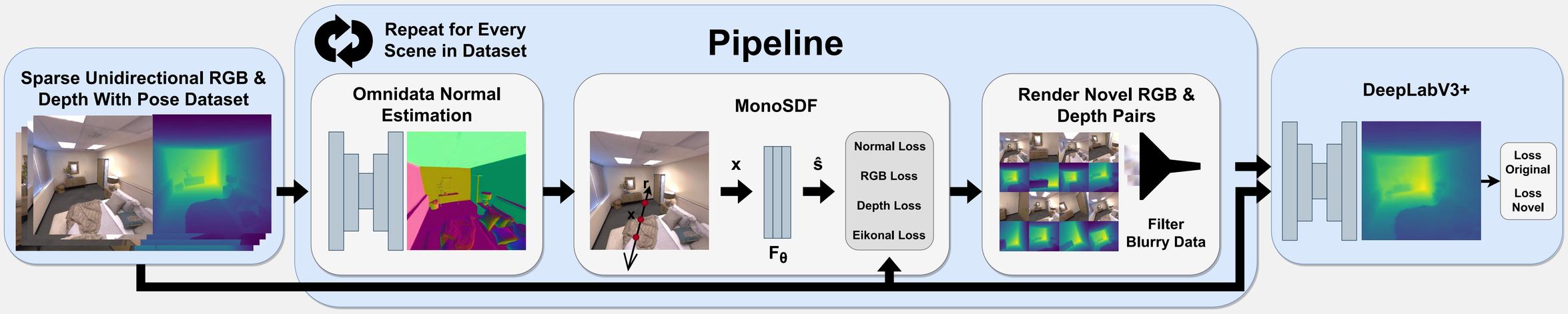
Monocular Depth Estimation with Virtual-view Supervision

Nikolas Hars[1*], Casimir Feldmann[1*], Rong Zou[1*], Kirat Virmani[1*], Dr. Zuria Bauer[1], Dr. Martin R. Oswald[1 2]
Computer Vision and Geometry Group, ETH Zürich
{nihars, cfeldmann, ronzou, kvirmani, zuria.bauer, moswald}@ethz.ch
*equal contributions



Motivation

Problem:
Gathering large amounts of RGBD data necessary for training Monocular Depth Estimation (MDE) networks is laborious and expensive

Contribution:

- Train a Neural Implicit Surface Reconstruction network (NISR) on existing sparse unidirectional RGBD pose data
- Generate novel views and use them as additional geometric supervisory signals during training to improve MDE

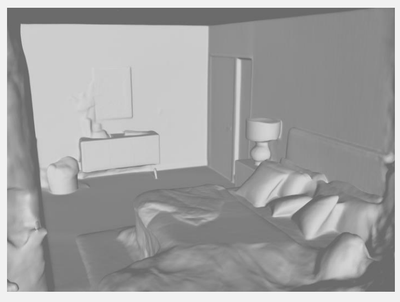
Method Overview

Custom Replica Dataset Split:

- Tiny dataset with constrained viewpoints
- Unidirectionality emulates Monocular Depth-Collection Pipeline, e.g., Autonomous Driving, Phone Sweep
- 2,000 Train / 1,234 Validation / 273 Test

Novel Viewpoint Synthesis (NVS) Pipeline:

- MonoSDF based model using the SDF Studio framework is trained on every scene separately
- Correct metric depth scaling enforced by ground truth depth loss
- Novel viewpoint poses are generated with random uniform translations and rotations of training poses
- The MDE network DeepLabV3+ is trained on a mix of original and filtered novel data



DeepLabV3+ Loss:

$$\mathcal{L} = \mathcal{L}_{berHu}$$

MonoSDF Loss: $\lambda_1 = 1, \lambda_2 = \lambda_3 = 0.1, \lambda_4 = 0.05$

$$\mathcal{L} = \mathcal{L}_{RGB} + \lambda_1 \mathcal{L}_{SensorDepth} + \lambda_2 \mathcal{L}_{Eikonal} + \lambda_3 \mathcal{L}_{Normal}$$

Key Insights

- **Up to 12% higher depth estimation accuracy from identical dataset**
- **Model agnostic → Can be used with off the shelf MDE network**
- **SDF convergence can be difficult and small details are lost**

References: 3: https://github.com/nianticlabs/monodepth2
4: https://github.com/facebookresearch/Replica-Dataset

Quantitative Comparison / Ablation Study

Comparison of Different Augmentation Techniques (Validation)

Technique	RMSE↓	RMSE LOG↓	SQ REL↓	$\delta_{0.5}$↑	δ_1↑	δ_2↑
Original	0.390	0.169	0.063	0.524	0.801	0.976
+ NISR	0.373	0.158	0.059	0.557	0.846	0.986
+ NISR - pretraining	0.380	0.167	0.059	0.560	**0.856**	0.981
+ Masked NISR	**0.356**	**0.154**	**0.054**	**0.565**	0.850	**0.990**

Effect of Relative Number of Extra NISR Images (Validation)

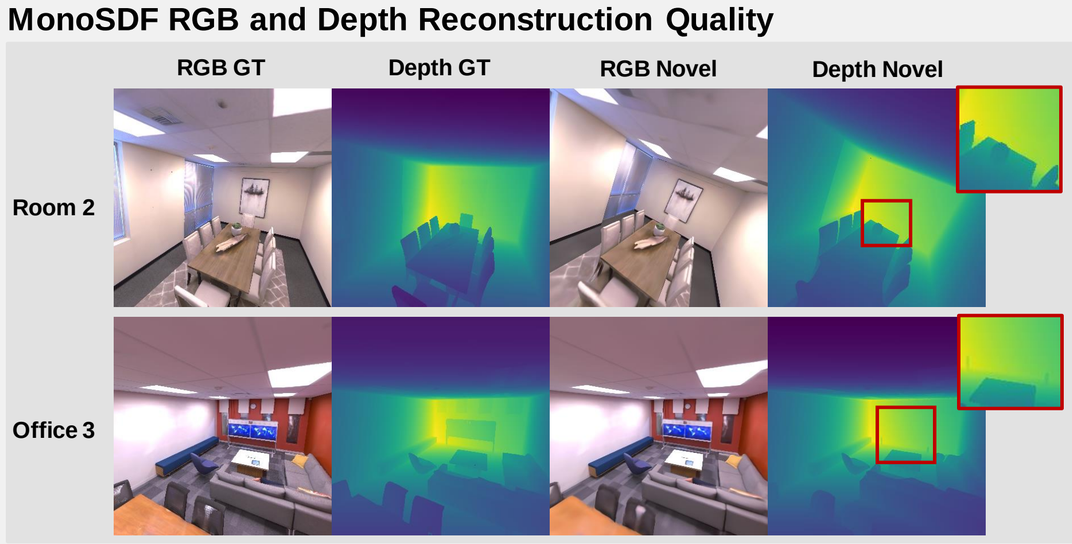
Extra NISR images	RMSE↓	RMSE LOG↓	SQ REL↓	$\delta_{0.5}$↑	δ_1↑	δ_2↑
+0%	0.390	0.169	0.063	0.5239	0.801	0.976
+40%	0.370	0.160	0.061	0.5646	0.845	0.982
+80%	**0.361**	0.159	**0.057**	0.5553	0.836	0.985
+200%	0.379	0.160	0.062	**0.5650**	**0.849**	0.981
+400%	0.373	**0.158**	0.059	0.5572	0.846	**0.986**

Result (Test)

Method	RMSE↓	RMSE LOG↓	SQ REL↓	$\delta_{0.5}$↑	δ_1↑	δ_2↑
Original	0.677	0.352	0.196	0.261	0.450	0.756
+ Masked NISR 400%	**0.562**	**0.318**	**0.179**	**0.276**	**0.504**	**0.825**

Qualitative Comparison

MonoSDF RGB and Depth Reconstruction Quality



Monocular Depth Estimation Comparison