CVPR
SEATTLE, WA JUNE 17-21, 2024

GIST

Depth Prompting for Sensor-Agnostic Depth Estimation

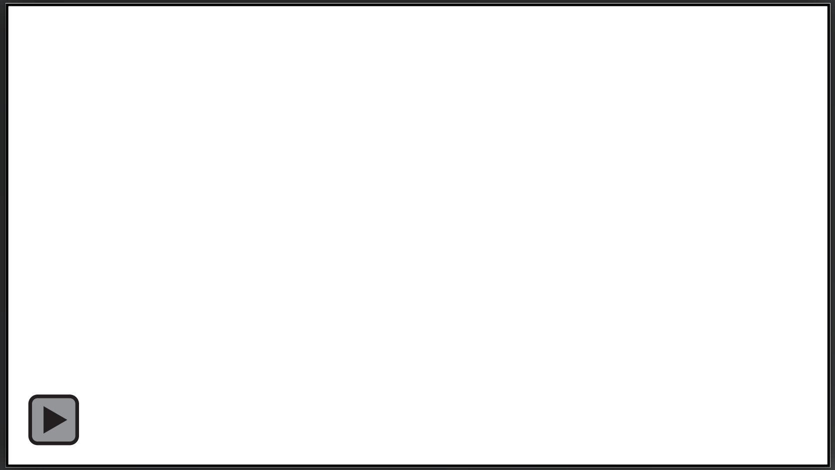
Jin-Hwi Park, Chanhwi Jeong, Junoh Lee and Hae-Gon Jeon

Computer Vision and Pattern Recognition (CVPR) 2024

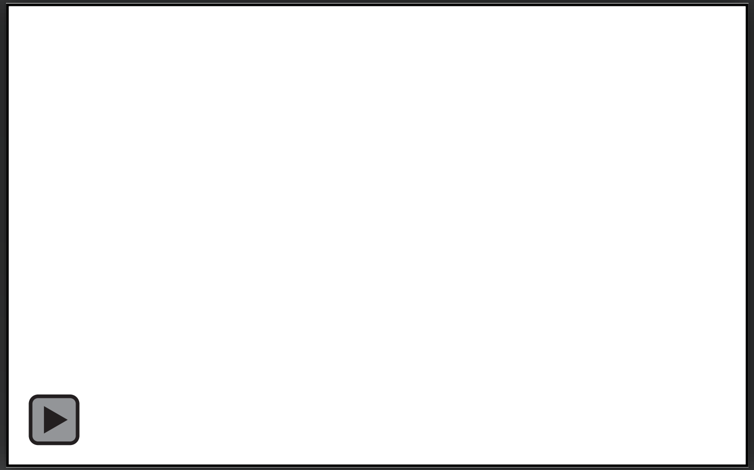
Depth from Active Sensors

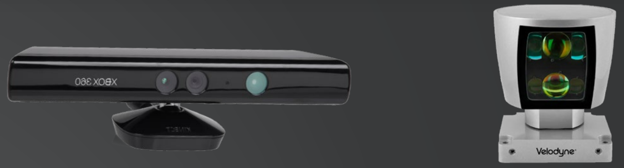
Depth from Active Sensors

Depth Estimation with Sparse Measurement



Apple iPhone & iPad





Microsoft Kinect & Velodyne LiDAR

*video source: developer.apple.com/videos/play/wwdc2020

Depth from Active Sensors

Diverse Type of Depth Sensors

*Public Dataset

Kinect V1
(Structured Light)

Kinect V2
(ToF)

iPad Pro
(LiDAR)

ZED
(Stereo Camera)

*Public Dataset

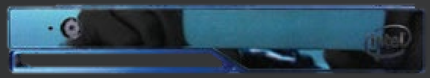
Intel RealSense Series

Velodyne LiDAR Series

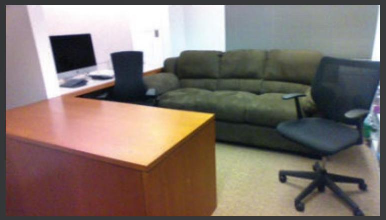
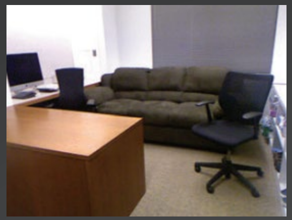
*Public Dataset

Depth from Active Sensors

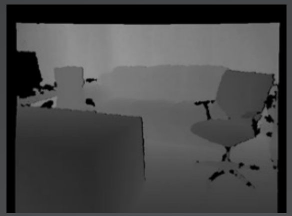
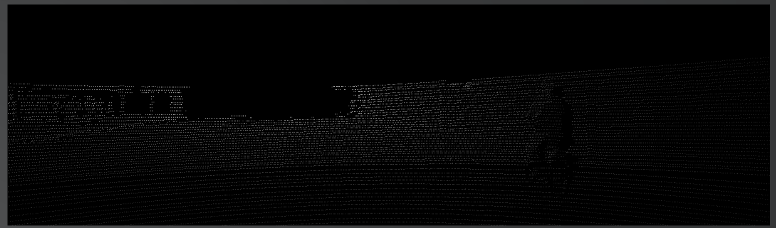
Diverse Type of Depth Sensors



Intel Realsense | Kinect v1 | Velodyne HDL-64E LiDAR

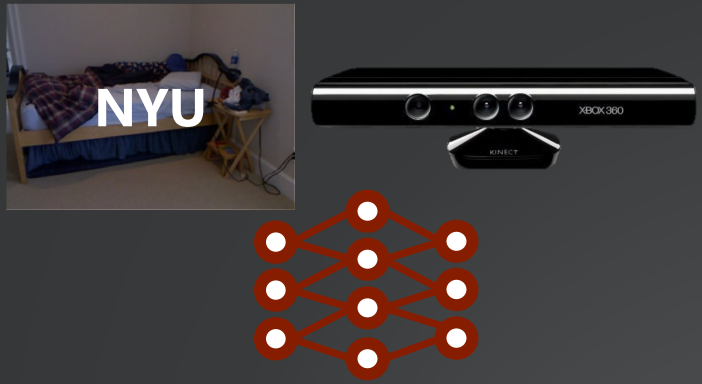


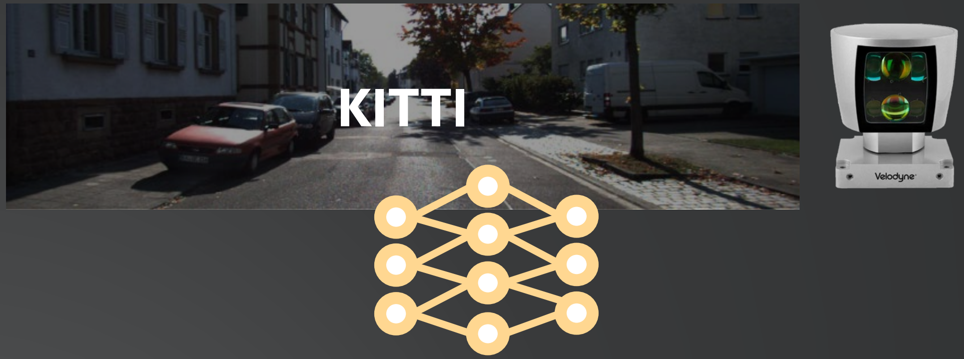



"Each sensor have different density, sensing pattern, and scan range."

Problem Definition

- Due to the challenges associated with acquiring high-resolution depth information, only two datasets, namely *NYU* and *KITTI*, have been used as standard benchmarks for more than 10 years.
  - [NYU, 2012]: indoor dataset - Kinect v1
  - [KITTI, 2012]: outdoor dataset - Velodyne LiDAR (64-Line)

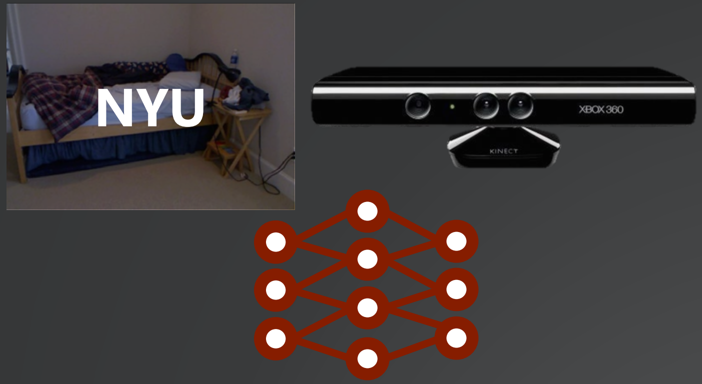


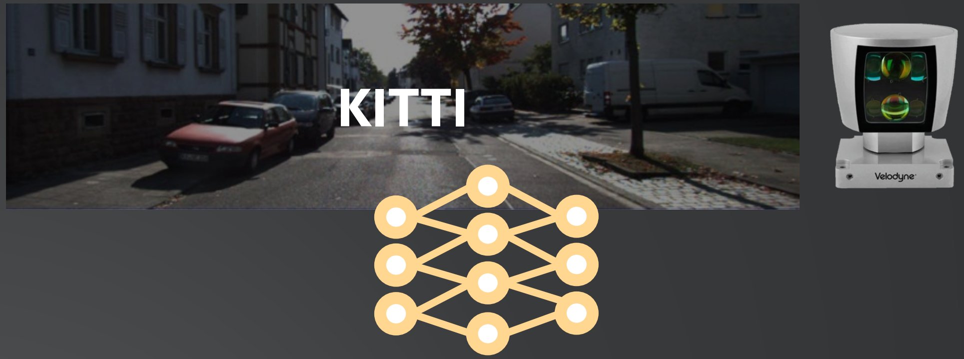


Silberman, et al., Are we ready for Autonomous Driving? Indoor Segmentation and Support Inference From RGBD Images. ECCV 2012
Geiger, et al., Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. CVPR 2012

Problem Definition

- Due to the challenges associated with acquiring high-resolution depth information, only two datasets, namely *NYU* and *KITTI*, have been used as standard benchmarks for more than 10 years.
  - [NYU, 2012]: indoor dataset - Kinect v1
  - [KITTI, 2012]: outdoor dataset - Velodyne LiDAR (64-Line)

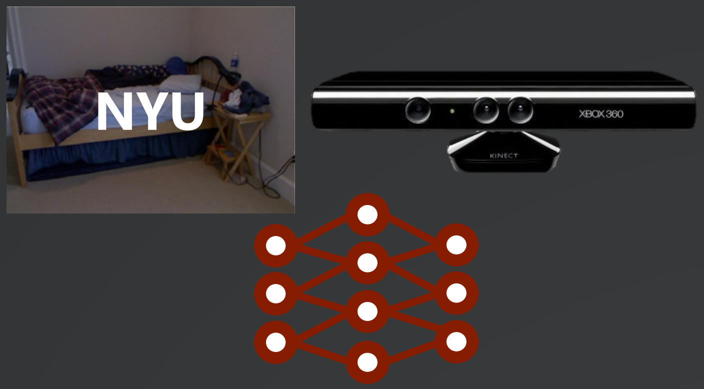


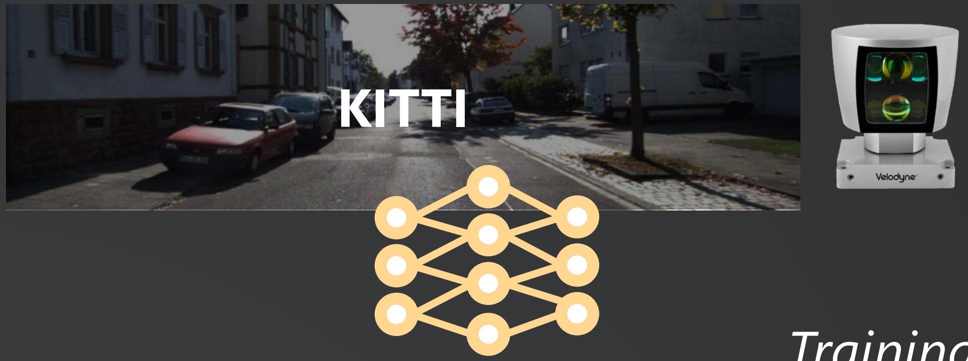


"A Model is biased toward to specific domain and sensor."

Silberman, et al., Are we ready for Autonomous Driving? Indoor Segmentation and Support Inference From RGBD Images. ECCV 2012
Geiger, et al., Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. CVPR 2012

Problem Definition



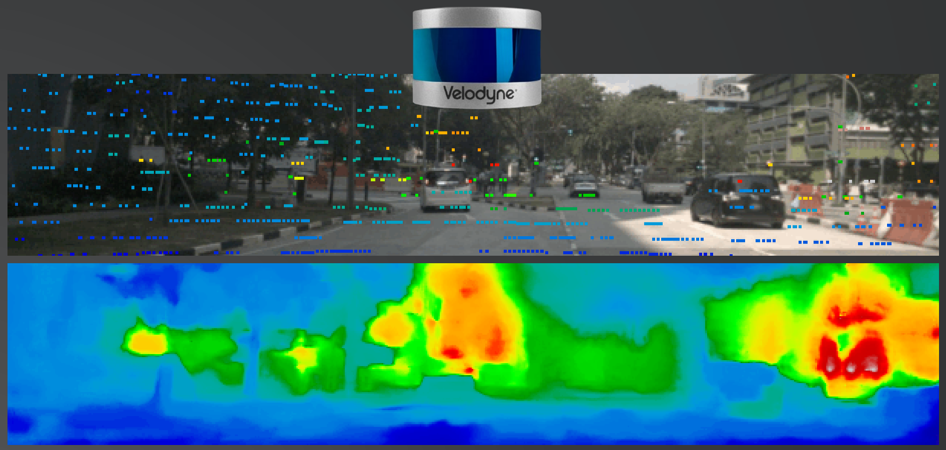


Training

Zero-shot Inference

Intel RealSense (2014) | Asus Xtion (2014) | Kinect v2 (2013) | Apple iPad (2022)

32-Line LiDAR (2018)



*Inferred by the state-of-the-art CompletionFormer method (CVPR23).

"A Model is biased toward to specific domain and sensor."

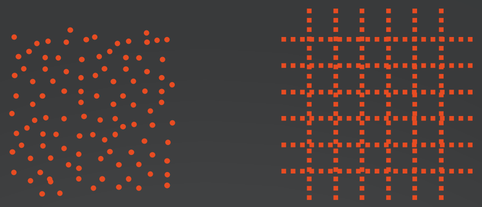
Introduction

Exploring Diverse Sensor Bias Problems

Train
(#500)

Test
(#50)

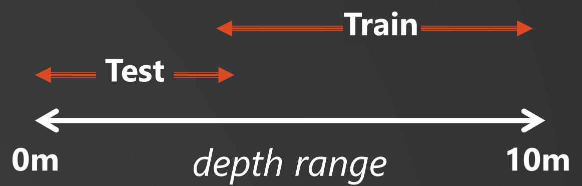
Sparsity Bias



Train
(Random)

Test
(Grid)

Pattern Bias



Train
(0m ~ 3m)

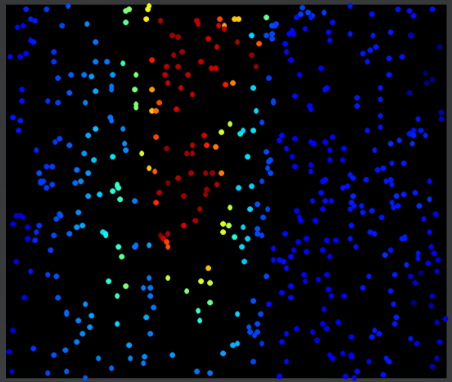
Test
(3m~10m)

Range Bias

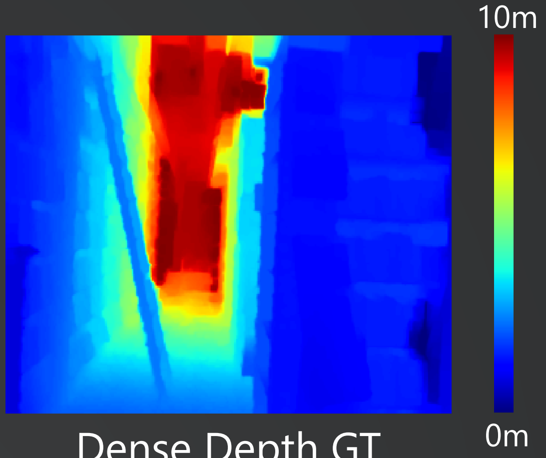
Introduction

Exploring Diverse Sensor Bias Problems

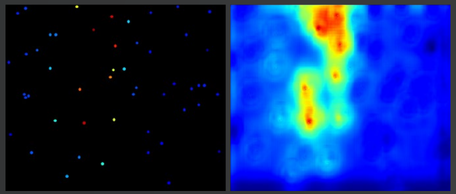
RGB Image



Sparse Depth

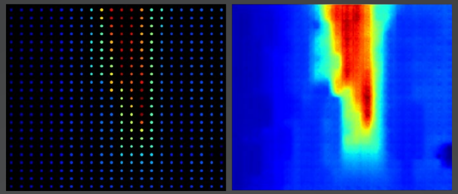


Dense Depth GT



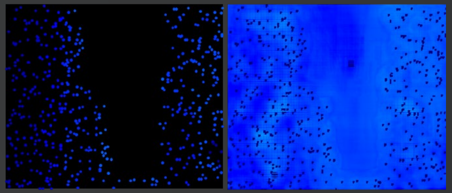
Train *(#500)* *Test* *(#50)*

Sparsity Bias



Train *(Random)* *Test* *(Grid)*

Pattern Bias



Train *(0m ~ 3m)* *Test* *(3m~10m)*

Range Bias

*Inferred by the state-of-the-art CompletionFormer method (CVPR23).

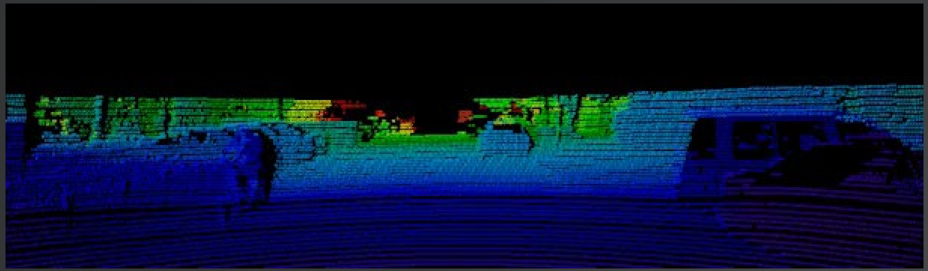
Introduction

Exploring Diverse Sensor Bias Problems

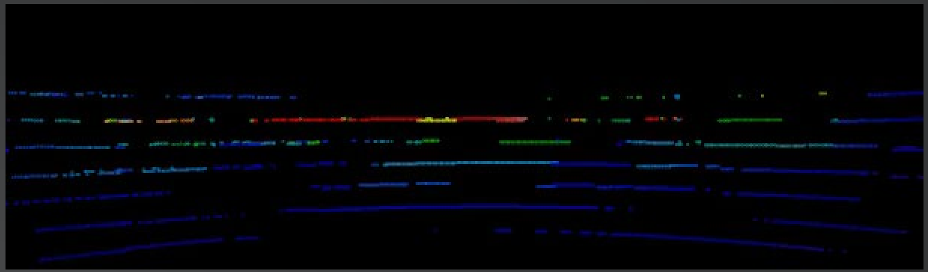
- Sparsity Bias (64-Line ↔ 8-Line)

RGB

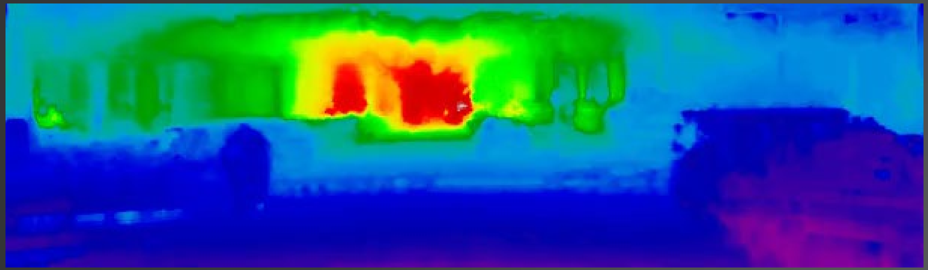
RGB



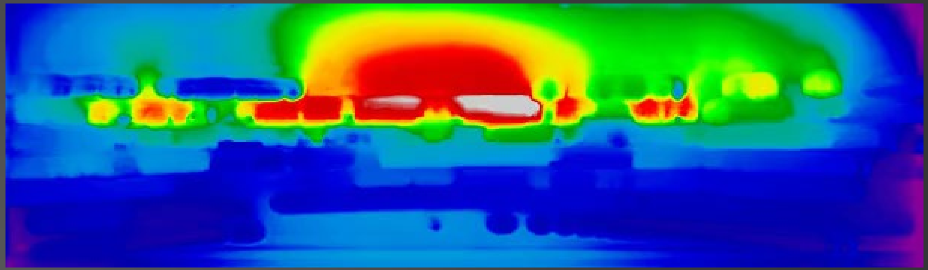
Sparse Depth [64-Line]



Sparse Depth [8-Line]



CompletionFormer (CVPR23)



CompletionFormer (CVPR23)

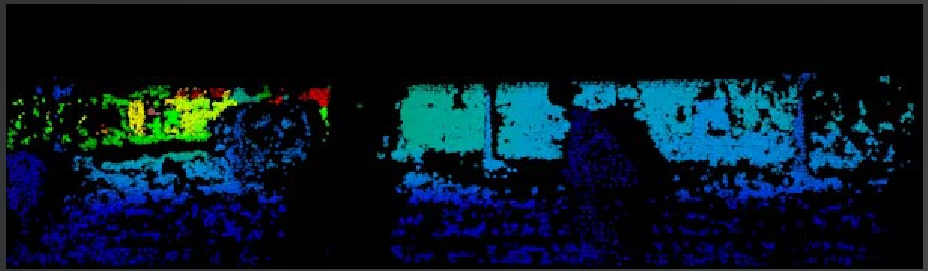
Introduction

Exploring Diverse Sensor Bias Problems

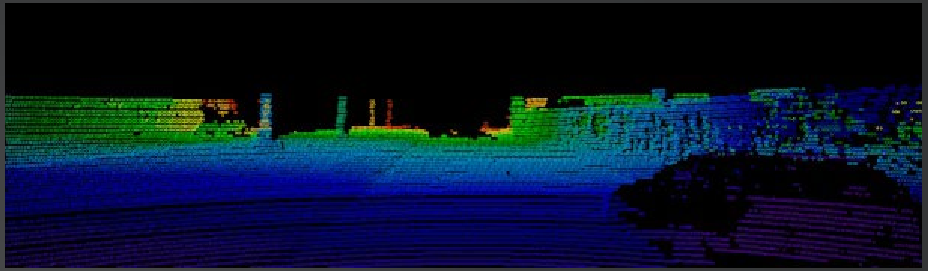
- Pattern Bias (Random ↔ Line)

RGB

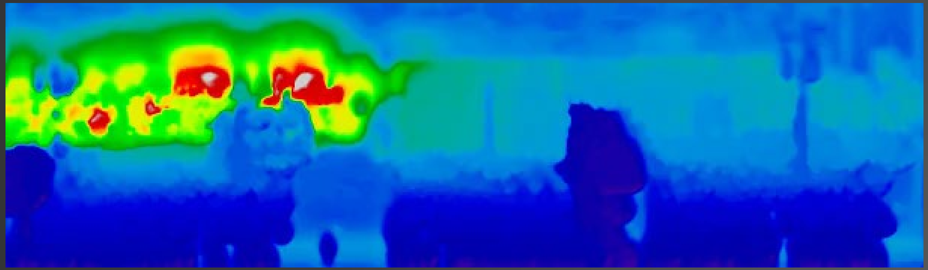
RGB



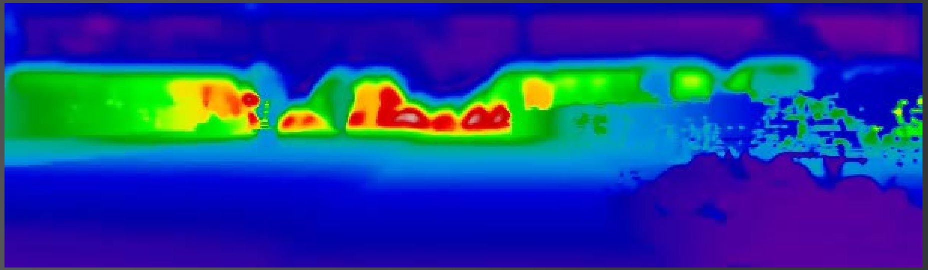
Sparse Depth [Random Sampling]



Sparse Depth [Line Sampling]



CompletionFormer (CVPR23)



CompletionFormer (CVPR23)

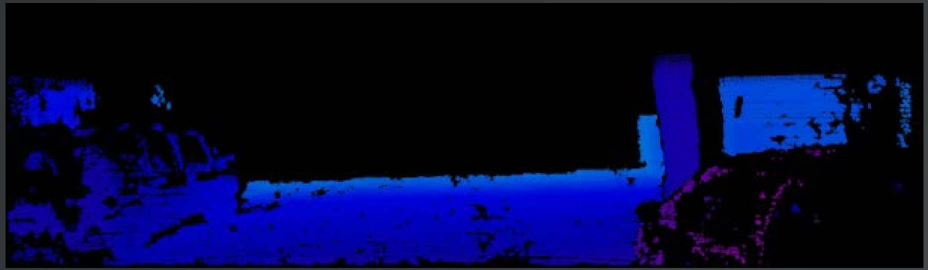
Introduction

Exploring Diverse Sensor Bias Problems

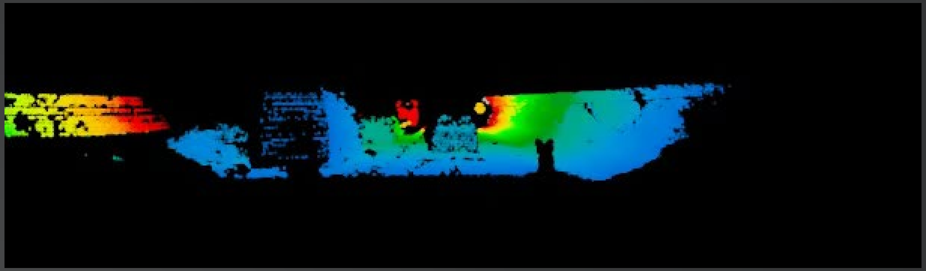
- Range Bias (under 15m ↔ over 15m)

RGB

RGB

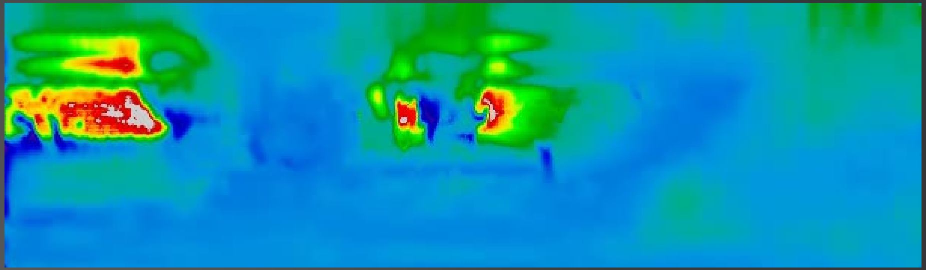


Sparse Depth [under 15m]



Sparse Depth [over 15m]

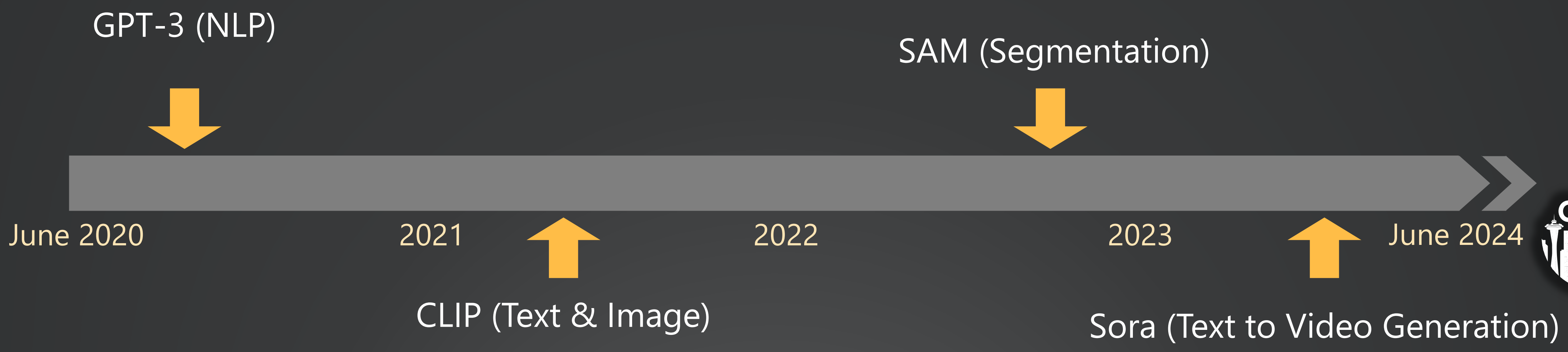
CompletionFormer (CVPR23)



CompletionFormer (CVPR23)

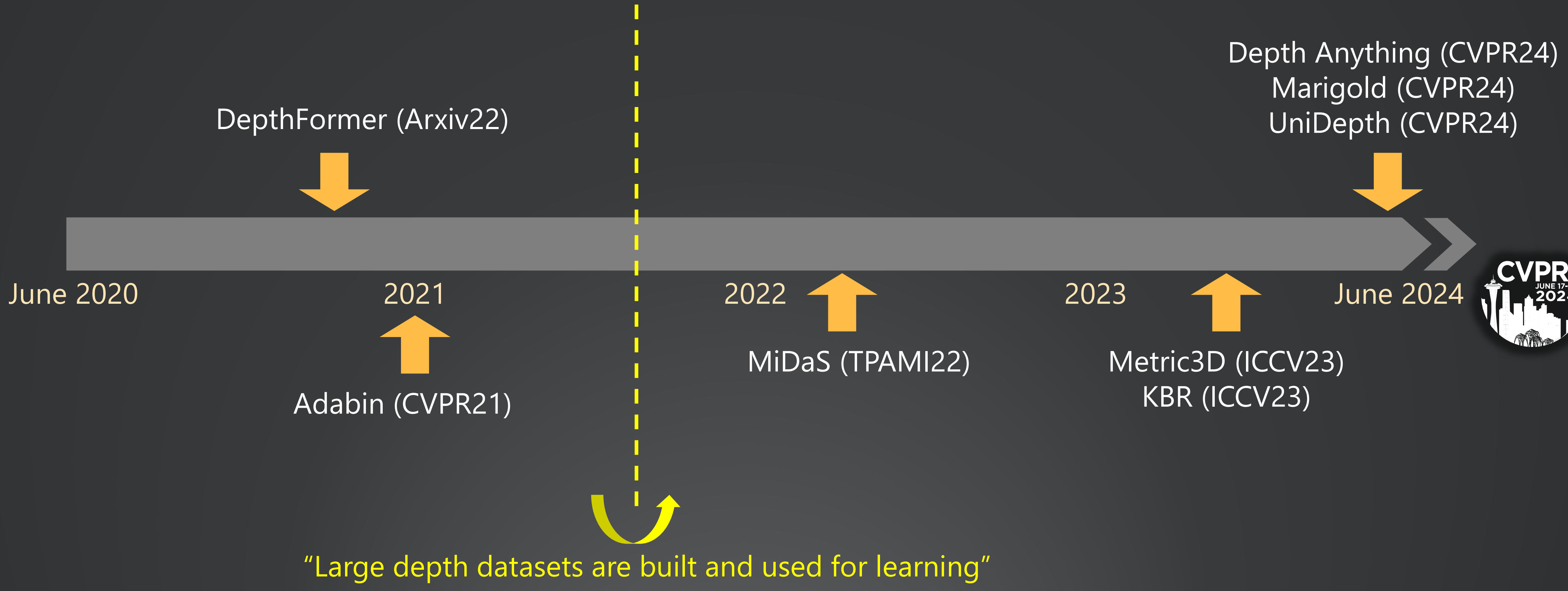
Motivation

The Era of Foundation Models



Motivation

Monocular Depth (Foundation) Models



Motivation

"Utilize the Depth (Foundation) Models to Mitigate Sensor Bias Problem."

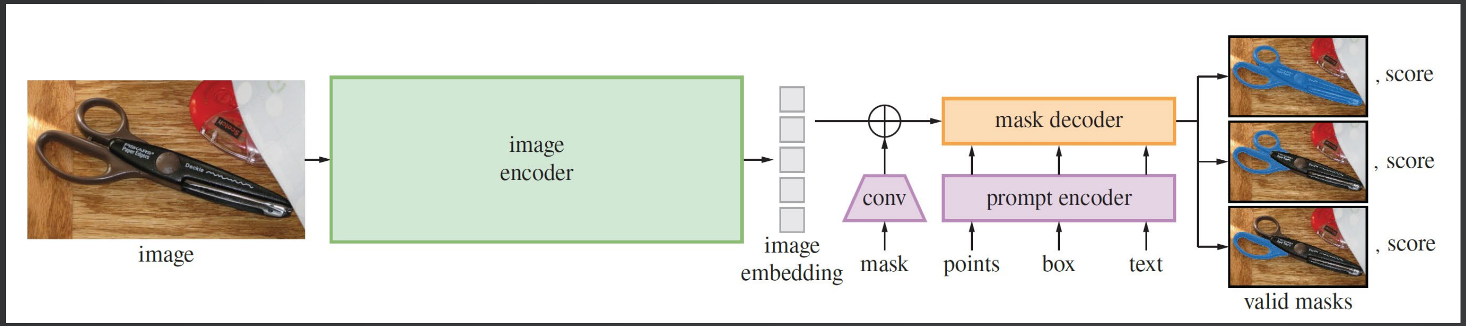
- ✓ Monocular depth models provides a comprehensive understanding of 3D structures.
- ✓ Our aim is to transfer the knowledge of depth model into our sensor agnostic model.

Baseline Models:

- DepthFormer (Arxiv22, in-domain monocular depth estimation)
- MiDaS (TPAMI22, out-domain monocular depth estimation)
- KBR (ICCV23, out-domain monocular depth estimation)

Motivation

Prompting Engineering Method



[SAM] Segment Anything (ICCV23, Best Paper Honorable Mention)

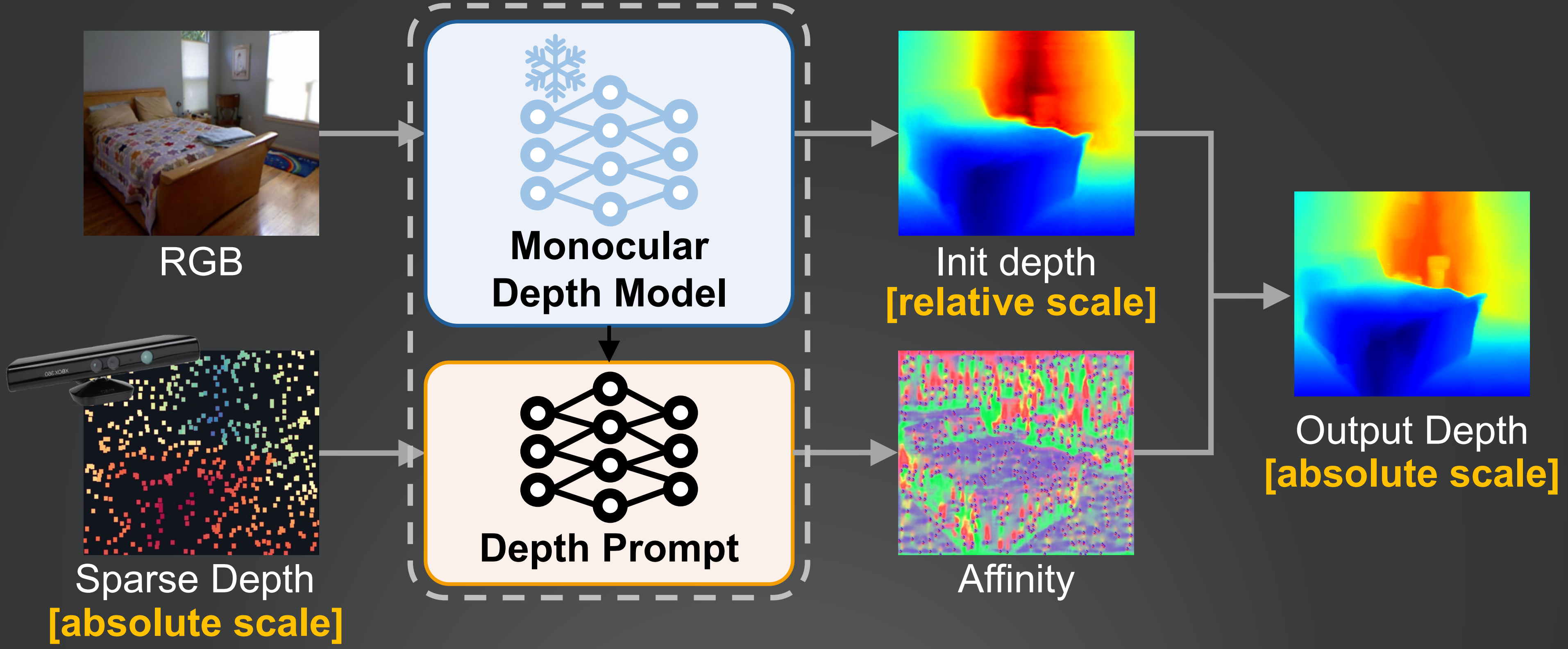
Demo

✓ Inspired by pioneering visual prompt work (SAM), we use prompting method in depth estimation to achieve adaptive output according to unseen sensor configuration.

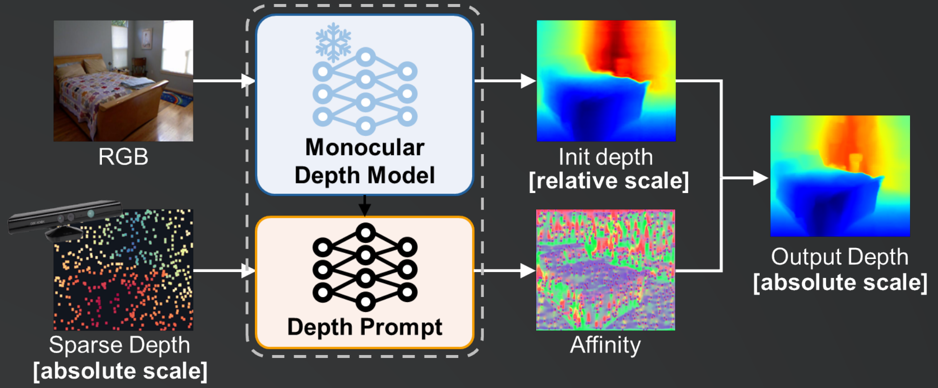
✓ Depth prompt with monocular depth models enables relative depth embedding freeing models from depth scan range limits and providing absolute scale depth maps.

Method

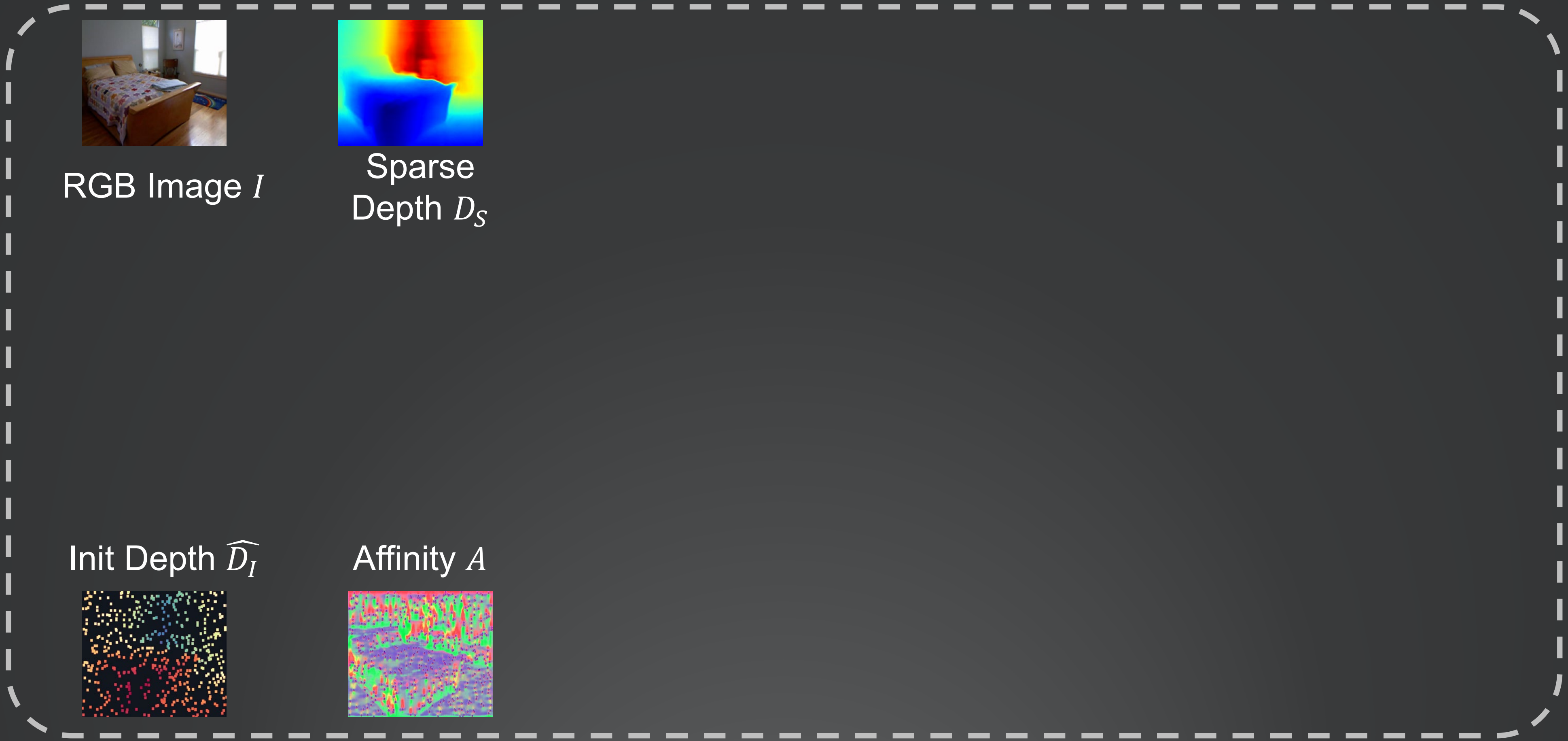
Monocular Depth Model & Prompting Engineering Method



Method

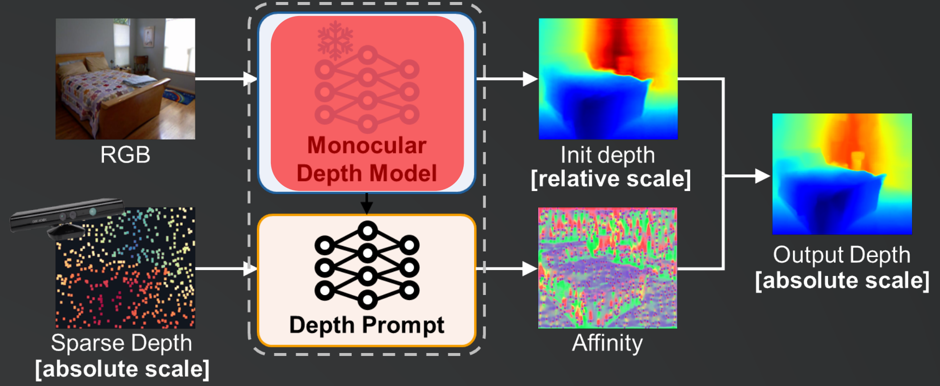


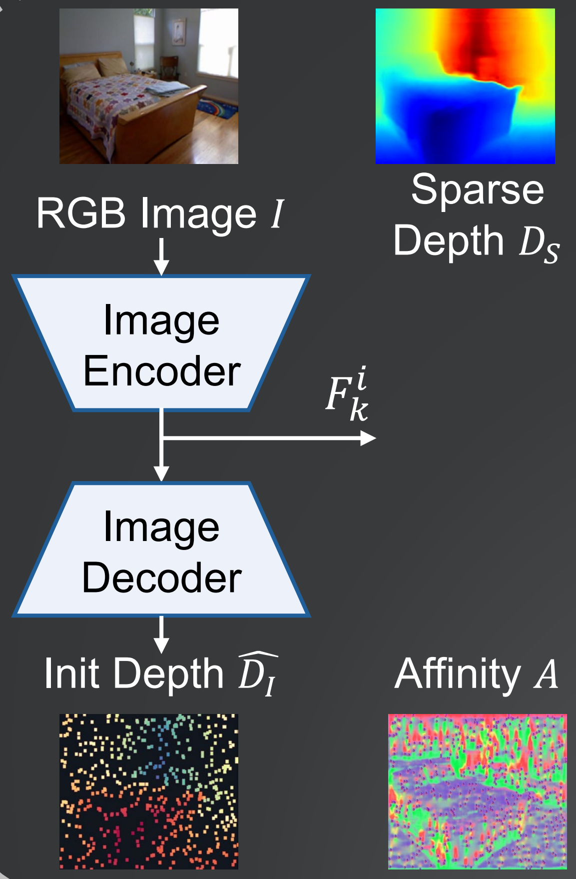
RGB
Monocular
Depth Model
Init depth
[relative scale]
Output Depth
[absolute scale]
Depth Prompt
Affinity
Sparse Depth
[absolute scale]



RGB Image I
Sparse Depth D_S
Init Depth $\widehat{D_I}$
Affinity A

Method



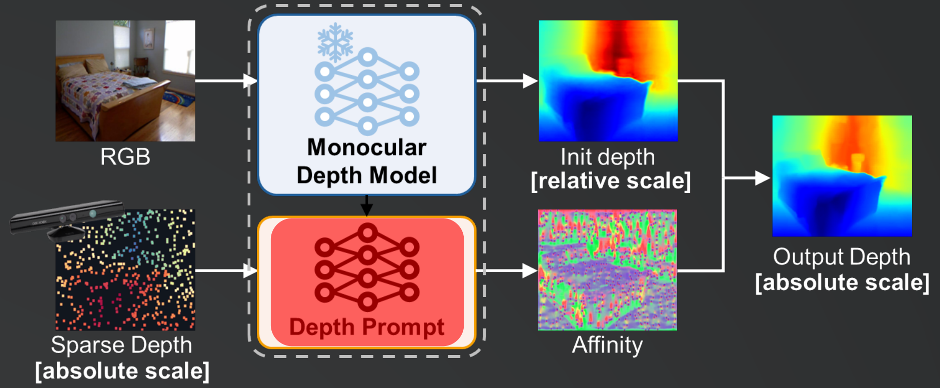


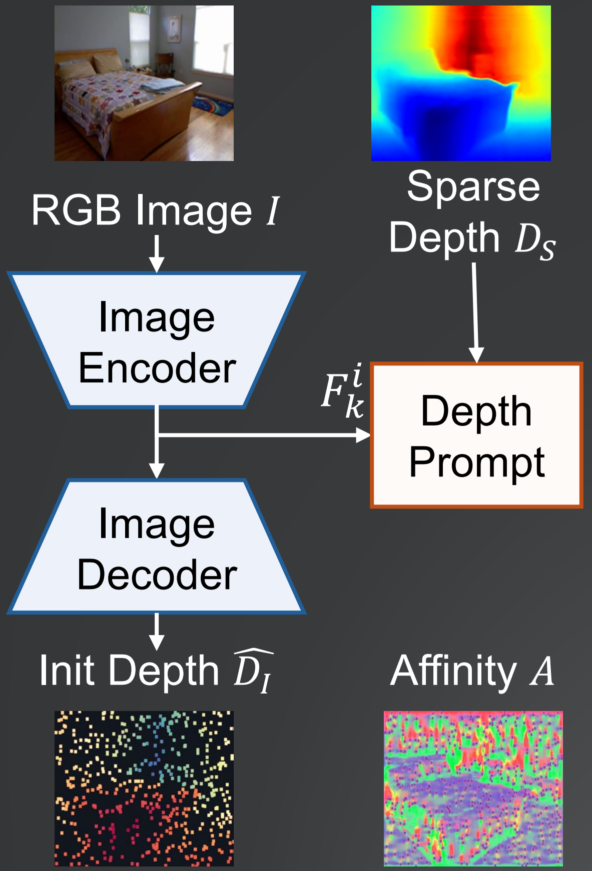
$$\widehat{D_I}, F_k^i = f_{\mathcal{F}}(I)$$

F_k^i: multi-scale image features

$f_{\mathcal{F}}$: monocular depth model

Method

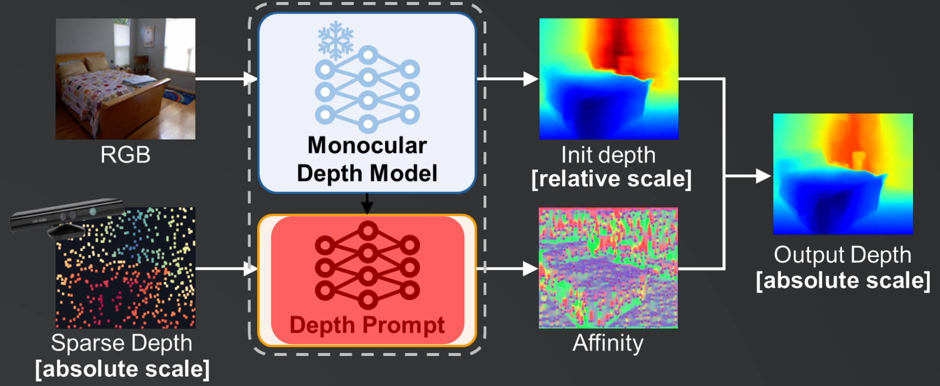


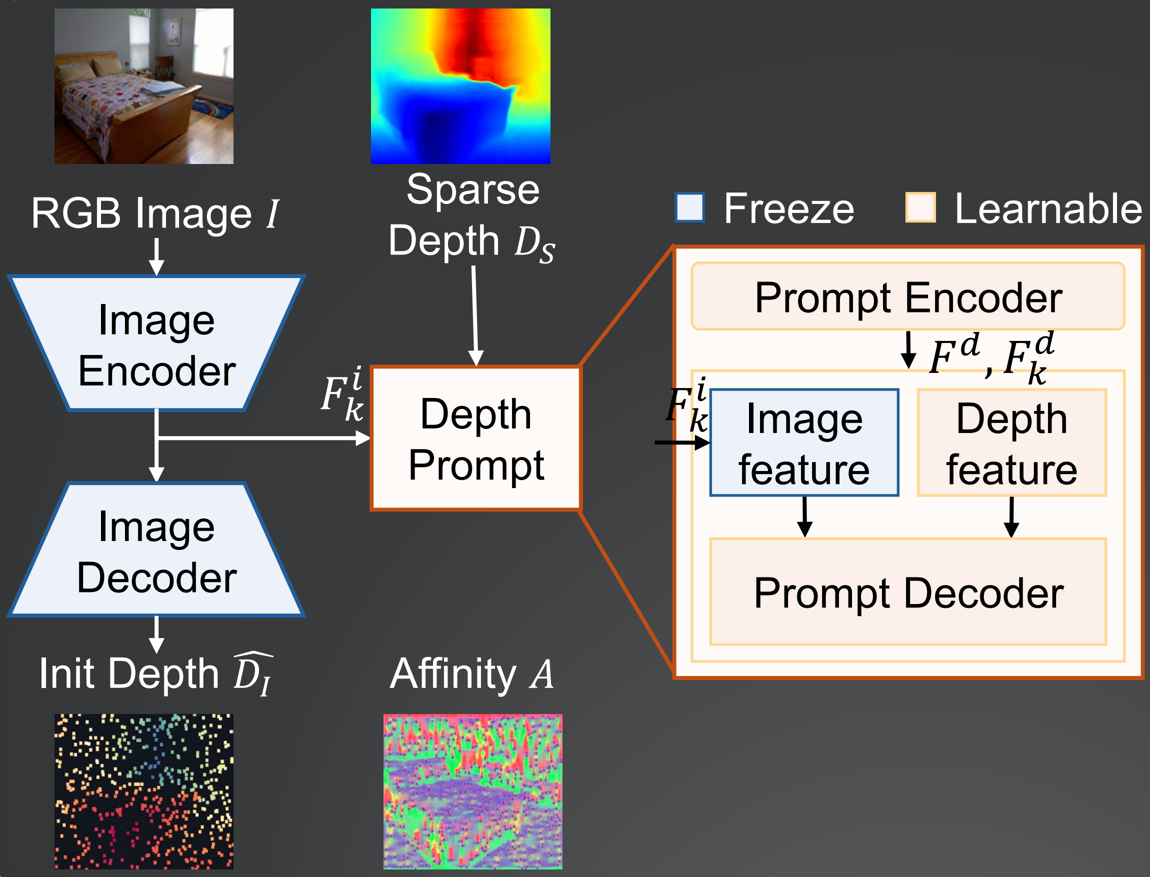


$$F^d, F_k^d = f_{\mathcal{E}}(D_S)$$

$f_{\mathcal{E}}$: depth prompt encoder
F^d: prompt embedding
F_k^d: multi-scale depth features

Method

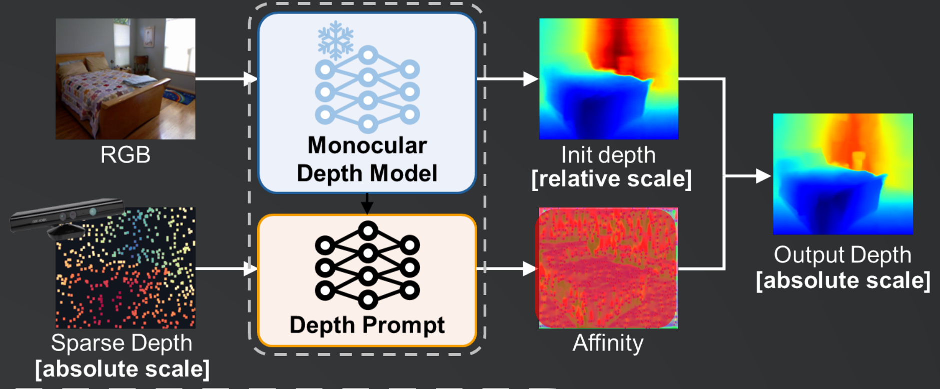


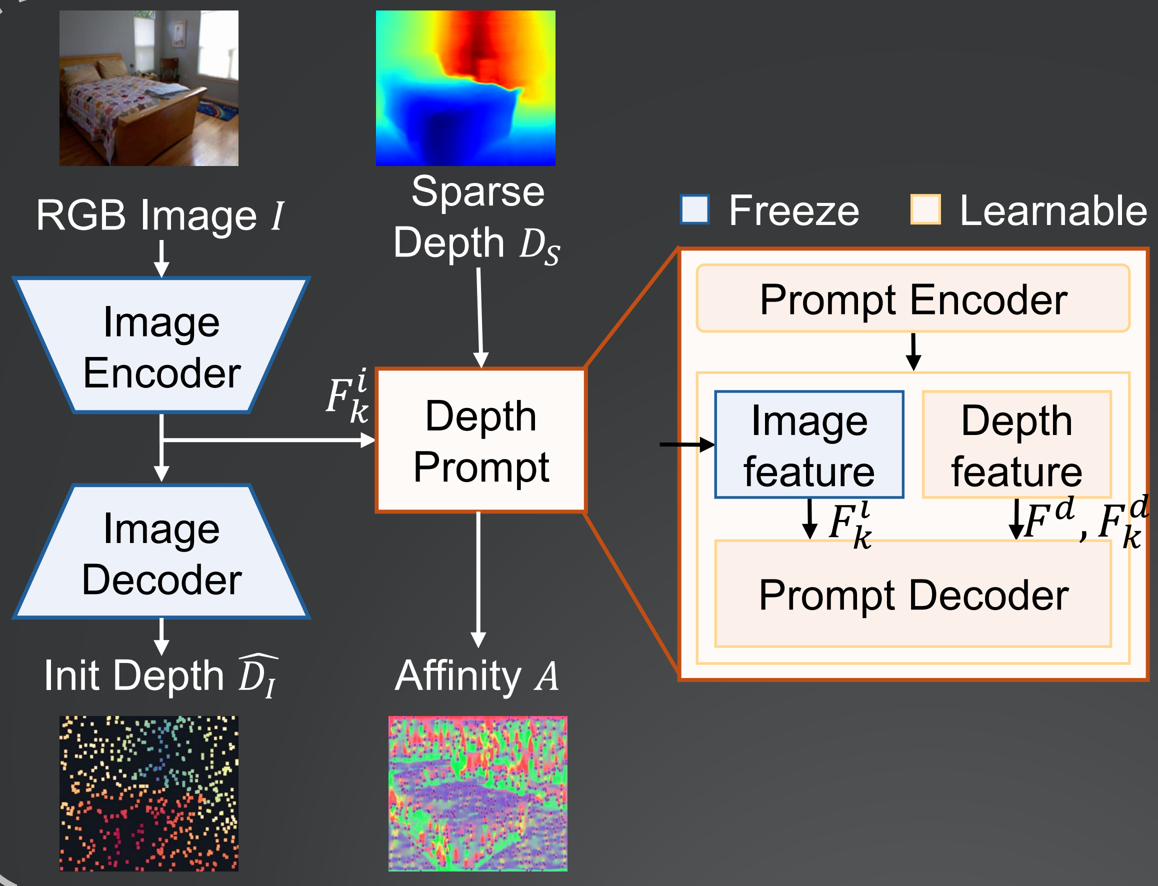


$$F^d, F_k^d = f_{\mathcal{E}}(D_S)$$

$f_{\mathcal{E}}$: depth prompt encoder
F^d: prompt embedding
F_k^d: multi-scale depth features

Method

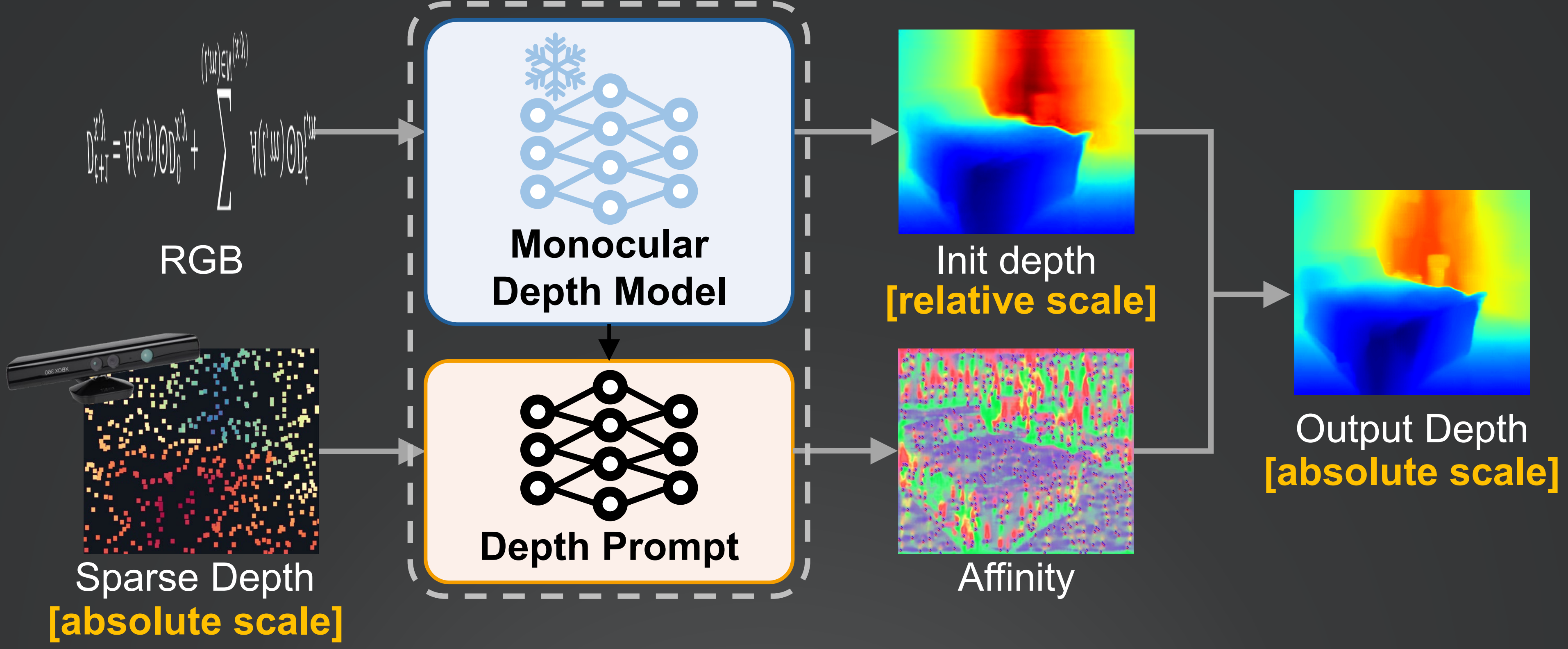




$$A = f_D(F^d, F_k^d, F_k^i)$$

f_D: depth prompt decoder

Method



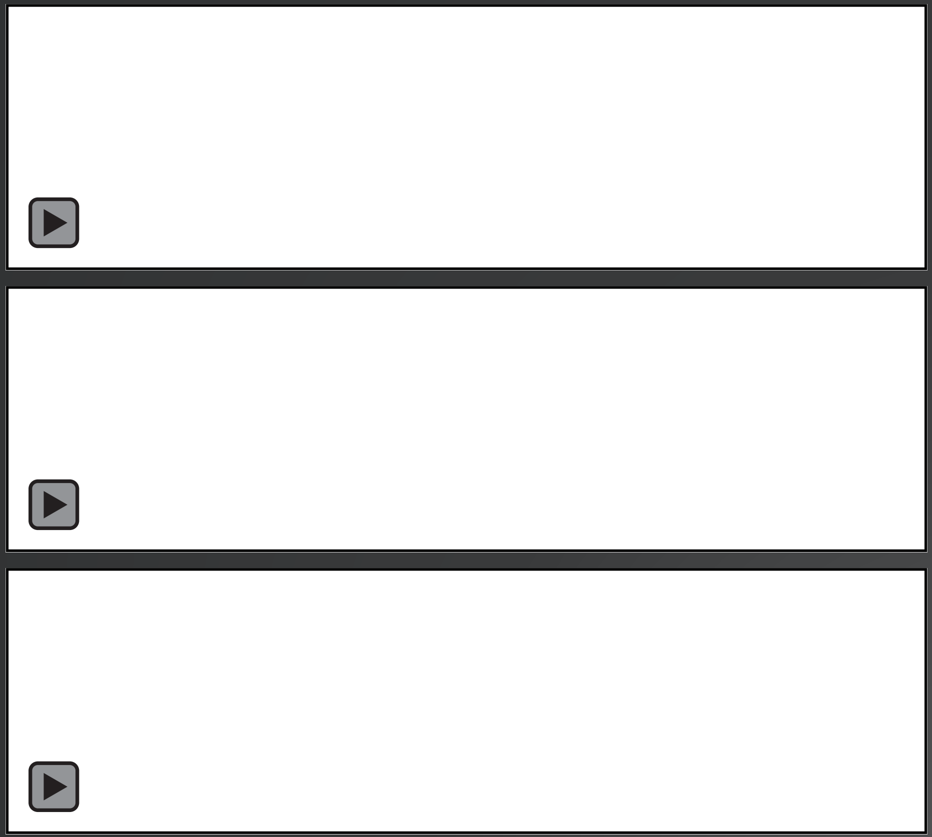
$$D^{t+1}_{x,y} = A(x,y)\odot D^{0}_{x,y} + \sum_{(l,m)\in N_{(x,y)}} A(l,m)\odot D^{t}_{l,m}$$

x, y: spatial coordinate
D: depth value
$(l, m) \in N_{(x,y)}$: neighboring pixels
$\odot$: element-wise product
t: propagation step

Experiments

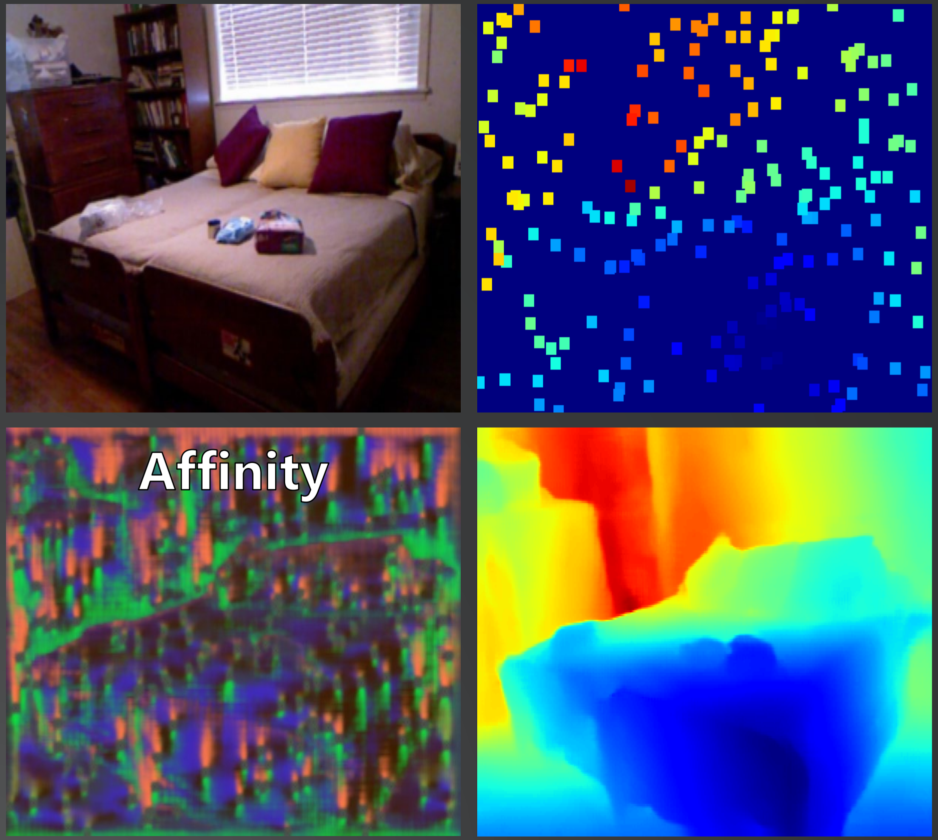
Qualitative Results

KITTI



Line: 32

NYU



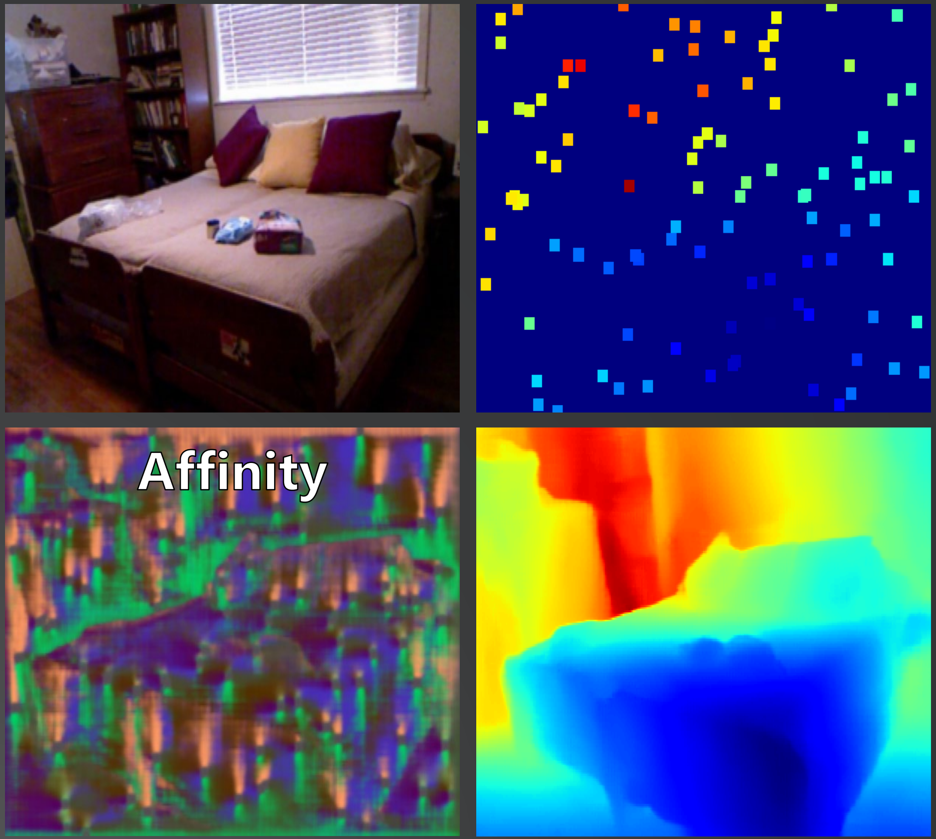
Sample: 200

Experiments

Qualitative Results

KITTI

NYU



Line: 16

Sample: 100

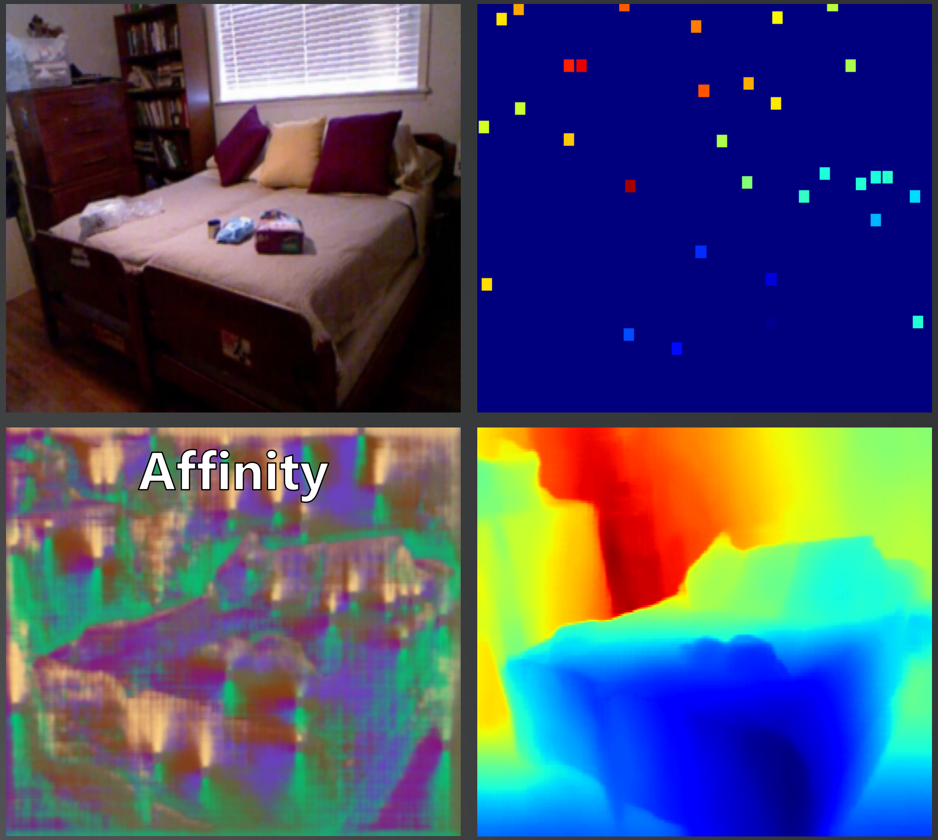
Experiments

Qualitative Results

KITTI

Line: 8

NYU

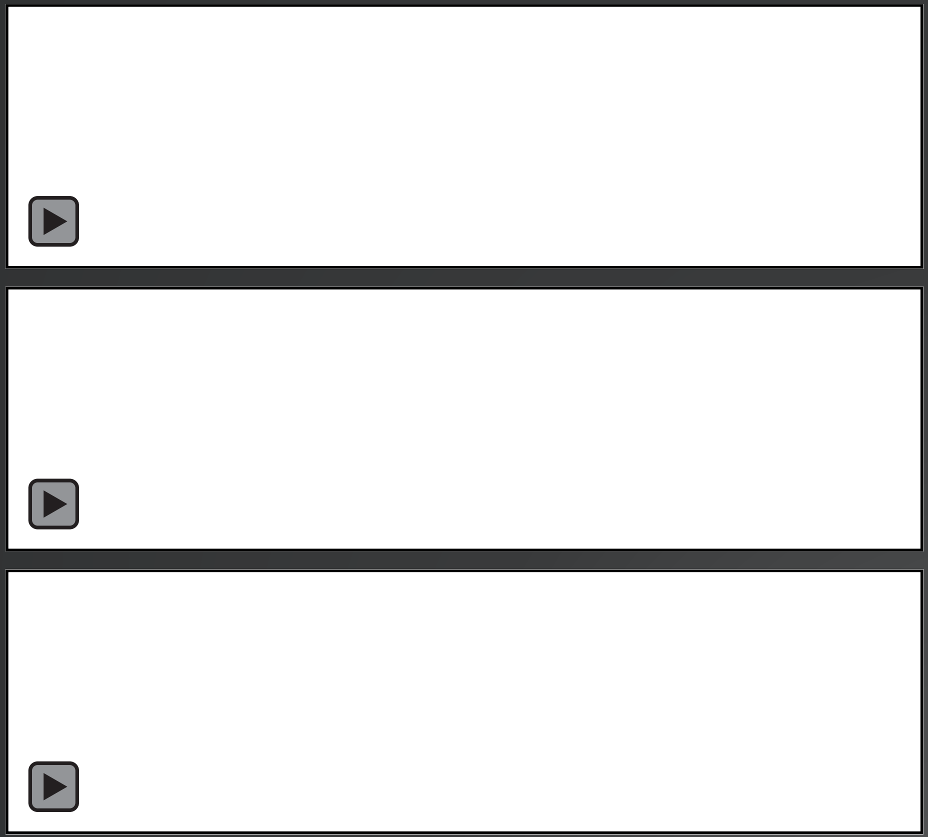


Sample: 32

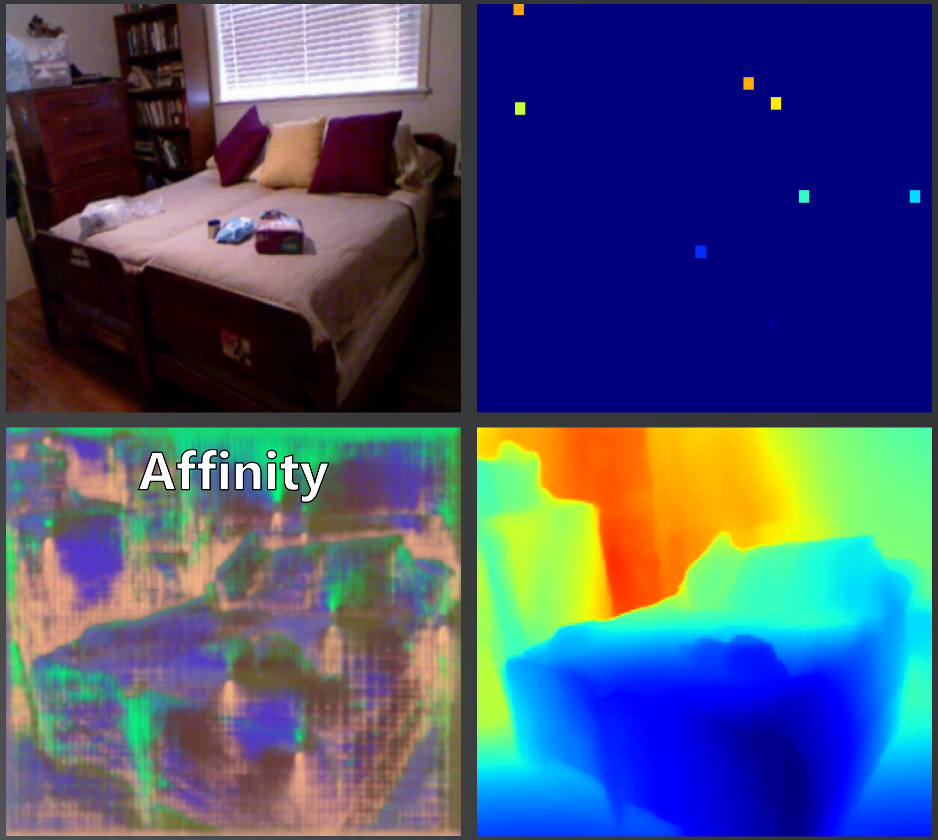
Experiments

Qualitative Results

KITTI



NYU



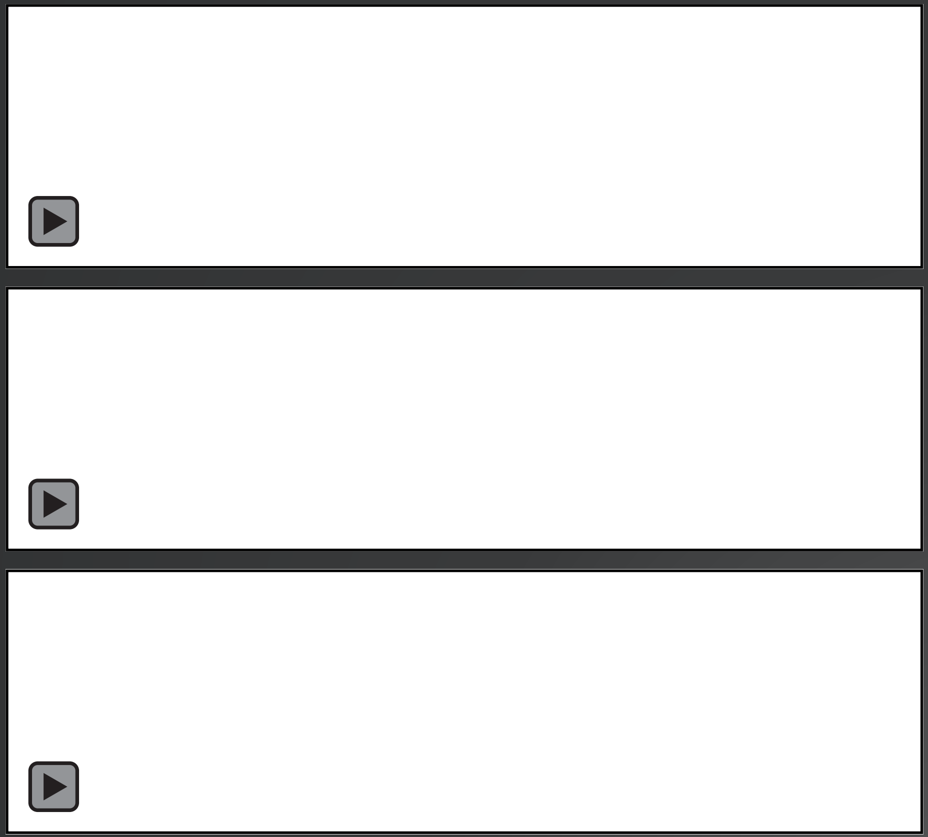
Line: 4

Sample: 8

Experiments

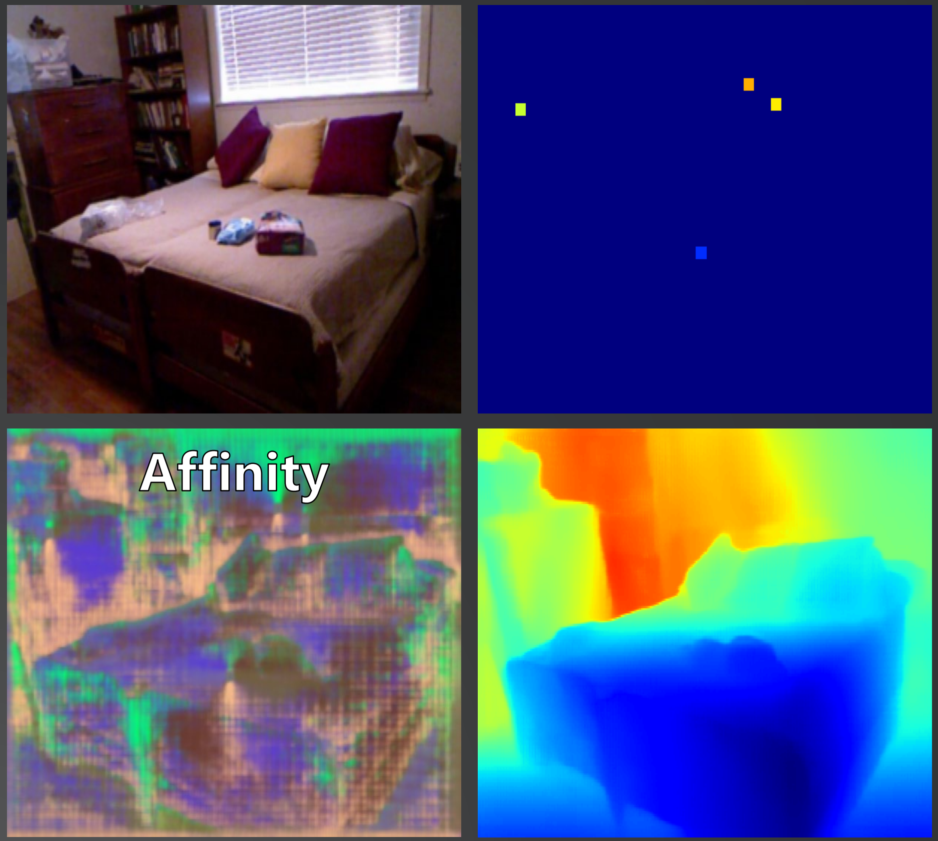
Qualitative Results

KITTI



Line: 2

NYU

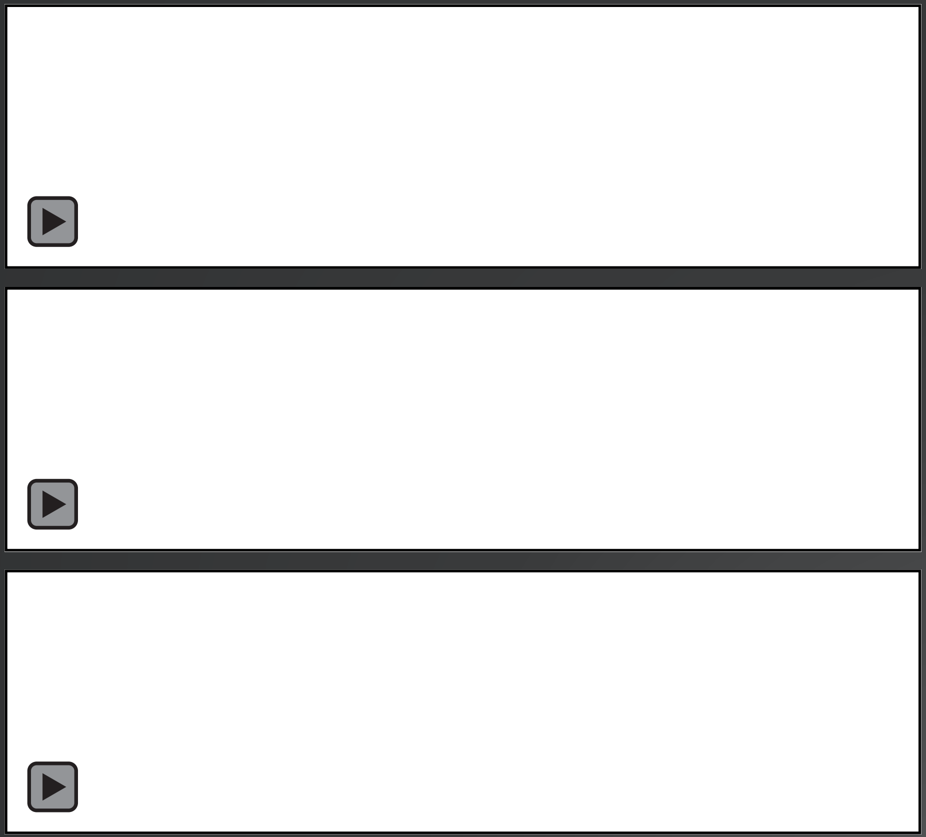


Sample: 4

Experiments

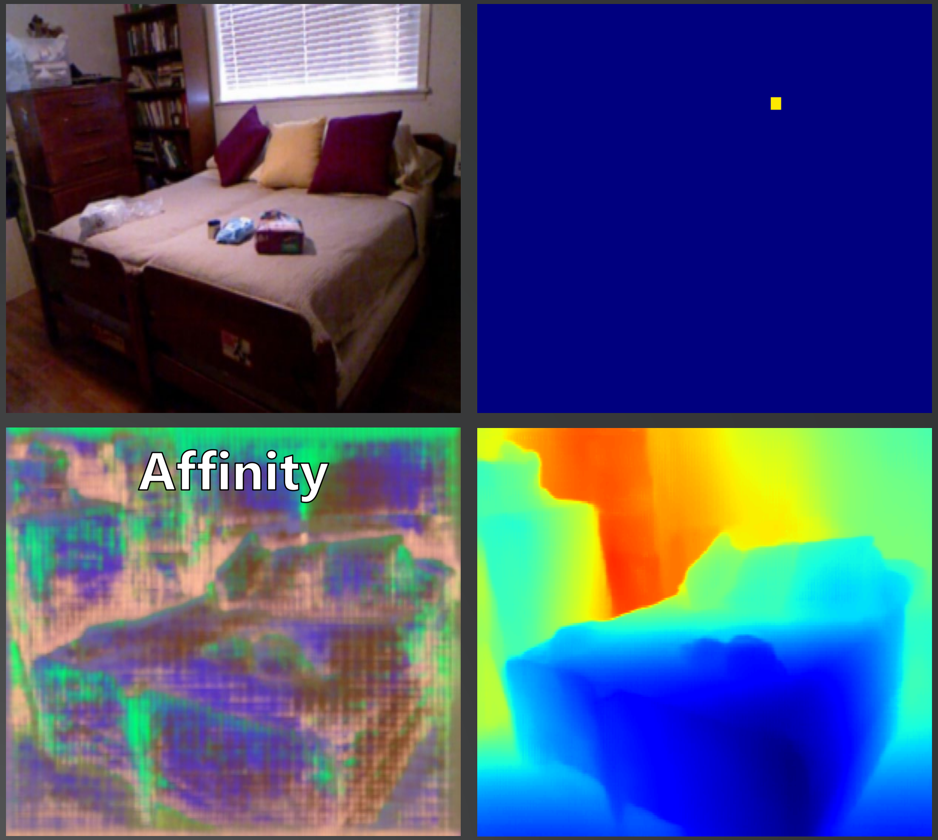
Qualitative Results

KITTI



Line: 1

NYU



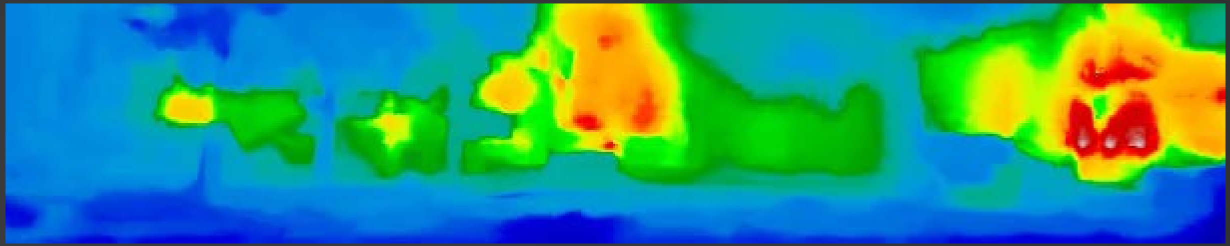
Sample: 1

Experiments

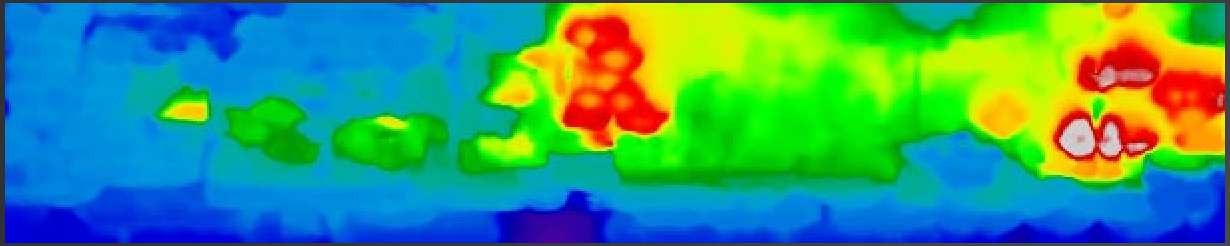
Qualitative Results

- Zero-shot Inference on nuScenes Dataset (Trained on KITTI dataset)

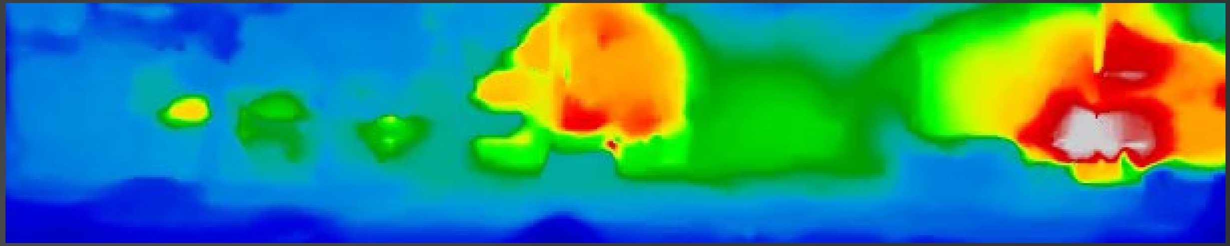
RGB and Sparse Depth



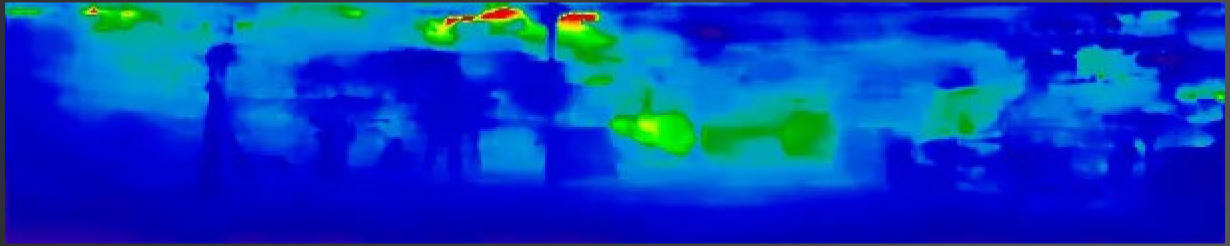
CompletionFormer (CVPR23)



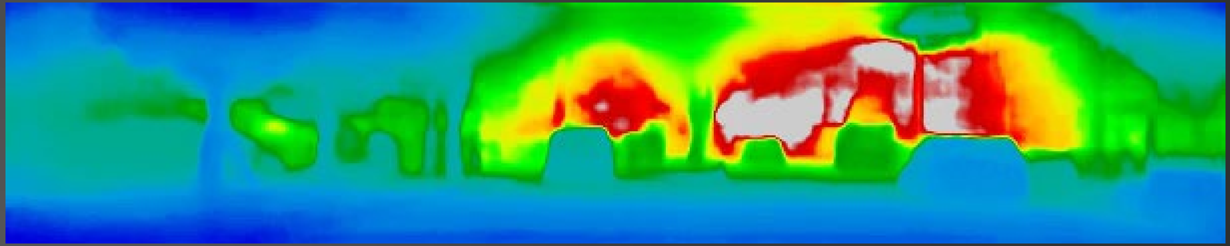
CSPN (ECCV18)



NLSPN (ECCV20)



SAN (CVPR21)

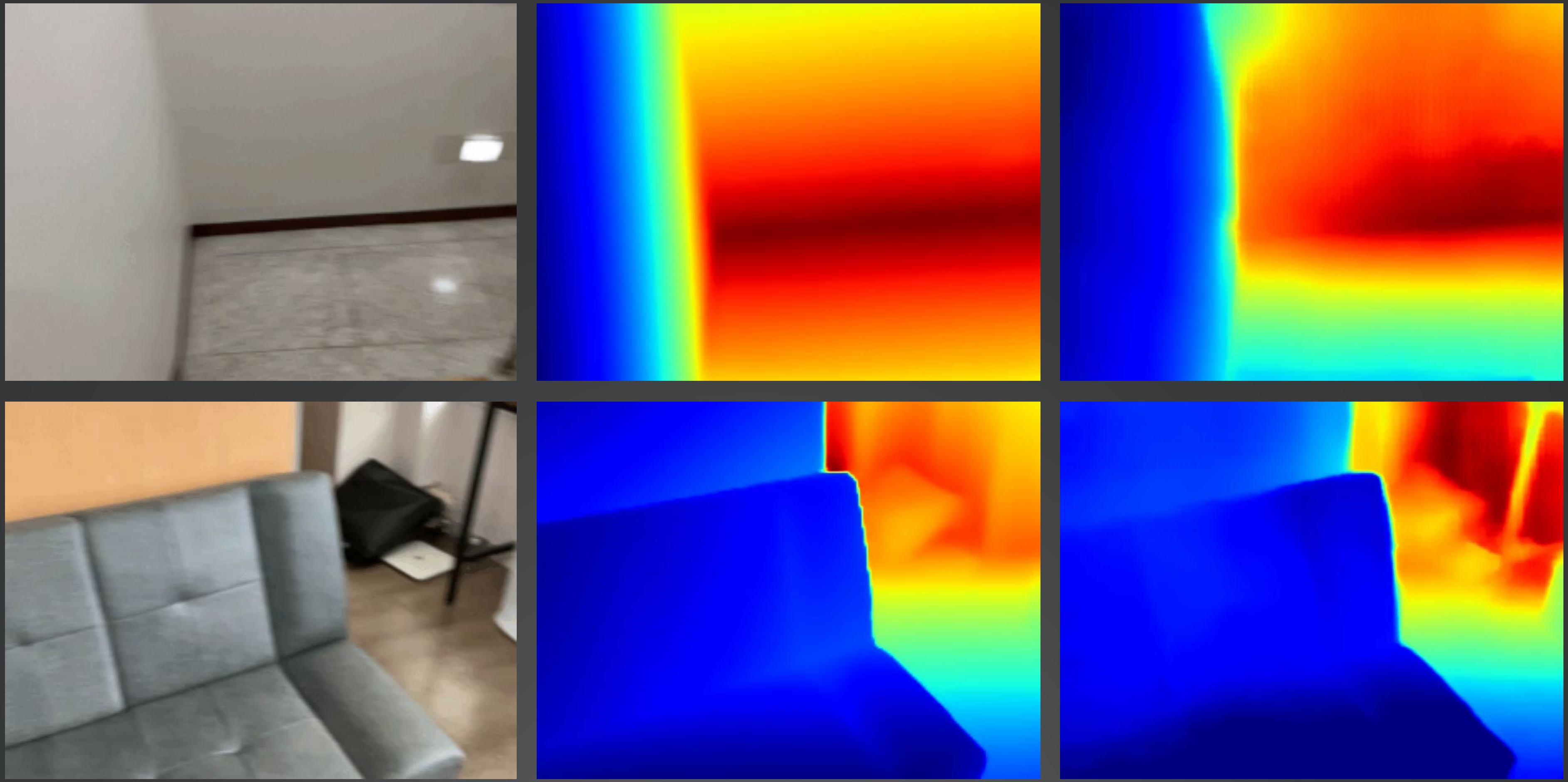


Ours (CVPR24)

Experiments

Qualitative Results

- Zero-shot Inference on iPad Dataset (Trained on NYU dataset)



RGB iPad Pro 6th Ours

Thanks for your attention!

Depth Prompting for
Sensor-Agnostic Depth Estimation

[Source code]