SAPIENZA
UNIVERSITÀ DI ROMA

Dipartimento di Ingegneria informatica, automatica e gestionale Antonio Ruberti

Depth Perception

Lorenzo Papa

papa@diag.uniroma1.it

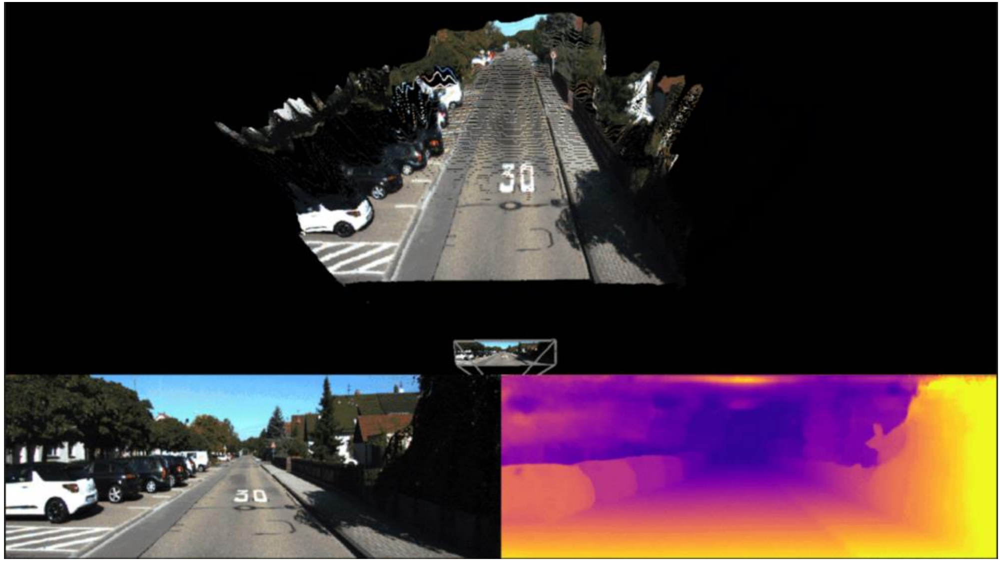
ALCOR Lab

What is the first information lost when an image is captured from a camera sensor?

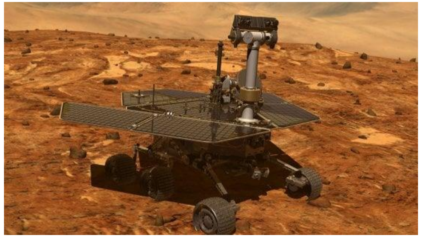
ALCOR Lab

Overview:

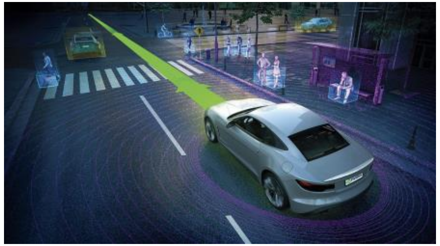
- Why depth?
- Active depth sensing
- Passive depth sensing
    - Binocular, stereo vision
    - Monocular depth estimation
- State-of-the-art/Examples
- Evaluation metrics
- Datasets
- AlcorLab research projects
    - PhD projects
    - Master thesis
    - Open challenges



Applications of depth sensing



Robotic



Autonomous driving

Biometric

Drones

Games

Augmented Reality

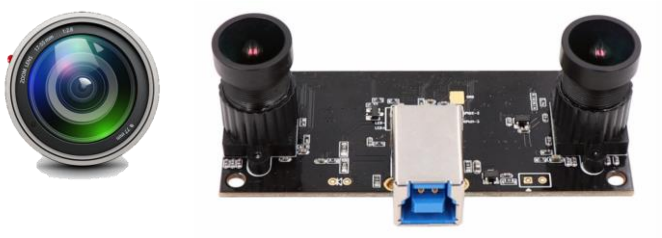
Depth sensing

Active depth sensing

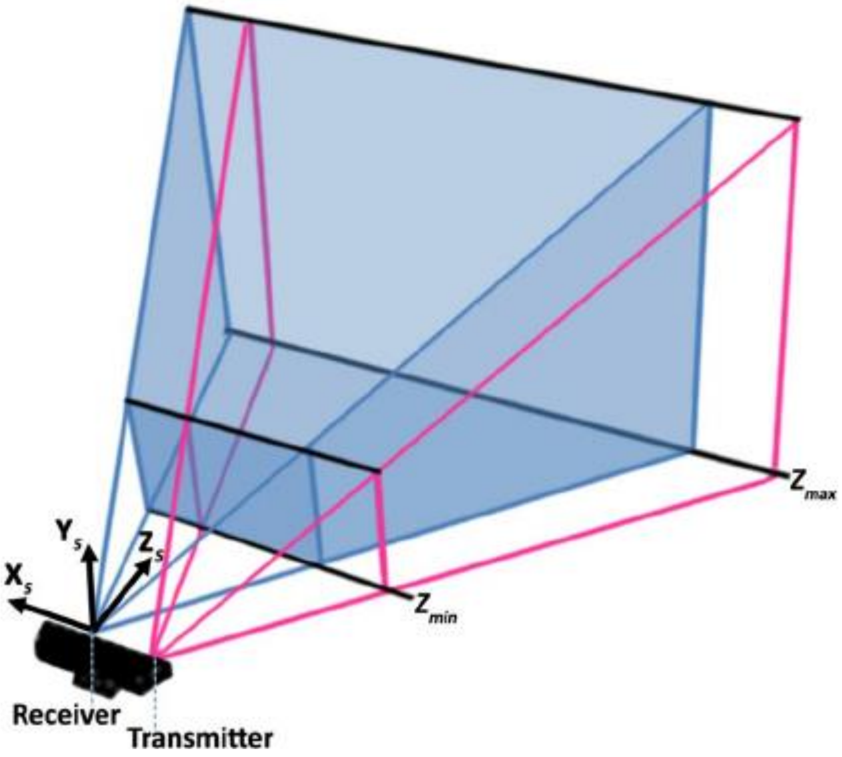
- Structured light (e.g., Kinect 1)
- ToF - Time of Flight (e.g., Kinect 2)
- LiDAR (e.g., Velodyne)

Passive depth sensing

- Binocular stereo
- Monocular
- Multi-view (e.g., Structure For Motion)



Active depth sensing

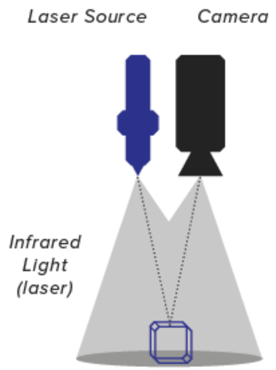


Depth is perceived by perturbing the environment

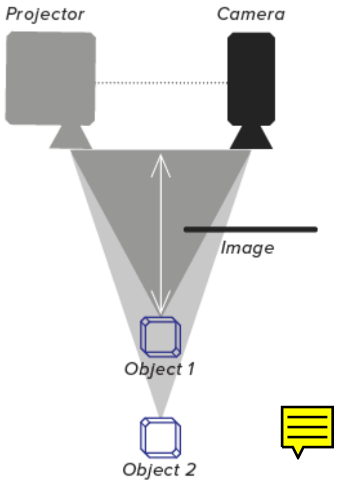
- LiDAR (e.g., Velodyne)
- Time of Flight (e.g., Kinect V2)
- Structured light (e.g., Kinect V1)
- Active stereo (e.g., Intel RealSense)

Examples of Active depth sensing

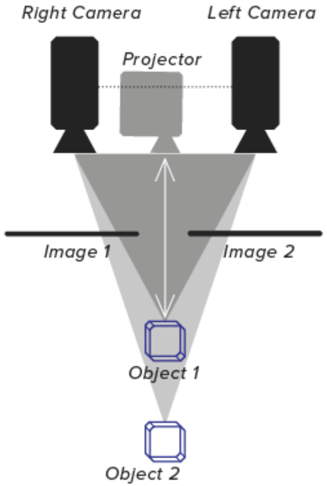
TIME OF FLIGHT



STRUCTURED LIGHT



ACTIVE STEREO

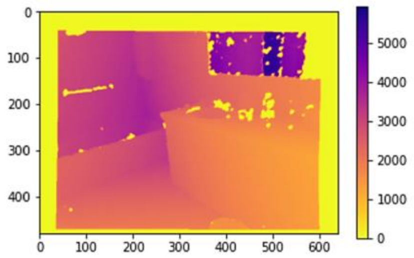


Active depth sensing

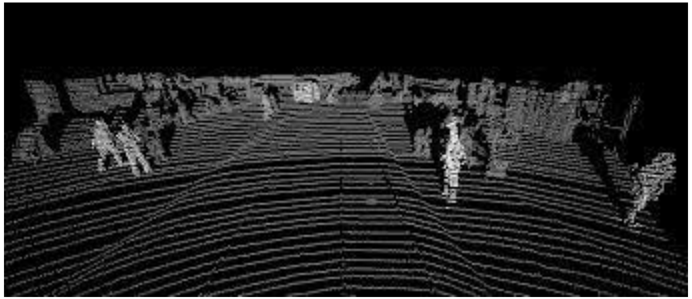
Cons

- Not suited for all environments
- Sometimes really expensive
- Cumbersome
- Not filled depth map
- LiDAR returns a point-cloud

RAW depth map



LiDAR point cloud



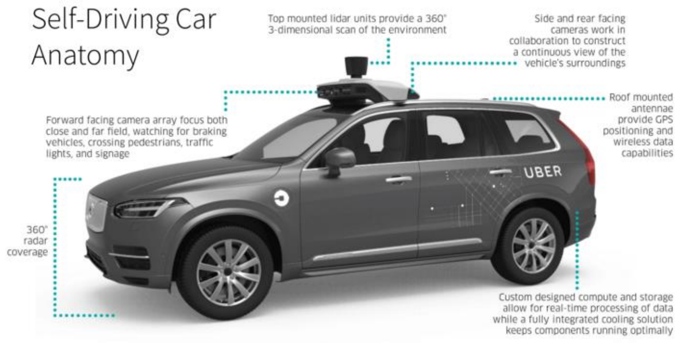
Active depth sensing

Cons

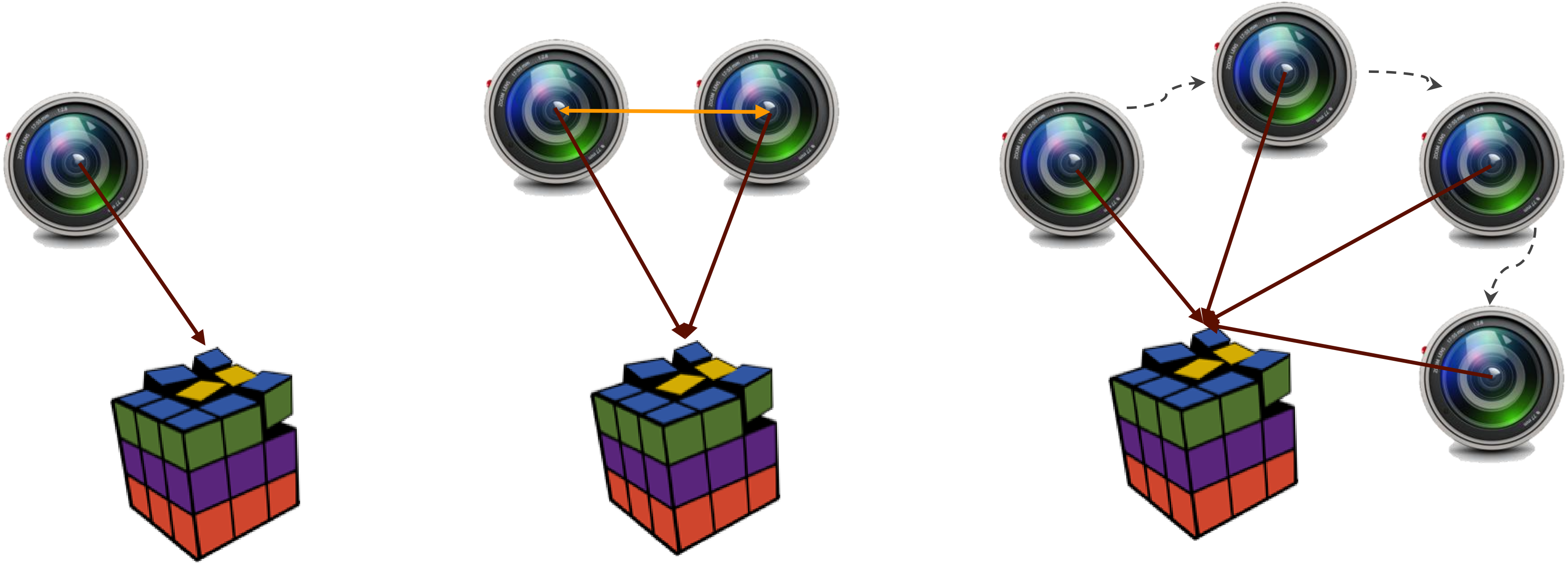
- Not suited for all environments
- Sometimes really expensive
- Cumbersome
- Not filled depth map
- LiDAR returns a point-cloud

Pros

- Very popular
- Used for multiple applications
- Effective depth measurements



Passive depth sensing

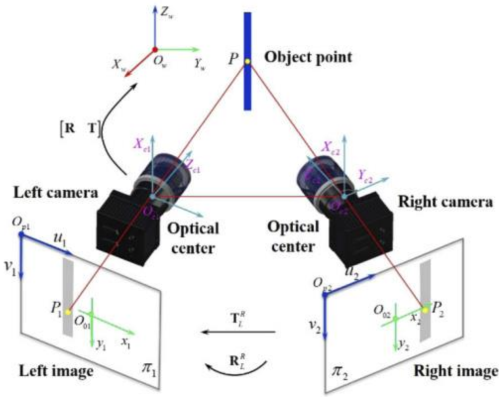


Monocular Binocular stereo Multi-view

Passive depth sensing

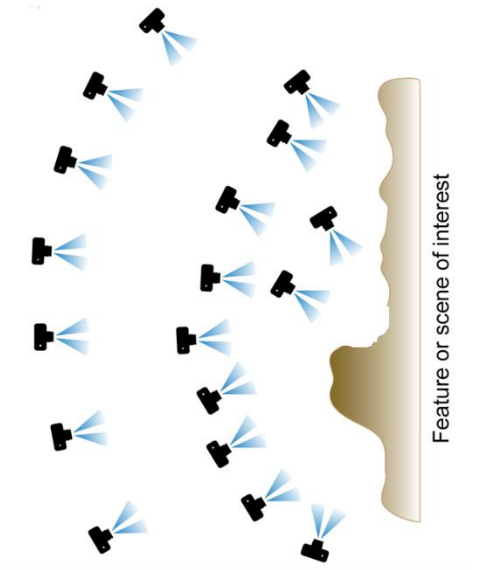
Monocular

Binocular stereo



Multi-view

(Structure For Motion)



Passive depth sensing

Cons

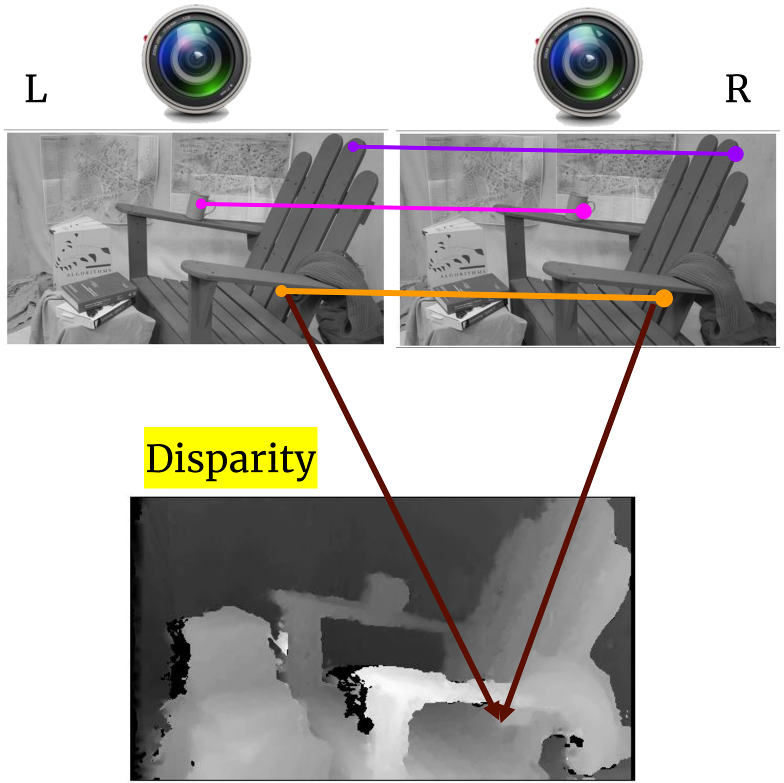
- Complexity is moved to **algorithms!!**
- Depth is reconstructed or estimated

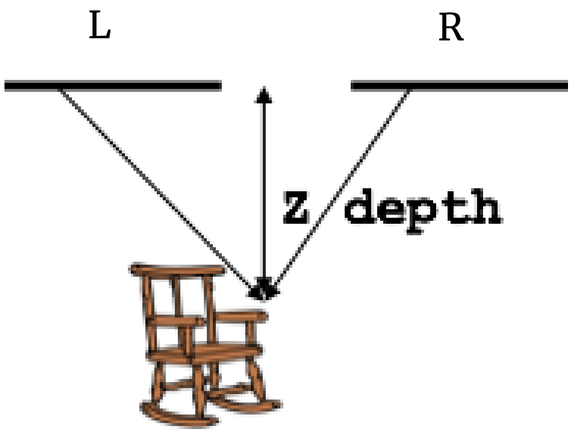
Pros

- Standard cameras, usually cheap, lightweight, fast, etc..
- Suitable for both indoor and outdoor environments

Potentially they can remove all the active sensors issues

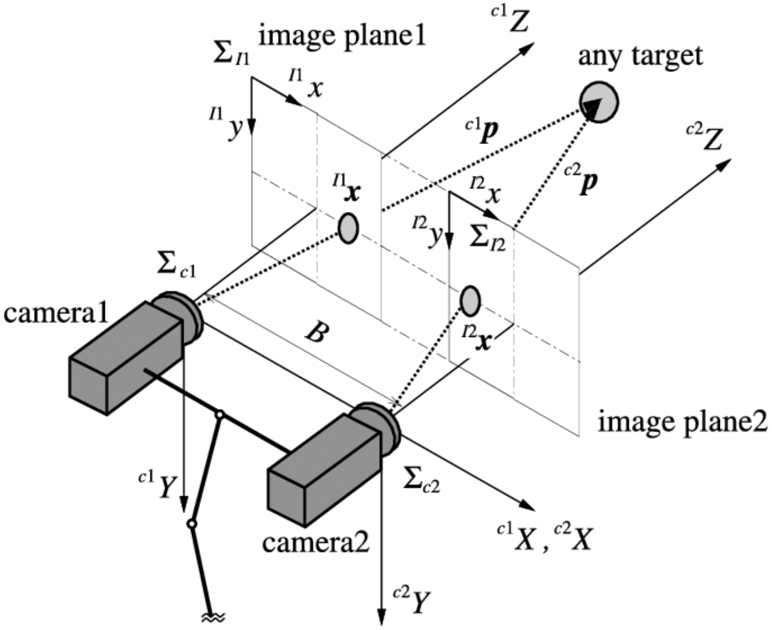
Binocular, Stereo vision





Binocular, Stereo vision

Given two images/cameras, if we are able to find corresponding (homologous) point in the two images we can infer depth by triangulation



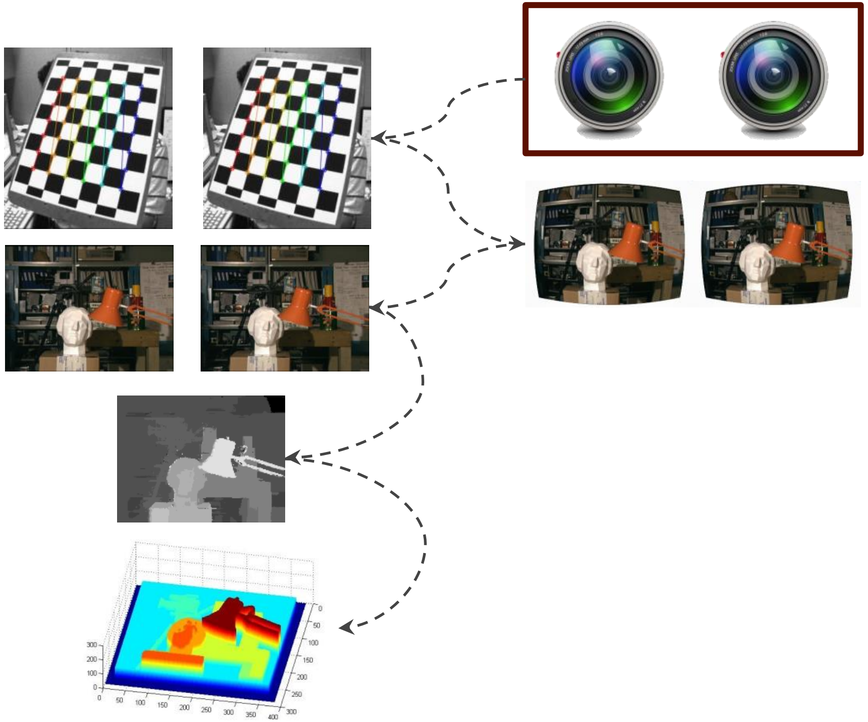
$$D = x(I1) - x(I2) = B*f / Z$$

$$Z = B*f / (x(I1) - x(I2)) = B*f / D$$

Binocular, Stereo vision

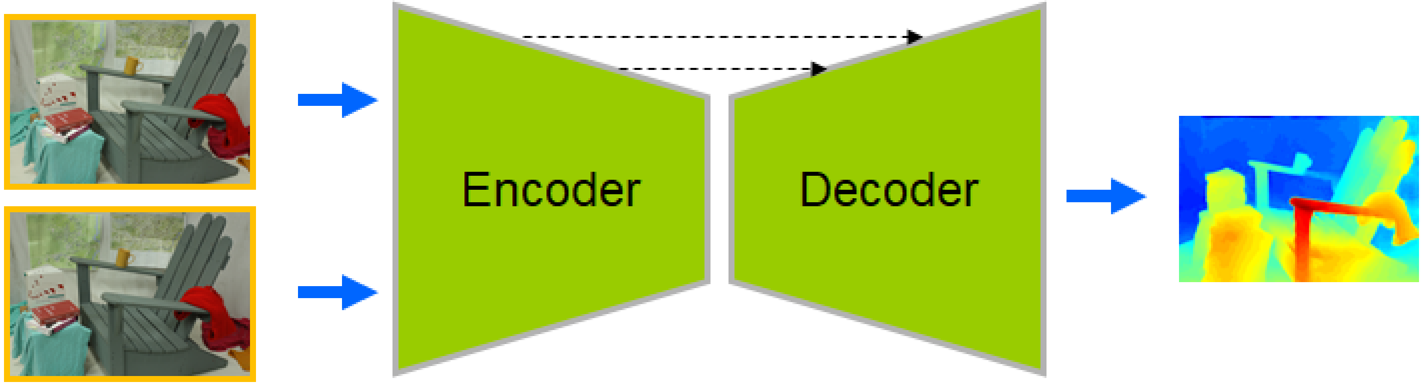
A general Overview

1. Cameras calibration (offline)
2. Rectification
3. Disparity map
4. Depth map



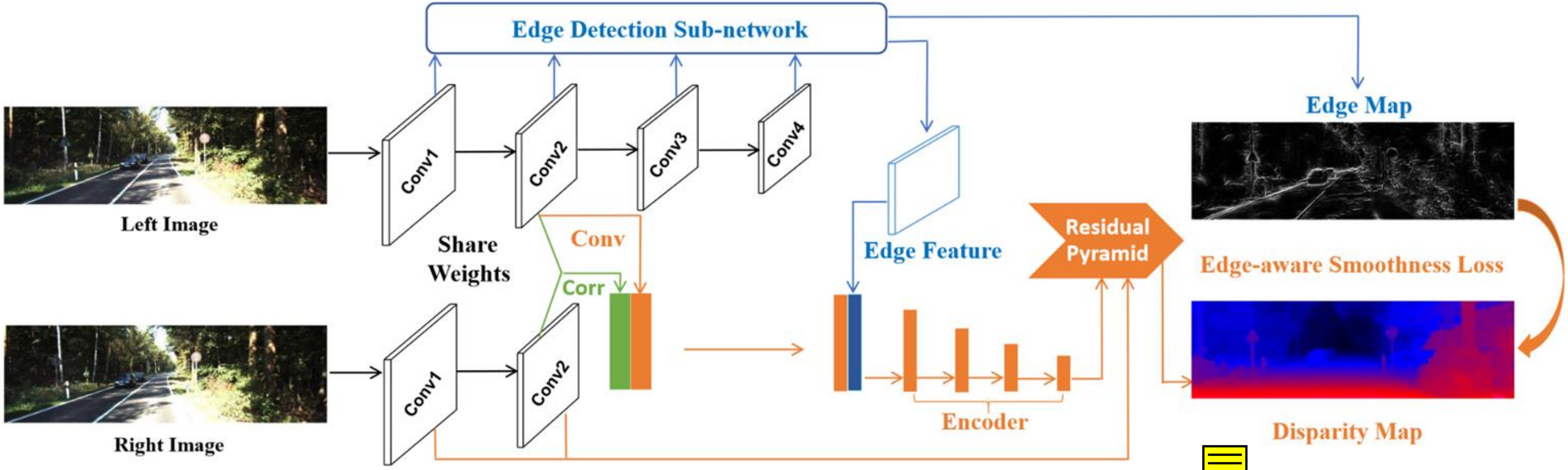
Binocular, Stereo vision

From Stereo-triangulation to Deep-Stereo



Examples (CNN)

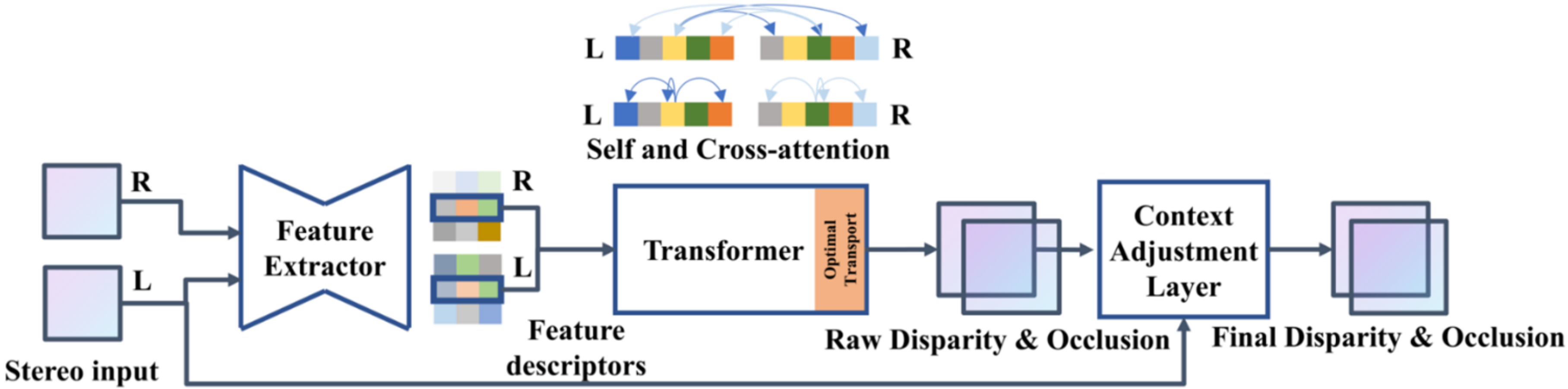
EdgeStereo: An Effective Multi-Task Learning Network for Stereo Matching and Edge Detection



Xiao Song, Xu Zhao, Liangji Fang, and Hanwen Hu. Edgestereo: An effective multi-task learning network for stereo matching and edge detection. arXiv:1903.01700, 2019. [LINK]

Examples (Hybrid ViT)

Revisiting Stereo Depth Estimation From a Sequence-to-Sequence Perspective with Transformers



Zhaoshuo Li, Xingtong Liu, Francis X Creighton, Russell H. Taylor, and Mathias Unberath. Revisiting stereo depth estimation from a sequence-to-sequence perspective with transformers. arXiv preprint, 2020. [LINK]

Monocular Depth Estimation

Motivations

Lightweight Robotic

ADAS

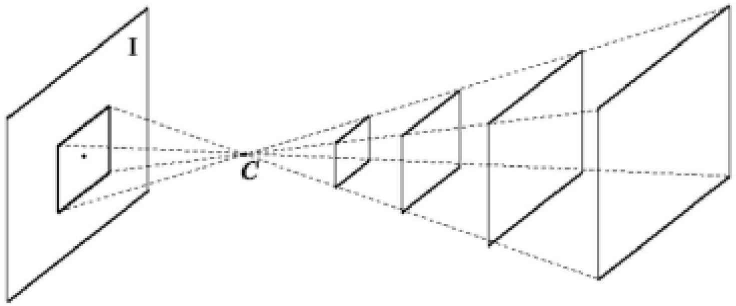
Augmented Reality/Mobile

Monocular Depth Estimation

Problem: Given a single RGB image as input, predict a dense depth map for each pixel

Perspective projection:

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



Monocular Depth Estimation

Problem: Given a single RGB image as input, predict a dense depth map for each pixel

Perspective projection:

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

Depth is an intrinsic information into the 2D space

Monocular Depth Estimation

Meaningful monocular cues:

- Linear Perspective
- Relative Size
- Superimposition
- Texture Gradient

Monocular Depth Estimation

Meaningful monocular cues:

- Linear Perspective
- Relative Size
- Superimposition
- Texture Gradient

... however ... (optical illusions)

Monocular Depth Estimation

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

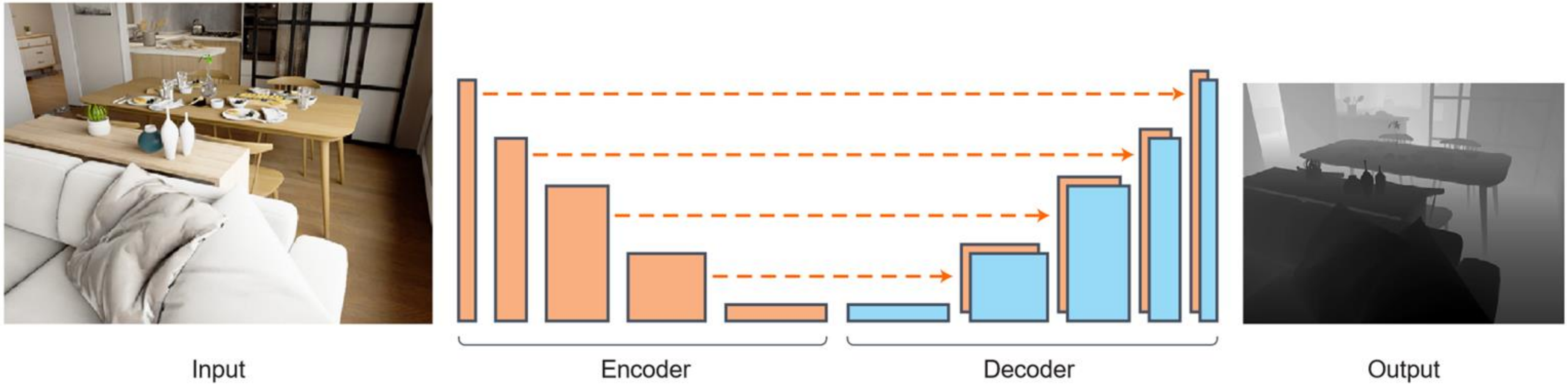
- Sparse ground-truth depth + image reconstruction

Unsupervised

- Image reconstruction (from monocular videos/stereo pairs/stereo sequences)

Examples (CNN)

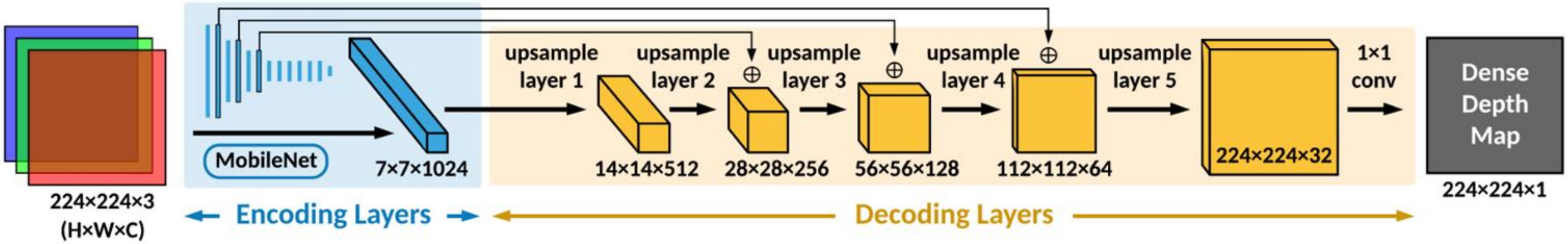
High Quality Monocular Depth Estimation via Transfer Learning



Ibraheem Alhashim and Peter Wonka. High quality monocular depth estimation via transfer learning. arXiv preprint arXiv:1812.11941, 2018.
LINK

Examples (Ligthweigth CNN)

FastDepth: Fast Monocular Depth Estimation on Embedded Systems

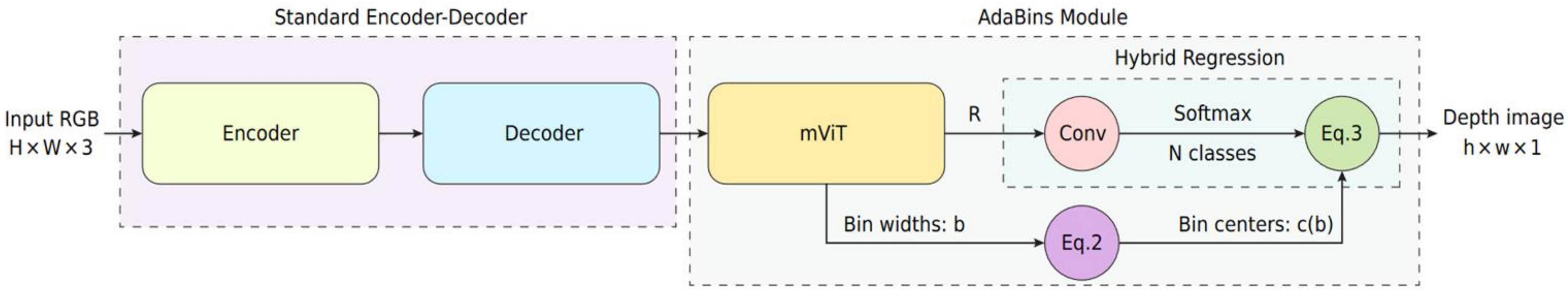


	Before Pruning	After Pruning	Reduction
Weights	3.93M	1.34M	2.9×
MACs	0.74G	0.37G	2.0×
RMSE	0.599	0.604	-
δ_1	0.775	0.771	-
CPU [ms]	66	37	1.8×
GPU [ms]	8.2	5.6	1.5×

Wofk D, Ma F, Yang T J, et al. Fastdepth: Fast monocular depth estimation on embedded systems. In: 2019 International Conference on Robotics and Automation (ICRA). Montreal: IEEE, 2019. 6101–6108 LINK

Examples (Hybrid ViT)

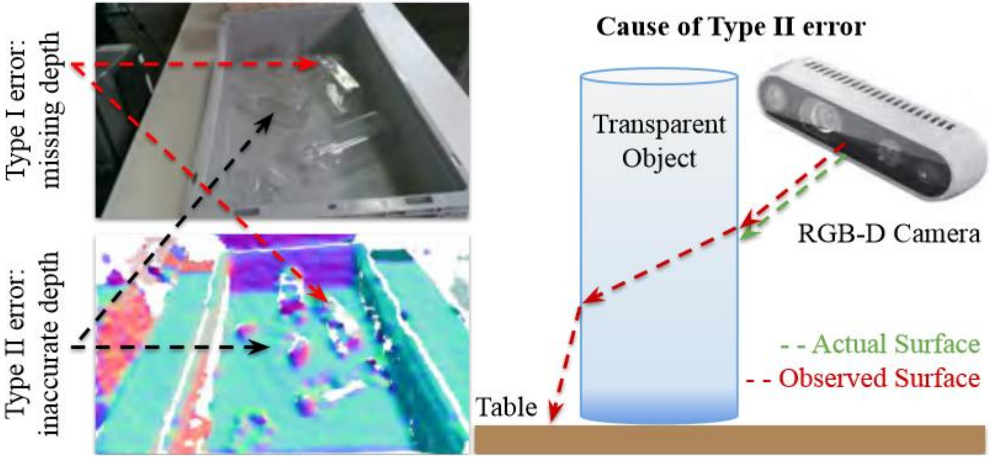
AdaBins: Depth Estimation using Adaptive Bins



Bhat, Shariq Farooq, Ibraheem Alhashim, and Peter Wonka. "Adabins: Depth estimation using adaptive bins." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

Examples (for robotics enthusiast)

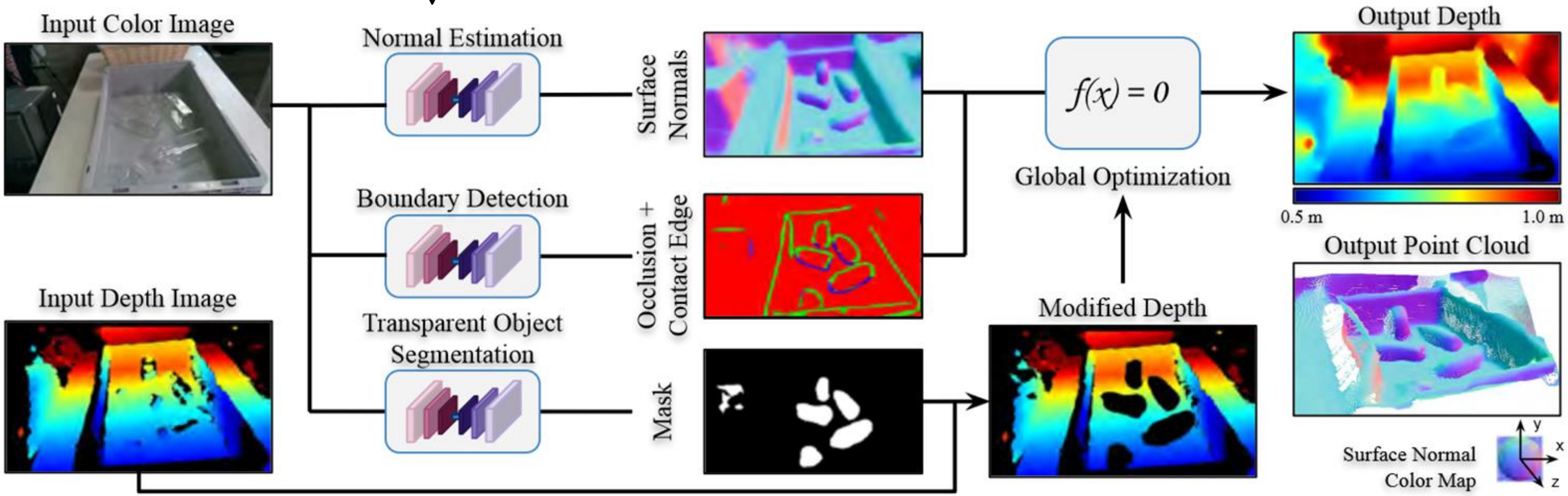
ClearGrasp: 3D Shape Estimation of Transparent Objects for Manipulation



Sajjan SS, Moore M, Pan M, Nagaraja G, Lee J, Zeng A, Song S (2019) Cleargrasp: 3d shape estimation of transparent objects for manipulation. Preprint arXiv:1910.02550. [LINK]

Examples (for robotics enthusiast)

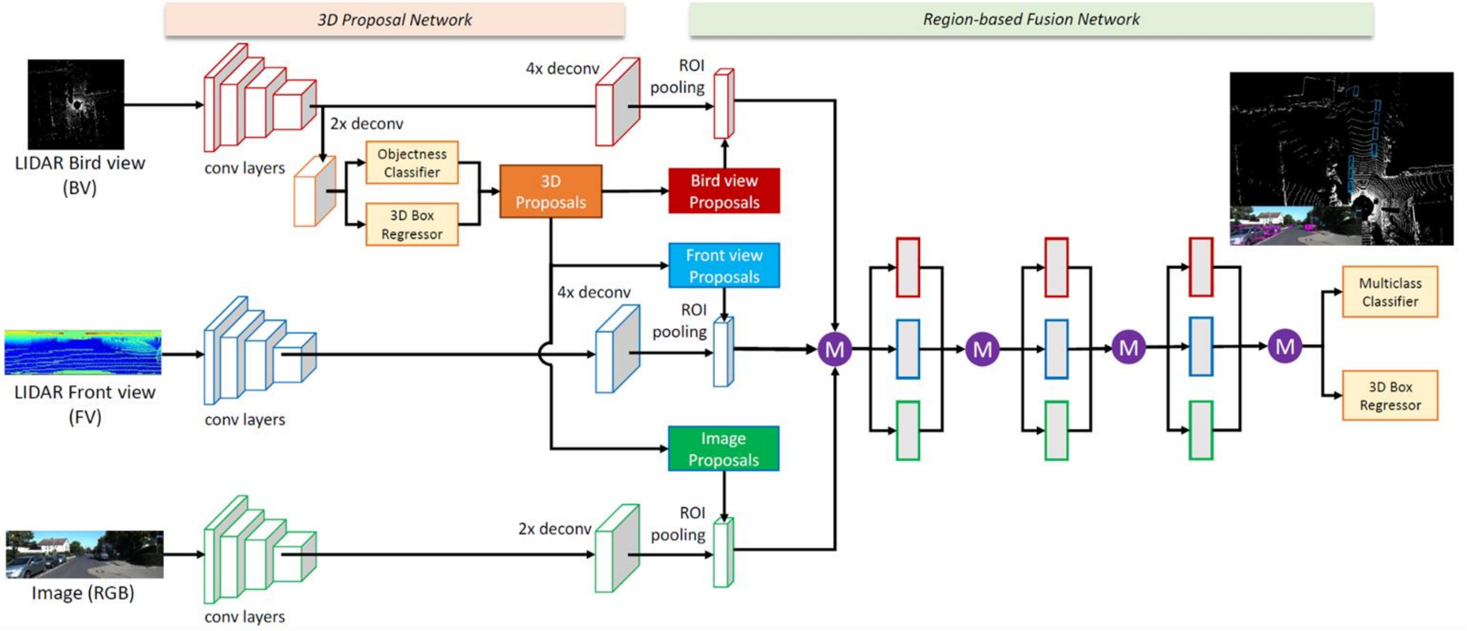
ClearGrasp: 3D Shape Estimation of Transparent Objects for Manipulation



Sajjan SS, Moore M, Pan M, Nagaraja G, Lee J, Zeng A, Song S (2019) Cleargrasp: 3d shape estimation of transparent objects for manipulation. Preprint arXiv:1910.02550. [LINK]

Examples (multi-view)

Multi-View 3D Object Detection Network for Autonomous Driving



X. Chen, H. Ma, J. Wan, B. Li, T. Xia. (2017) Multi-View 3D Object Detection Network for Autonomous Driving. Preprint arXiv:1611.07759.

Metrics

Given a predicted D-map p_i and its Grundtruth g_i:

5 Errors:

- Mean Absolute Error → $$mae = \frac{1}{|P|}\sum_{i \in P} ||p_i - g_i||$$
- Root Mean Squared Error → $$rmse = \sqrt{\frac{1}{|P|}\sum_{i \in P} ||p_i - g_i||^2}$$
- Relative Absolute Error → $$abs_{rel} = \frac{1}{|P|}\sum_{i \in P} \frac{|p_i - g_i|}{g_i}$$
- Logged Errors → log_{mae} & log_{rmse}

Metrics

Given a predicted D-map p_i and its grundtruth g_i:

3 Accuracy:

- Indicate the number of correctly predicted data points out of all the data points

$$d_1 = \frac{1}{|P|} \sum_{i \in P} \max\left(\frac{p_i}{g_i}, \frac{g_i}{p_i}\right) < thr = 1.25$$

$$d_2 = \frac{1}{|P|} \sum_{i \in P} \max\left(\frac{p_i}{g_i}, \frac{g_i}{p_i}\right) < thr = 1.25^2$$

$$d_3 = \frac{1}{|P|} \sum_{i \in P} \max\left(\frac{p_i}{g_i}, \frac{g_i}{p_i}\right) < thr = 1.25^3$$

Datasets

Two main benchmark datasets:

NYU Depth V2

- **Range:** 0.5 – 10 meters
- **Samples:** 50K
- **Type:** depth image

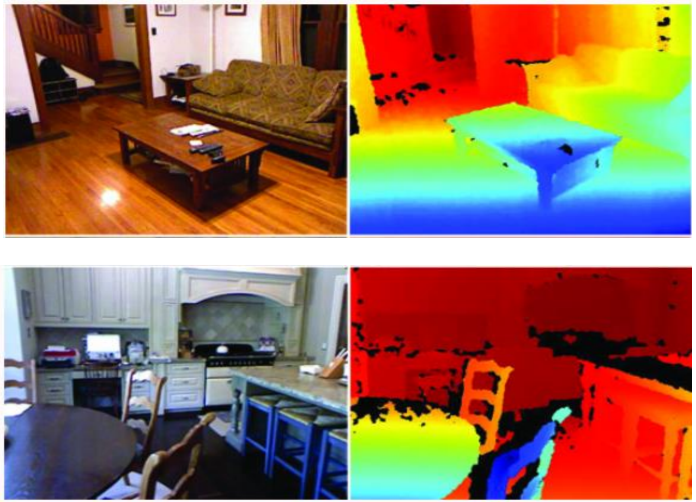
KITTI

- **Range:** 0.9 – 80 meters
- **Samples:** 25K
- **Type:** LiDAR point cloud

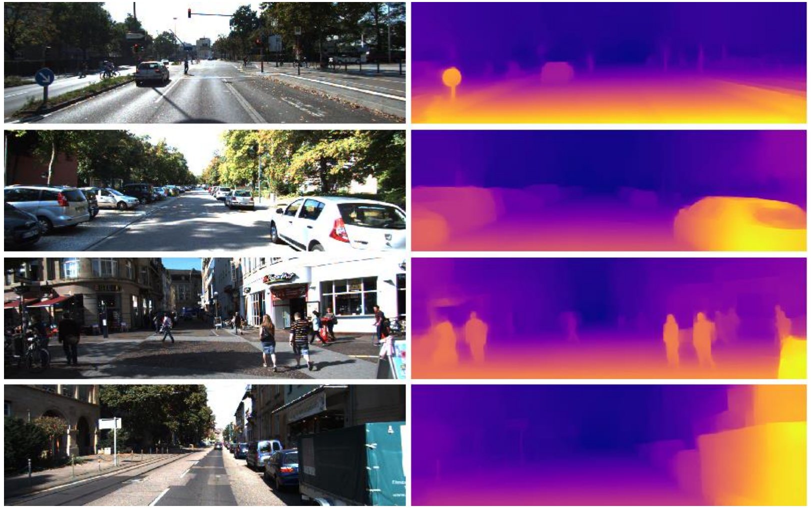
Datasets

Two main benchmark datasets:

NYU Depth V2



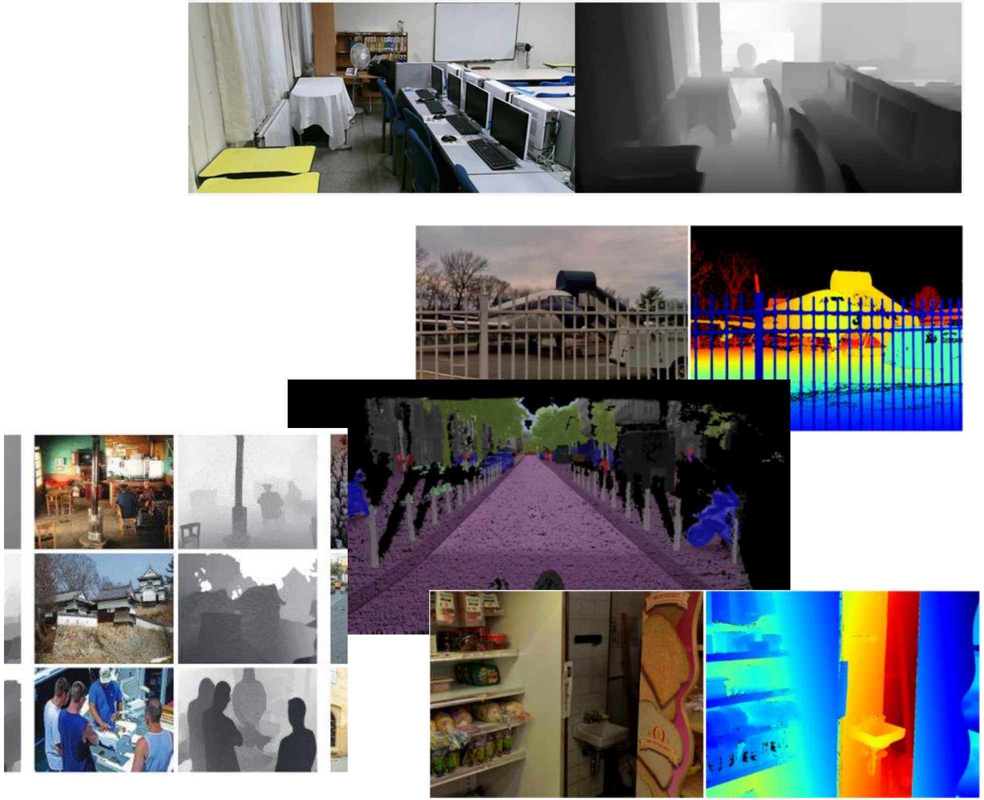
KITTI



Datasets

Other datasets:

- Cityscapes
- SYNTHIA
- Dense Indoor and Outdoor *DEpth*
- DIML/CVL RDB+D
- ReDWeb2018
- YouTube 3D
- Mid-Air
- ...

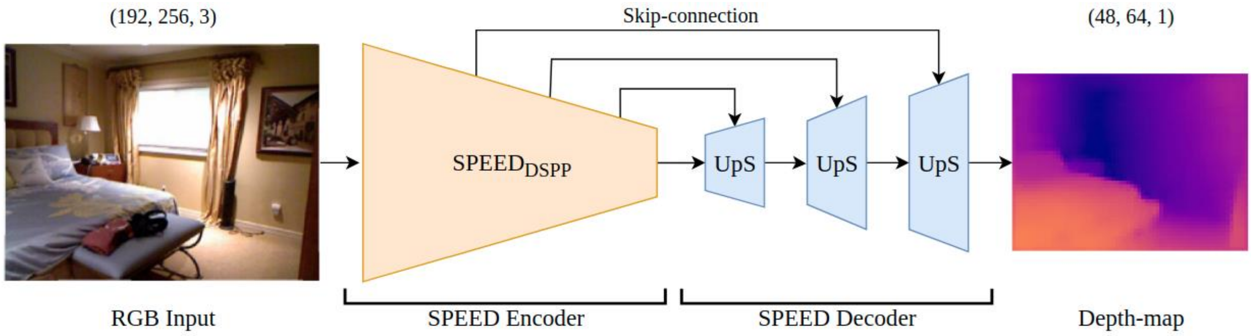


AlcorLab research projects

Research projects: SPEED (CNN)

ALCOR Lab

SPEED: Separable Pyramidal Pooling EncodEr-Decoder for Real-Time Monocular Depth Estimation on Low-Resource Settings

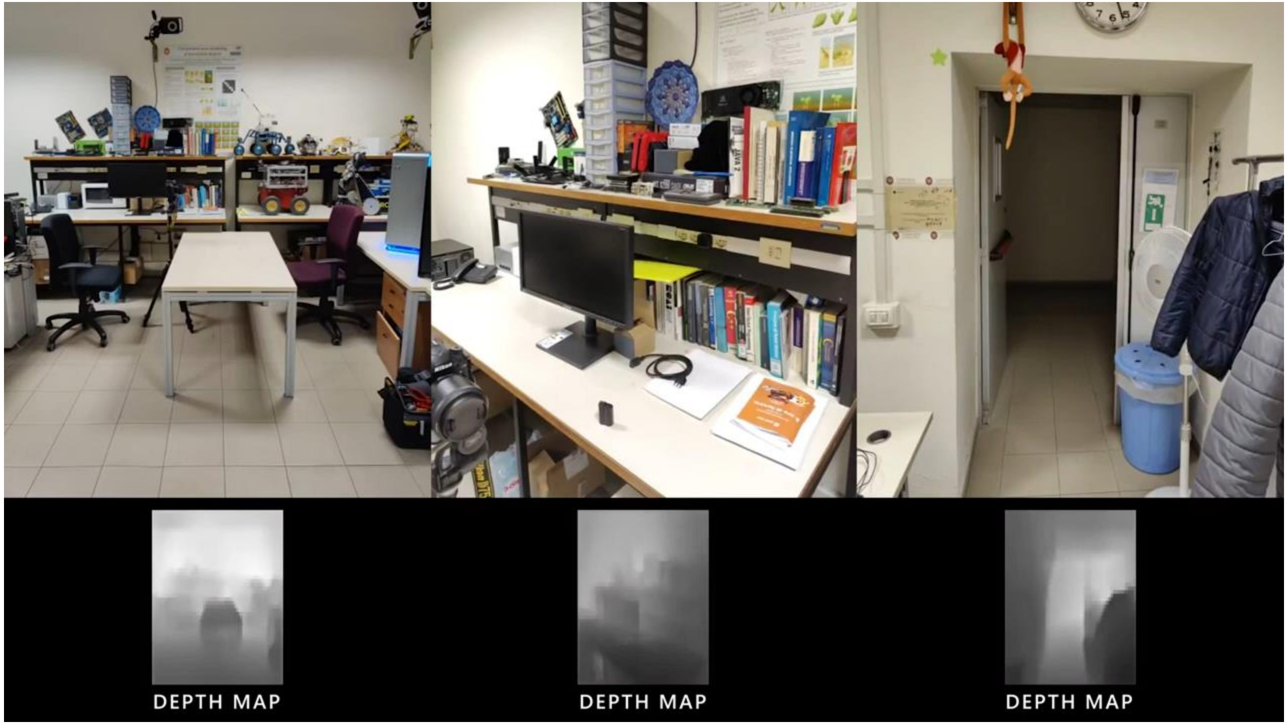


- Novel Depthwise Separable Pyramidal Pooling layers
- Real-Time frequency performances over CPU, TPU workstation and low-power GPU
- Achieve state-of-the-art accuracy performances compared with related works

L. Papa, E. Alati, P. Russo and I. Amerini, "SPEED: Separable Pyramidal Pooling EncodEr-Decoder for Real-Time Monocular Depth Estimation on Low-Resource Settings," in *IEEE Access*, vol. 10, pp. 44881-44890, 2022, doi: 10.1109/ACCESS.2022.3170425.

Research projects: SPEED (CNN)

ALCOR Lab

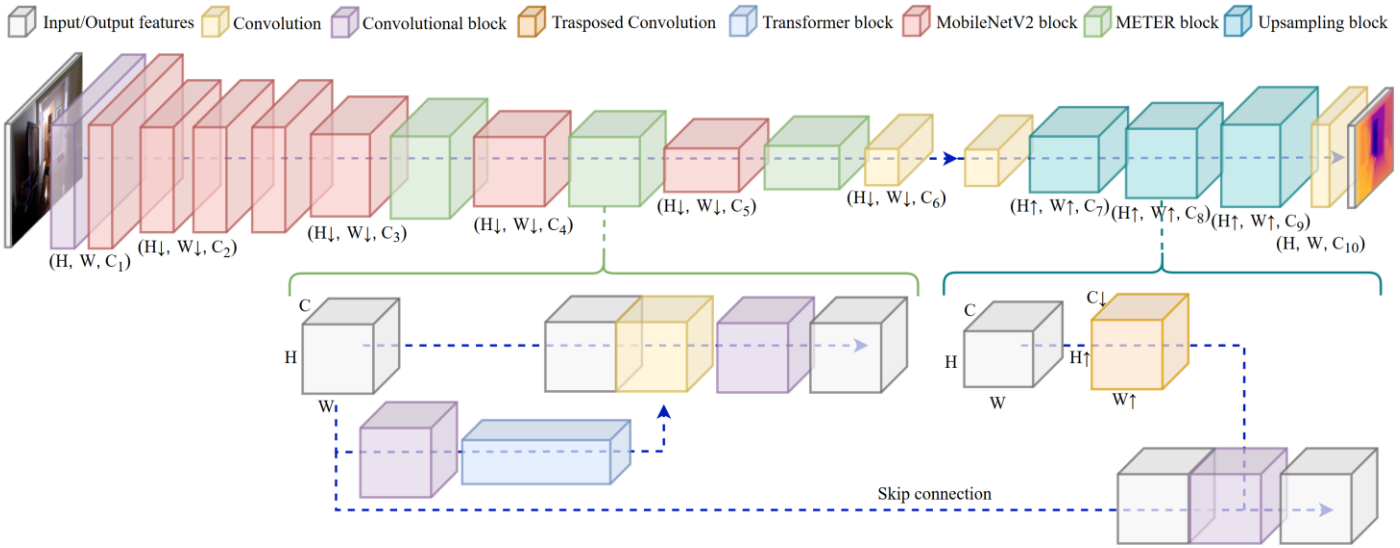


L. Papa, E. Alati, P. Russo and I. Amerini, "SPEED: Separable Pyramidal Pooling EncodEr-Decoder for Real-Time Monocular Depth Estimation on Low-Resource Settings," in *IEEE Access*, vol. 10, pp. 44881-44890, 2022, doi: 10.1109/ACCESS.2022.3170425.

Research projects: METER (Hybrid ViT)

ALCOR Lab

METER: a mobile vision transformer architecture for monocular depth estimation

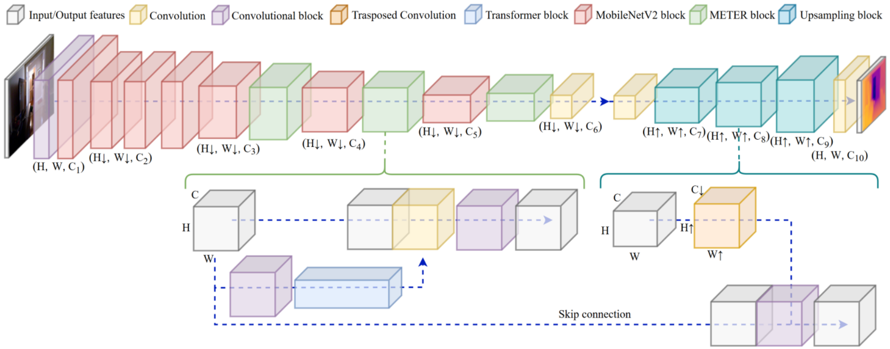


L. Papa, P. Russo and I. Amerini, "METER: a mobile vision transformer architecture for monocular depth estimation," in *IEEE Transactions on Circuits and Systems for Video Technology*, doi: 10.1109/TCSVT.2023.3260310.

Research projects: METER (Hybrid ViT)

ALCOR Lab

METER: a mobile vision transformer architecture for monocular depth estimation

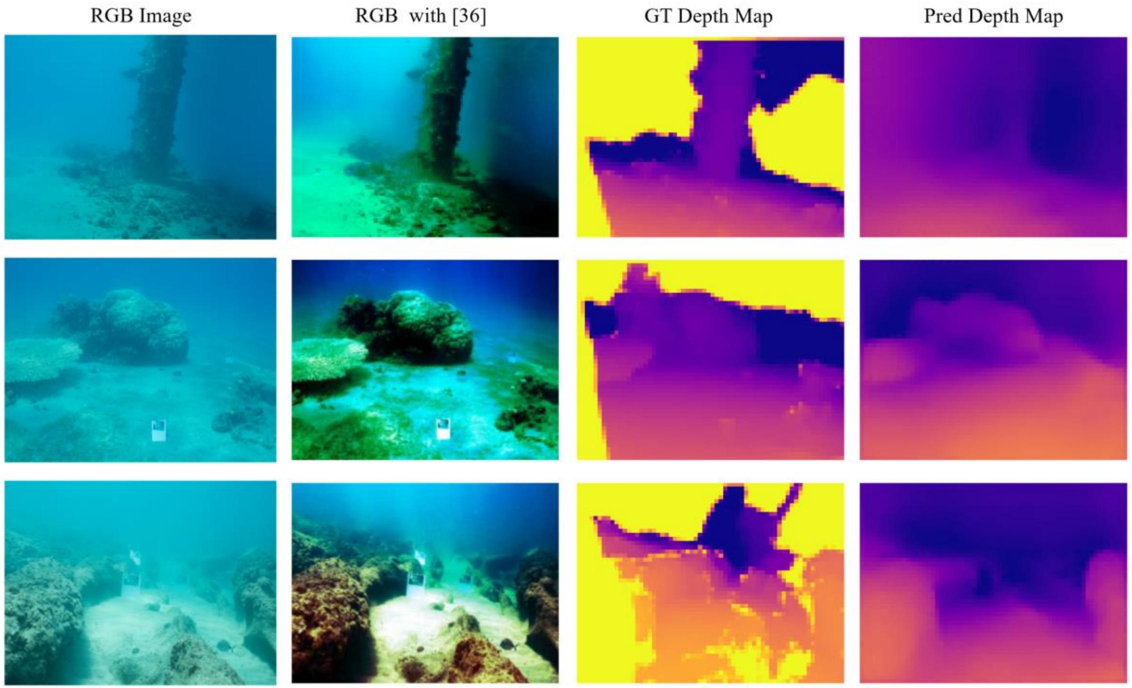


- Novel METER block and different architectures (S, XS, and XXS)
- Balanced Loss function & Specific data augmentation for MDE
- Achieve state-of-the-art accuracy performances compared with related works

L. Papa, P. Russo and I. Amerini, "METER: a mobile vision transformer architecture for monocular depth estimation," in *IEEE Transactions on Circuits and Systems for Video Technology*, doi: 10.1109/TCSVT.2023.3260310.

Research projects: Underwater MDE

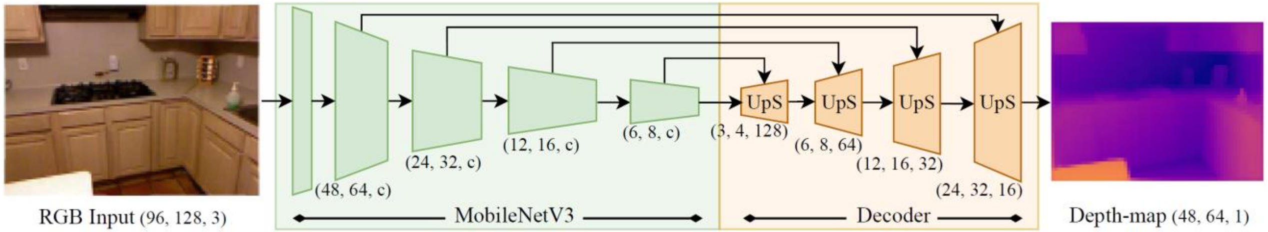
ALCOR Lab

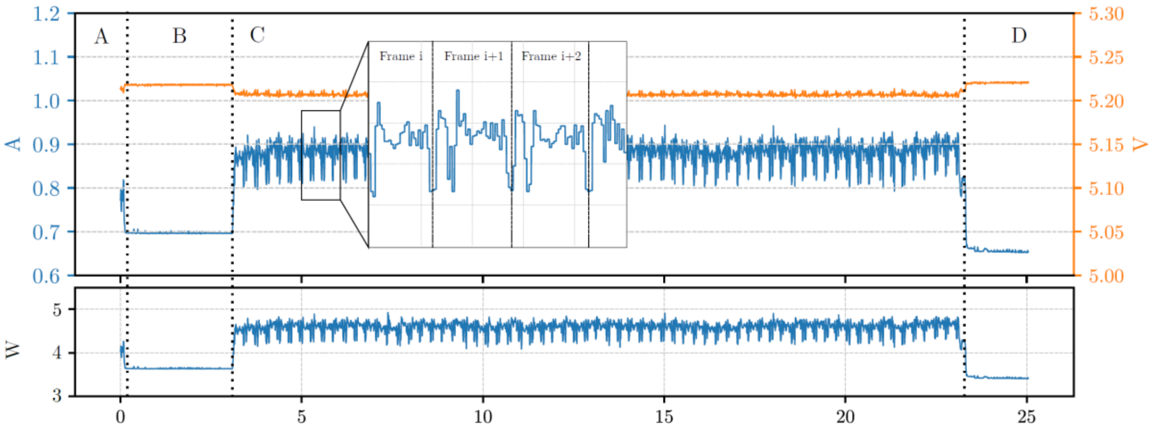


L. Papa, P. Russo and I. Amerini, "Real-time monocular depth estimation on embedded devices: challenges and performances in terrestrial and underwater scenarios," *2022 IEEE International Workshop on Metrology for the Sea; Learning to Measure Sea Health Parameters (MetroSea)*, Milazzo, Italy, 2022, pp. 50-55, doi: 10.1109/MetroSea55331.2022.9950812.

Research projects: Energy-aware MDE

ALCOR Lab



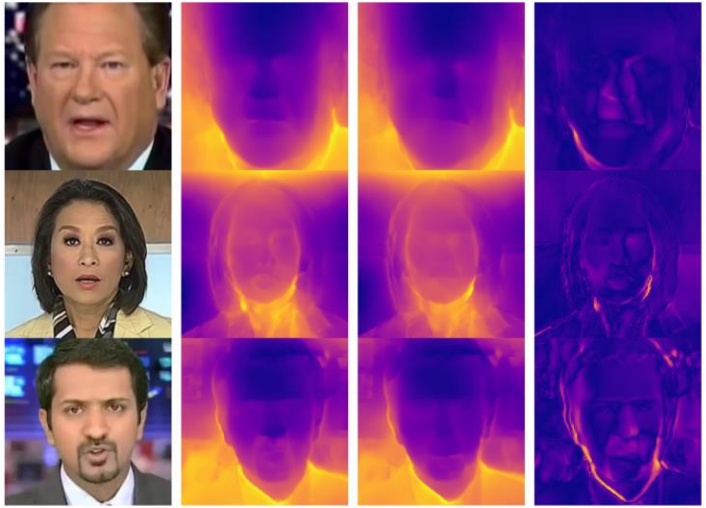


Papa L, Proietti Mattia G, Russo P, Amerini I, Beraldi R. Lightweight and Energy-Aware Monocular Depth Estimation Models for IoT Embedded Devices: Challenges and Performances in Terrestrial and Underwater Scenarios. *Sensors*. 2023; 23(4):2223. https://doi.org/10.3390/s23042223

Research projects: Master thesis

ALCOR Lab

DepthFake



(a) Original face (b) Real depth (c) Fake depth (d) Difference

From point-cloud to 3D mesh

L. Maiano, L. Papa, K. Vocaj and I. Amerini, "DepthFake: a depth-based strategy for detecting Deepfake videos," in Workshop on Artificial Intelligence for Multimedia Forensics and Disinformation Detection (AI4MFDD) at ICPR, 2022, [Accepted]

Open challenges

Promising research directions:

- Domain adaptation / Transferability: Synthetic to real scenarios
- Lightweight / energy-aware networks for mobile / edge applications
- Learning efficient techniques: pruning, knowledge distillation, and quantization
- Temporal consistency: improve the estimation with sequence of predictions
- Multimodal learning (RGB +D)
- 3D mesh construction / From point cloud to filled depth

- ... and many others ...