ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA
DIPARTIMENTO DI INGEGNERIA CIVILE, CHIMICA, AMBIENTALE E DEI MATERIALI

PROFESSIONAL MASTER'S PROGRAMME 2ND LEVEL
SUSTAINABLE AND INTEGRATED MOBILITY IN URBAN REGIONS

Monocular Depth Estimation

Fabio Tosi:
Dipartimento di Informatica, Università di Bologna
MODULO:
Autonomous vehicles , Electric cars and recharging systems

Imola, 29/11/2019

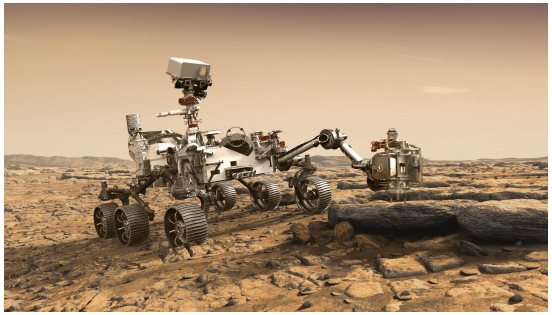
With the unconditional support:

FONDAZIONE Cassa di Risparmio di IMOLA
ORDINE INGEGNERI BOLOGNA
FAM FONDAZIONE ALMA MATER
architettibologna

With the contribution of:

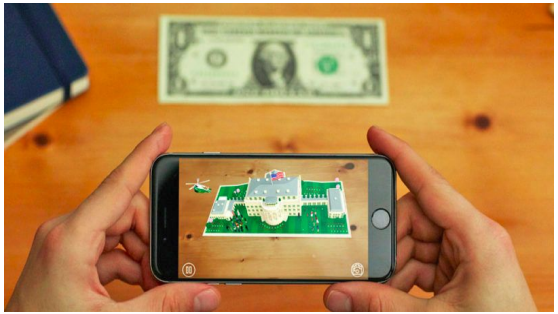
COOPERATIVA TRASPORTI IMOLA
TERCHIMICA
ZINIELIO

Monocular Depth Estimation - Motivation



Robotics

ADAS

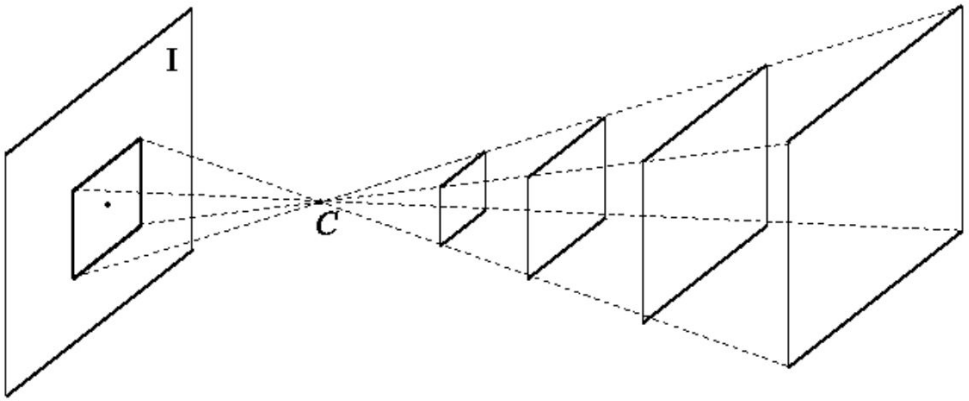


AR/VR

Medical Applications

Perspective Projections

- The image formation process deals with mapping a 3D space onto a 2D space
- Indeed, the mapping is not a bijection
- Estimating depth from a single image is an ill-posed problem



Perceiving 3D from 2D

- Humans excel at this task

Perceiving 3D from 2D

- Meaningful monocular cues:
  - ❏ Linear Perspective
  - ❏ Relative Size
  - ❏ Superimposition
  - ❏ Texture Gradient
  - ❏ Height in plane

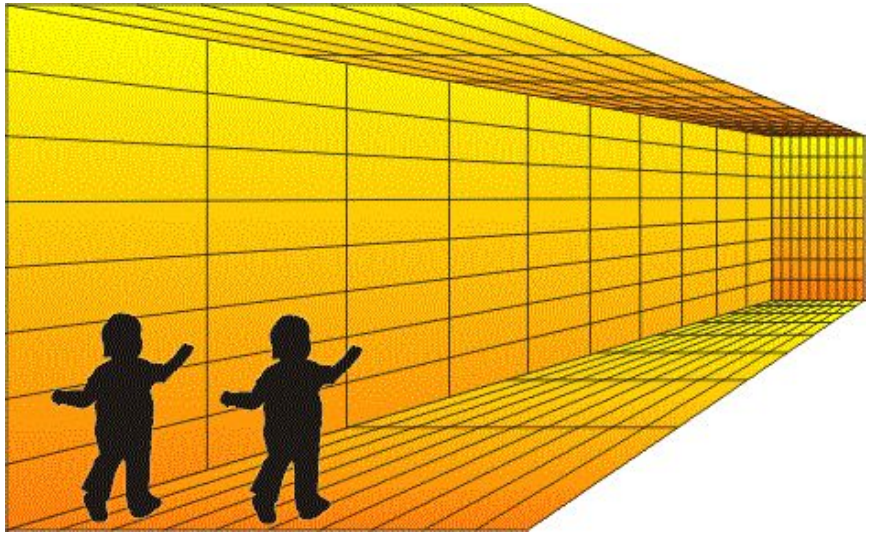
Optical illusions

- Monocular cues don’t always help :

Optical illusions

Optical illusions

- Ponzo illusion:



https://www.eruptingmind.com/depth-perception-cues-other-forms-of-perception/

https://www.moillusions.com/these-3-cars-are-same-in-size/

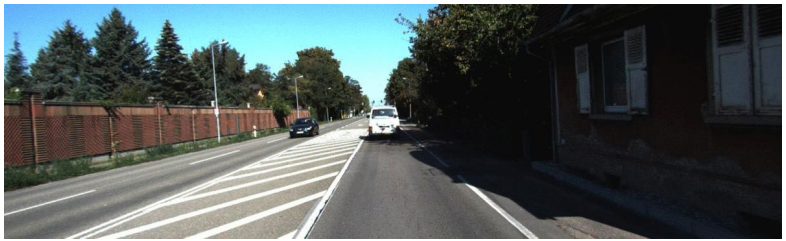
Depth estimation in monocular images

In Computer Vision, existing solutions to depth estimation from a single image usually rely on deep learning based approaches:

- Supervised
  - ground-truth depth data (RGB-D cameras, 3D laser scanners)
- Semi-Supervised
  - sparse ground-truth depth + image reconstruction
- **Self-Supervised**
  - image reconstruction (from monocular videos/stereo pairs/stereo sequences)
- **Proxy-Supervised**
  - depth annotations generated from traditional algorithms using stereo or video sequences

Depth estimation in monocular images

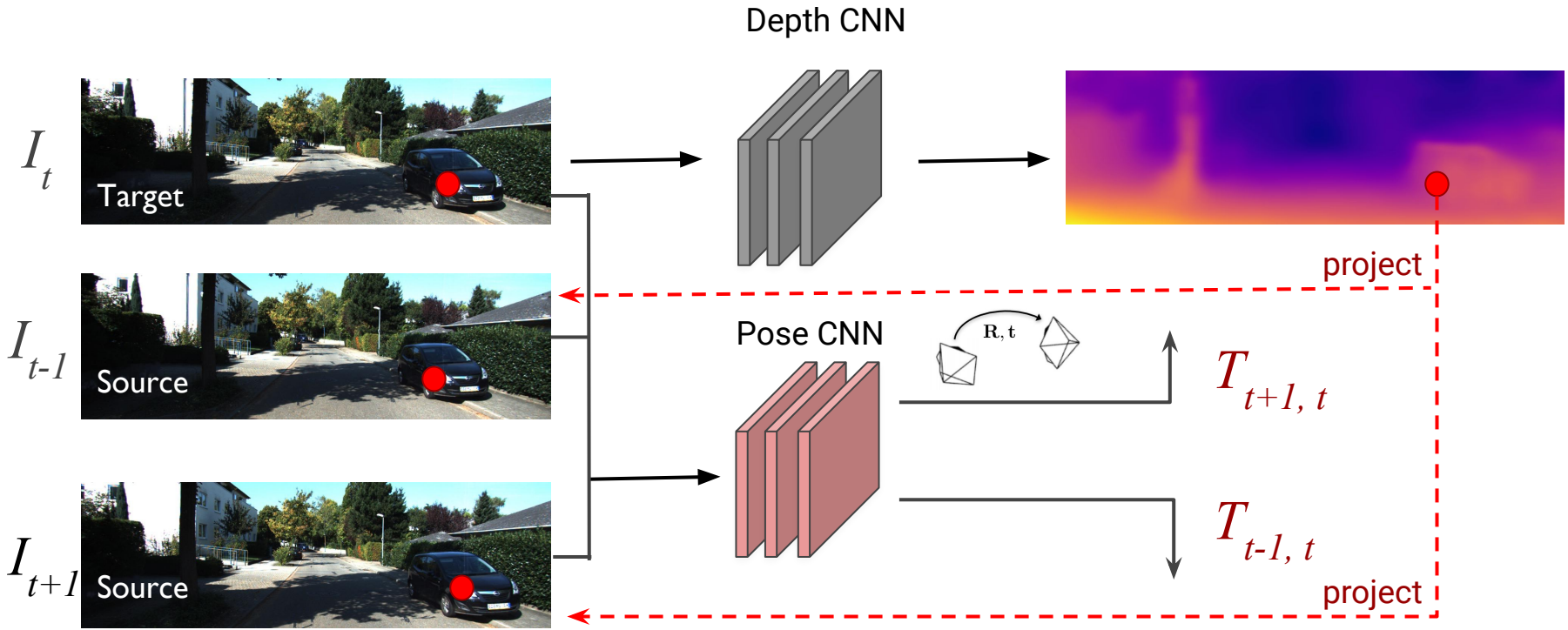
At training time:



Monocular videos

Stereo pairs

Depth from videos

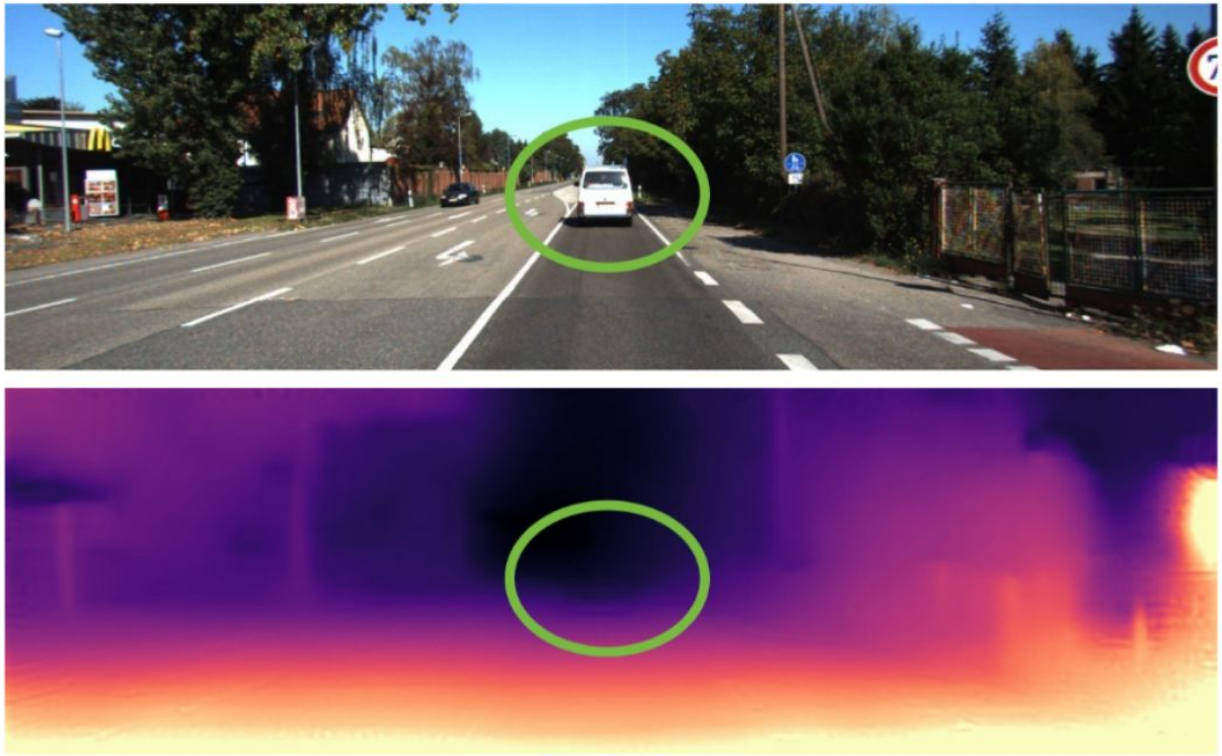


Depth from videos - Assumptions

- The view synthesis formulation implicitly assumes:
  - Sufficient illumination in the environment
  - The scene is **static** without moving objects
  - Sufficient motion parallax in successive frames
  - Sufficient scene overlap between consecutive frames
  - There is **no occlusion** between the target view and the source views
  - The surface is **Lambertian**

Car Following Scenario

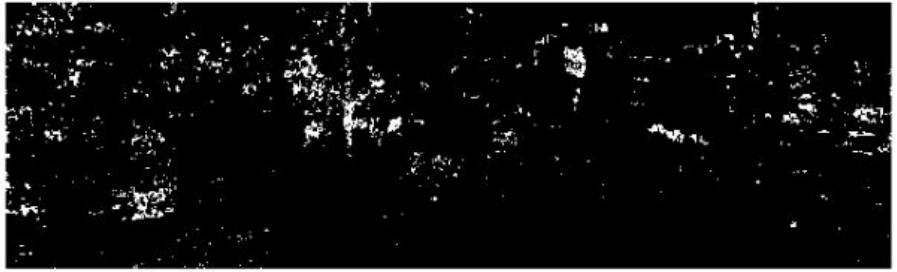
Car Following Scenario



Auto-masking

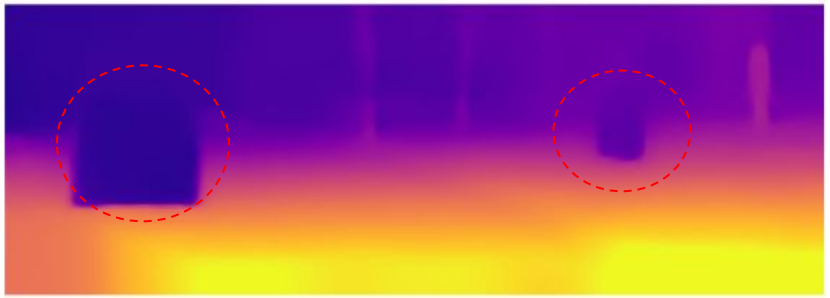
Goal: Filter out pixels which do not change appearance from one frame to the next in the sequence.



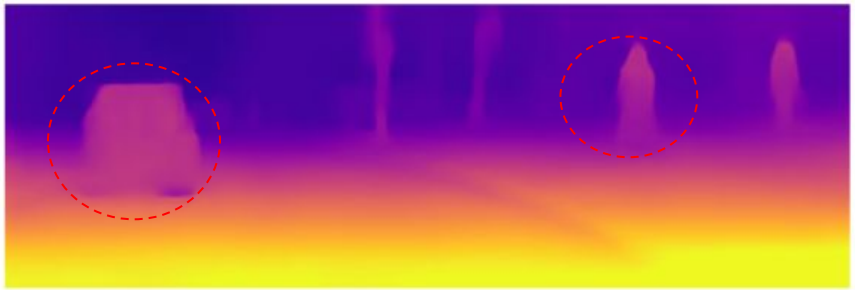



Auto-masking

Single-view RGB image



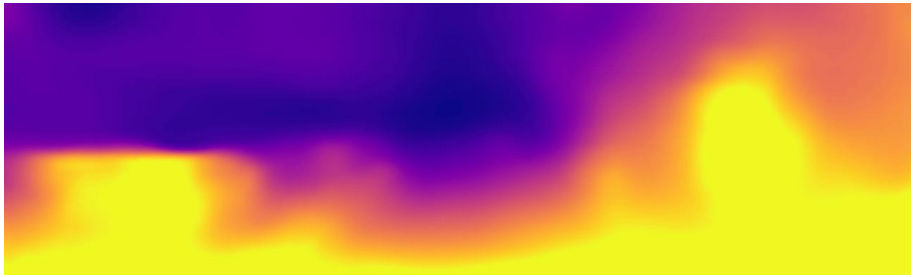
Depth prediction - W/O auto-mask



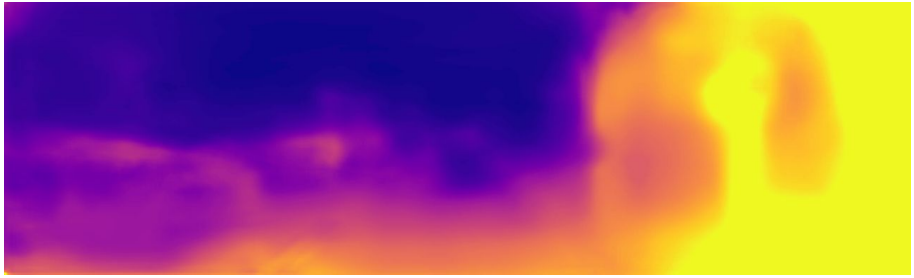
Depth prediction - With auto-mask

Single-view depth from videos - Evolution

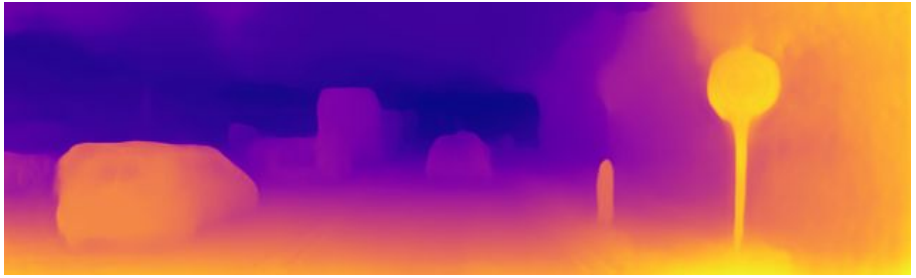
Single-view input



Zhou et al. 2017



Yin et al. 2018



Godard et al. 2019

Depth from videos

Pros:

- It is possible to train a monocular depth network on an infinite number of unconstrained images (e.g. YouTube videos)

Cons:

- Depth is defined up to a scale factor
- The camera pose is unknown
- Dynamic scenes are problematic

Depth from synchronized stereo pairs

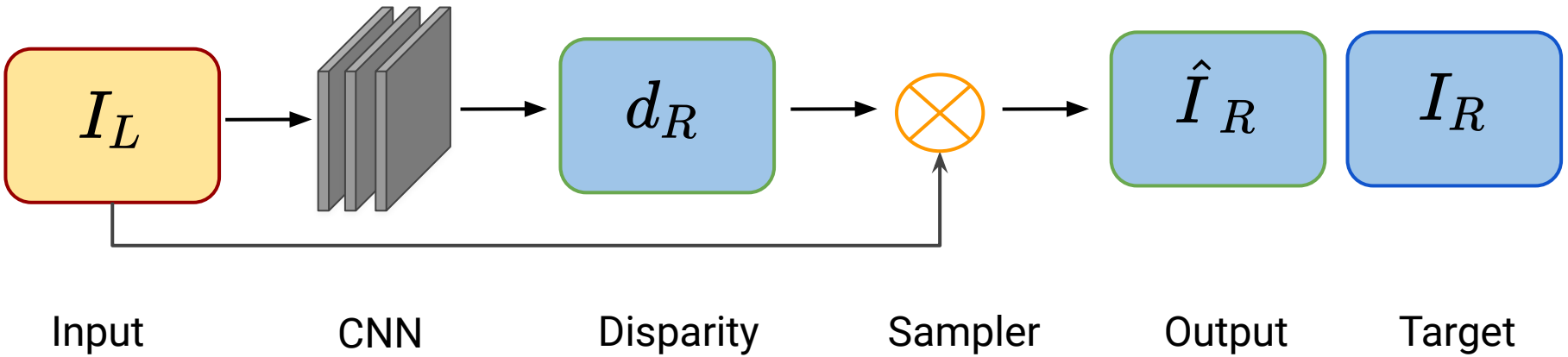
- Using stereo data for training makes the camera-pose estimation a one-time offline calibration

- Given a calibrated stereo pair at training time, the goal is to find a dense correspondence field (***disparity***) that, when applied to the left/right image, would enable to reconstruct the right/left image (Garg, 2016)

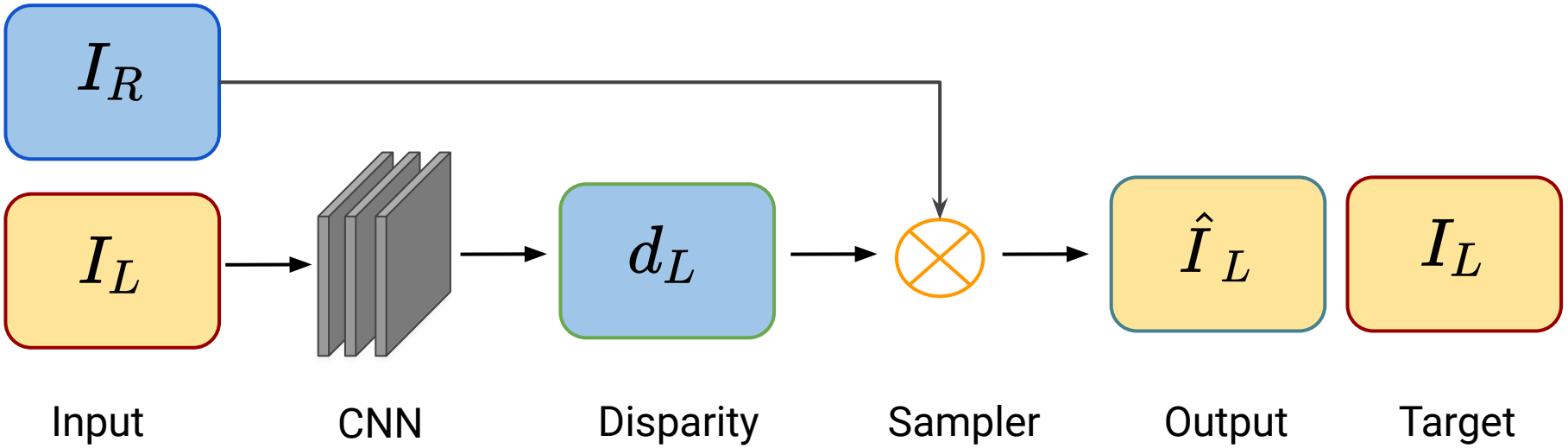
- Given the predicted disparity, the baseline and the focal length, we can trivially recover the depth as:

$$z = \frac{bf}{d}$$

Depth from stereo images - Naive approach



Depth from stereo images



Depth from synchronized stereo pairs

Pros:

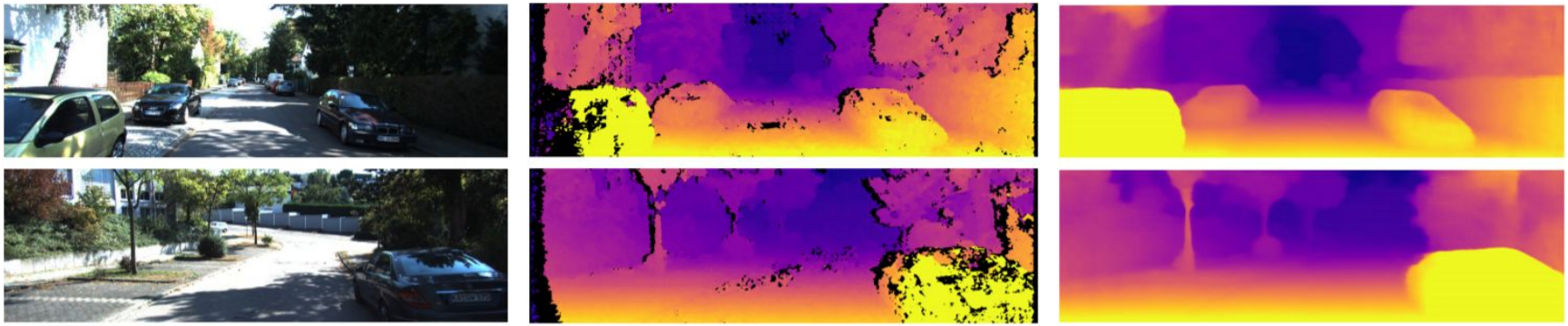
- It works on dynamic scenes
- Inference without any scale ambiguity (but..)
- The camera pose is known
- Better monocular depth prediction (at the moment)

Cons:

- Stereo pairs are not always available

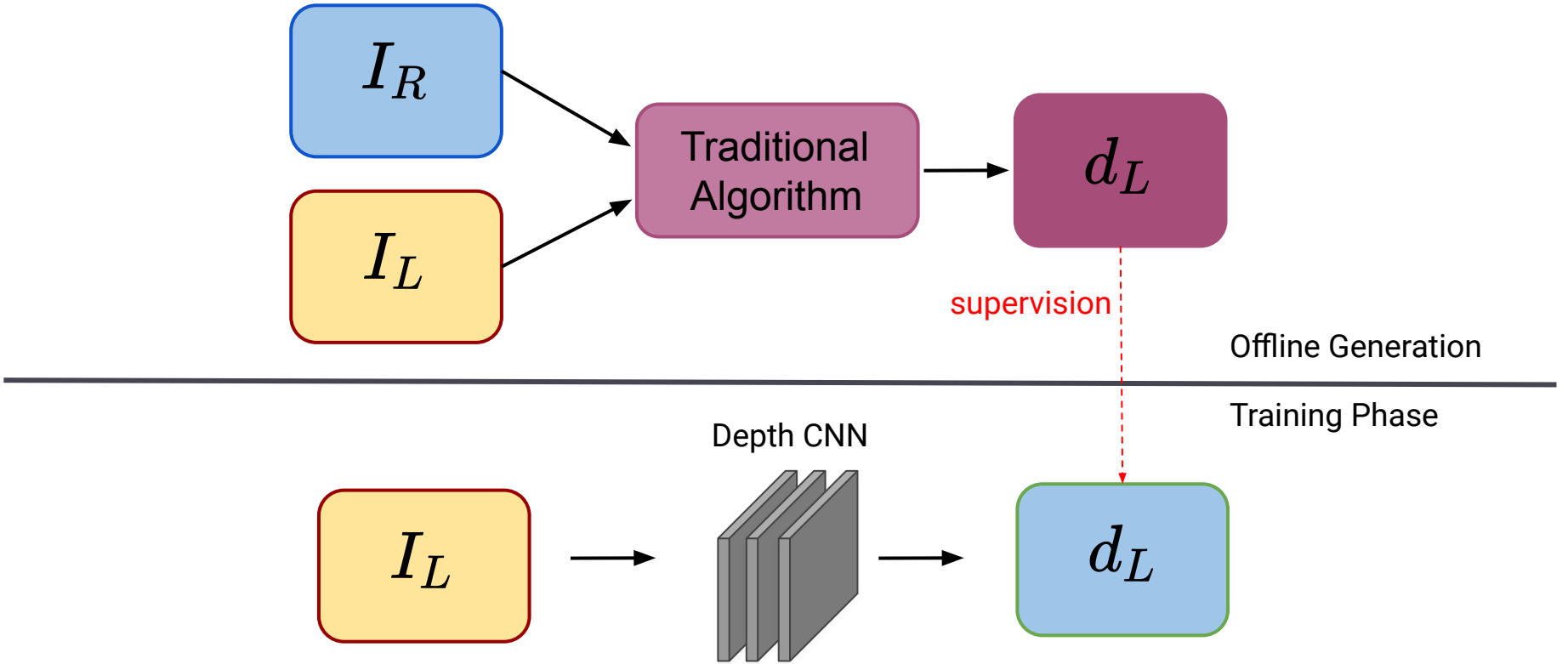
Proxy-supervision from stereo algorithms

- Before deep-learning era, stereo algorithms were good indeed at estimating semi-dense disparity maps

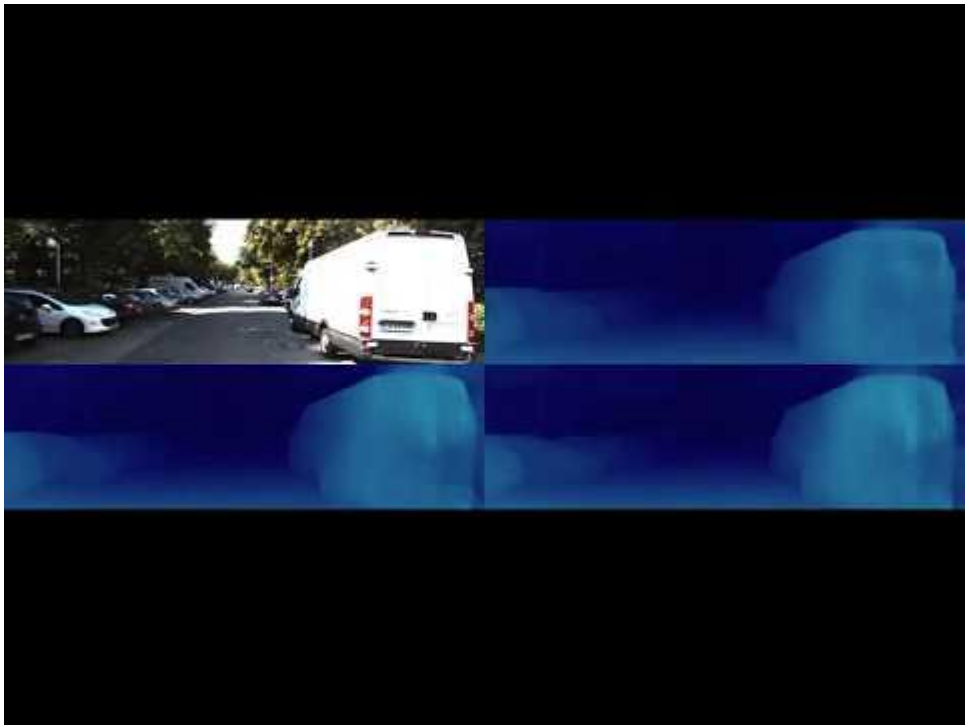


- They are effective source of proxy-supervision for both deep stereo networks (Tonioni et al., ICCV 2017) as well as depth-from-mono frameworks (Tosi et al., CVPR 2019)

Monocular Depth from Proxy Supervision



Monocular Depth from Proxy Supervision



Single-view depth estimation

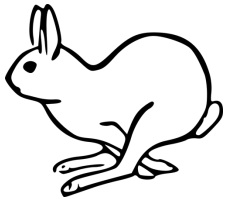
- Qualitative comparison

Towards real-time unsupervised monocular depth estimation on CPU

- Current architectures for monocular depth estimation are very deep and complex; for these reasons they require dedicated hardware such as high-end and power-hungry GPUs.

- This fact precludes to infer depth from a single image in many interesting applications fields characterized by low-power constraints (e.g. UAVs, wearable devices, ...)

Can we run such systems everywhere?



High-end GPU (i.e. nVidia Titan X)

- Power hungry (250 Watt) Nearly 30 fps (~ 0,035s per frame)

Average CPU (i.e., Intel i7)

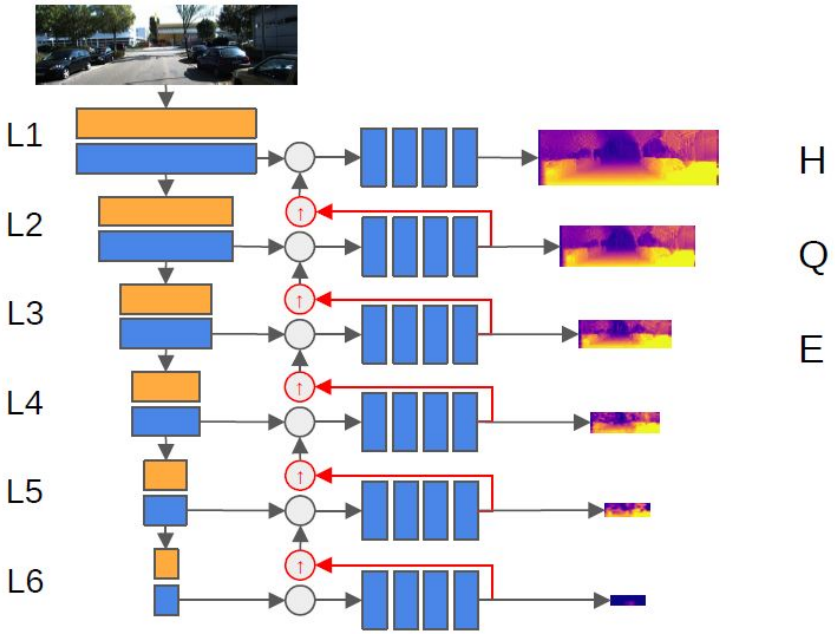
- Lower energy requirements (~90 Watt)
- Less than 2 fps (~ 0,60s per frame)

Embedded CPU (i.e., Raspberry Pi 3)

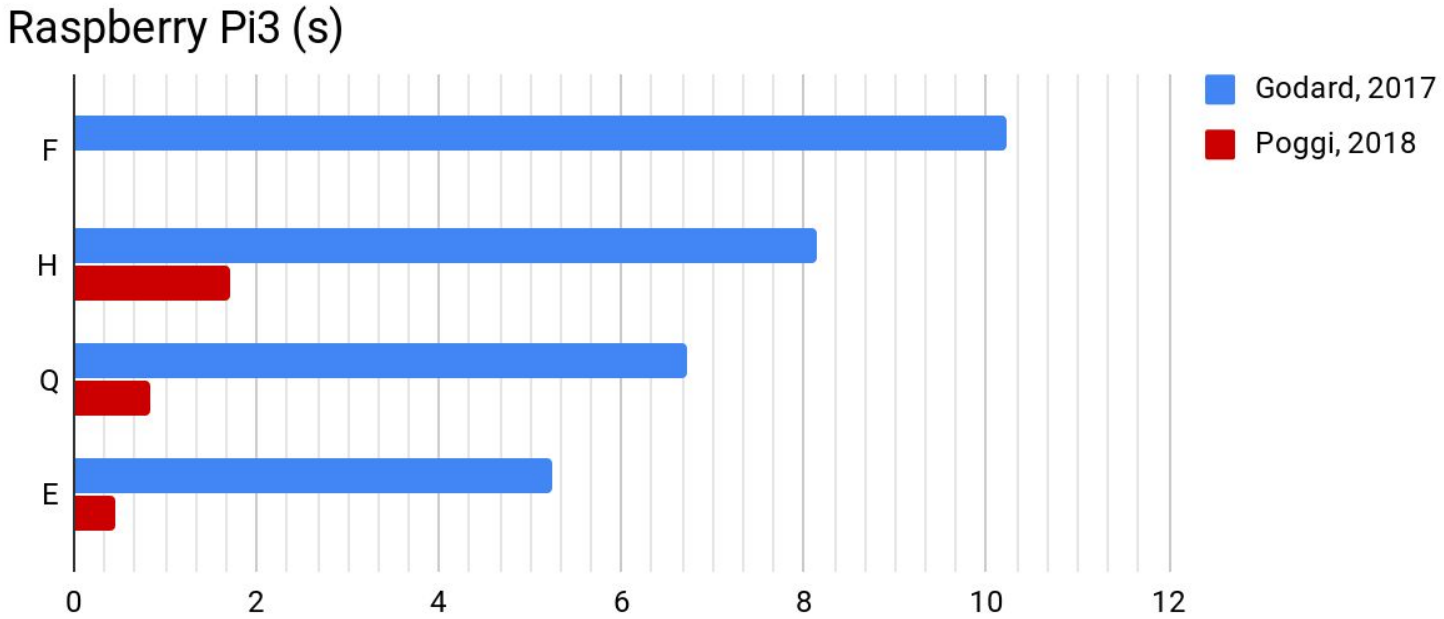
- Extremely low consumption (~3,5 Watt)
- Incredibly SLOW (~ 10s per frame)

Towards real-time unsupervised monocular depth estimation on CPU

- Shallow, **pyramidal features** encoder
- **Coarse-to-fine** strategy: depth is estimated from lower to higher resolution by lightweight decoders
- Each decoder outputs depth, so as we can **early stop** to trade accuracy for efficiency
- About **6% complexity** compared to Godard et al., CVPR 2017 (1.9M vs 31.6M params)



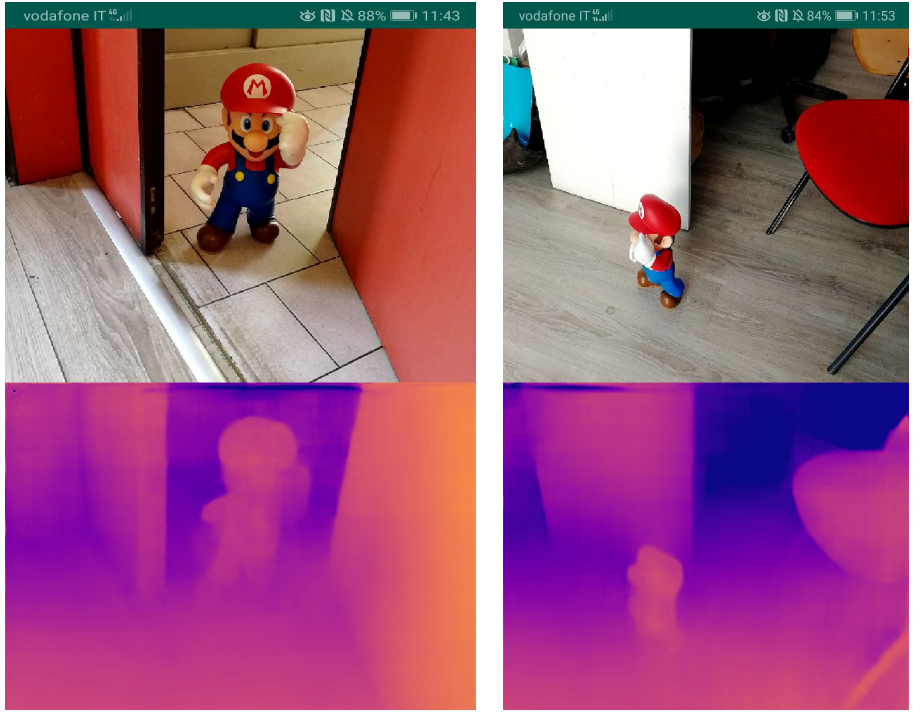
Towards real-time unsupervised monocular depth estimation on CPU



Towards real-time unsupervised monocular depth estimation on CPU

Monocular Depth Estimation on mobile devices

- iOS and Android
- The network has been trained on indoor scenes (Matterport) dataset for 1.2 M steps on Microsoft Kinect depth labels offered by the dataset as supervision



https://github.com/FilippoAleotti/mobilePydnet

Thank you

Fabio Tosi
Dipartimento di Informatica, Università di Bologna
https://vision.disi.unibo.it/~ftosi/