SCADE: NeRFs from Space Carving with Ambiguity-aware Depth Estimates

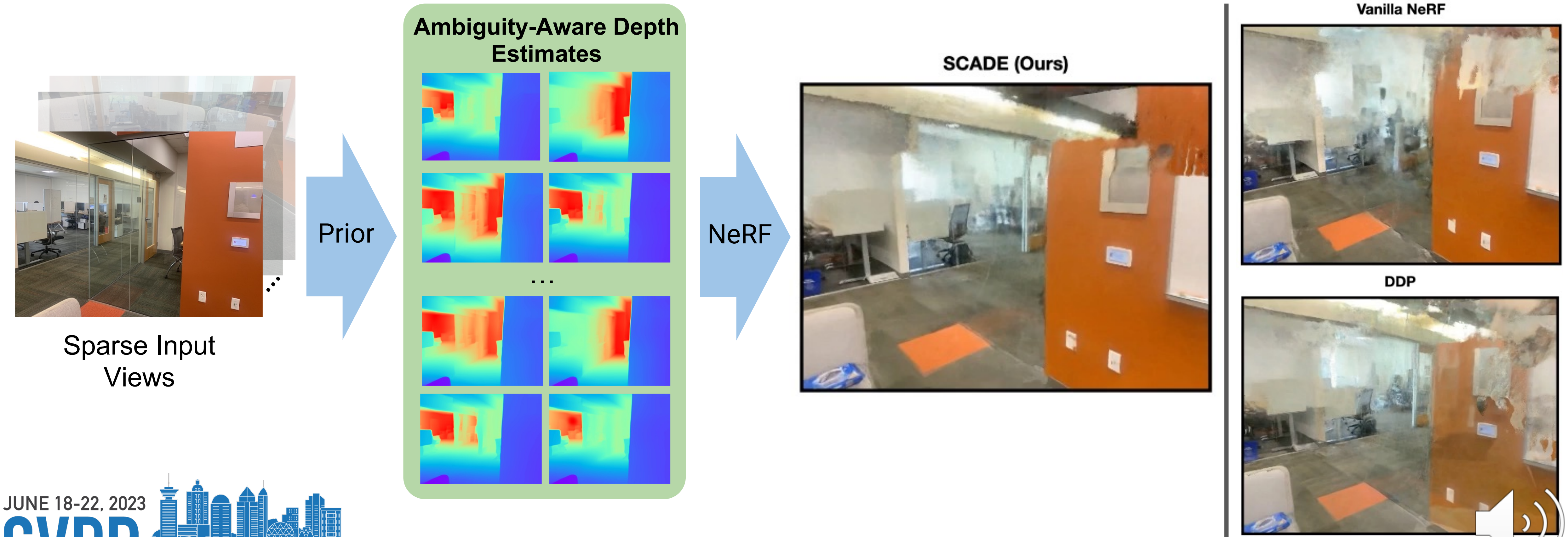
Mikaela Angelina Uy [1,2], Ricardo Martin-Brualla [2], Leonidas Guibas [1,2] , Ke Li [2,3]

Stanford University [1]

Google [2]

SIMON FRASER UNIVERSITY

Simon Fraser University [3]



JUNE 18-22, 2023

CVPR

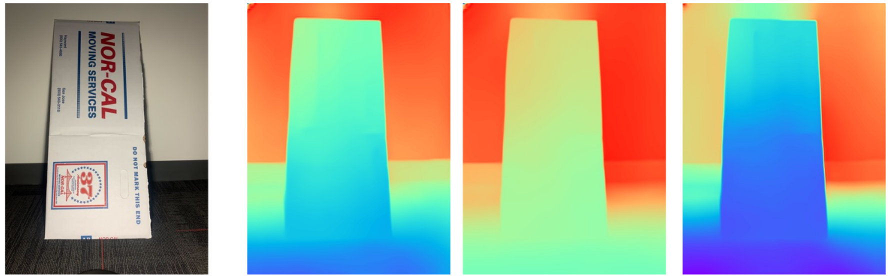
VANCOUVER, CANADA

Visit us at Poster: THU-AM-004

Overview

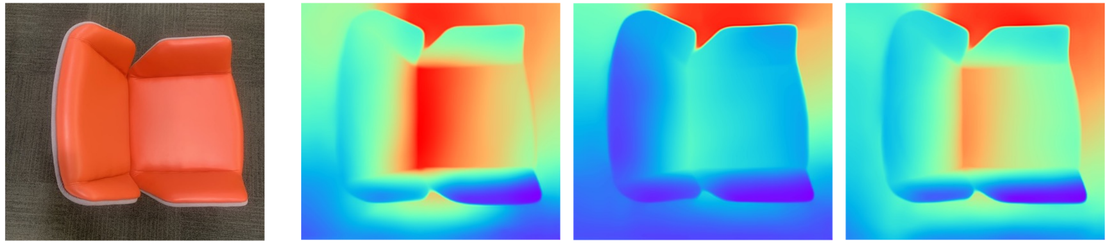
- There can be multiple, equally valid depth estimates given a single image.
- I.e. Monocular depth is inherently ambiguous.

Albedo vs Shading



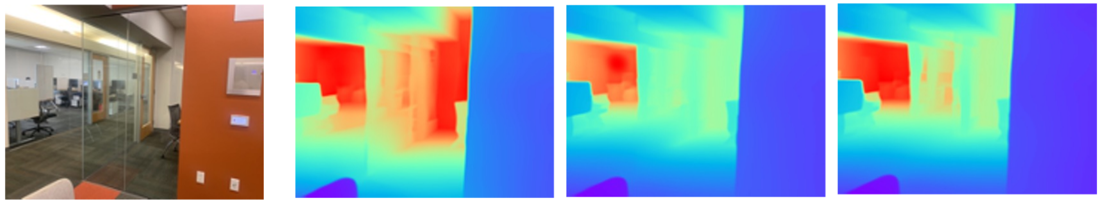
Possible depth maps

Scale / Degree of Convexity



Possible depth maps

Non-opaque surfaces

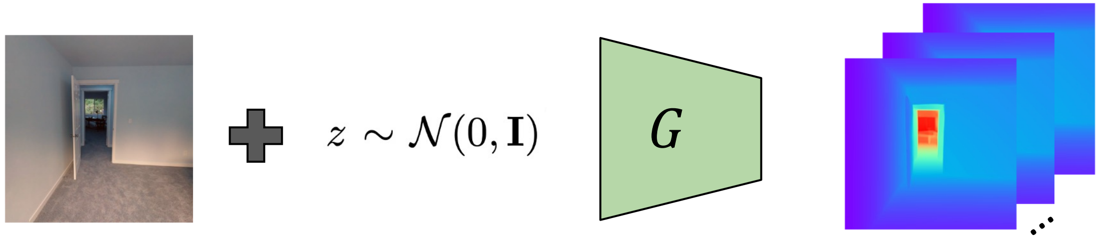


Possible depth maps

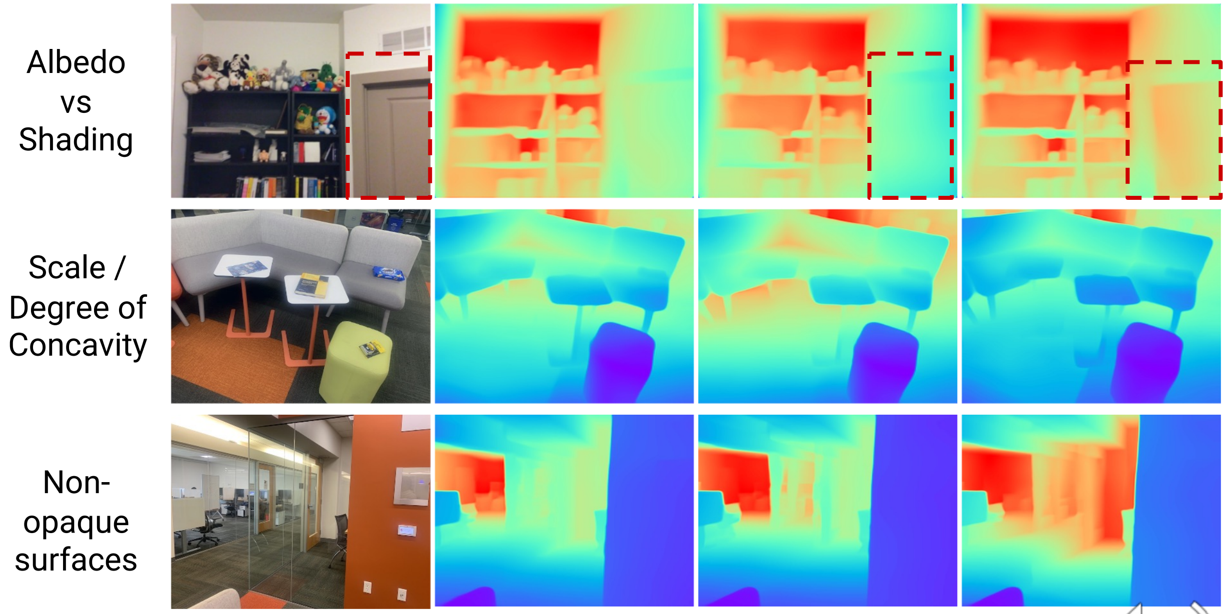
[1] The Bas-Relief Ambiguity. P. N. Belhumeur, et. al., IJCV 1999.

Overview

- Our prior represents depth as a distribution, to handle ambiguity.
    - This distribution can be **multimodal**.



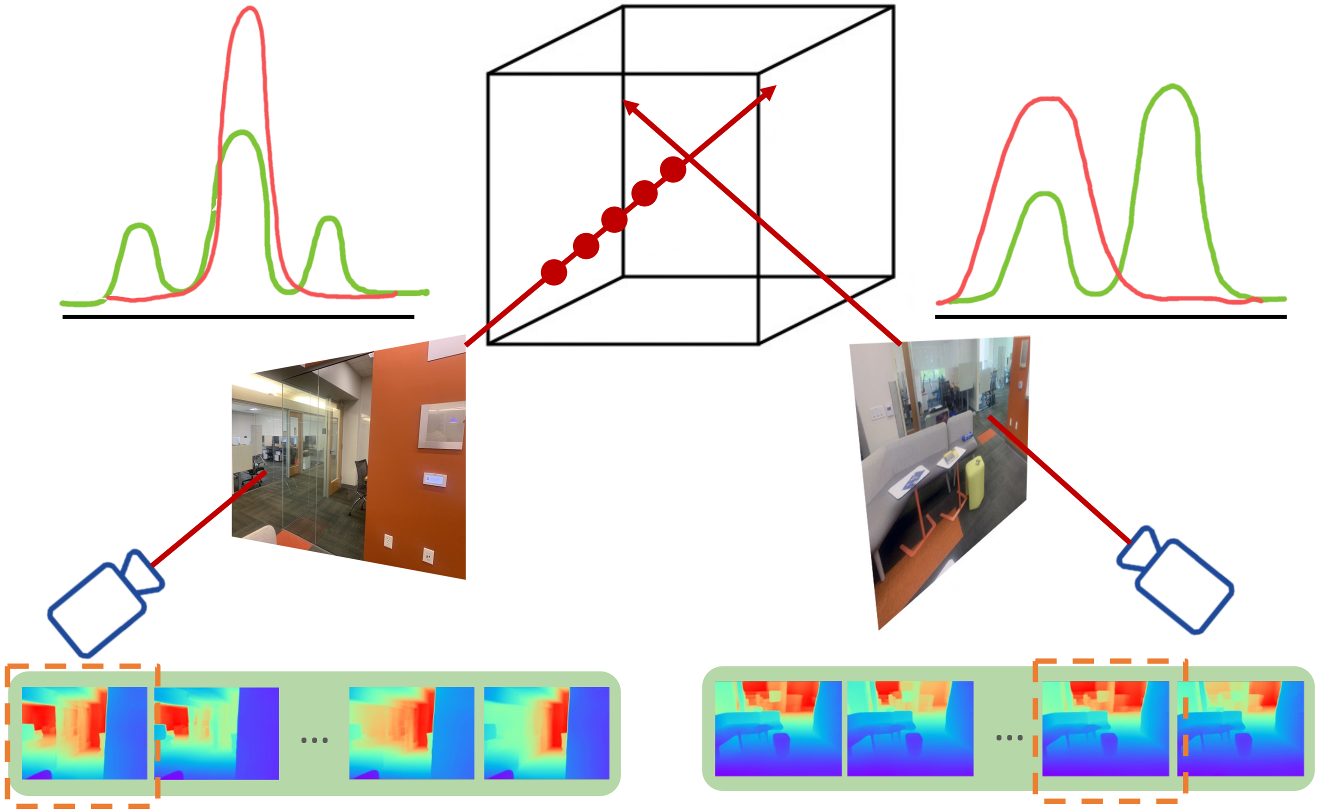
- Represent ambiguities and capture variable modes through **samples** via conditional Implicit Maximum Likelihood Estimation (cIMLE).



[2] Multimodal Image Synthesis with Conditional Implicit Maximum Likelihood Estimation. K. Li, et. al., IJCV 2020.

Overview

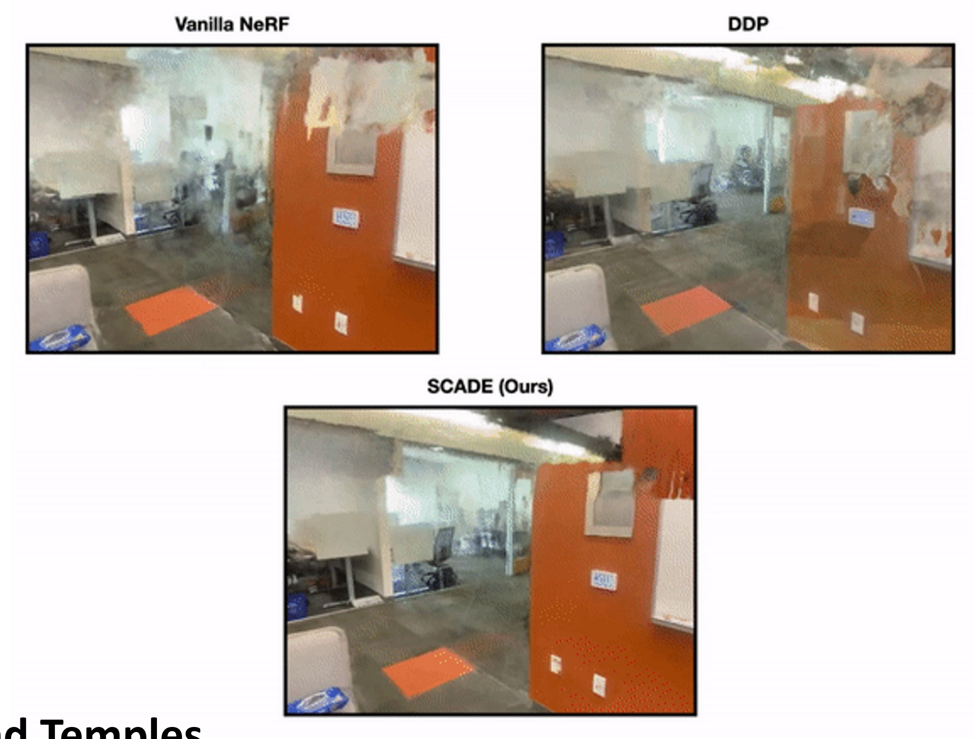
- Resolve ambiguities by **fusing** together information from multiple views.
- **Mode seeking**: finds the consistent agreement across views.
- **Sample-based loss** on the distribution instead of the moments leads to supervision in 3D instead of 2D.



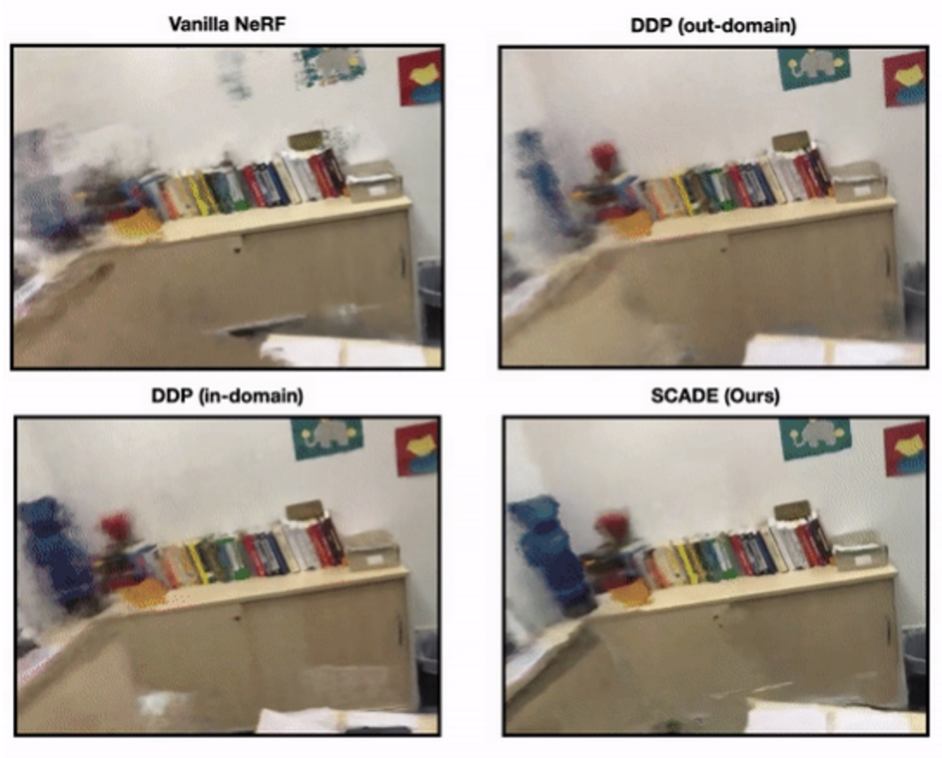
[3] A Theory of Shape by Space Carving. K. Kutulakos and S. Seitz, IJCV 2000.

Overview

In-the-Wild Scenes



Scannet



Tanks and Temples

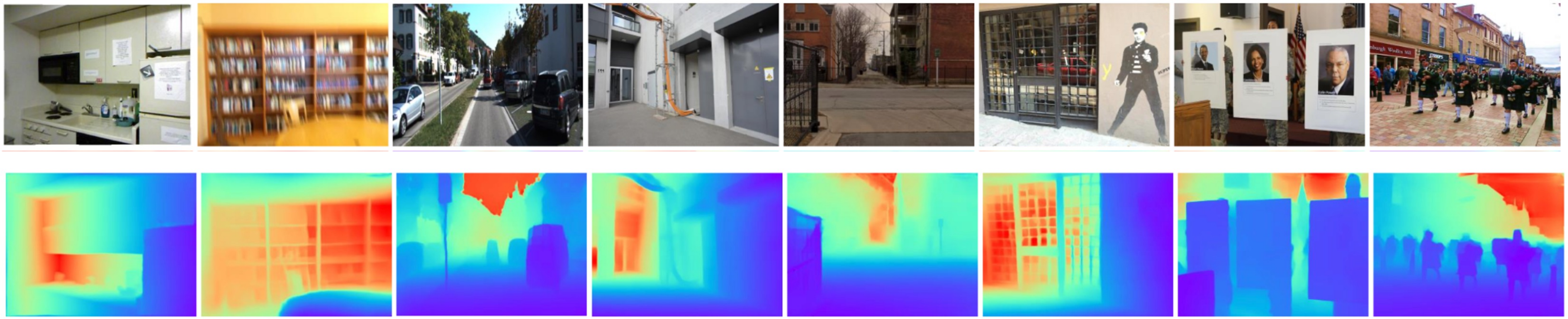
Idea

Prior: Depth	Fuse: Space Carving	Ambiguity	Distribution	Generalize Space Carving

Idea

- **Monocular Depth Estimation**
  - Category agnostic
  - Generalizes to in-the-wild scenes

Image taken from [1]

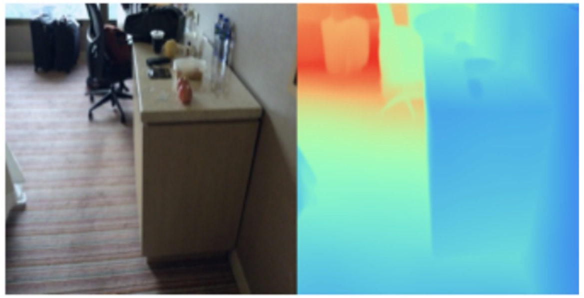


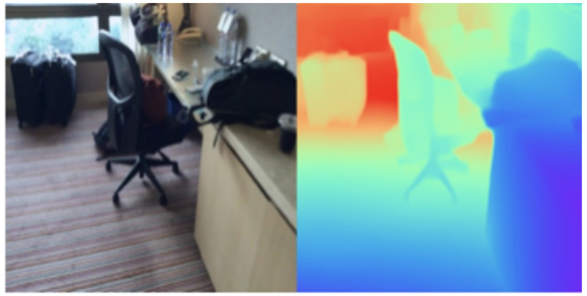
NYU Scannet KITTI ETH3D DIODE In-the-Wild Scenes

[4] Learning to Recover 3D shape from a Single Image. W. Yin, et. al., CVPR 2021.

Idea

- **Fuse**
  - How do we fuse depths from multiple views?
  - Space Carving!





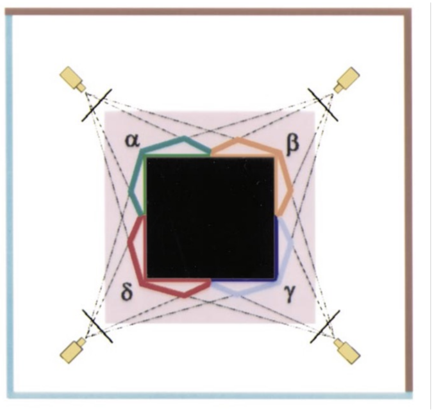
[3] A Theory of Shape by Space Carving. K. Kutulakos and S. Seitz, IJCV 2000.

Idea

- **Classical Space Carving**
  - Finds the geometry that satisfies the different views.
  - “Carves” out empty space

Image taken from [3]

(a) (b) (c)



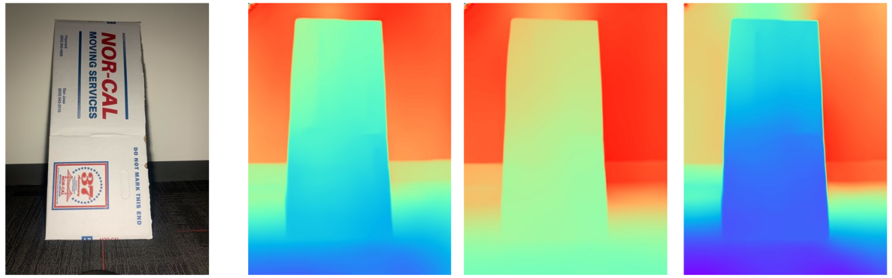
  - Works great with ground truth depth. But…

[3] A Theory of Shape by Space Carving. K. Kutulakos and S. Seitz, IJCV 2000.

Idea

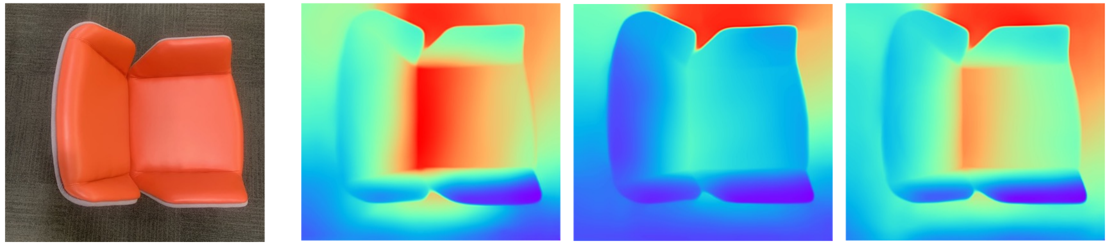
- Monocular depth is inherently ambiguous.

Albedo vs Shading



Possible depth maps

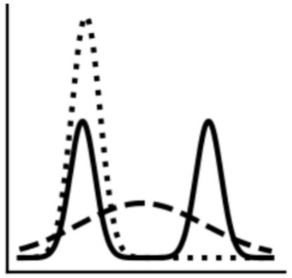
Scale / Degree of Convexity



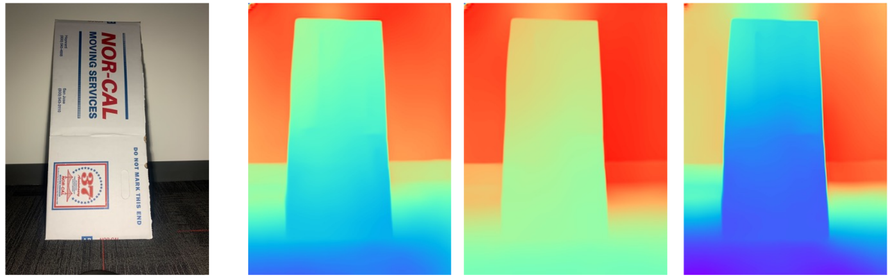
Possible depth maps

Idea

- Represent depth as a distribution.
  - Distribution can be **multimodal**.

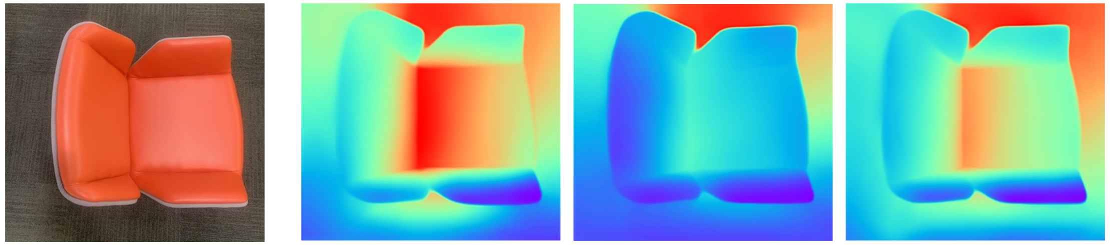


Albedo vs Shading



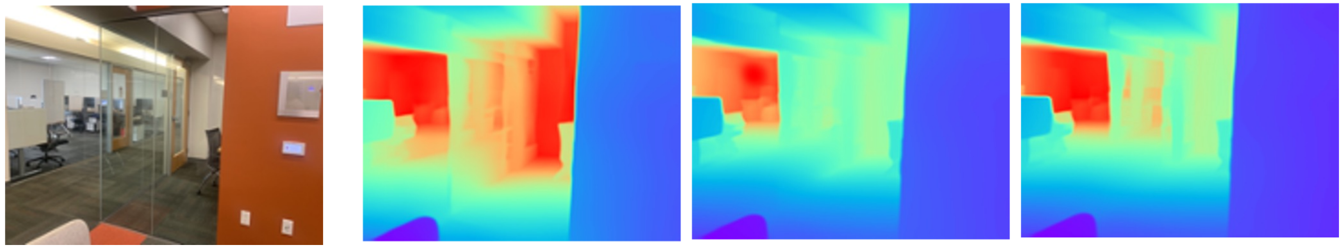
Possible depth maps

Scale / Degree of Convexity



Possible depth maps

Multimodal Example



Possible depth maps

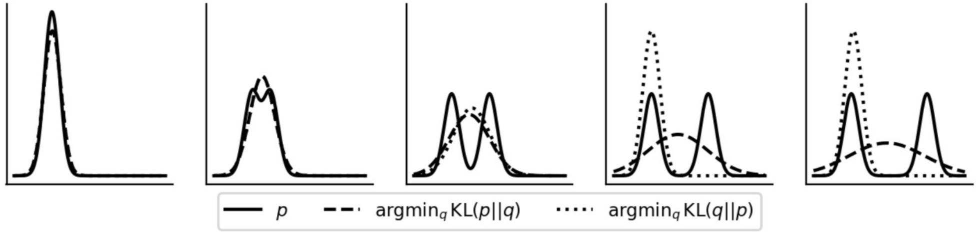
Idea

- Generalized space carving
  - Classical space carving only works with point estimates, i.e. no uncertainties.

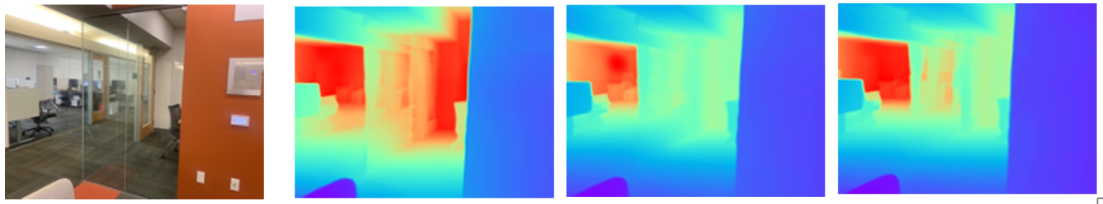
Idea

- Generalized space carving
  - Classical space carving only works with point estimates, i.e. no uncertainties.
  - Probabilistic analogue: Ambiguities are only **resolved** once information on **multiple views are fused together**.
  - Pick the mode that satisfies the different views.

- **Mode seeking** vs mean seeking:
  - Expected depth would fall to the mean of multimodal distributions. The mean is not necessarily a valid depth.
  - We instead want to find a consistent mode, which is valid.



Multimodal Example

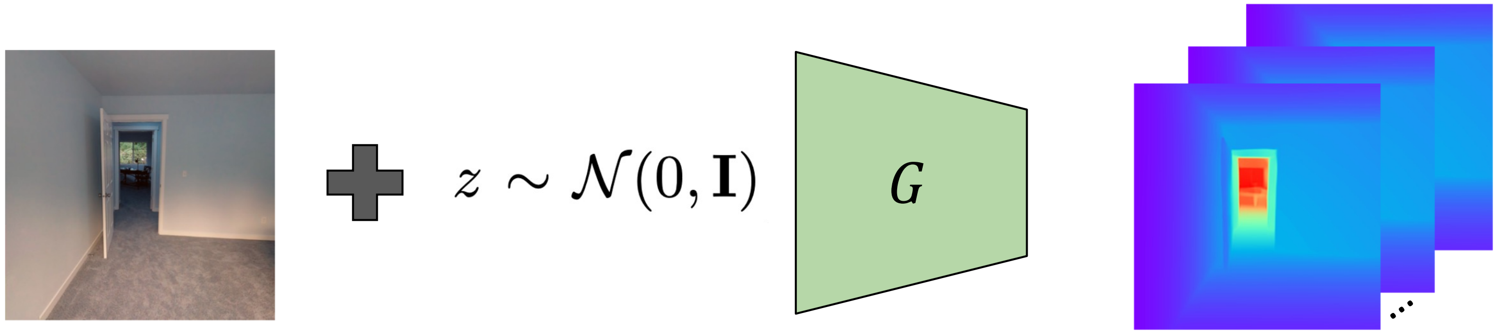


Possible depth maps

Our Ambiguity-Aware Prior

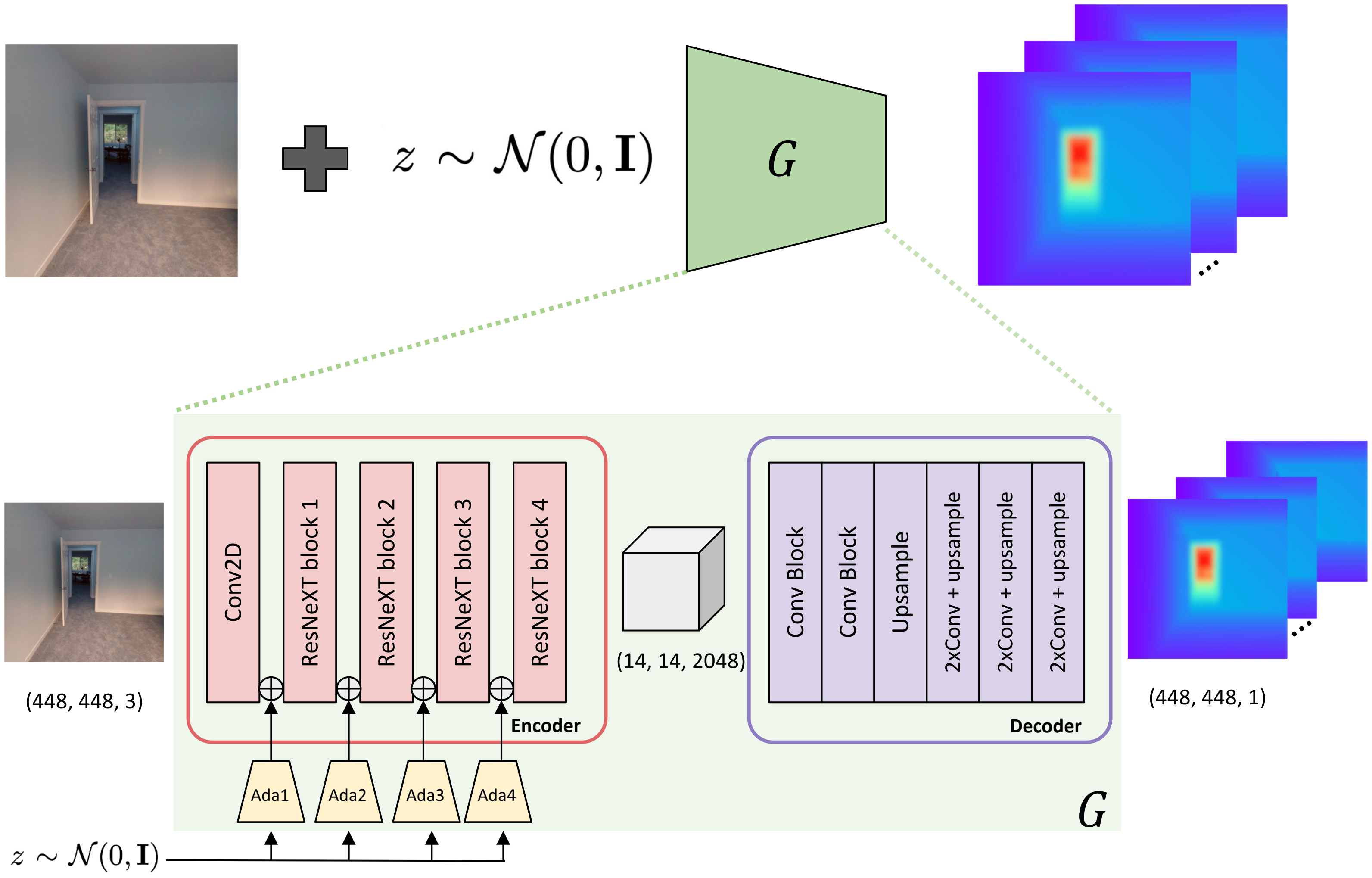
- Our prior represents depth as a distribution, to handle ambiguity.
  - This distribution can be **multimodal**.

- Represent ambiguities and capture variable modes through **samples** via conditional Implicit Maximum Likelihood Estimation (cIMLE).



[2] Multimodal Image Synthesis with Conditional Implicit Maximum Likelihood Estimation. K. Li, et. al., IJCV 2020.

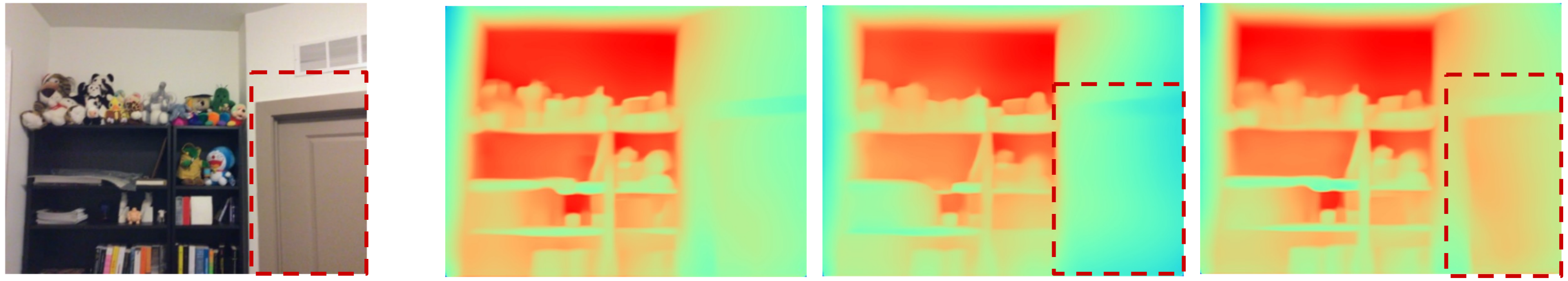
Our Ambiguity-Aware Prior



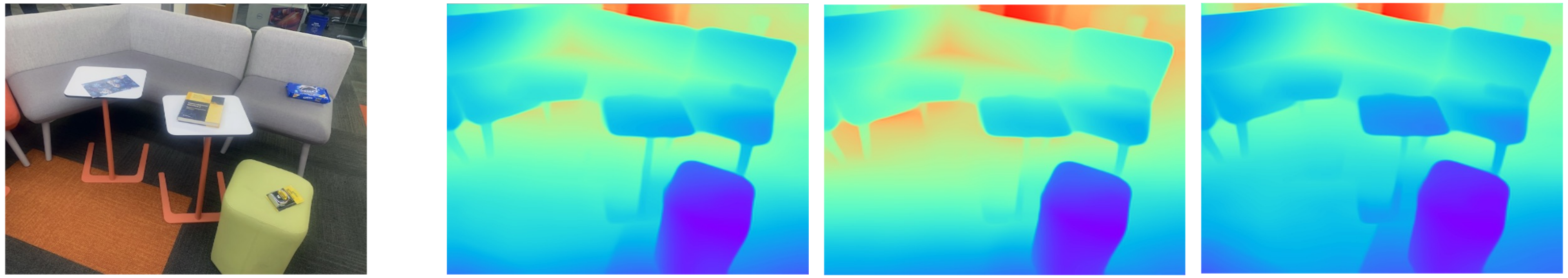
[2] Multimodal Image Synthesis with Conditional Implicit Maximum Likelihood Estimation. K. Li, et. al., IJCV 2020.
[4] Learning to Recover 3D shape from a Single Image. W. Yin, et. al., CVPR 2021.

Our Ambiguity-Aware Depth Estimates

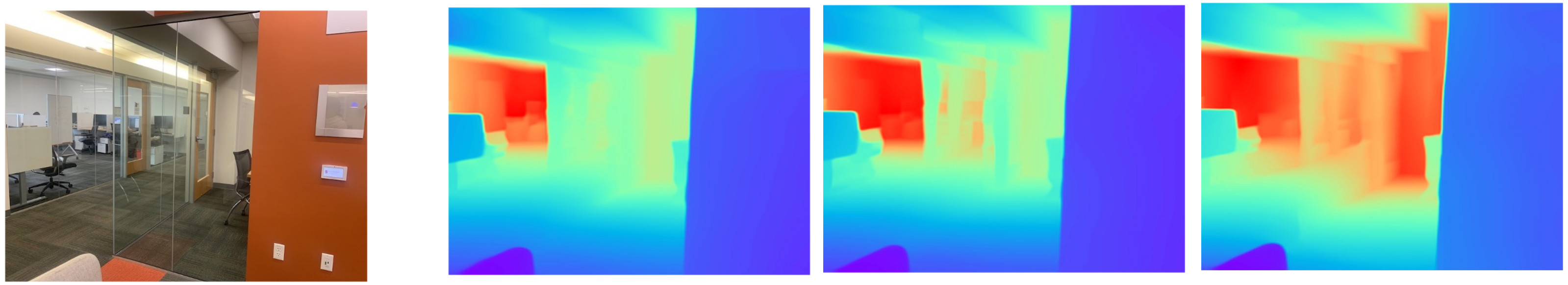
Albedo
vs
Shading



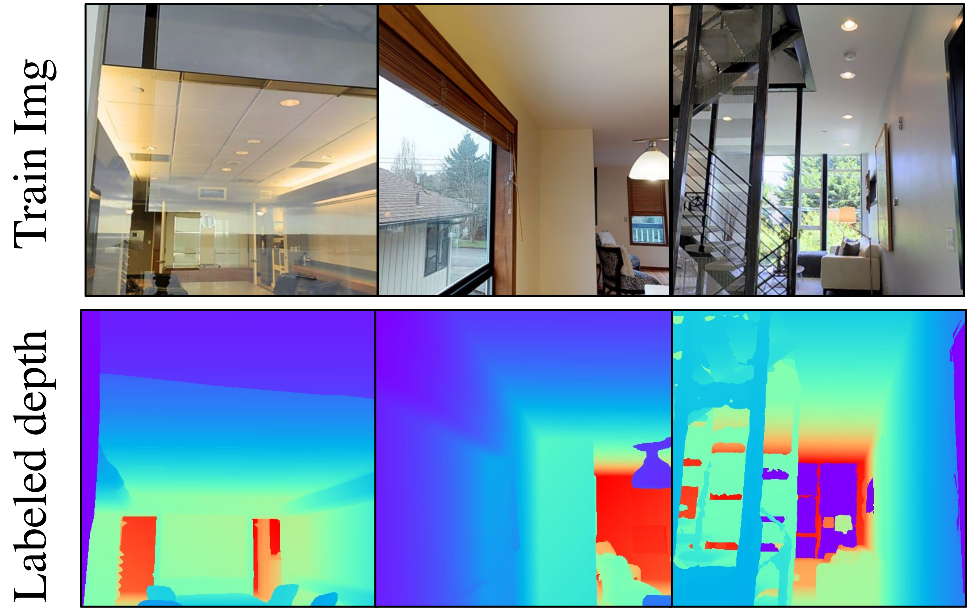
Scale /
Degree of
Concavity

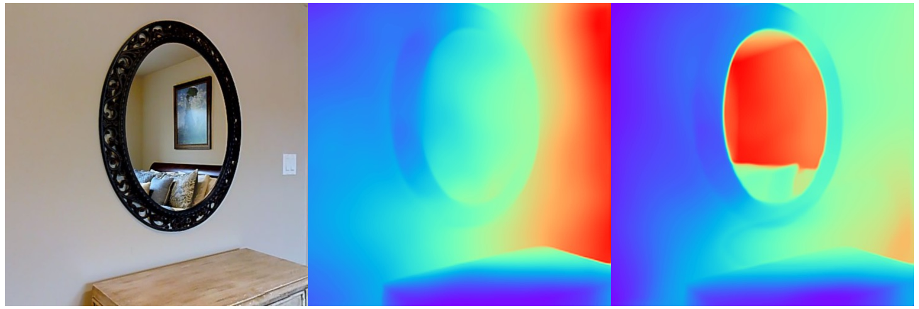


Non-opaque
surfaces



Why does it work?





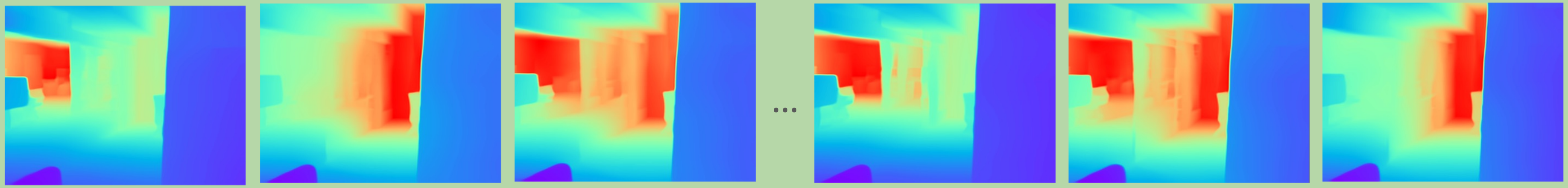
Test Img

Samples from our ambiguity-aware prior

SCADE

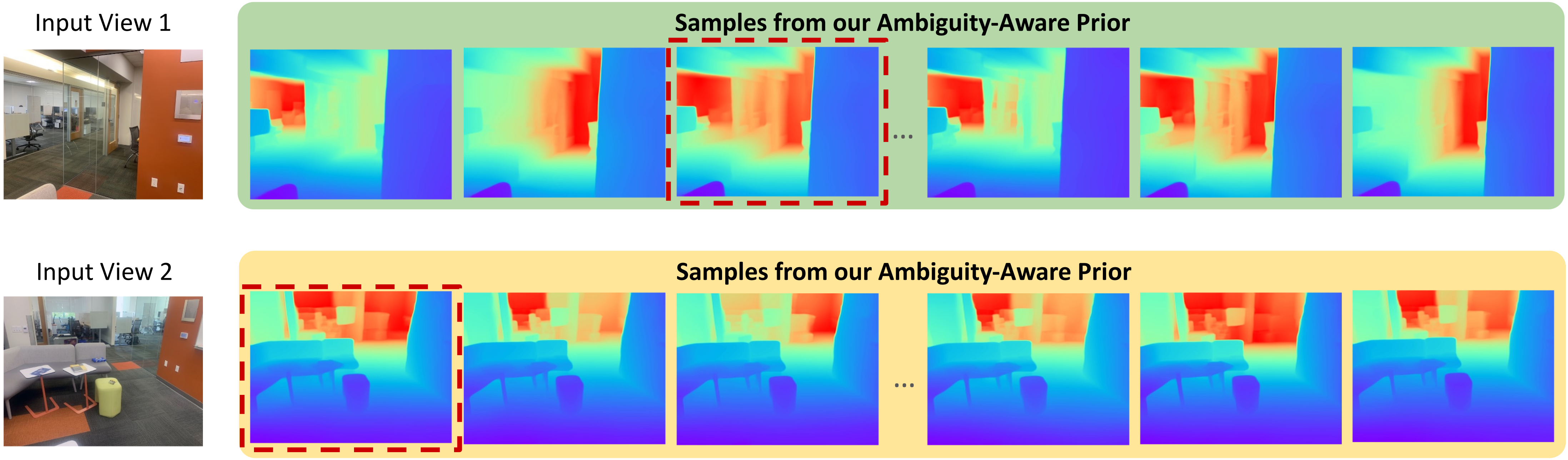
Input View 1

Samples from our Ambiguity-Aware Prior



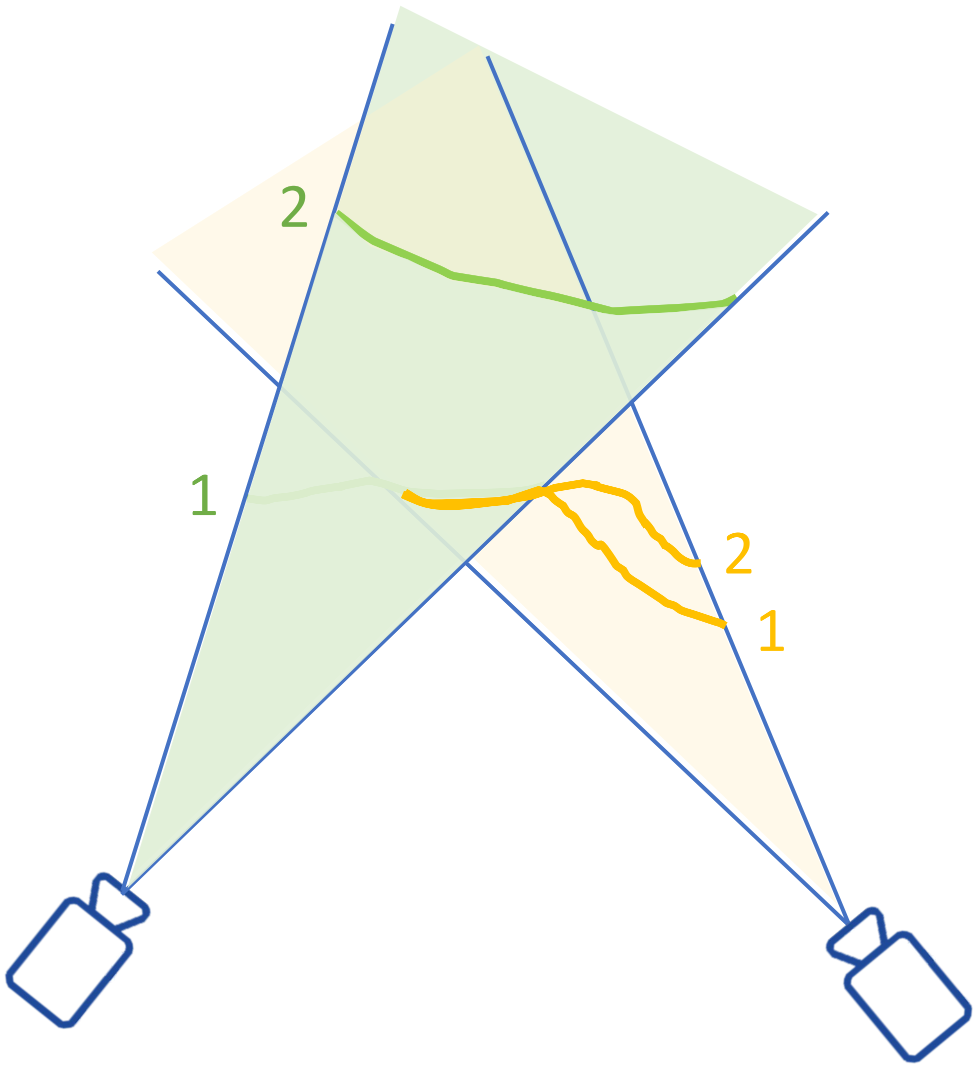
Ambiguous!

SCADE



- Resolve ambiguities by **fusing** together information from multiple views.

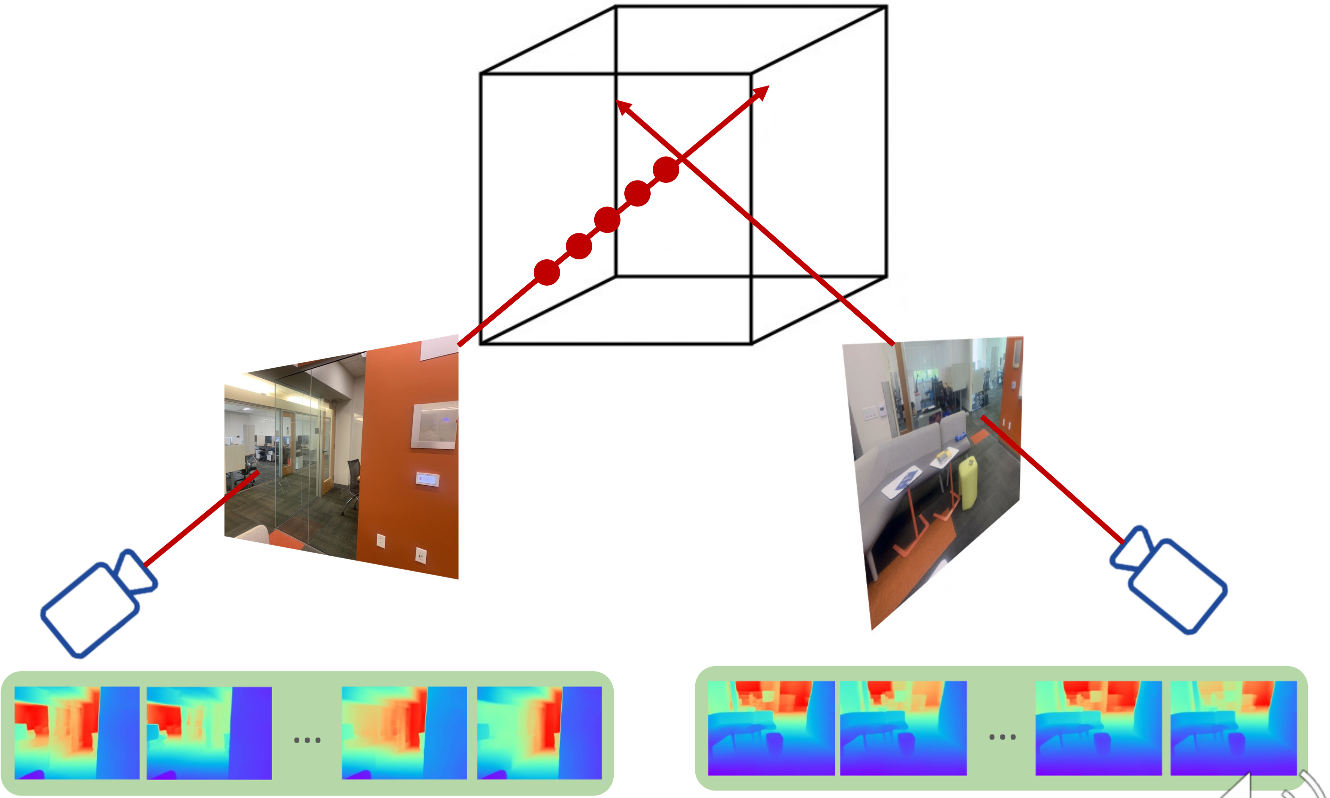
Space Carving Intuition



SCADE

- We **distill** the **consistent hypotheses** for each view into a global 3D geometry represented with a **NeRF**.

- We introduce our novel **space carving loss** on the two distributions:
    1. Ambiguity-aware prior
    2. Ray termination distance from NeRF

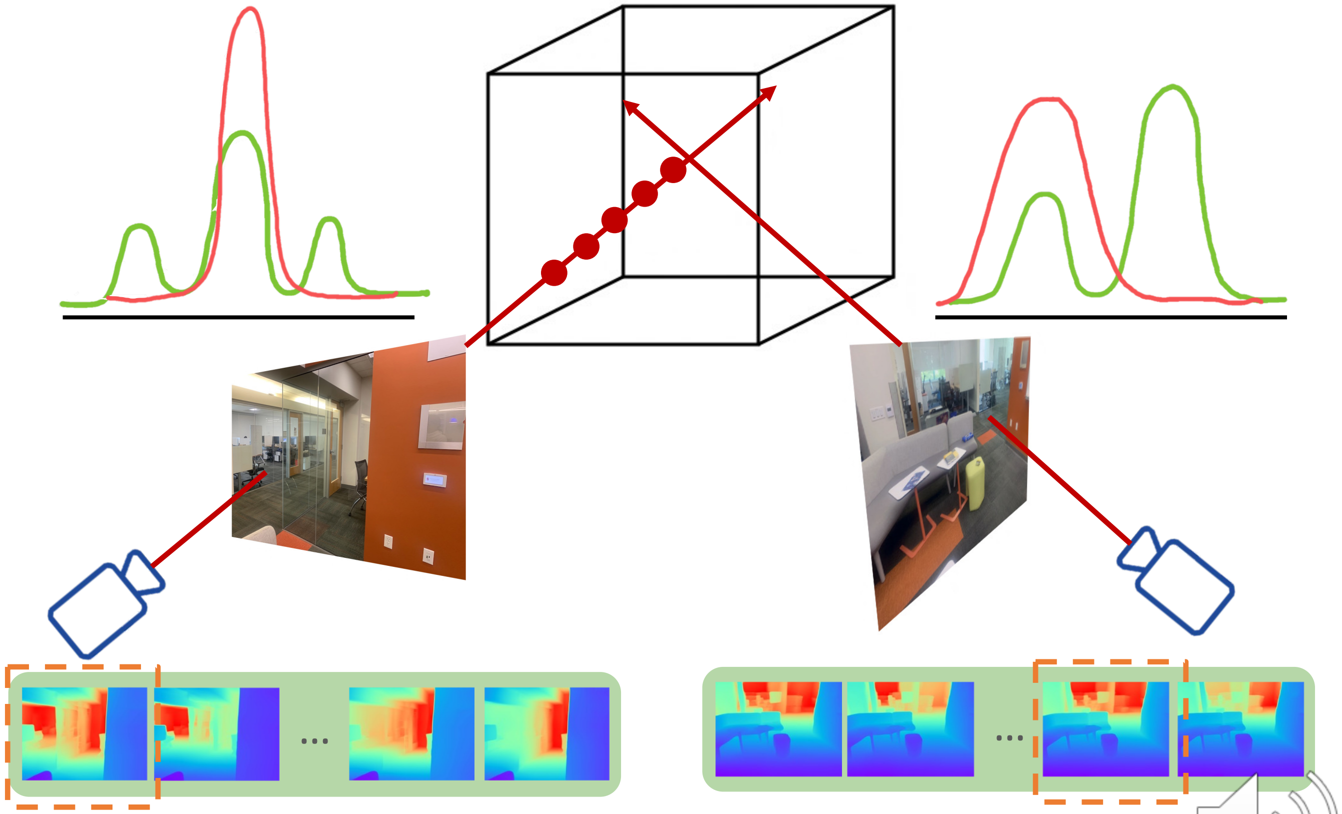


SCADE

Our Space Carving Loss

- The learned depth distribution should be **consistent** with **some** depth hypothesis in **every** view.

$$\mathcal{L}_{\text{space\_carving}}(\mathbf{r}) = \sum_{i \in [N]} \min_{j \in [M]} \| x_i - y_j \|_2^2$$



- **Mode seeking** : finds the consistent agreement across views.
- **Sample-based loss** on the distribution *instead of moments* leads to supervision in 3D instead of 2D.

Results – In-the-Wild Demo

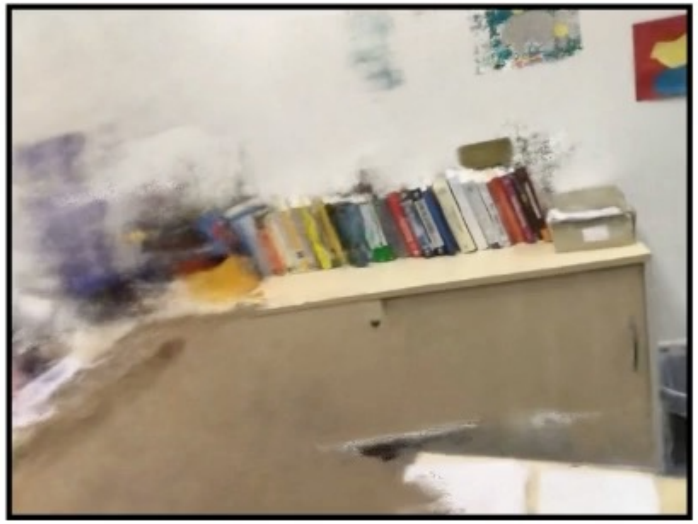
Vanilla NeRF

DDP

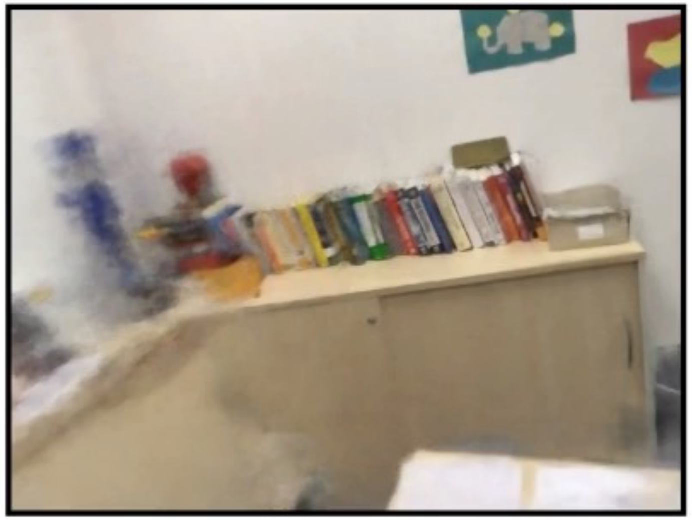
SCADE (Ours)

Results – Scannet Demo

Vanilla NeRF



DDP (out-domain)



DDP (in-domain)

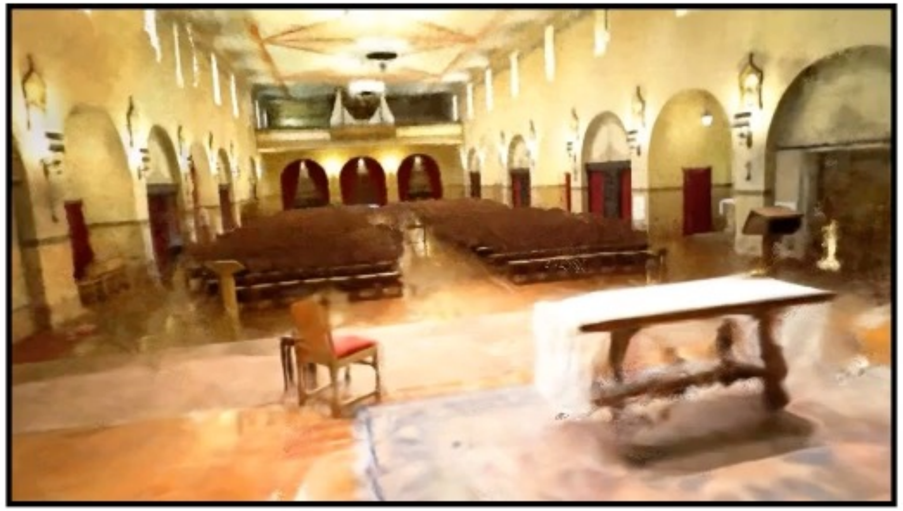
SCADE (Ours)

Results – Tanks and Temples Demo

Vanilla NeRF

DDP

SCADE (Ours)



Results

	PSNR ↑	SSIM ↑	LPIPS ↓
Vanilla NeRF [24]	19.03	0.670	0.398
NerfingMVS [47]	16.29	0.626	0.502
IBRNet [41]	13.25	0.529	0.673
MVSNeRF [3]	15.67	0.533	0.635
DS-NeRF [6]	20.85	0.713	0.344
DDP [32]	19.29	0.695	0.368
SCADE (Ours)	**21.54**	**0.732**	**0.292**

Table 1. **ScanNet Results**. Results for DS-NeRF and NerfingMVS follow what was reported in prior literature [32]. Because our setting requires out-of-domain priors, the results for DDP are with out-of-domain priors. The results of DDP with in-domain priors are (20.96, 0.737, 0.236) for PSNR, SSIM and LPIPS, respectively.

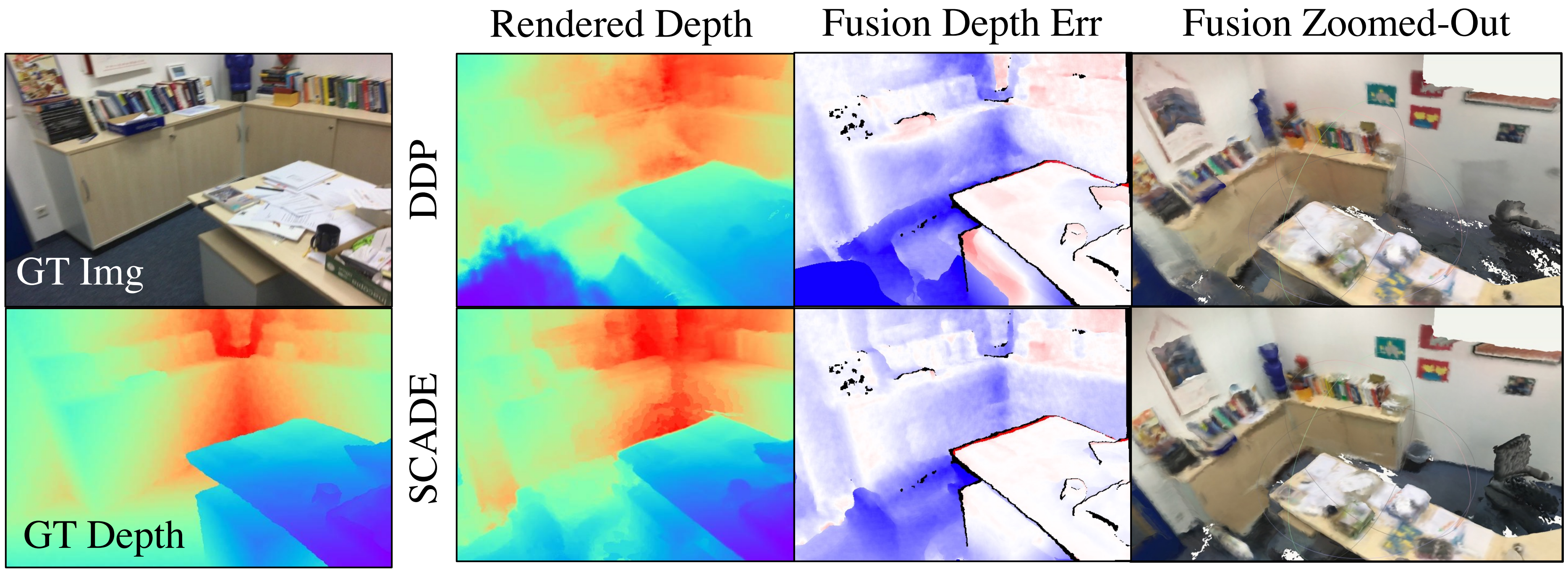
	PSNR ↑	SSIM ↑	LPIPS ↓
Vanilla NeRF [25]	19.09	0.700	0.437
DDP [33]	19.84	0.727	0.382
SCADE	**21.48**	**0.736**	**0.356**

Table 2. **In-the-wild Results**.

	PSNR ↑	SSIM ↑	LPIPS ↓
Vanilla NeRF [6]	17.19	0.559	0.457
DDP [7]	18.23	0.631	0.377
SCADE	**20.32**	**0.663**	**0.348**

Table 1. **Quantitative results for the Tanks and Temples [3] dataset.**

Results


Rendered Depth
Fusion Depth Err
Fusion Zoomed-Out
GT Img
DDP
GT Depth
SCADE

Thank you!

Visit our project page!

Poster: THU-AM-004