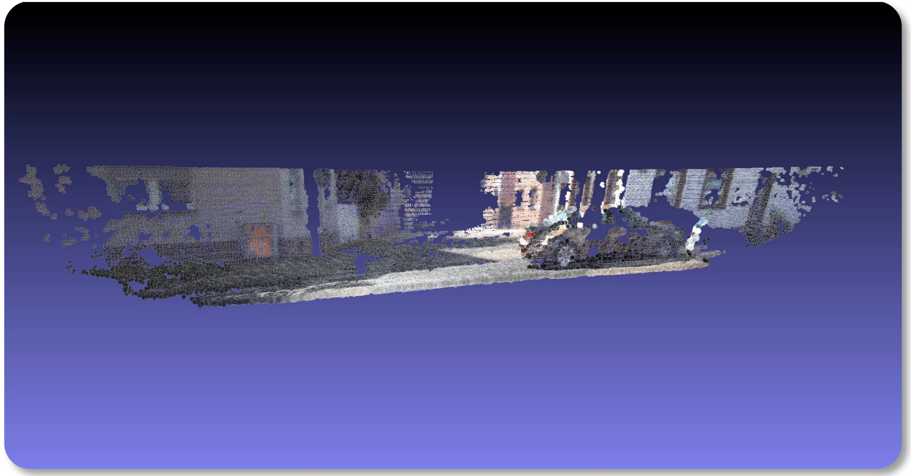
Introduction to MDE

Jaime Spencer
@jaimespencer06 @depthchallenge

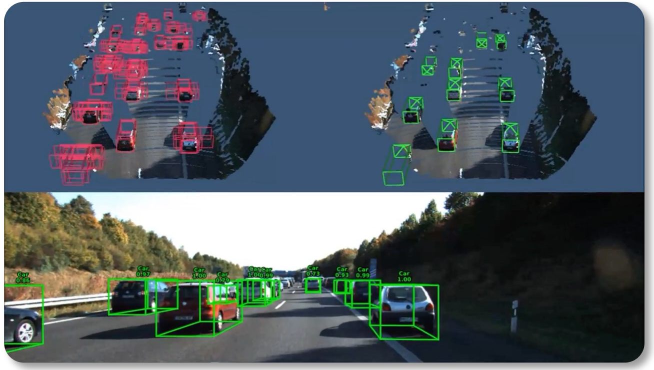
CVSSP Centre for Vision, Speech and Signal Processing

1. What is Depth Estimation?
2. Stereo vs. Monocular
3. Self-supervision via View Synthesis
4. State-of-the-Art Review

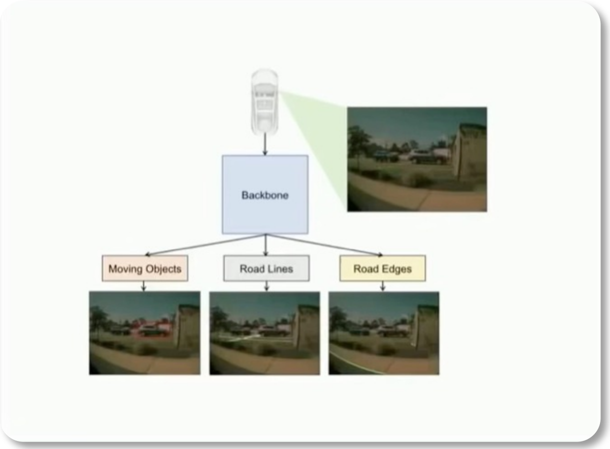
– Depth estimation is the process of reconstructing the **3D geometry** of the scene from its **2D image projection(s)**



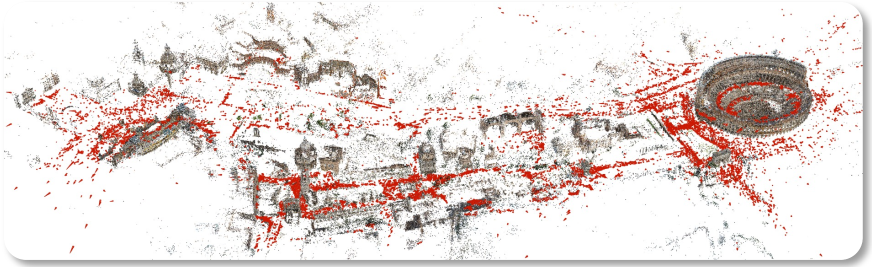
– Core component of **mid/high-level** computer vision tasks



3D Object Detection (AVOD)



BEV Mapping (Tesla)



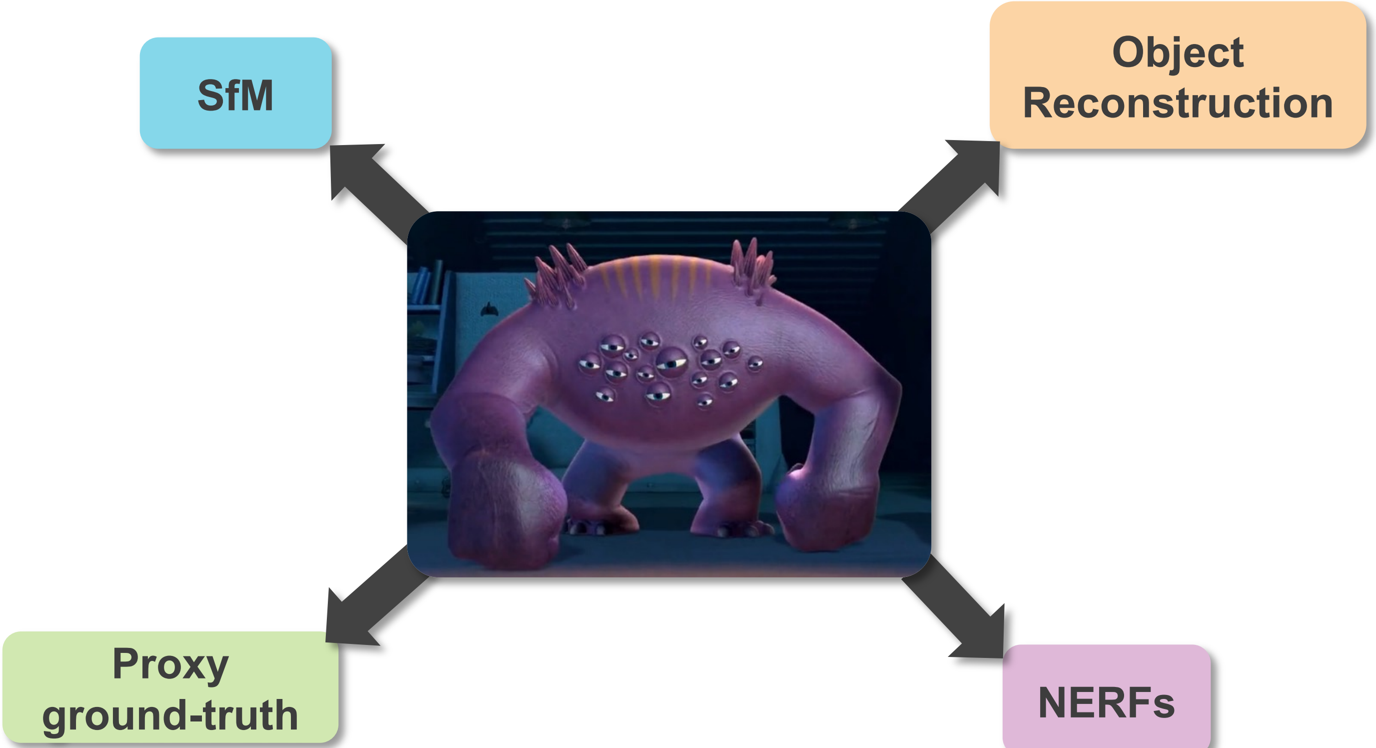
Structure-from-Motion (COLMAP)

– Depth estimation comes in **many forms!**

Multi-view

Stereo

Monocular


SfM
Object Reconstruction
Proxy ground-truth
NERFs

Disparity

Optical Flow

Proxy ground-truth

SLAM

AR

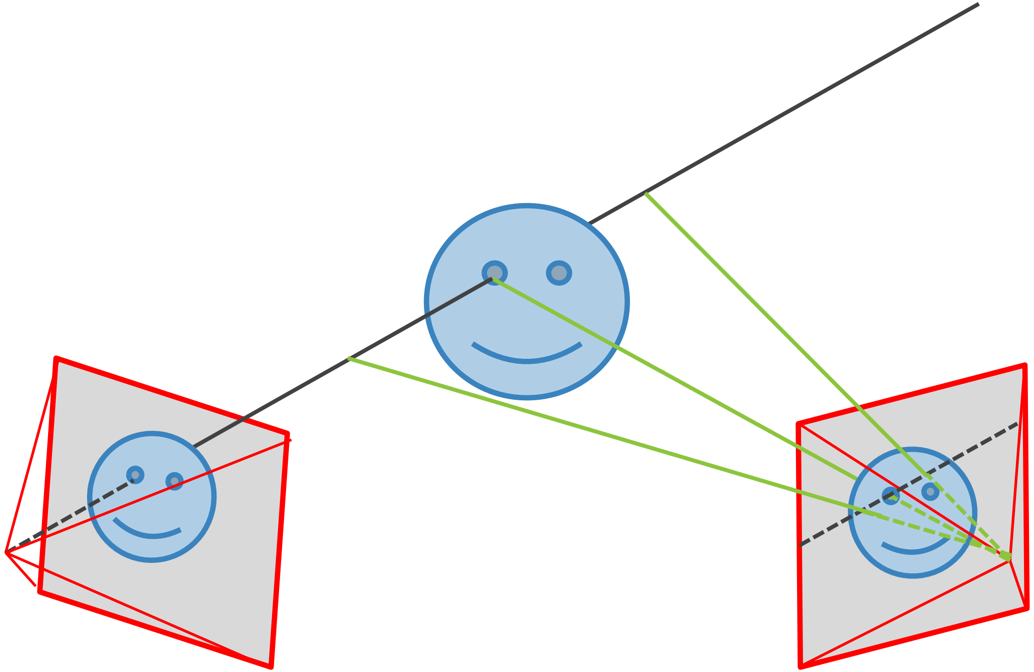
SLAM

Proxy LiDAR

Sensor Fusion

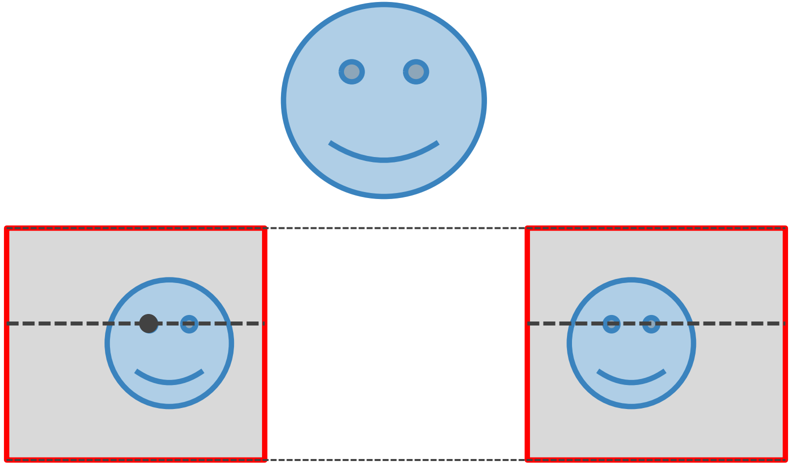
Stereo

– Depth estimation as **correspondence estimation** and **triangulation**



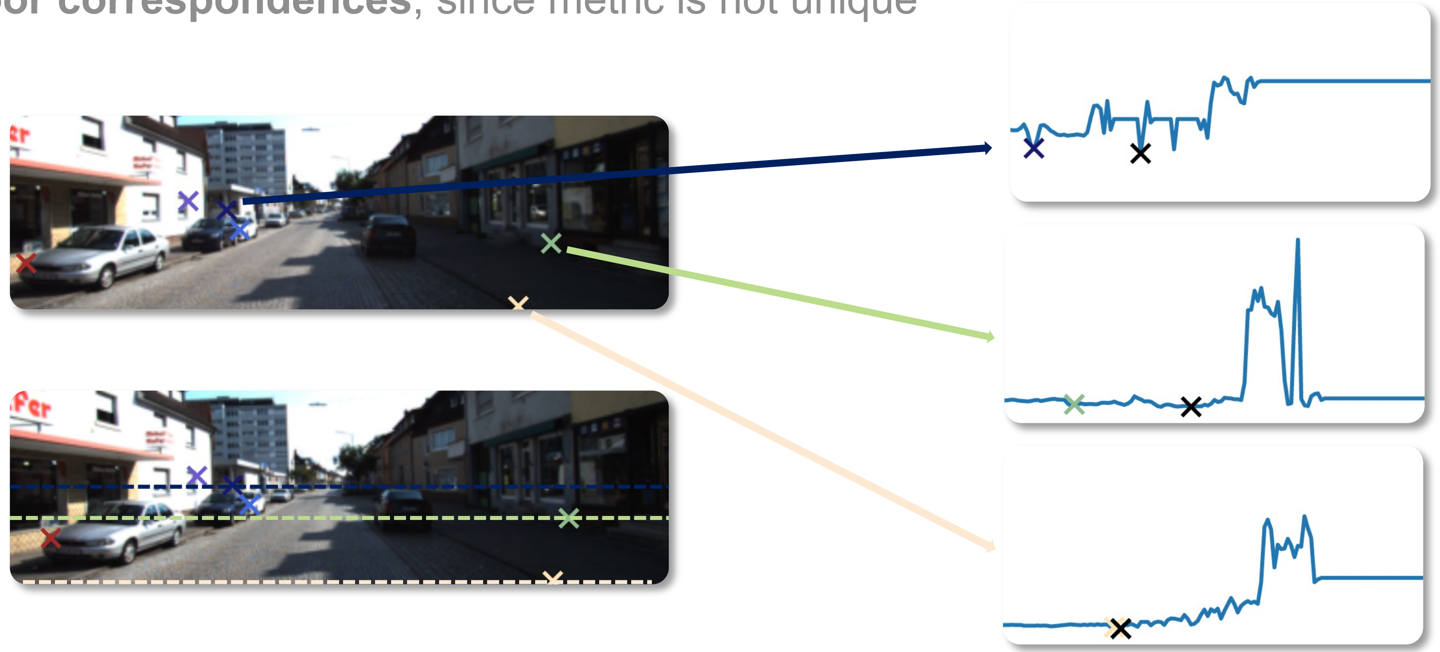
– Depth estimation as **correspondence estimation** and **triangulation**

➢ **Stereo rectified** ⟶ Correspondence lies in **horizontal scanline**!



– Simplest matching uses **photometric error** between pixels

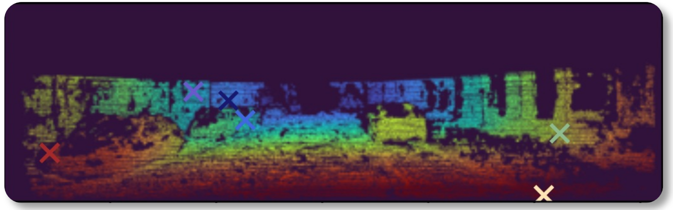
➢ **Poor correspondences**, since metric is not unique

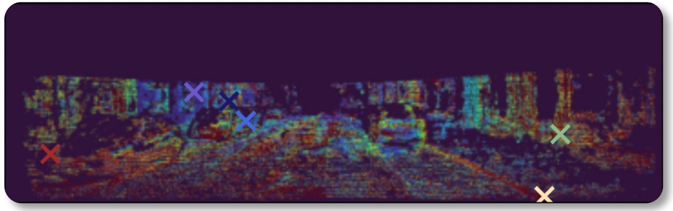


– Simplest matching uses **photometric error** between pixels

➢ **Poor correspondences**, since metric is not unique

$$|I(p) - \hat{I}(p + h)|$$





– How to deal with this?

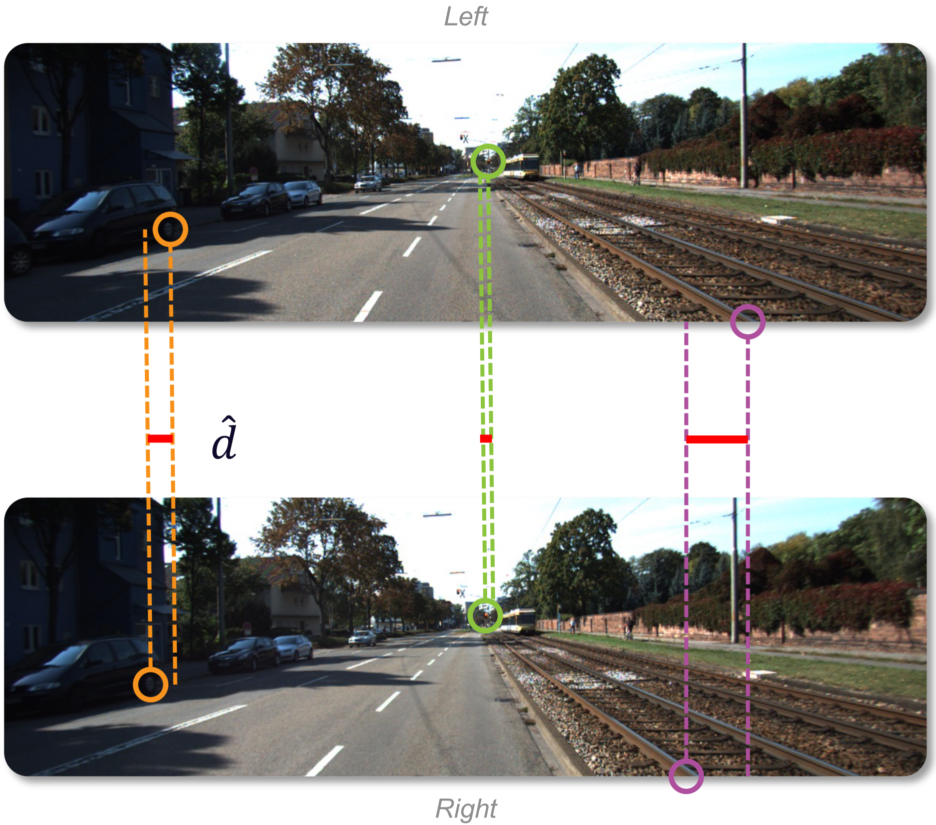
- Improve **similarity** metric → SSIM, descriptors…
- Add **priors** to cost volume → Smoothness, surface normals…

➢ **Semi-Global Block Matching** is commonly used to generate proxy depth

- Technically, we are predicting **pixel disparity**
  - ➢ **Inverse parametrization** of depth, more stable

Focal Length *Stereo Baseline*

$$d = \frac{fb}{\hat{d}}$$



- DeepStereo: Learning to Predict New Views from the World's Imagery

  Flynn et al., CVPR16

- Pyramid Stereo Matching Network

  Chang & Chen, CVPR18

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

  Li et al., ICCV19

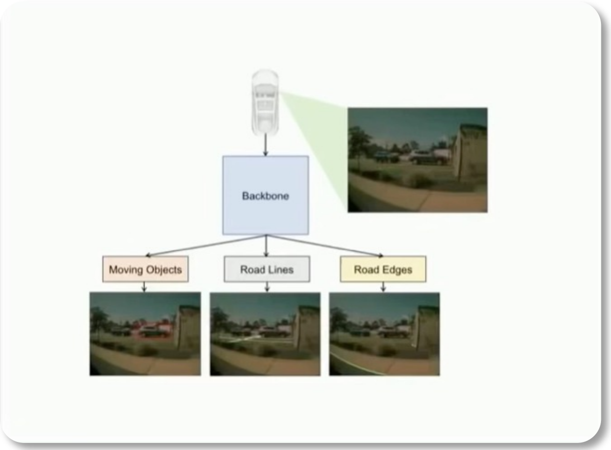
- Attention Concatenation Volume for Accurate and Efficient Stereo Matching

  Xu et al., CVPR22

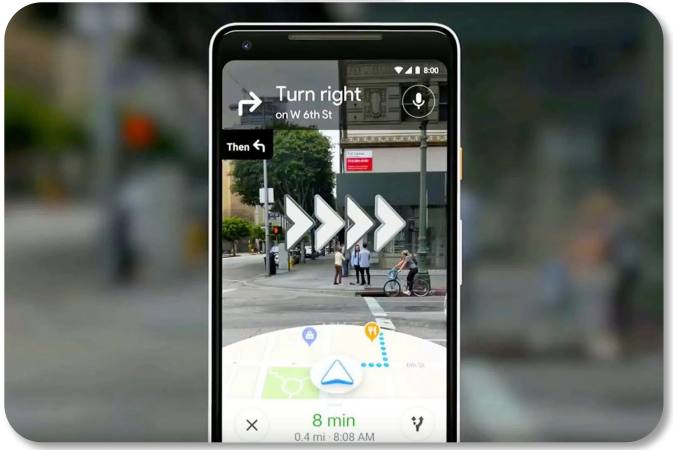
Monocular

– Why **monocular**?

- Cheap & flexible
- Real-world deployment

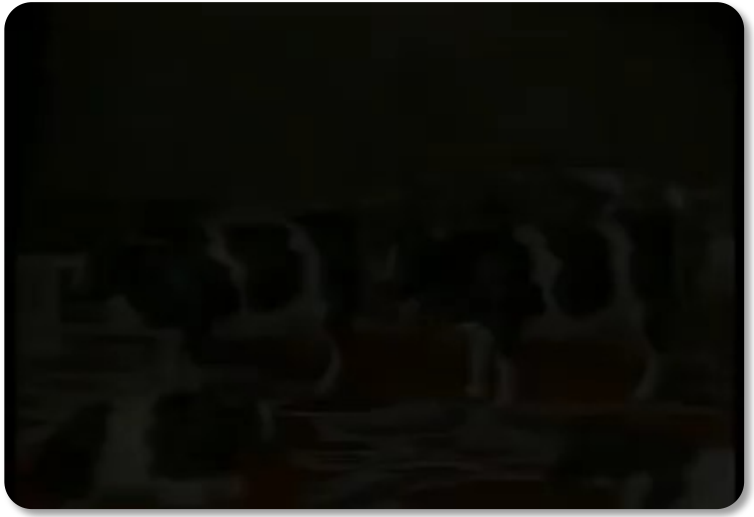


Tesla

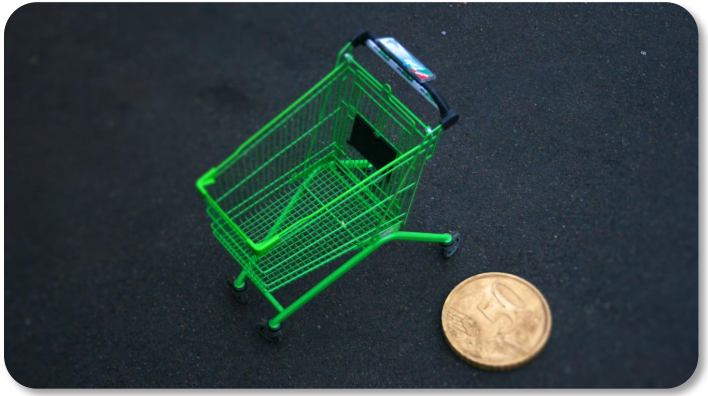


Google AR Maps

– Monocular depth estimation is an **ill-posed problem**

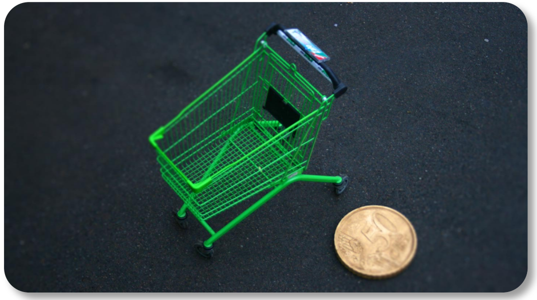


– Monocular depth estimation is an **ill-posed problem**



– Monocular depth estimation is an **ill-posed problem**

Skrekkogle



- Humans can take advantage of **priors**
    - Absolute/relative object size
    - Elevation
    - Perspective and horizon
    - Stereo/motion parallax
    - Texture gradient

- Network must **learn these geometric priors!**
  Not just rely on correspondences

- Humans can take advantage of **priors**
    - Absolute/relative object size
    - Elevation
    - Perspective and horizon
    - Stereo/motion parallax
    - Texture gradient

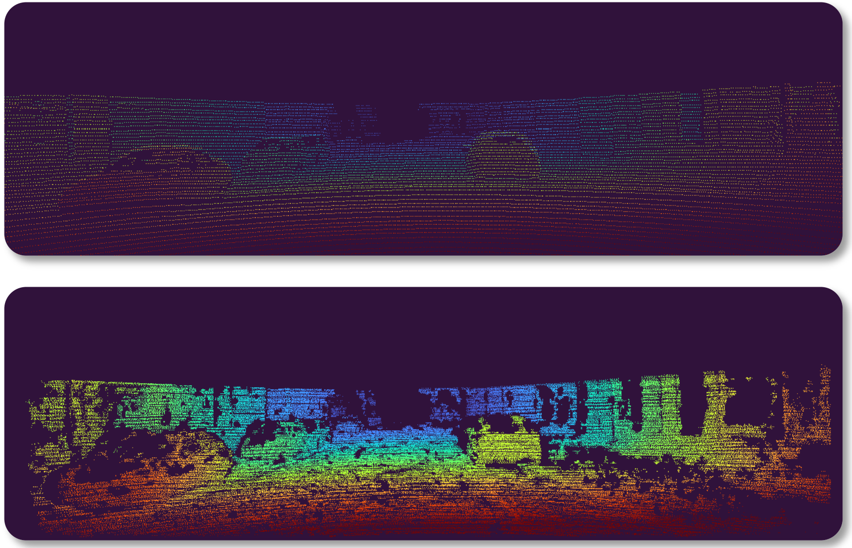
- Network must **learn these geometric priors!**
  Not just rely on correspondences

- Humans can take advantage of **priors**
  - Absolute/relative object size
  - Elevation
  - Perspective and horizon
  - Stereo/motion parallax
  - Texture gradient

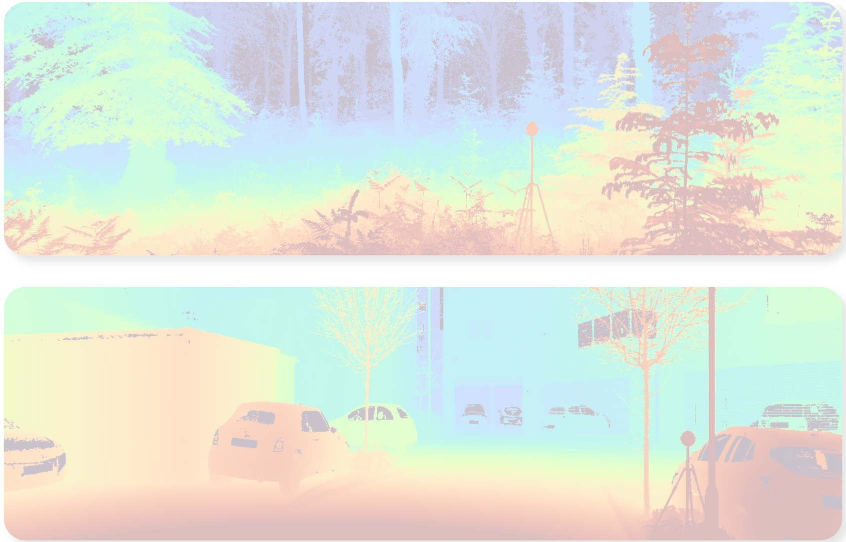
- Network must **learn these geometric priors!**
  Not just rely on correspondences

– Supervised learning with LiDAR, SfM, SLAM…

➢ Collecting this data is **challenging** and **expensive**



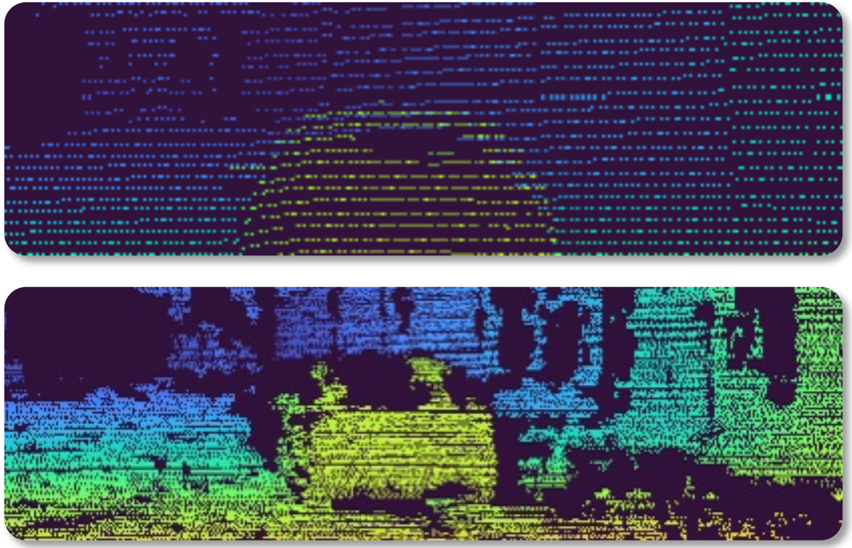
Kitti



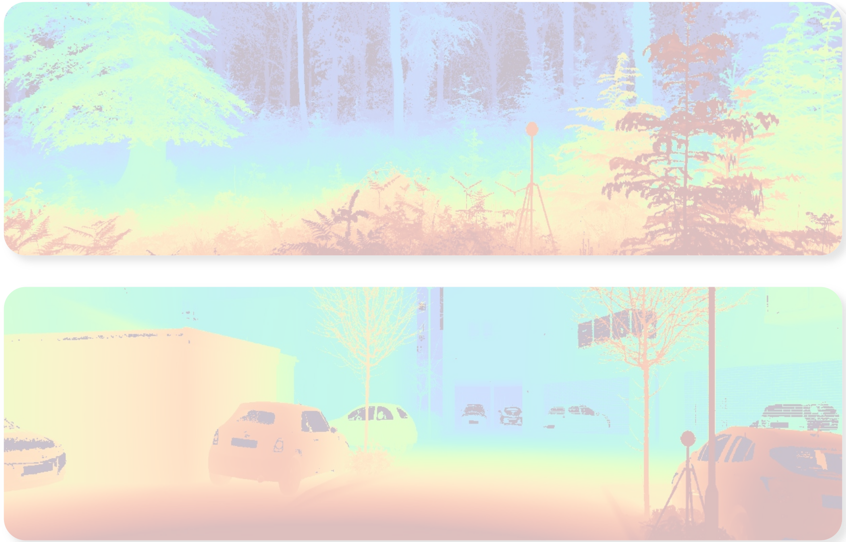
SYNS-Patches

– Supervised learning with LiDAR, SfM, SLAM…

➢ Collecting this data is **challenging** and **expensive**



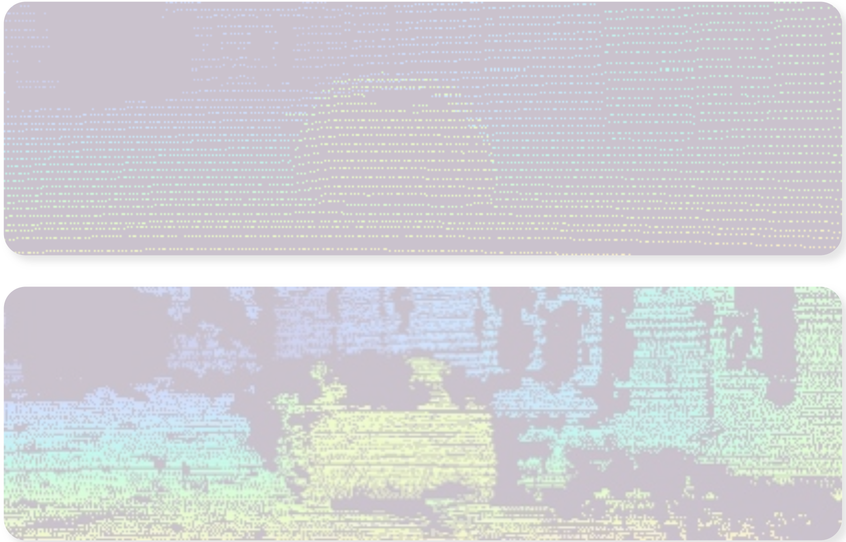
Kitti



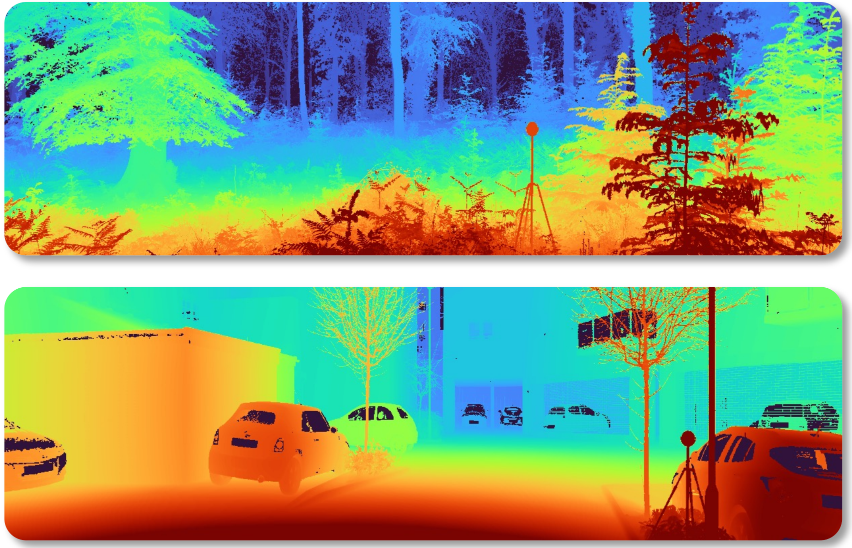
SYNS-Patches

– Supervised learning with LiDAR, SfM, SLAM…

➢ Collecting this data is **challenging** and **expensive**



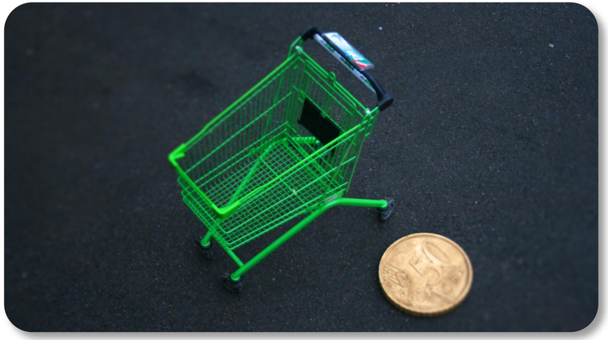
Kitti



SYNS-Patches

- Supervised learning with LiDAR, SfM, SLAM…
  - ➢ Collecting this data is **challenging** and **expensive**

- Let’s go **self-supervised!**
  - ➢ In both cases we predict **sigmoid disparity**, applying arbitrary scale



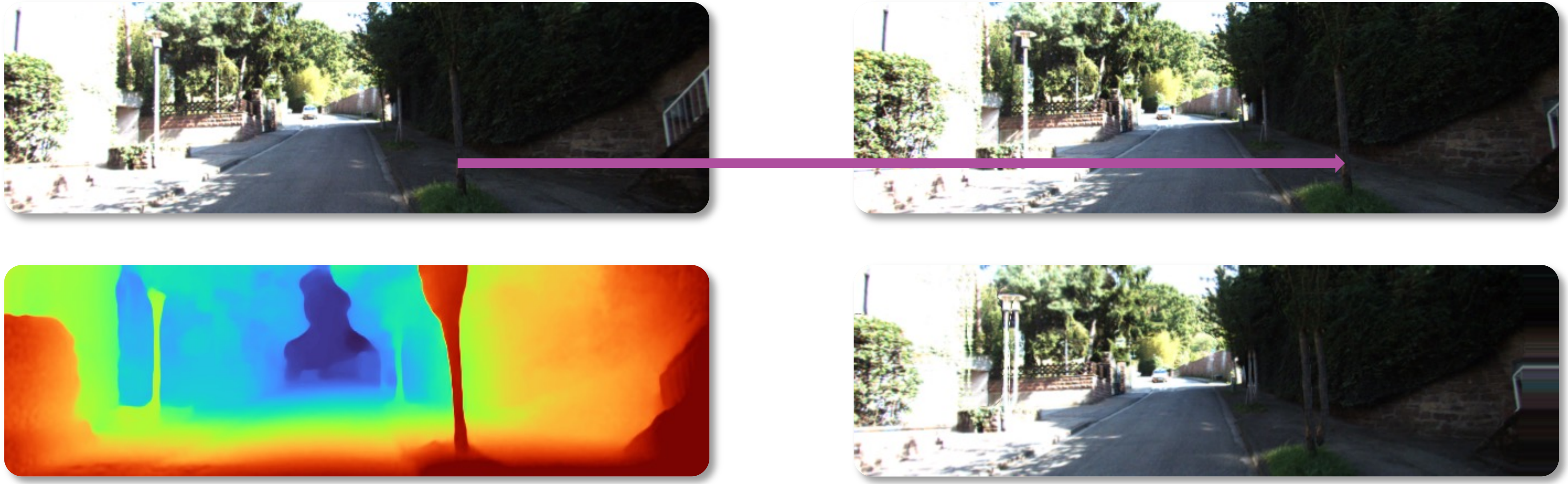
– How to train self-supervised then?

- Stereo/motion **parallax**
- Reconstruct target view + **photometric error**

- In stereo, correspondence **must** lie on **horizontal scanline**
- Disparity ⟶ Correspondence ⟶ Photometric relationship

$$I(p) = \hat{I}(p + \hat{d})$$

Left *Right*



- In stereo, correspondence **must** lie on **horizontal scanline**
- Disparity $\longrightarrow$ Correspondence $\longrightarrow$ Photometric relationship

$$|I(p) - \hat{I}(p + \hat{d})|$$

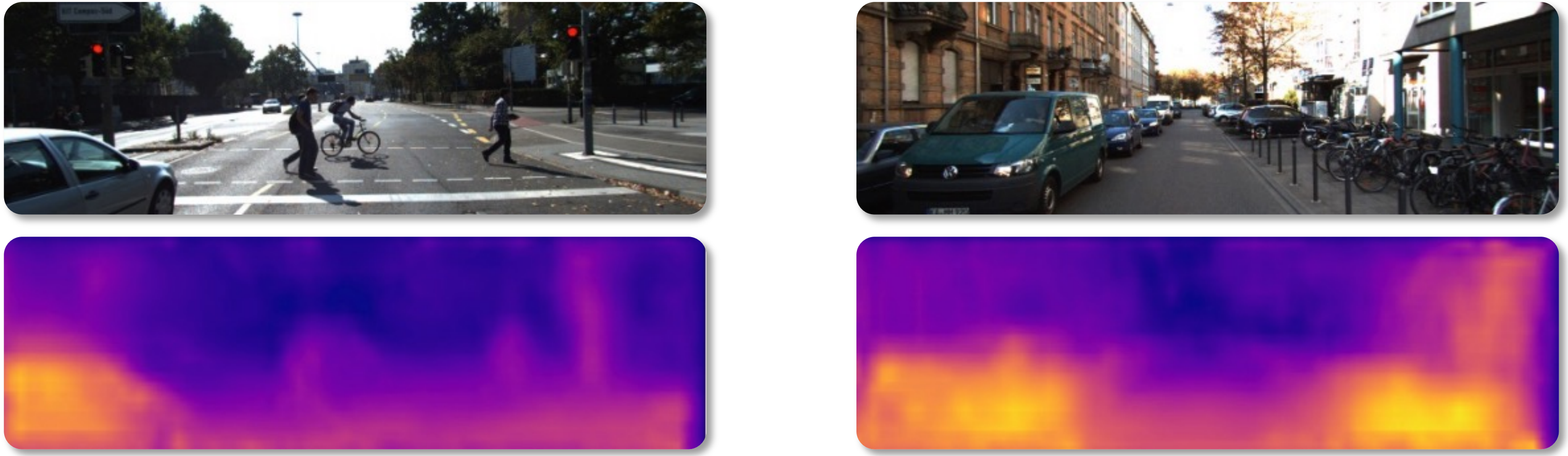
– **Garg et al.** used this procedure to train first self-supervised CNN

➢ U-Net based on **AlexNet**, implemented in **Caffe** with custom layers

Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue, Garg et al, ECCV16

– **Garg et al.** used this procedure to train first self-supervised CNN

➢ U-Net based on **AlexNet**, implemented in **Caffe** with custom layers



Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue, Garg et al, ECCV16

– **Monodepth** modernized Garg + other contributions

- Spatial Transformer Networks (**STN**)
- Structural similarity error (**SSIM**)

Unsupervised Monocular Depth Estimation with Left-Right Consistency, Godard et al, CVPR17

A Digression

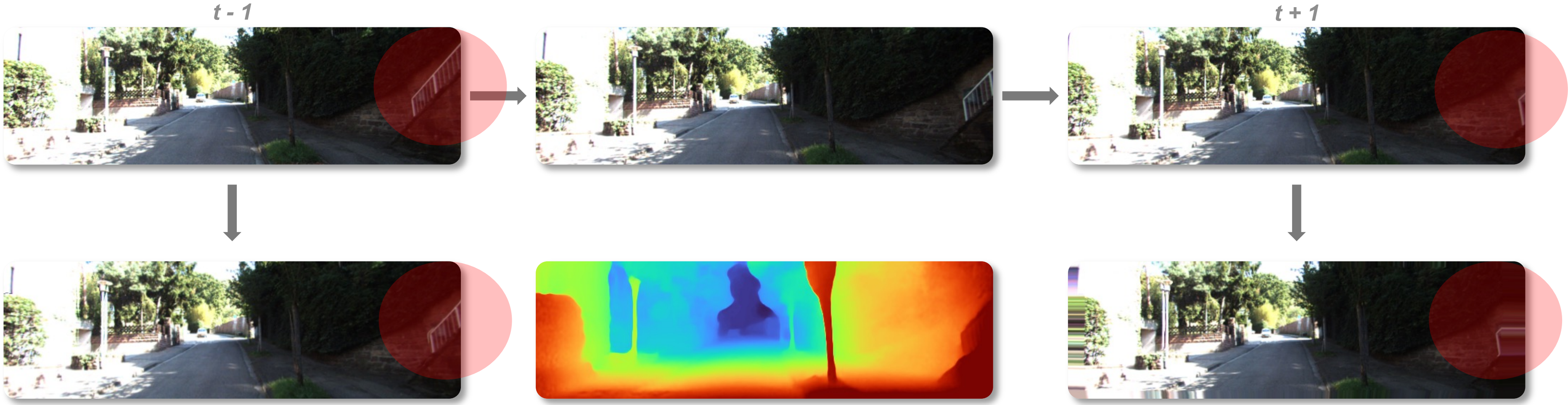
– How to generalize to **monocular video** streams?

➢ Replace known stereo baseline with **pose prediction network**!

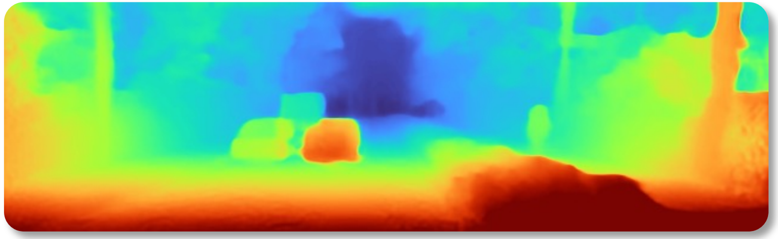
– Correspondences & view synthesis now depend on…

projective geometry

– Requires knowledge of **camera intrinsics**
– Plus network to predict **relative motion** between frames (VO)
➢ Sensitive to **dynamic objects**!

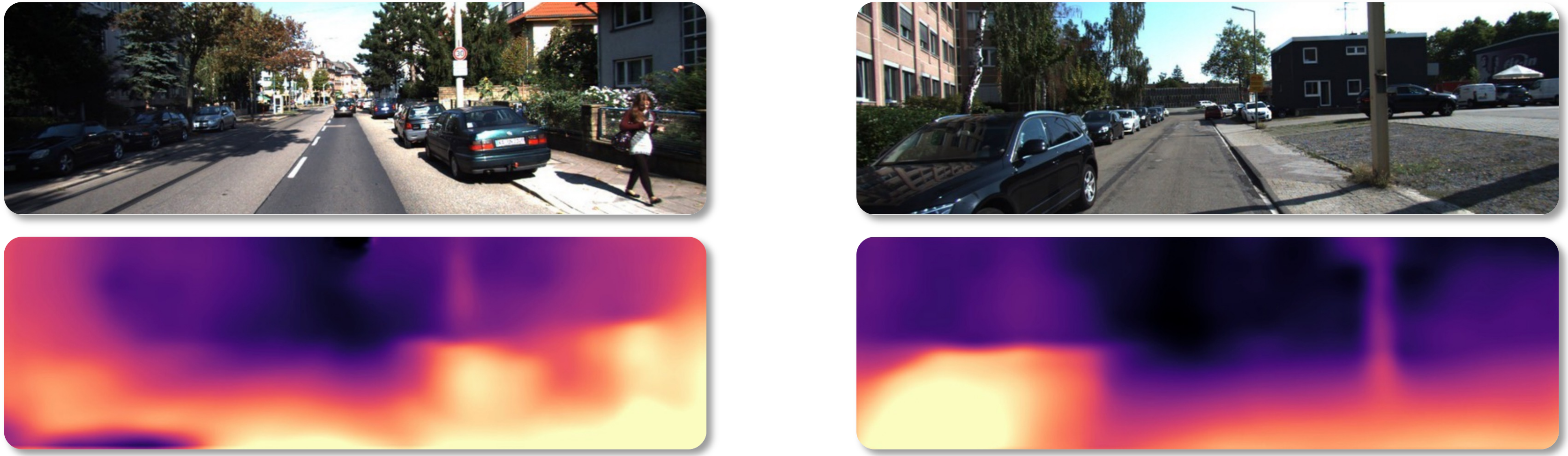


– Requires knowledge of **camera intrinsics**
– Plus network to predict **relative motion** between frames (VO)
➢ Sensitive to **dynamic objects**!



– **SfM-Learner** was the first to apply these concepts

➢ Added **explainability mask** to account for dynamic objects



Unsupervised Learning of Depth and Ego-Motion from Video, Zhou et al, CVPR17

Above Us Only Sky

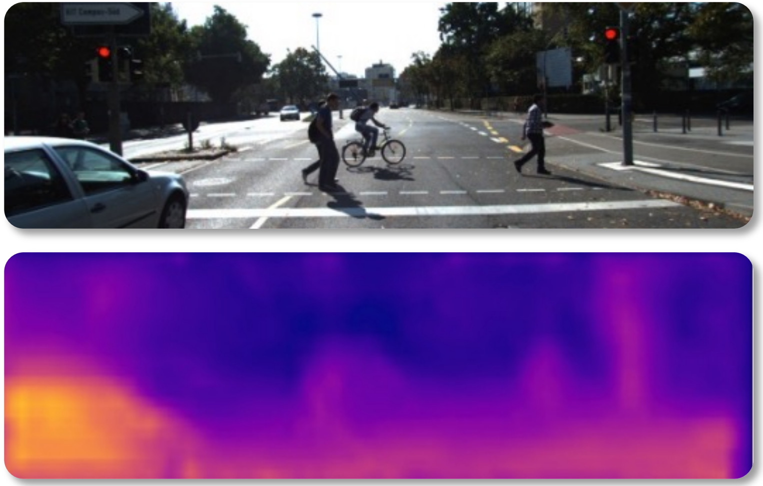
Base
- Garg
- Monodepth
- SfM-Learner

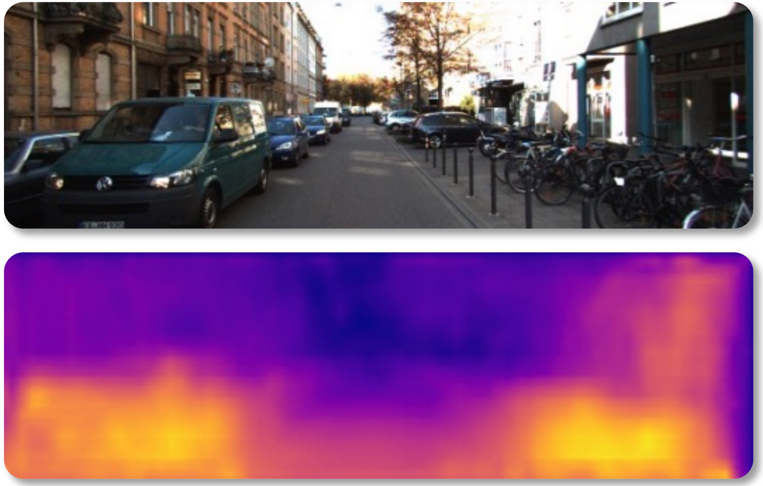
Proxy Supervision
- Kuznietsov
- DVSO
- MonoResMatch
- DepthHints

Improved Photometric
- Monodepth2
- D3VO
- Depth-VO-Feat
- DeFeat-Net
- Feat-Depth

Architecture
- PackNet
- CADepth
- Johnston
- DiffNet
- HR-Depth

- **Garg:** Stereo view synthesis (S) + **Smoothness** prior
- U-Net based on **AlexNet**, implemented in **Caffe** with custom layers

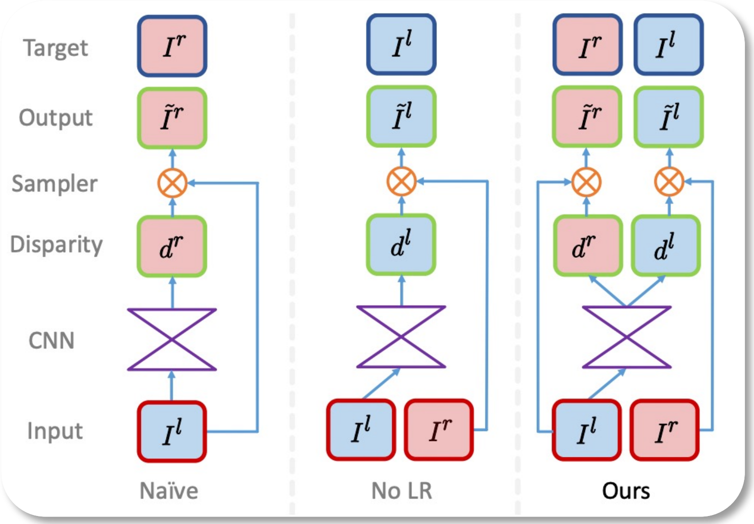




Unsupervised CNN for Single View Depth Estimation: Geometry to the Rescue, Garg et al, ECCV16

– **Monodepth:** S + **Edge-aware** smoothness + **Virtual** stereo

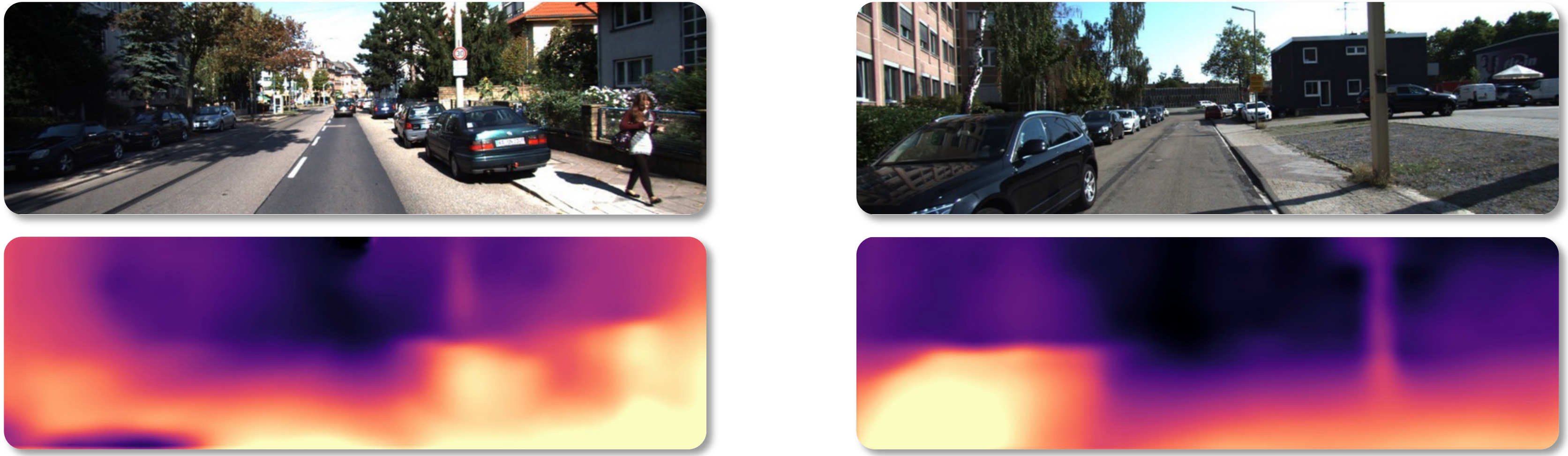
➢ Spatial Transformers + SSIM play large role in improvements



Unsupervised Monocular Depth Estimation with Left-Right Consistency, Godard et al, CVPR17

– **SfM-Learner:** Mono (M) + **PoseNet** + **Explainability** mask

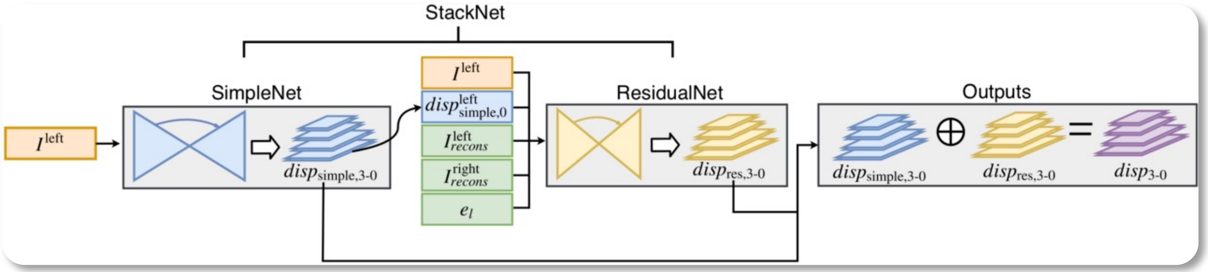
➢ Pose representation as **Euler** & “bug” in smoothness prior



Unsupervised Learning of Depth and Ego-Motion from Video, Zhou et al, CVPR17

- **DVSO**: S + **Refinement** + Virtual Stereo + Proxy regression (**SLAM**) + **Occlusion** regularization

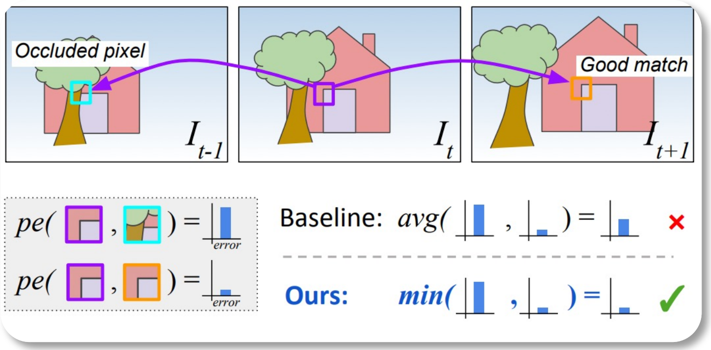
- **MonoResMatch** : S + **Refinement** + Virtual Stereo + Proxy regression (**SGBM**)

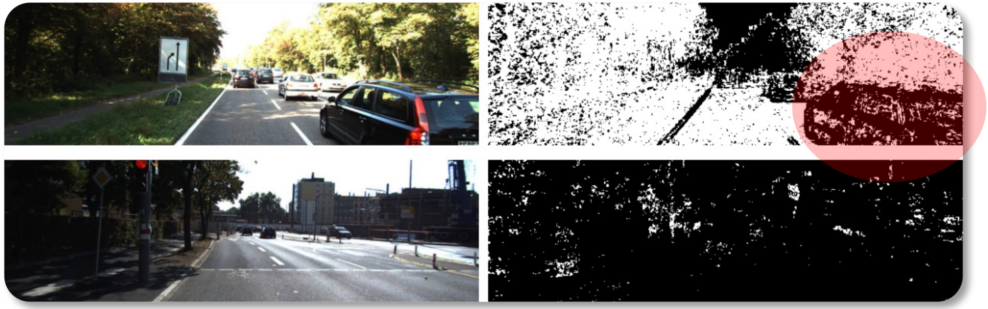


Deep Virtual Stereo Odometry: Leveraging Deep Depth Prediction for Monocular Direct Sparse Odometry, Yang et al, ECCV18
Learning monocular depth estimation infusing traditional stereo knowledge, Tosi et al, CVPR19

– **Monodepth2:** MS + **Min reconstruction** loss + **Automasking**

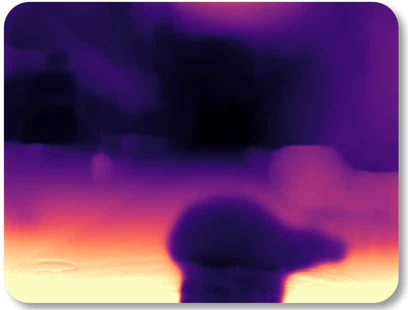
➢ **Upsampled multi-scale** losses



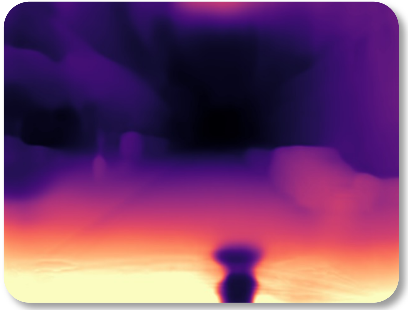


Digging into Self-Supervised Monocular Depth Estimation, Godard et al, ICCV19

- **Depth-VO-Feat:** MS + Feature synthesis (**pretrained**)
- **DeFeat-Net:** MS + Feature synthesis (**co-trained**)
- **FeatDepth:** MS + Feature synthesis (**autoencoder**) + Feature smoothness



Monodepth2

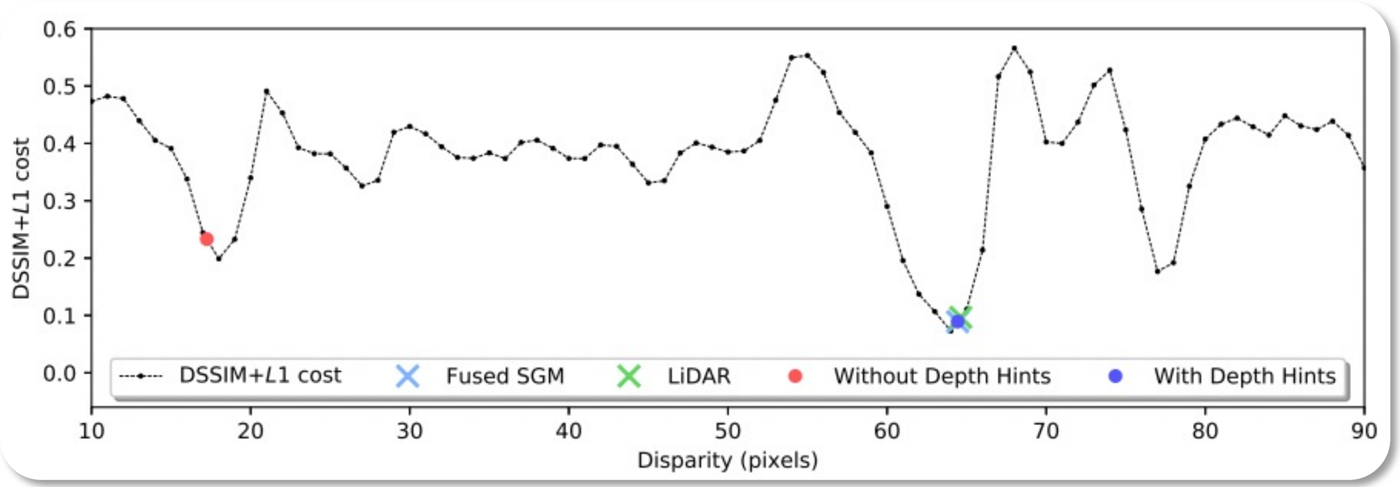


DeFeat-Net

Unsupervised Learning of Monocular Depth Estimation and Visual Odometry with Deep Feature Reconstruction, Zhan et al, CVPR18
DeFeat-Net: General Monocular Depth via Simultaneous Unsupervised Representation Learning, Spencer et al, CVPR20
Feature-metric Loss for Self-supervised Learning of Depth and Egomotion, Shu et al, ECCV20

– **DepthHints:** MS + Proxy regression (**fused SGBM**) + Automasking

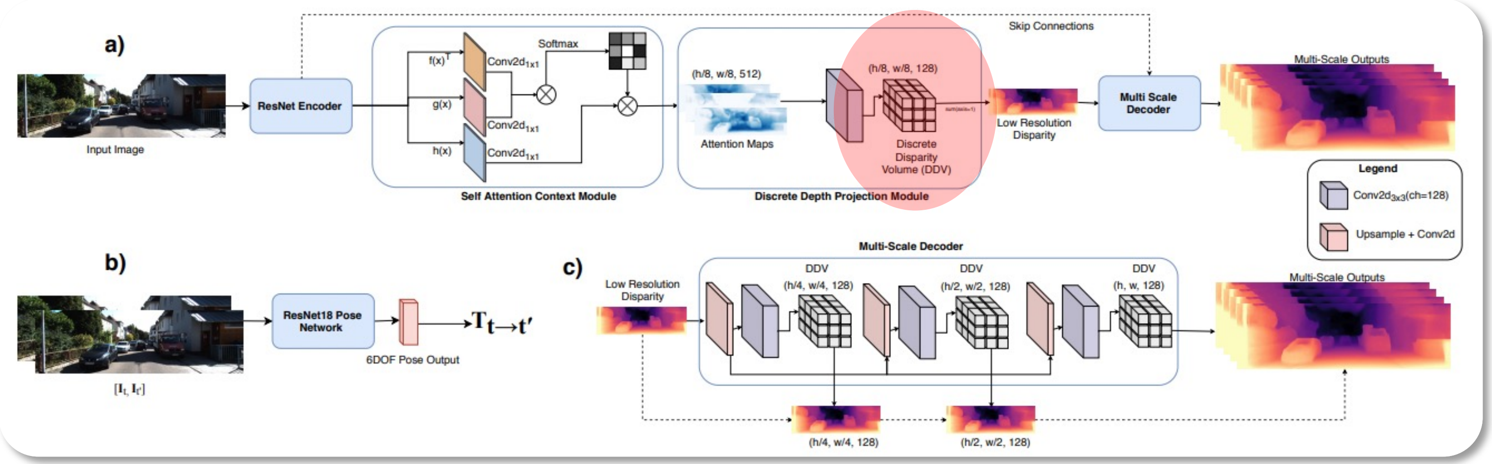
➢ Incorporate min reconstruction into proxy ground-truth generation



Self-Supervised Monocular Depth Hints, Watson et al, ICCV19

– **Johnson:** M + **Discrete** disparity volume + **Self-attention**

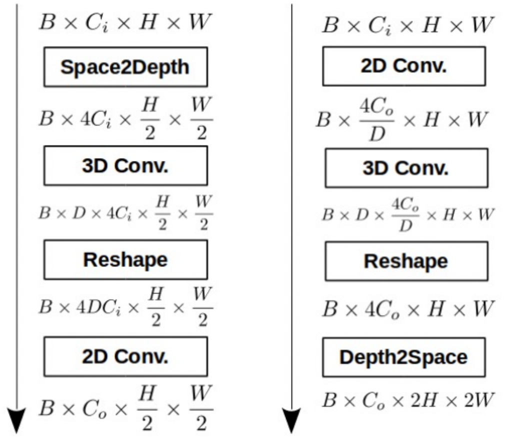
➢ Final disparity given by **Expected** value (weighted sum)



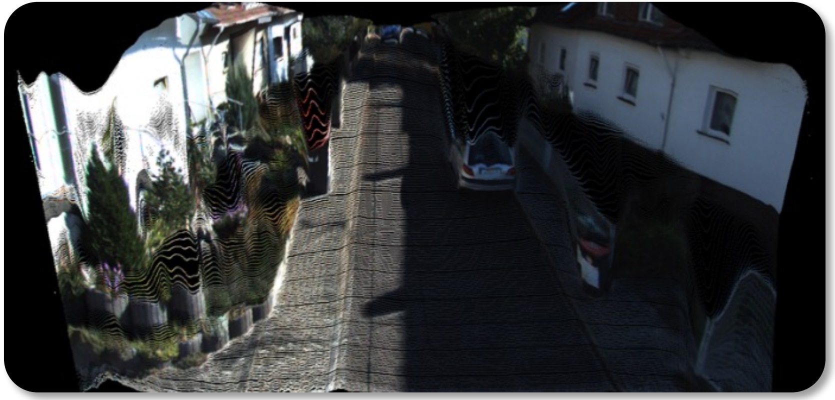
Self-supervised Monocular Trained Depth Estimation using Self-attention and Discrete Disparity Volume, Johnston & Carneiro, CVPR20

– **PackNet:** M + **Speed** loss + **3D (un)packing** architecture

➢ Speed loss as a cheap way of constraining **metric scale**!



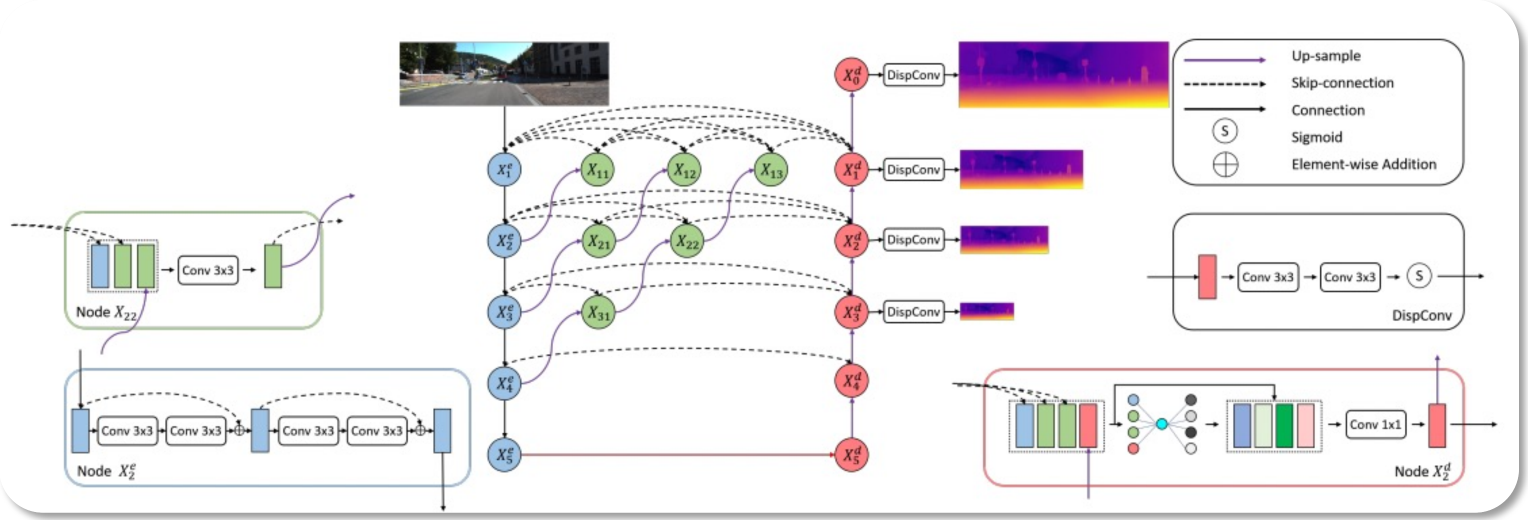
(a) Packing (b) Unpacking



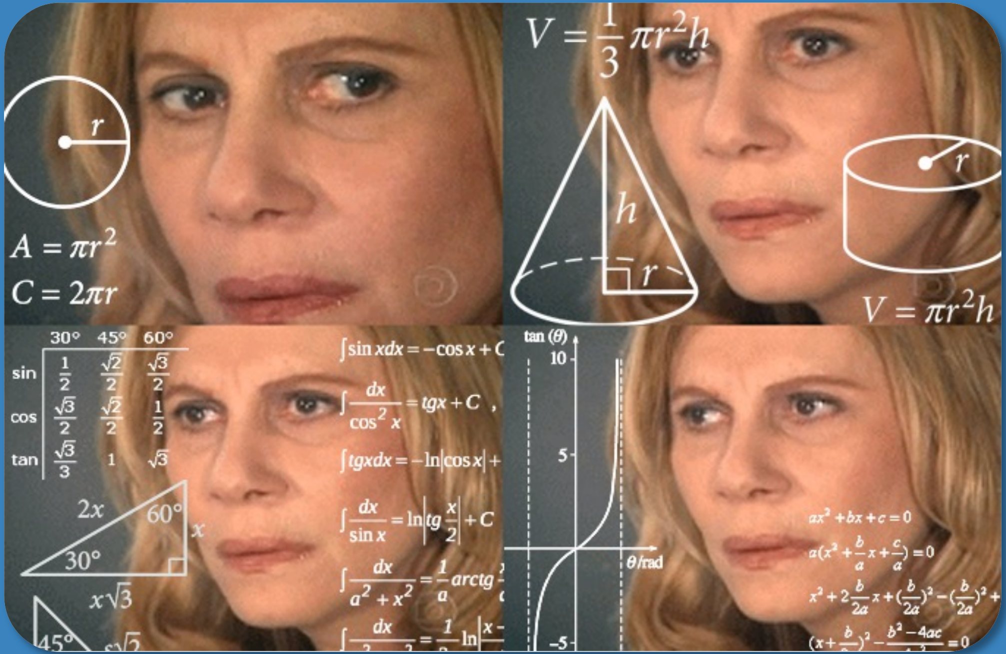
PackNet-SfM: 3D Packing for Self-Supervised Monocular Depth Estimation, Guizilini et al, CVPR20

– **HR-Depth:** MS + **Progressive** skip connection + **SqueezeExcite**

➢ **10x parameter reduction** w.r.t. PackNet + better performance!



HR-Depth: High Resolution Self-Supervised Monocular Depth Estimation, Lyu et al, CAI21



How to Evaluate?

Questions?

CVSSP Centre for Vision, Speech and Signal Processing