An ODE to MonODEpth

Vitor Campagnolo Guizilini

Toyota Research Institute

In realms where pixels dance with light's embrace,
There lies a quest, profound, in cyberspace.
Monocular depth, thou art the key,
To unlock realms unseen, for all to see.

So here's to thee, in ode we sing,
To monocular depth, eternal spring.
In algorithms' dance, forever we'll trace,
The wonders of depth, in digital space.

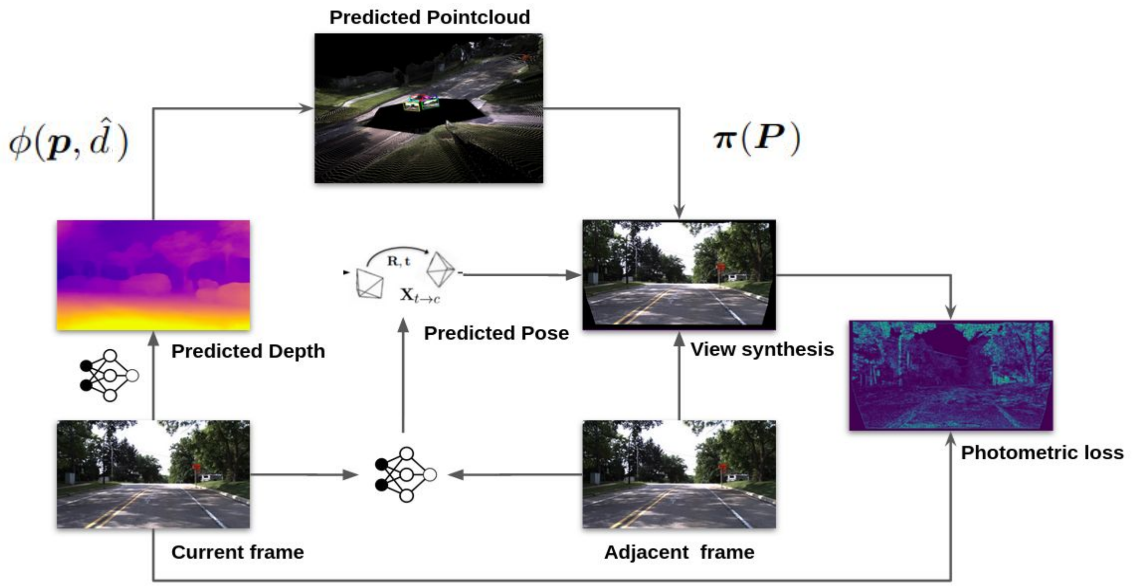
- *ChatGPT*

PackNet

3D Packing for Self-Supervised Monocular Depth Estimation

V Guizilini, R Ambrus, S Pillai, A Raventos, A Gaidon ***(CVPR'20)***

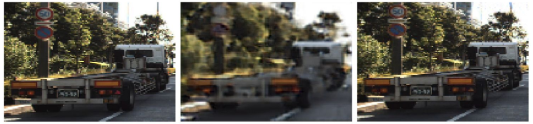
Self-supervised depth and ego-motion estimation



PackNet

3D Packing for Self-Supervised Monocular Depth Estimation

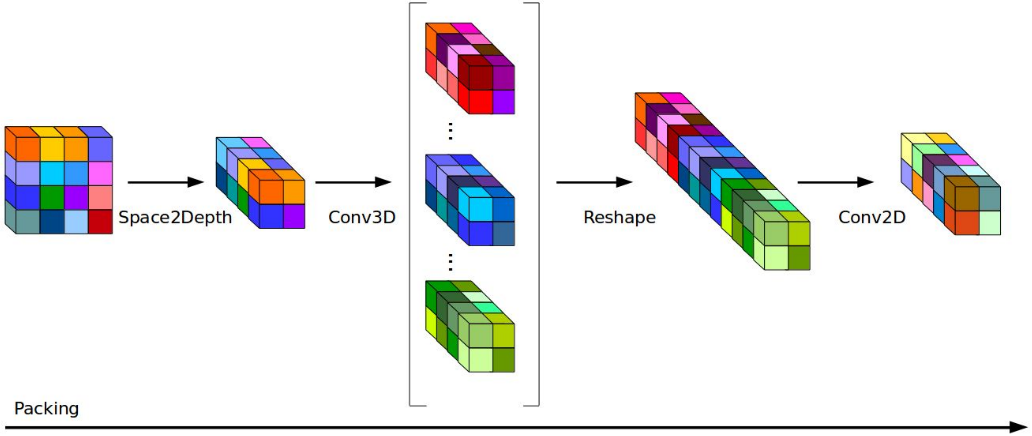
V Guizilini, R Ambrus, S Pillai, A Raventos, A Gaidon ***(CVPR'20)***

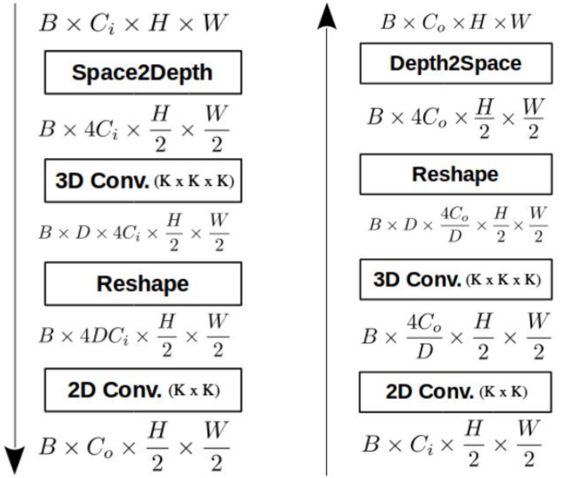


(a) Input Image (b) Max Pooling + Bilinear Upsample (c) Pack + Unpack

Packing and unpacking operations

Preserve spatial information during the encoding and decoding stages





(a) Packing (b) Unpacking

PackNet

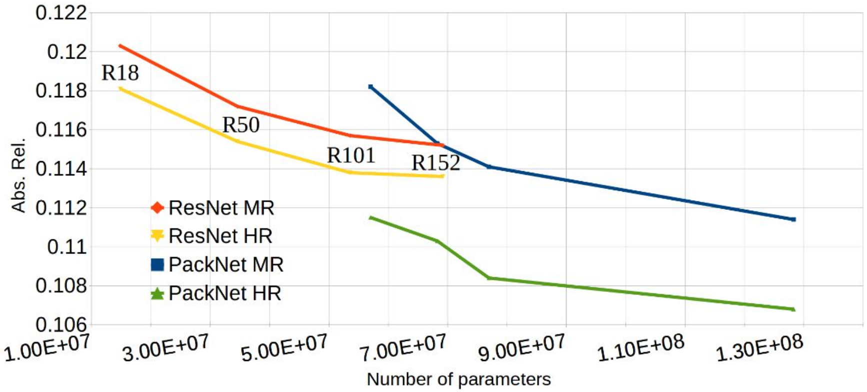
3D Packing for Self-Supervised Monocular Depth Estimation

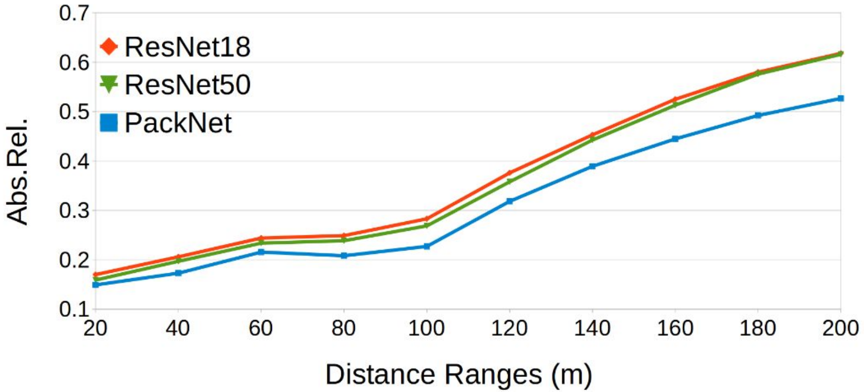
V Guizilini, R Ambrus, S Pillai, A Raventos, A Gaidon ***(CVPR'20)***

Better scalability at:

Larger network sizes (128M parameters 🤯)

Longer depth ranges



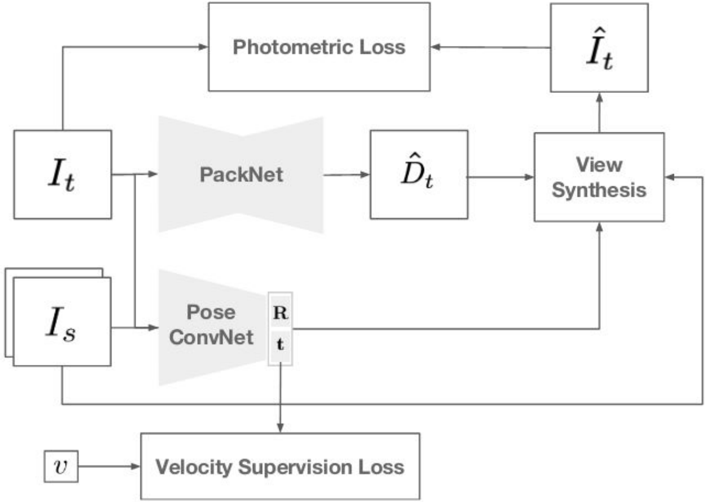


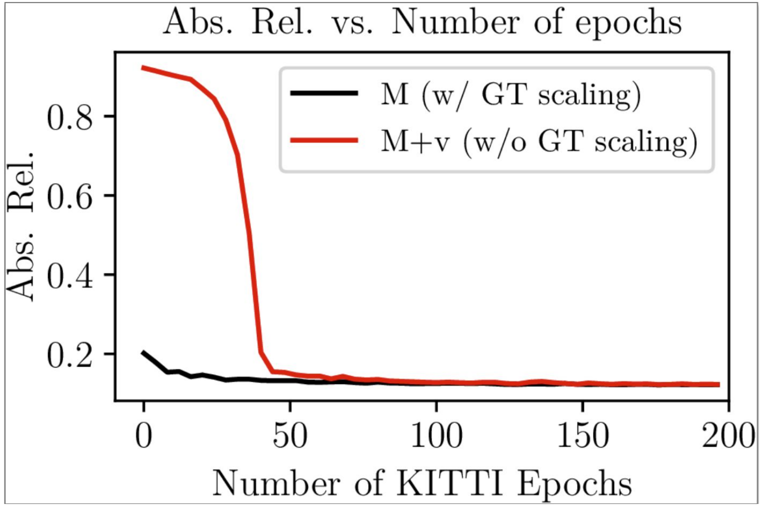
Metric Velocity Supervision

3D Packing for Self-Supervised Monocular Depth Estimation

V Guizilini, R Ambrus, S Pillai, A Raventos, A Gaidon ***(CVPR'20)***

Scale-aware depth estimates by supervising on translation speed





Dense Depth for Automated Driving (DDAD)

3D Packing for Self-Supervised Monocular Depth Estimation

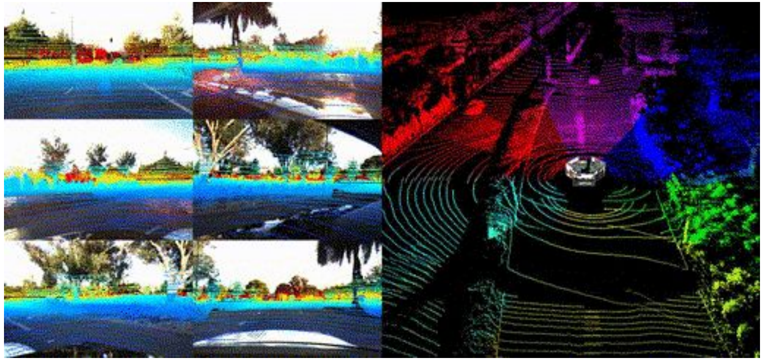
V Guizilini, R Ambrus, S Pillai, A Raventos, A Gaidon ***(CVPR'20)***

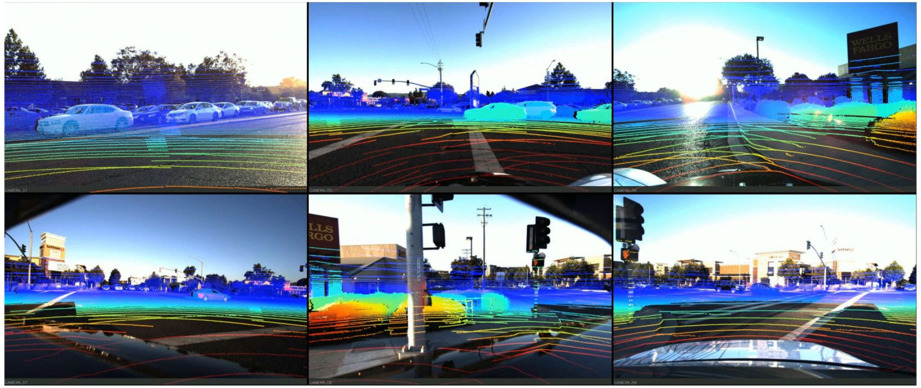
Depth estimation driving benchmark

6 cameras with **360º coverage** and **high-density** ground-truth up to **250m**

Training: 150 scenes -> 12650 samples x 6 cameras = 75900 frames

Validation: 50 scenes -> 3950 samples x 6 cameras = 23700 frames



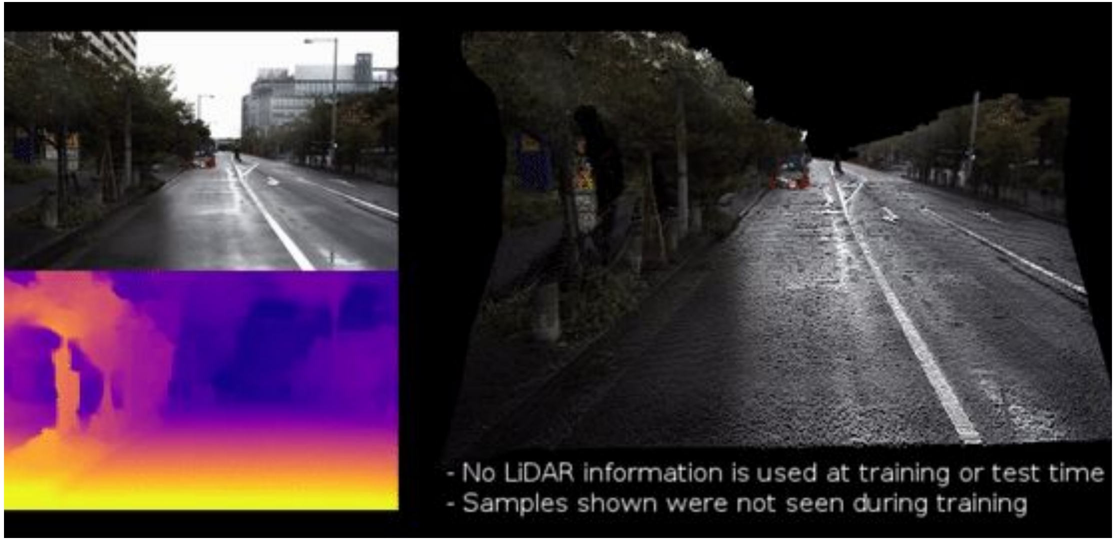


Dense Depth for Automated Driving (DDAD)

3D Packing for Self-Supervised Monocular Depth Estimation

V Guizilini, R Ambrus, S Pillai, A Raventos, A Gaidon ***(CVPR'20)***

PackNet results on DDAD (self-supervised)



Semantic Guidance

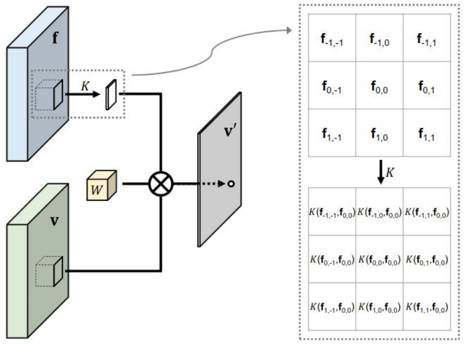
Semantically-Guided Representation Learning for Self-Supervised Monocular Depth

V Guizilini, R Hou, J Li, R Ambrus, A Gaidon ***(ICLR'20)***

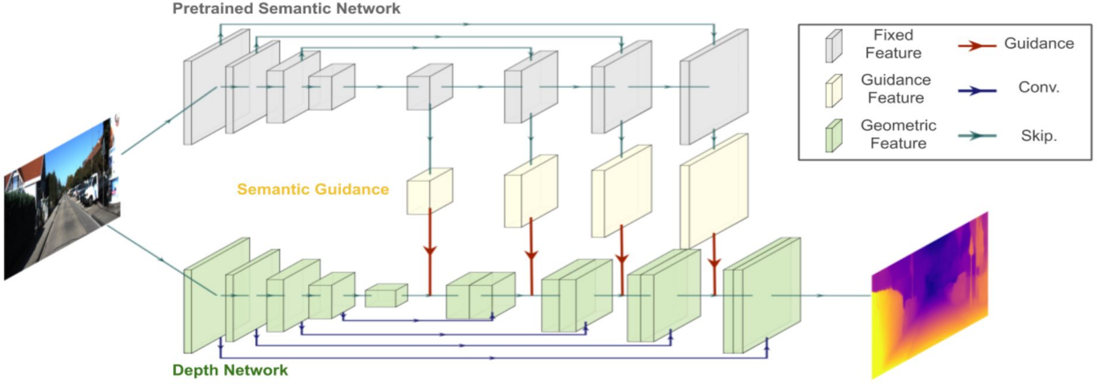
Pixel-Adaptive Convolutions*

Semantic segmentation is injected into the depth network

Source of object boundaries and scale priors



(Su et al., CVPR'19)



**Pixel-Adaptive Convolutional Neural Networks. Su et al., CVPR 2019.*

The Infinite Depth Problem

Semantically-Guided Representation Learning for Self-Supervised Monocular Depth

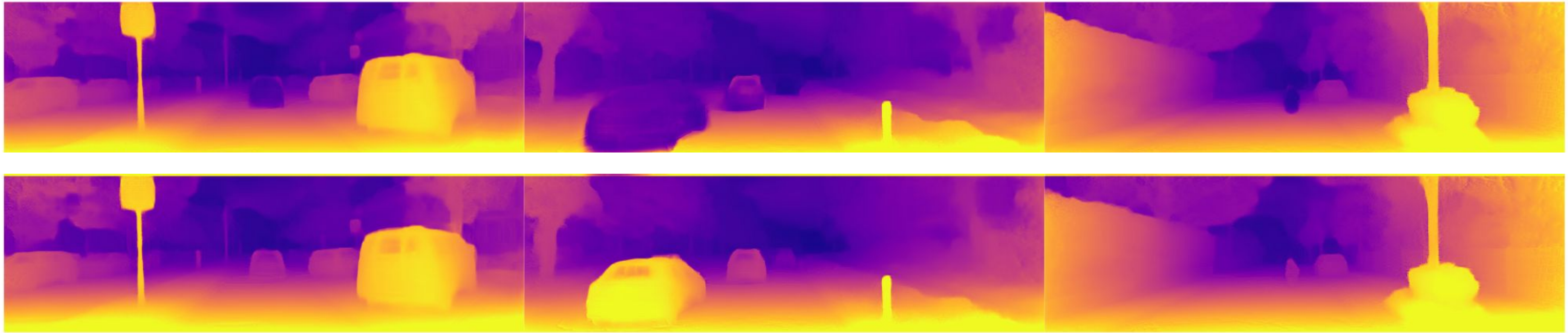
V Guizilini, R Hou, J Li, R Ambrus, A Gaidon ***(ICLR'20)***

Two-Stage Training: infinite depth as a dataset bias problem

1) Model is trained using all the data

    Ground-plane assumption: no predictions below (dominant) ground plane

2) Train a second model on filtered dataset



Sparse Semi-Supervision

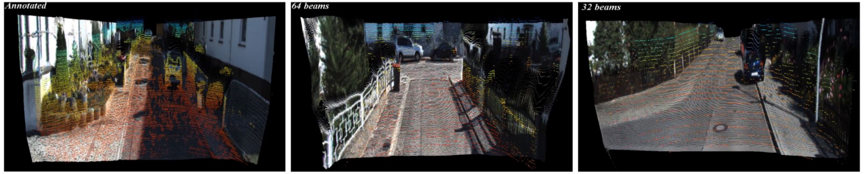
CVPR JUNE 17-21, 2024 SEATTLE, WA

Robust Semi-Supervised Monocular Depth Estimation With Reprojected Distances

V Guizilini, J Li, R Ambrus, S Pillai, A Gaidon ***(CoRL'19)***

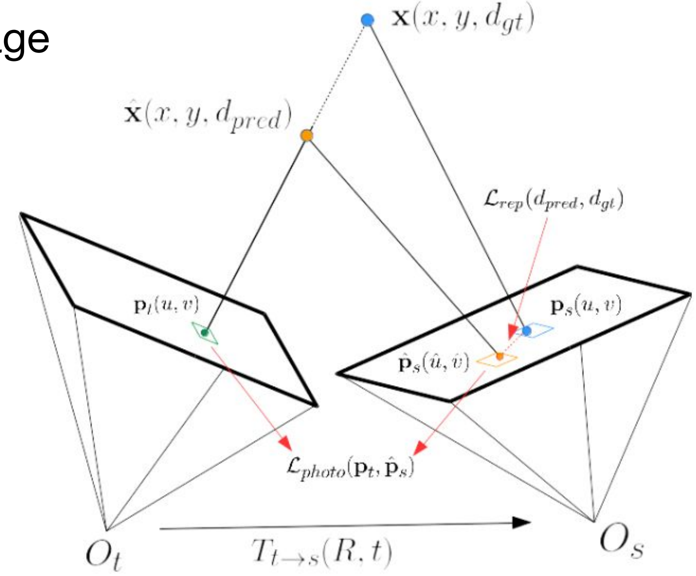
Self-Supervision + Sparse Supervision

Target supervised error reprojected to context image



(a) *Annotated* depth maps (b) 64 beams (c) 32 beams

(d) 16 beams (e) 8 beams (f) 4 beams



TOYOTA RESEARCH INSTITUTE

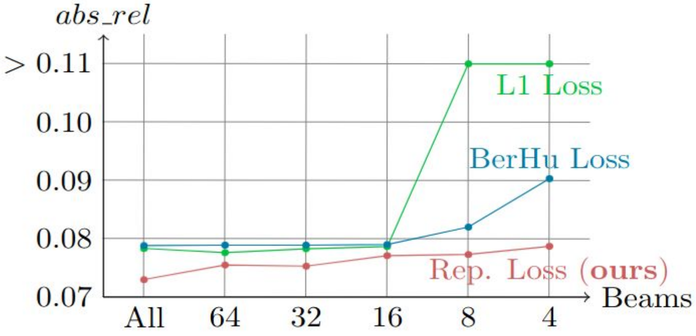
Sparse Semi-Supervision

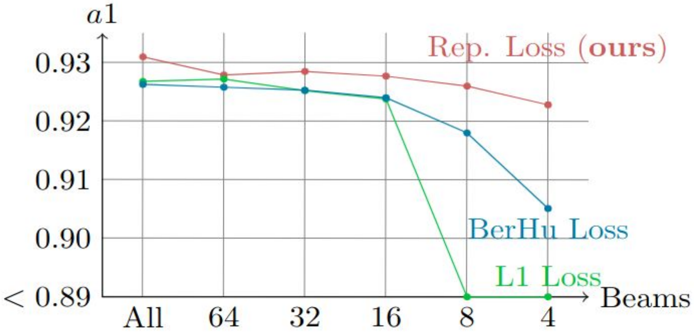
Robust Semi-Supervised Monocular Depth Estimation With Reprojected Distances

V Guizilini, J Li, R Ambrus, S Pillai, A Gaidon ***(CoRL'19)***

(a) *Annotated* depths (18288 points) (b) 64 beams (1427 points) (c) 32 beams (711 points)

(d) 16 beams (347 points) (e) 8 beams (171 points) (f) 4 beams (77 points)





Depth Completion

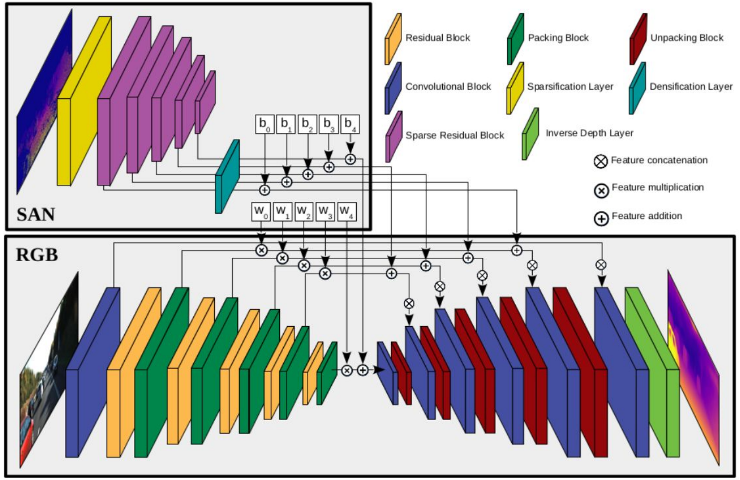
Sparse Auxiliary Networks for Unified Monocular Depth Prediction and Completion

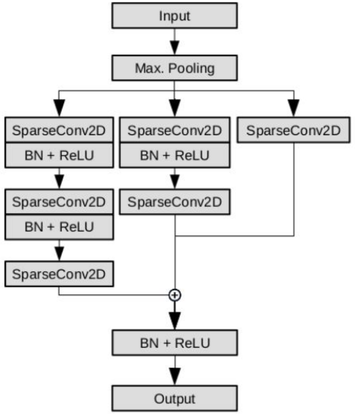
V Guizilini, R Ambrus, W Burgard, A Gaidon ***(CVPR'21)***

Dialable Perception

Depth **prediction** and **completion** with the same model

Depth features injected into RGB features





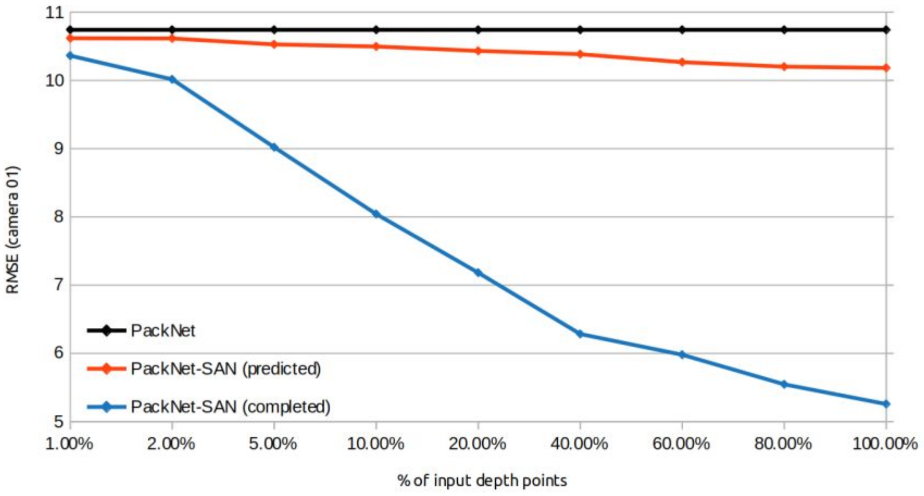
Depth Completion

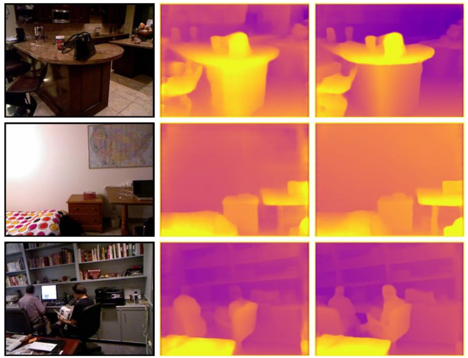
Sparse Auxiliary Networks for Unified Monocular Depth Prediction and Completion

V Guizilini, R Ambrus, W Burgard, A Gaidon ***(CVPR'21)***

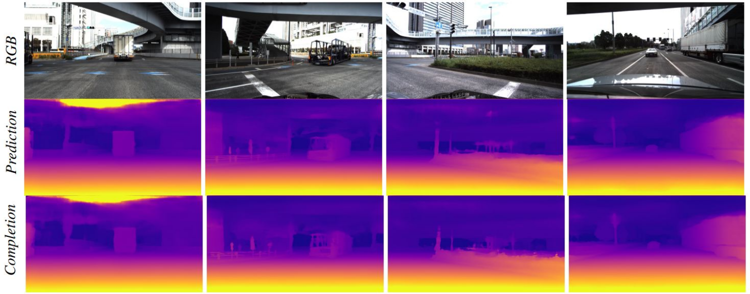
Experiments with varying amounts of depth density

Prediction results improve when jointly trained





(a) Input (b) Predicted (c) Completed



Pre-Trained Features

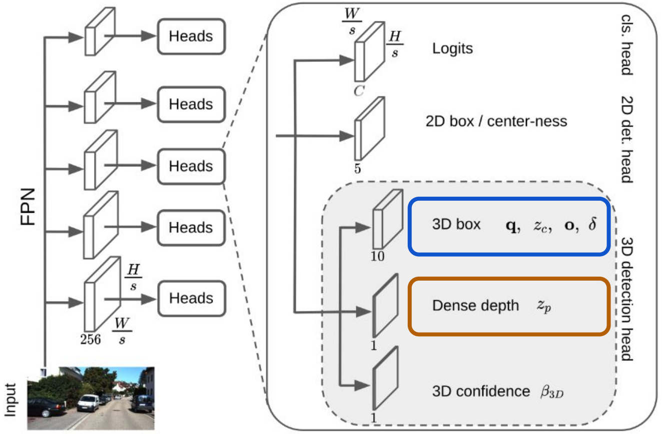
Is Pseudo-Lidar Needed for Monocular 3D Object Detection?

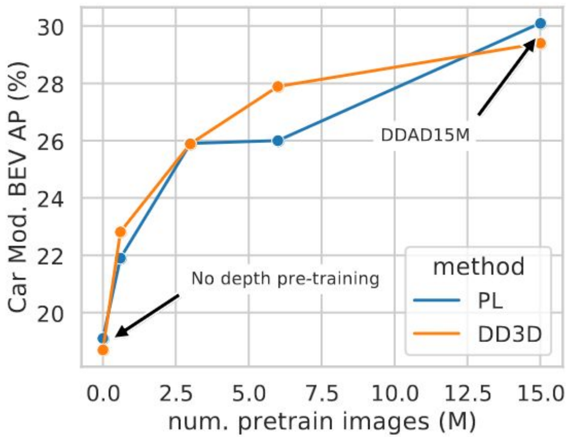
D Park, R Ambrus, V Guizilini, J Li, A Gaidon ***(ICCV'21)***

Depth estimation as a pre-training task for 3D detection

Maximize sharing of weights

Consistent improvements with more data





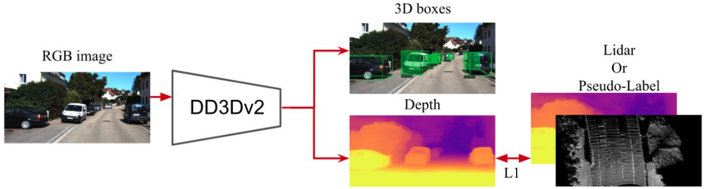
Pre-Trained Features

Depth Is All You Need for Monocular 3D Detection

D Park, J Li, D Chen, V Guizilini, A Gaidon ***(ICRA'23)***

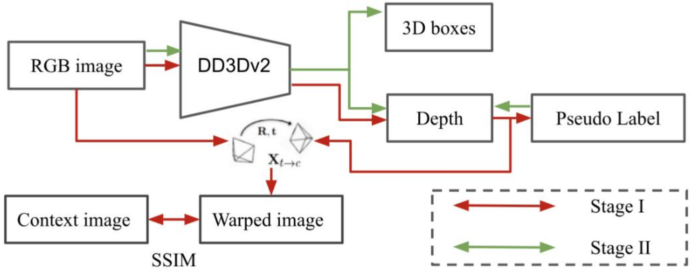
Augment depth pre-training with self-supervision

Pseudo-labeled supervision works better



Methods	Depth Sup.	Car BEV AP Easy	Car BEV AP Med	Car BEV AP Hard	Car 3D AP Easy	Car 3D AP Med	Car 3D AP Hard
SMOKE [27]	-	20.83	14.49	12.75	14.03	9.76	7.84
MonoPair [48]	-	19.28	14.83	12.89	13.04	9.99	8.65
AM3D [26]	LiDAR	25.03	17.32	14.91	16.50	10.74	9.52
PatchNet† [12]	LiDAR	22.97	16.86	14.97	15.68	11.12	10.17
RefinedMPL [49]		28.08	17.60	13.95	18.09	11.14	8.96
D4LCN [50]	LiDAR	22.51	16.02	12.55	16.65	11.72	9.51
Kinematic3D [51]	Video	26.99	17.52	13.10	19.07	12.72	9.17
Demystifying [5]	LiDAR	-	-	-	23.66	13.25	11.23
CaDDN [30]	LiDAR	27.94	18.91	17.19	19.17	13.41	11.46
MonoEF [52]	Video	29.03	19.70	17.26	21.29	13.87	11.71
MonoFlex [53]	-	28.23	19.75	16.89	19.94	13.89	12.07
GUPNet [54]	-	-	-	-	20.11	14.20	11.77
PGD [42]	-	30.56	23.67	20.84	24.35	18.34	16.90
DD3D [1]	-	30.98	22.56	20.03	23.22	16.34	14.20
Ours	LiDAR	35.70	24.67	21.73	26.36	17.61	15.32

NuScenes test set



Methods	Depth Sup.	Backbone	AP[%]↑	ATE[m]↓	ASE[1-IoU]↓	AOE[rad]↓	NDS↑
MonoDIS [40]	-	R34	30.4	0.74	0.26	0.55	0.38
FCOS3D [3]	-	R101	35.8	0.69	0.25	0.45	0.43
PGD[42]	-	R101	37.0	0.66	0.25	0.49	0.43
DD3D [1]	-	V2-99	41.8	0.57	0.25	0.37	0.48
DETR3D [43]	-	V2-99	41.2	0.64	0.26	0.39	0.48
DD3Dv2-selfsup	Video	V2-99	43.1	0.57	0.25	0.38	0.48
DD3Dv2	LiDAR	V2-99	**46.1**	**0.52**	**0.24**	**0.36**	**0.51**

KITTI test set

TOYOTA RESEARCH INSTITUTE

Unsupervised Domain Adaptation

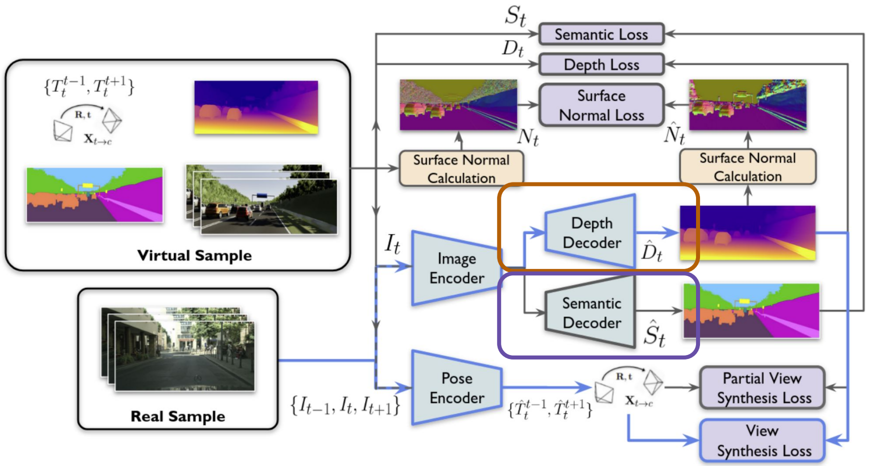
Geometric unsupervised domain adaptation for semantic segmentation

V Guizilini, J Li, R Ambruș, A Gaidon ***(ICCV'21)***

Unsupervised semantic segmentation via self-supervised depth estimation

Real-world self-supervision + synthetic supervision

Shared depth and semantic encoder



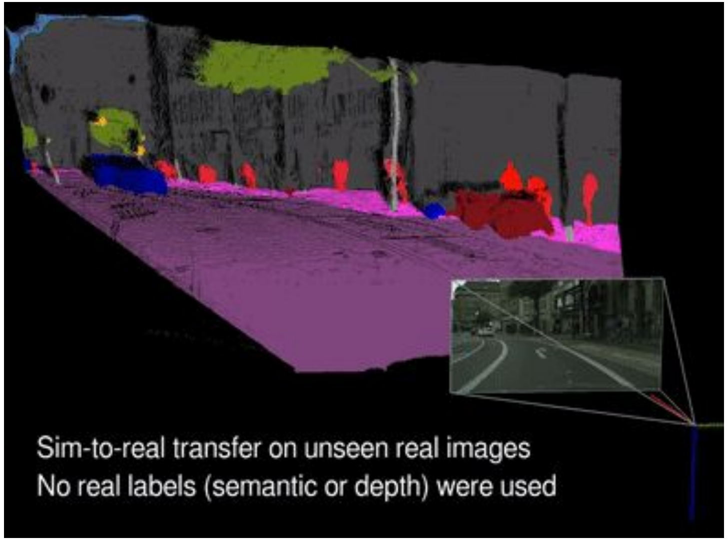
Unsupervised Domain Adaptation

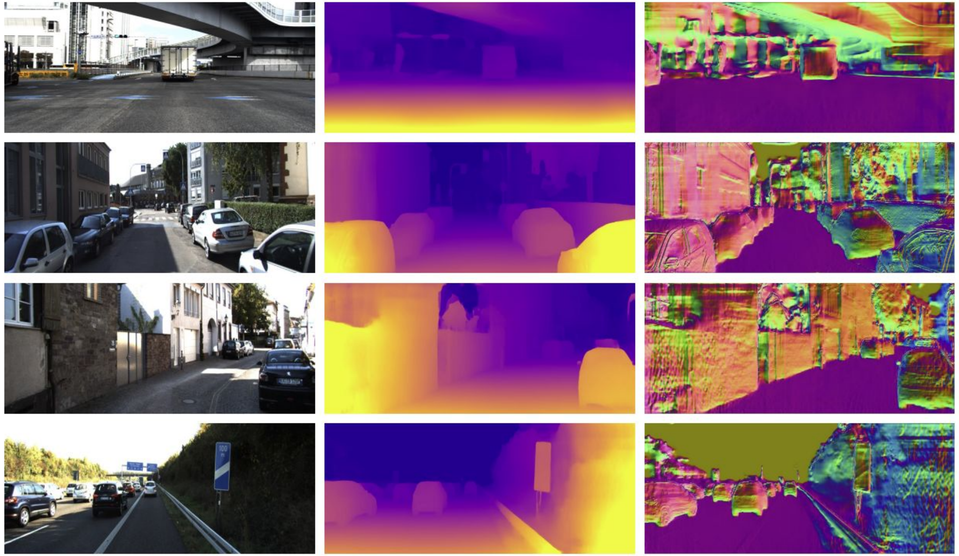
Geometric unsupervised domain adaptation for semantic segmentation

V Guizilini, J Li, R Ambruș, A Gaidon ***(ICCV'21)***

State of the art unsupervised domain adaptation **with no bells and whistles**

Improvements in depth estimation as well





Multi-Frame Depth Estimation

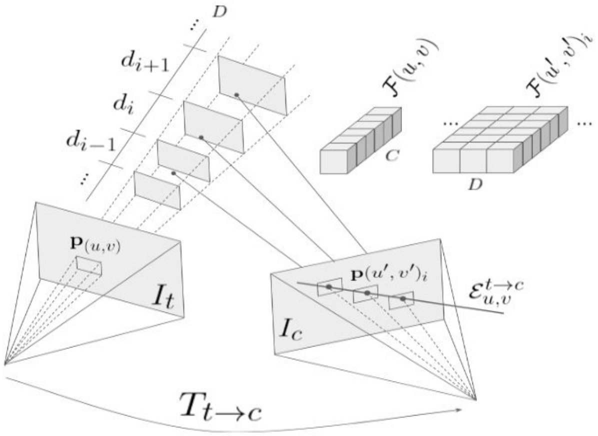
Multi-frame Self-Supervised Depth with Transformers

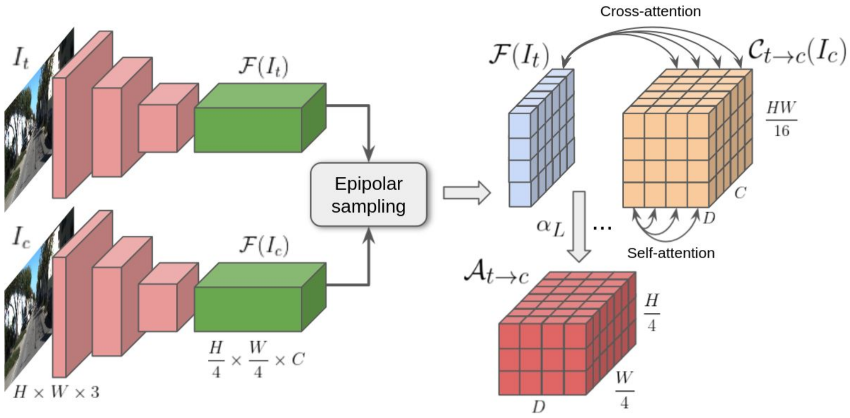
V Guizilini, R Ambruș, D Chen, S Zakharov, A Gaidon ***(CVPR'22)***

Feature matching module

Depth-discretized epipolar constraints (matching candidates)

Attention-based feature matching (self- and cross-attention between candidates)





Multi-Frame Depth Estimation

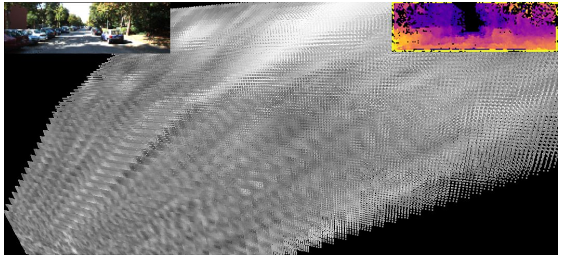
Multi-frame Self-Supervised Depth withTransformers

V Guizilini, R Ambruș, D Chen, S Zakharov, A Gaidon ***(CVPR'22)***

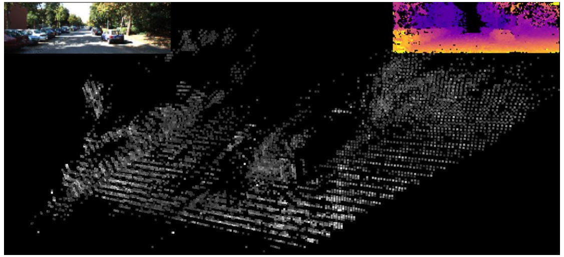
Sharper matching distributions

Better reasoning over photometric ambiguities

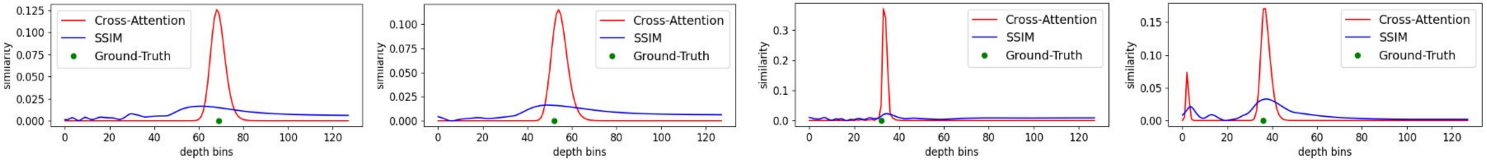
SSIM



DepthFormer



Per-pixel matching probability



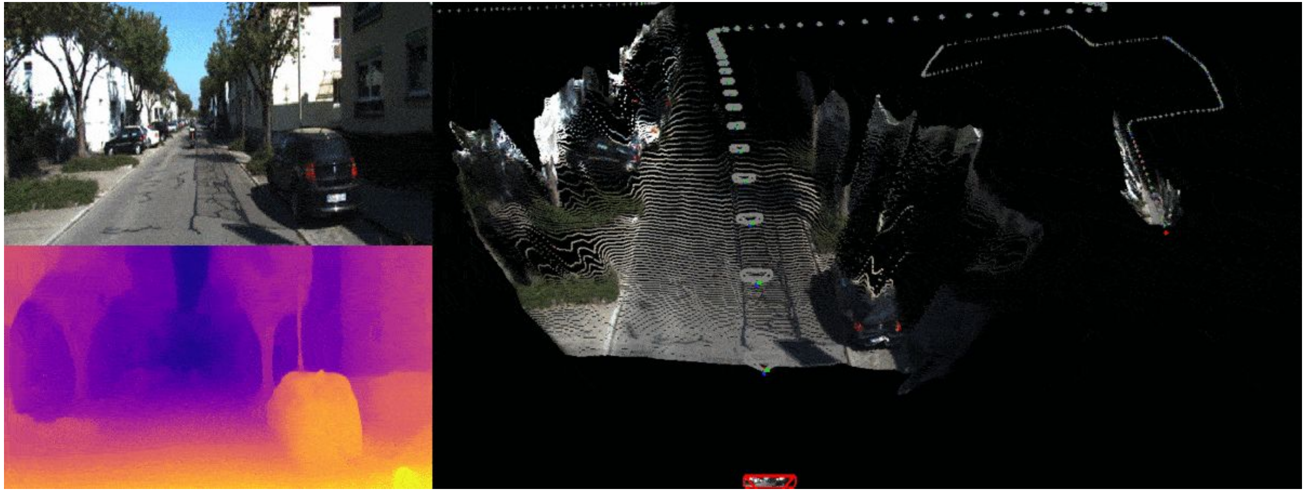
Multi-Frame Depth Estimation

Multi-frame self-supervised depth with transformers

V Guizilini, R Ambruș, D Chen, S Zakharov, A Gaidon ***(CVPR'22)***

Joint multi-frame depth and pose estimation

Better temporal consistency

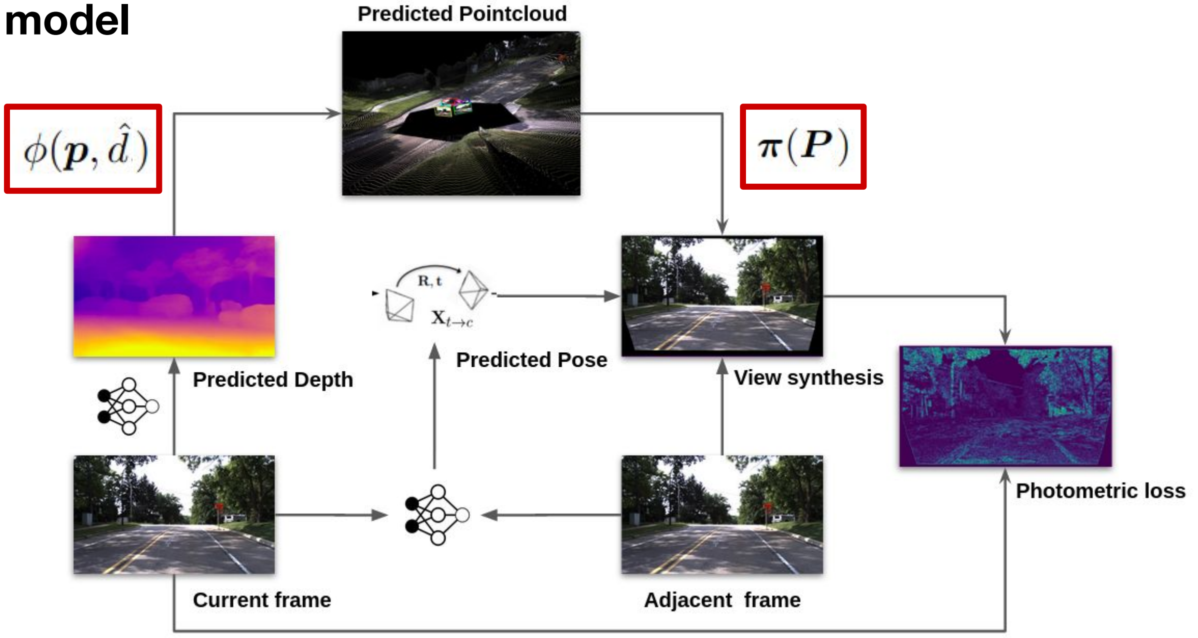


Neural Ray Surfaces

Neural Ray Surfaces for Self-Supervised Learning of Depth and Ego-Motion

I Vasiljevic, V Guizilini, R Ambrus, S Pillai, W Burgard, G Shakhnarovich, A Gaidon ***(3DV'20)***

Hidden label: camera model



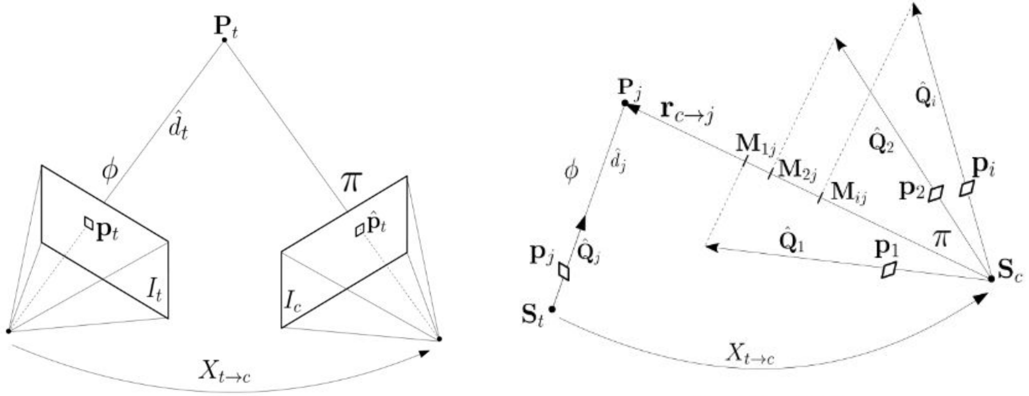
Neural Ray Surfaces

Neural Ray Surfaces for Self-Supervised Learning of Depth and Ego-Motion

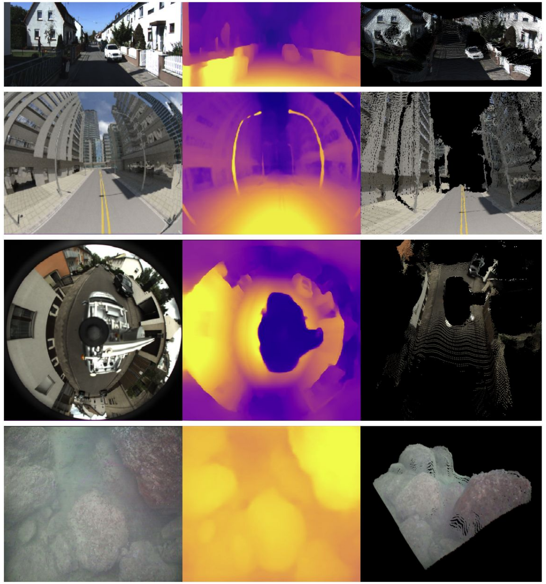
I Vasiljevic, V Guizilini, R Ambrus, S Pillai, W Burgard, G Shakhnarovich, A Gaidon ***(3DV'20)***

Dense ray surface network

- Closed form unprojection (ray x depth)
- Cosine similarity matching for projection



(a) Pinhole (b) Generic



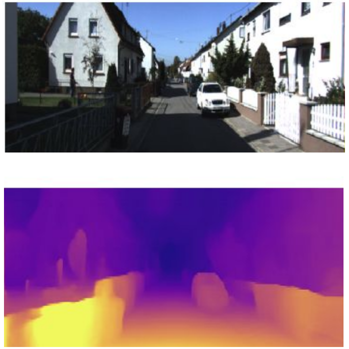
Neural Ray Surfaces

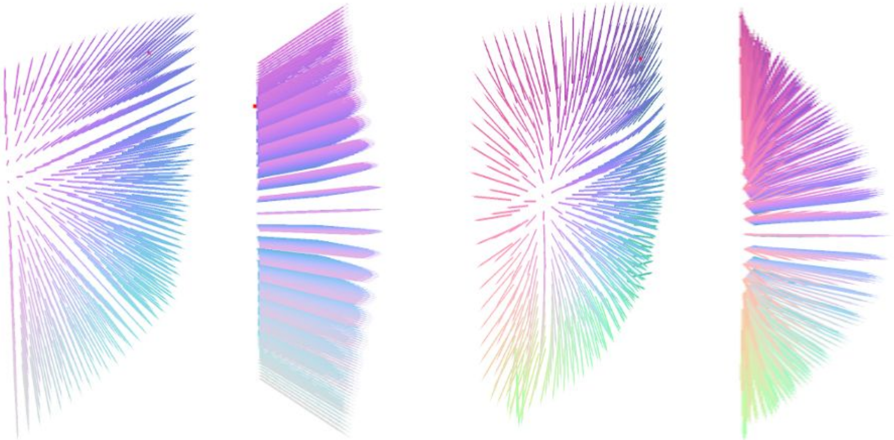
Neural Ray Surfaces for Self-Supervised Learning of Depth and Ego-Motion

I Vasiljevic, V Guizilini, R Ambrus, S Pillai, W Burgard, G Shakhnarovich, A Gaidon ***(3DV'20)***

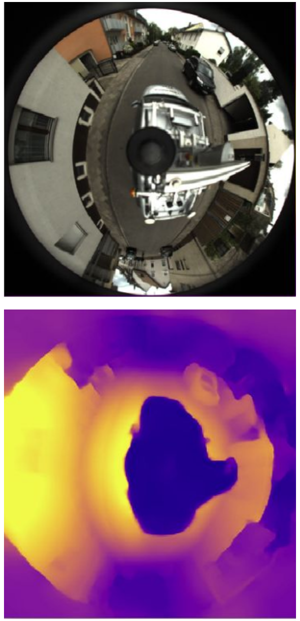
Self-supervised depth, ego-motion, and camera model

Adaptation to different geometries





(a) Pinhole (KITTI) (b) Catadioptric (OmniCam)

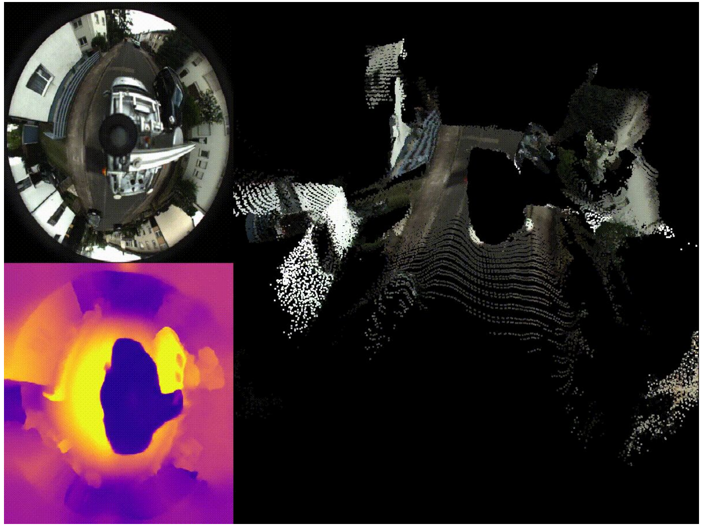


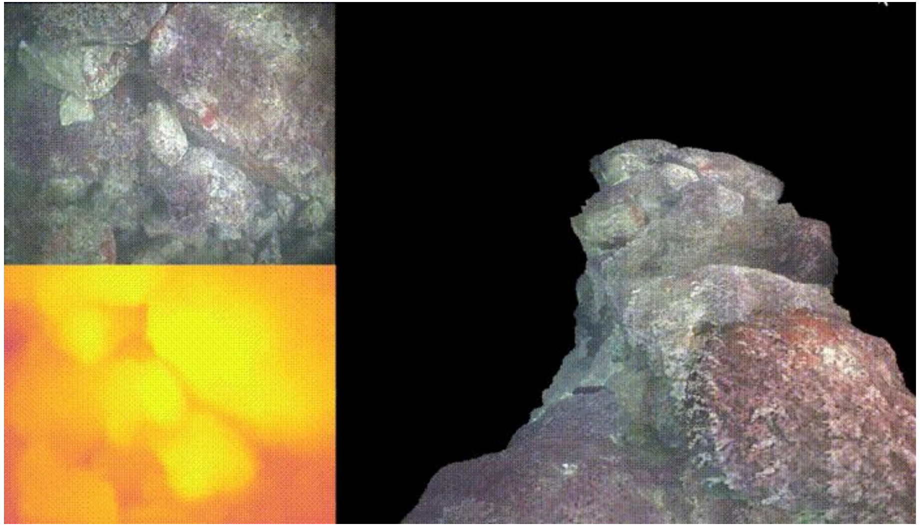
Neural Ray Surfaces

Neural Ray Surfaces for Self-Supervised Learning of Depth and Ego-Motion

I Vasiljevic, V Guizilini, R Ambrus, S Pillai, W Burgard, G Shakhnarovich, A Gaidon ***(3DV'20)***

It works even underwater!





Intrinsics Self-Calibration

Self-Supervised Camera Self-Calibration from Video

J Fang, I Vasiljevic, V Guizilini, R Ambrus, G Shakhnarovich, A Gaidon, MR Walter ***(ICRA'22)***

Unified Camera Model

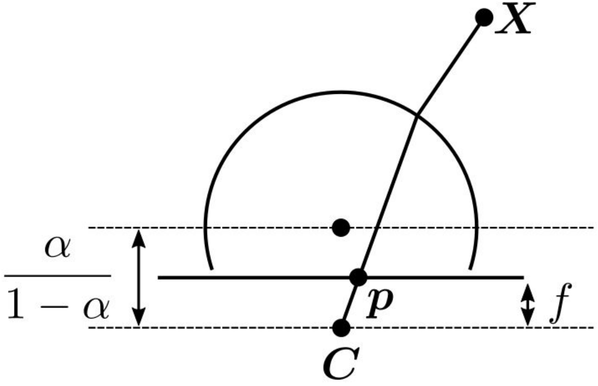
- Closed-form projection and unprojection operations
- Only one extra parameter over the pinhole model

$$\boldsymbol{\pi}(\boldsymbol{P}, \boldsymbol{i}) = \begin{bmatrix} f_x \frac{x}{\alpha d + (1-\alpha) z} \\ f_y \frac{y}{\alpha d + (1-\alpha) z} \end{bmatrix} + \begin{bmatrix} c_x \\ c_y \end{bmatrix}$$

$$\phi(\boldsymbol{p}, \hat{d}, \boldsymbol{i}) = \hat{d} \frac{\xi + \sqrt{1 + (1 - \xi^2) r^2}}{1 + r^2} \begin{bmatrix} m_x \\ m_y \\ 1 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ \hat{d}\zeta \end{bmatrix}$$

$$m_x = \frac{u - c_x}{f_x}(1 - \alpha) \qquad m_y = \frac{v - c_y}{f_y}(1 - \alpha)$$

$$r^2 = m_x^2 + m_x^2 \qquad \zeta = \frac{\alpha}{1 - \alpha}$$



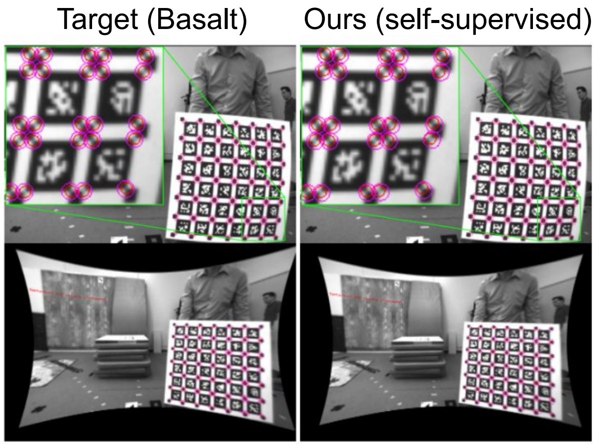
Intrinsics Self-Calibration

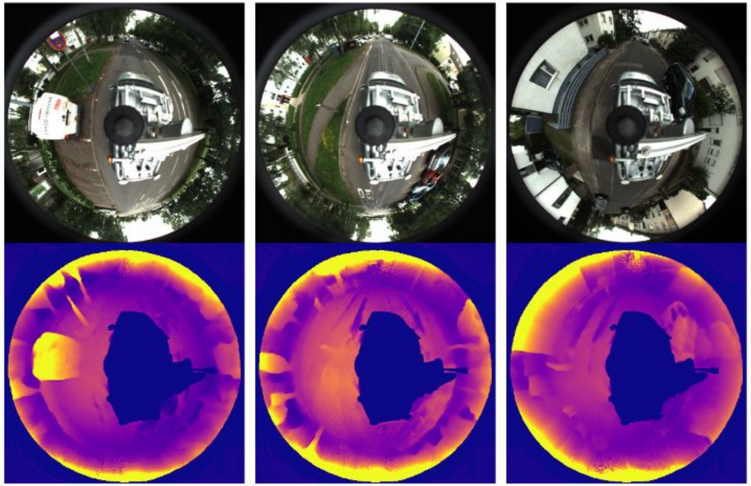
Self-Supervised Camera Self-Calibration from Video

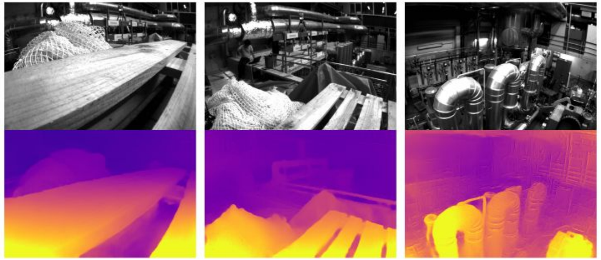
*J Fang, I Vasiljevic, V Guizilini, R Ambrus, G Shakhnarovich, A Gaidon, MR Walter **(ICRA'22)***

Sub-pixel calibration accuracy

Self-supervised depth **from any central camera**



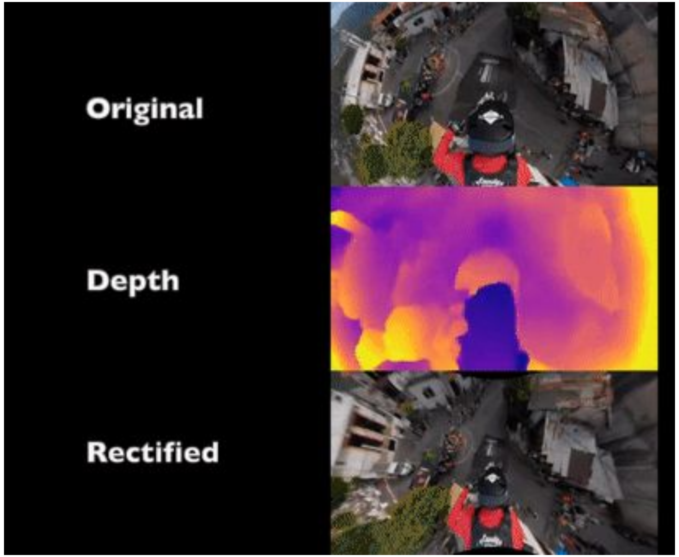


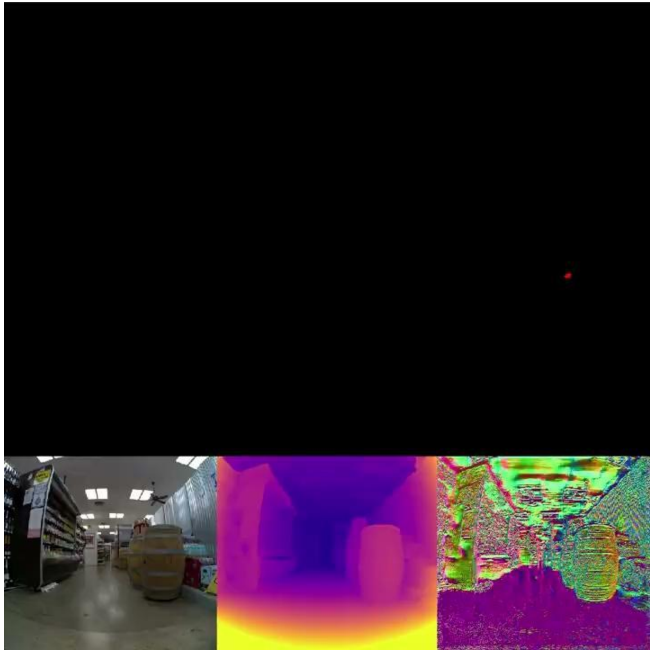


Intrinsics Self-Calibration

Self-Supervised Camera Self-Calibration from Video

J Fang, I Vasiljevic, V Guizilini, R Ambrus, G Shakhnarovich, A Gaidon, MR Walter ***(ICRA'22)***



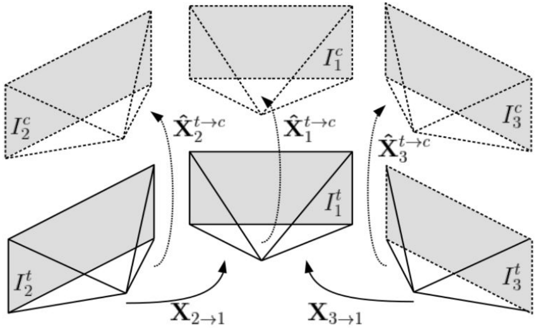


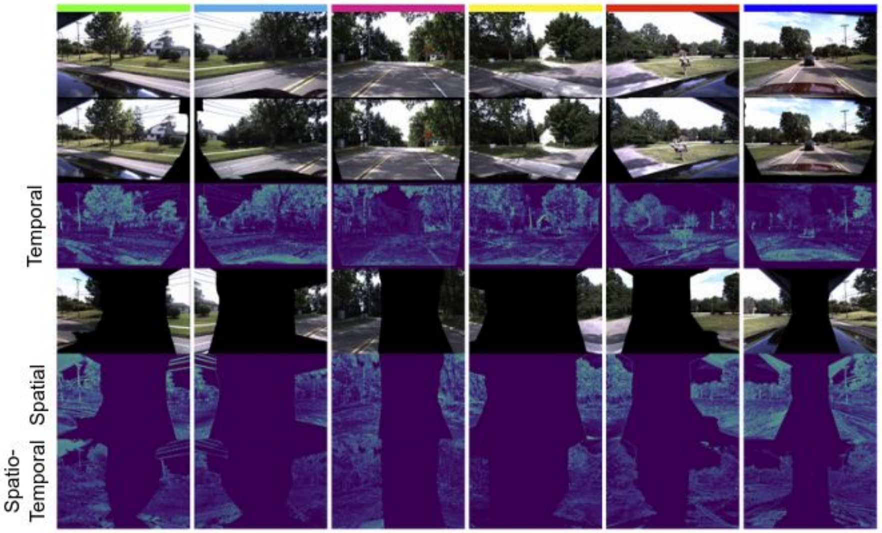
Full Surround Monodepth

Full Surround Monodepth from Multiple Cameras
V Guizilini, I Vasiljevic, R Ambrus, G Shakhnarovich, A Gaidon ***(ICRA'22)***

Spatio-Temporal photometric loss

- Same camera, different timesteps
- Different cameras, same timesteps
- Different cameras, different timesteps





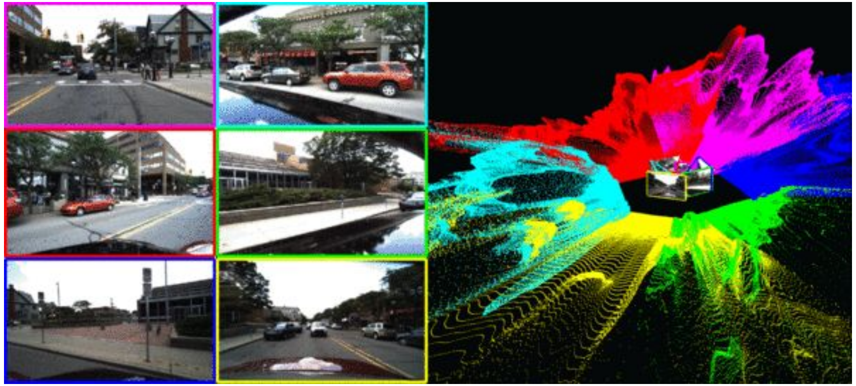
Full Surround Monodepth

Full Surround Monodepth from Multiple Cameras

V Guizilini, I Vasiljevic, R Ambrus, G Shakhnarovich, A Gaidon ***(ICRA'22)***

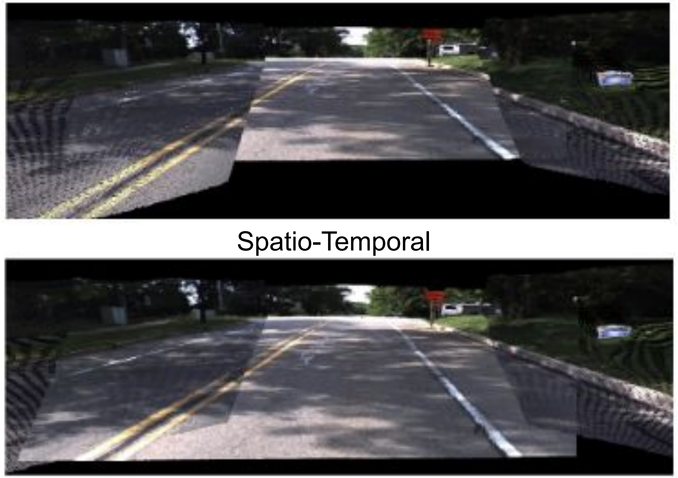
Scale-aware models

- Known extrinsics used to learn metric depth (and pose)
- Better cross-camera pointcloud consistency



Temporal

Spatio-Temporal



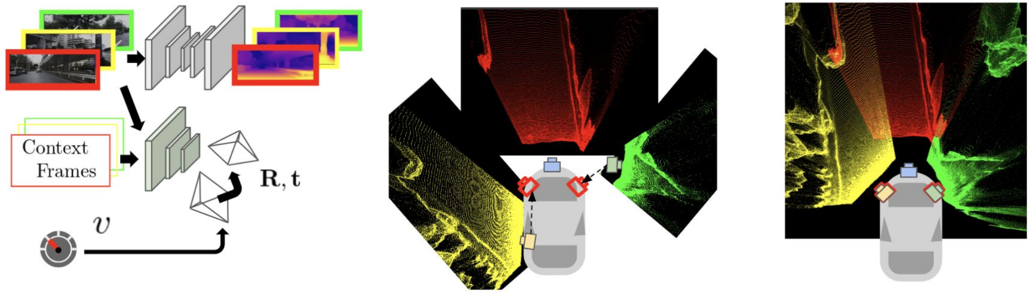
Extrinsics Self-Calibration

Robust Self-Supervised Extrinsic Self-Calibration

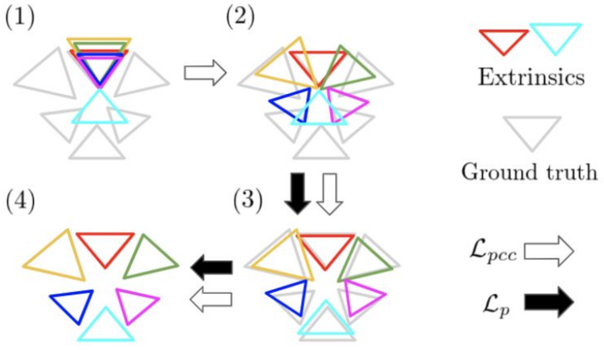
T Kanai, I Vasiljevic, V Guizilini, A Gaidon, R Ambrus ***(IROS'23)***

Joint depth, ego-motion, intrinsics, and extrinsics estimation

- Multi-stage curriculum learning
- Further improvements to depth estimation



(a) Self-supervised learning with velocity supervision (b) Extrinsic estimation (c) Self-calibration via joint optimization



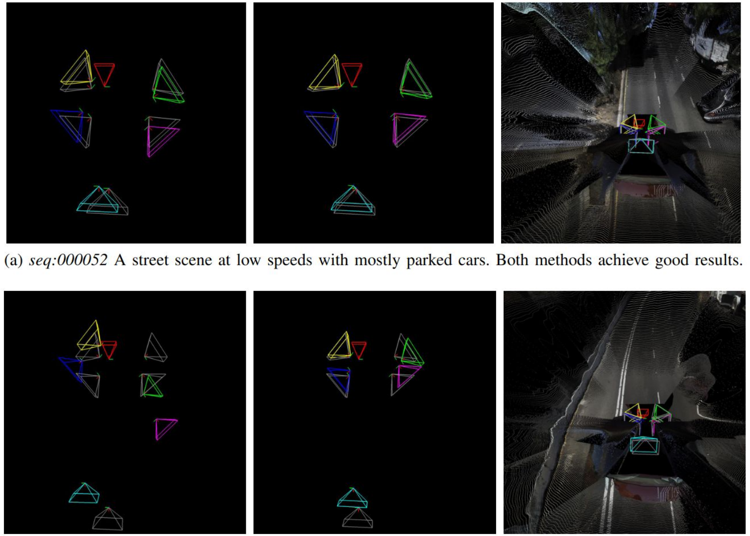
Stage	Optimization			Loss	
	depth	*ego-motion*	*extrinsics*	*Photo*	*Pose*
Monodepth Pretraining	✓	✓	-	✓	✓
Rotation Estimation	-	*Fix*	✓	-	✓
Extrinsic Estimation	*Fix*	✓	✓	✓	✓
End-to-end Training	✓	✓	✓	✓	✓

Extrinsics Self-Calibration

Robust Self-Supervised Extrinsic Self-Calibration

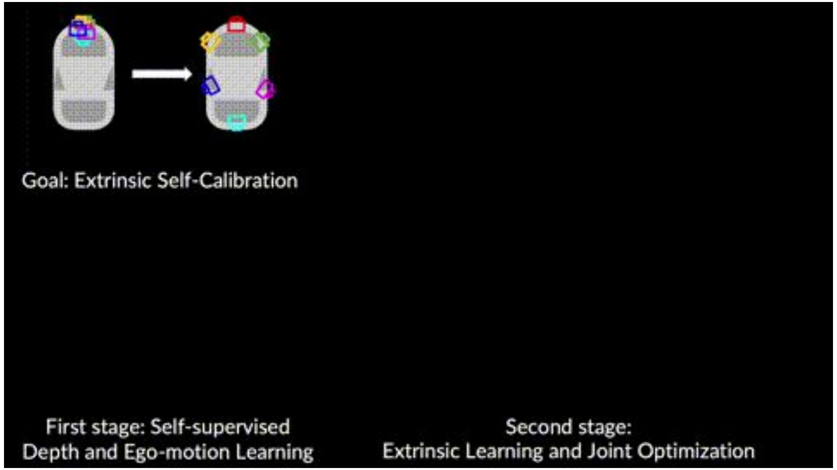
T Kanai, I Vasiljevic, V Guizilini, A Gaidon, R Ambrus ***(IROS'23)***

Improves over COLMAP for dynamic scenes



(a) *seq:000052* A street scene at low speeds with mostly parked cars. Both methods achieve good results.

(b) *seq:000016*: A highway scene at high speeds with many dynamic objects. COLMAP fails while SESC still achieves competitive results.



Geometry-Guided Visual Odometry

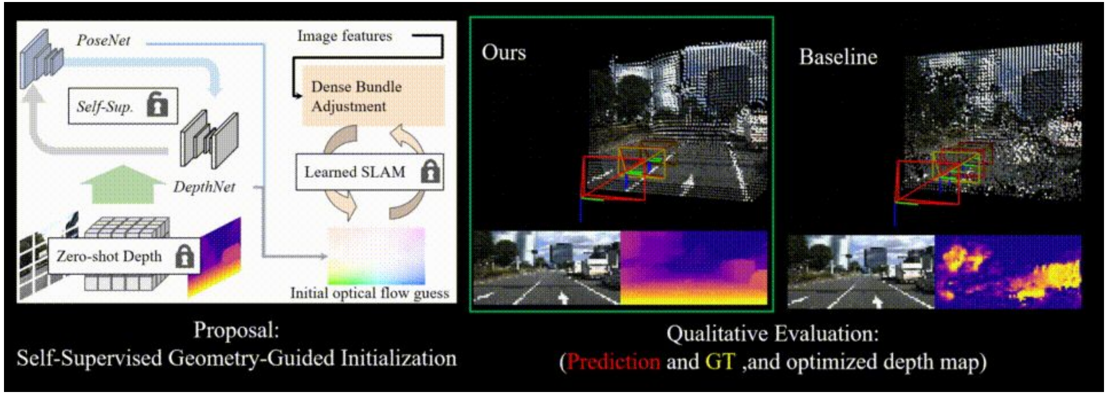
Self-Supervised Geometry-Guided Initialization for Robust Monocular Visual Odometry

T Kanai, I Vasiljevic, V Guizilini, K Shintani ***(arXiv, 2024)***

Self-supervised depth as initialization for bundle adjustment

Optical flow refinement based on depth and ego-motion estimation

Frozen zero-shot monocular depth network as additional source of priors



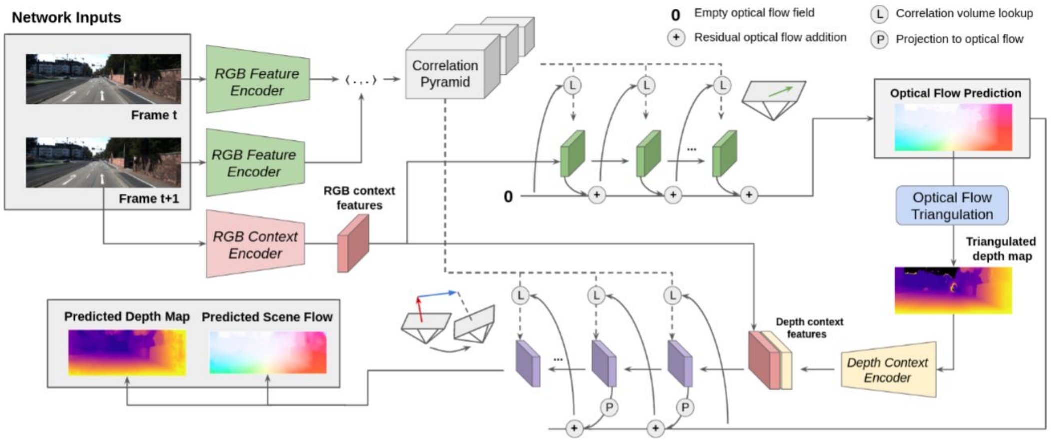
Self-Supervised Scene Flow

Learning Optical Flow, Depth, and Scene Flow Without Real-World Labels

V Guizilini, KH Lee, R Ambruş, A Gaidon ***(RA-L'22)***

Self-supervised depth and scene flow is an ill-posed problem

- Domain transfer via real-world self-supervision and synthetic supervision
- Joint multi-task optical flow initialization

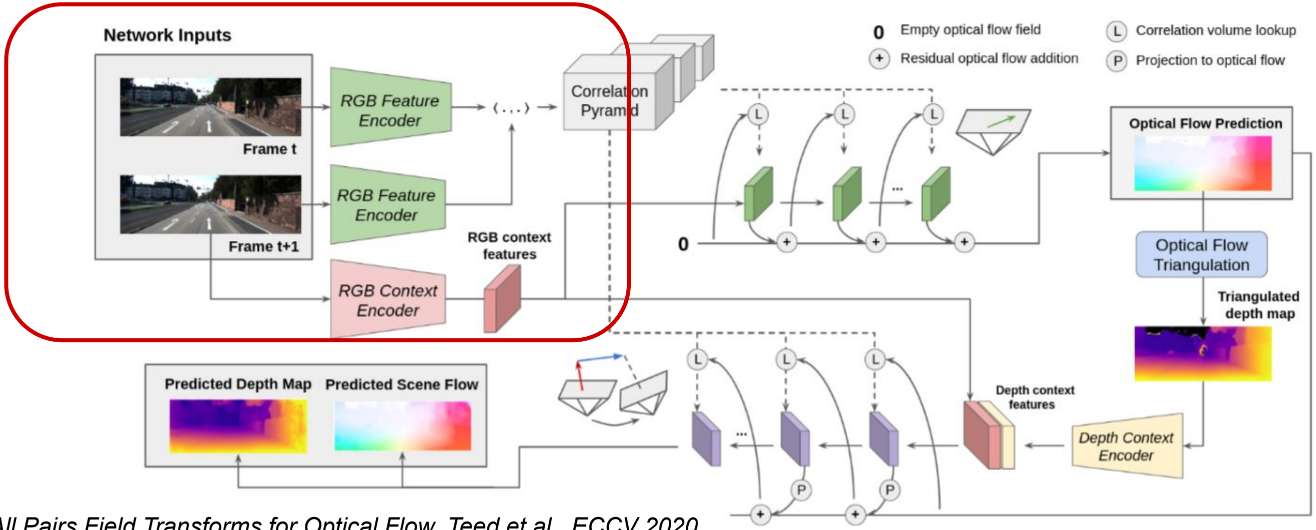


Self-Supervised Scene Flow

Learning Optical Flow, Depth, and Scene Flow Without Real-World Labels

V Guizilini, KH Lee, R Ambruş, A Gaidon ***(RA-L'22)***

Correlation pyramid* generated from target and context images



**RAFT: Recurrent All Pairs Field Transforms for Optical Flow. Teed et al., ECCV 2020.*

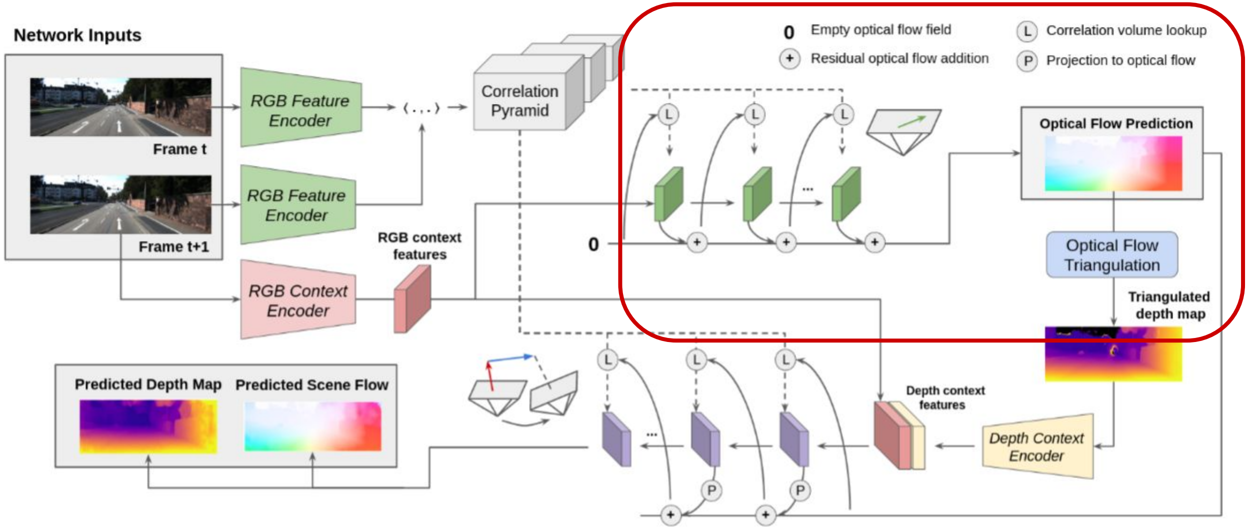
Self-Supervised Scene Flow

Learning Optical Flow, Depth, and Scene Flow Without Real-World Labels

V Guizilini, KH Lee, R Ambruş, A Gaidon ***(RA-L'22)***

Multi-stage residual optical flow estimation

Triangulation into depth maps



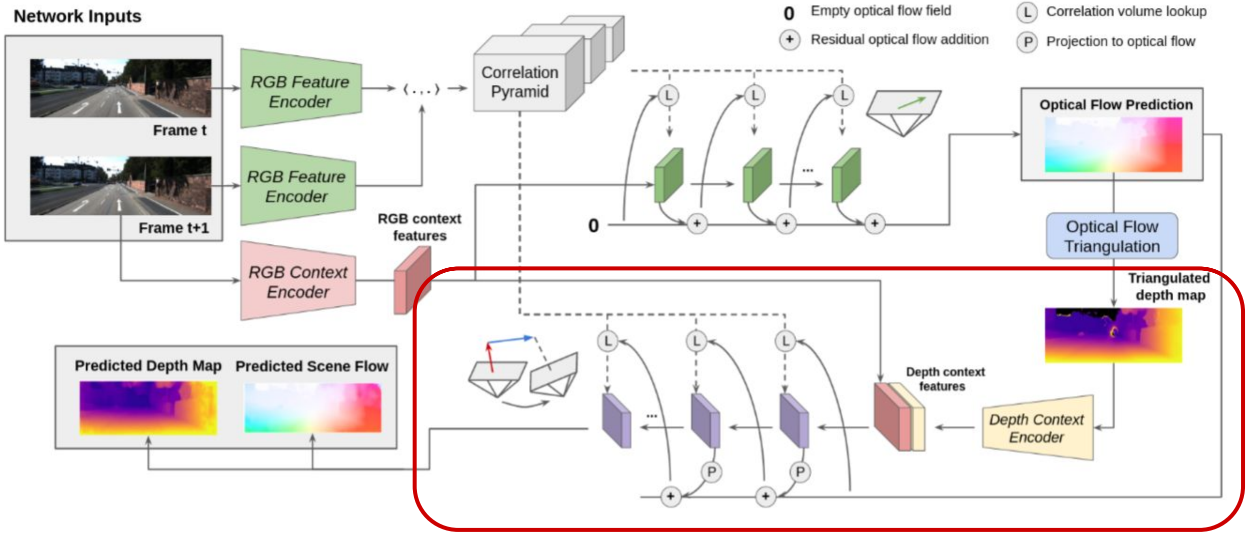
Self-Supervised Scene Flow

Learning Optical Flow, Depth, and Scene Flow Without Real-World Labels

V Guizilini, KH Lee, R Ambruş, A Gaidon ***(RA-L'22)***

Multi-stage depth and scene flow estimation

Triangulated depth features are used jointly with image features



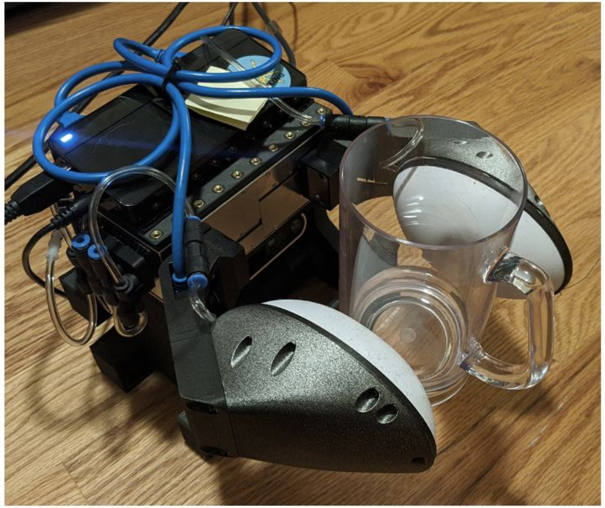
Tactile Sensors

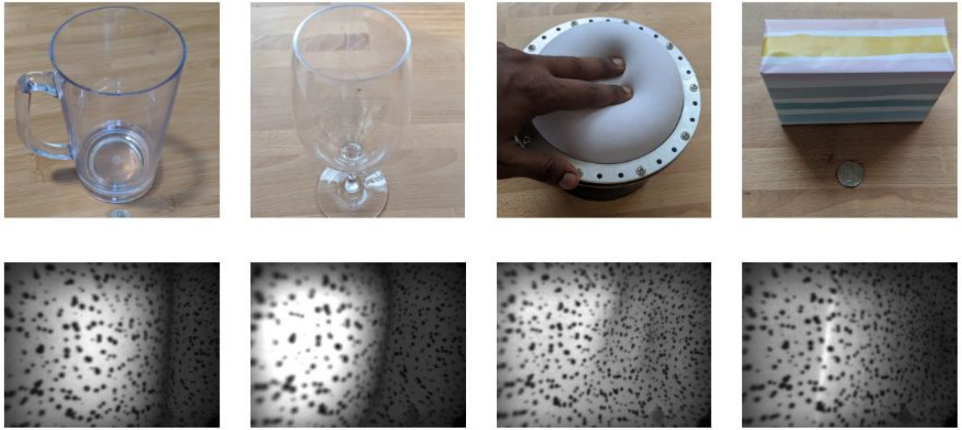
Monocular Depth Estimation for Soft Visuotactile Sensors

R Ambrus, V Guizilini, N Kuppuswamy, A Beaulieu, A Gaidon, A Alspach ***(RoboSoft'21)***

Depth estimation in a new domain: inside a bubble

Replace range sensors for object pose estimation (1-100mm ranges)





(a) Mug (b) Wine Glass (c) Fingers (d) Box

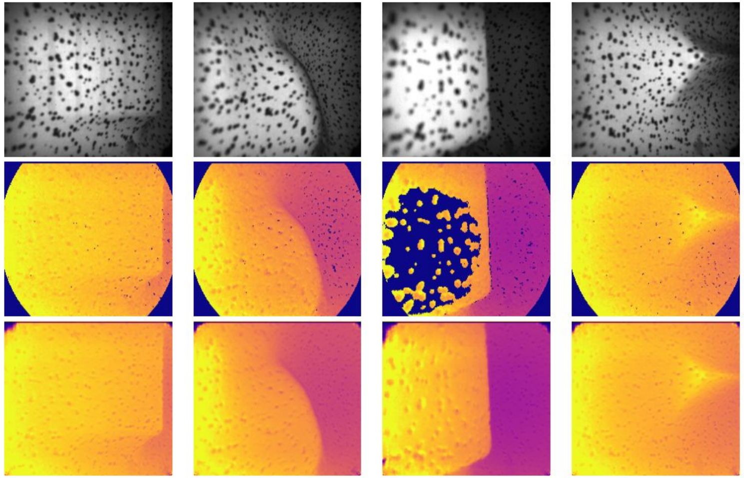
Tactile Sensors

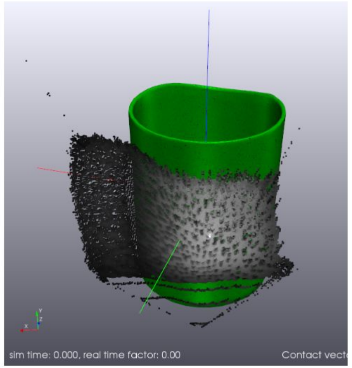
Monocular Depth Estimation for Soft Visuotactile Sensors

R Ambrus, V Guizilini, N Kuppuswamy, A Beaulieu, A Gaidon, A Alspach ***(RoboSoft’21)***

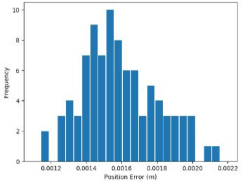
Depth estimation in a new domain: inside a bubble

Replace range sensors for object pose estimation

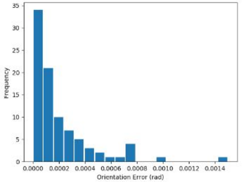




(a) Pose estimation on monocular depth maps



(b) Position norm error histogram



(c) Orientation error histogram

Depth Field Networks

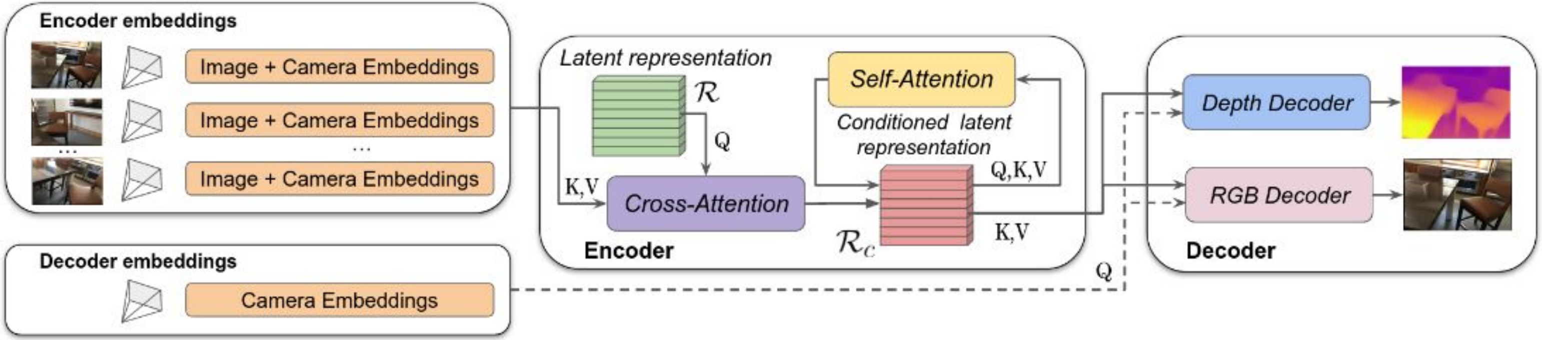
Depth Field Networks for Generalizable Multi-View Scene Representation

V Guizilini, I Vasiljevic, J Fang, R Ambrus, G Shakhnarovich, MR Walter, A Gaidon **(ECCV'22)**

Implicit learning of multi-view geometry

Condition a learned latent representation* using image and camera information

Decoding using only camera information



Perceiver IO: A General Architecture for Structured Inputs & Outputs. Jaegle et al., ICLR 2022.

Depth Field Networks

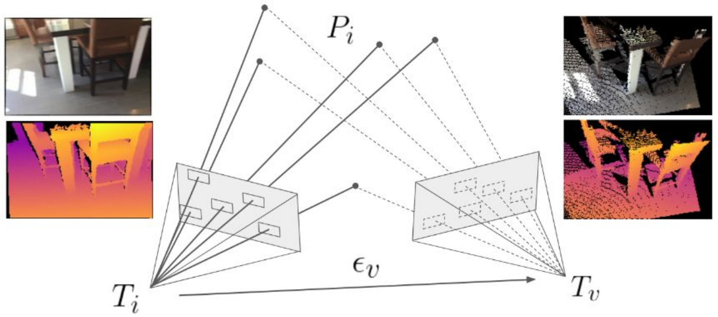
Depth Field Networks for Generalizable Multi-View Scene Representation

V Guizilini, I Vasiljevic, J Fang, R Ambrus, G Shakhnarovich, MR Walter, A Gaidon **(ECCV'22)**

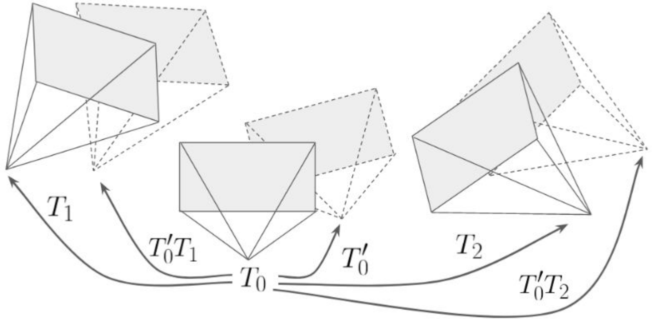
Geometry-preserving 3D augmentations

Increase scene diversity during training

Enforce equivariance in the learned latent representation



(a) Virtual Camera Projection.



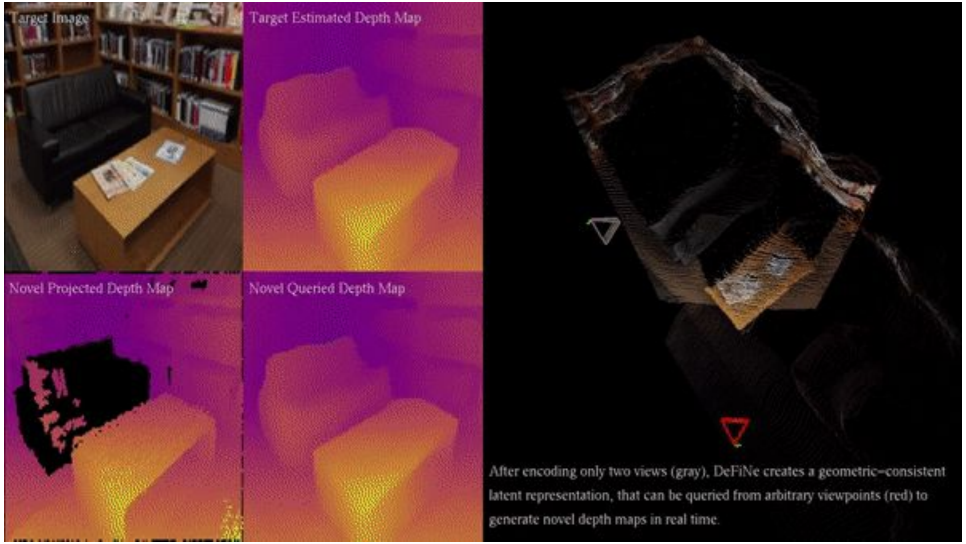
(b) Canonical Jittering.

Depth Field Networks

Depth Field Networks for Generalizable Multi-View Scene Representation

V Guizilini, I Vasiljevic, J Fang, R Ambrus, G Shakhnarovich, MR Walter, A Gaidon **(ECCV’22)**

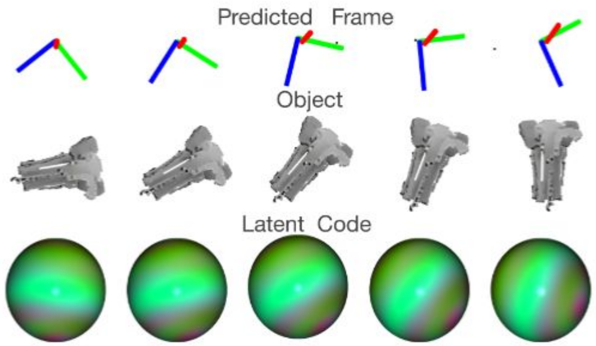
Novel depth synthesis by decoding from arbitrary viewpoints

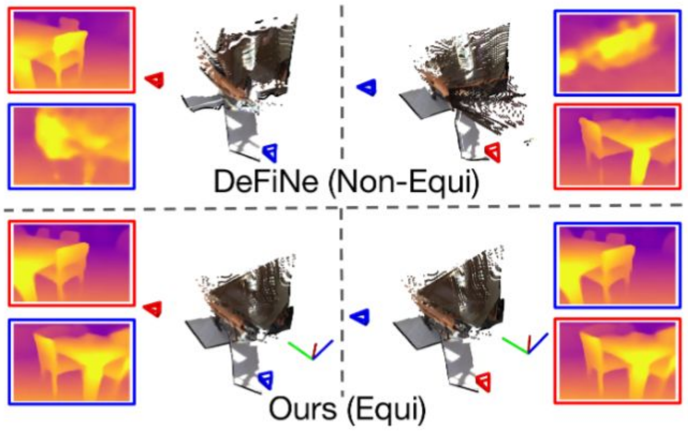


Equivariant Perceiver IO

Latent representation equivariance by design

- **Spherical harmonics** used to encode camera information
- Equivariant encoding -> **invariant latent representation**
- **Standard decoders** can be used





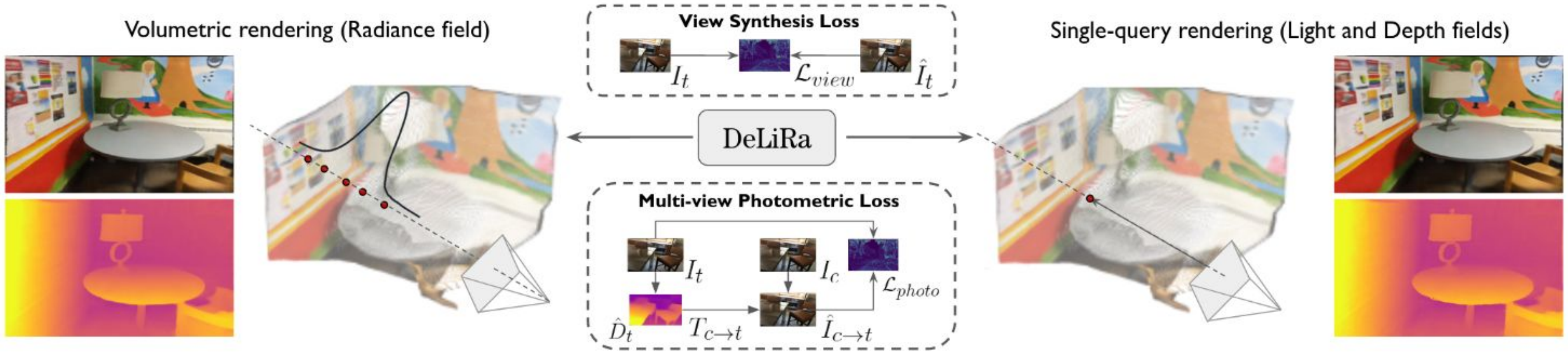
Depth, Light, and Radiance Fields

DeLiRa: Self-Supervised Depth, Light, and Radiance Fields

V Guizilini, I Vasiljevic, J Fang, R Ambrus, S Zakharov, V Sitzmann, A Gaidon ***(ICCV'23)***

Self-supervised photometric warping to eliminate **shape-radiance ambiguity**

Joint decoding of **volumetric** (radiance) **and single-query** (depth and light) heads



Depth, Light, and Radiance Fields

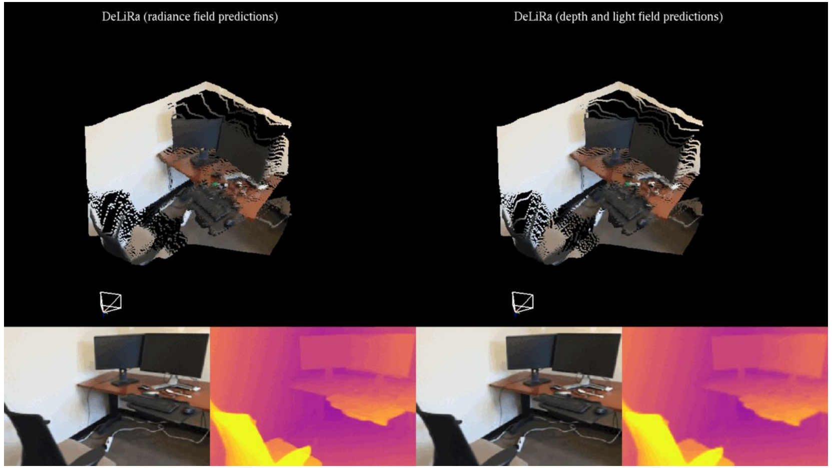
DeLiRa: Self-Supervised Depth, Light, and Radiance Fields

V Guizilini, I Vasiljevic, J Fang, R Ambrus, S Zakharov, V Sitzmann, A Gaidon ***(ICCV'23)***

Synergies between representations

Volumetric predictions increase diversity for single-query training

Depth predictions improve volumetric importance sampling



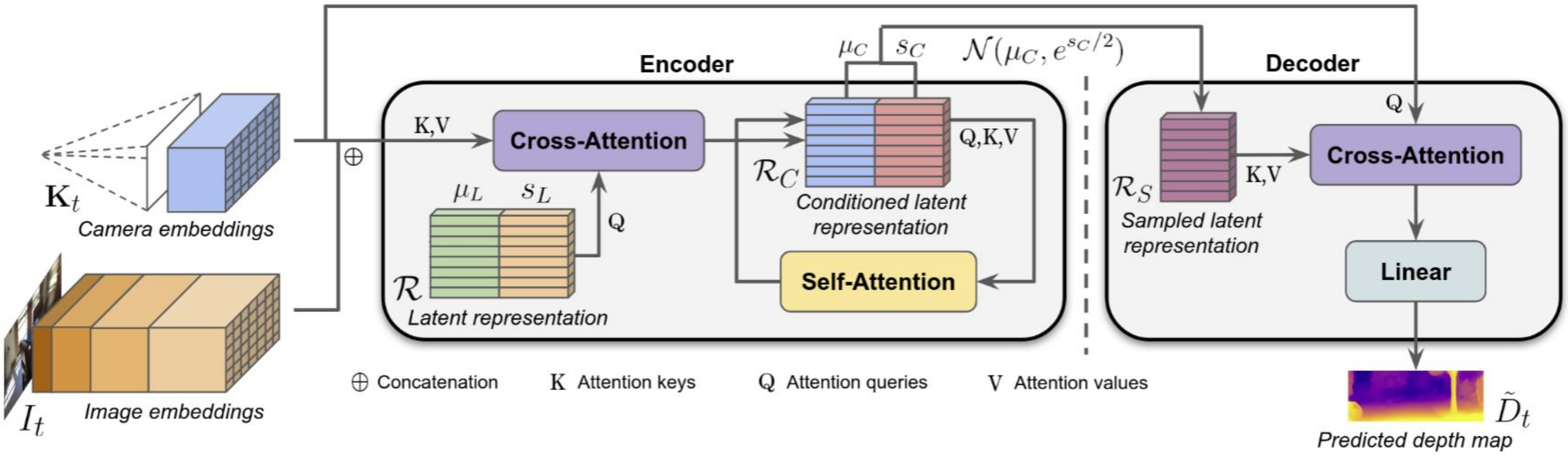
Scale-Aware Metric Depth

Towards zero-shot scale-aware monocular depth estimation

*V Guizilini, I Vasiljevic, D Chen, R Ambruș, A Gaidon **(ICCV'23)***

Metric monocular depth estimation

Camera embeddings used to learn scale priors



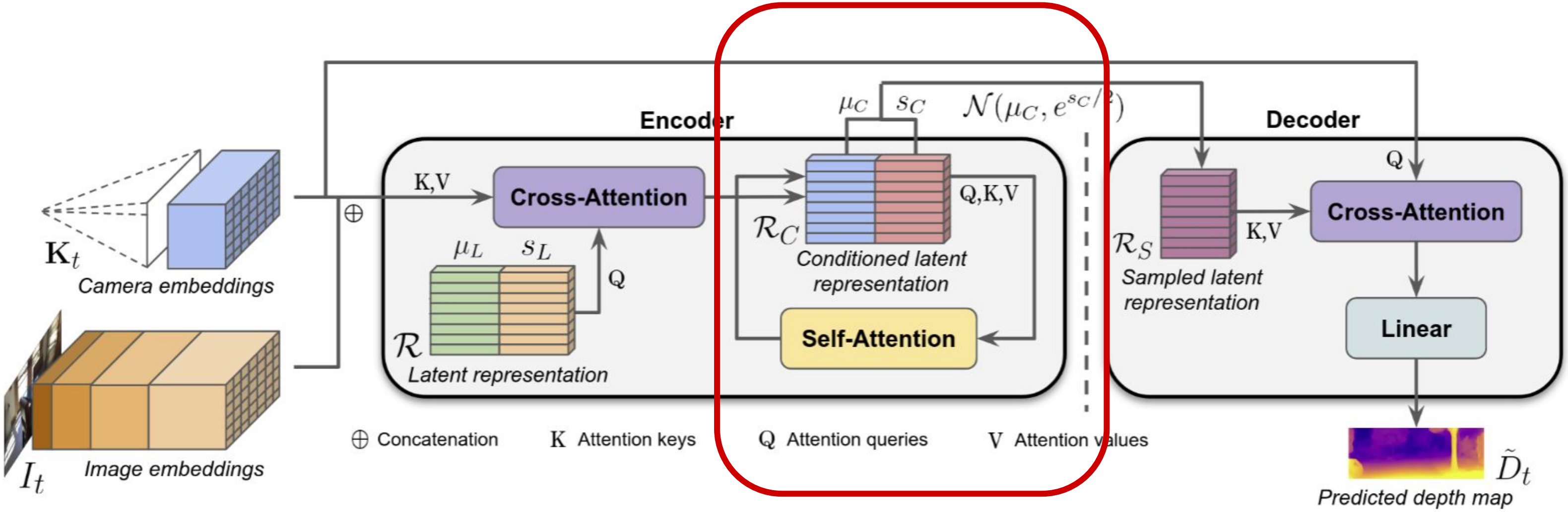
Scale-Aware Metric Depth

Towards zero-shot scale-aware monocular depth estimation

V Guizilini, I Vasiljevic, D Chen, R Ambruș, A Gaidon ***(ICCV'23)***

Variational latent representation

Samples from variational distribution are decoded

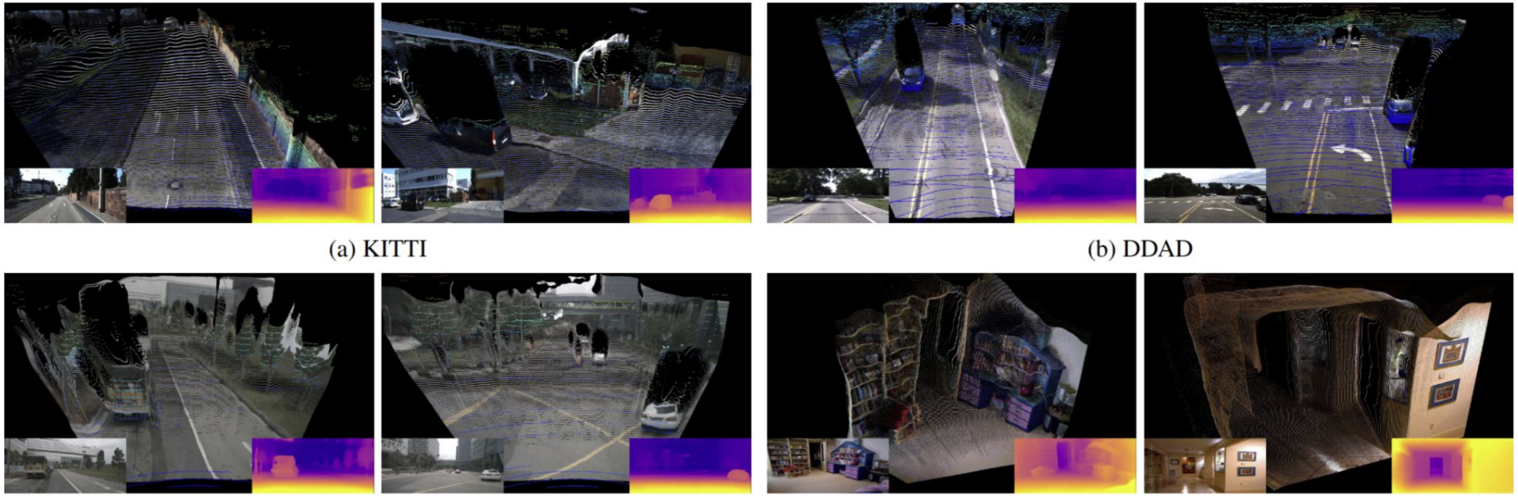


Scale-Aware Metric Depth

Towards zero-shot scale-aware monocular depth estimation

V Guizilini, I Vasiljevic, D Chen, R Ambruș, A Gaidon ***(ICCV'23)***

Zero-shot transfer across both indoor and outdoor domains



(a) KITTI (b) DDAD

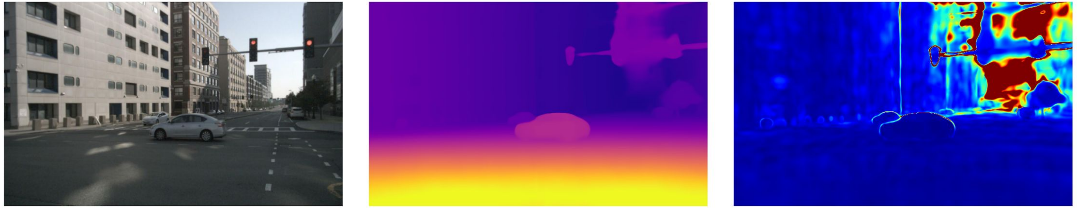
(c) nuScenes (d) NYUv2

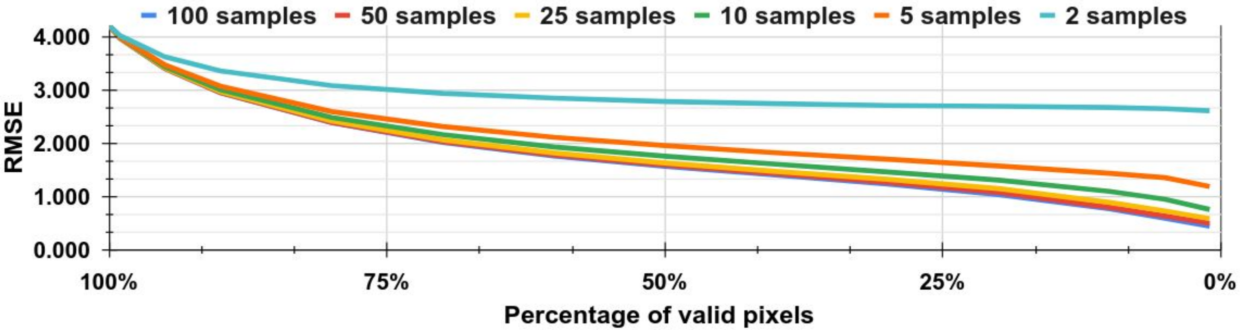
Scale-Aware Metric Depth

Towards zero-shot scale-aware monocular depth estimation

V Guizilini, I Vasiljevic, D Chen, R Ambruș, A Gaidon ***(ICCV'23)***

Improvements in depth estimation by filtering out pixels with high uncertainty

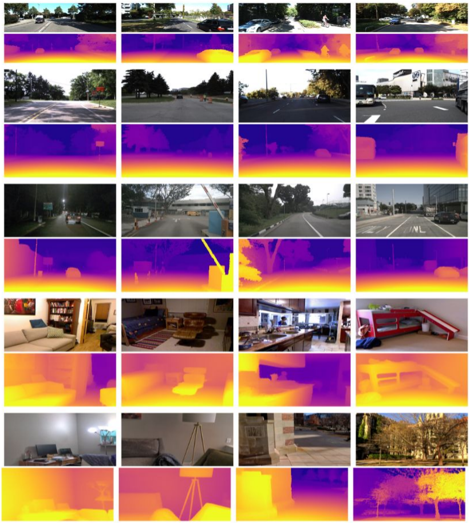


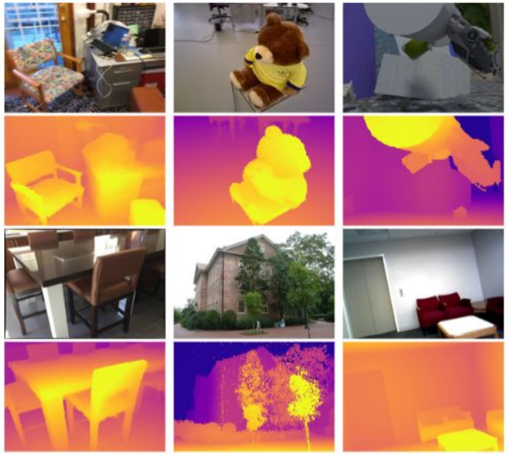


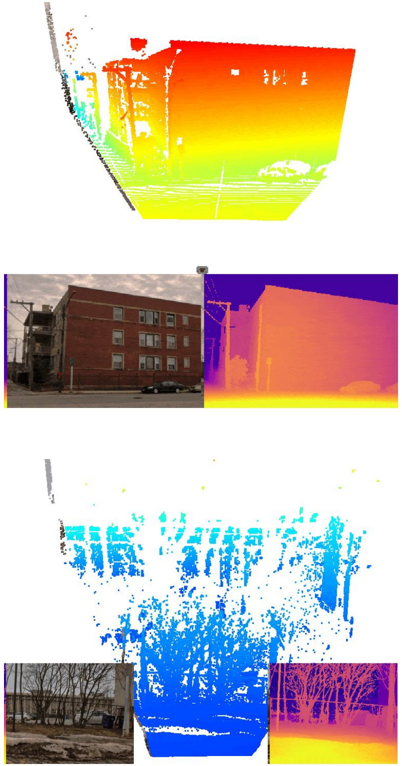
Scale-Aware Metric Depth

Efficient pixel-level diffusion with sparse training data

Improvements over ZeroDepth (and others)







LiDAR Generation

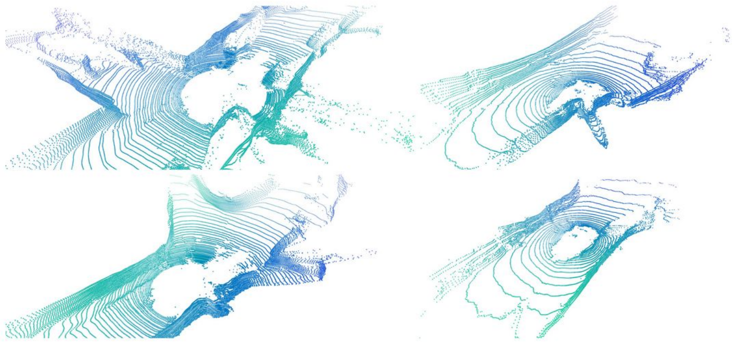
Towards Realistic Scene Generation with LiDAR Diffusion Models

H Ran, V Guizilini, Y Wang ***(CVPR'24)***

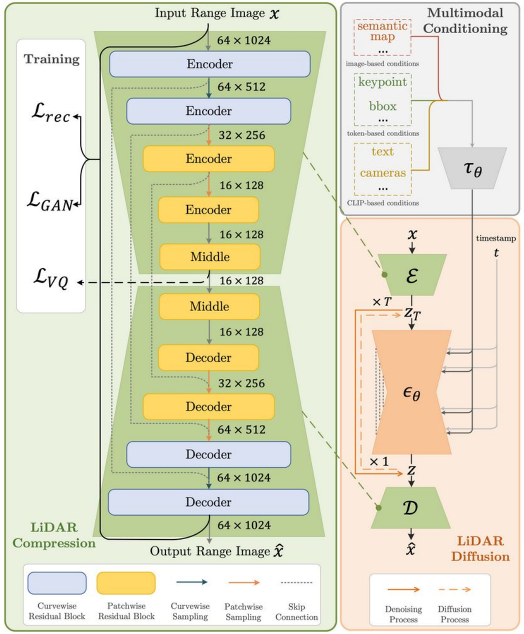
Realistic LiDAR Generation

Latent autoencoder designed to capture LiDAR patterns

- **Patterns:** Curve-wise compression
- **Geometry:** point-wise coordinate supervision
- **Objects:** patch-wise encoding



64-beam   32-beam



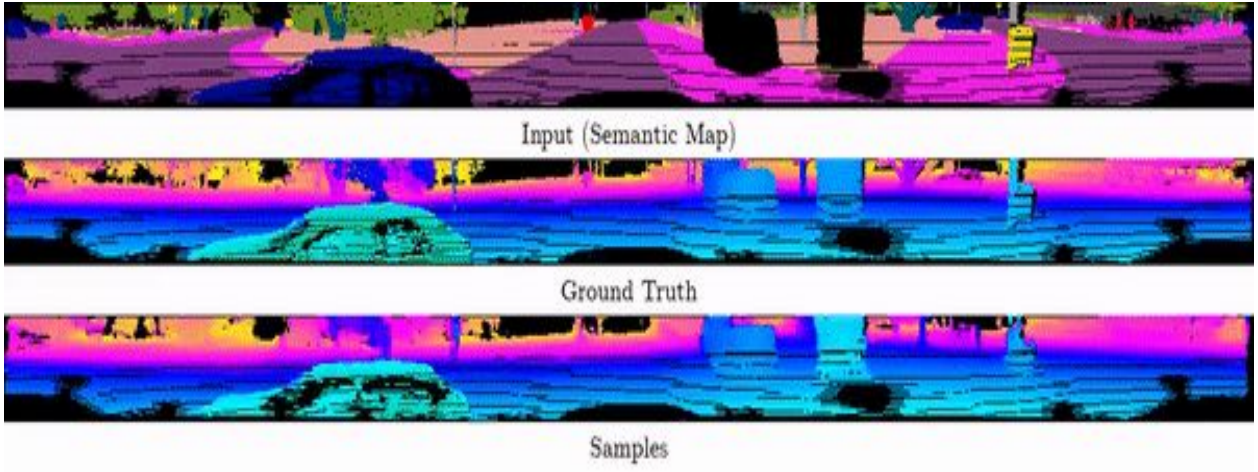
LiDAR Generation

Towards Realistic Scene Generation with LiDAR Diffusion Models

H Ran, V Guizilini, Y Wang ***(CVPR'24)***

Conditional LiDAR generation

Images / semantic maps / bounding boxes / text



Input (Semantic Map)

Ground Truth

Samples

Thank You!

PackNet-SfM: [https://github.com/tri-ml/**packnet-sfm**](https://github.com/tri-ml/packnet-sfm)

Vidar: [https://github.com/tri-ml/**vidar**](https://github.com/tri-ml/vidar)

DDAD: [https://github.com/tri-ml/**ddad**](https://github.com/tri-ml/ddad)

Camviz: [https://github.com/tri-ml/**camviz**](https://github.com/tri-ml/camviz)

https://vitorguizilini.github.io

tri.global/careers