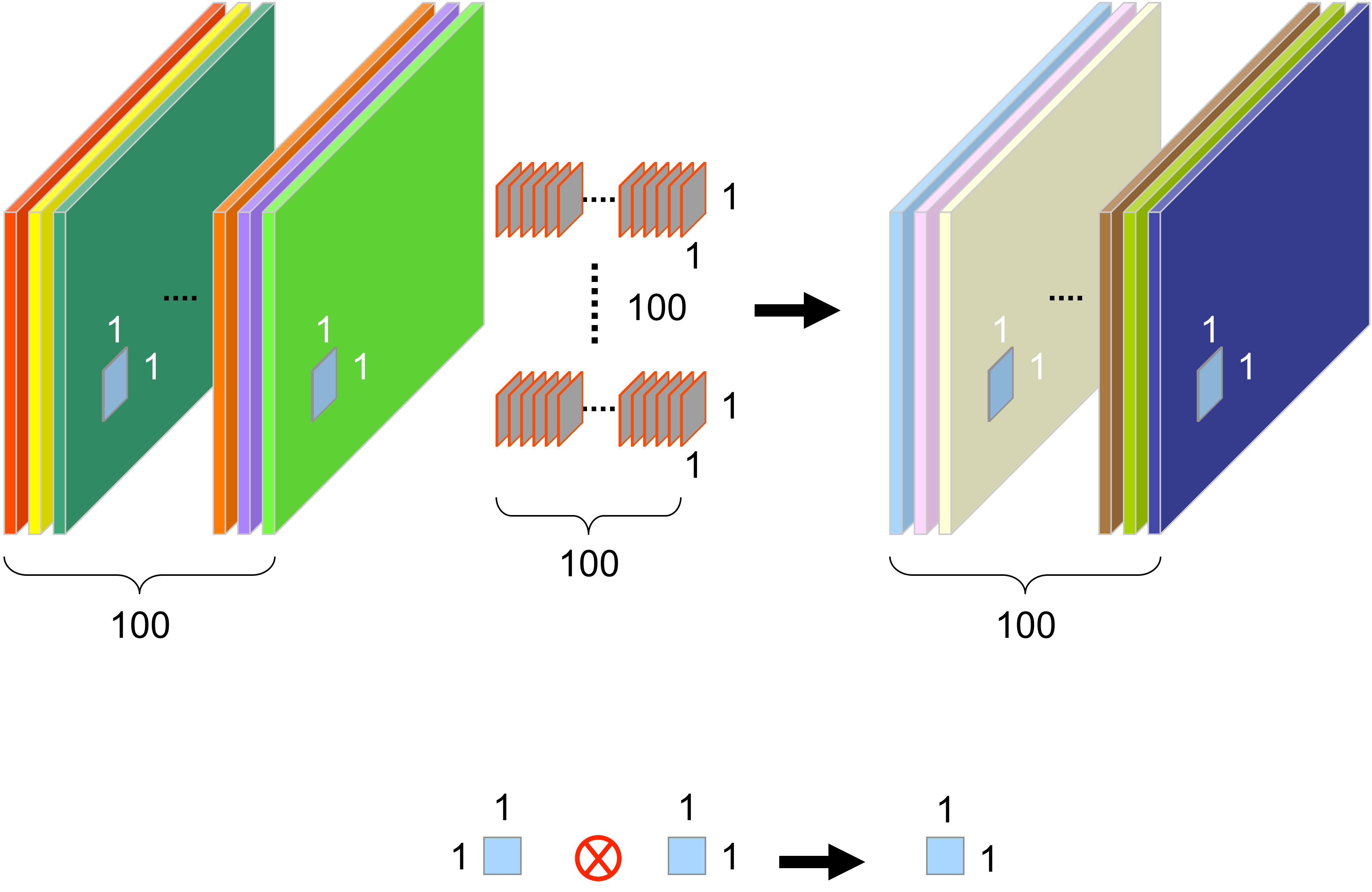
Layer 4: conv + ReLU



Layer 5: Fully connected

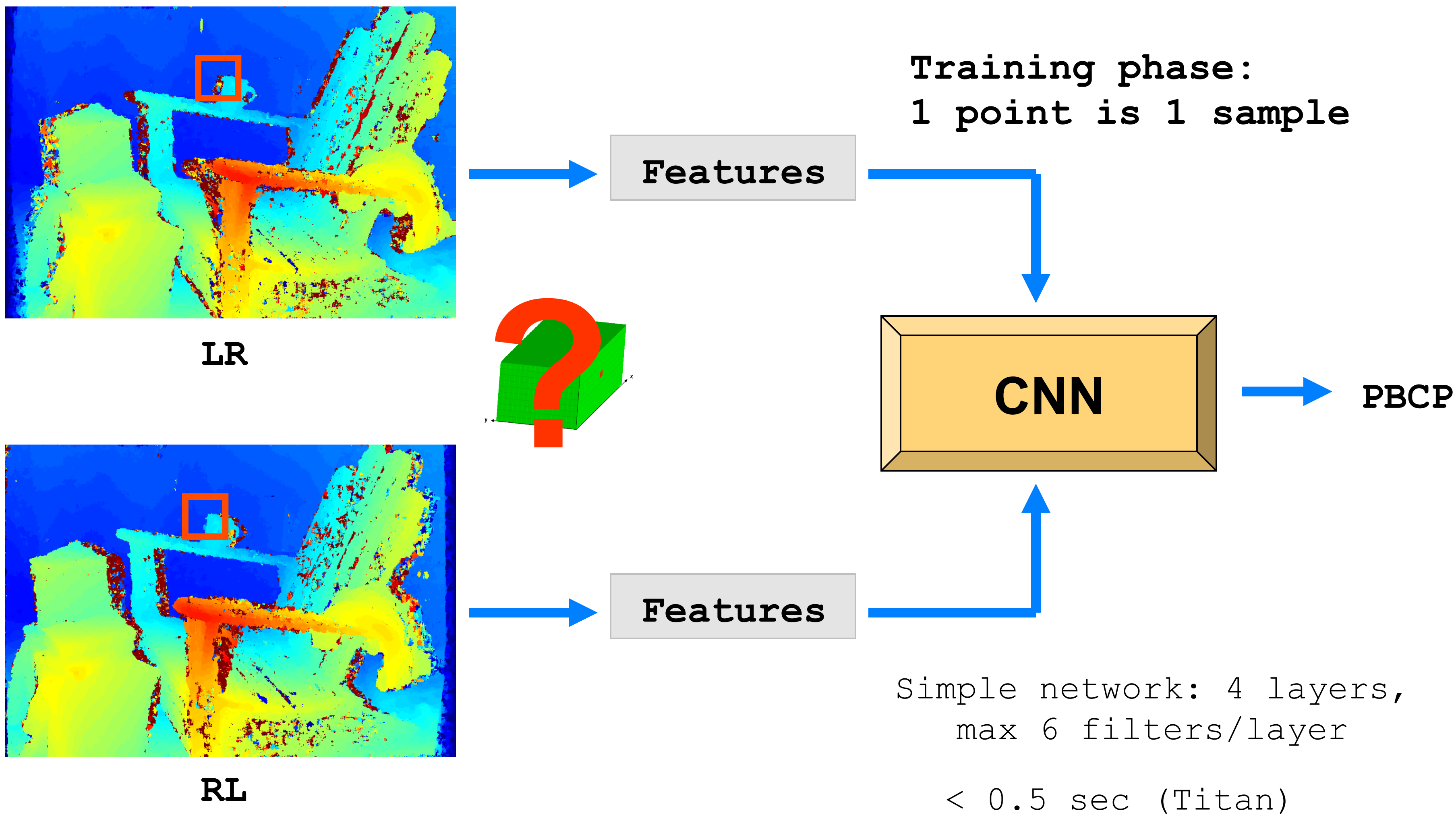


Layer 6: Fully connected

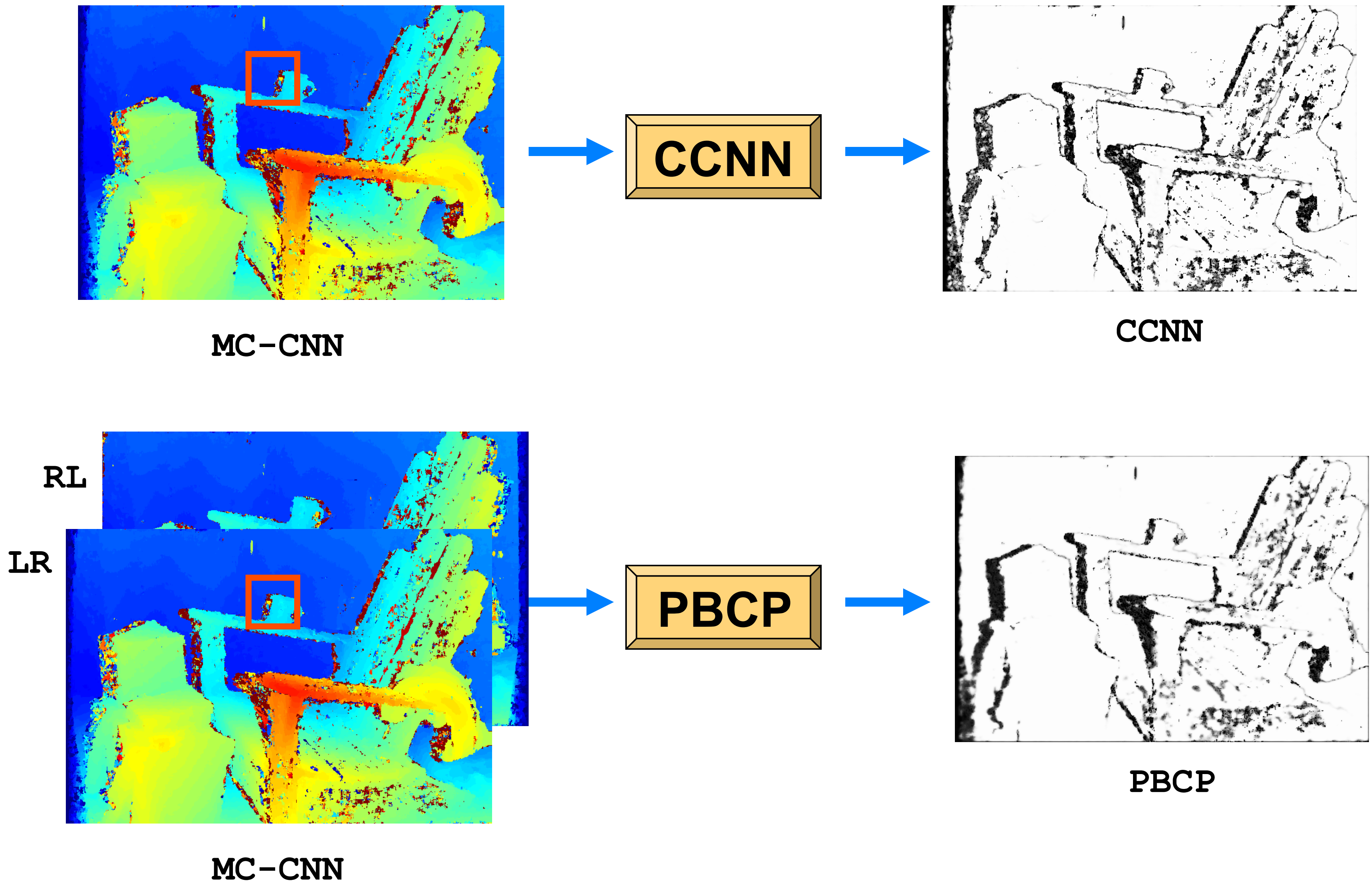


Patch Based Confidence Prediction

- At the same conference was proposed a similar strategy [PBCP] based on hand-crafted features from LR and RL maps



[PBCP] Seki and Pollefeys, Patch based confidence prediction for dense disparity map, BMVC 2016


MC-CNN
CCNN
CCNN
RL
LR
MC-CNN
PBCP
PBCP

Evaluation of confidence measure

- State-of-the-art confidence measures have been exhaustively evaluated in [CONF]
- What is the outcome of this evaluation?
  - ML-based are more effective wrt conventional ones
  - Most effective: those not using the DSI* [CCNN,O1,PBCP]
  - CCNN always outperforms any other method
  - Training is an additional issue but [CCNN,PBCP and O1] generalize quite well to new data

1. Can we train confidence measures without GT data?
2. Can we adapt stereo algorithms to new environments without GT data?

[CONF] Poggi, Tosi, Mattoccia, "Quantitative evaluation of confidence measures in a machine learning world", ICCV 2017

Census (a)

	K12 ($\varepsilon = 38.82\%$)			K15 ($\varepsilon = 35.41\%$)			M14 ($\varepsilon = 37.78\%$)		
Category	measure	rank	AUC	measure	rank	AUC	measure	rank	AUC
3.1	$APKR_{11}$	4^{12}	0.1806	$APKR_{11}$	4^{12}	0.1541	$APKR_{11}$	4^{7}	0.1355
3.2	WMNN	7^{34}	0.2215	WMN	7^{34}	0.2024	WMN	6^{23}	0.1579
3.3	LRD	5^{20}	0.1946	LRD	6^{28}	0.1825	LRD	5^{21}	0.1519
3.4	DA_{11}	3^{8}	0.1668	DA_{11}	3^{7}	0.1399	DA_{11}	3^{4}	0.1294
3.5	DB	8^{65}	0.3446	DB	8^{66}	0.3103	DLB	8^{69}	0.3333
3.6	SAMM	6^{25}	0.2030	SAMM	5^{20}	0.1715	DSM	7^{40}	0.1798
3.7.1	O1	2^{3}	0.1309	O1	2^{3}	0.1128	O1	2^{3}	0.1211
3.7.2	**CCNN**	$\mathbf{1^{1}}$	**0.1223**	**CCNN**	$\mathbf{1^{1}}$	**0.1041**	**CCNN**	$\mathbf{1^{1}}$	**0.1128**
Optimal			0.1067			0.0884			0.0899

Categories 3.7.1 and 3.7.2			
Measure	K12	K15	M14
ENS_c	7	11	44
ENS_r	5	5	33
GCP	6	6	8
LEV	4	4	5
O1	3	3	3
PBCP	2	2	2
CCNN	**1**	**1**	**1**

(b)

MC-CNN (c)

	K12 ($\varepsilon = 17.10\%$)			K15 ($\varepsilon = 15.37\%$)			M14 ($\varepsilon = 26.70\%$)		
Category	measure	rank	AUC	measure	rank	AUC	measure	rank	AUC
3.1	$APKR_{11}$	4^{11}	0.0566	$APKR_{11}$	4^{11}	0.0508	$APKR_{11}$	3^{5}	0.0728
3.2	WMN	6^{30}	0.0748	WMN	6^{31}	0.0654	WMN	4^{13}	0.0763
3.3	LRD	7^{31}	0.0748	LRD	7^{32}	0.0712	UCC	5^{22}	0.0896
3.4	DS_9	3^{8}	0.0542	DS_9	3^{8}	0.0477	DS_{11}	6^{35}	0.1061
3.5	DLB	8^{66}	0.1543	HGM	8^{67}	0.1439	DLB	8^{68}	0.2260
3.6	SAMM	5^{16}	0.0598	SAMM	5^{21}	0.0557	DSM	7^{40}	0.1228
3.7.1	O1	2^{2}	0.0317	O1	2^{2}	0.0324	O1	2^{3}	0.0680
3.7.2	**CCNN**	$\mathbf{1^{1}}$	**0.0297**	**CCNN**	$\mathbf{1^{1}}$	**0.0297**	**CCNN**	$\mathbf{1^{1}}$	**0.0637**
Optimal			0.0231			0.0213			0.0459

Categories 3.7.1 and 3.7.2			
Measure	K12	K15	M14
ENS_c	7	7	24
ENS_r	5	5	17
GCP	6	6	14
LEV	4	4	4
O1	2	2	3
PBCP	3	3	2
CCNN	**1**	**1**	**1**

(d)

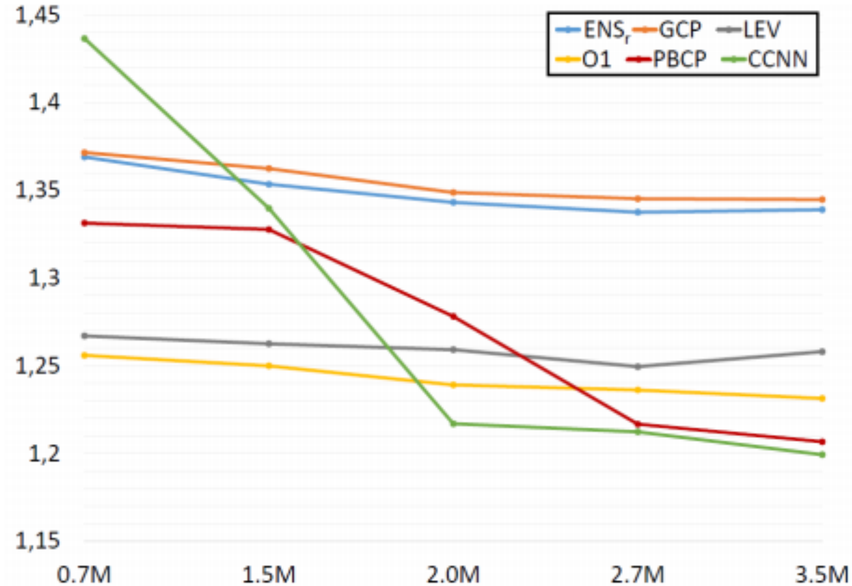
SGM (e)

	K12 ($\varepsilon = 16.78\%$)			K15 ($\varepsilon = 13.68\%$)			M14 ($\varepsilon = 25.91\%$)		
Category	measure	rank	AUC	measure	rank	AUC	measure	rank	AUC
3.1	$APKR_{11}$	3^{7}	0.0492	$APKR_{11}$	3^{7}	0.0457	$APKR_9$	2^{2}	0.0739
3.2	WMN	4^{11}	0.0554	WMN	5^{12}	0.0502	WMN	4^{8}	0.0.779
3.3	UCC	6^{21}	0.0735	UCC	6^{19}	0.0640	UCC	6^{23}	0.0959
3.4	DS_{11}	5^{12}	0.0554	DS_{11}	4^{11}	0.0501	DS_{11}	5^{13}	0.0884
3.5	DB	9^{67}	0.1378	DB	9^{68}	0.1265	DLB	9^{70}	0.2157
3.6	DSM	7^{36}	0.0811	DSM	7^{28}	0.0679	DSM	7^{32}	0.1041
3.7.1	LEV	2^{2}	0.0358	O1	2^{2}	0.0323	O1	3^{6}	0.0777
3.7.2	**CCNN**	$\mathbf{1^{1}}$	**0.0358**	**CCNN**	$\mathbf{1^{1}}$	**0.0302**	**CCNN**	$\mathbf{1^{1}}$	**0.0736**
3.8	SCS	8^{41}	0.0851	SCS	8^{48}	0.0790	SCS	8^{36}	0.1080
Optimal			0.0227			0.0184			0.0431

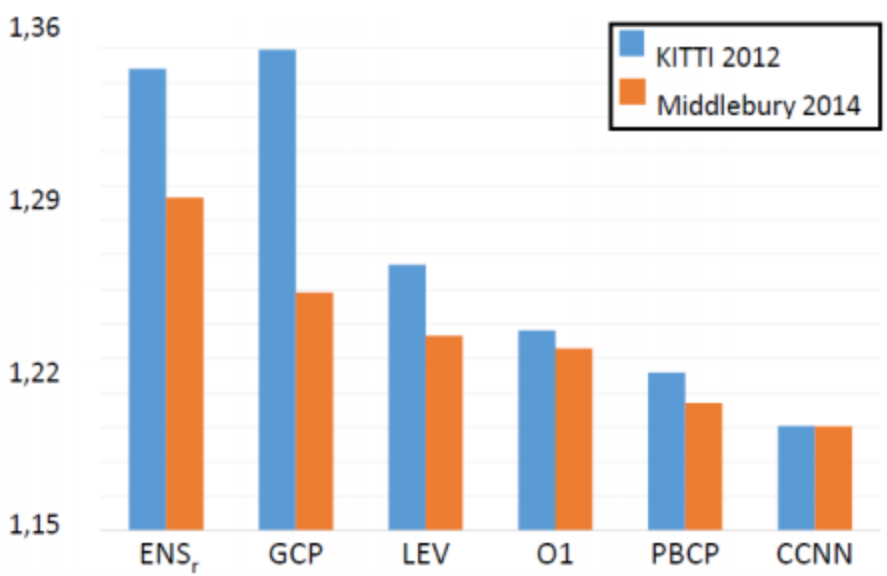
Categories 3.7.1 and 3.7.2			
Measure	K12	K15	M14
ENS_c	27	31	44
ENS_r	5	5	11
GCP	6	6	28
LEV	2	4	19
O1	3	2	6
PBCP	4	3	7
CCNN	**1**	**1**	**1**

(f)

Training on first 20 Kitti 12 (dataset): testing o K12, K15 and M14



Impact of training data:
5, 10, 15, 20, 25 images

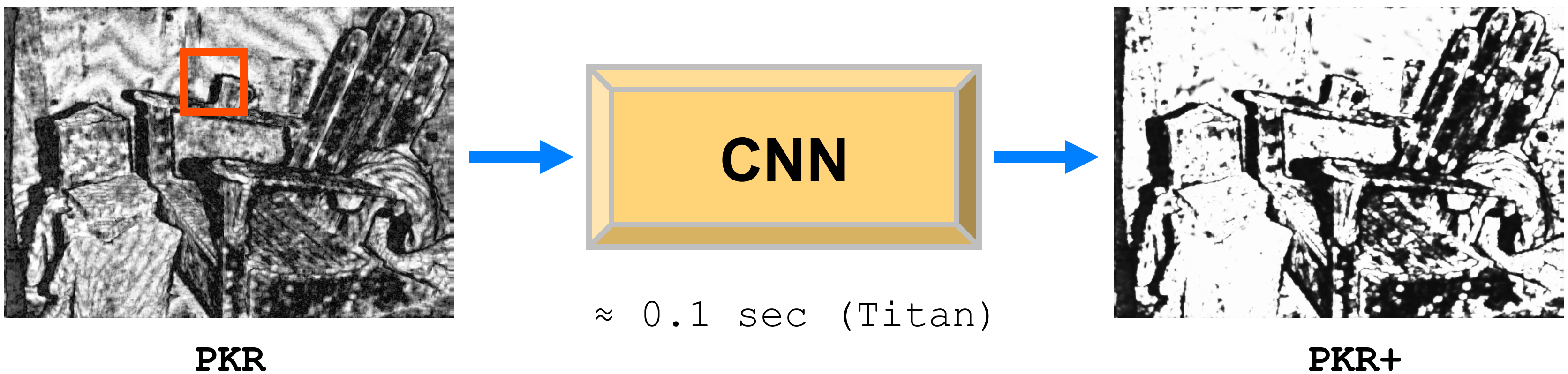


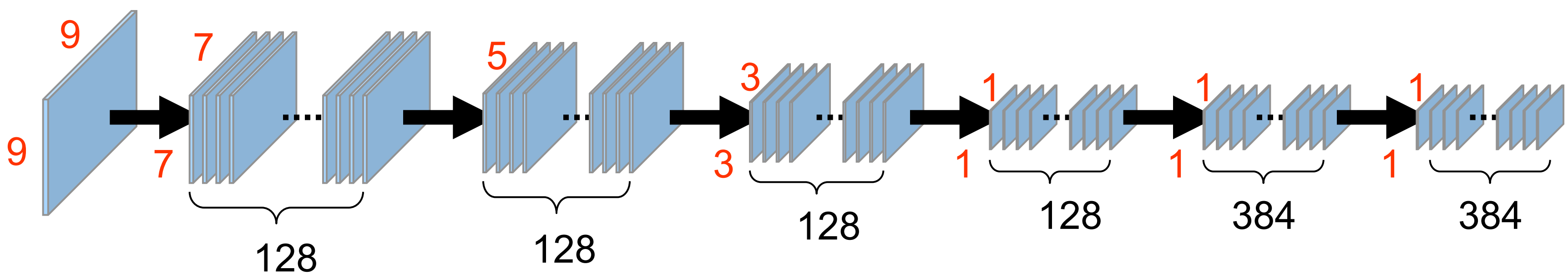
Generalization

$AUC_{KITTI\ (training)}$ / $AUC_{MIDD\ (testing)}$

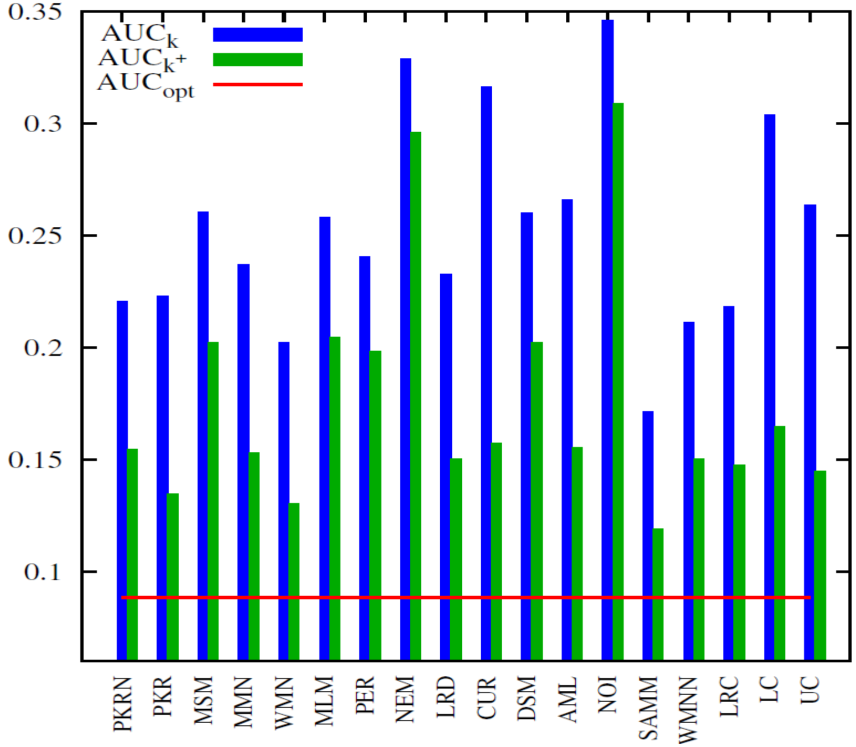
Enforcing local consistency with a CNN

- Given any (conventional or ML-based) confidence measure, a CNN is trained to improve its accuracy by exploiting local consistency [PLUS]
- Always notable improvements in terms of AUC (up to ≈75%)

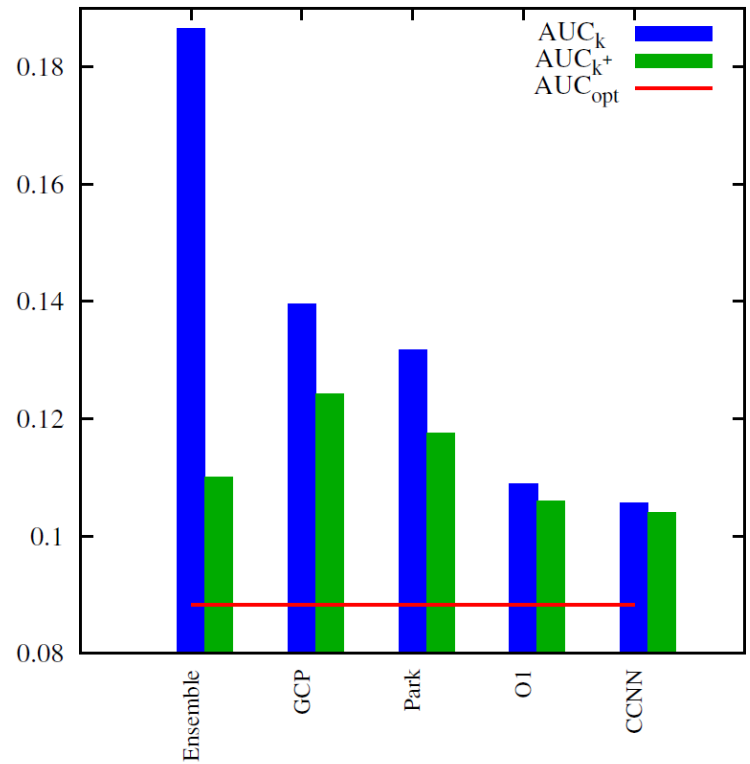




[PLUS] Poggi and Mattoccia, "Learning to predict stereo reliability enforcing local consistency of confidence maps", CVPR 2017



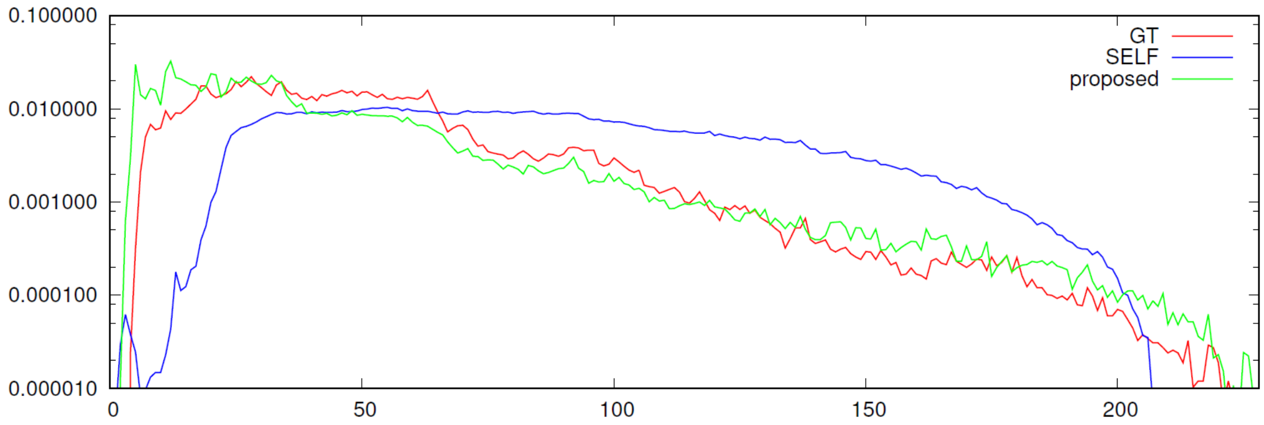
Conventional



ML-based

Unsupervised training of confidence measures

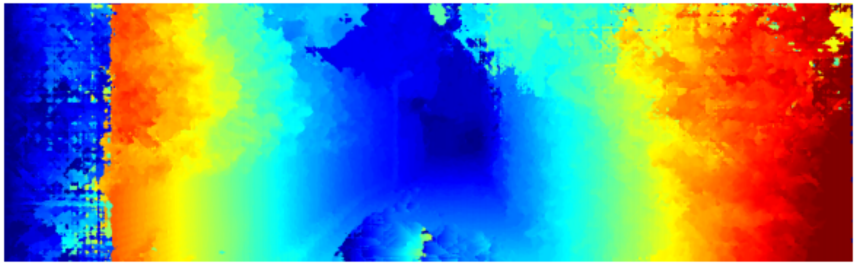
- **Top performing confidence measures rely on ML**
- **Datasets are seldom available**
- **Self-labelling strategy based on a pool of conventional confidence measures [BMVC17]**
- **This method enables to improve state-of-the-art [SELF] without any constraint (i.e., sequences, only ego-motion)**



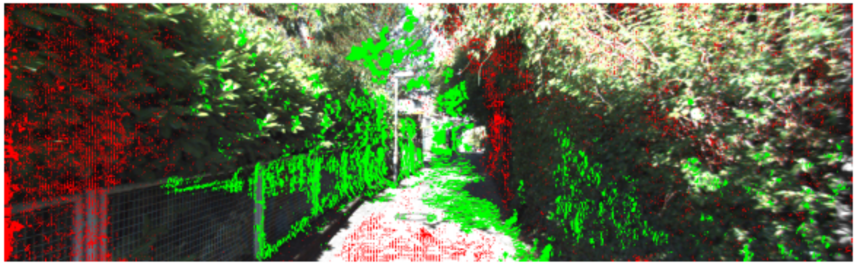
[SELF] Mostegel, Rumpler, Fraundorfer, Bischof, “Using Self- Contradiction to Learn Confidence Measures in Stereo Vision”, CVPR 2016

[BMVC17] Tosi, Poggi, Tonioni, Di Stefano, Mattoccia, “Learning confidence measures in the wild”, BMVC 2017

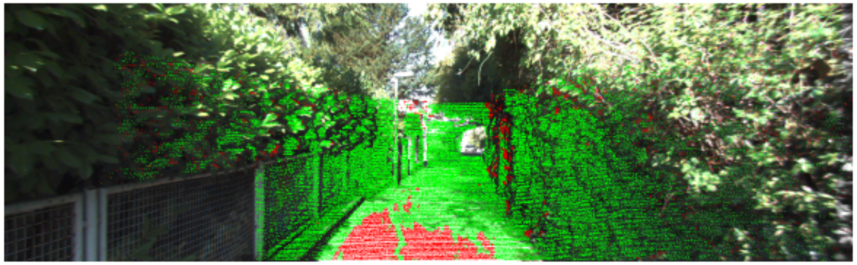
(a)



(b)



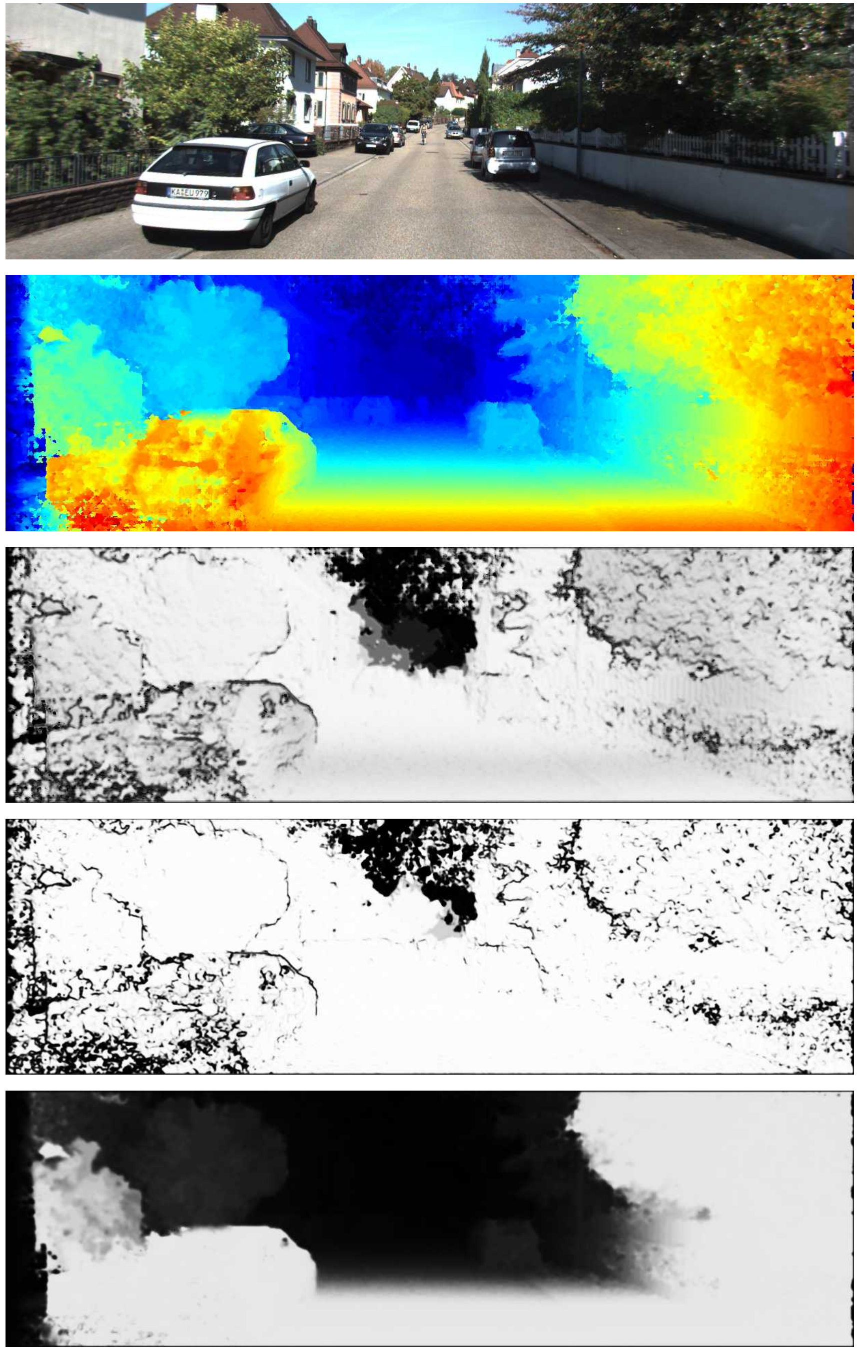
(c)



(d)

Self-supervised [BMVC17]

GT



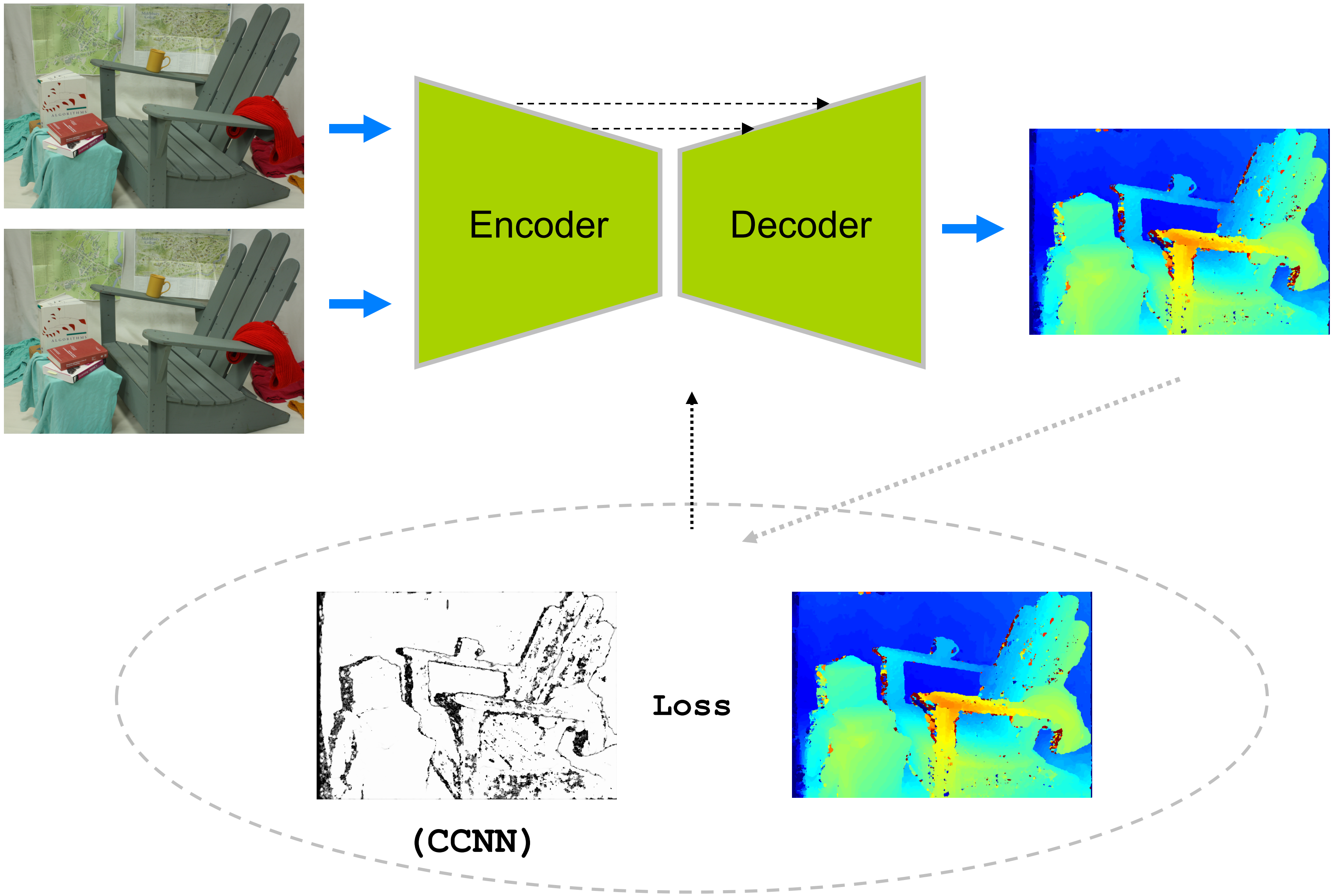
SGM

CCNN with GT

CCNN with [BMVC17]

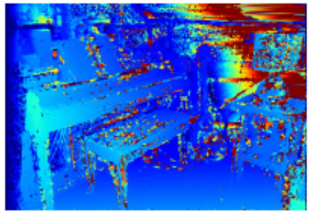
CCNN with [SELF]

Unsupervised adaptation for DispNet

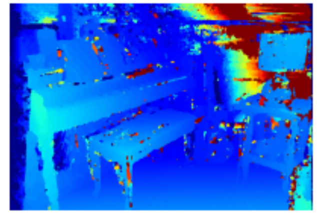


[ADAPT] Tonioni, Poggi, Mattoccia, Di Stefano, “Unsupervised Adaptation for Deep Stereo”, ICCV 2017

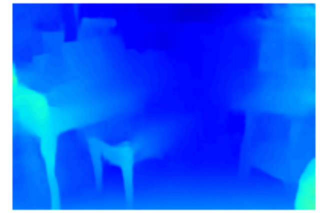
GT



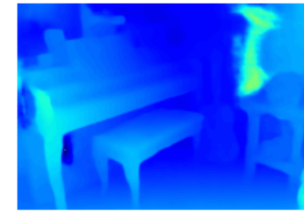
AD-CENSUS (24.89)



SGM (18.08)



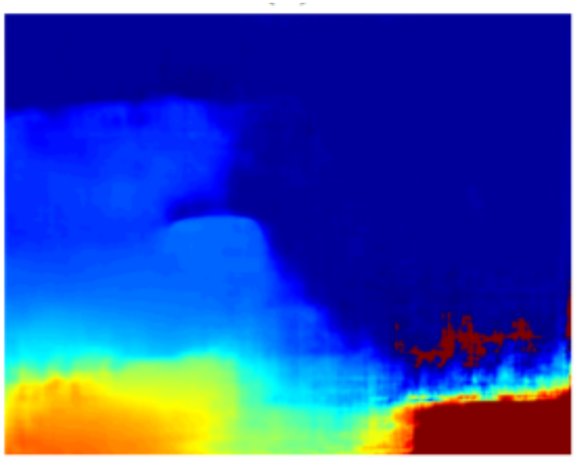
DispNet K12-GT (29.55)



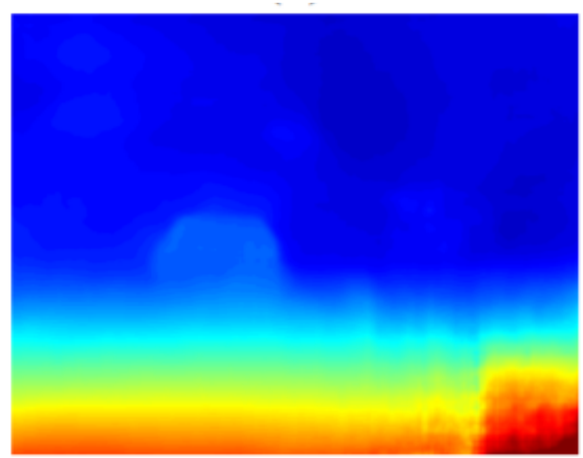
DispNet SGM (15.12)

Left

Right



DispNet

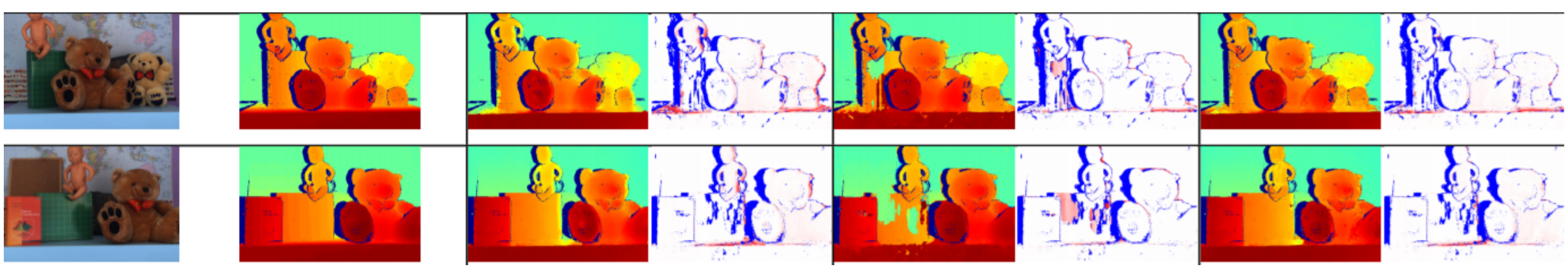


DispNet and [ADAPT]

Sensor fusion

- Confidence measures can be useful for other purposes
- In [FUSION] was proposed a method to combine the depth maps provided by two sensors:
  - Stereo (SGM algorithm)
  - ToF (Mesa)
- Each depth measurements is weighted by its confidence within a local disparity optimization framework
- The resulting disparity combines the strengths of the two sensors

Left GT ToF SGM Fusion

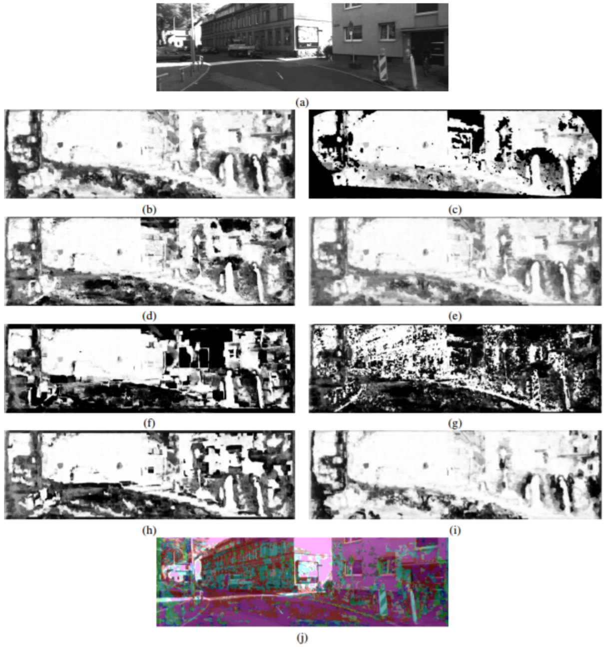


[FUSION] Marin, Zanuttigh, Mattoccia, “Reliable fusion of ToF and stereo depth driven by confidence measures”, ECCV 2016

Disparity fusion

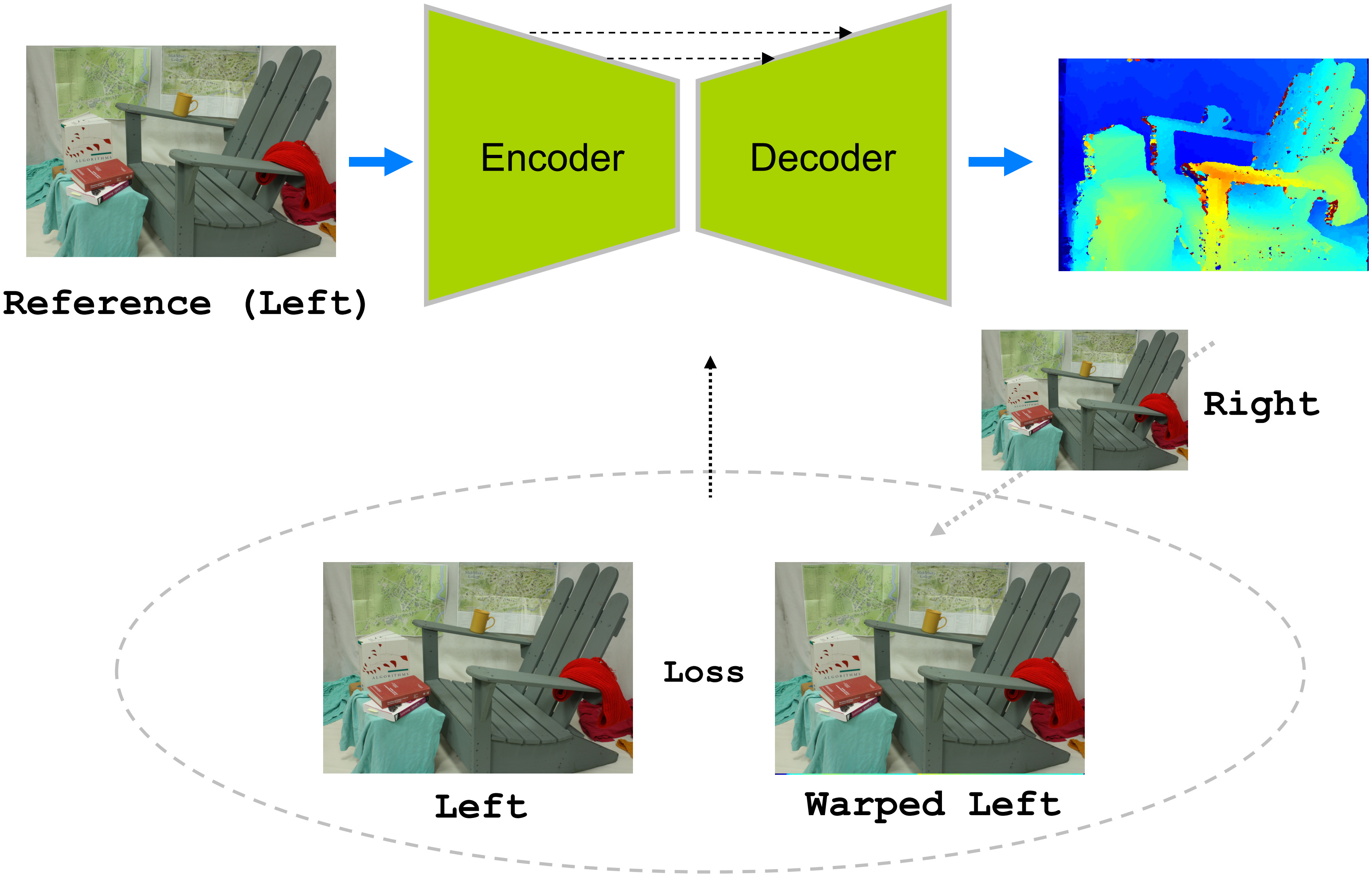
Given N disparity maps:

- CCNN-like architecture to combine multiple disp. maps
- the network selects the most confident disparity
- more effective than a comparable RF-based strategy



M. Poggi, S. Mattoccia, “Deep Stereo Fusion: combining multiple disparity hypotheses with deep-learning”, 3DV 2016

Unsupervised monocular depth* estimation

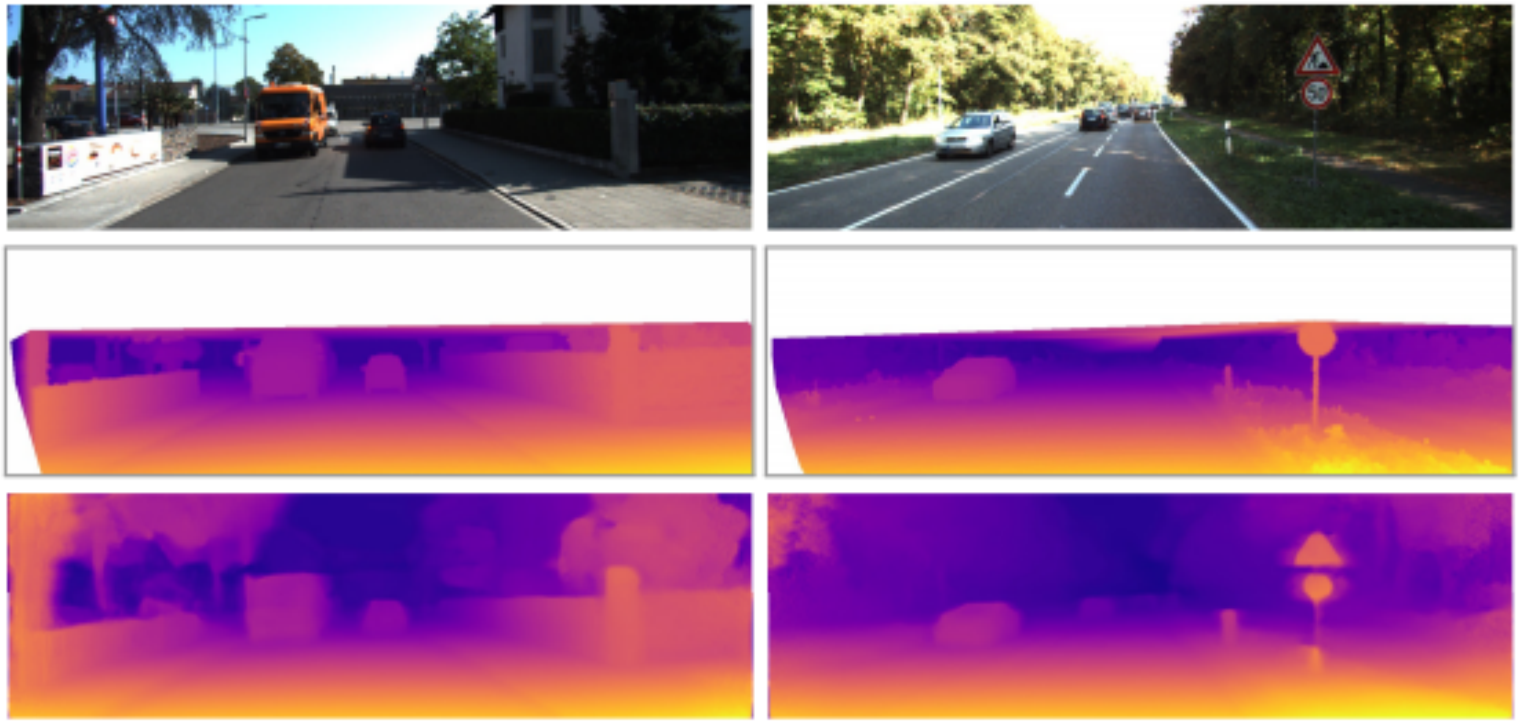


[MONO] Godard, Aodha, Brostow, "Unsupervised Monocular Depth Estimation with Left-Right Consistency", CVPR 2017

Left

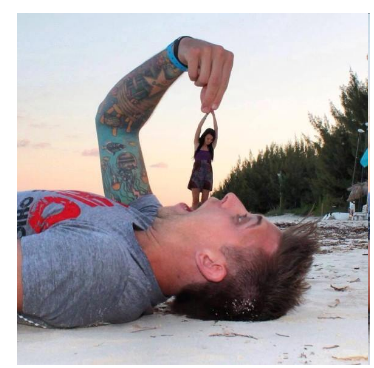
GT

[MONO]



[MONO] Godard, Aodha, Brostow, "Unsupervised Monocular Depth Estimation with Left-Right Consistency", CVPR 2017

Impressive results! (but remember ->)



Conclusions

- **Low-level vision problems recently tackled with ML**
- **Depth sensing and confidence measures: state-of-the-art**
- **Unsupervised training and monocular depth estimation very interesting topics for future research**

Acknowledgements*

Paolo Di Febbo
Matteo Poggi
Fabio Tosi

We gratefully acknowledge the support of NVIDIA Corporation with the donation of a Titan X Pascal GPU.

** alphabetical order*