Technische Universität München
Munich Center for Machine Learning
UNIVERSITY OF OXFORD
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Behind the Scenes
Density Fields for Single View Reconstruction

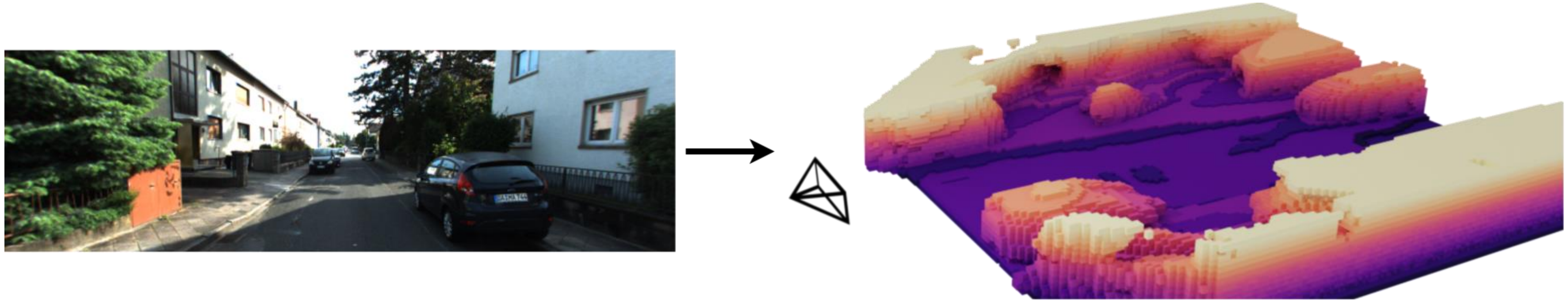
CVPR 2023 – Tag: WED-AM-081

`fwmb.github.io/bts`



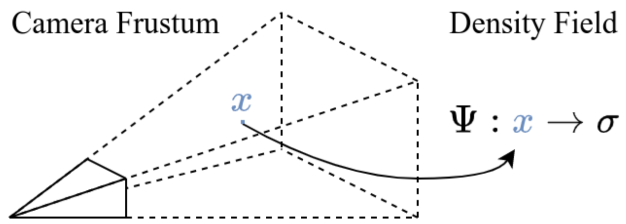

Felix Wimbauer[1,2] Nan Yang[1] Christian Rupprecht[3] Daniel Cremers[1,2,3]

[1]Technical University of Munich [2]MCML [3]University of Oxford



Behind the Scenes

A **self-supervised** method for **volumetric reconstruction** of a scene from a **single image**.



Density Field

A function ψ that maps **every location** $\boldsymbol{x}$ in the camera frustum to **volumetric density** $\boldsymbol{\sigma}$.

Training

Self-supervised from **only (stereo) video data**.

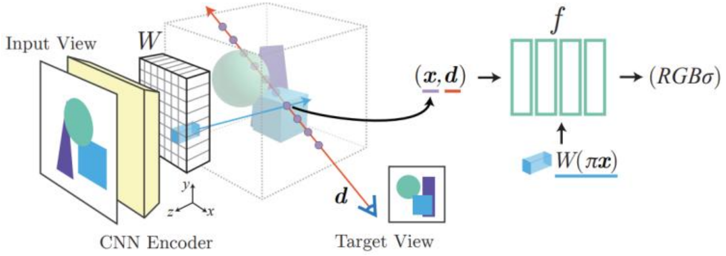
---

vs. Monocular Depth Prediction
e.g. Monodepth 2[1]

☑ We can reason about **occluded areas.**

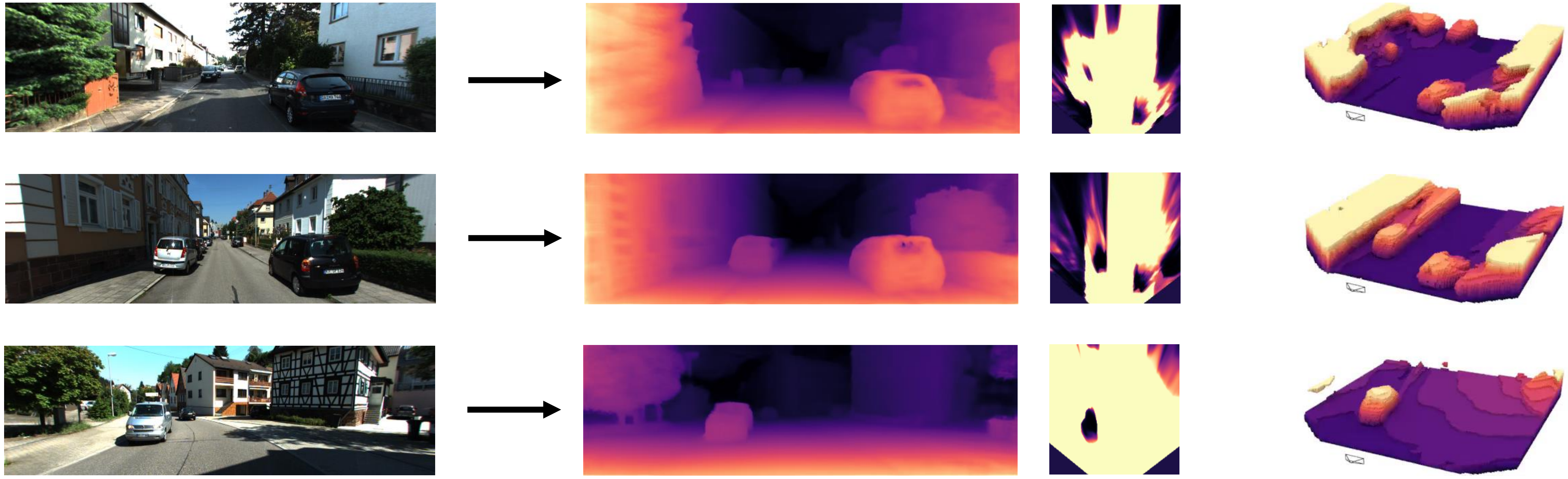
vs. Learnable NeRFs
e.g. PixelNeRF[2]

☑ We achieve **better generalization.**



[1] Godard et al., Digging into Self-Supervised Monocular Depth Prediction, ICCV 2019
[2] Yu, et al. Pixelnerf: Neural radiance fields from one or few images, CVPR 2021

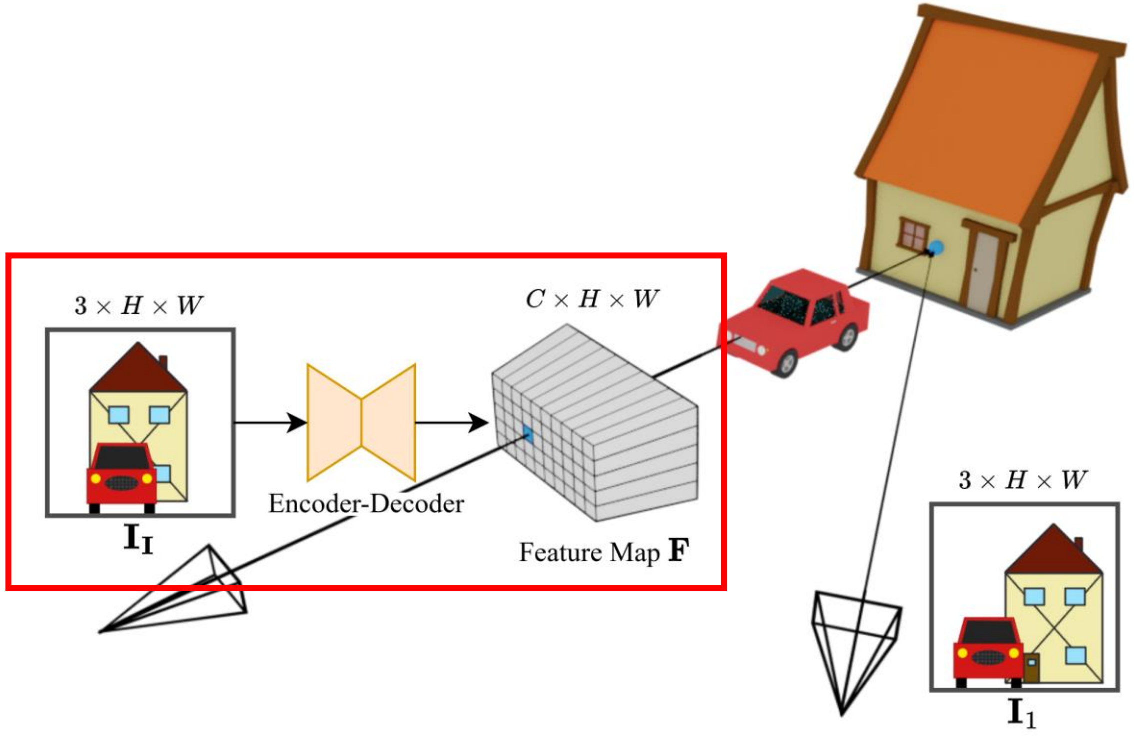
Results


Input Image
Expected Ray Termination Depth
Birds-Eye View
Voxelization
Volumetric Reconstruction on **KITTI-360**

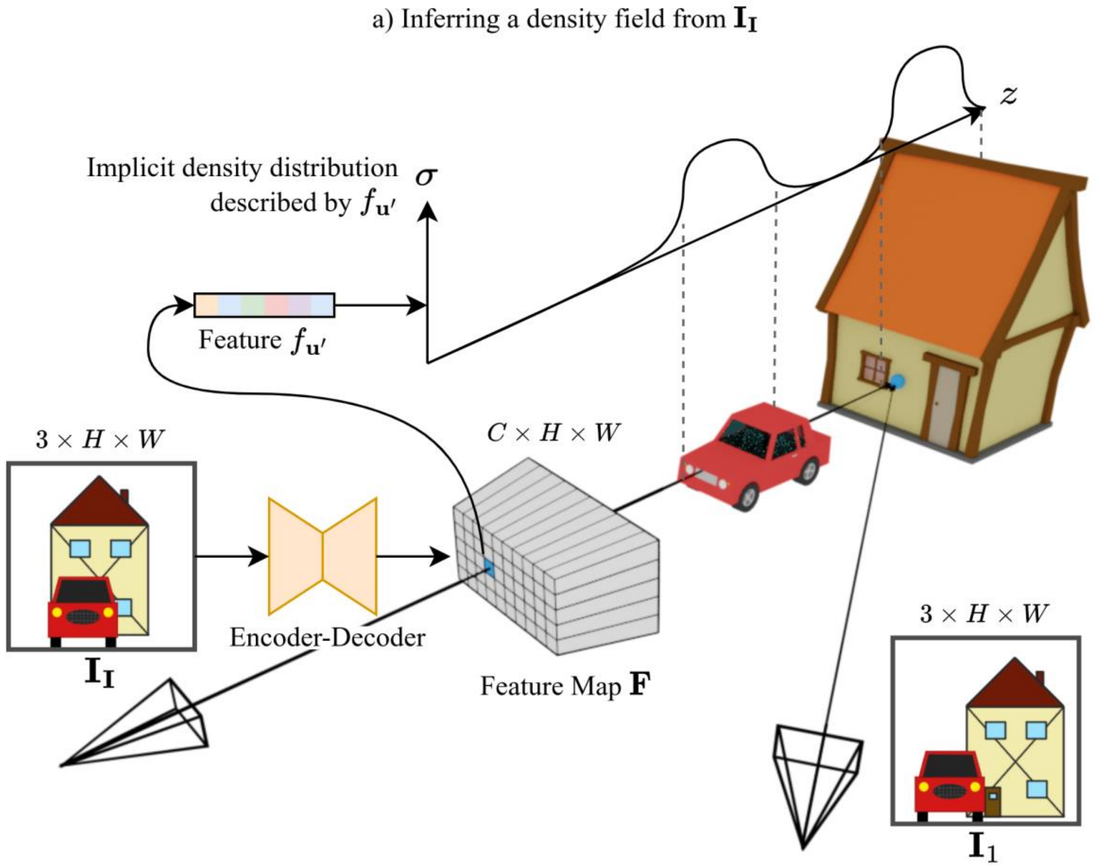

Novel View Synthesis
KITTI
RealEstate10K

Model Architecture

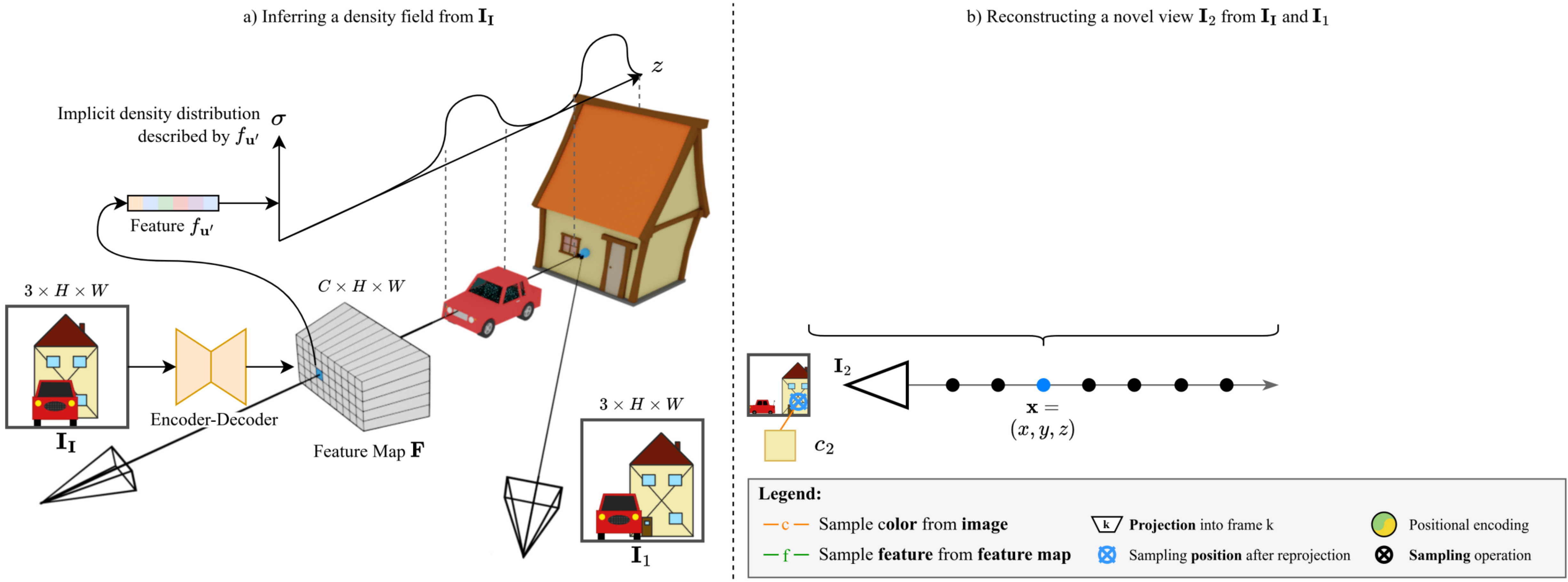
a) Inferring a density field from $\mathbf{I_I}$



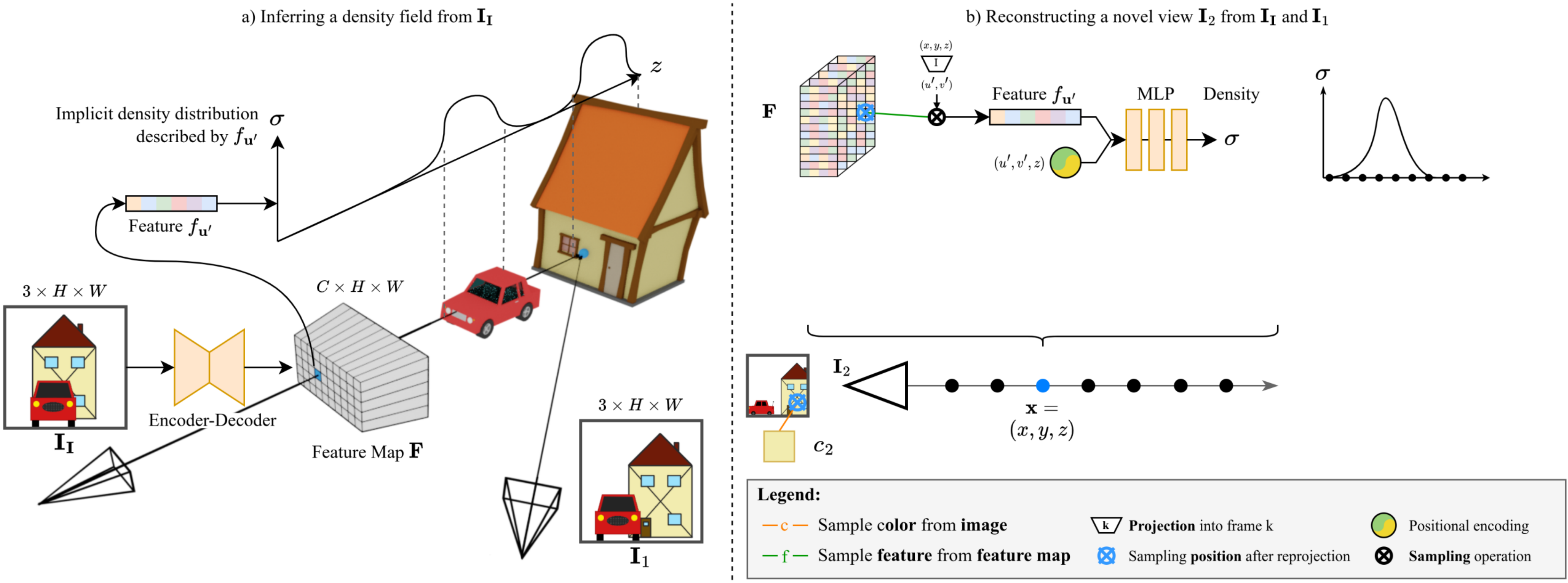
Model Architecture



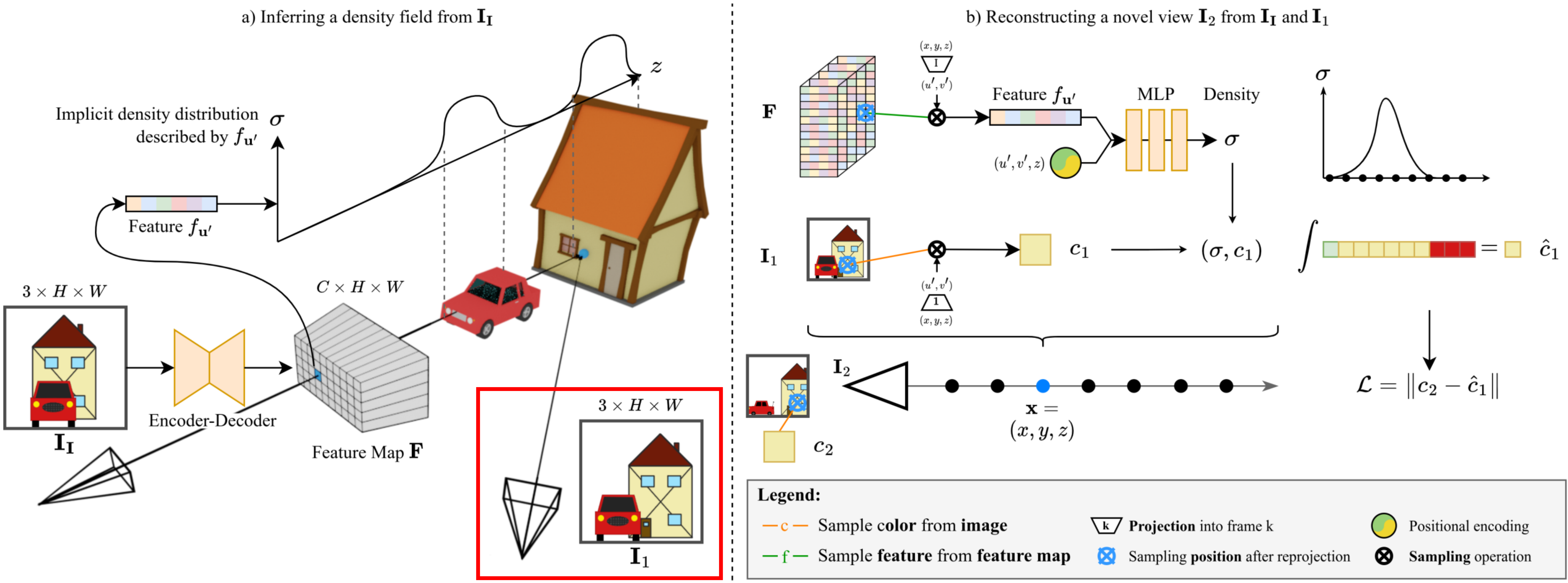
Model Architecture



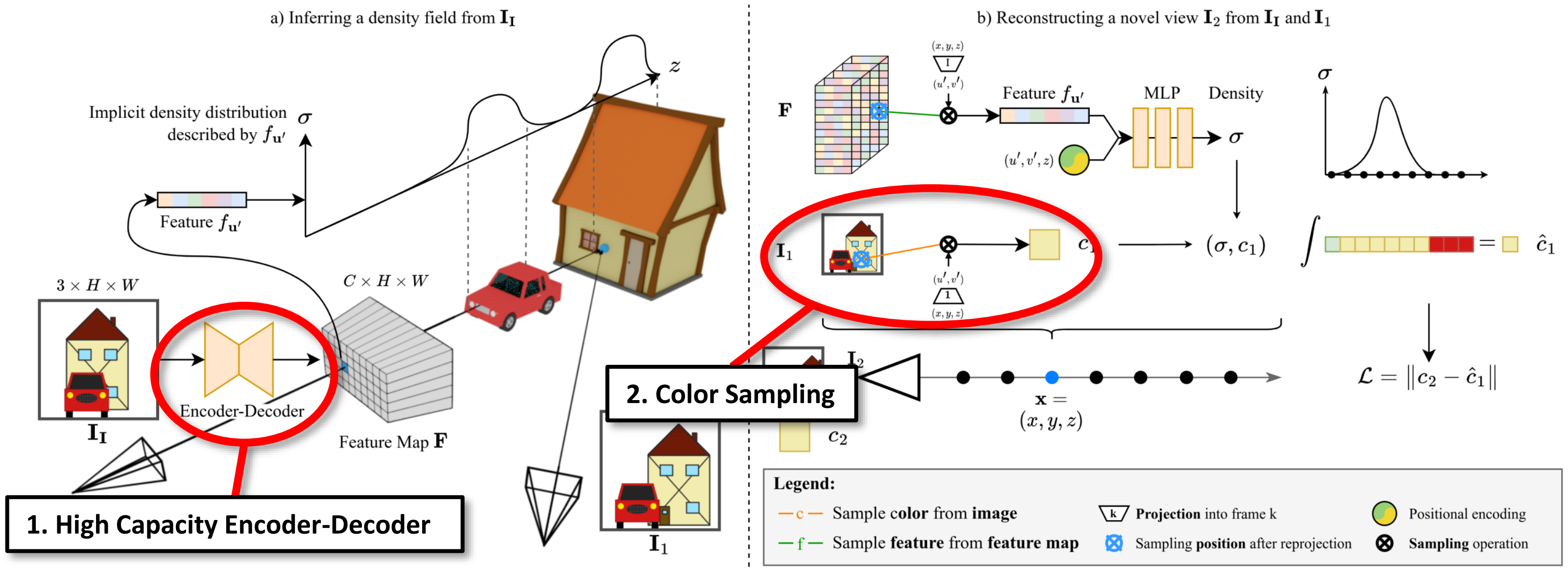
Model Architecture



Model Architecture



Model Architecture



1. Shift capacity from MLP to feature extractor

→ **MLP** can only reason about **local geometry**
→ **Encoder-Decoder** has to capture **entire scene**
→ Better **generalization**

2. Sample color instead of the MLP predicting color

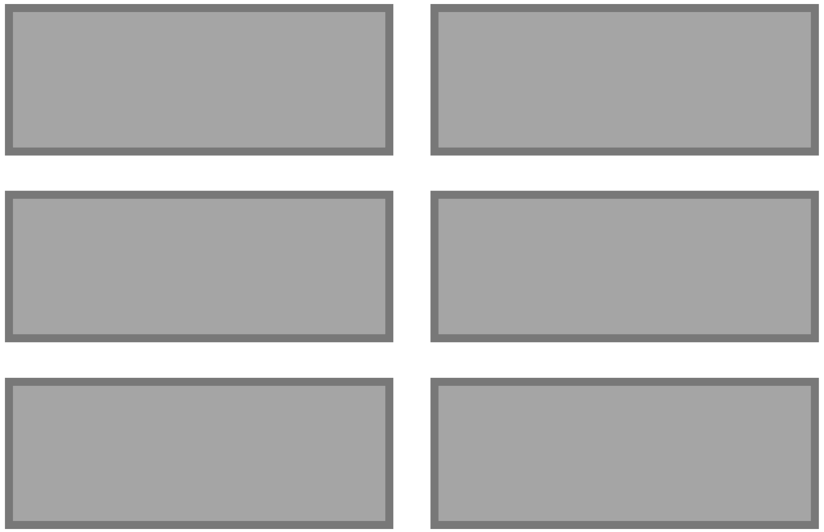
→ Implicit field function becomes **simpler**
→ Enforces **multi-view consistency**
→ More **training stability**, **fewer artifacts**

Self-Supervised Training

During training, multiple views are available:

- One view is considered the **input image**
- All views are partitioned into **Loss** and **Render** views

Available views during training



t = 1

t = 0

t =-1

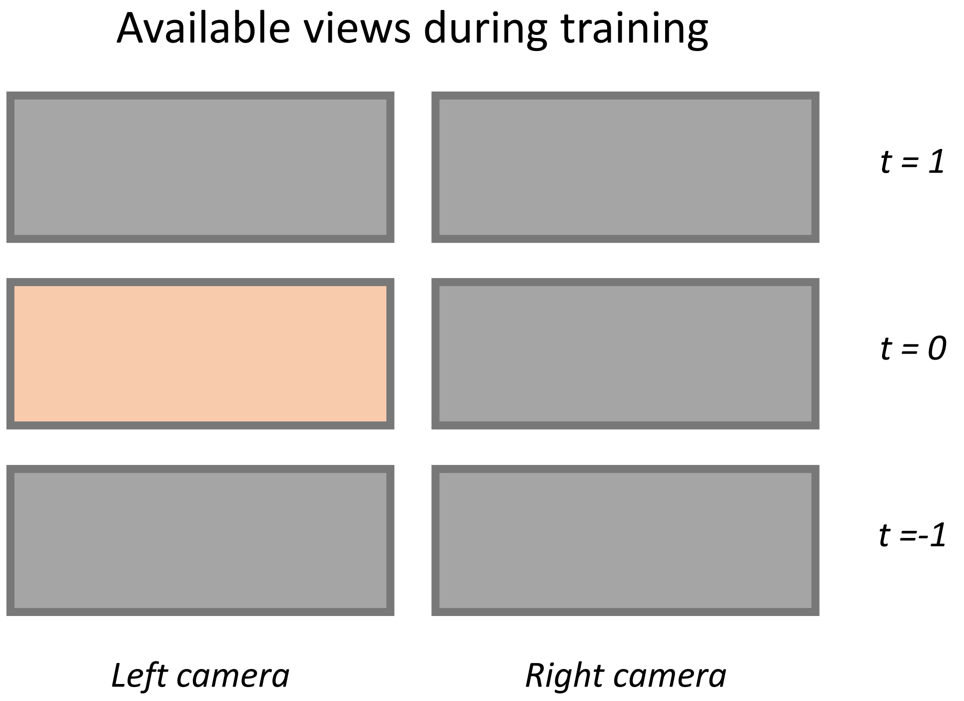
Left camera

Right camera

Self-Supervised Training

During training, multiple views are available:

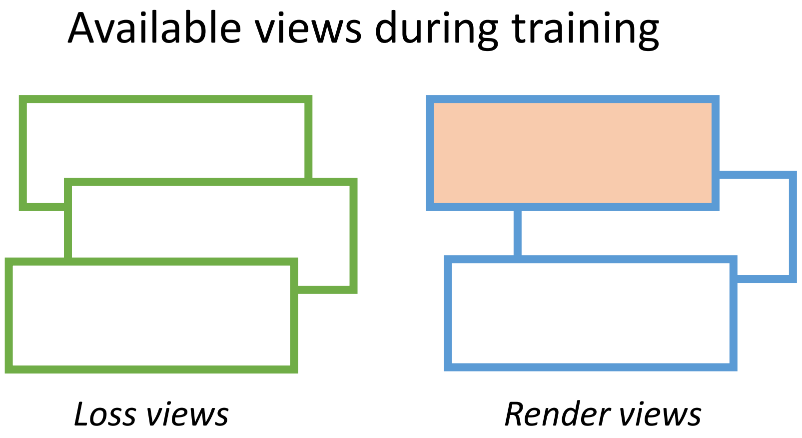
- One view is considered the **input image**
- All views are partitioned into **Loss** and **Render** views



Self-Supervised Training

During training, multiple views are available:

- One view is considered the **input image**
- All views are partitioned into **Loss** and **Render** views

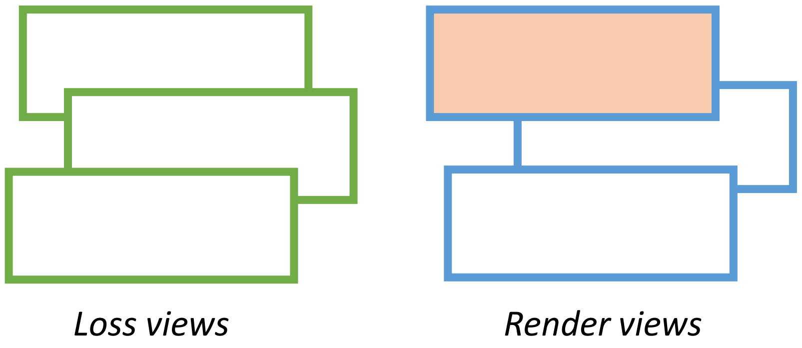


Self-Supervised Training

During training, multiple views are available:

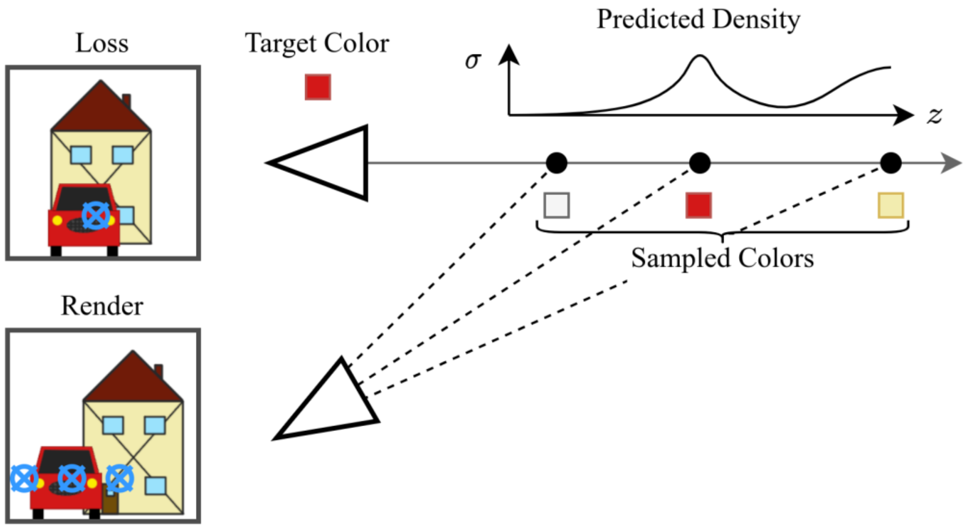
- One view is considered the **input image**
- All views are partitioned into **Loss** and **Render** views

Available views during training

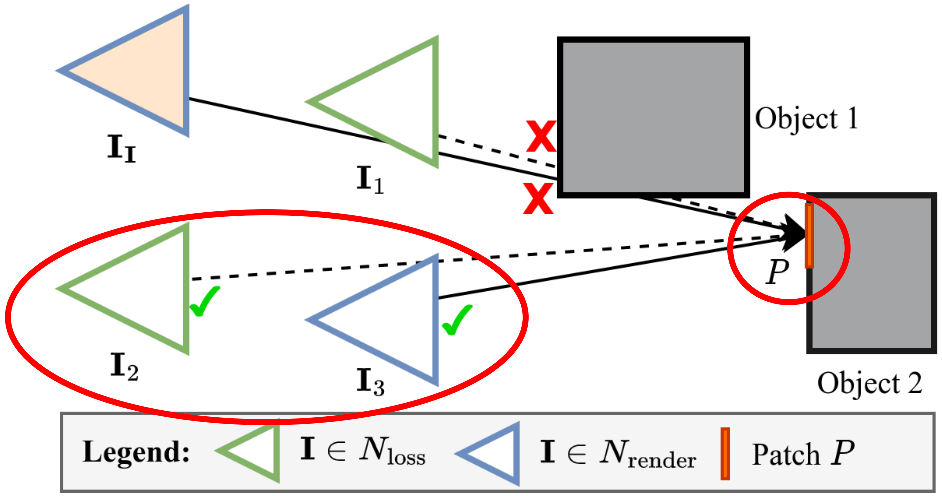


Reconstruction loss:

- Perform volume rendering to reconstruct **Loss** views based on the **predicted density**
- Sample color from **Render** views
- Use **photometric consistency** as supervision signal



Self-Supervised Training



Learning Geometry in Occluded Regions

Traditional reprojection loss formulations do not give training signals for areas occluded in the input image.

→ Our density field allows reconstructing **any frame** from **any other frame**

→ We can reconstruct **P** in view $\mathbf{I_2}$ by sampling colors from $\mathbf{I_3}$

→ To minimize the loss, our network has to predict correct geometry for **P**, even though **P** is occluded in $\mathbf{I_I}$

→ This requires at **least two extra views** other than the input view.

Datasets

Datasets

KITTI-360[1]

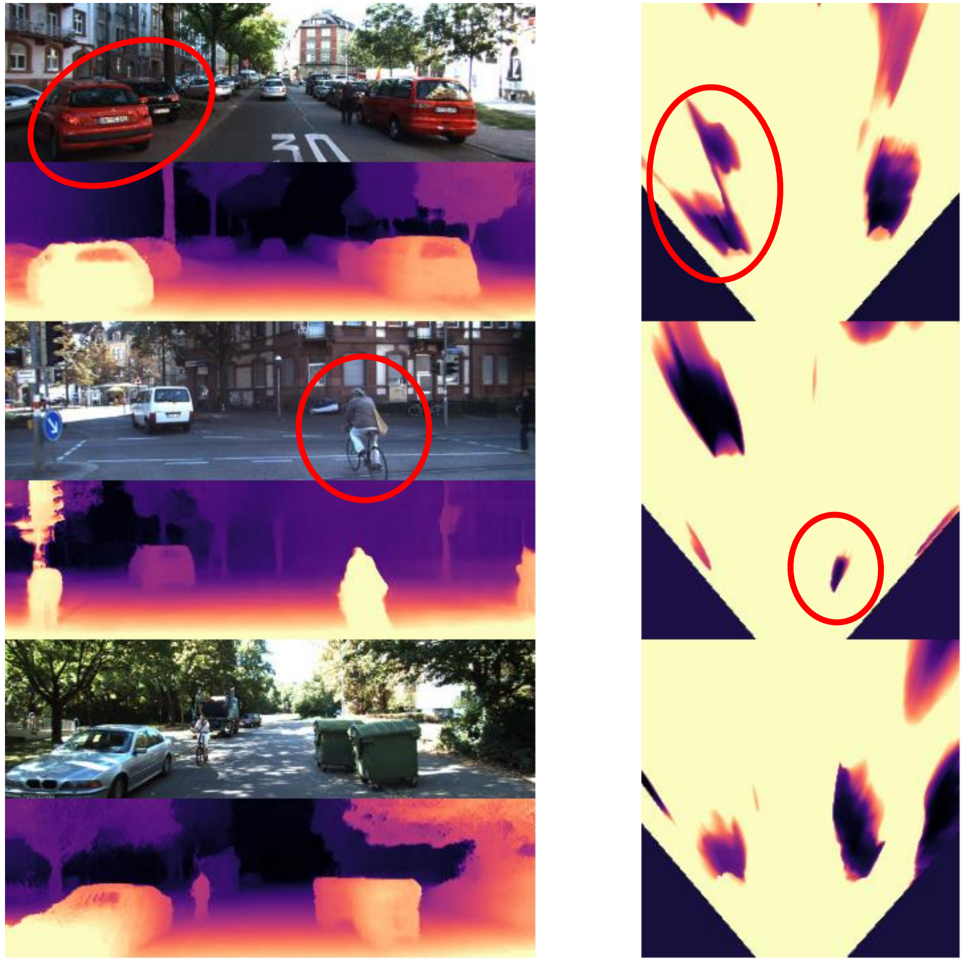
KITTI[2]

RealEstate10K[3]

[1] Liao et al., KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d , TPAMI 2022
[2] Geiger et al., Vision meets Robotics: The KITTI Dataset, IJRR 2013
[3] Zhou et al., Stereo magnification: Learning view synthesis using multiplane images, SIGGRAPH 2018

Occupancy Estimation - KITTI

Birds-Eye View (dark = high density)



Input & Predicted Depth

Ours

[1] **Monodepth2**: Godard et al., Digging into Self-Supervised Monocular Depth Prediction, ICCV 2019
[2] **PixelNeRF**: Pixelnerf: Neural radiance fields from one or few images, CVPR 2021
[3] **MINE**: Li et al., Mine: Towards continuous depth mpi with nerf for novel view synthesis, ICCV 2021

Occupancy Estimation - KITTI

Method	O_{acc} ↑	IE_{acc} ↑	IE_{rec} ↑
Depth† [14]	**0.94**	n/a	n/a
Depth† + *4m* [14]	0.91	0.63	0.22
PixelNeRF† [57]	0.92	0.63	**0.43**
Ours (No *S*, *F*)	**0.94**	0.70	0.06
Ours (No *F*)	**0.94**	0.71	0.09
Ours	**0.94**	**0.77**	**0.43**

Occupancy Estimation against aggregated LiDAR Scans form multiple timesteps.

Model	Volum.	Split	Abs Rel ↓	RMSE ↓	$\alpha < 1.25$ ↑
PixelNeRF [57]	✓		0.130	5.134	0.845
EPC++ [29]	✗		0.128	5.585	0.831
MonoDepth2 [14]	✗		0.106	4.750	0.874
PackNet [16]	✗	Eigen [10]	0.111	4.601	0.878
DepthHint [51]	✗		0.105	4.627	0.875
FeatDepth [44]	✗		0.099	4.427	0.889
DevNet [60]	(✓)		**0.095**	**4.365**	**0.895**
Ours	✓		0.102	4.407	0.882
MINE [23]	✓	Tuls. [49]	0.137	6.592	0.839
Ours	✓		**0.132**	**6.104**	**0.873**

Depth prediction against state-of-the-art monocular depth prediction methods.

[1] **Monodepth2**: Godard et al., Digging into Self-Supervised Monocular Depth Prediction, ICCV 2019
[2] **PixelNeRF**: Pixelnerf: Neural radiance fields from one or few images, CVPR 2021
[3] **MINE**: Li et al., Mine: Towards continuous depth mpi with nerf for novel view synthesis, ICCV 2021

EPC++: Luo et al., Every pixel counts++: Joint learning of geometry and motion with 3d holistic understanding, TPAMI 2019
PackNet: Guizilini et al., 3d packing for self-supervised monocular depth estimation, CVPR 2020
DepthHint: Watson et al., Self-supervised monocular depth hints, ICCV 2019
FeatDepth: Shu et al., Feature-metric loss for self-supervised learning of depth and egomotion, ECCV 2020
DevNet: Zhou et al., Devnet: Self-supervised monocular depth learning via density volume construction, ECCV 2022

Qualitative Results – KITTI-360

Inference per frame on test sequences from KITTI-360. We show smooth transitions between expected ray termination depth, novel view synthesis, and birds-eye view.

Novel View Synthesis – KITTI & RealEstate10K

Technische Universität München

Munich Center for Machine Learning

UNIVERSITY OF OXFORD

Come and visit our poster at June 21th 10:30am – 12:30pm!

Behind the Scenes
Density Fields for Single View Reconstruction

- ✓ **Volumetric reconstruction** from a **single image,** even in occluded areas.
- ✓ New **density field formulation** and **improved architecture** enable training on challenging datasets and improve generalization.
- ✓ A self-supervised training scheme from **only (stereo) video**.

For **code, pretrained models and more,**
please visit our project page at fwmb.github.io/bts