DEEP VISION CONSULTING

A journey on 3D vision in the deep learning era

(without forgetting industrial exploitation)

Giovanni Gualdi, CEO
giovanni.gualdi@deepvisionconsulting.com

6 Marzo 2024

from challenge to success in computer vision

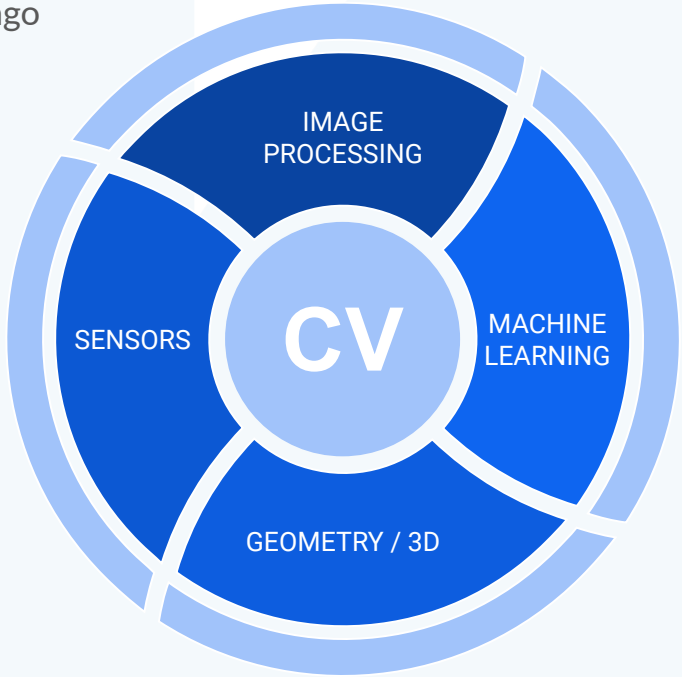
since 2011

A journey on 3D vision in the deep learning era

A journey on ~~3D~~ vision in the deep learning era

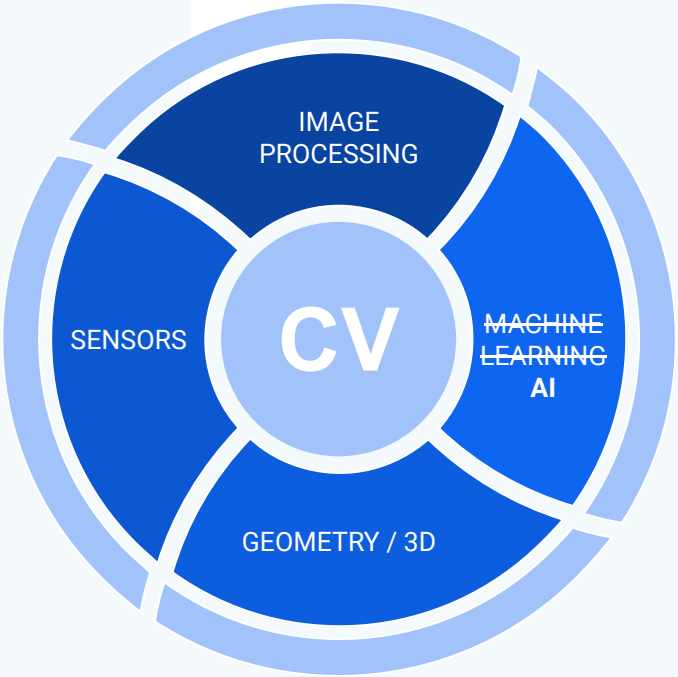
The computer vision paradigm

Paradigm valid until 10 yrs ago



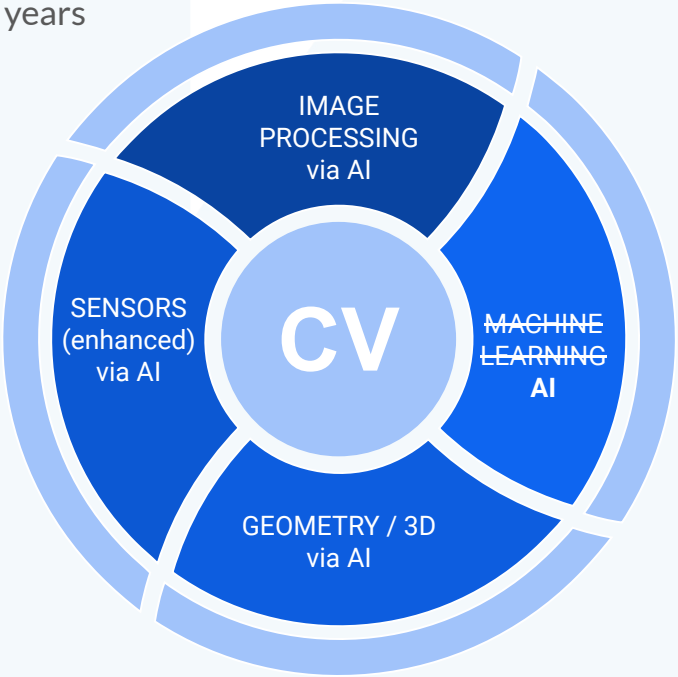
The computer vision paradigm

Valid in the last decade
by the introduction of
Deep Learning



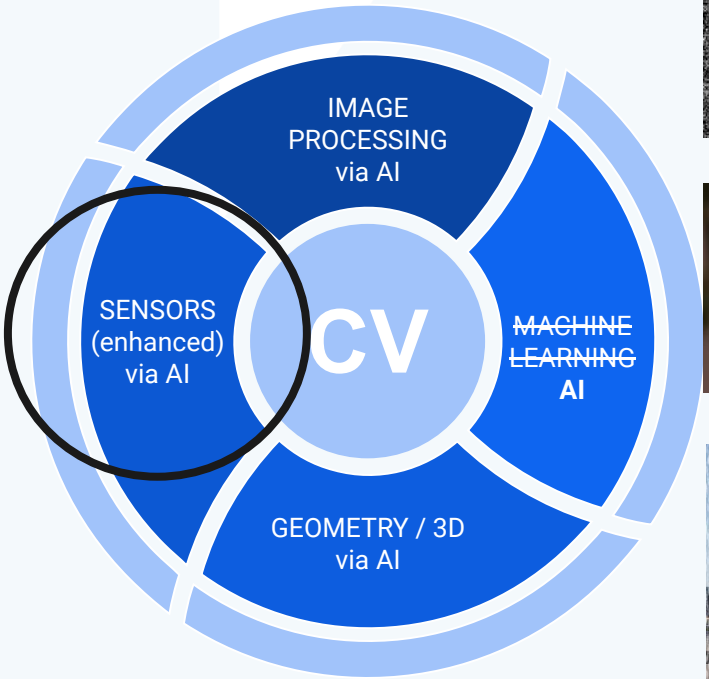
The computer vision paradigm

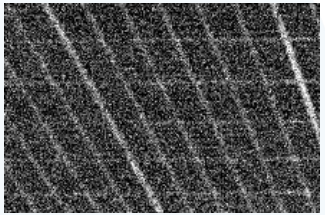
Trend started in the last few years

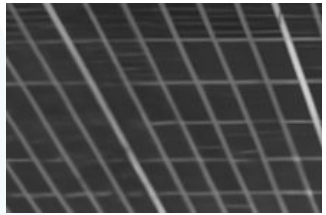


The computer vision paradigm

Denoising, deblurring,
image restoration,
tonemapping, super-res, ...

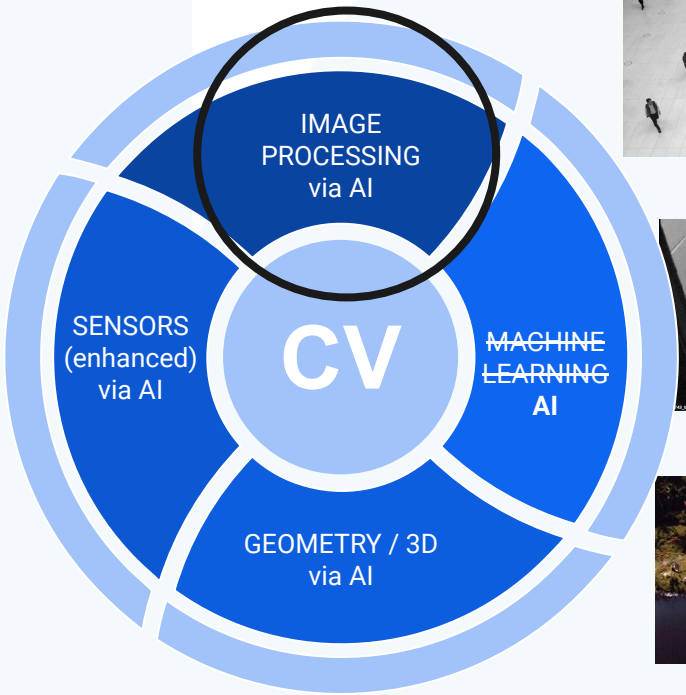


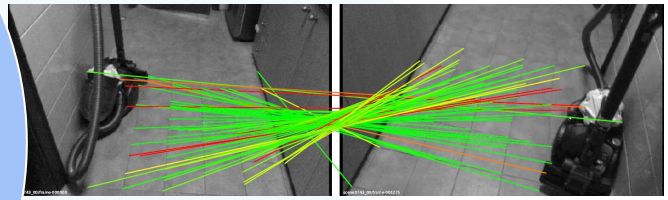


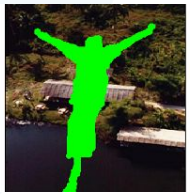


The computer vision paradigm

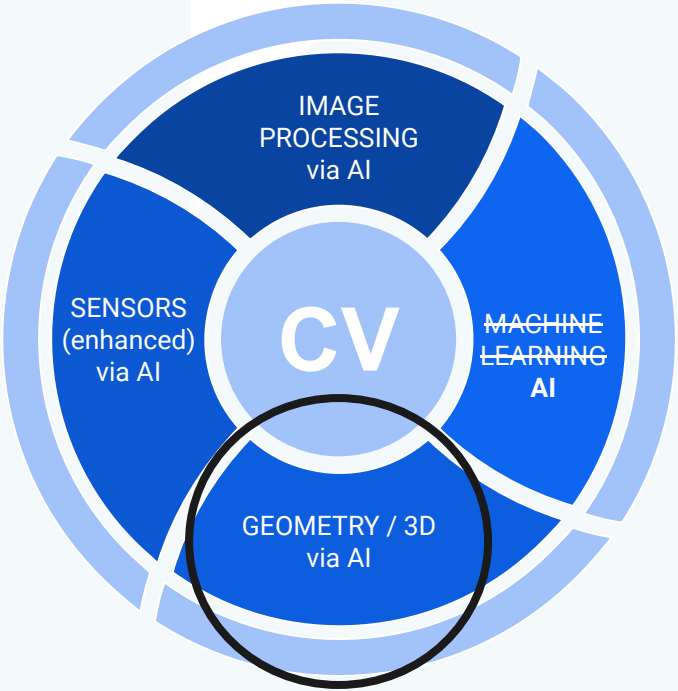
Optical flow, keypoint extraction and matching, inpainting…



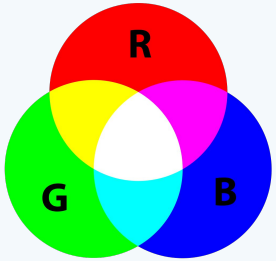




What about 3D?



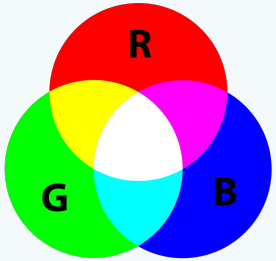
computer vision in the **visible spectrum**



computer vision in the **visible spectrum**

most common tasks tackled with computer vision

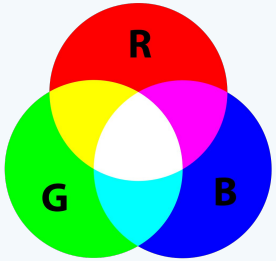
- classification
- detection
- segmentation
- recognition / re-id



computer vision in the **visible spectrum**

most common tasks tackled with computer vision

- classification
- detection
- segmentation
- recognition / re-id
  - action
  - tracking



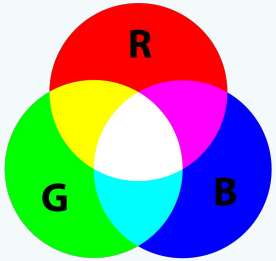
+ TIME

computer vision in the **visible spectrum**

most common tasks tackled with computer vision

- classification
- detection
- segmentation
- recognition / re-id
  - action
  - tracking

what
where
who
how



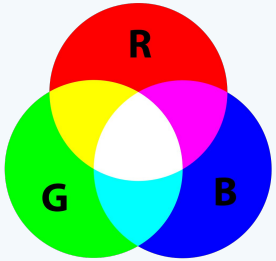
+ TIME

computer vision in the **visible spectrum**

most common tasks tackled with computer vision

- classification
- detection
- segmentation
- recognition / re-id
  - action
  - tracking

what
where
who
how



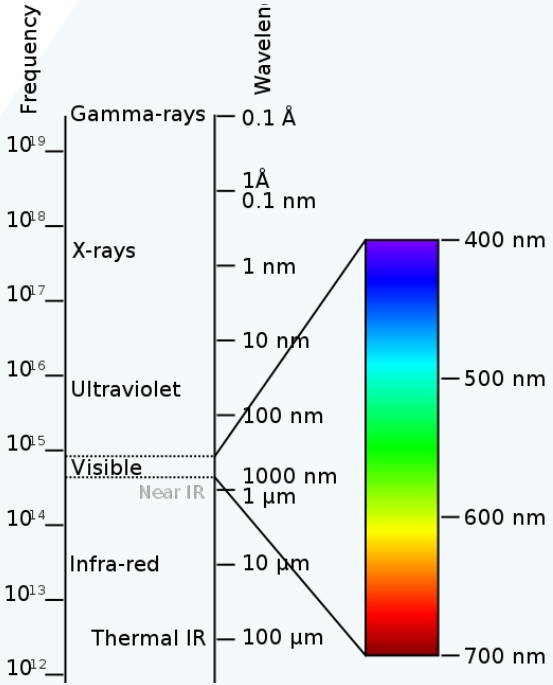
+ TIME

RGB is not enough for these tasks in all use cases
no matter how powerful is your AI behind

alternatives to the visible spectrum

other spectrums

- x ray
- far ir / thermal
- near ir
- uv



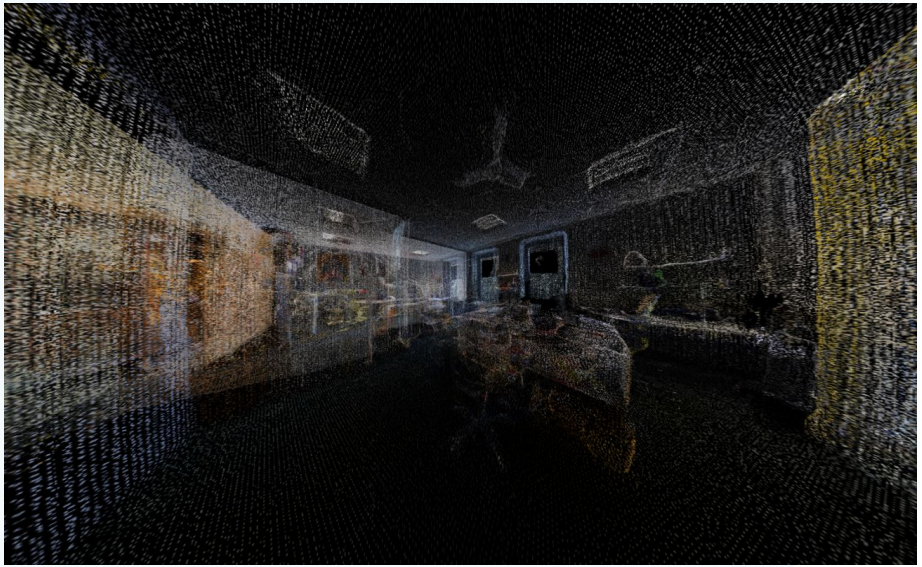
alternatives to the visible spectrum

other spectrums

- x ray
- far ir / thermal
- near ir
- uv

different domains

- 3D



alternatives to the visible spectrum

other spectrums

- x ray
- far ir / thermal
- near ir
- uv

different domains

- 3D
- events

Conventional sensor

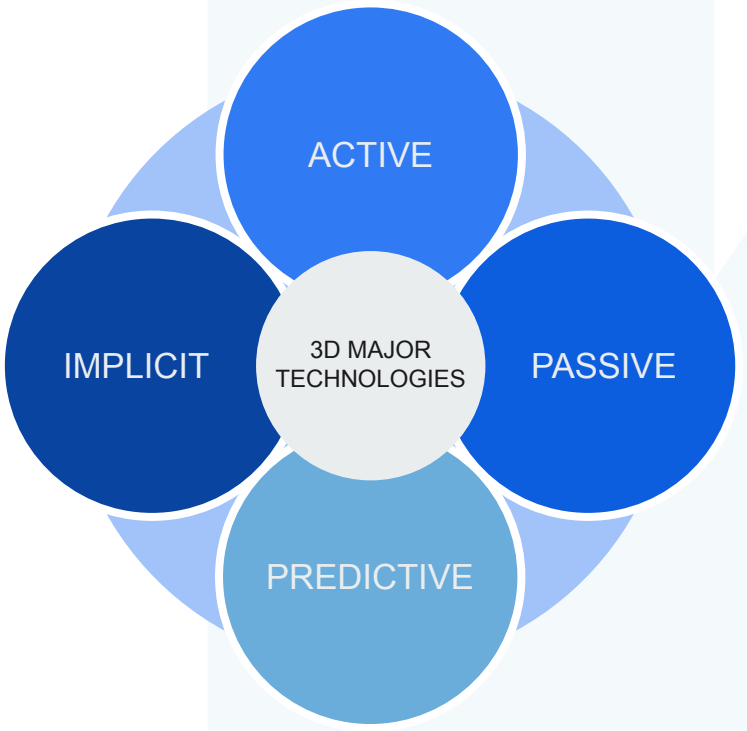
Event-based sensor

a few industries and use cases where **3D is relevant**

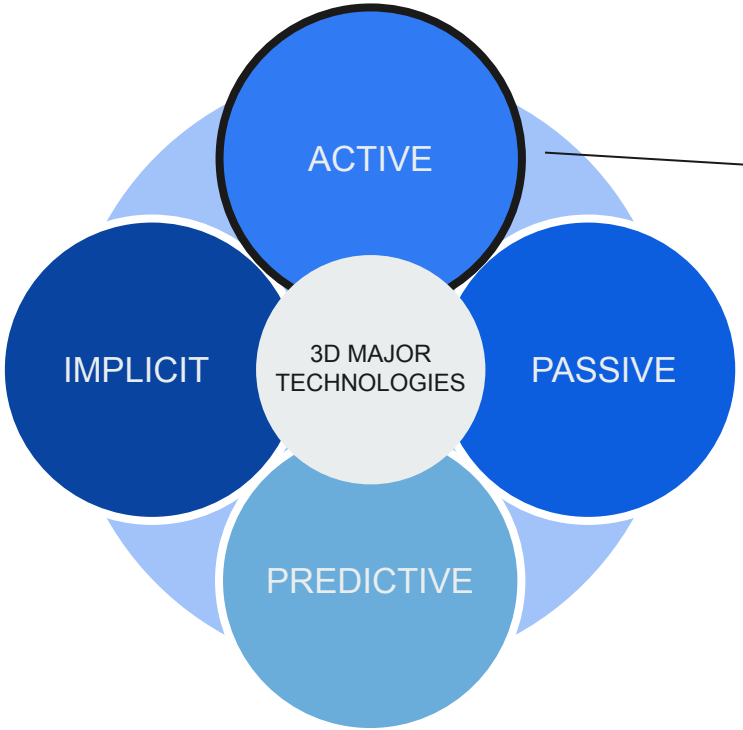
safety	collision avoidance
security	face reco/verification
fashion	data augmentation
autonomous vehicles	agv/mowers/service robots
industrial robots	bin picking
logistics	packing optimization
agri / livestock	measure volumes/growth/vegetation

and many others

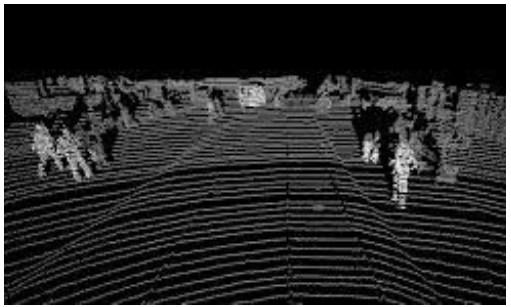
review of **3D sensor technologies**

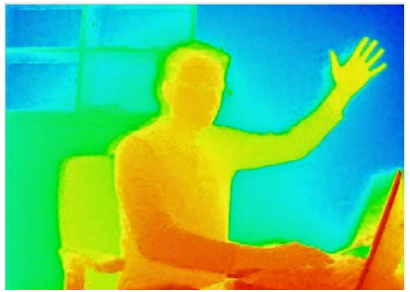


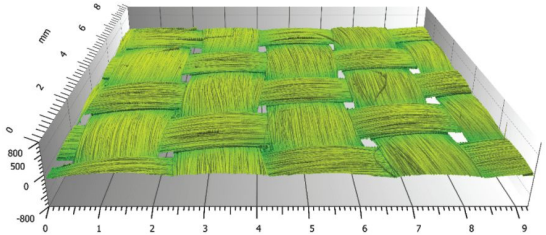
review of **3D sensor technologies**



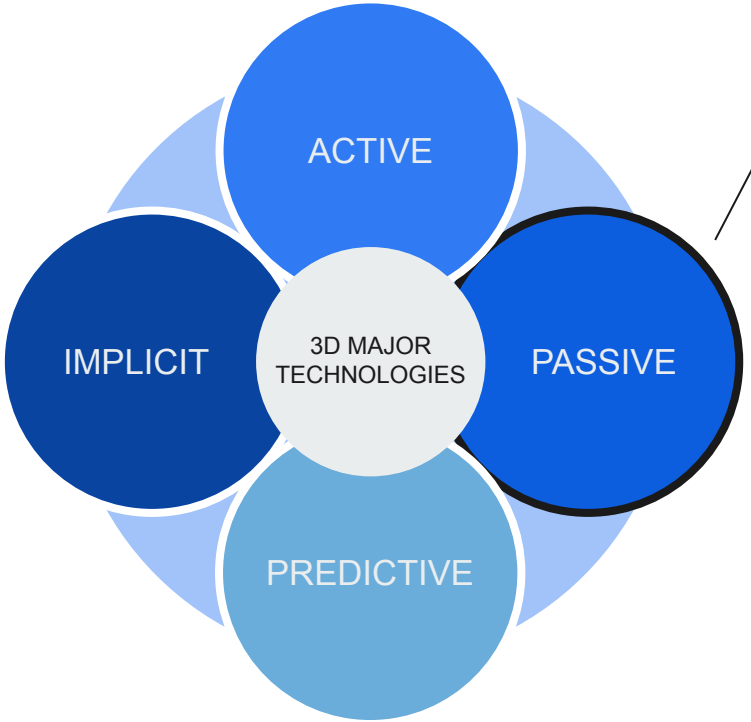
- lidar
- tof
- structured light
- photometric stereo
- profilometers
- sonar
- radar



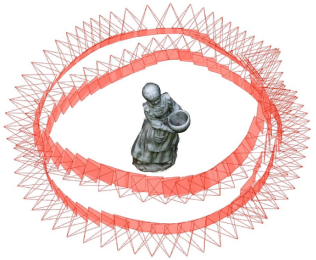


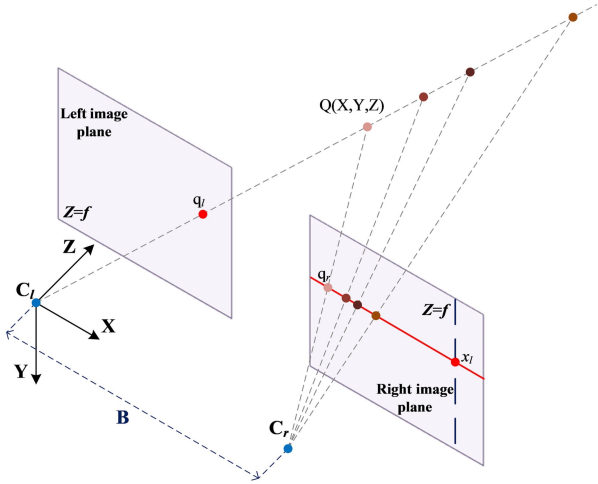


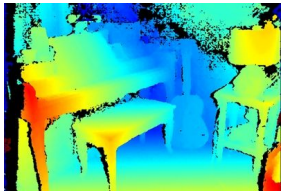
review of **3D sensor technologies**



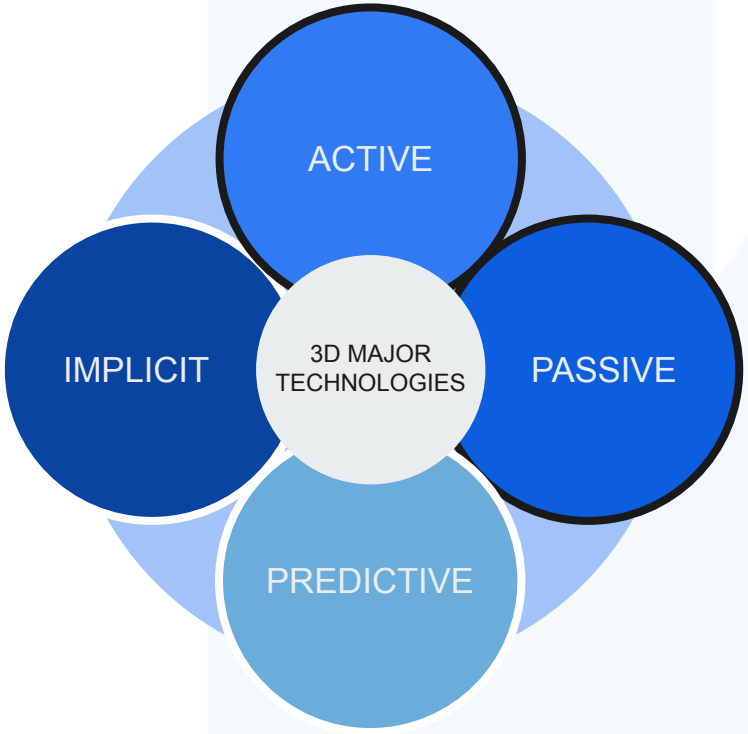
- stereo camera triangulation
- structure from motion
- depth from focus/defocus





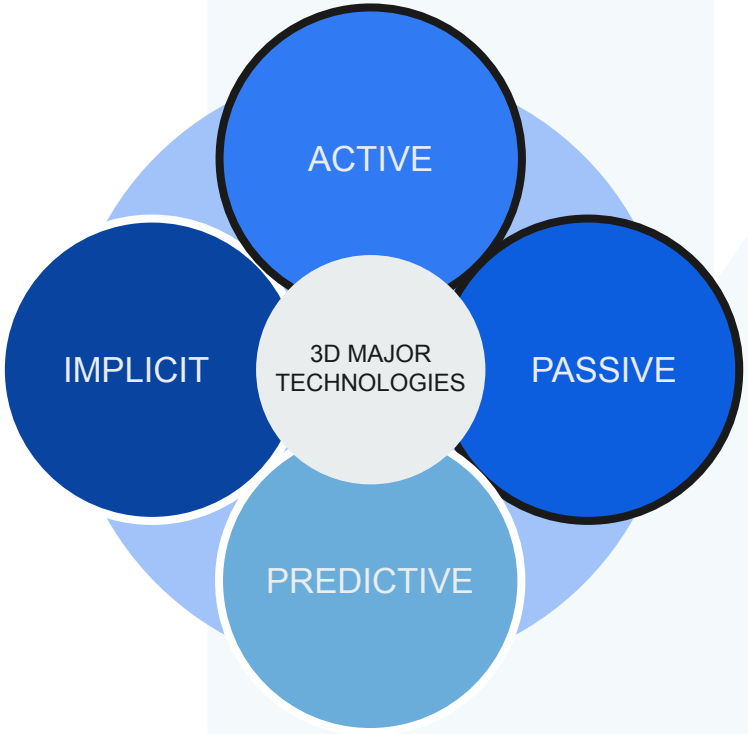


review of **3D sensor technologies**



by exploiting properties of
physics, geometry and optics
active and passive 3D sensors
measure distance
and provide
metric values of x,y,z
for each sensed point

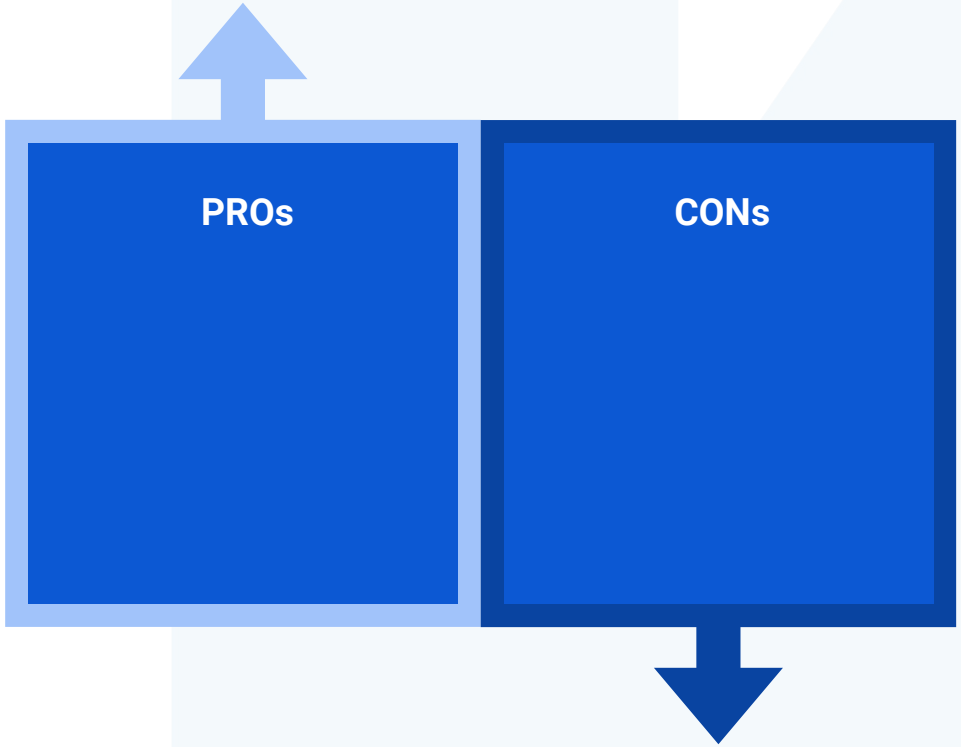
review of **3D sensor technologies**



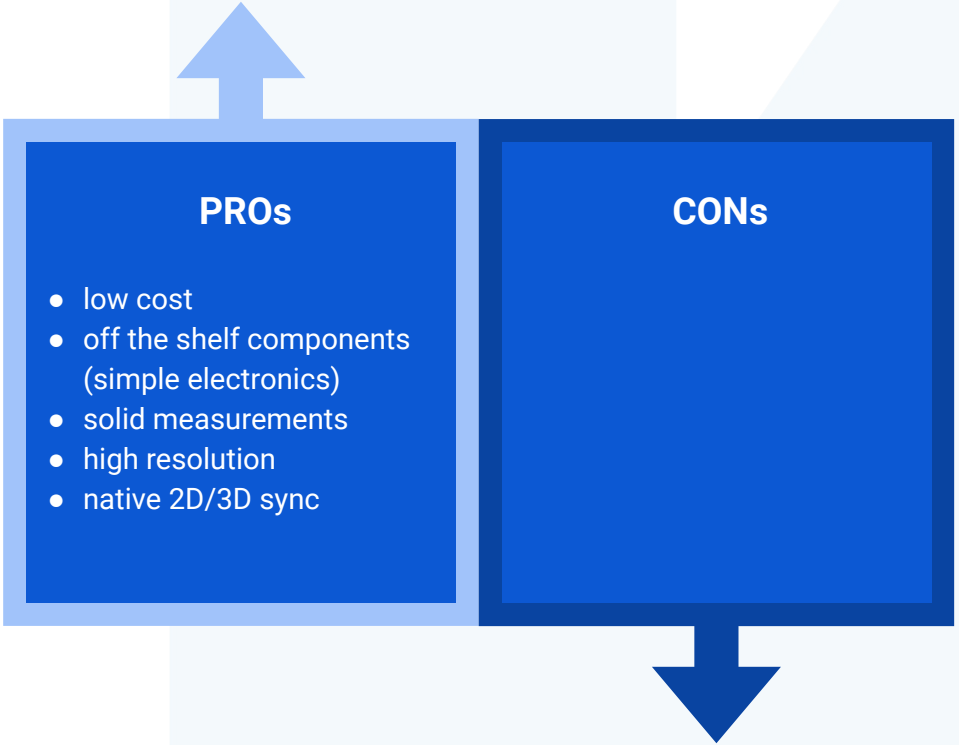
by exploiting properties of
physics, geometry and optics
active and passive 3D sensors
measure distance
and provide
metric values of x,y,z
for each sensed point

what can AI do for these technologies?

on **stereo cameras**



on **stereo cameras**



on **stereo cameras**

PROs

- low cost
- off the shelf components (simple electronics)
- solid measurements
- high resolution
- native 2D/3D sync

CONs

- mid/low range
- sensitive to mechanical or thermal stress
- risk for sparse 3D reconstruction
- computationally intensive

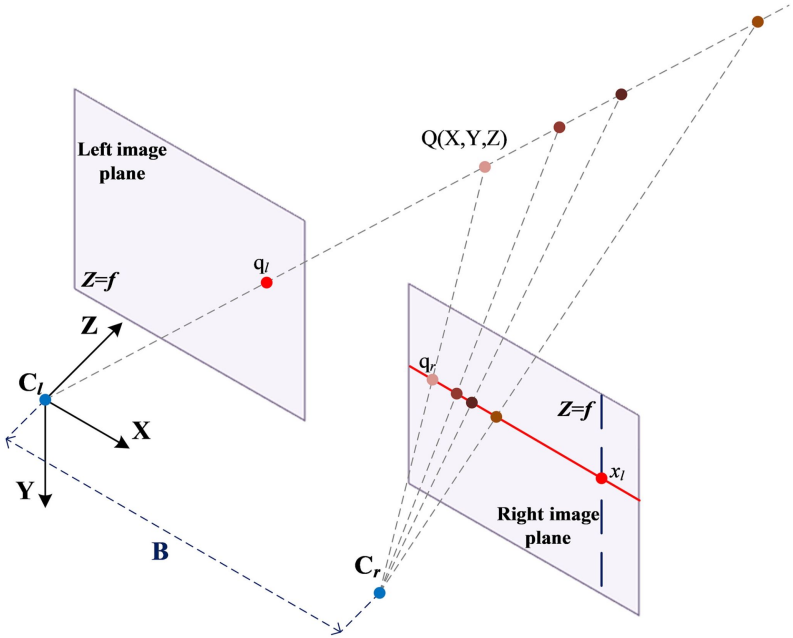
on **stereo cameras**

PROs

- low cost
- off the shelf components (simple electronics)
- solid measurements
- high resolution
- native 2D/3D sync

CONs

- mid/low range
- sensitive to mechanical or thermal stress
- risk for sparse 3D reconstruction
- computationally intensive



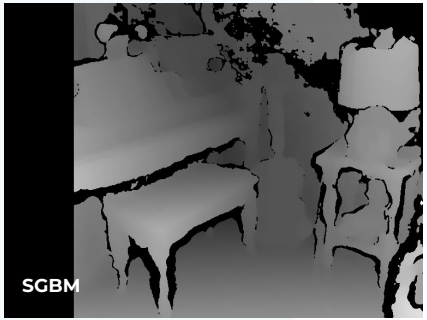
on **stereo cameras**

PROs

- low cost
- off the shelf components (simple electronics)
- solid measurements
- high resolution
- native 2D/3D sync

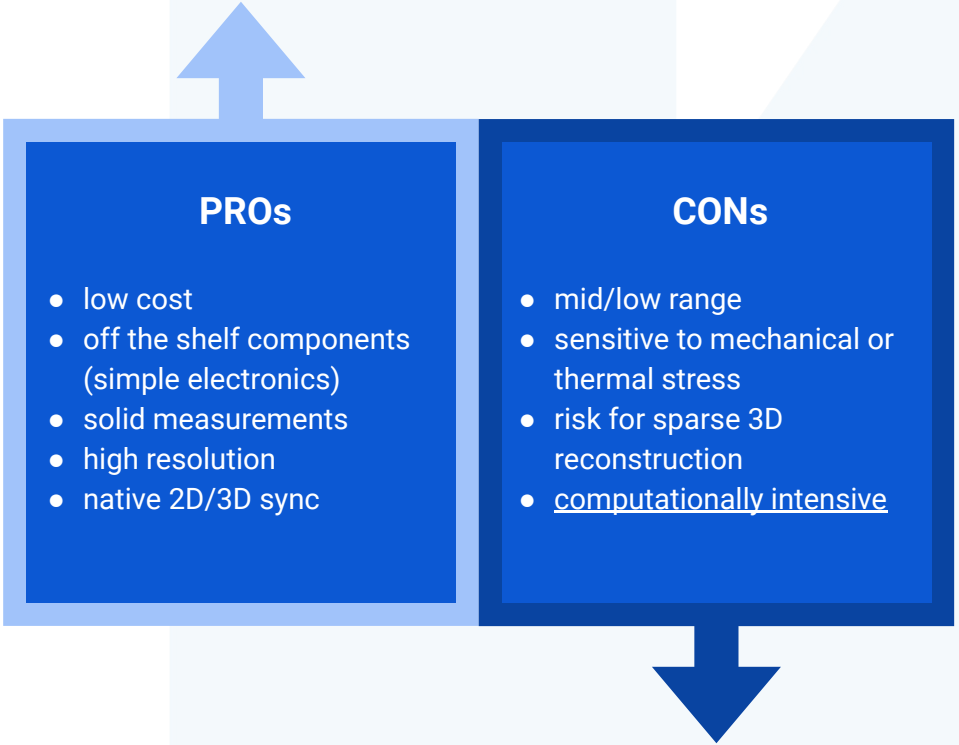
CONs

- mid/low range
- sensitive to mechanical or thermal stress
- risk for sparse 3D reconstruction
- computationally intensive

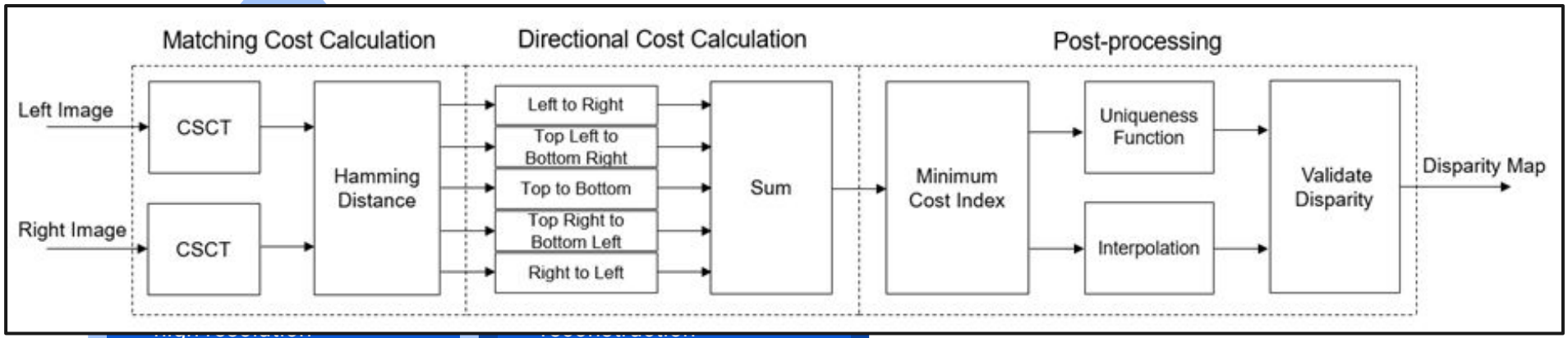


- missing texture
- blur (motion, out of focus)
- visual adversities (dust, rain, noise)

on **stereo cameras**



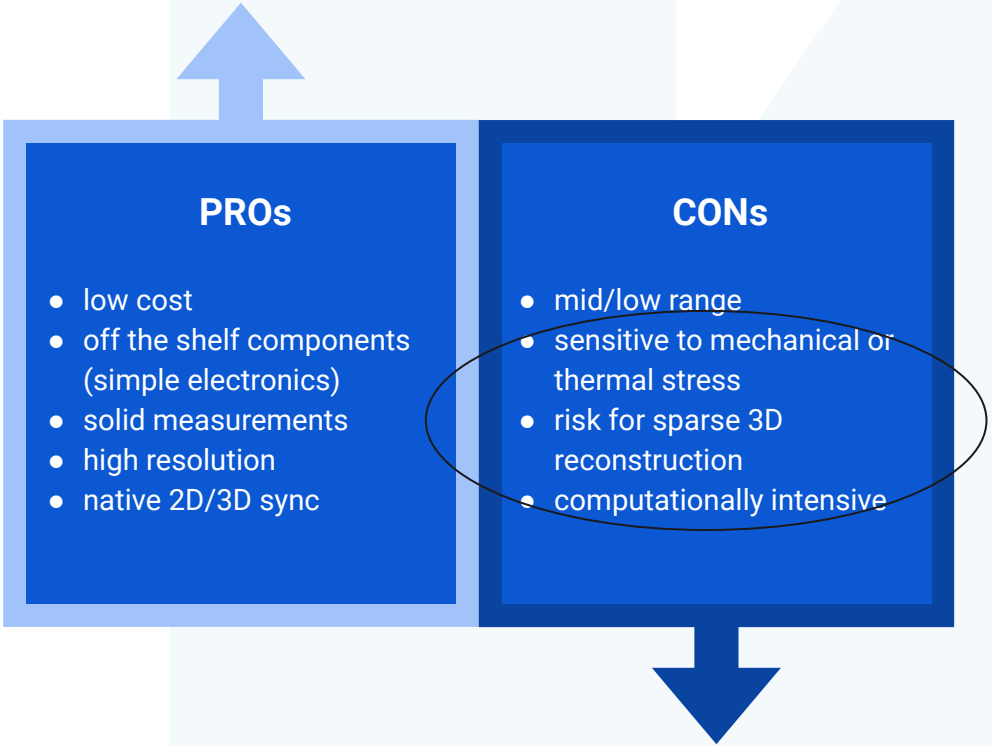
SGBM workflow



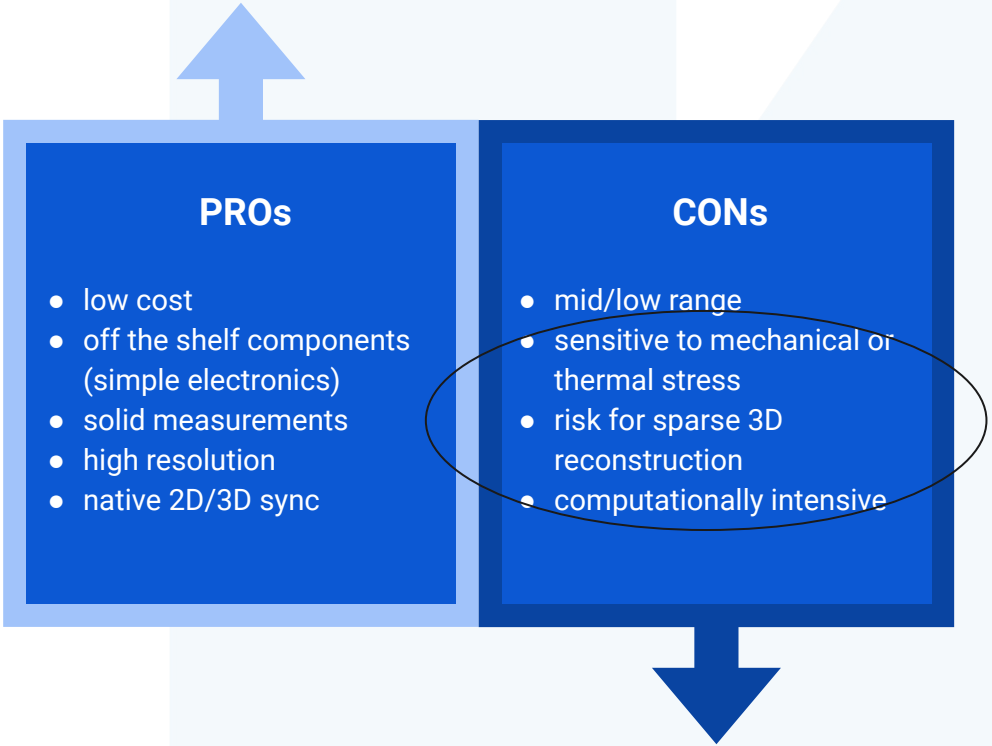
- native 2D/3D sync
- computationally intensive

Image taken from “Stereo Disparity Using Semi-Global Block Matching”, Matlab documentation

AI for stereo matching



AI for stereo matching



1) exploit the best of **geometry** and **prior knowledge**
2) cast the stereo matching to a **DNN computing paradigm**

AI for stereo matching

exploit the best of **geometry** and **prior knowledge**

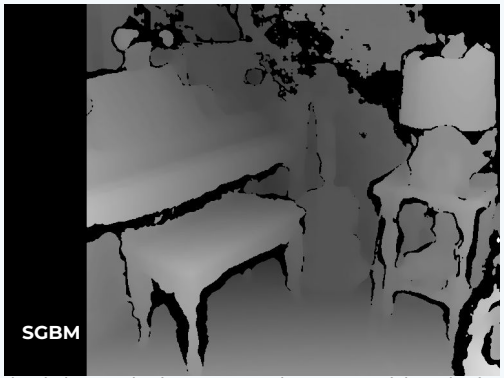
- geometry => metric measurements

AI for stereo matching

exploit the best of **geometry** and **prior knowledge**

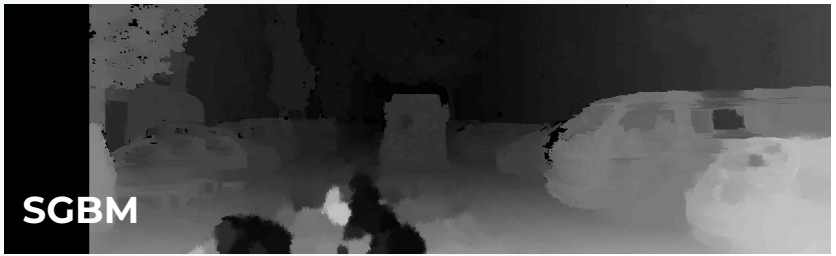
- geometry => metric measurements
- prior knowledge => (1) solve ambiguities
  (2) increase robustness to stress

AI for stereo matching



Images are taken from Middelbury Stereo Dataset, *D. Scharstein, et al. High-resolution stereo datasets with subpixel-accurate ground truth. GCPR 2014*

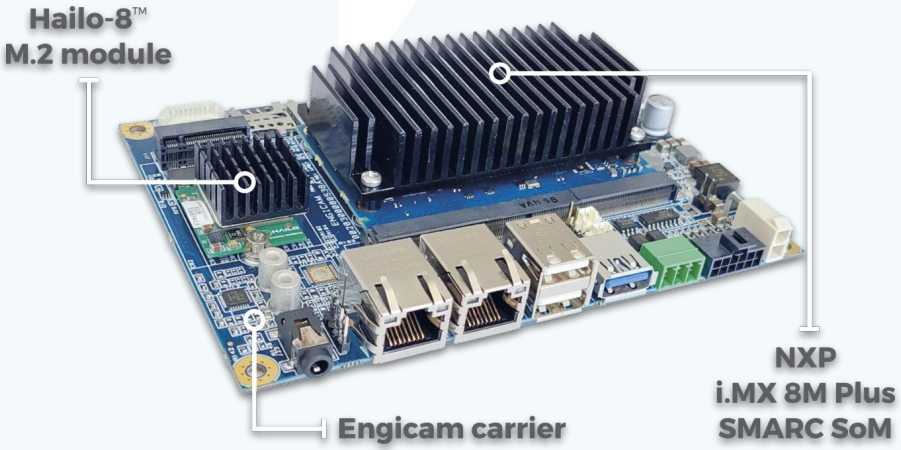
AI for stereo matching



Images are taken from Kitti Stereo Dataset 2015, *M. Menze, A. Geiger; Object Scene Flow for Autonomous Vehicles. CVPR 2015*

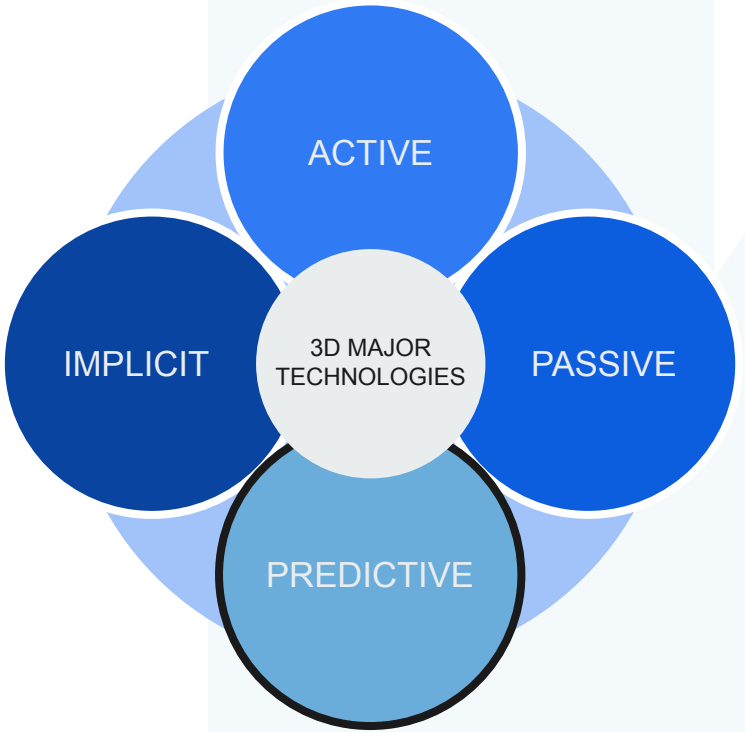
AI for stereo matching

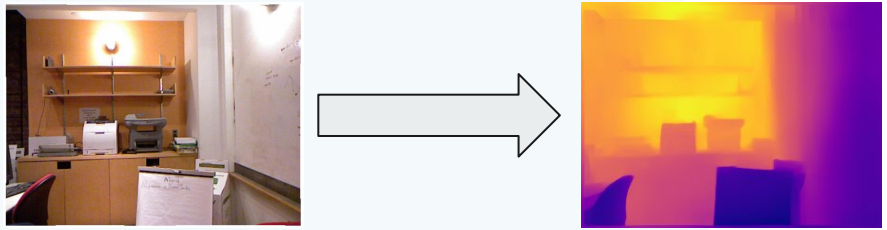
by casting the stereo matching task to a DNN network, there is an additional benefit: get access to the huge computational resources of AI accelerators



SGBM @640x480: ~ 2fps (CPU at 100%)
3D+AI @640x480: **~27fps (CPU at 50%)**

review of some **3D sensor technologies**



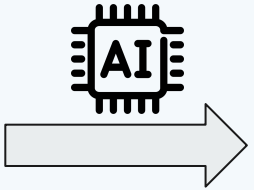


“monocular depth” or
“depth estimation” or
“depth prediction”

on **monocular depth**

close  far

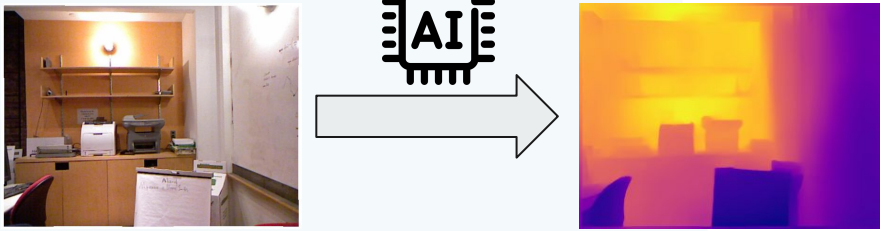
- no use of physics, geometry or optics, **just pure AI**
- embed into the AI model prior knowledge
- estimate distance based on visual appearance just like the humans with a single eye



on **monocular depth**

close  far

- no use of physics, geometry or optics, **just pure AI**
- embed into the AI model prior knowledge
- estimate distance based on visual appearance just like the humans with a single eye



monocular depth does not measure but infers;

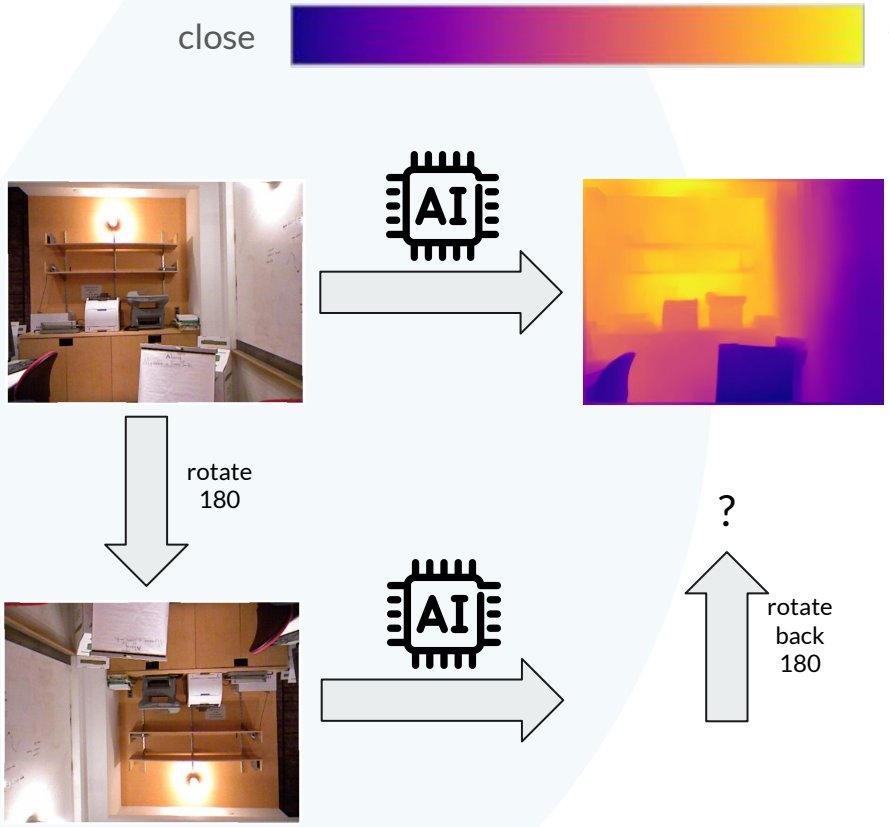
- it is **context dependent**
- it measures **up to a scale factor** (SiLog)

on **monocular depth**

- no use of physics, geometry or optics, **just pure AI**
- embed into the AI model prior knowledge
- estimate distance based on visual appearance just like the humans with a single eye

monocular depth does not measure but infers;

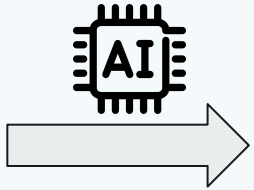
- it is **context dependent**
- it measures **up to a scale factor** (SiLog)



on **monocular depth**

close  far

- no use of physics, geometry or optics, **just pure AI**
- embed into the AI model prior knowledge
- estimate distance based on visual appearance just like the humans with a single eye



monocular depth does not measure but infers;

- it is **context dependent**
- it measures **up to a scale factor** (SiLog)

on **monocular depth**

- no use of physics, geometry or optics, **just pure AI**
- embed into the AI model prior knowledge
- estimate distance based on visual appearance just like the humans with a single eye

monocular depth does not measure but infers;

- it is **context dependent**
- it measures **up to a scale factor** (SiLog)

on **monocular depth**

- no use of physics, geometry or optics, **just pure AI**
- embed into the AI model prior knowledge
- estimate distance based on visual appearance just like the humans with a single eye

monocular depth does not measure but infers;

- it is **context dependent**
- it measures **up to a scale factor** (SiLog)

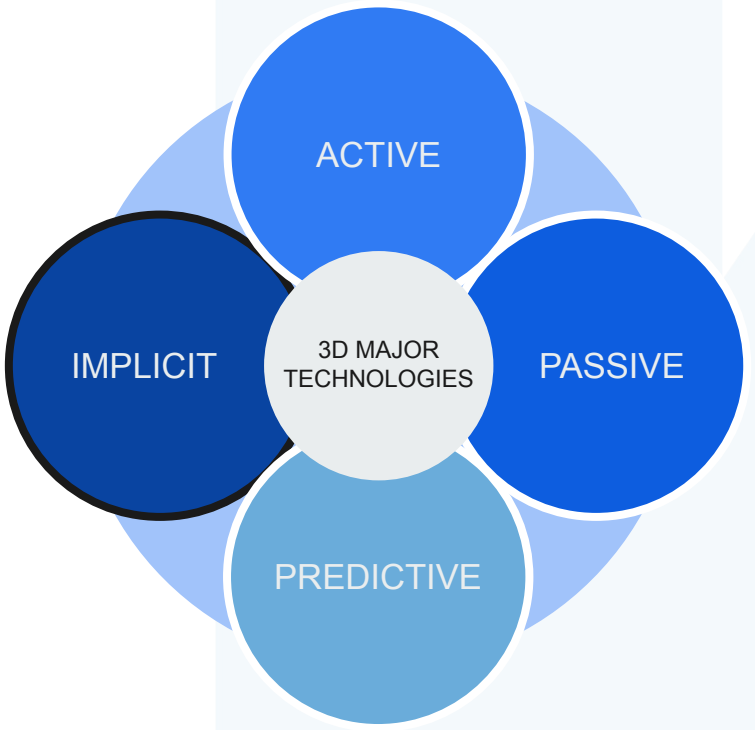
on **monocular depth**

- no use of physics, geometry or optics, **just pure AI**
- embed into the AI model prior knowledge
- estimate distance based on visual appearance just like the humans with a single eye

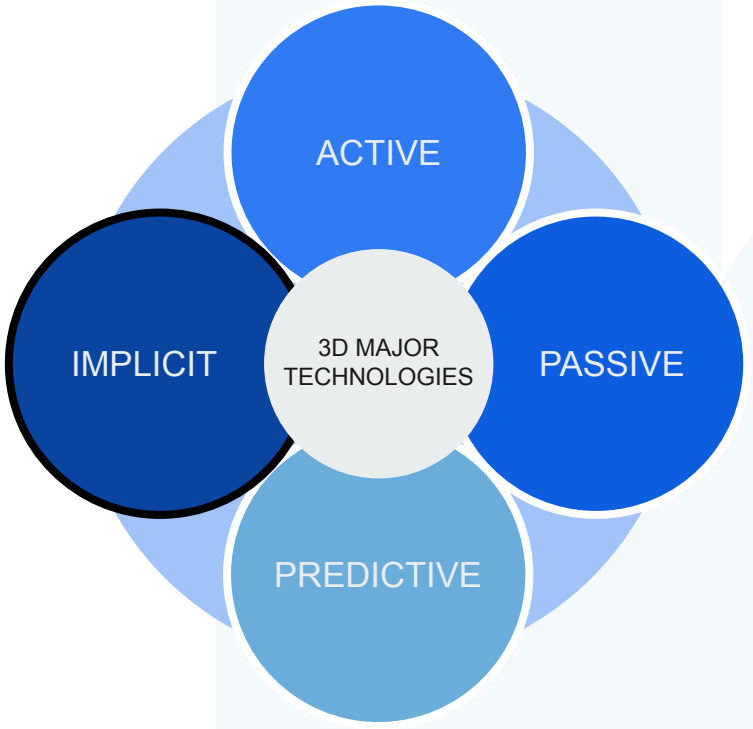
monocular depth does not measure but infers;

- it is **context dependent**
- it measures **up to a scale factor** (SiLog)

review of **3D sensor technologies**

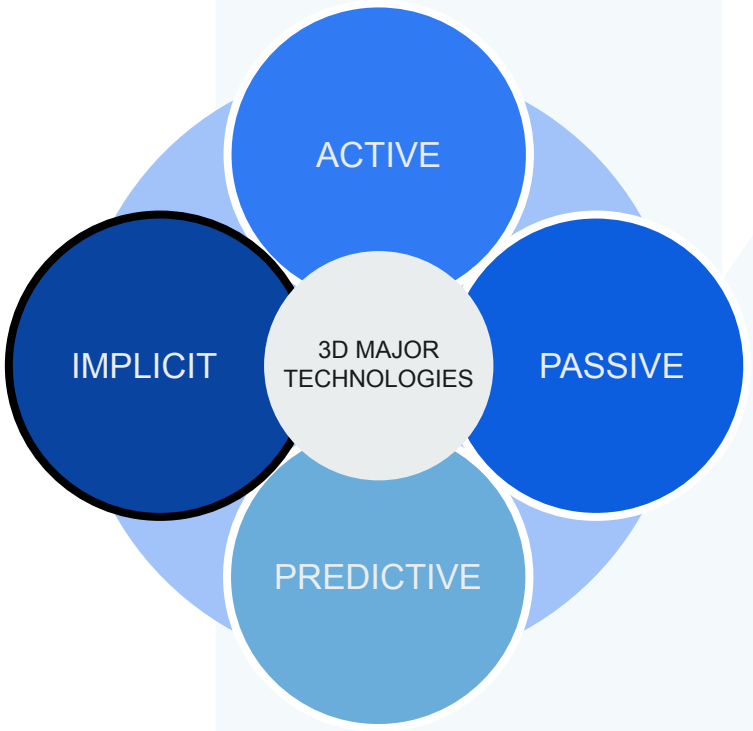


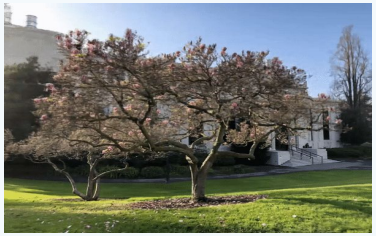
review of **3D sensor technologies**



- Implicit means that the 3D is encoded in the weights of a NN and not directly accessible.
- The most remarkable example of implicit models are **NERFs** that include texture and viewing angle

review of **3D sensor technologies**

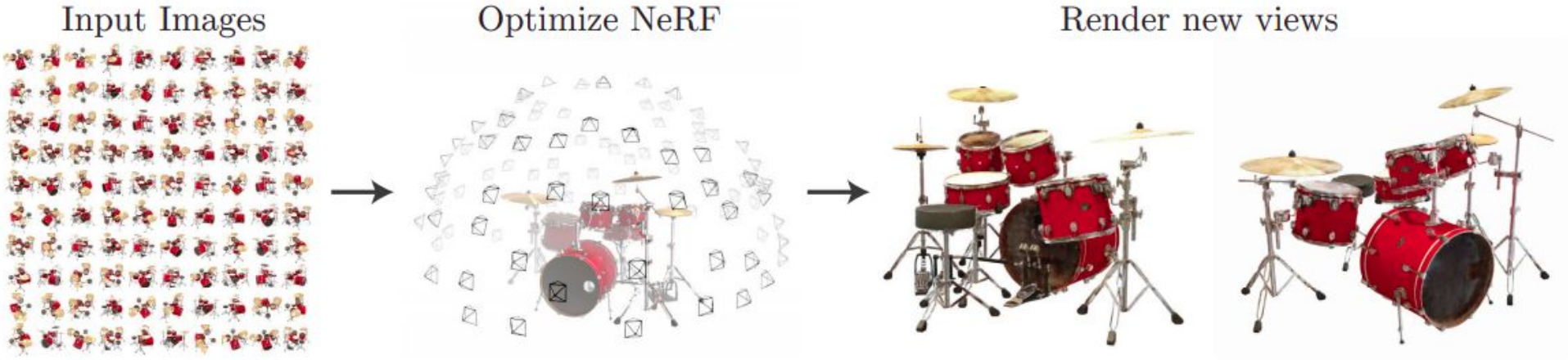




- **The network itself “somehow is” the 3D model**
- An inference step is like a camera taking a picture.
- The model weights must represent:
  - the geometry of the object / scene
  - how the texture looks from different viewpoints

NeRF

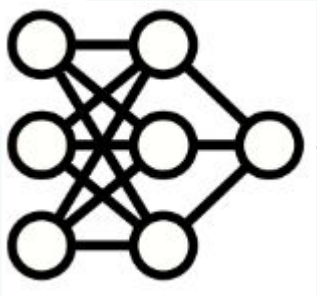
- Technique for **Novel View Synthesis** from a **sparse** set of input views



NeRF - Overview

3D point and viewing direction

(x,y,z,ϑ,φ)

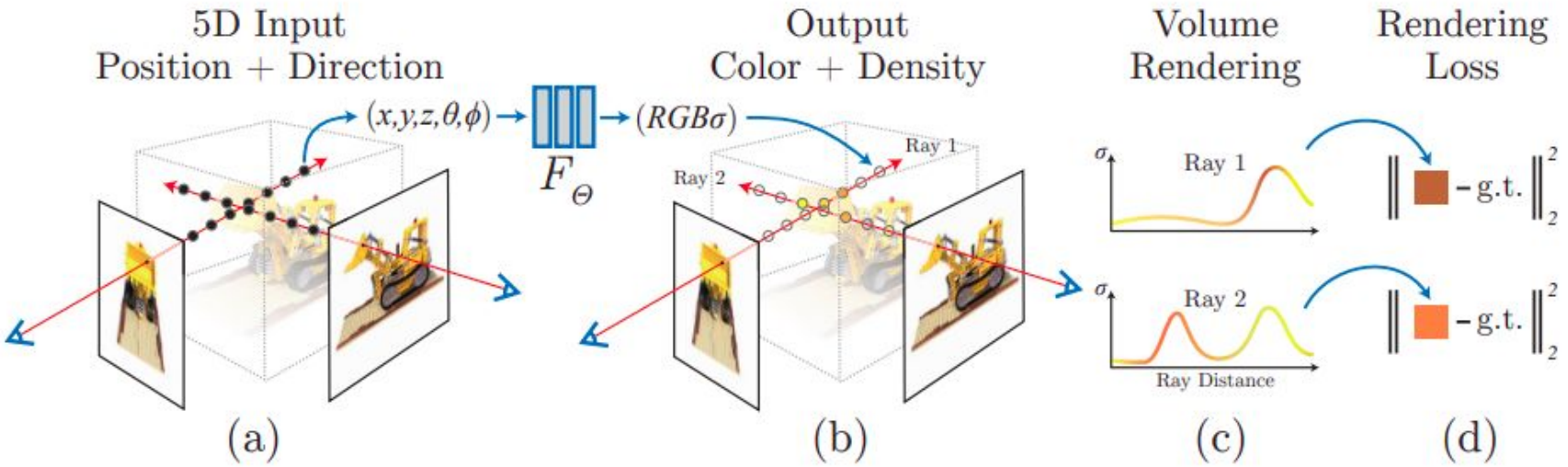


MLP

Color RGB and volume density σ

(RGBσ)

NeRF - Overview



Pros:

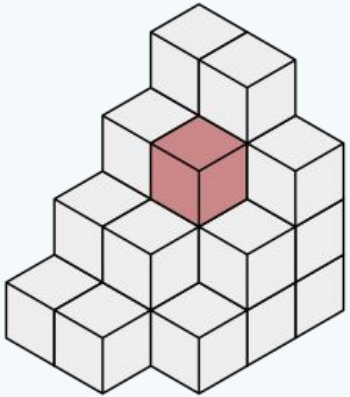
- a compact and continuous way to represent a 3D scene
- once trained it can generate a point cloud similar to SFM, but completely dense
- it takes a few seconds to encode a scene

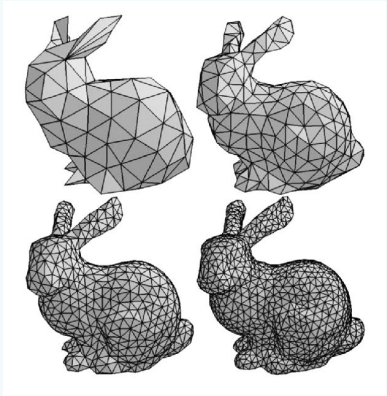
Cons:

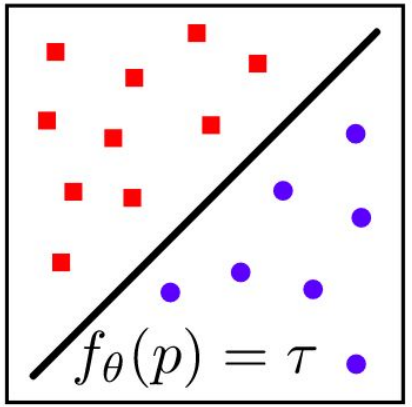
- not handy for downstream tasks (e.g. segmentation or detection), but researchers are working on it
- sometimes a few seconds are too much
- it still requires a large number of images to work well (in the tens)

The Problem of **3D Representation**

How can we represent the 3D space in a suitable way for learning-based methods?

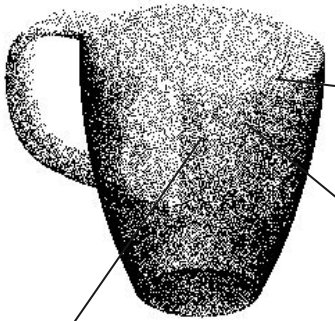


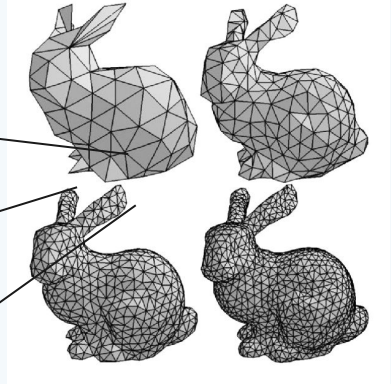


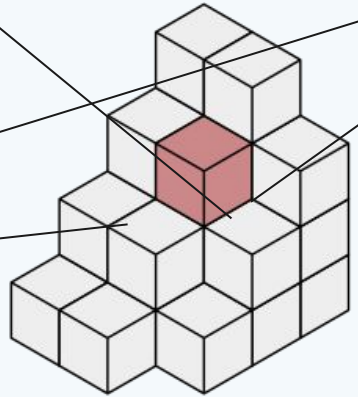


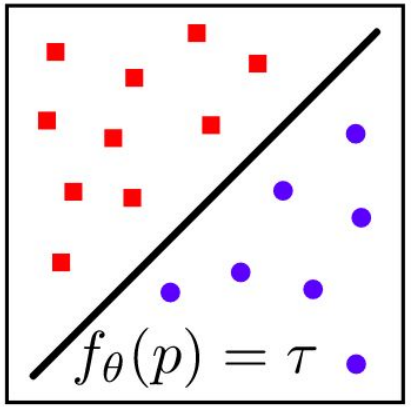
The Problem of **3D Representation**

How can we represent the 3D space in a suitable way for learning-based methods?









Depth image representation

Represent a scene using a matrix of distances

Depth image representation

Pros:

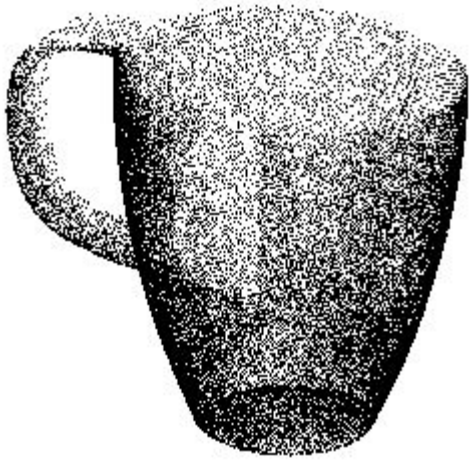
- Simple, compact and understandable representation
- can be tackled with regular 2D DNN approaches

Cons:

- 2.5D, not a real 3D;
- discretization of 3D points
- lacks the connectivity structure of the underlying surfaces

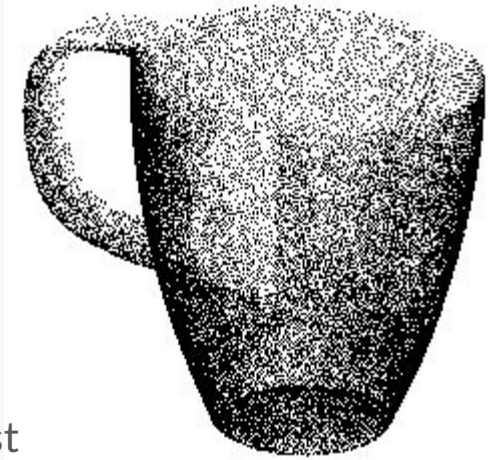
Point-Based Representation

Represent a scene using a set of 3D points



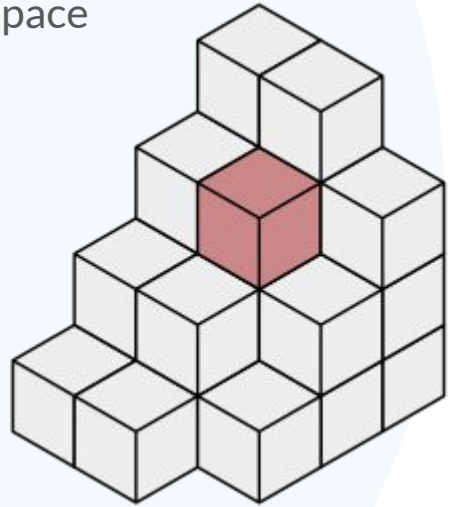
Point-Based Representation

- **Pro:**
  - Simple and compact representation
  - efficient for sparseness
- **Con:**
  - Lacks the connectivity structure of the underlying surfaces
  - To be invariant to density and order, points must be processed independently



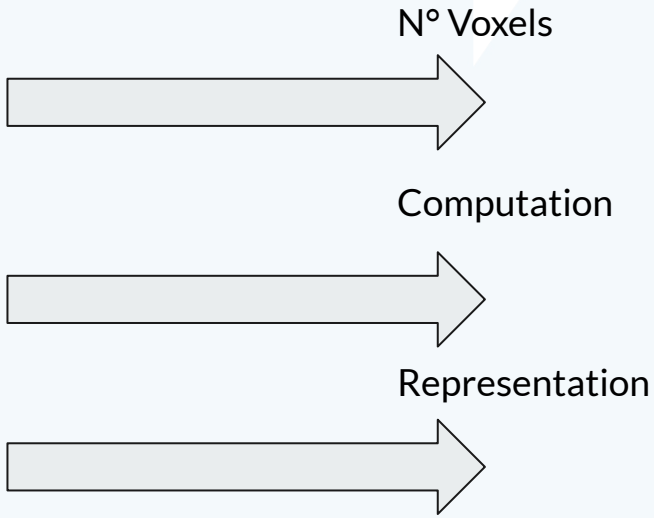
Voxel Representation

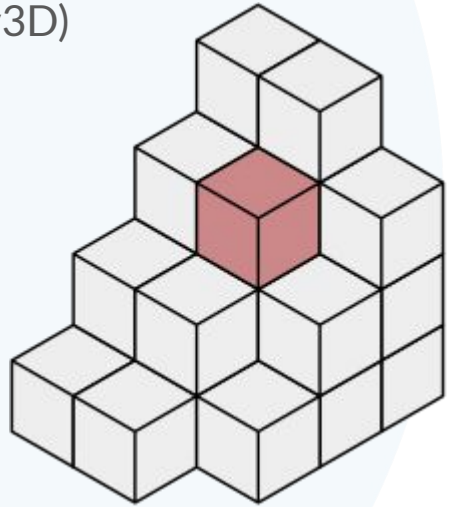
- A voxel represent a value on a **regular grid** on 3D space
- Represents the 3D version of a pixel



Voxel Representation

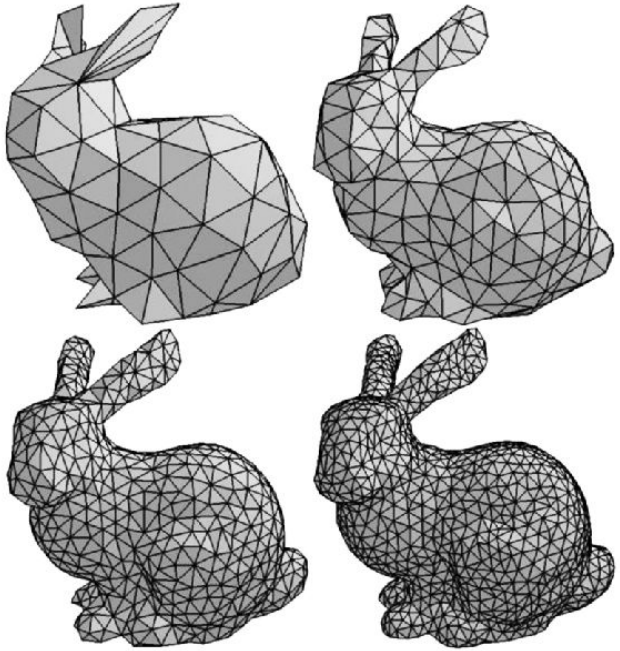
- **Pro**: suitable for learning based methods (e.g. conv3D)
- **Cons**: Computationally expensive





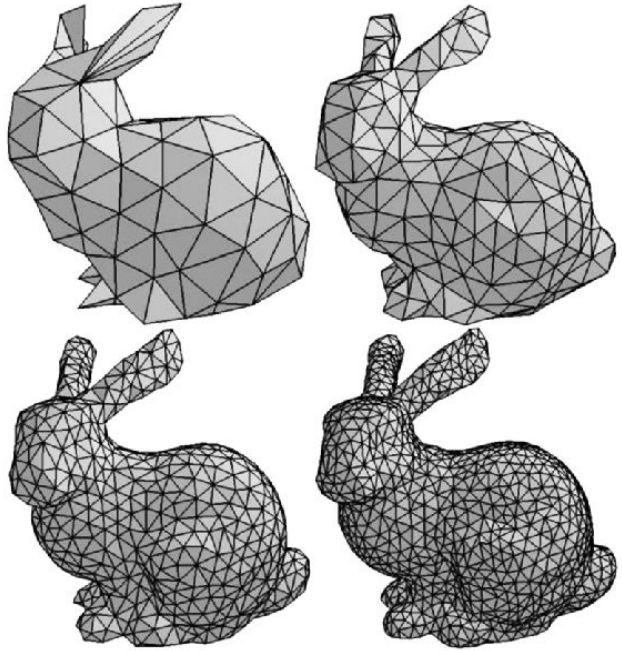
Mesh Representation

- Represent a scene using triangles and vertices



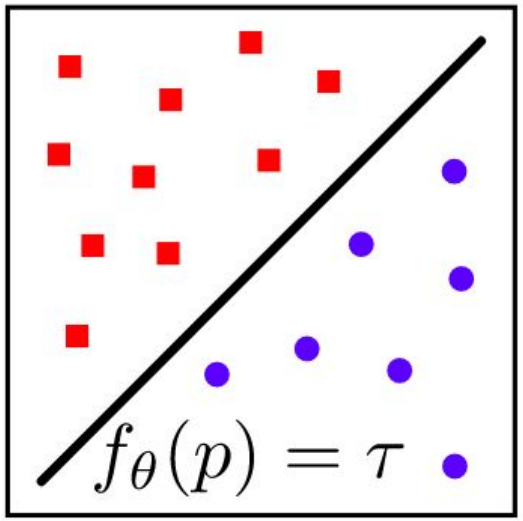
Mesh Representation

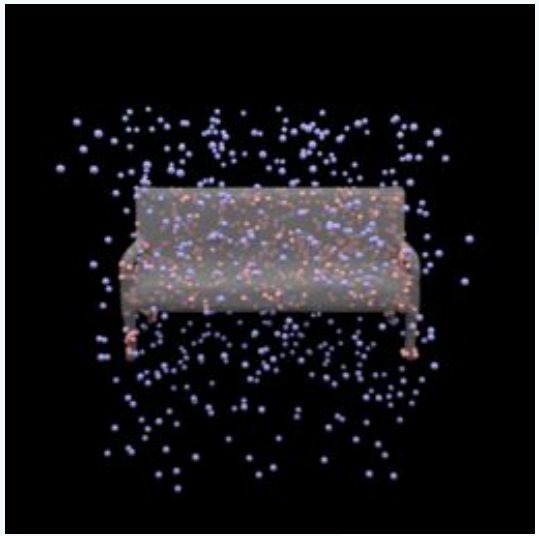
- **Pro:**
  - Efficient
  - Structured representation
- **Cons:**
  - Based on deforming a template mesh -> Do not allow arbitrary topologies



Implicit Representation

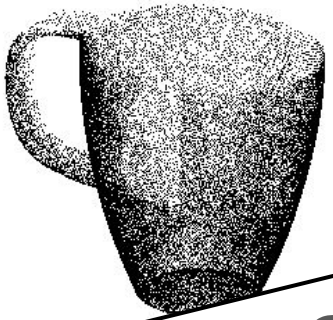
Describing 3D geometry **implicitly**, e.g., as the decision boundary of a **binary classifier**

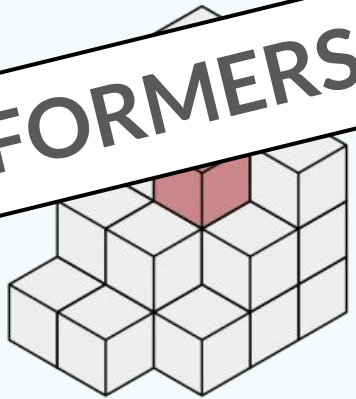




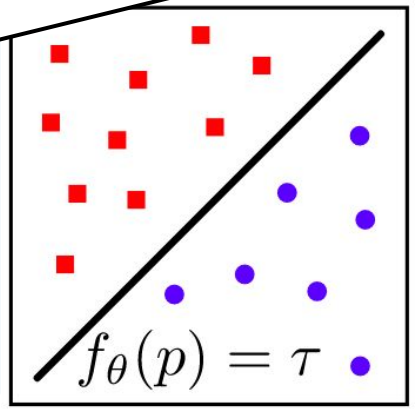
The Problem of **3D Representation**

How can we represent the 3D space in a suitable way for learning-based methods?





TRANSFORMERS TAKE THEM ALL



Wrapping up…

thanks for your attention!
questions are welcome!

drop us an email

Headquarter: Via Capilupi 21, 41122 Modena, Italy
Phone: +39 059 8678417
Mail: info@deepvisionconsulting.com
Web: www.deepvisionconsulting.com
Linkedin:www.linkedin.com/company/deep-vision-consulting