Confidence estimation and early machine learning for stereo

Matteo Poggi[*], Fabio Tosi[*], Konstantinos Batsos[**], Philippos Mordohai[**], Stefano Mattoccia[*]

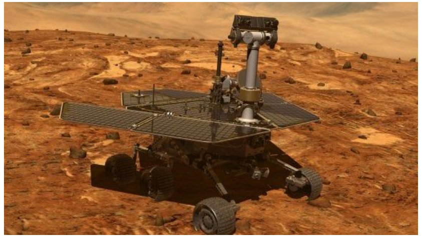
[*] University of Bologna
[**] Stevens Institute of Technology

Outline

- Introduction to depth sensing and stereo basics
- Confidence measures
- Learning based confidence measures
- Some applications of confidence measures
- Conclusions and open problems

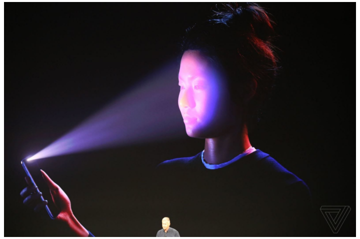
Depth sensing

Depth is a crucial cue for many computer vision applications



Robotic (NASA)

Autonomous driving (Google)



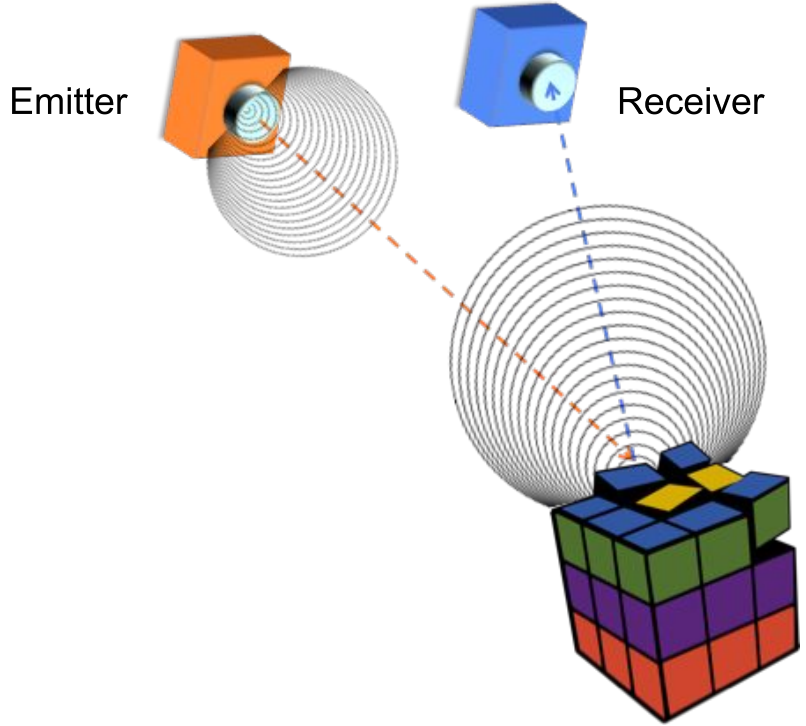
Biometric (Apple)

Drones (DJI)

Gaming (Microsoft)

Augmented Reality (Microsoft)

Active depth sensing 主动深度感知



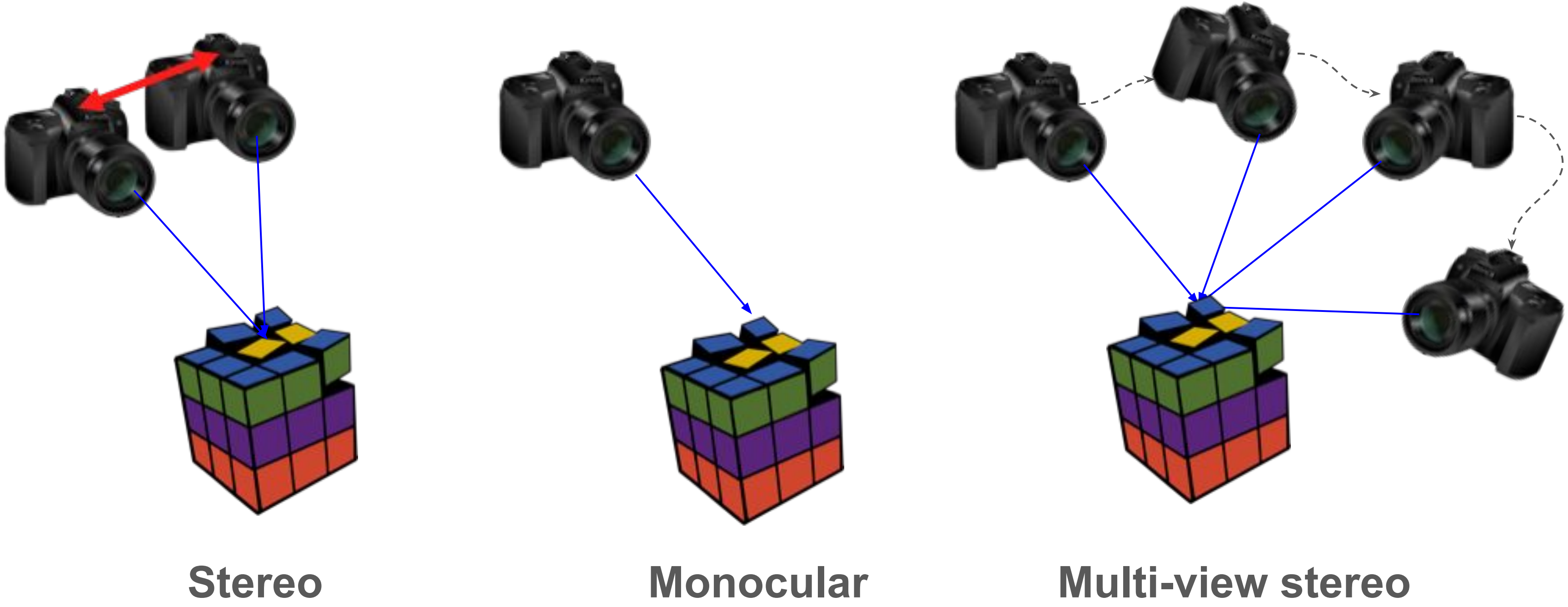
Depth is perceived by perturbing the sensed environment:

- LiDAR (e.g., Velodyne)
- Structured light (e.g., Kinect 1)
- Active stereo (e.g., Intel RealSense)

雷达
结构光
主动立体匹配

Passive depth sensing

被动深度感知：立体诗句，单目视觉，多视角立体视觉

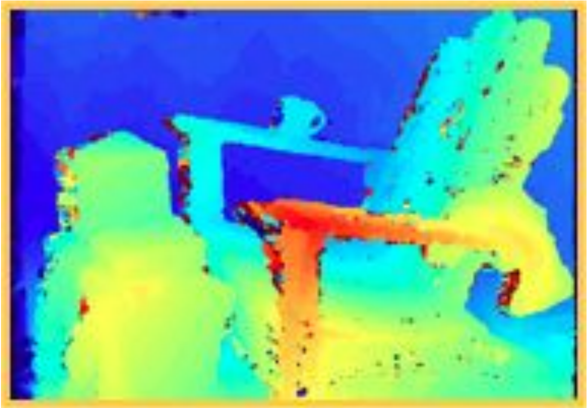


Stereo vision

- Given two (or more) images of the same scene, aims at inferring depth
- The **disparity** is the **difference between x coordinates of corresponding points** 视差=对应点x坐标的差值

Left (Reference)

Right (Target)

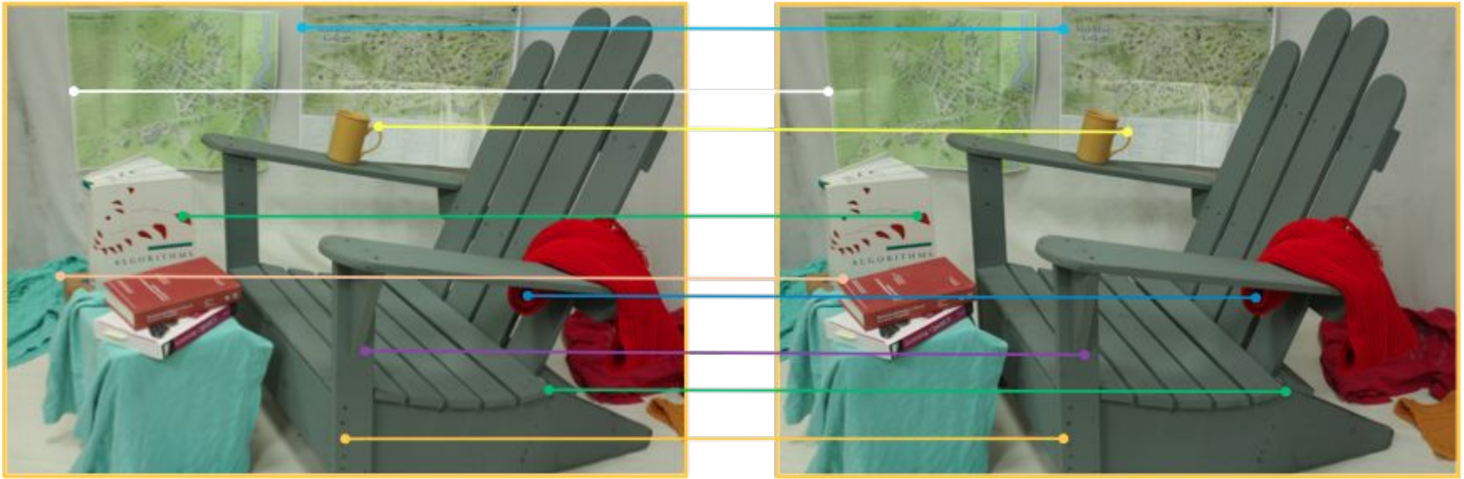


Disparity map

Correspondence problem

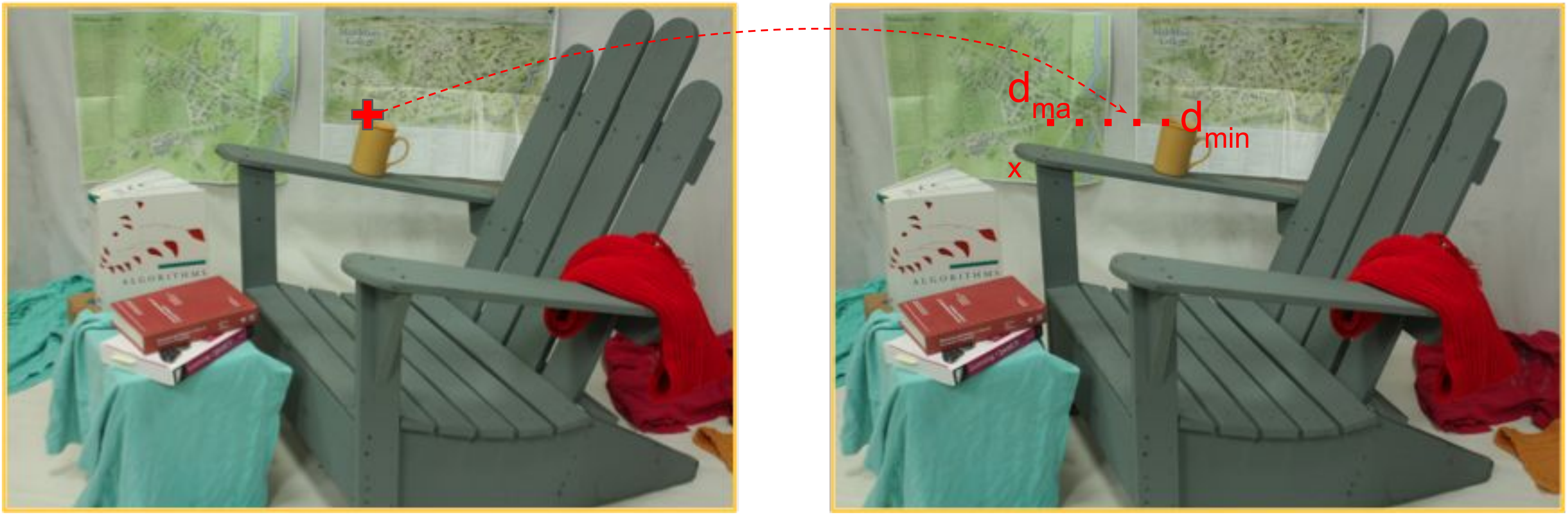
- Finding homologous points is crucial (and challenging) 同源点
- Stereo pairs are typically rectified (homologous points into the same scanline)
- Once found corresponding points, depth is inferred by a simple triangulation

矫正使得同源点在同一扫描线上，深度通过简单的三角化得到



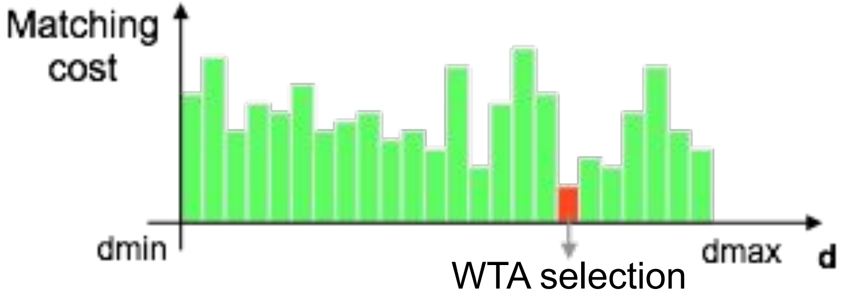
How to find homologous points?

- Looking for *similar* points/patches along scanlines 在扫描线上预设范围内寻找相似的点/块
- Corresponding points are sought within a prefixed (disparity) range **[d_{min},d_{max}]**



How to evaluate similarity between two points?

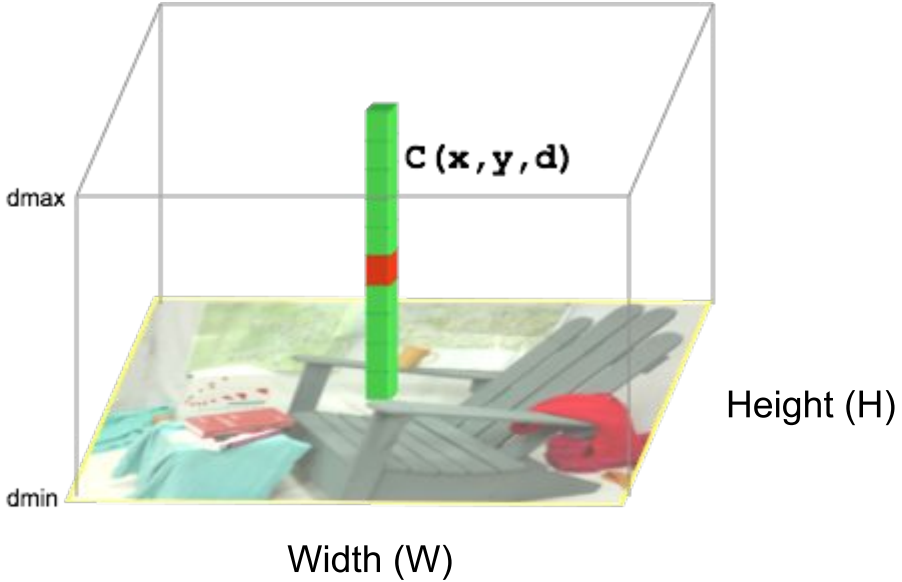
- Given a point p_R in the reference image, at each potential correspondence p_T in $[d_{min}, d_{max}]$ in the target image is associated a *score*
- Such score is referred to as ***matching cost*** $\mathbf{C(p_R, p_T, d)}$, with d in $[d_{min}, d_{max}]$
  - Pointwise matching cost (e.g., $|I(p_R) - I(p_T)|$) 绝对值匹配成本
  - Patch based matching cost (e.g., average $|I(p_R) - I(p_T)|$ on a patch) 块匹配成本



- Each p_R is assumed as uncorrelated to its neighbors 选择最小的误差
- Often, disparity selection consists in selecting the minimum score (**WTA**)

Cost volume or DSI (Disparity Space Image)

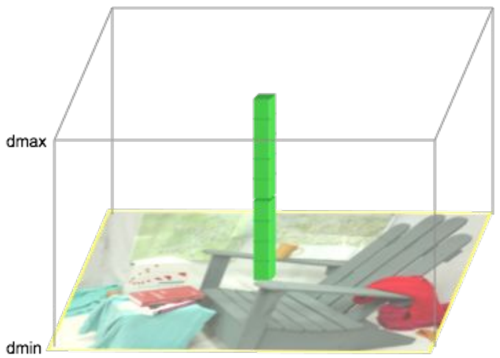
- The data structure containing all matching costs $C(p_R, p_T, d)$, with d in $[d_{min}, d_{max}]$



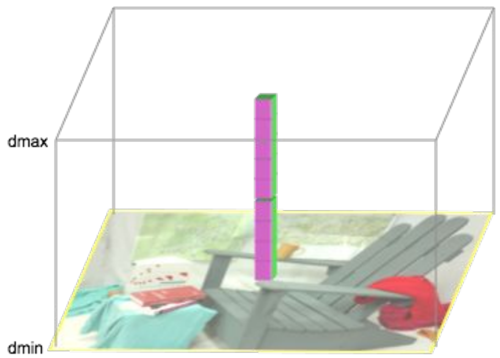
Cost volume optimization 1/2

- Often, the raw matching costs are further processed
- The outcome is a refined cost volume (e.g., to enforce smoothness)
- Yields better accuracy

光滑性refine成本体



Raw cost volume



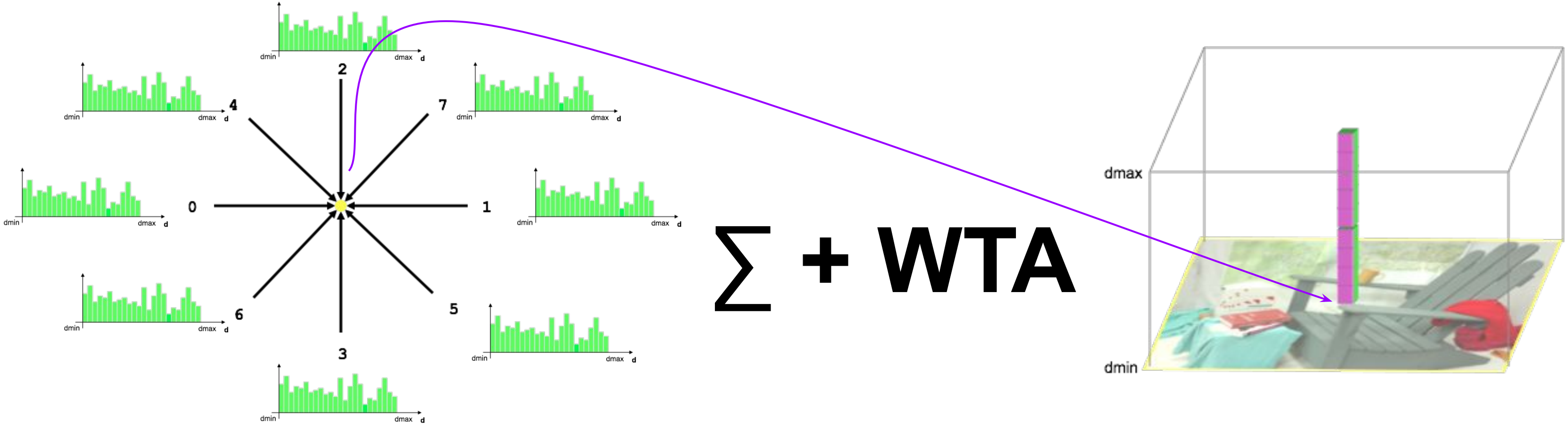
Refined cost volume

Disparity optimization 2/2

- Real scenes are piecewise smooth and often the disparity optimization step aims at enforcing such behaviour (among nearby pixels)
- The smoothing term is relaxed in proximity of (unknown) depth discontinuities
- Among disparity optimization methods, SGM (Hirschmuller, 2008) is a popular cost effective choice

Semi Global Matching (SGM) optimization

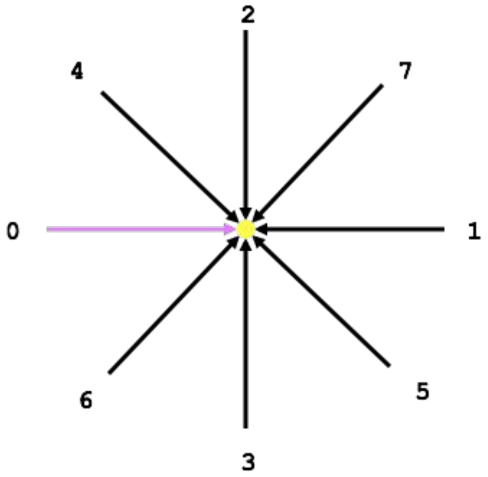
- Smoothness is enforced along multiple paths/scanlines (e..g, 4, 8 or 16) converging into the same target point
- Matching costs along each path are computed independently according to the *scanline optimization* (SO) approach
- The raw cost volume is replaced averaging the outcome of SOs, then WTA



Scanline Optimization (SO)

- Along each path (e.g. 0) , the raw matching cost **C** is refined (**R**) to enforce smoothness:

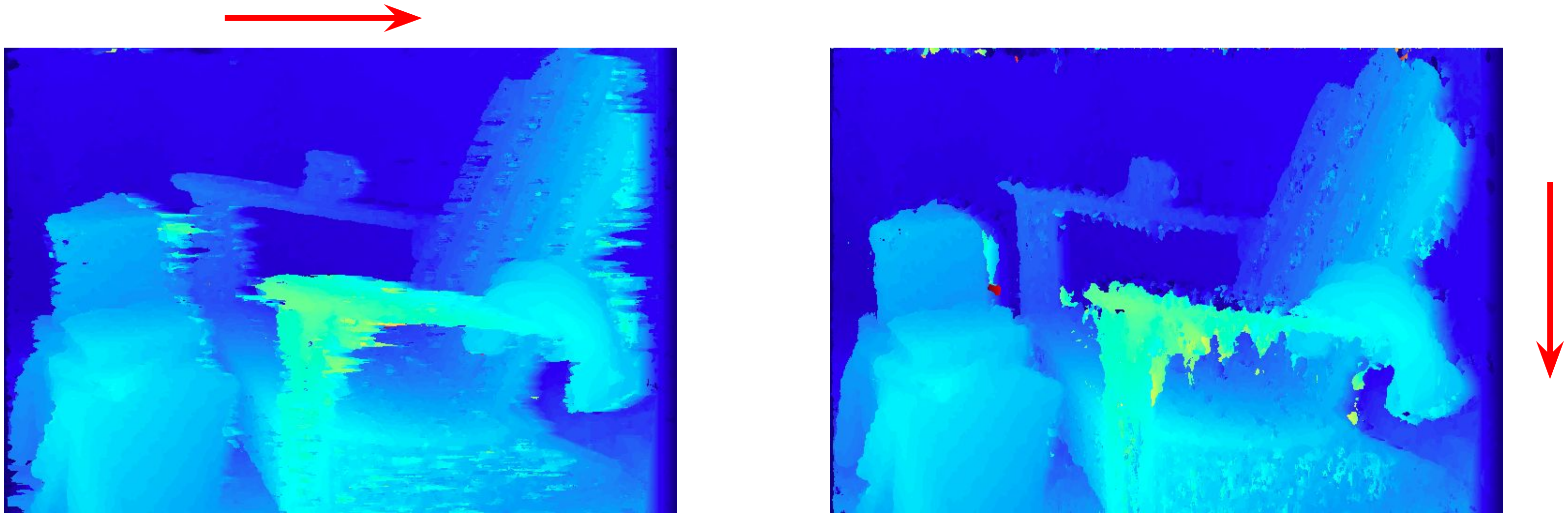
- **R**(current_p,**d**) = **C**(current_p,**d**) + min{ **R**(previous_p,**d**),
  R(previous_p,**d-1**) + **P1**,
  R(previous_p,**d+1**) + **P1**,
  R(previous_p,**d ≠ d, d-1, d+1**) + **P2** }



Smoothness penalties **P1** and **P2** (**P1**<**P2**) *discourage* disparity changes wrt previous disparity assignment along the scanline

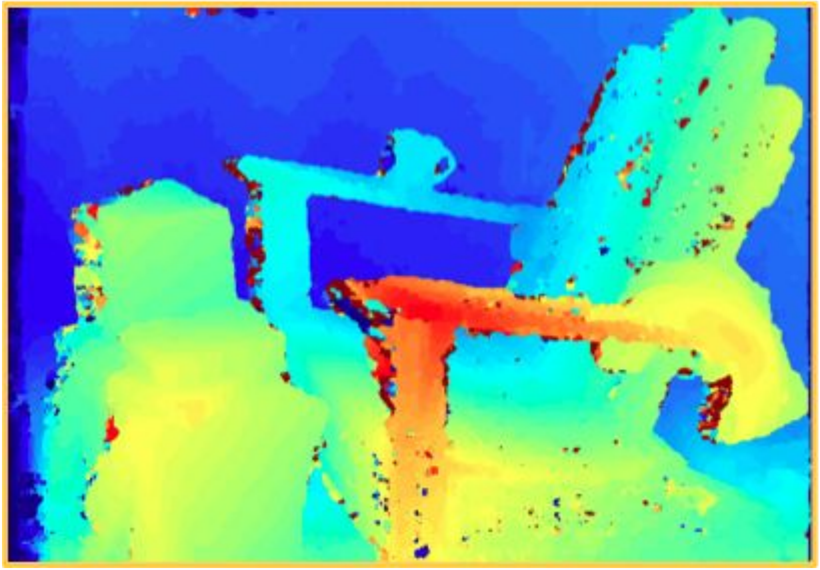
Scanline Optimization (SO) & SGM

- The outcome of each SO contains artifacts (*streaking*)
- Averaging the DSI computed along each scanline yields much better results
- P1 and P2 setting is crucial

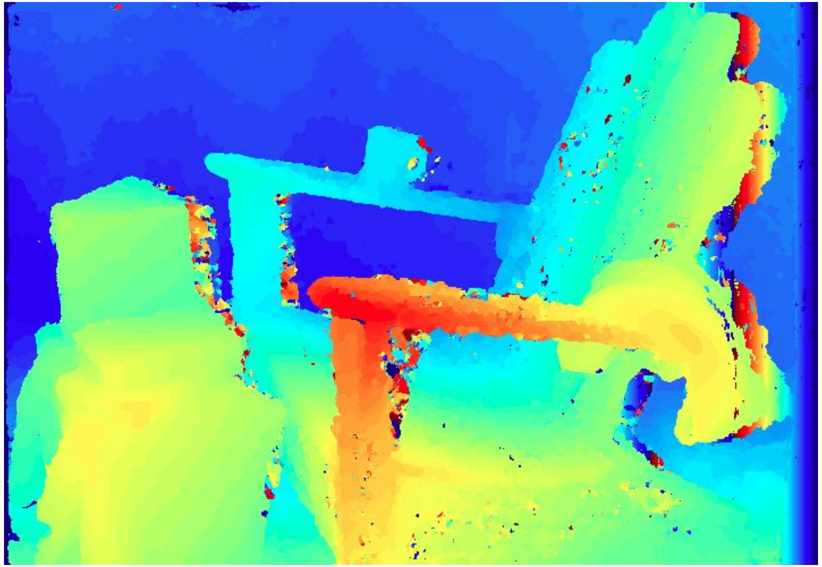


The role of the reference image

- The disparity map is computed according to the image assumed as reference
- Given a stereo pair, we can obtain two disparity maps: $\mathbf{D}_{\mathbf{LEFT}}$ and $\mathbf{D}_{\mathbf{RIGHT}}$



$\mathbf{D}_{\mathbf{LEFT}}$

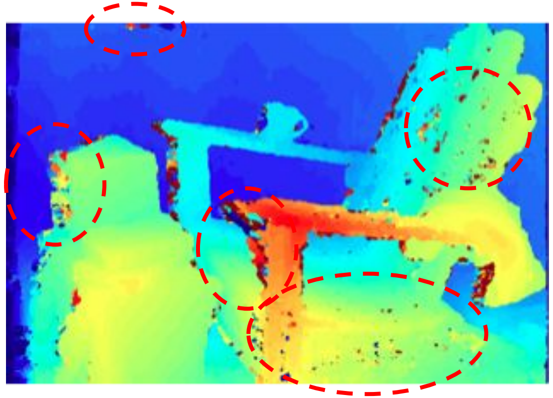


$\mathbf{D}_{\mathbf{RIGHT}}$

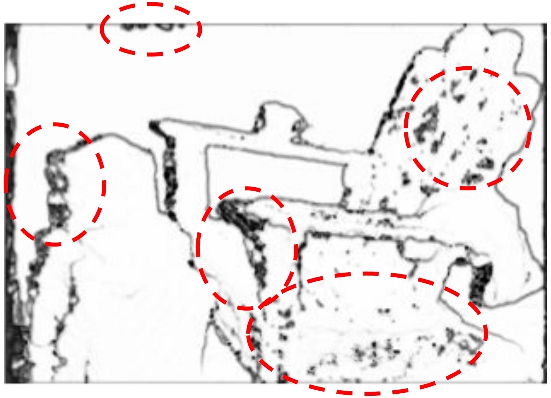
Confidence measures (CM)

- Regardless of the stereo algorithm, disparity maps contain outliers
- Confidence estimation aims at detecting unreliable depth assignments

Reference image



Disparity map (SGM)

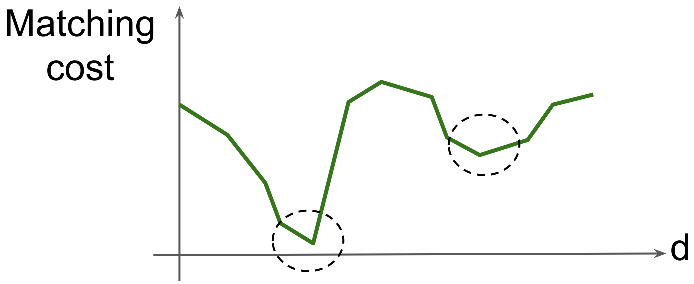


Confidence map
(the brighter, the more reliable)

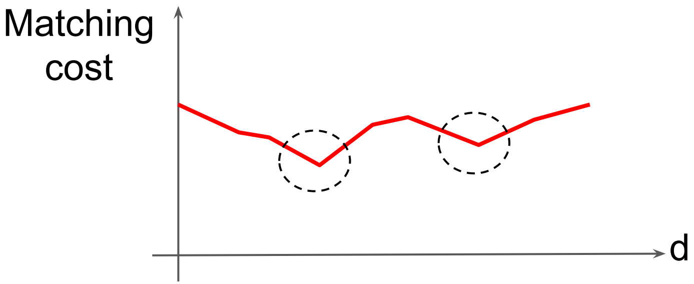
Confidence estimation basics

- Conventional methods, reviewed and evaluated in (Hu, 2012), relies on assumptions mostly based on matching cost analysis
- For instance, the matching costs on the left are assumed to be more likely to yield a more reliable correspondence compared to the right ones
- Many other heuristics have been proposed in the literature
- Standard evaluation metric: the Area Under the Curve (AUC)

左图比右图置信度更高



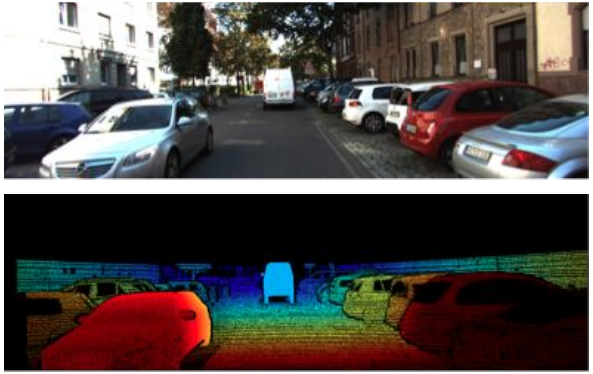
High confidence (?)



Low confidence (?)

Evaluation datasets

- For evaluation datasets with groundtruth (GT) depth labels are required
- For confidence evaluation: KITTI 2012, KITTI 2015 and Middlebury v3
- KITTI 2012 and 2015, respectively, 194 and 200 stereo pairs with GT
- Recently, made available longer KITTI sequences with GT depth labels
- Middlebury, (only) 23 stereo pairs with GT labels



(Geiger, 2012)

LiDAR

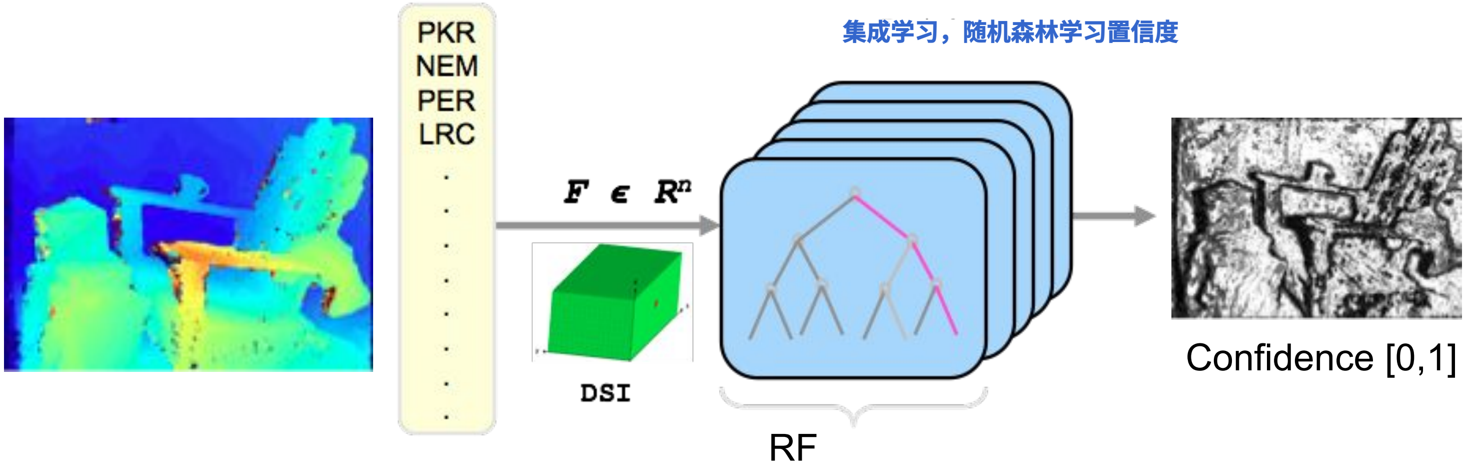
Structured light

(Scharstein, 2014)

Confidence measures and machine learning (ML)

开创性的

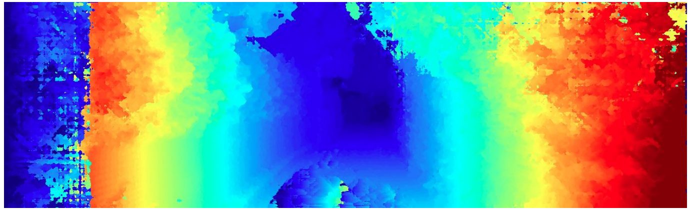
- A seminal work in this field is represented by ENSEMBLE (Haeusler, 2013)
- Idea: feeding a random forest (RF) with a pool of (23) standard CMs
- Much better accuracy compared to any CM included in the pool
- Other methods: (Spyropoulos, 2014) and (Park, 2015)

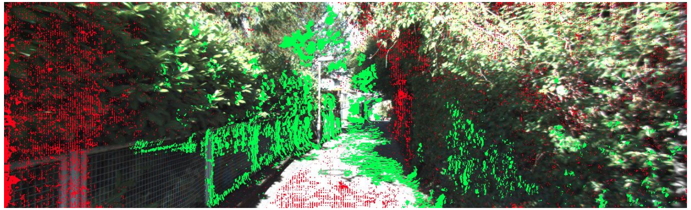


Training of confidence measures

- Learning based CMs require training data
- The ML framework is trained on a balanced number of samples
- Few stereo pairs (e.g. 10 or 20) with GT labels
- Unsupervised training of CMs is feasible: (Mostegel, 2016) and (Tosi,2017)

SGM



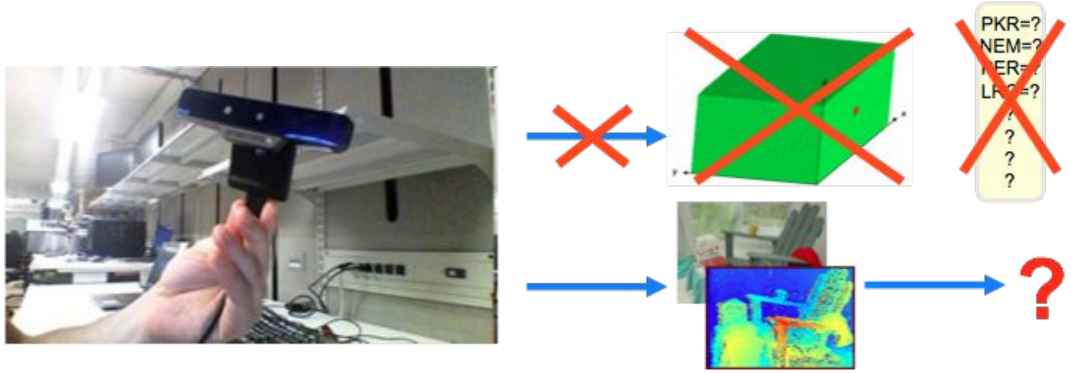


Green, *good* samples
Red, *bad* samples

Can we learn CMs without a DSI?

- Previous ML methods rely on features extracted from the DSI
- DSI is not always available
    - Off-the-shelf stereo camera (e.g., Intel Real Sense)
    - Closed source stereo algorithms or pre-computed maps
    - Deep stereo and monocular networks
    - Depth sensors based on other technologies? (e.g., Kinect?, LiDAR?)

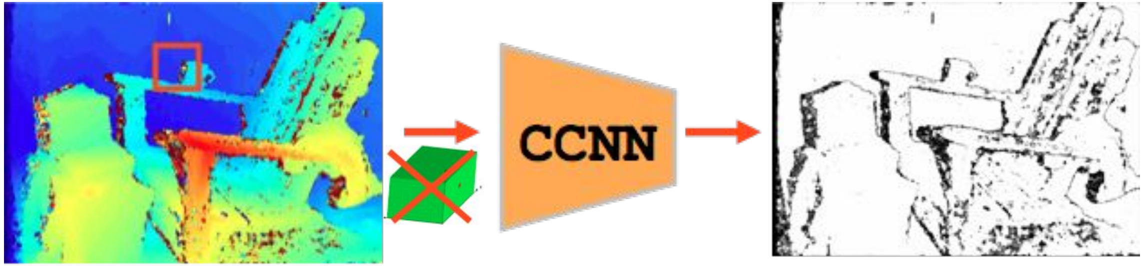
Intel Real Sense



Can we learn a CM in the disparity domain?

Learning CMs with a CNNs in the disparity domain

- Input, the reference disparity map, **no DSI**
- End-to-end learning of a CM in the disparity domain: CCNN (Poggi, 2016b) requires only D_L, PBCP (Seki, 2016) requires D_L and D_R
- Better results vs methods based on hand-crafted features and RF
- Exhaustive evaluation of learning-based CMs in (Poggi, 2017c)



Left disparity

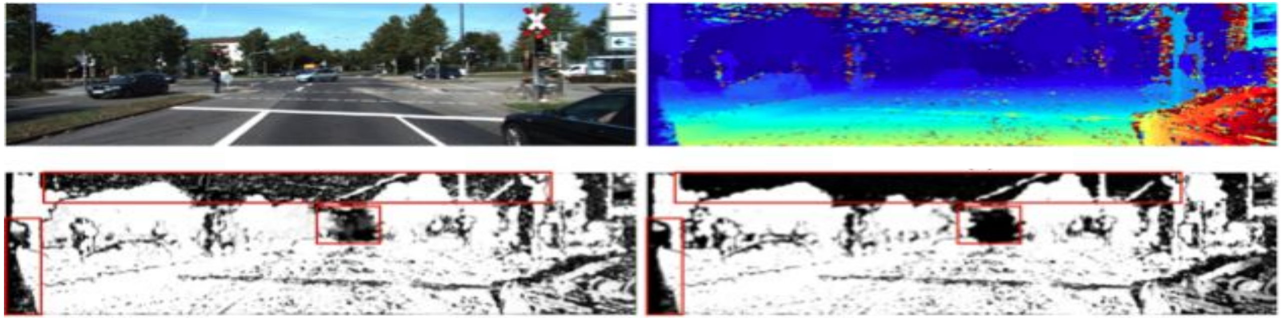
CCNN

CCNN Source code: https://github.com/fabiotosi92/CCNN-Tensorflow

How to further improve confidence prediction?

Despite the excellent performance of learning based methods, confidence prediction can be further improved:

1. Given an existing CM, by learning to enforce its local consistency (Poggi, 2017a)
2. From scratch, moving beyond local reasoning with LGCNet (Tosi, 2018)



CCNN LGCNet

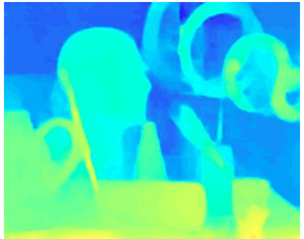
Applications of CMs

- Self supervised training of CMs (Tosi, 2017)
- Outliers detection and disparity refinement (Tosi, 2019)
- Improving stereo accuracy (Spyropoulos, 2014), (Park,2015), (Poggi, 2016a)
- Disparity fusion (Spyropoulos, 2015), (Poggi, 2016c)
- Sensor fusion (e.g., Time of Flight and stereo vision) (Marin, 2016)
- Domain shift adaptation for deep stereo (Tonioni, 2017)

Other applications discussed later

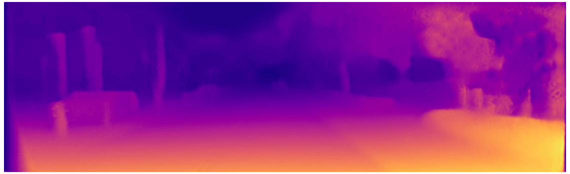
Conclusions and open problems

- CMs are useful and effective when dealing with traditional stereo algorithms
- Can we detect outliers in disparity maps generate by deep architectures for depth estimation? Yes (e.g., CCNN), but there’s room for improvements

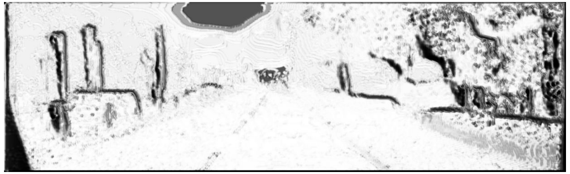


Deep stereo

CCNN



Monocular depth estimation



CCNN

References 1/7

(Hirschmuller, 2008) Hirschmuller, Stereo processing by semiglobal matching and mutual information, PAMI 2008

(Hu, 2012) Hu and Mordohai, A Quantitative Evaluation of Confidence Measures for Stereo Vision, PAMI 2012

(Geiger, 2012) Geiger, Lenz, Urtasun, “Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite”, CVPR 2012

(Scharstein, 2014) Scharstein, Hirschmüller, Kitajima, Krathwohl, Nesic, Wang, Westling, “High-resolution stereo datasets with subpixel-accurate ground truth”, GCPR 2014

References 2/7

(Haeusler, 2013) Haeusler, Nair, and Kondermann, “Ensemble learning for confidence measures in stereo vision”, CVPR 2013

(Spyropoulos, 2014) Spyropoulos, Komodakis, Mordohai, “Learning to detect ground control points for improving the accuracy of stereo matching”, CVPR 2014

(Park, 2015) Park and Yoon, “Leveraging stereo matching with learning-based confidence measures”, CVPR 2015

(Spyropoulos, 2015) A. Spyropoulos and P. Mordohai. Ensemble classifier for combining stereo matching algorithms, 3DV 2015

References 3/7

(Poggi, 2016a) 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
Code: http://vision.deis.unibo.it/~mpoggi/code/3DV2016_O1.zip

(Poggi, 2016b) Poggi and Mattoccia, "Learning from scratch a confidence measure", BMVC 2106
Code: https://github.com/fabiotosi92/CCNN-Tensorflow

(Poggi, 2016c) Poggi and Mattoccia, "Deep Stereo Fusion: combining multiple disparity hypotheses with deep-learning", 3DV 2016
Code: http://vision.deis.unibo.it/~mpoggi/code/3DV2016_test.zip
http://vision.deis.unibo.it/~mpoggi/code/3DV2016_train_and_test.zip

References 4/7

(Poggi, 2017a) Poggi and Mattoccia, “Learning to predict stereo reliability enforcing local consistency of confidence maps”, CVPR 2017
Code: http://vision.deis.unibo.it/~mpoggi/code/CVPR2017_test.zip
http://vision.deis.unibo.it/~mpoggi/code/CVPR2017_train_and_test.zip

(Poggi, 2017b) Poggi, Tosi, Mattoccia, “Even More Confident predictions with deep machine-learning”, EVW 2017

(Poggi, 2017c) Poggi, Tosi, Mattoccia, “Quantitative evaluation of confidence measures in a machine learning world”, ICCV 2017
Code: http://vision.deis.unibo.it/~mpoggi/code/ICCV2017.zip

References 5/7

(Tosi, 2017) Tosi, Poggi, Tonioni, Di Stefano, Mattoccia, “Learning confidence measures in the wild”, BMVC 17
Code: https://github.com/fabiotosi92/Unsupervised-Confidence-Measures

(Tosi, 2018) Tosi, Poggi, Benincasa, Mattoccia, “Beyond local reasoning for stereo confidence estimation with deep learning”, ECCV 2018
Code: https://github.com/fabiotosi92/LGC-Tensorflow

(Tonioni, 2017) Tonioni, Poggi, Mattoccia, Di Stefano, “Unsupervised Adaptation for Deep Stereo”, ICCV 2017
Code: https://github.com/CVLAB-Unibo/Unsupervised-Adaptation-for-Deep-Stereo

References 6/7

(Poggi, 2017d) Poggi, Tosi, Mattoccia, “Efficient confidence measures for embedded stereo”, ICIAP 2017

(Fu, 2017) Fu and Ardabilian, “Stereo matching confidence learning based on multi-modal convolution neural networks”, RFMI 2017

(Marin, 2016) Marin, Zanuttigh, Mattoccia, “Reliable fusion of ToF and stereo depth driven by confidence measures”, ECCV 2016

(Schoenberger, 2018) Schoenberger, Sinha, Pollefeys, Learning to Fuse Proposals from Multiple Scanline Optimizations in Semi-Global Matching, ECCV 2018

References 7/7

(Mostegel, 2016) Mostegel, Rumpler, Fraundorfer, Bischof, “Using self-contradiction to learn confidence measures in stereo vision”, CVPR 2016

(Seki, 2016) Seki and Pollefeys, “Patch Based Confidence Prediction for Dense Disparity Map”, BMVC 2016

(Tosi, 2019) Tosi, Poggi, Mattoccia, “Learning to detect and take advantage of reliable anchor points for embedded stereo refinement”, EVW 2019