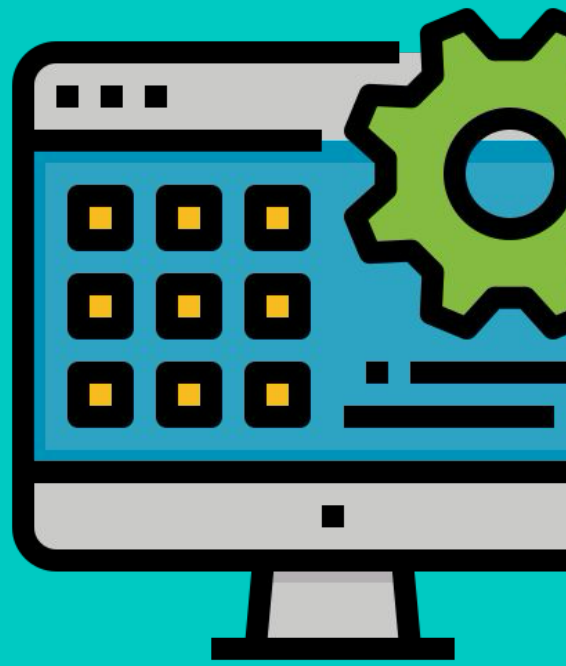
Monocular Depth Estimation: an eye on real-world applications



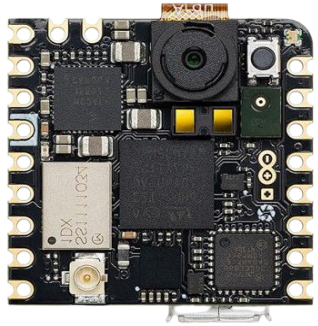
A.A. 2024-2025

Claudio Schiavella, Ph.D Student

Supervisor: Prof: Irene Amerini

SAPIENZA
UNIVERSITÀ DI ROMA

ALCOR Lab



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

Depth

Why?

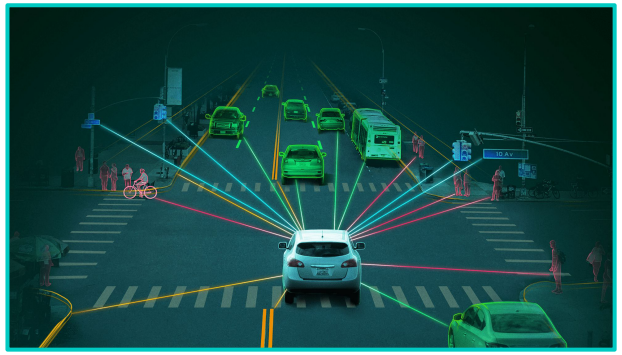
Overview

- Depth: the 4th dimension
- Deep Learning for depth estimation
- Open challenges
- Efficiency is the key

Depth: the 4th dimension

Where Depth Makes the Difference

Robotics



Autonomous driving

Biometrics

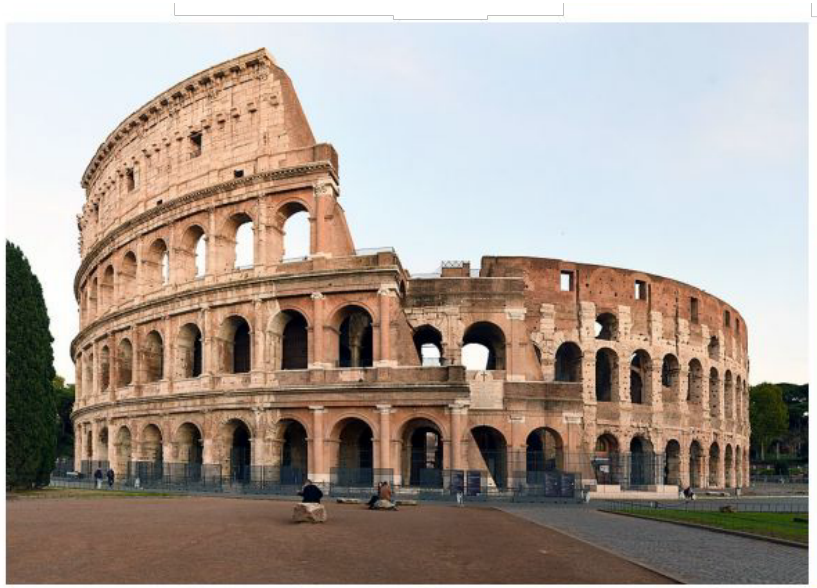
Drones

Games

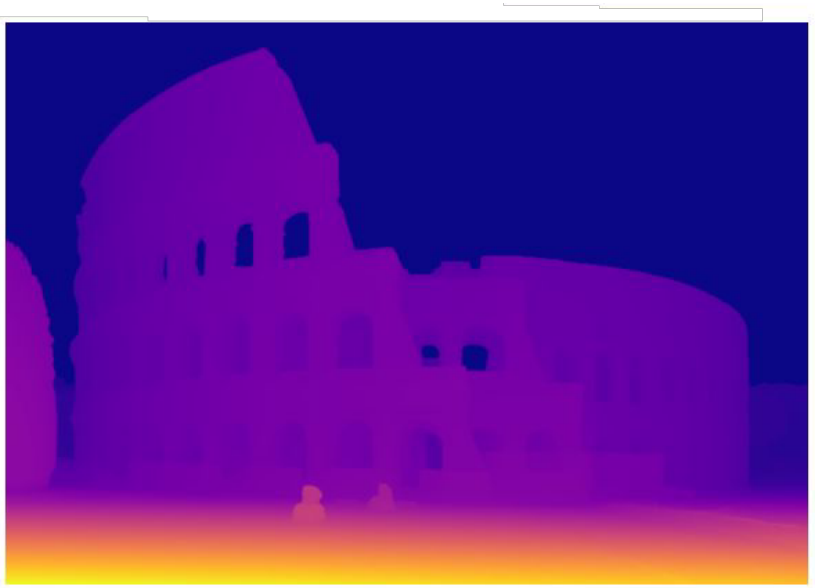
Augmented reality

RGB(D) image

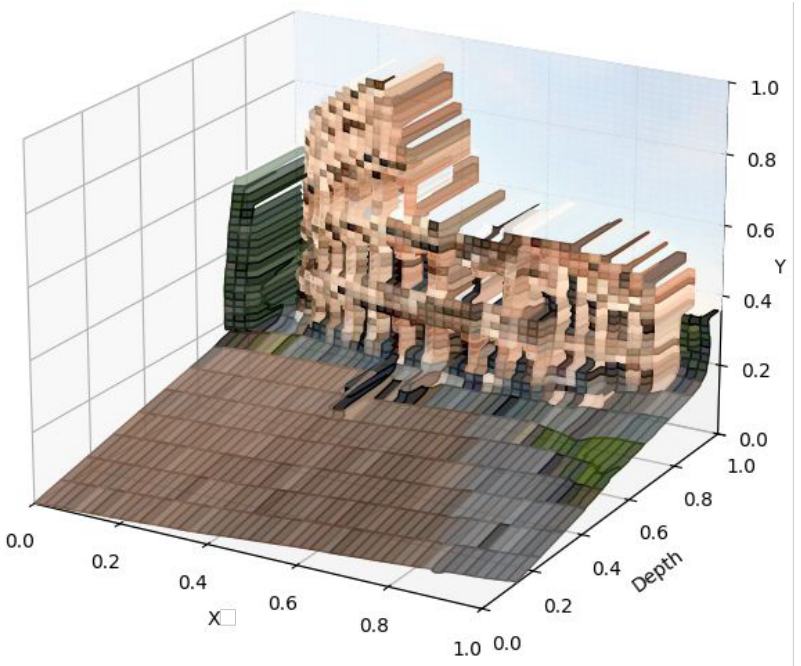
Image



Depth

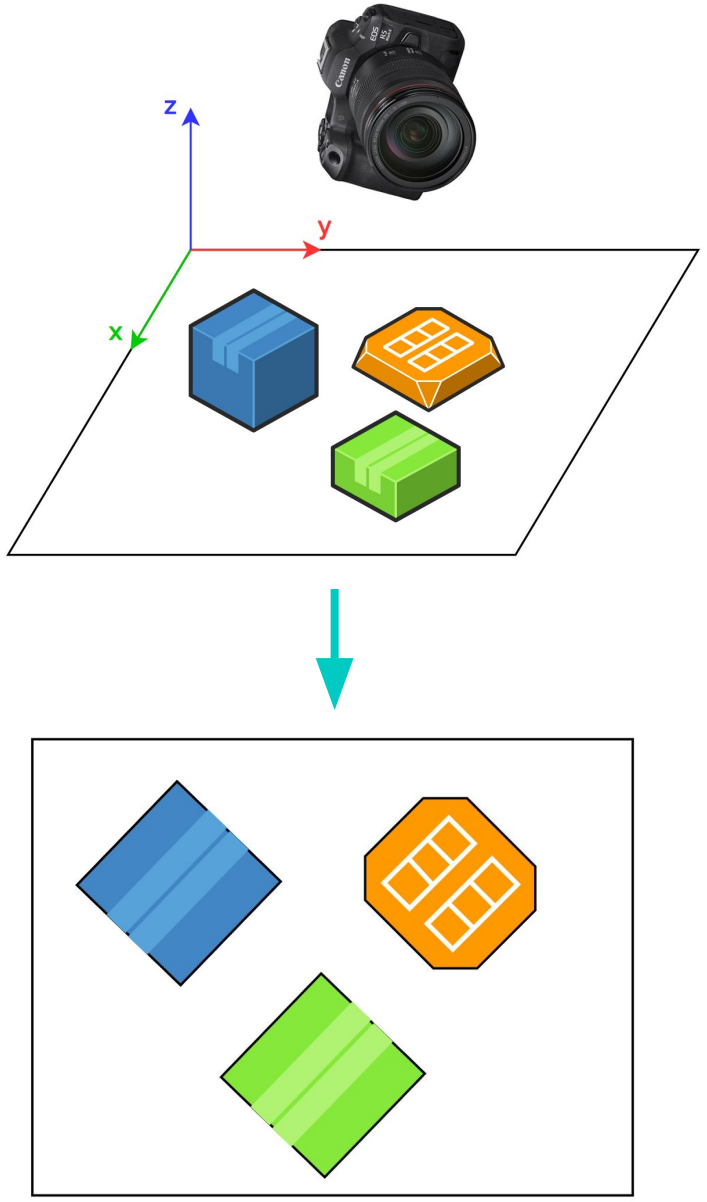


3D reconstruction



Sometimes we lose

- Images project **3D scene** into a **2D representation**
  - This results in the **loss** of depth information
- Without depth, we miss:
  - Spatial relationships between objects
  - Accurate object proportions and scale
- Depth estimation is **necessary**
  - Many systems rely on it to interact with the world
  - Enables machines to reason about scene geometry



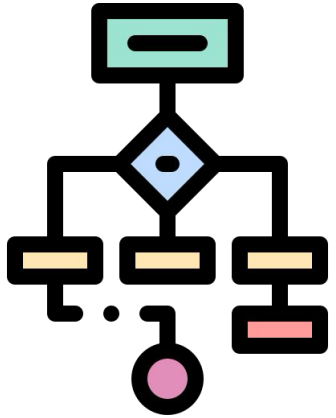
Depth sensing

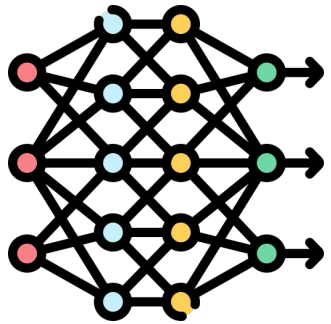
Active depth sensing

Measure the depth

Passive depth sensing

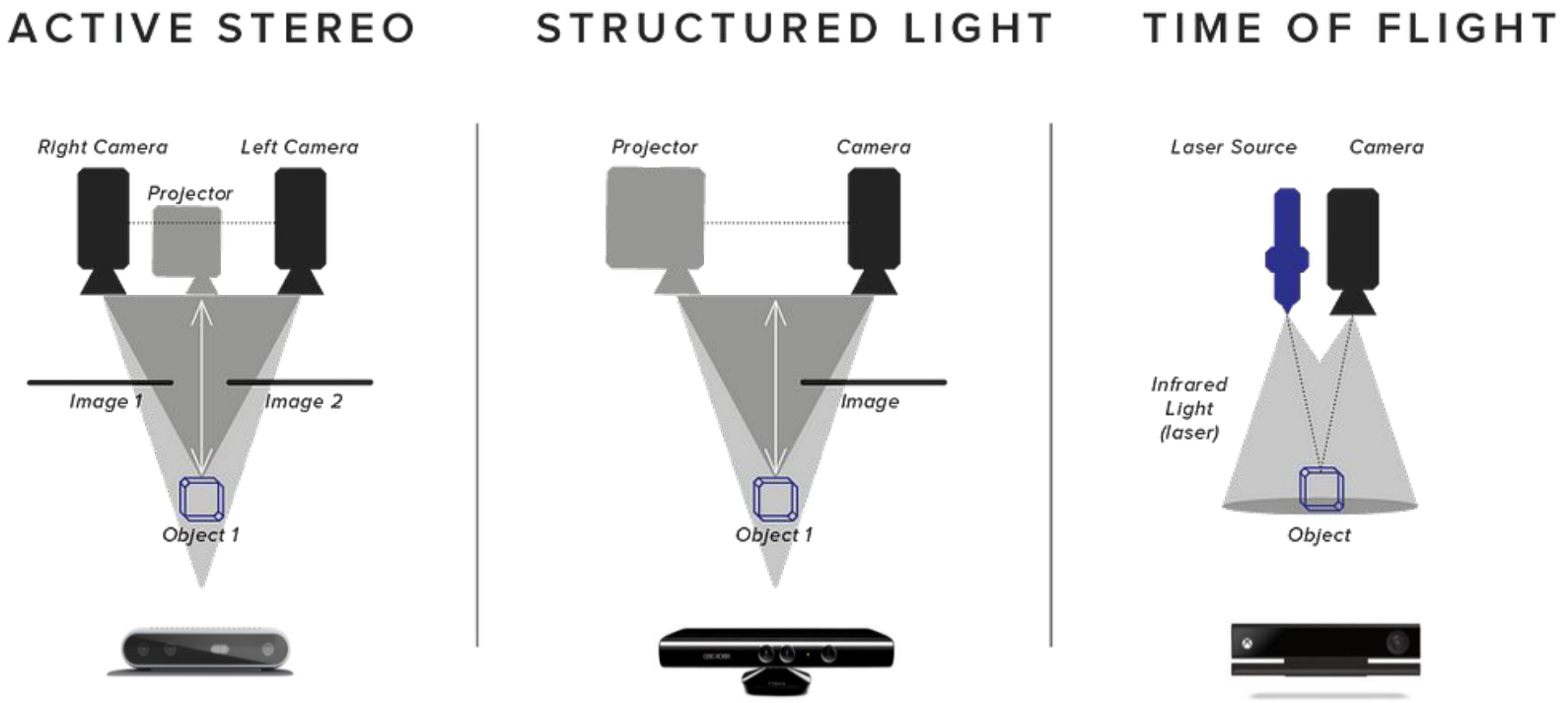
Estimate the depth





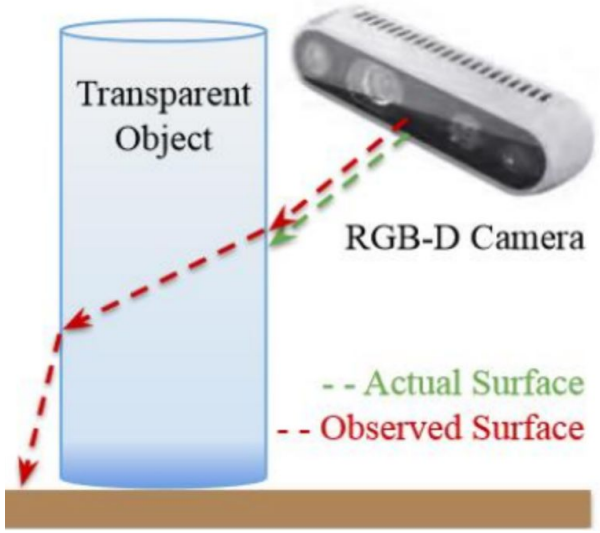
In symbiosis with the environment…

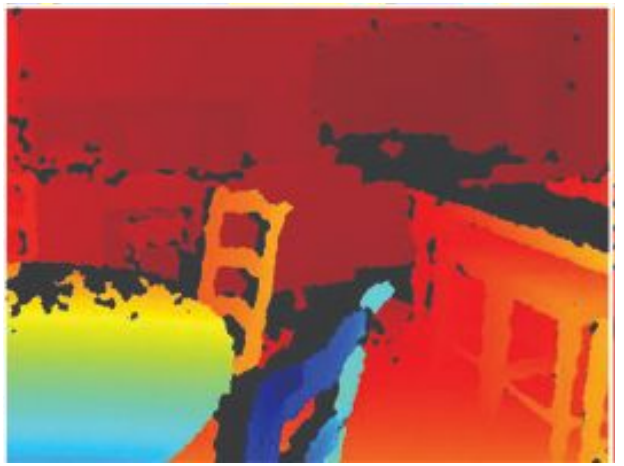
- Depth is perceived **actively** by **perturbing** the environment
  - Active sensors emit and capture signals to **infer distance**
  - Effective and popular depth measurement



… But not with the hardware

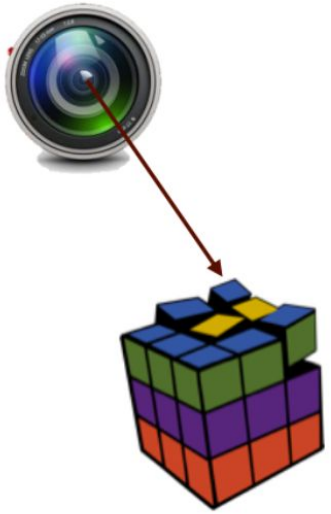
- **Sensitive** to environmental conditions
  - Emitted signals can be disrupted
- **Limited** range and resolution
  - Active sensors typically work only at short distances
- **Issues** with reflective or transparent surfaces
  - Signals can be misdirected or not returned
- **Energy** and hardware demands
  - Emitting signals requires specialized components
  - Limiting use in mobile or low-power devices



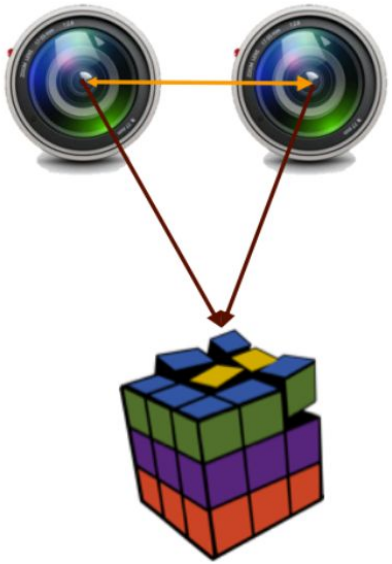


Change the point of view

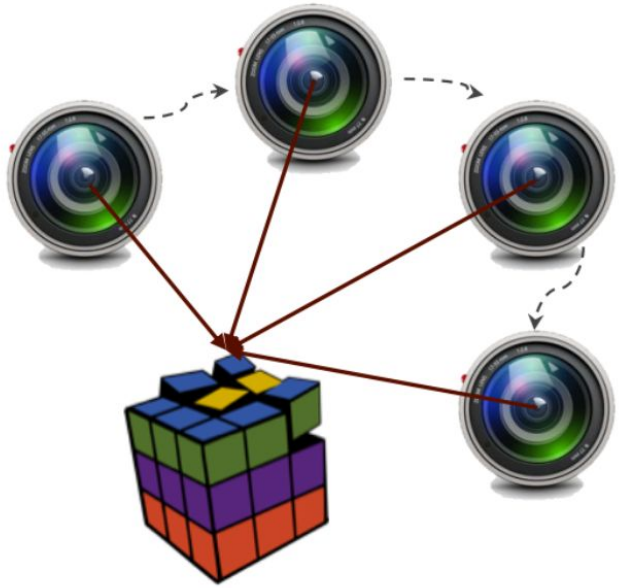
- Depth is inferred passively by considering one or a few images
  - Depth is reconstructed is estimated, **not measured**
  - Complexity is moved to **algorithms**



Monocular



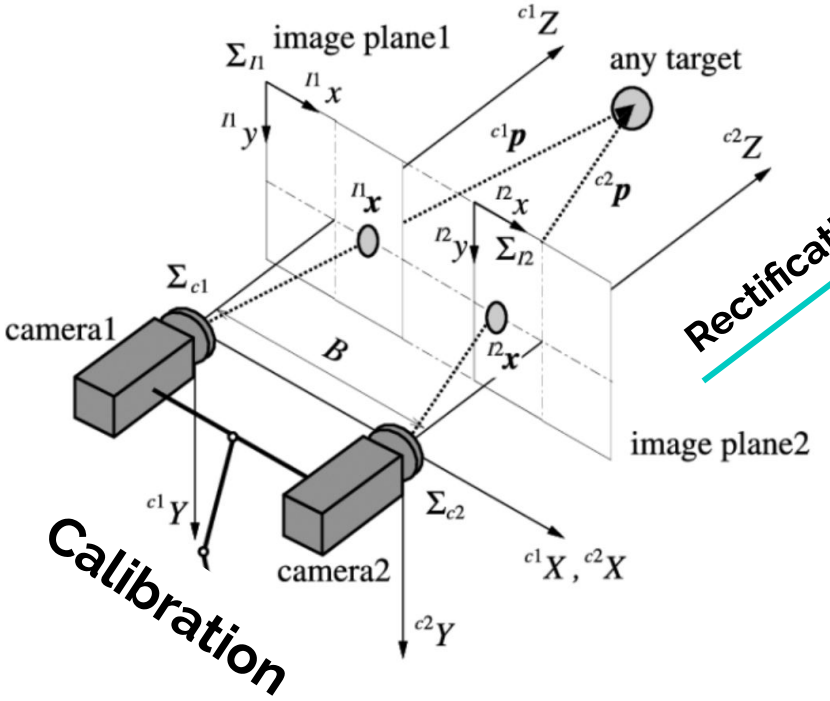
Binocular stereo



Multi-view

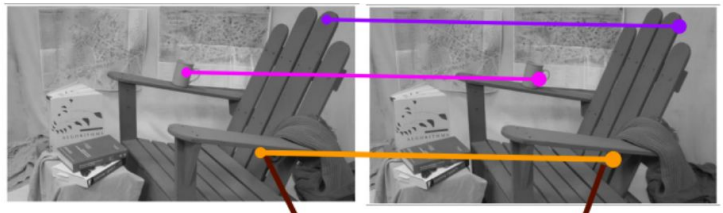
Two eyes better than one

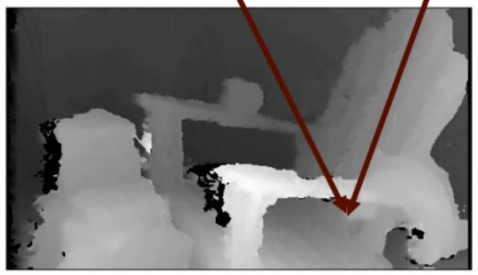
- Given more images, if we are able to find **corresponding point** in the images
  - We can infer depth by **triangulation**



Rectification

Matching





Disparity map

Triangulation

Depth Map

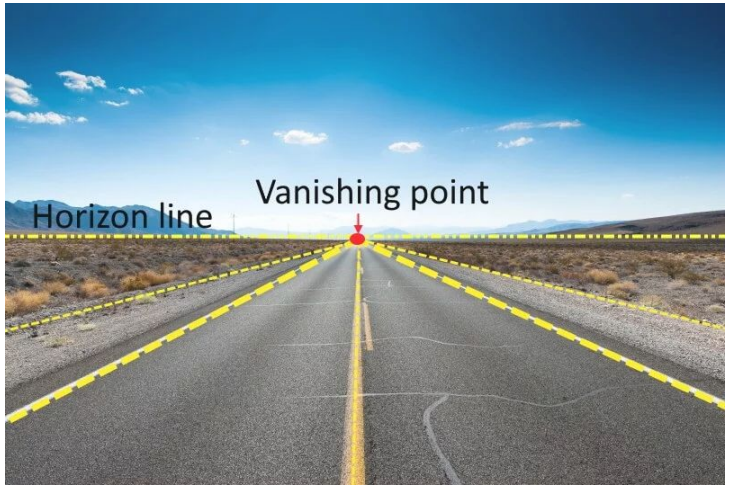
Two eyes better than one… or not?

- **Monocular Depth Estimation (MDE)**
  - Given a single image, predict a dense depth map for each pixel
- Requires **only one camera**
  - Reducing cost, size, and calibration complexity
  - Ideal for edge computing and not limited by camera placement
- Enables depth estimation from **existing photos** or videos
  - Can fill gaps when fail to find correspondences (e.g., occlusions)

An ill-posed problem

- 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
  - There are some meaningful **cues**
  - But they aren't alway strong anchor



Linear perspective

Relative size

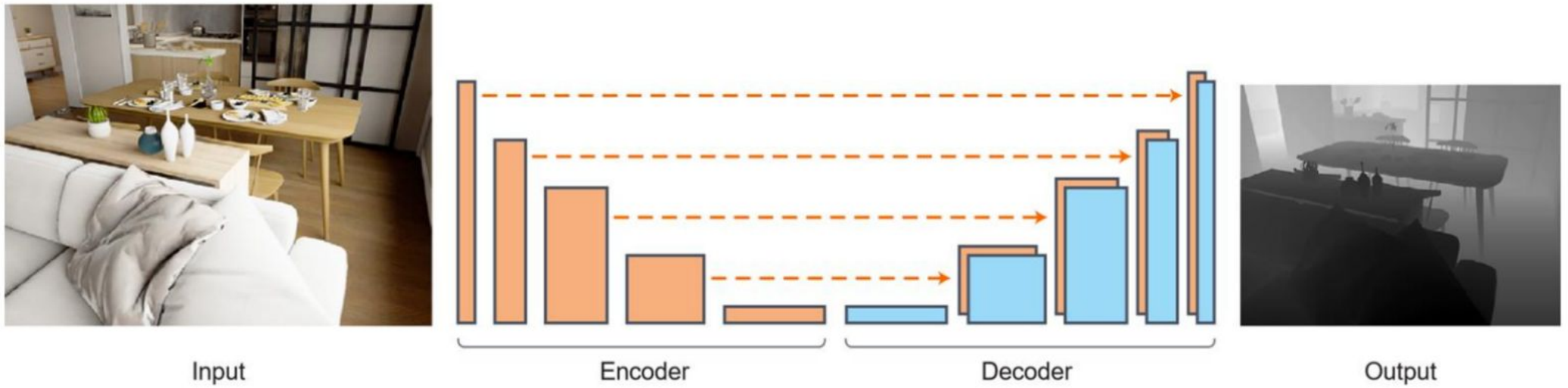
Depth pitfalls

Ambiguous perspective

Ambiguous pattern

How approach MDE

- In Computer Vision, existing MDE solutions rely on **Deep Learning**
  - **Supervised:** Ground-truth depth data
  - **Semi-Supervised:** Sparse ground-truth depth + image reconstruction
  - **Unsupervised:** Image reconstruction



Alhashim, I., & Wonka, P. (2019). High Quality Monocular Depth Estimation via Transfer Learning

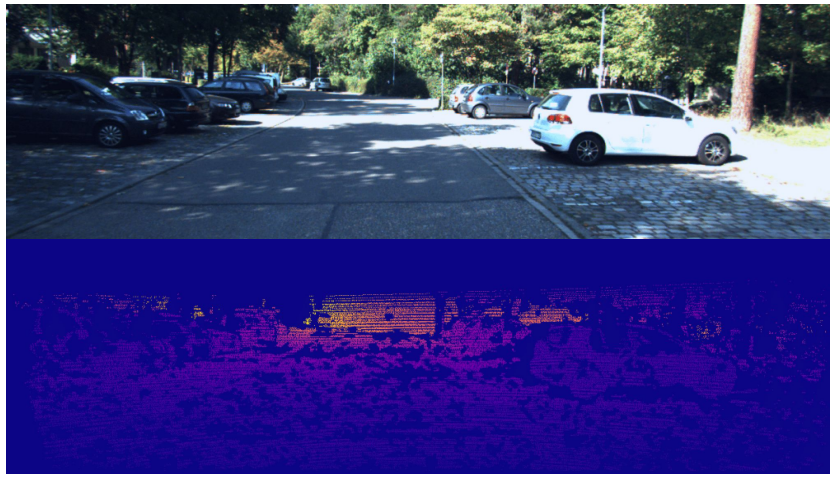
Benchmark datasets

NYU Depth V2

- **Range:** 0.5 - 10 meters
- **Samples**: 120K
- **Type:** depth image

KITTI

- **Range:** 0.9 - 120 meters
- **Samples:** 93K
- **Type:** LiDAR point cloud



Evaluation metrics - Errors

- Given a predicted depth map p_i , its ground truth g_i for an image P
  - **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** $$Abs_{rel} = \frac{1}{|P|} \sum_{i \in P} \frac{|p_i - g_i|}{g_i}$$

Evaluation metrics - Accuracy

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

$$\delta_1 = \frac{1}{|P|} \sum_{i \in P} max\left(\frac{p_i}{g_i}, \frac{g_i}{p_i}\right) < threshold = 1.25$$

$$\delta_2 = \frac{1}{|P|} \sum_{i \in P} max\left(\frac{p_i}{g_i}, \frac{g_i}{p_i}\right) < threshold = 1.25^2$$

$$\delta_3 = \frac{1}{|P|} \sum_{i \in P} max\left(\frac{p_i}{g_i}, \frac{g_i}{p_i}\right) < threshold = 1.25^3$$

Deep Learning for depth estimation

Problem definition

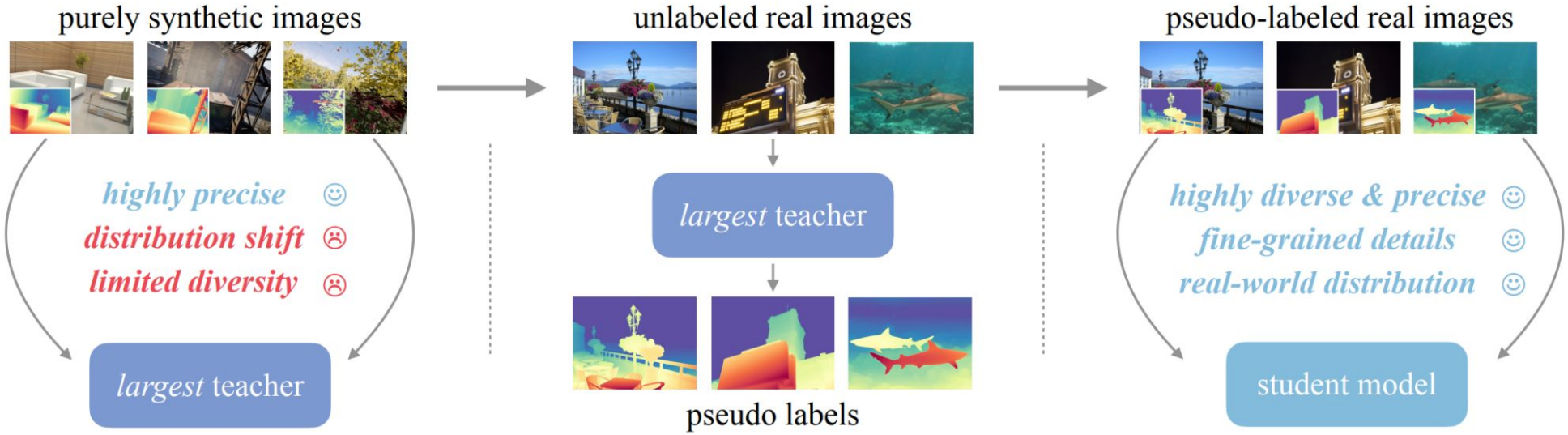
- **Goal:** Estimate the depth of each pixel in a single image
  - **Input:** A single 2D image
  - **Output:** A depth map
    - Each pixel's value represents its distance from the camera.

- **Regression Problem**
  - Predict continuous depth values for every pixel in the image

- MDE losses focus on **distance**
  - Combining different losses helps balance accuracy and visual quality

Depth Anything v2

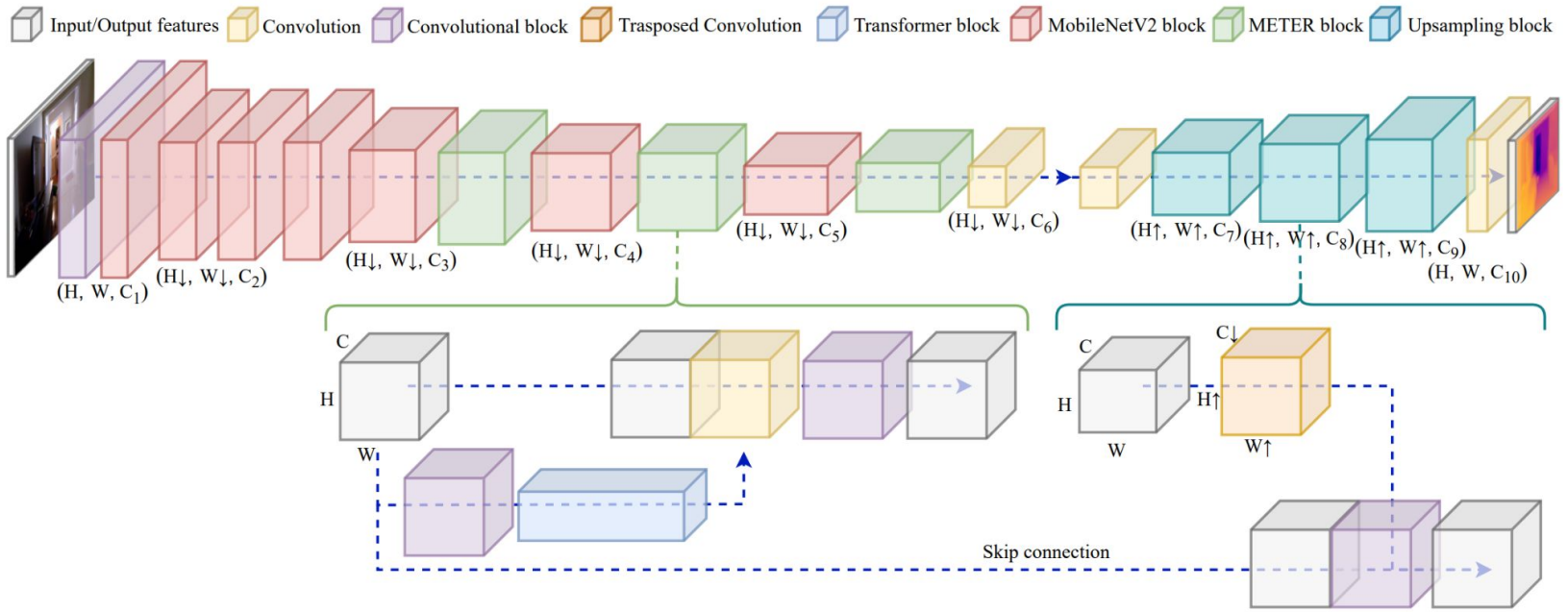
- **Teacher Model:** trained on synthetic images, but limited by distribution shift
- **Pseudo-labeling:** Teacher generates pseudo-labels for real images, training the student model to generalize better.
- The approach achieves **high accuracy** on real-world data



Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., & Zhao, H. (2024). Depth Anything V2

METER

- **Novel Transformer** block in order to improve efficiency and limit the computation
- **Novel Loss** function and Data augmentation strategy
- State-of-the-art estimation **performances** w.r.t. comparable MDE models

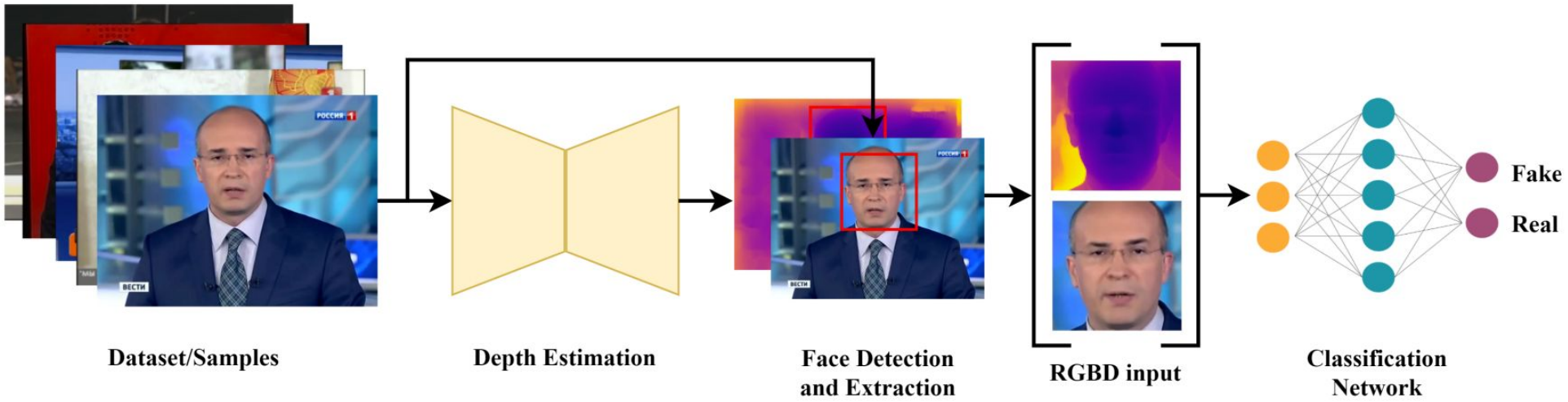


Papa, L., Russo, P., & Amerini, I. (2023). METER: A Mobile Vision Transformer Architecture for Monocular Depth Estimation. *IEEE Transactions on Circuits and Systems for Video Technology*

RGBD for a general task

- Model combines **RGB and depth** maps extracted using MDE
  - Depth is hypothesized to reveal **inconsistencies** from deepfake manipulation

- RGBD approach **significantly improves** deepfake detection



L. Maiano, L. Papa, K. Vocaj and I. Amerini, "DepthFake: a depth-based strategy for detecting Deepfake videos," AI4MFDD Workshop at ICPR, 2022

Open challenges

Promising research directions

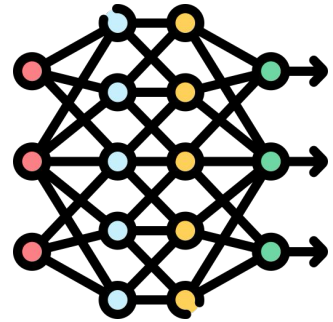
- **Domain** adaptation & Transferability
  - Merge synthetic datasets with real ones
  - Learn camera parameters
- **Lightweight** networks for mobile & edge applications
- **Temporal** consistency
  - Improve the estimation with sequence of predictions
- **Multimodal** learning (RGB +D)
- … and many others …

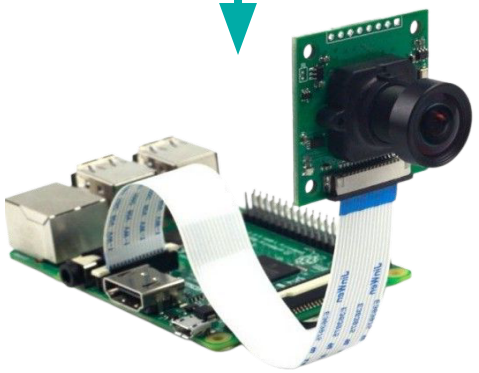
Better, Faster, Smaller

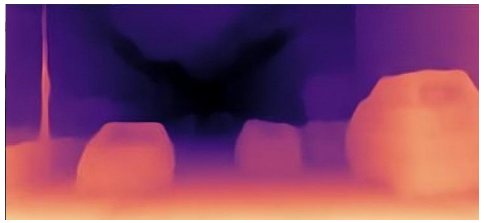
- **Limited resources** make deploying models on edge devices challenging
  - Need for lightweight and efficient architectures

- My research focus on **improving MDE**
  - Challenging due to its need for high accuracy and fine-grained pixel-level depth predictions

- Ensuring that models are fast and accurate
  - Focusing on **real-time performance** while still maintaining high depth estimation **quality**
  - Across various environments and devices.



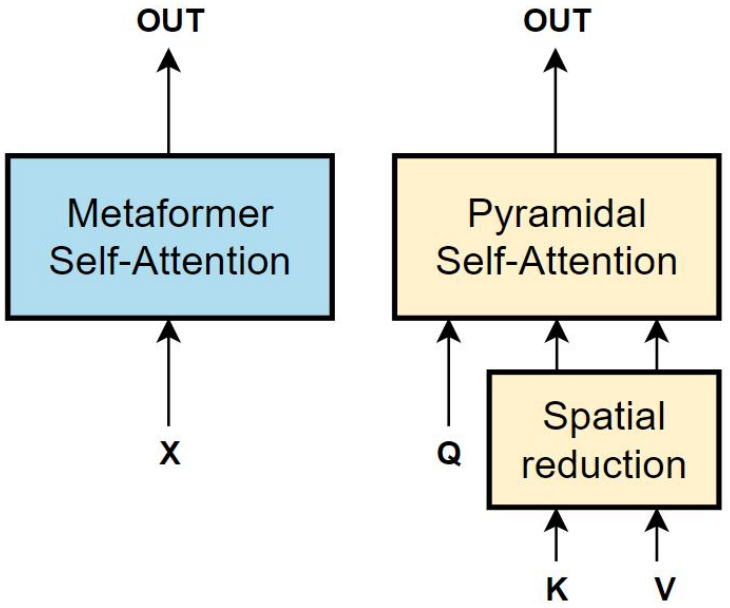




Exploring Our Research - 1

Optimize ViT Architecture via Efficient Attention Modules

- Exploring use of **transformers** models for MDE
  - Customizing METER architecture
- Addressing the attention **quadratic complexity**
  - More challenging when dealing with dense tasks
  - Modifications to the attention mechanism
- Promising results
  - Achieving **faster inference** speeds
  - While maintaining or improving **performance**

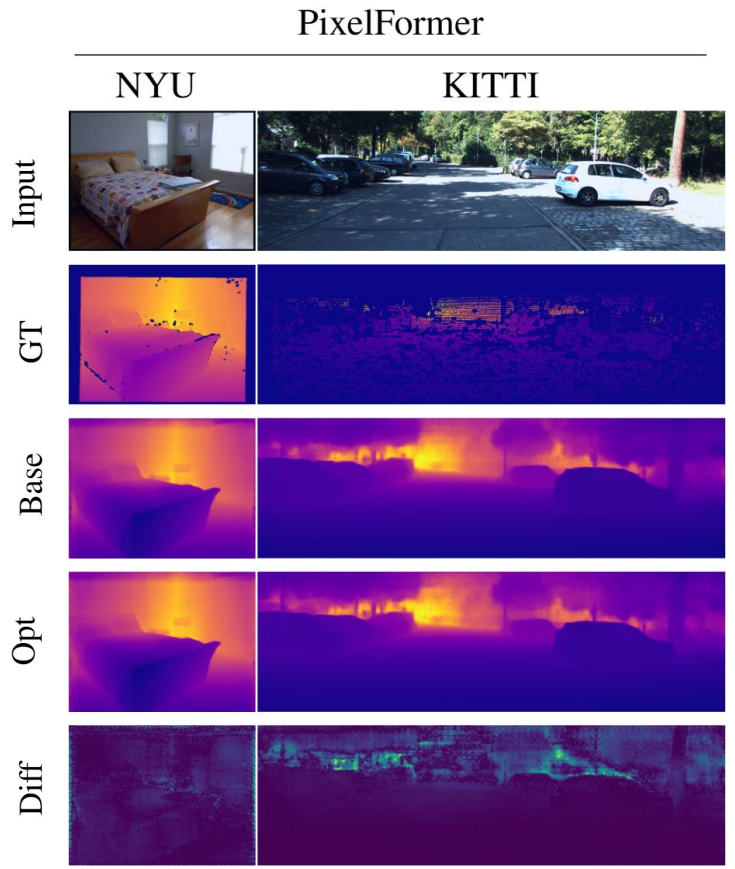


Schiavella, C., Cirillo, L., Papa, L., Russo, P., Amerini, I. Optimize Vision Transformer Architecture via Efficient Attention Modules: A Study on the Monocular Depth Estimation Task. In ICIAP 2023 Workshops

Exploring Our Research - 2

Towards Optimised Networks for MDE on Limited Resources Hardware

- Extending previous findings
  - Optimization of **deep** transformer models
- Studying the impact of attention modifications
  - Modifying **different parts** of the network
  - Building diverse solutions
- Proposing a **standardized** time-performance analysis
  - Optimized decoder often yields the best performance
  - Different optimized models lead to **optimal** trade-offs



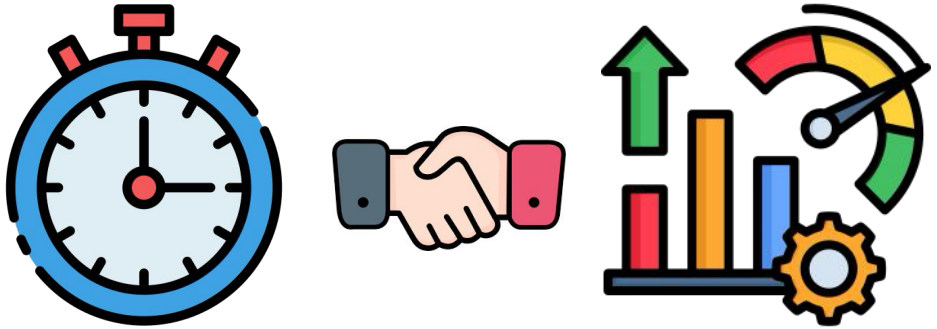
Schiavella, C. & Cirillo, L. & Papa, L: & Russo, P. & Amerini, I. “Towards Optimised Networks for Monocular Depth Estimation on Limited Resources Hardware: An Investigation of Efficient Attention Modules on Transformer-based Architectures”

Efficiency is the key

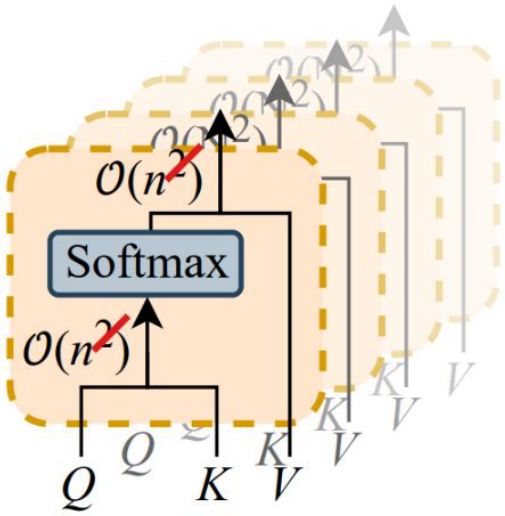
Efficiency in Deep Learning

- **Efficient Training**
  - Reduce the time required to train a model
  - While making the best use of available computational resources.
  - Parallelizing computation, hardware accelerators, optimizing training data, …

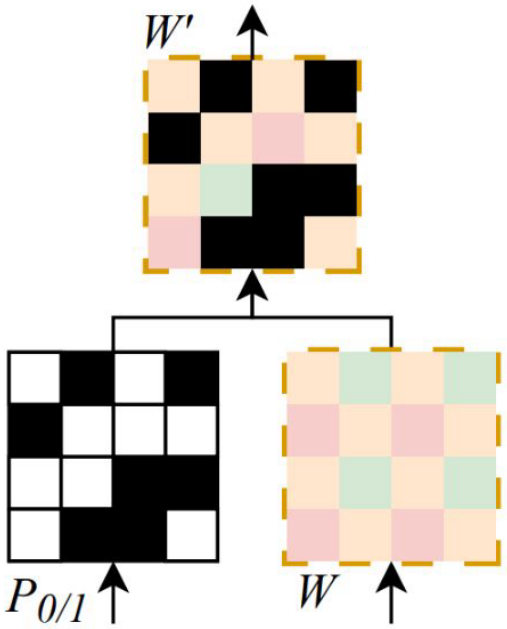
- **Efficient Inference**
  - Running a trained deep learning model on new data quickly
  - Crucial for the practical application of models on embedded systems



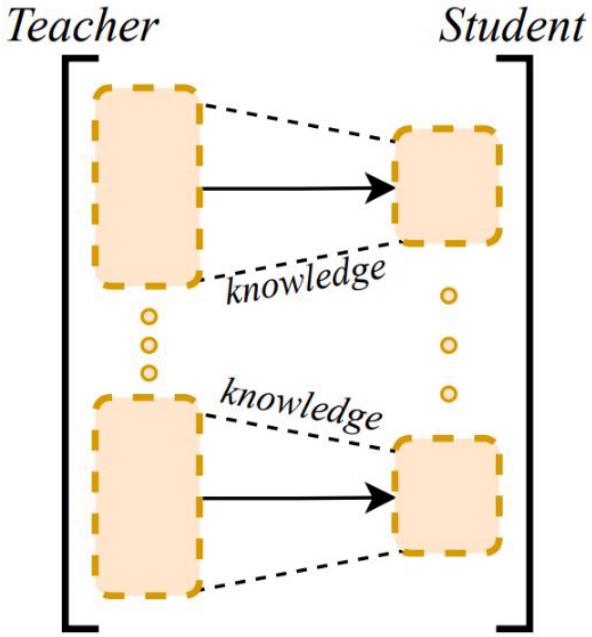
Efficient inference map



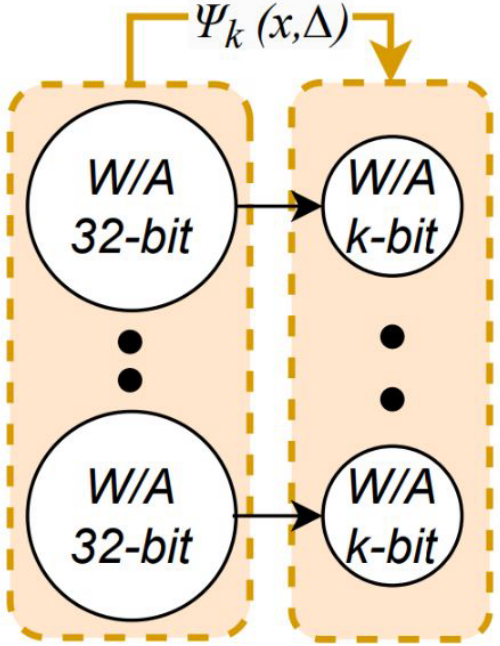
Compact Architecture



Pruning



Knowledge distillation



Quantization

L. Papa, P. Russo, I. Amerini and L. Zhou, "A Survey on Efficient Vision Transformers: Algorithms, Techniques, and Performance Benchmarking," in IEEE Transactions on Pattern Analysis and Machine Intelligence

Compact architecture

- Reduce **computational complexity** and memory requirements
  - Focus on **simplifying** the approach
  - Rather than reducing the size

Attention

$$Sofmax\left(\frac{QK^T}{\sqrt{d_h}}\right)V$$

Meta-Attention

$$Pooling_k\left(x\right)$$

Pyra-Attention

$$Sofmax\left(\frac{Q\ SR(K)^T}{\sqrt{d_h}}\right)\ SR(V)$$

Quadratic

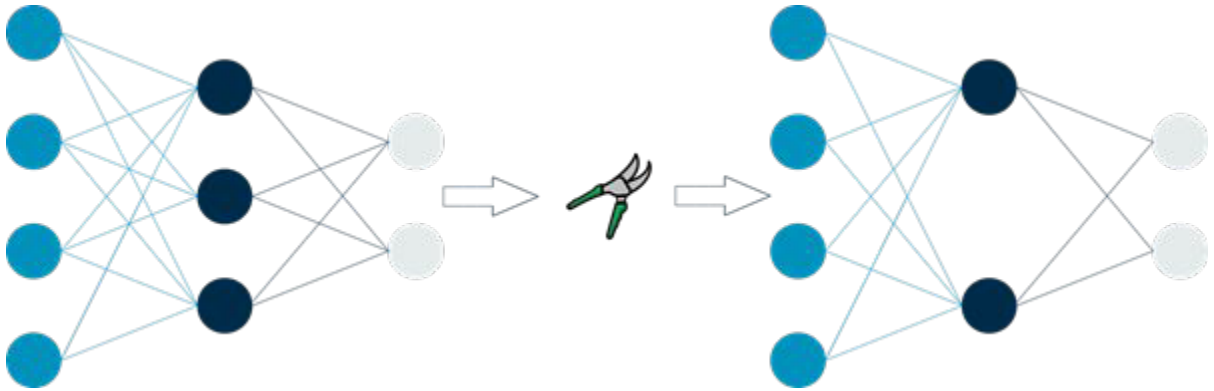
Linear

W. Yu et al., "MetaFormer is Actually What You Need for Vision," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022
W. Wang et al., "Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions," 2021 IEEE/CVF International Conference on Computer Vision (ICCV)

Pruning

- Selectively **removing** neurons or connections
  - That **contribute less** to the network's performance
  - While maintaining or even improving its accuracy

- Create a more compact and efficient network **without sacrificing performance**
  - Remove redundant connections or neurons

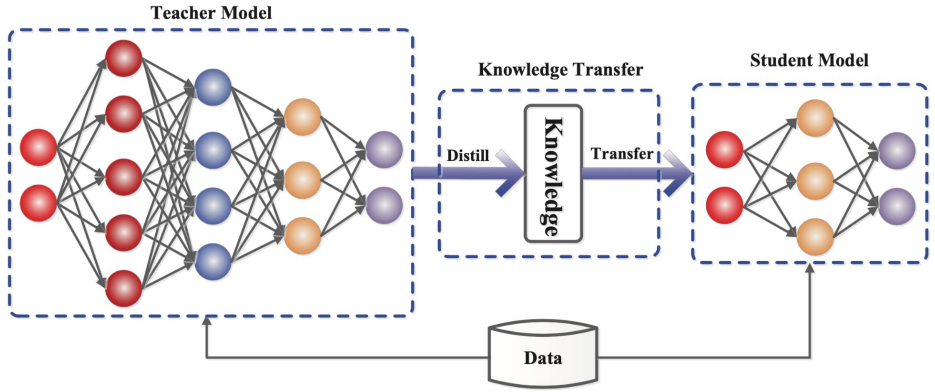


H. Cheng, M. Zhang and J. Q. Shi, "A Survey on Deep Neural Network Pruning: Taxonomy, Comparison, Analysis, and Recommendations," in IEEE Transactions on Pattern Analysis and Machine Intelligence

Knowledge distillation

- Process of **transferring** knowledge
  - From a **large**, complex model (teacher model)
  - To a **smaller**, more compact model (student model).

- Train the student model to mimic the behavior of the teacher model
  - Capturing the essential knowledge encoded in the teacher model
  - With reduced computational complexity and memory footprint.

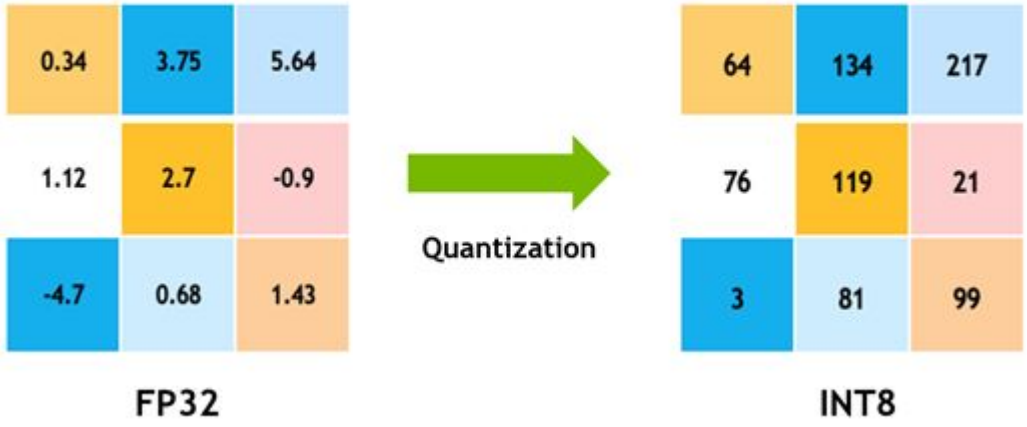


Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network.

Quantization

- **Reducing** numerical precision of parameters and activations within the network
  - **Converting** floating-point values to lower precision fixed-point or integer

- Represent the network parameters and activations using **fewer bits**
  - While minimizing the loss in model accuracy.
  - And enabling significant reductions in memory usage and complexity



Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., … Kalenichenko, D. (2018, June). Quantization and Training of Neural Networks for Efficient Integer-Arithmetic-Only Inference. Proceedings of the IEEE CVPR

Metrics

- **Multiply and Accumulate operations (MACs)**
  - Number of multiplications and additions performed to compute the output
- **Floating point operations (FLOPs)**
  - Number of arithmetic operations involving floating-point numbers
- **Memory Footprint**
  - Amount of memory required to store all model parameters
- **Inference/Training time** on a specific hardware
  - When on GPU need *torch.cuda.synchronize()*
  - To ensure that all operations are completed before measuring the time

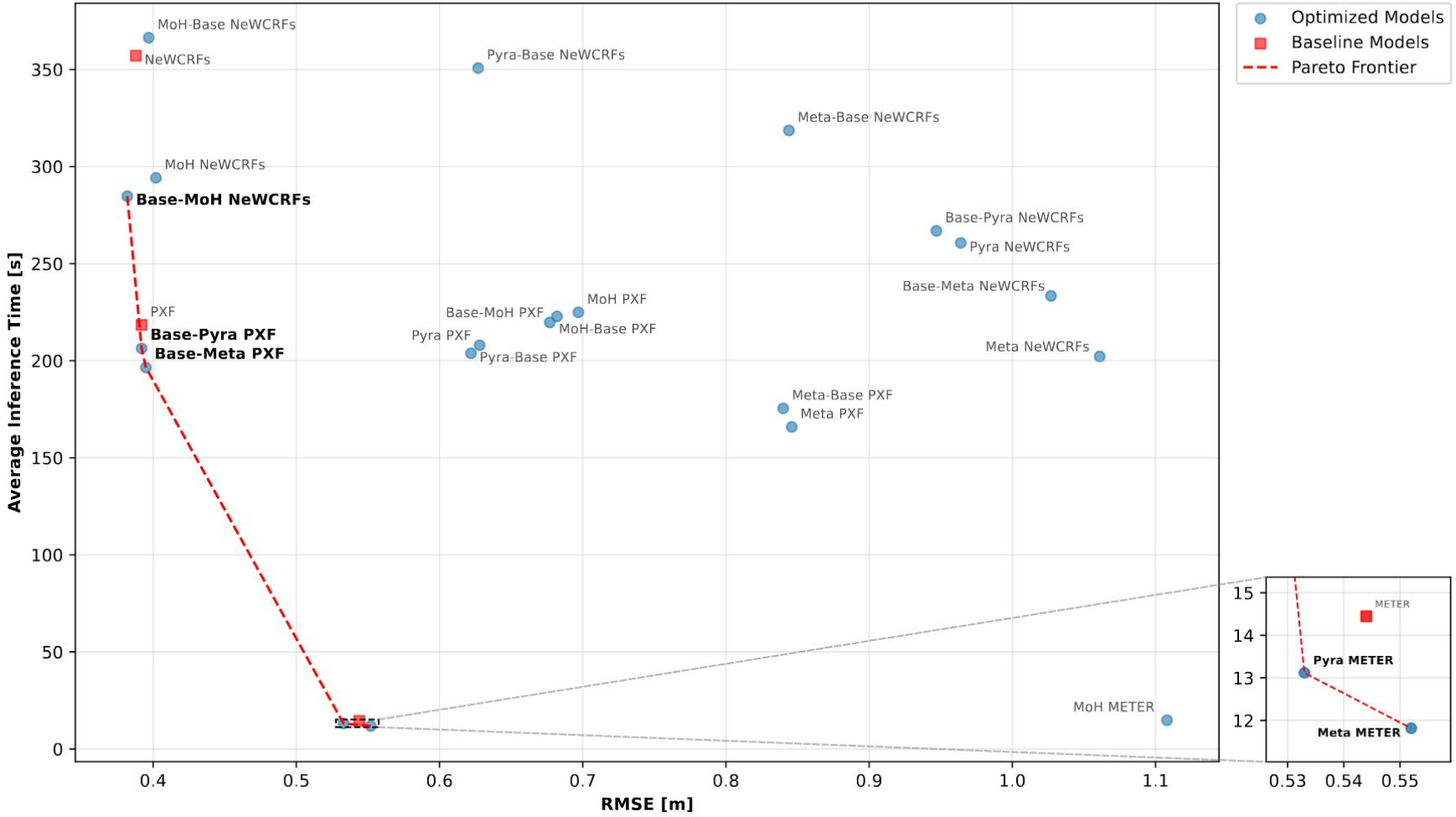
Time-Performance tradeoff

- Changing the **method** changes the **results**
  - Any modification to the model architecture directly impacts performance
  - Crucial to quantify the **trade-offs** between efficiency and quality
- **Pareto frontier** - **Revisited by ALCOR Lab**
  - Identify the optimal set of solutions (frontier)
  - That represent the best trade-offs in a multi-objective optimization problem.
  - Graphically, the further a model is from the frontier, the worse the trade-off
- **Efficient Error Rate (EER)** - **Developed by ALCOR Lab**
  - Combines all efficiency metrics into a single value for easier comparison.
  - Comparing an optimized model M_i with a reference model R_i

$$EER = \frac{1}{\|i\|} \cdot \sum_i \left(\frac{M_i}{R_i} \right) \quad i \in \{\#Par., \text{FLOPs}, \ldots\}$$

L. Papa, P. Russo, I. Amerini and L. Zhou, "A Survey on Efficient Vision Transformers: Algorithms, Techniques, and Performance Benchmarking," in IEEE Transactions on Pattern Analysis and Machine Intelligence

Time-Performance tradeoff



Schiavella, C. & Cirillo, L. & Papa, L: & Russo, P. & Amerini, I. “Towards Optimised Networks for Monocular Depth Estimation on Limited Resources Hardware: An Investigation of Efficient Attention Modules on Transformer-based Architectures”

That's all!
Thank you for the attention