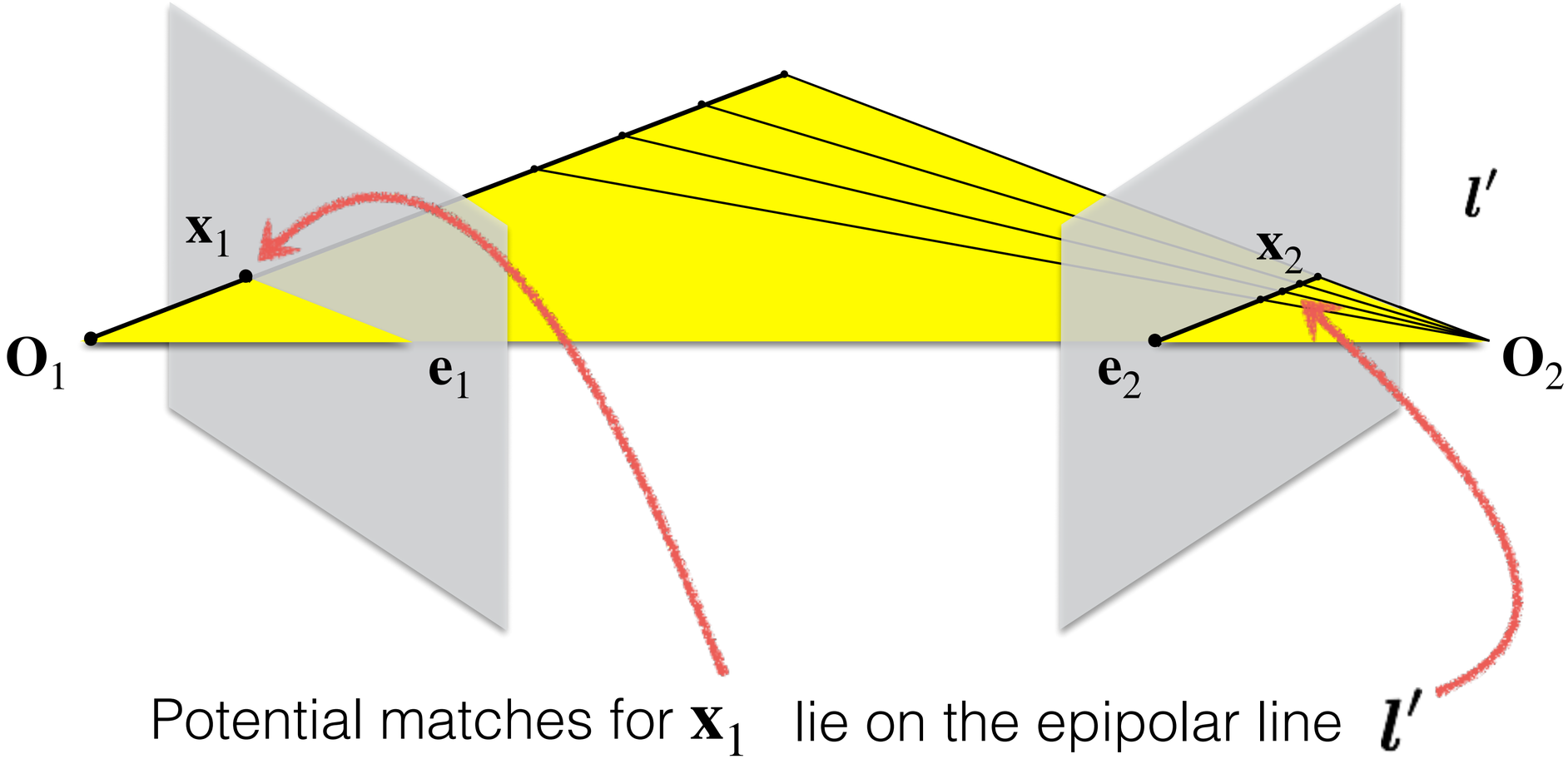
INVERSE GRAPHICS

MODULE 2:
3D Scene Representations & Neural Rendering

DEPTH AND EGOMOTION PREDICTION

6.S980

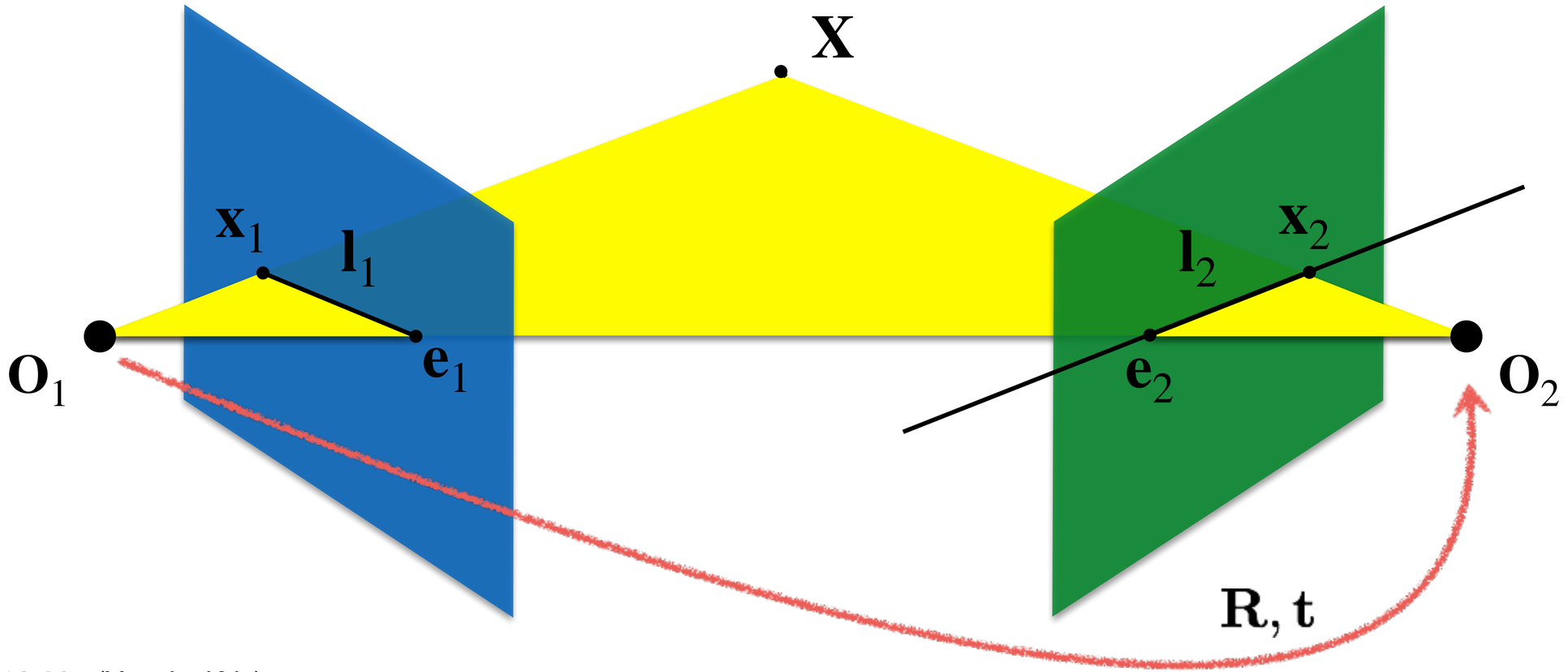
Recap: Epipolar Geometry



Potential matches for $\mathbf{x}_1$ lie on the epipolar line l'

$$\mathbf{F} = \mathbf{K}_2^{-T}(\mathbf{R}[\mathbf{t}]_\times)\mathbf{K}_1^{-1}$$

$$\mathbf{F}\tilde{\mathbf{x}}_1 = \mathbf{l}_2$$



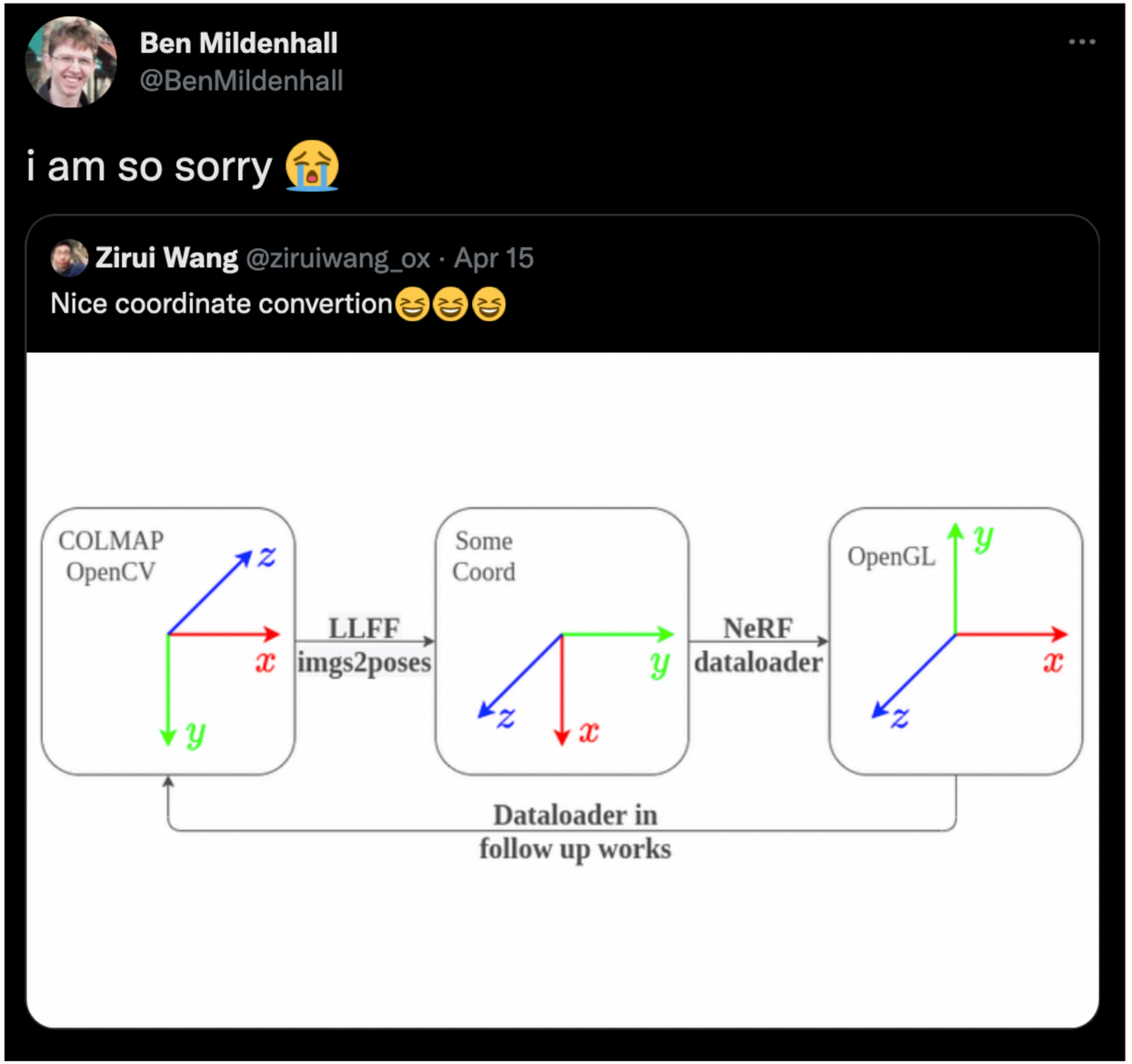
Adapted from: CMU 16-385 (Yannis, Kris)

Recap: Bundle-Adjustment

$$\Pi^*, \mathcal{X}_w^* = \underset{\Pi, \mathcal{X}_w}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{p=1}^{P} w_{ip} \left\| \mathbf{x}_{ip}^s - \pi_i(\mathbf{x}_p^w) \right\|_2^2$$

Adapted from: University of Tübingen: Computer Vision, Prof. Andreas Geiger

On camera conventions



3D Scene

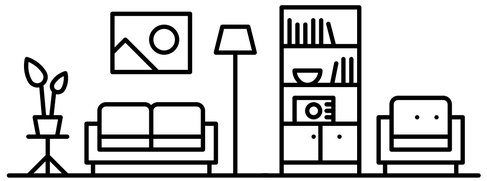


Image Formation

Graphics

3D Scene

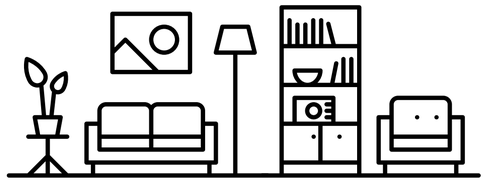


Image Formation

Neural Scene Representation

Graphics

3D Scene

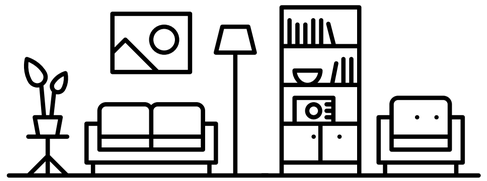


Image Formation

Inference

Neural Scene Representation

Graphics

3D Scene

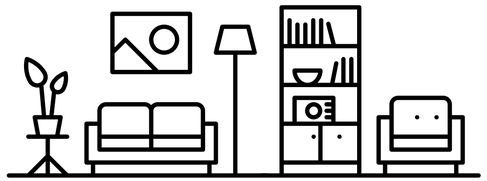


Image Formation

Inference

Neural Scene Representation

Graphics

“inverse graphics”…?

3D Scene

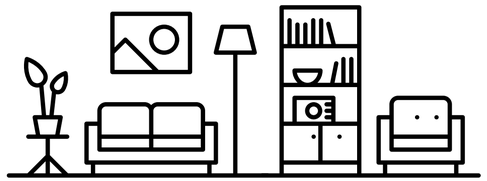


Image Formation

Inference

Neural Scene Representation

Downstream Tasks

Graphics

“inverse graphics”…?

Today: How to computationally represent 3D scenes.

3D Scene

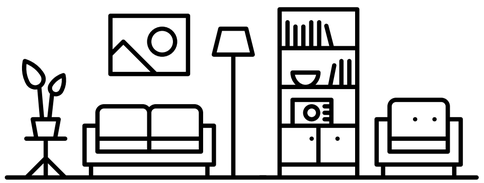


Image Formation

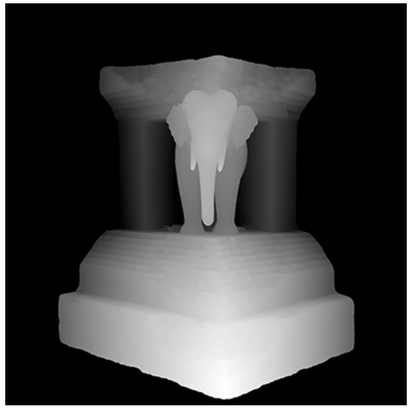
Inference

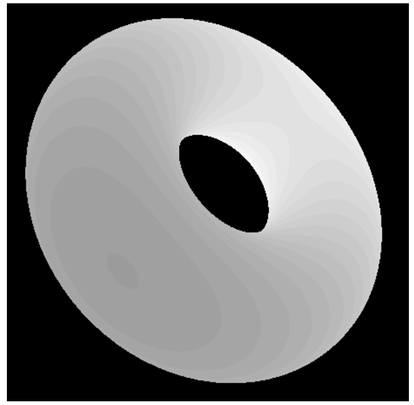
Neural Scene Representation

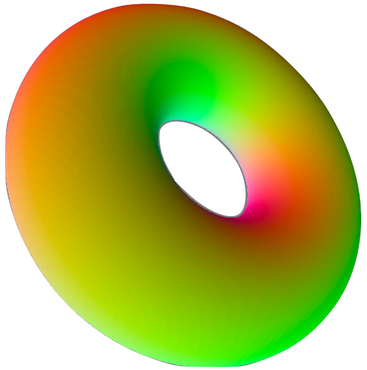
Downstream Tasks

Why?	At the end of the day, we want to make predictions about 3D scenes. For that, we need to know how we can represent 3D scenes computationally.
What you'll learn.	Surface-based representations, volumetric representations, discrete representations, continuous representations.

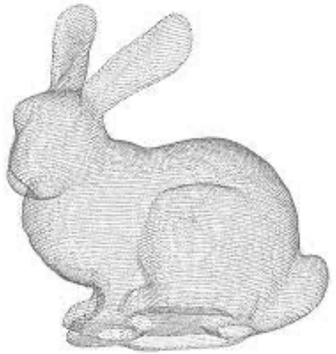
2.5D Representations

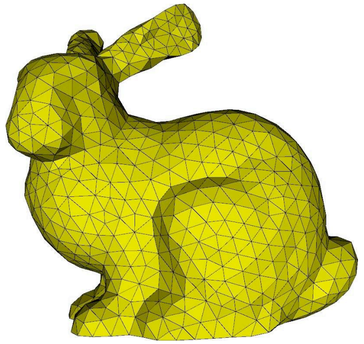




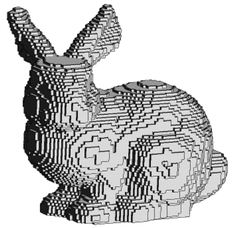


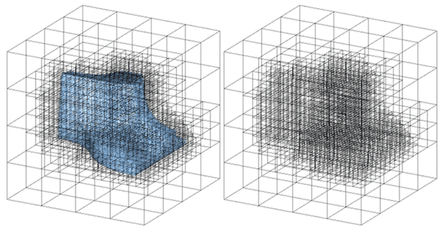
Surface Representations

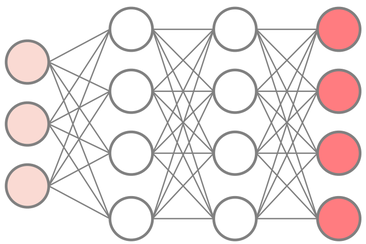


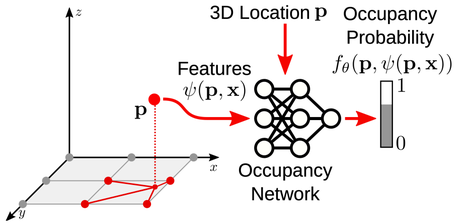


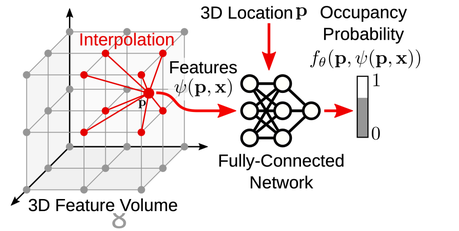
Field Parameterizations



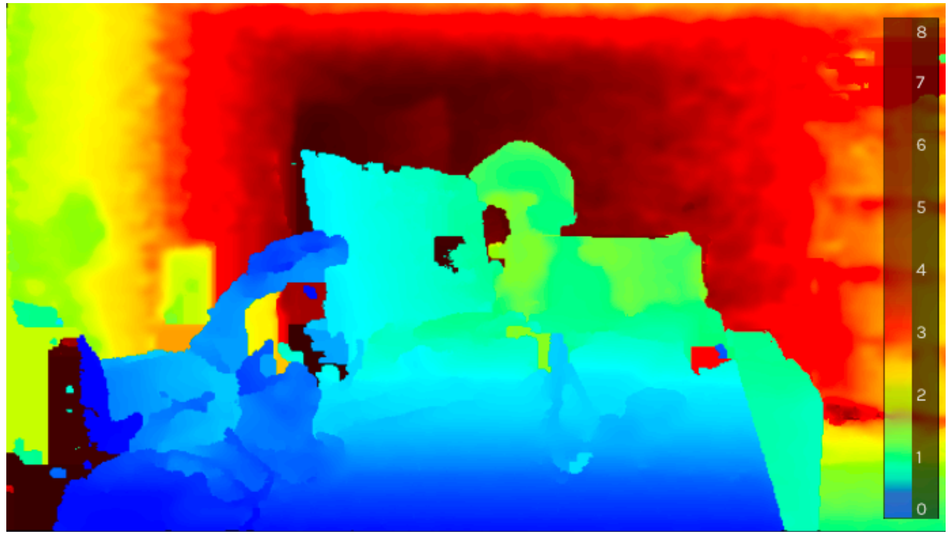








Today: all about depth!



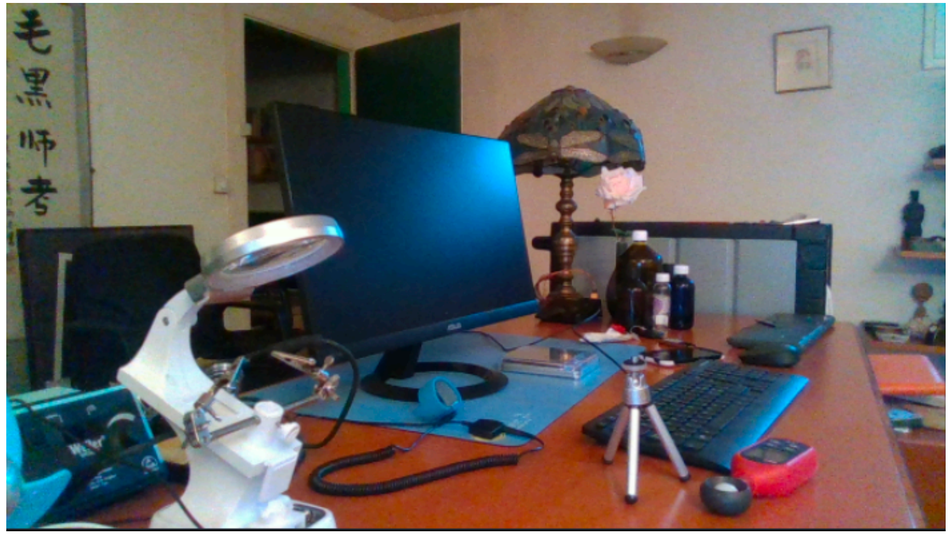
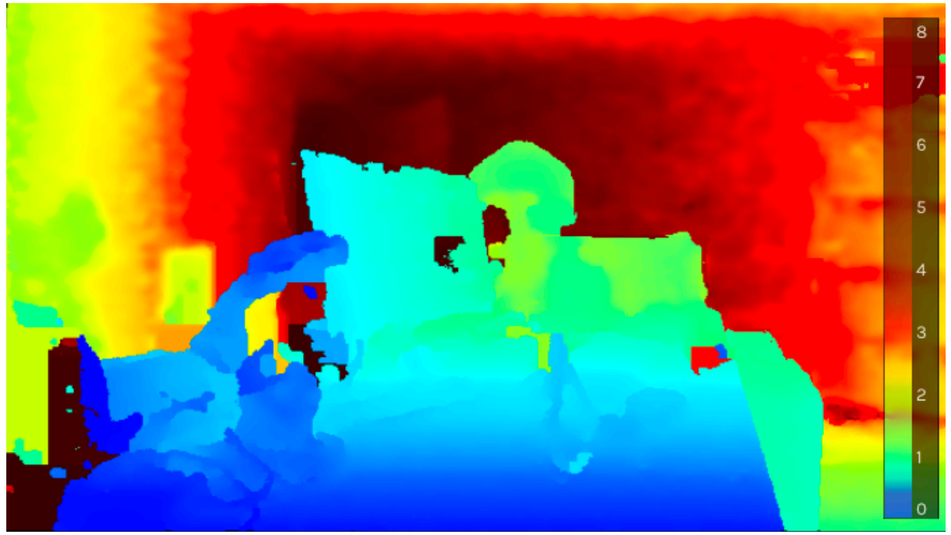
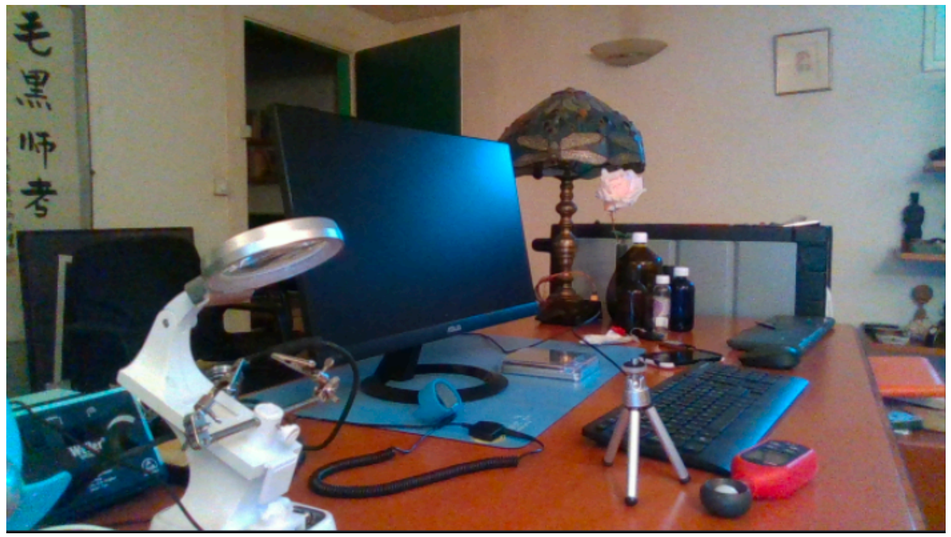


Image source: https://www.calvert.ch/maurice/2018/06/12/intel-realsense-d435-review/

Today: all about depth!





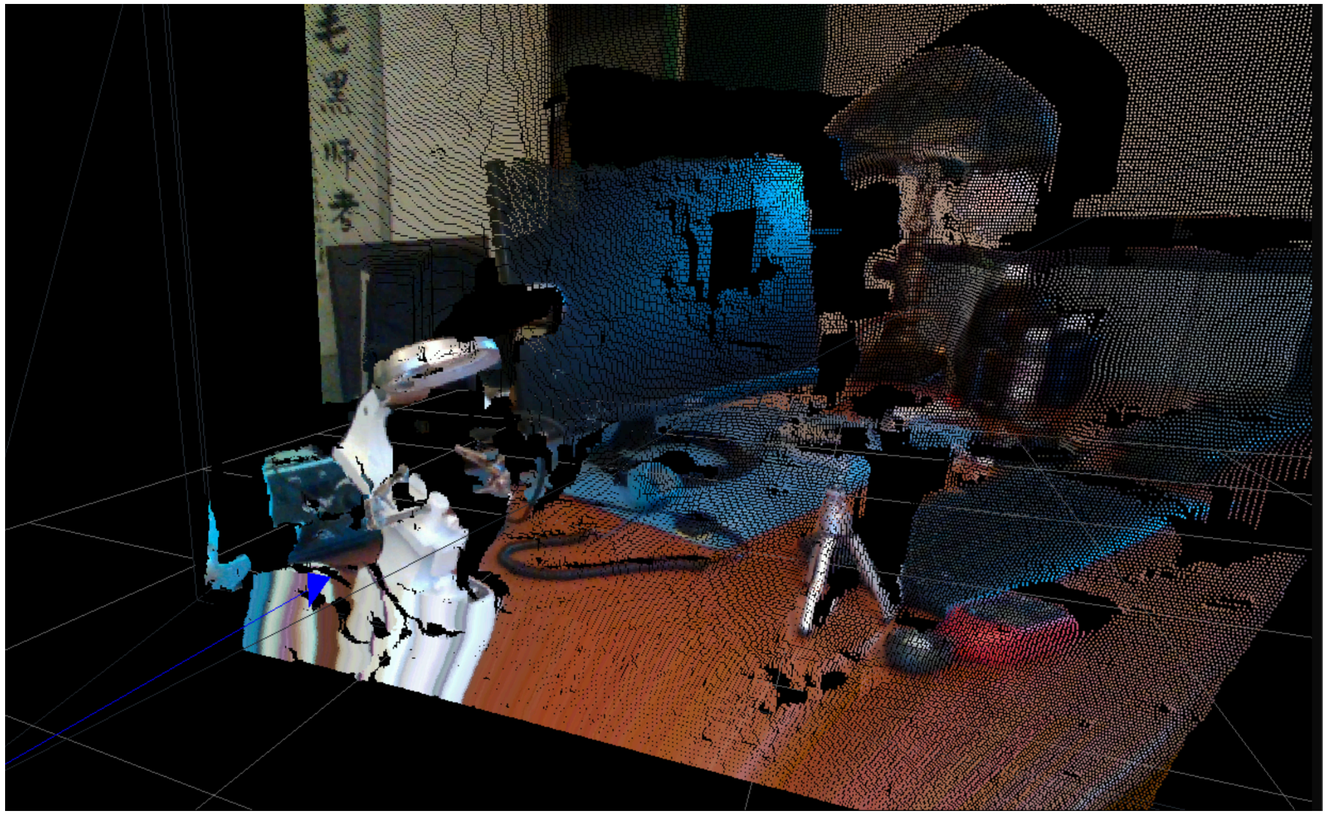
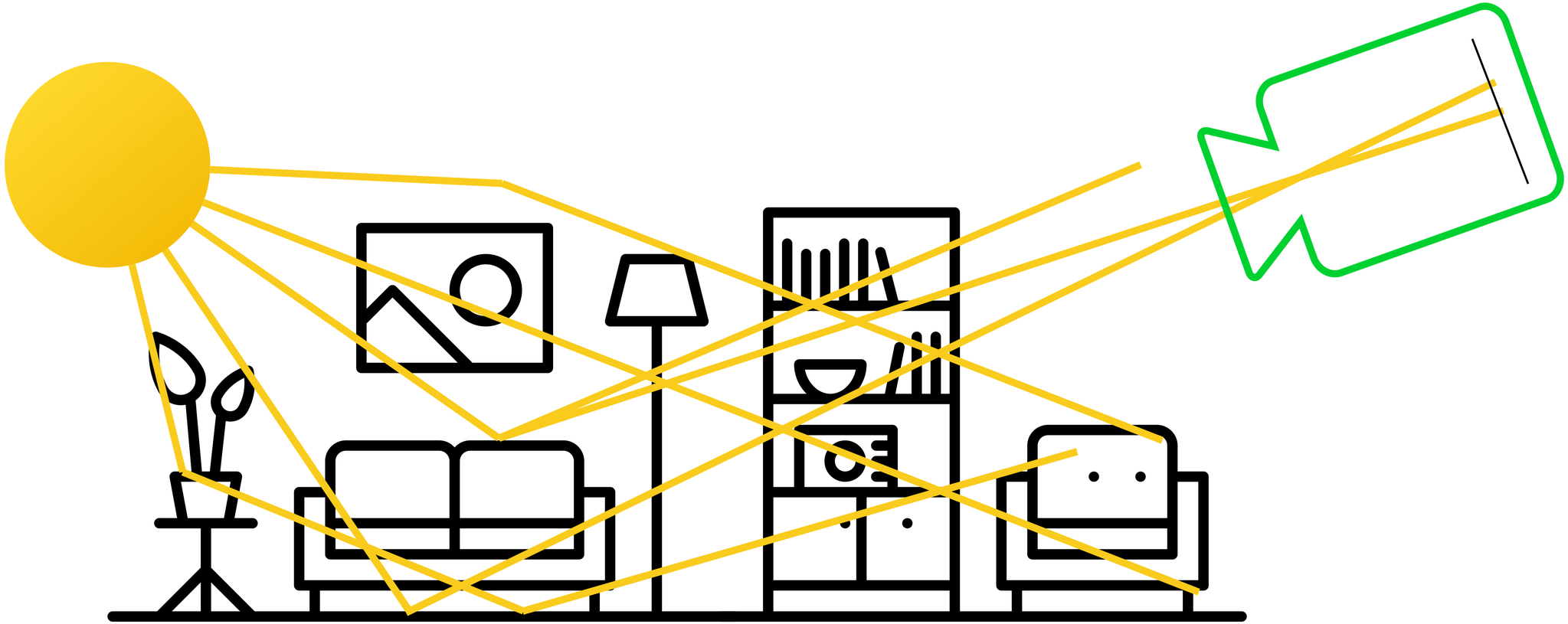
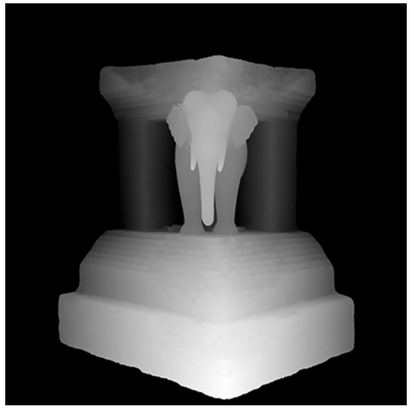


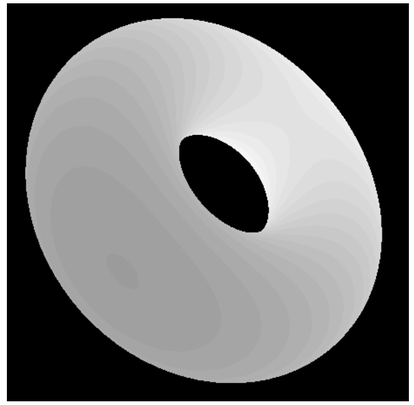
Image source: https://www.calvert.ch/maurice/2018/06/12/intel-realsense-d435-review/

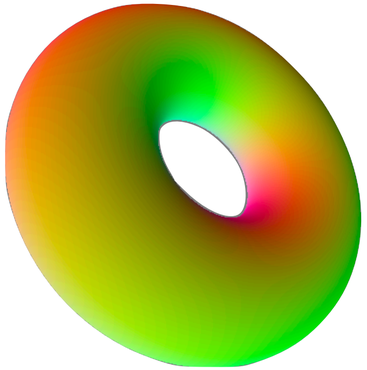
What does the camera see?



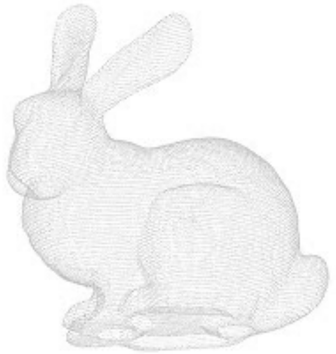
2.5D Representations

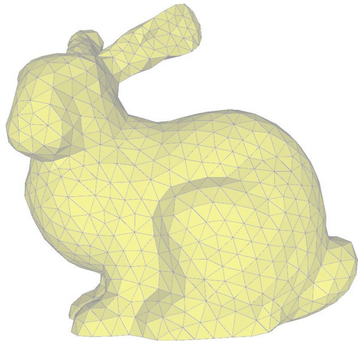




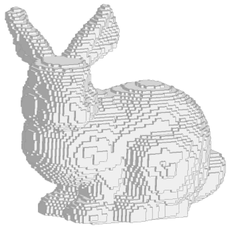


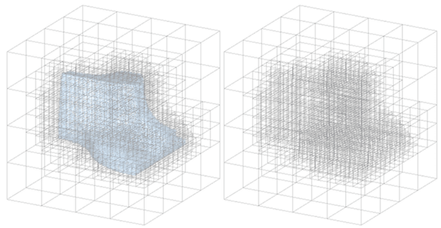
Surface Representations

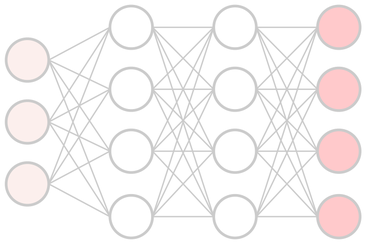


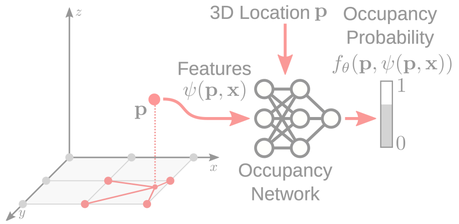


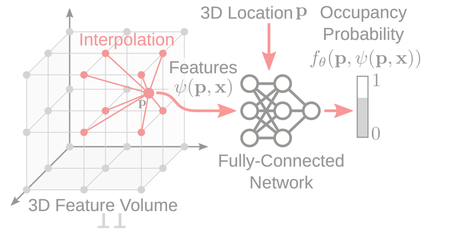
Field Parameterizations











Depth Maps

Slide credit: CMU 16-889: Learning for 3D Vision, Prof. Shubham Tulsiani

Depth Maps

Common modality, captured with Stereo, LIDAR, ….

Slide credit: CMU 16-889: Learning for 3D Vision, Prof. Shubham Tulsiani

Depth Maps

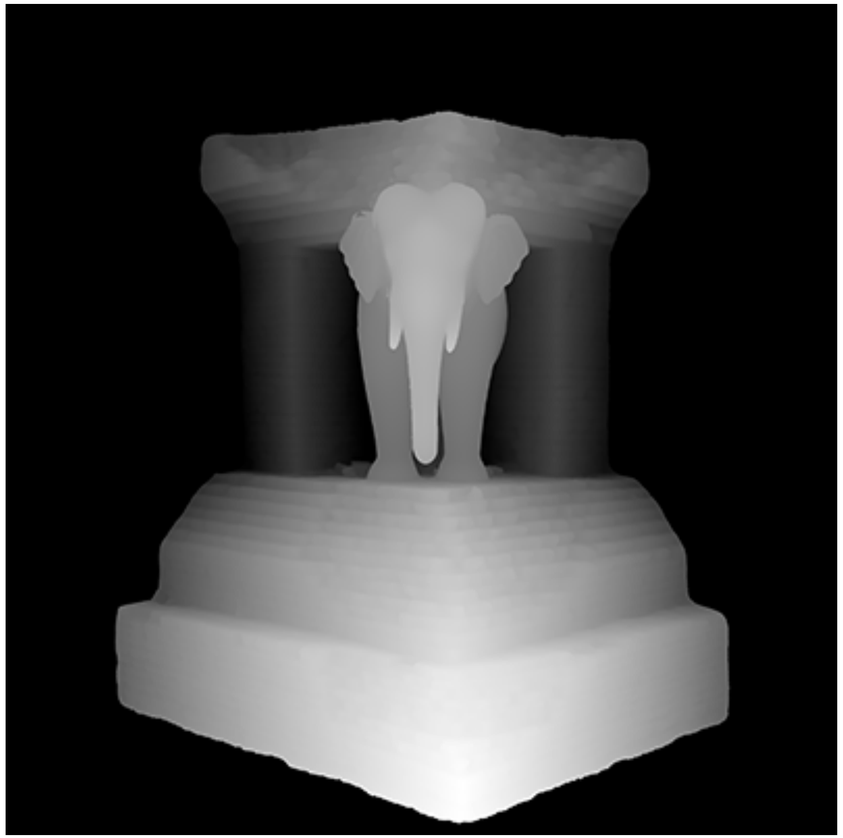
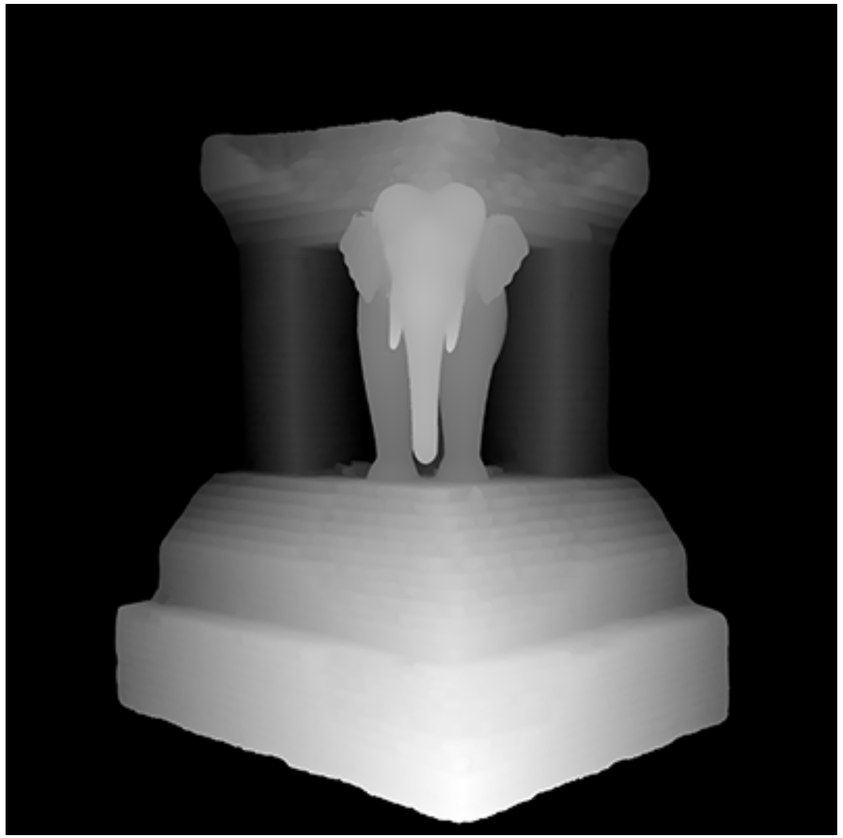


Image credits: Paul Bourke, Slide credit: CMU 16-889: Learning for 3D Vision, Prof. Shubham Tulsiani

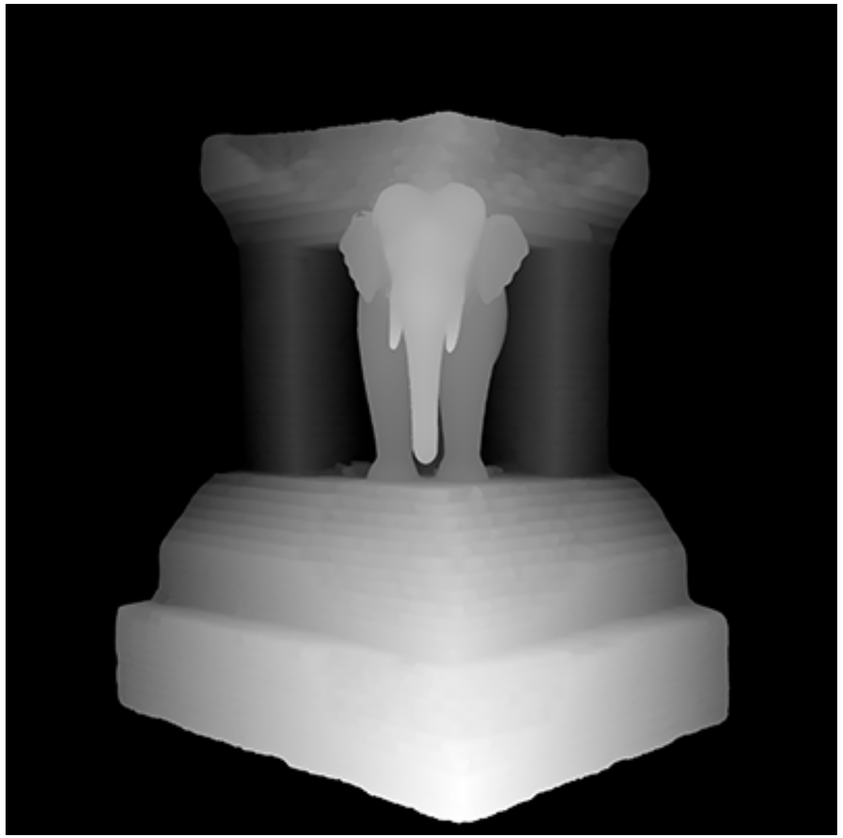
Depth Maps



An **image** that represents how far each pixel **p** is

Image credits: Paul Bourke, Slide credit: CMU 16-889: Learning for 3D Vision, Prof. Shubham Tulsiani

Depth Maps

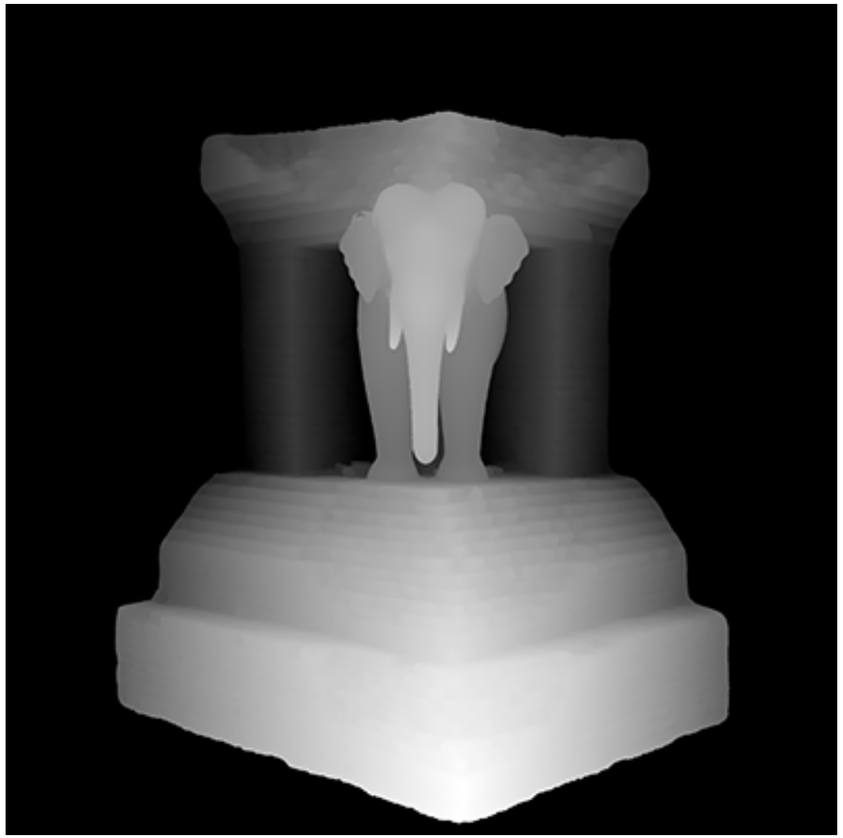


An **image** that represents how far each pixel **p** is

$$D[\mathbf{p}] \in \mathbb{R}^+$$

Image credits: Paul Bourke, Slide credit: CMU 16-889: Learning for 3D Vision, Prof. Shubham Tulsiani

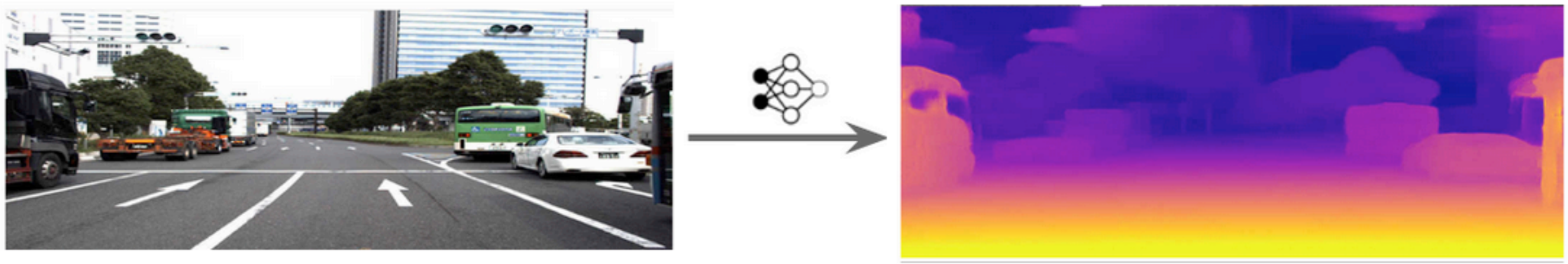
Depth Maps



The ‘standard’ image processing tools can be used

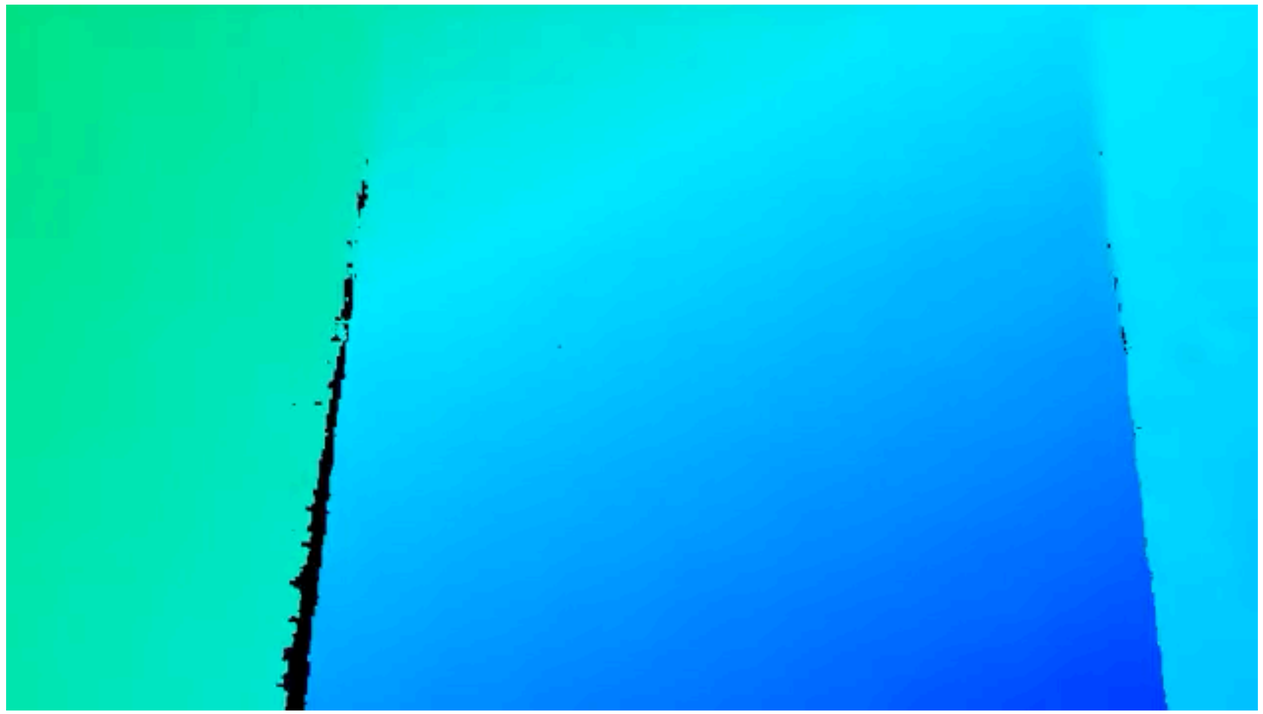
Image credits: Paul Bourke, Slide credit: CMU 16-889: Learning for 3D Vision, Prof. Shubham Tulsiani

Monocular Depth Estimation



https://medium.com/toyotaresearch/self-supervised-learning-in-depth-part-1-of-2-74825baaaa04

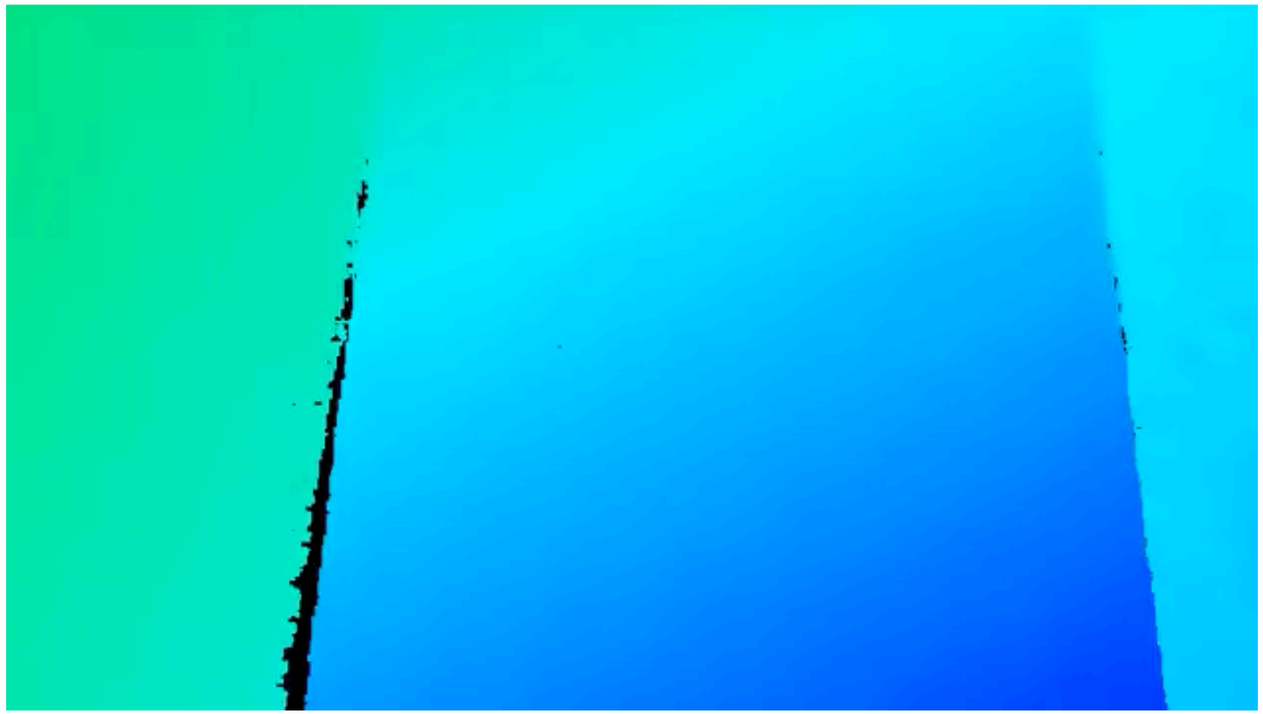
Depth Sensors



Output of Intel RealSense Camera

https://www.youtube.com/watch?v=t-X7tnT8CVY

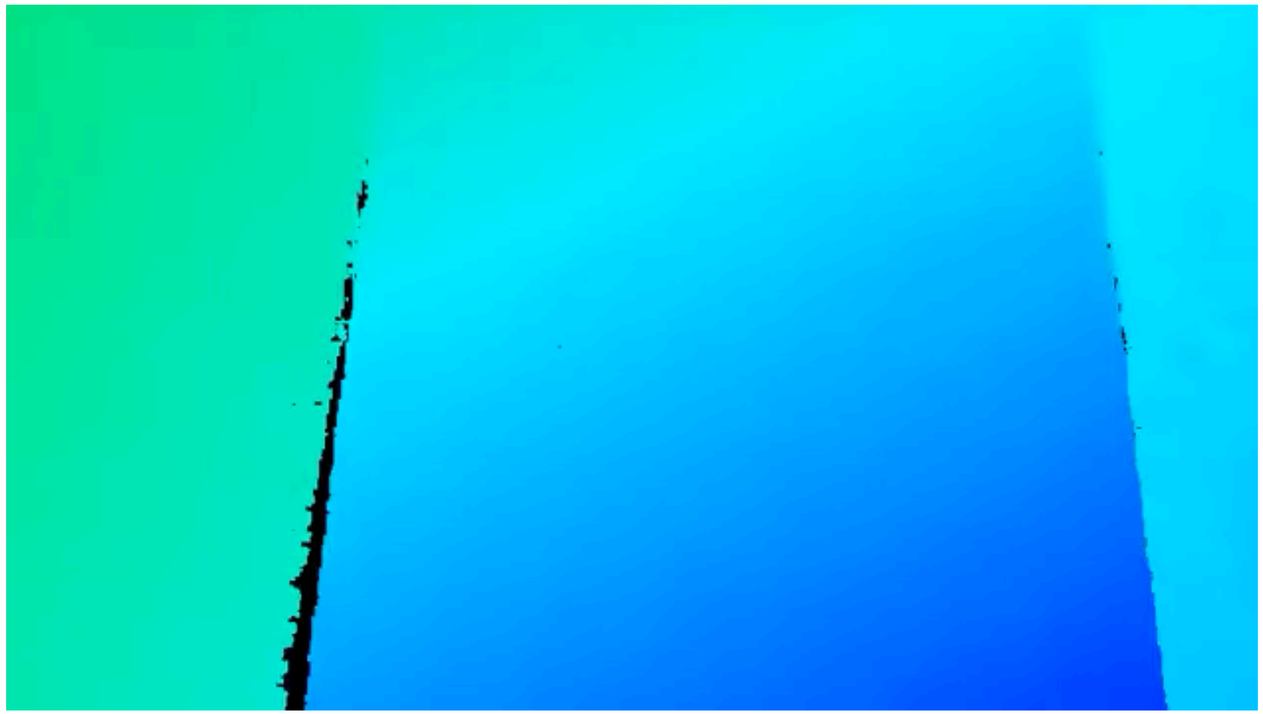
Depth Sensors



Output of Intel RealSense Camera

https://www.youtube.com/watch?v=t-X7tnT8CVY

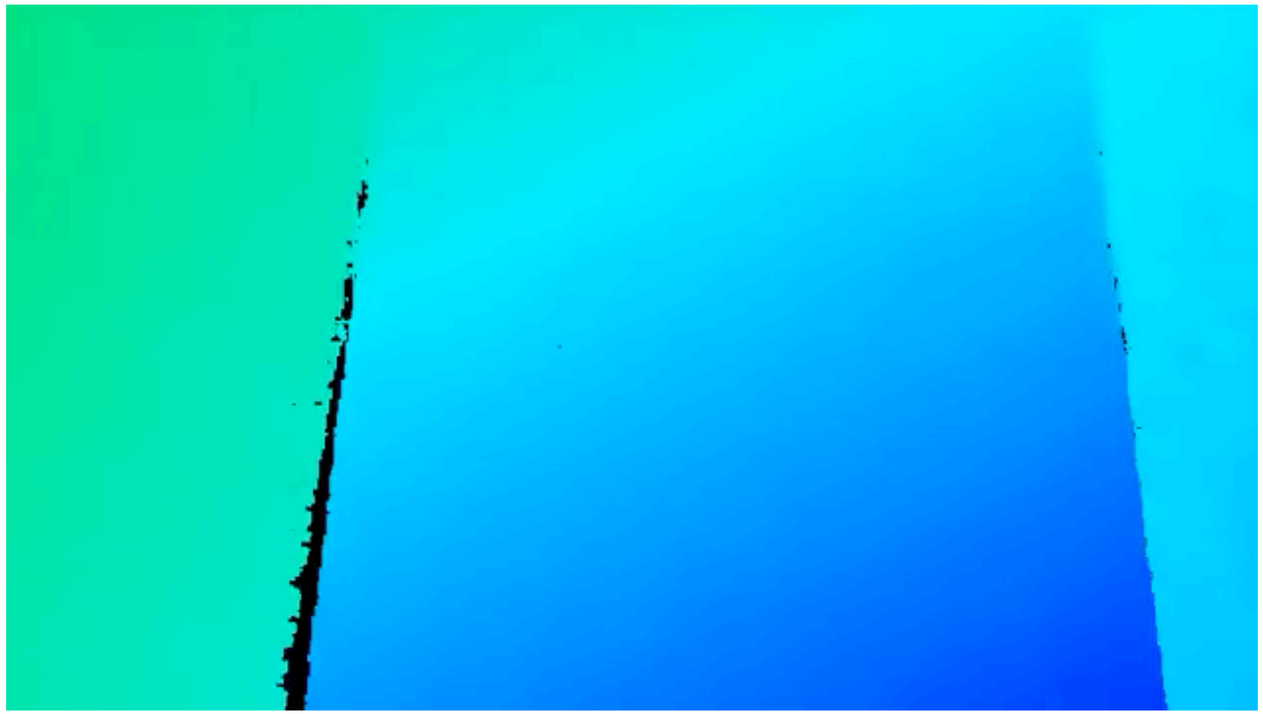
Depth Sensors: An easy way to collect “ground truth” supervision



Output of Intel RealSense Camera

https://www.youtube.com/watch?v=t-X7tnT8CVY

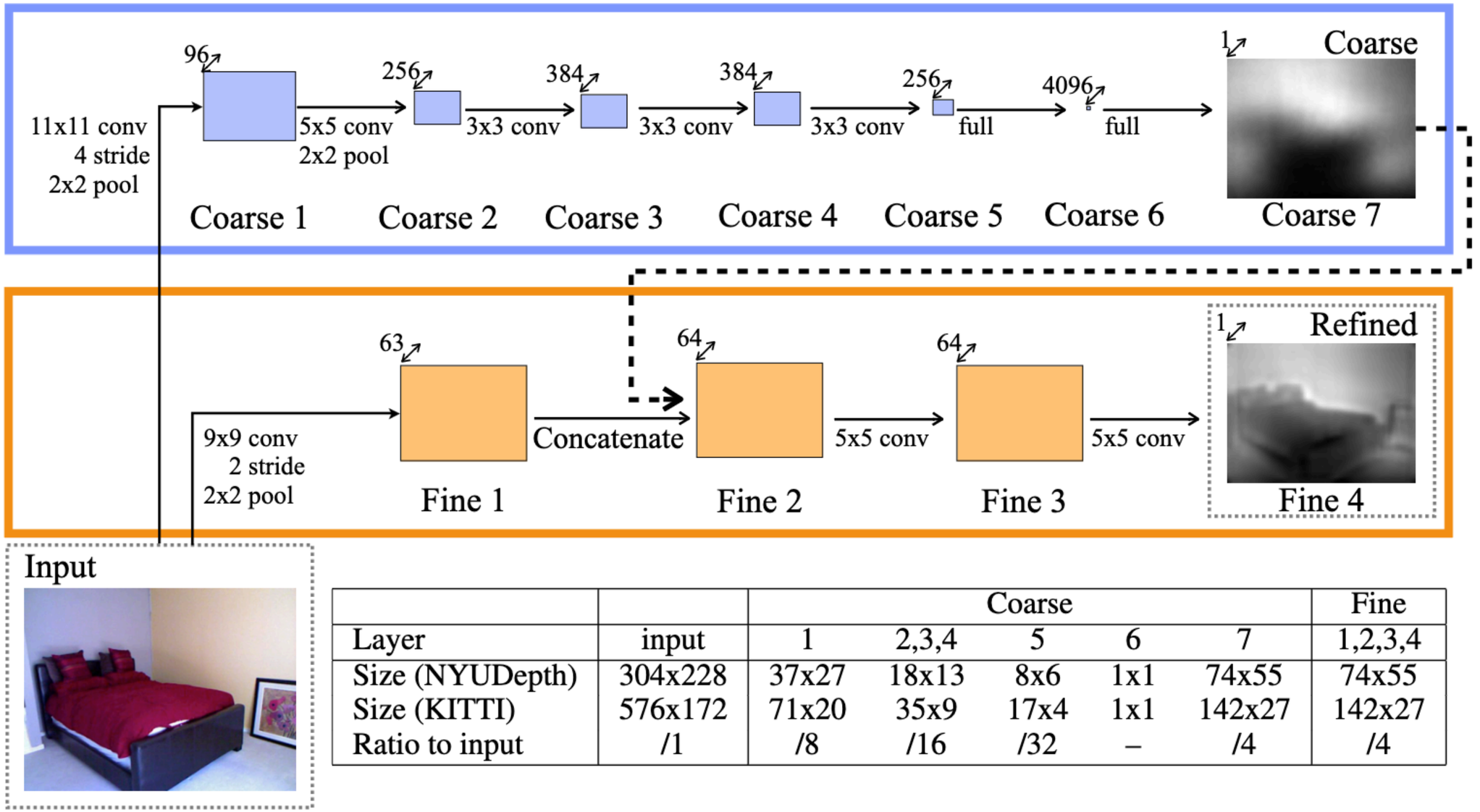
Depth Sensors: An easy way to collect “ground truth” supervision



Output of Intel RealSense Camera

https://www.youtube.com/watch?v=t-X7tnT8CVY

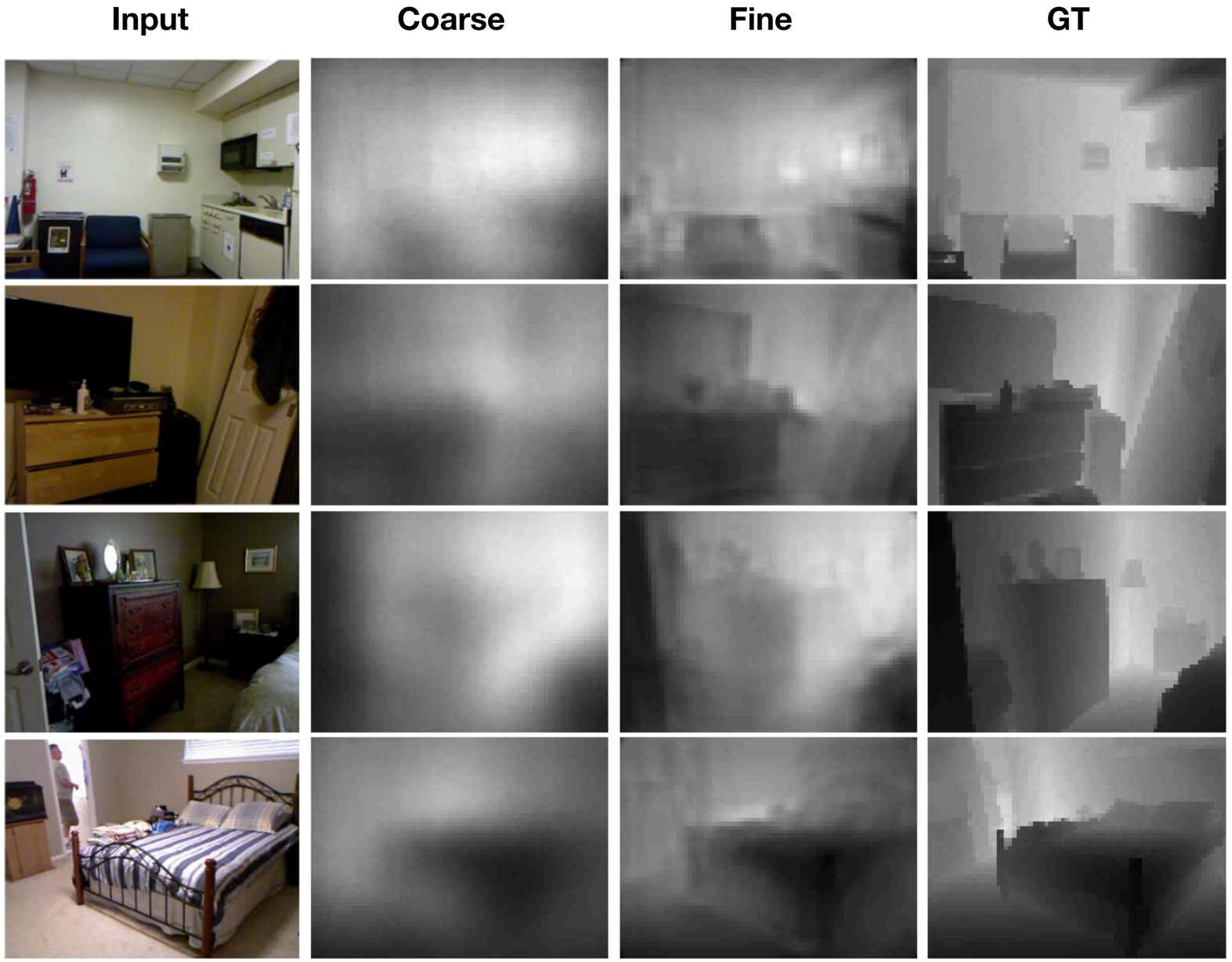
Supervised **Monocular** Depth Estimation:
Depth Map Prediction from a Single Image using a Multi-Scale Deep Network
Eigen et al. 2014



		Coarse					Fine
Layer	input	1	2,3,4	5	6	7	1,2,3,4
Size (NYUDepth)	304x228	37x27	18x13	8x6	1x1	74x55	74x55
Size (KITTI)	576x172	71x20	35x9	17x4	1x1	142x27	142x27
Ratio to input	/1	/8	/16	/32	–	/4	/4

Figure 1: Model architecture.

Depth Map Prediction from a Single Image using a Multi-Scale Deep Network
Eigen et al. 2014



“Ground-truth” depth

NYU Depth v2, Silberman et al.

Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer, Ranftl et al. 2022

Dataset	Indoor	Outdoor	Dynamic	Video	Dense	Accuracy	Diversity	Annotation	Depth	# Images
DIML Indoor [31]	✓			✓	✓	Medium	Medium	RGB-D	**Metric**	220K
MegaDepth [11]		✓	(✓)		(✓)	Medium	Medium	SfM	No scale	130K
ReDWeb [32]	✓	✓	✓		✓	Medium	**High**	Stereo	No scale & shift	3600
WSVD [33]	✓	✓	✓	✓	✓	Medium	**High**	Stereo	No scale & shift	1.5M
3D Movies	✓	✓	✓	✓	✓	Medium	**High**	Stereo	No scale & shift	75K
DIW [34]	✓	✓	✓			Low	**High**	User clicks	Ordinal pair	496K
ETH3D [35]	✓	✓			✓	**High**	Low	Laser	**Metric**	454
Sintel [36]	✓	✓	✓	✓	✓	**High**	Medium	Synthetic	(Metric)	1064
KITTI [28], [29]		✓	(✓)	✓	(✓)	Medium	Low	Laser/Stereo	**Metric**	93K
NYUDv2 [30]	✓		(✓)	✓	✓	Medium	Low	RGB-D	**Metric**	407K
TUM-RGBD [37]	✓		(✓)	✓	✓	Medium	Low	RGB-D	**Metric**	80K

Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer, Ranftl et al. 2022

Dataset	Indoor	Annotation	Depth	Diversity	Annotation	Depth	# Images
DIML Indoor [31]	✓	RGB-D	**Metric**	Medium	RGB-D	**Metric**	220K
MegaDepth [11]		SfM	No scale	Medium	SfM	No scale	130K
ReDWeb [32]	✓	Stereo	No scale & shift	**High**	Stereo	No scale & shift	3600
WSVD [33]	✓	Stereo	No scale & shift	**High**	Stereo	No scale & shift	1.5M
3D Movies	✓	Stereo	No scale & shift	**High**	Stereo	No scale & shift	75K
DIW [34]	✓	User clicks	Ordinal pair	**High**	User clicks	Ordinal pair	496K
ETH3D [35]	✓	Laser	**Metric**	Low	Laser	**Metric**	454
Sintel [36]	✓	Synthetic	(Metric)	Medium	Synthetic	(Metric)	1064
KITTI [28], [29]		Laser/Stereo	**Metric**	Low	Laser/Stereo	**Metric**	93K
NYUDv2 [30]	✓	RGB-D	**Metric**	Low	RGB-D	**Metric**	407K
TUM-RGBD [37]	✓	RGB-D	**Metric**	Low	RGB-D	**Metric**	80K

What does that mean for a mean-squared error loss?

$$\mathscr{L} = \|d - d^* \|^2$$

Depth Map Prediction from a Single Image using a Multi-Scale Deep Network
Eigen et al. 2014

$$D(y, y^*) = \frac{1}{n}\sum_{i=1}^{n}(\log y_i - \log y_i^* + \alpha(y, y^*))^2, \quad (1)$$

where $\alpha(y, y^*) = \frac{1}{n}\sum_i(\log y_i^* - \log y_i)$ is the value of α that minimizes the error for a given (y, y^*).

Depth Map Prediction from a Single Image using a Multi-Scale Deep Network
Eigen et al. 2014

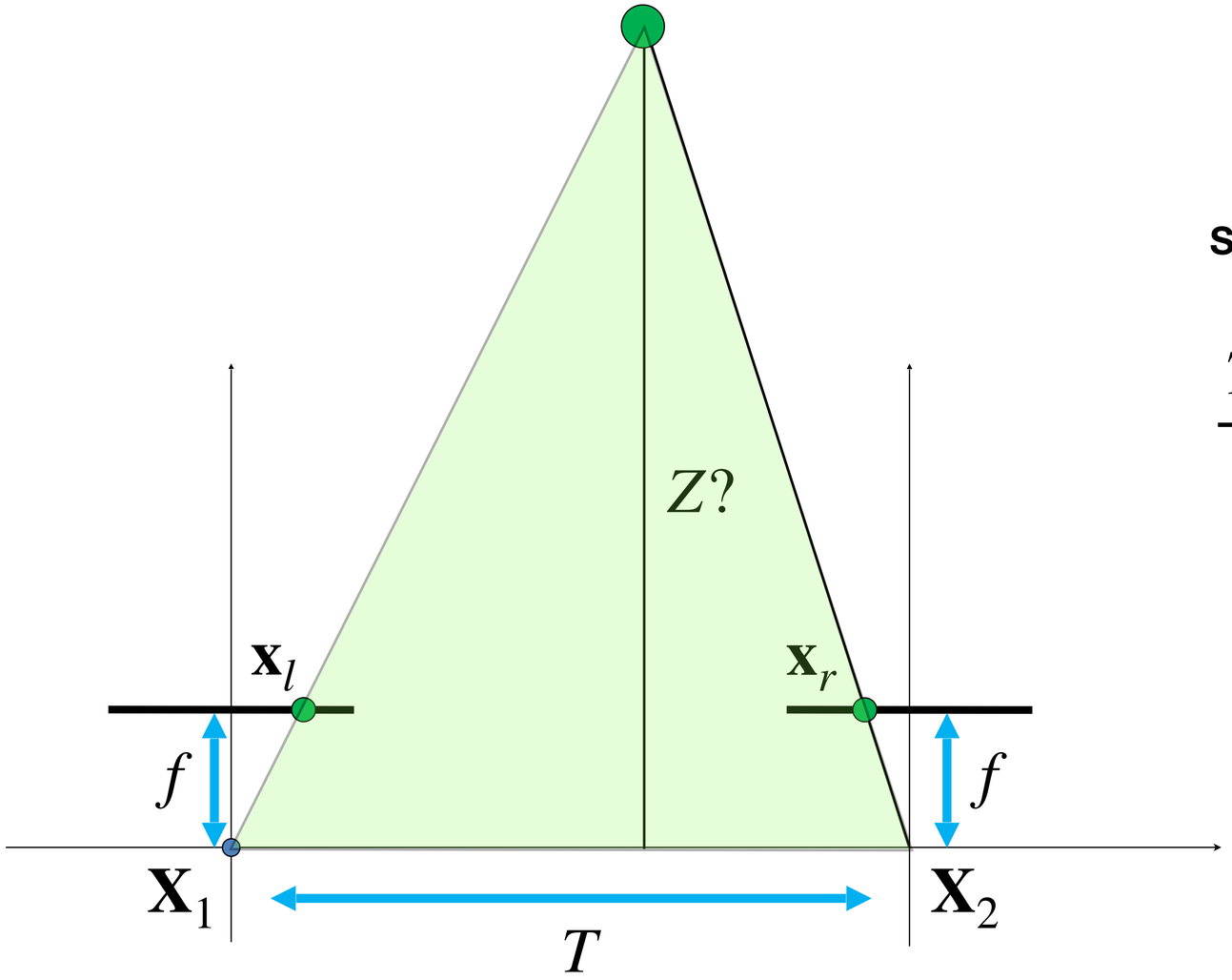
$$D(y, y^*) = \frac{1}{n}\sum_{i=1}^{n}(\log y_i - \log y_i^* + \alpha(y, y^*))^2, \quad (1)$$

where $\alpha(y, y^*) = \frac{1}{n}\sum_i(\log y_i^* - \log y_i)$ is the value of α that minimizes the error for a given (y, y^*). For any prediction y, e^α is the scale that best aligns it to the ground truth. All scalar multiples of y have the same error, hence the scale invariance.

Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer, Ranftl et al. 2022

Scale- and Shift-Invariant Losses. We propose to perform prediction in disparity space (inverse depth up to scale and shift) together with a family of scale- and shift-invariant dense losses …

Disparity map



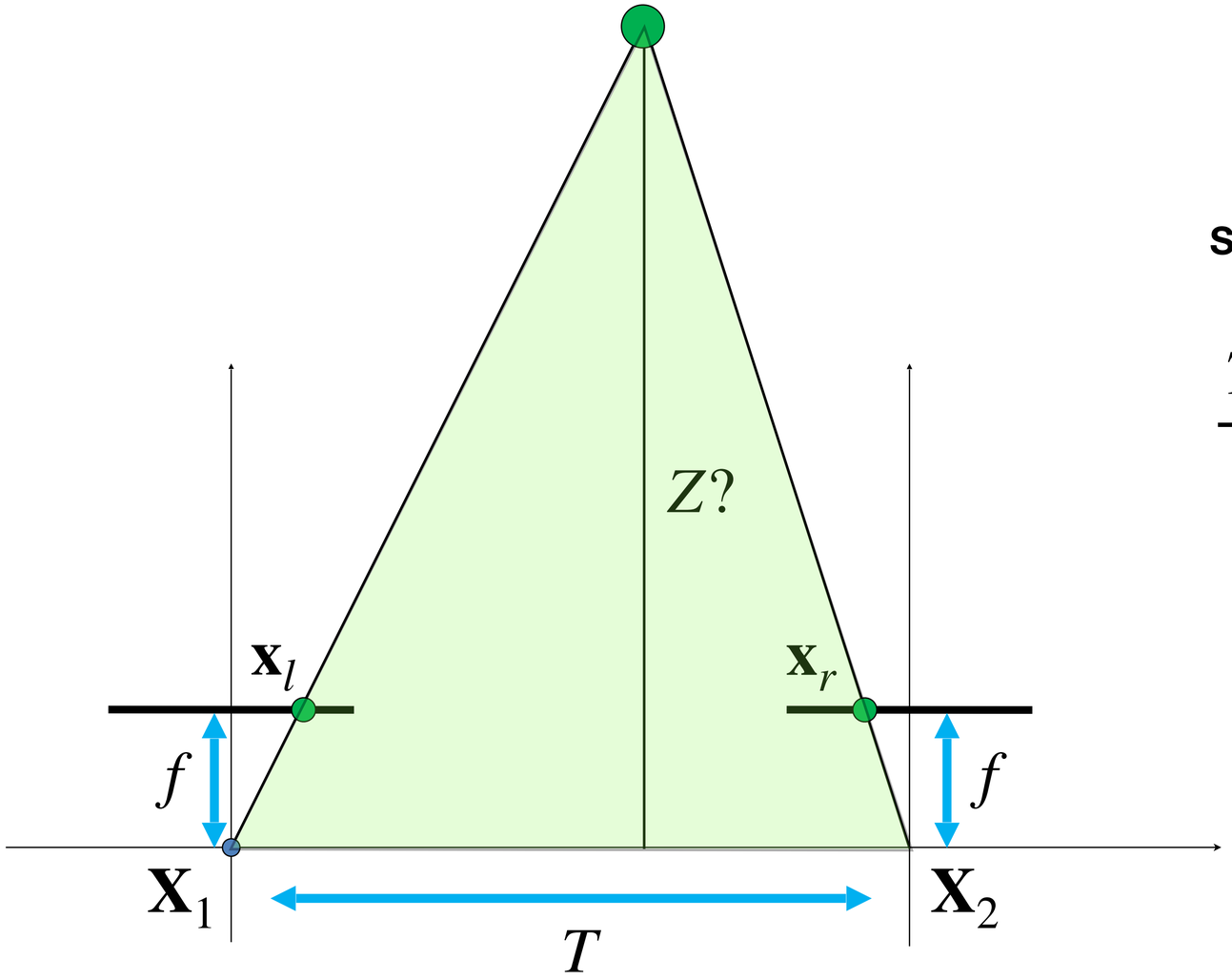
Similar Triangles:

$$\frac{T + \mathbf{x}_r - \mathbf{x}_l}{Z - f} = \frac{T}{Z}$$

Solve for Z:

$$Z = f\frac{T}{\mathbf{x}_l - \mathbf{x}_r}$$

Disparity map



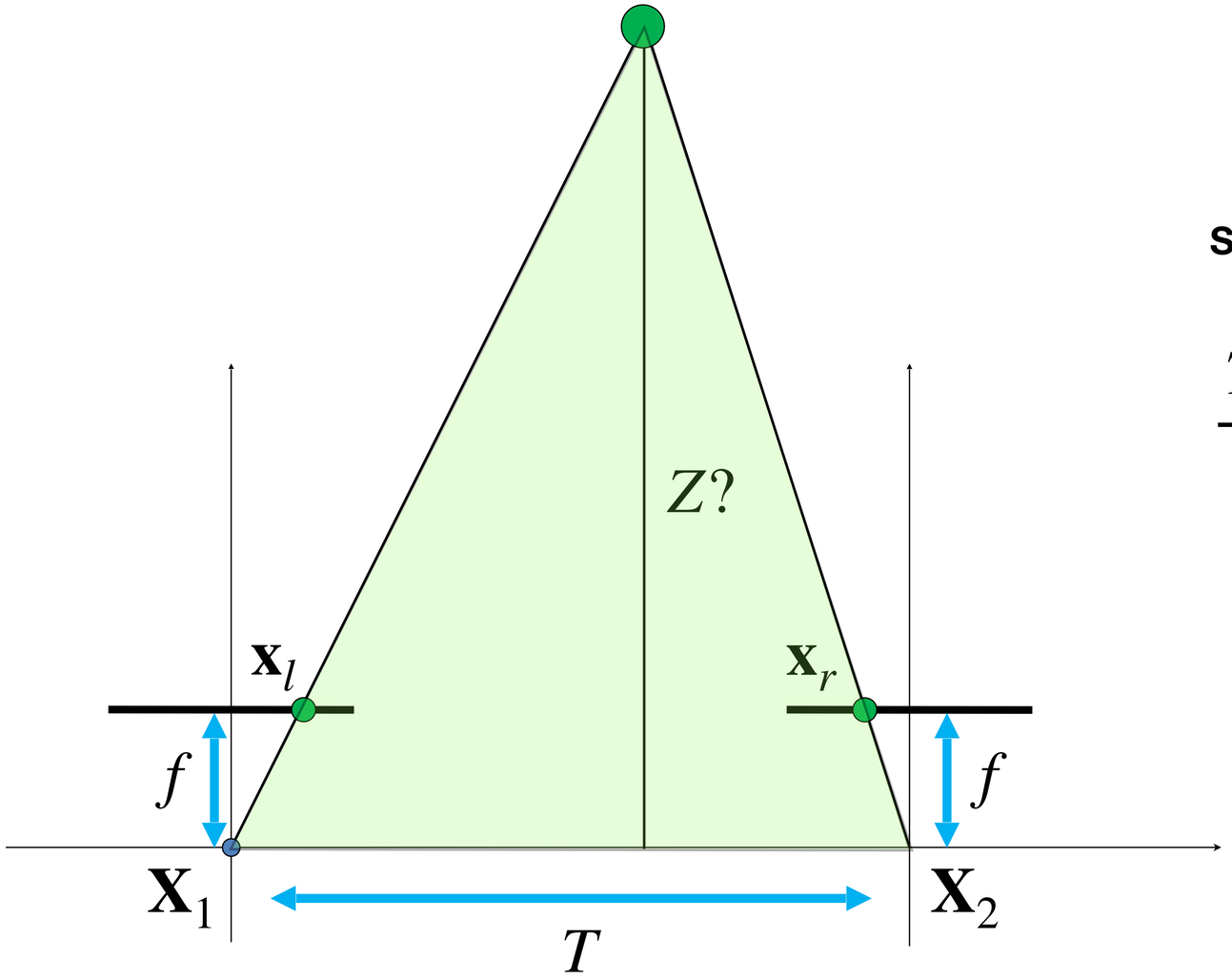
Similar Triangles:

$$\frac{T + \mathbf{x}_r - \mathbf{x}_l}{Z - f} = \frac{T}{Z}$$

Solve for Z:

$$Z = f\frac{T}{\mathbf{x}_l - \mathbf{x}_r}$$

Disparity map



Similar Triangles:

$$\frac{T + \mathbf{x}_r - \mathbf{x}_l}{Z - f} = \frac{T}{Z}$$

Solve for Z:

$$Z = f\frac{T}{\mathbf{x}_l - \mathbf{x}_r}$$

Disparity

Disparity map

$I(x, y)$

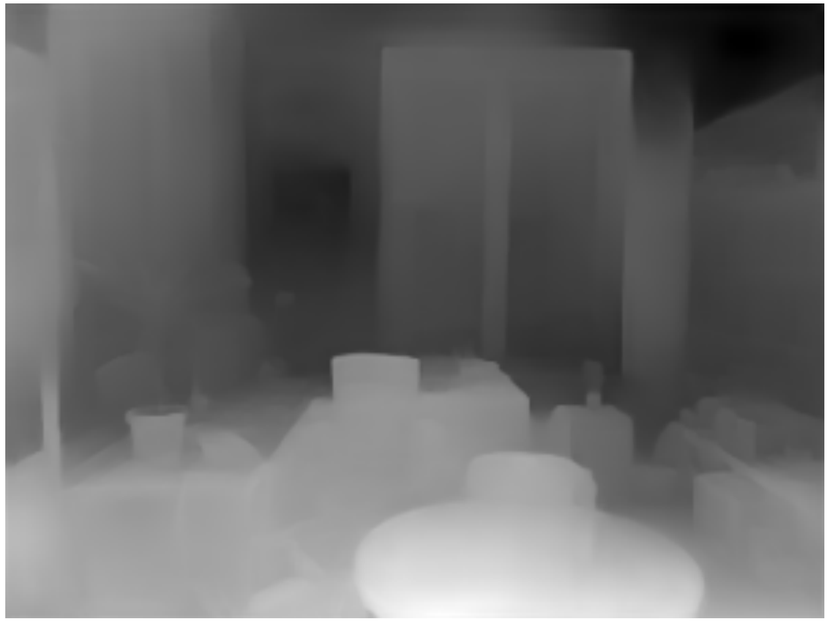
$I'(x, y) = I(x + D(x, y), y)$

Disparity map

$I(x, y)$

$I'(x, y) = I(x + D(x, y), y)$

$D(x, y)$

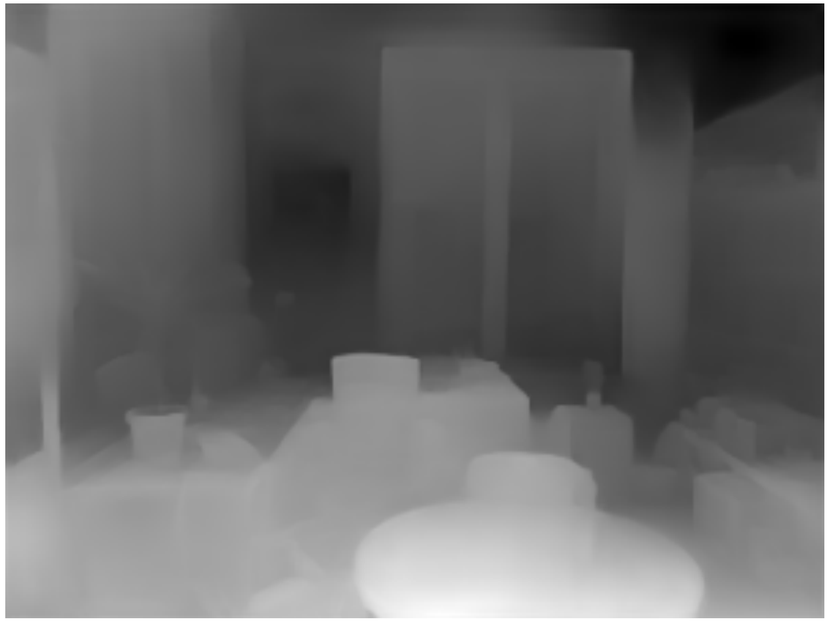


Disparity map

$I(x, y)$

$I'(x, y) = I(x + D(x, y), y)$

$D(x, y)$



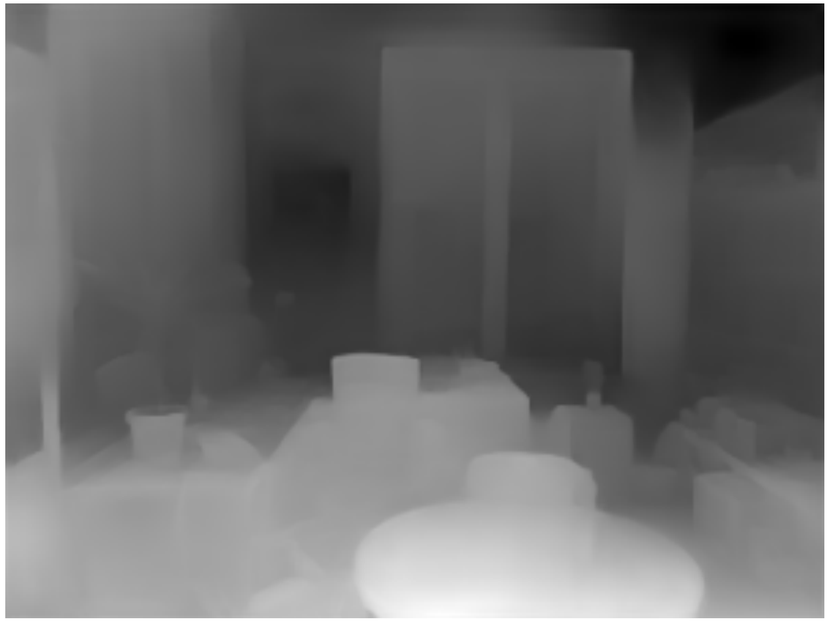
$$Z = f\frac{T}{D(x, y)}$$

Disparity map

$I(x, y)$

$I'(x, y) = I(x + D(x, y), y)$

$D(x, y)$



Depth map

$$Z = f\frac{T}{D(x, y)}$$

Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer, Ranftl et al. 2022

We define the scale- and shift-invariant loss for a single sample as

$$\mathcal{L}_{ssi}(\hat{\mathbf{d}}, \hat{\mathbf{d}}^*) = \frac{1}{2M} \sum_{i=1}^{M} \rho\left(\hat{\mathbf{d}}_i - \hat{\mathbf{d}}_i^*\right), \tag{1}$$

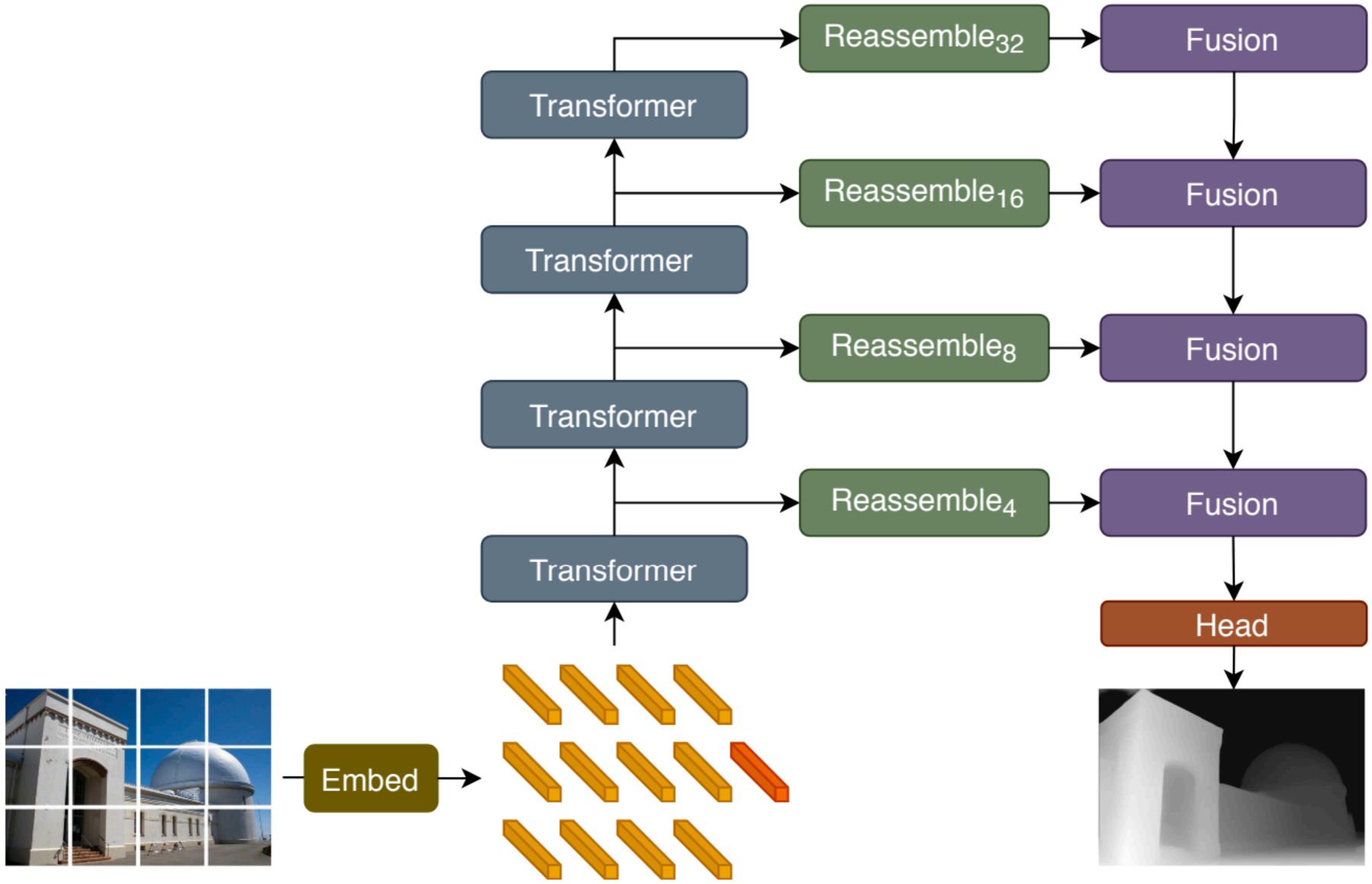
where $\hat{\mathbf{d}}$ and $\hat{\mathbf{d}}^*$ are scaled and shifted versions of the predictions and ground truth, and ρ defines the specific type of loss function.

Key idea: Solve for scale and shift for each (prediction, ground_truth) pair

$$(s,t) = \arg\min_{s,t} \sum_{i=1}^{M} \left(s\mathbf{d}_i + t - \mathbf{d}_i^*\right)^2,$$

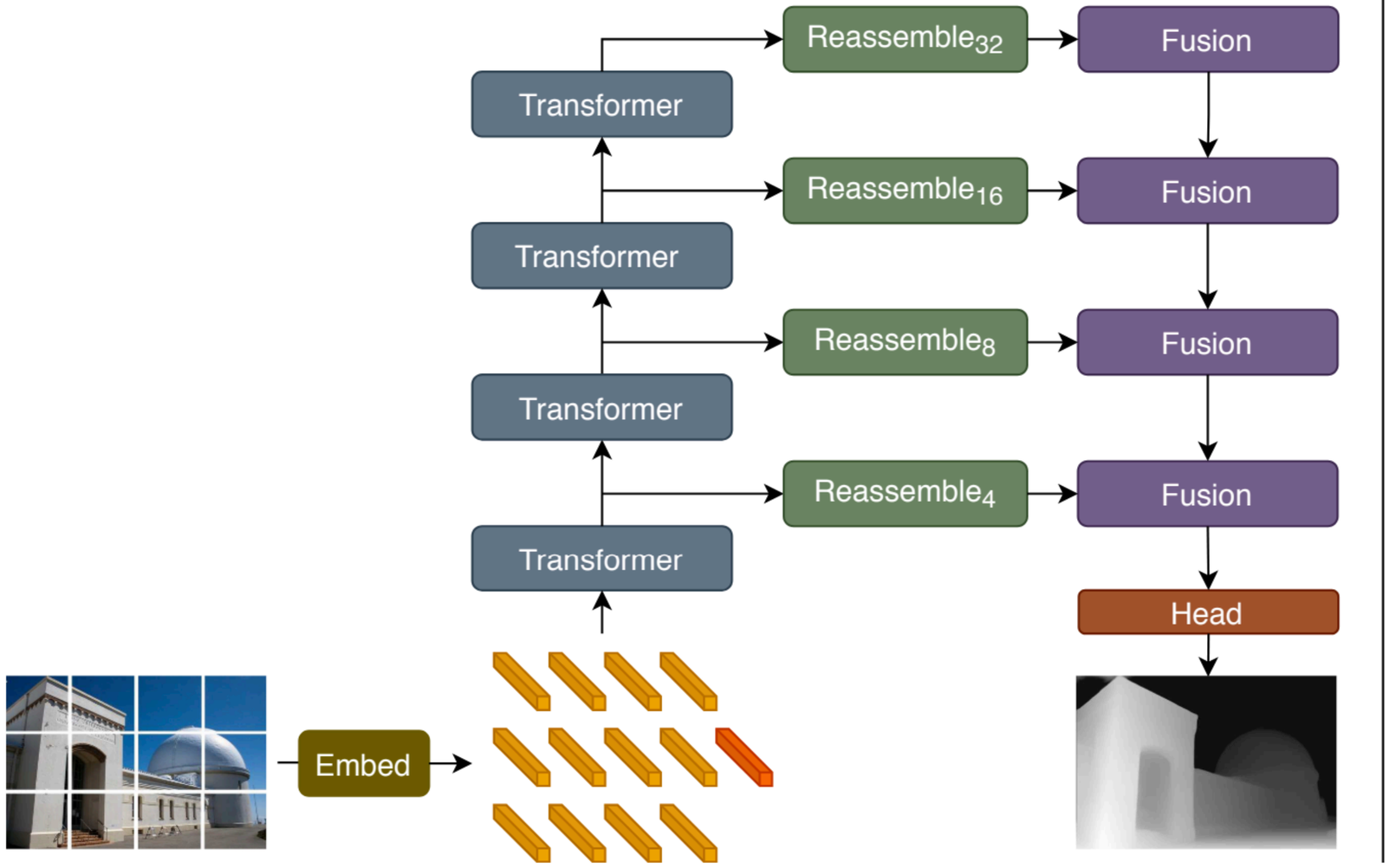
$$\hat{\mathbf{d}} = s\mathbf{d} + t, \quad \hat{\mathbf{d}}^* = \mathbf{d}^*,$$

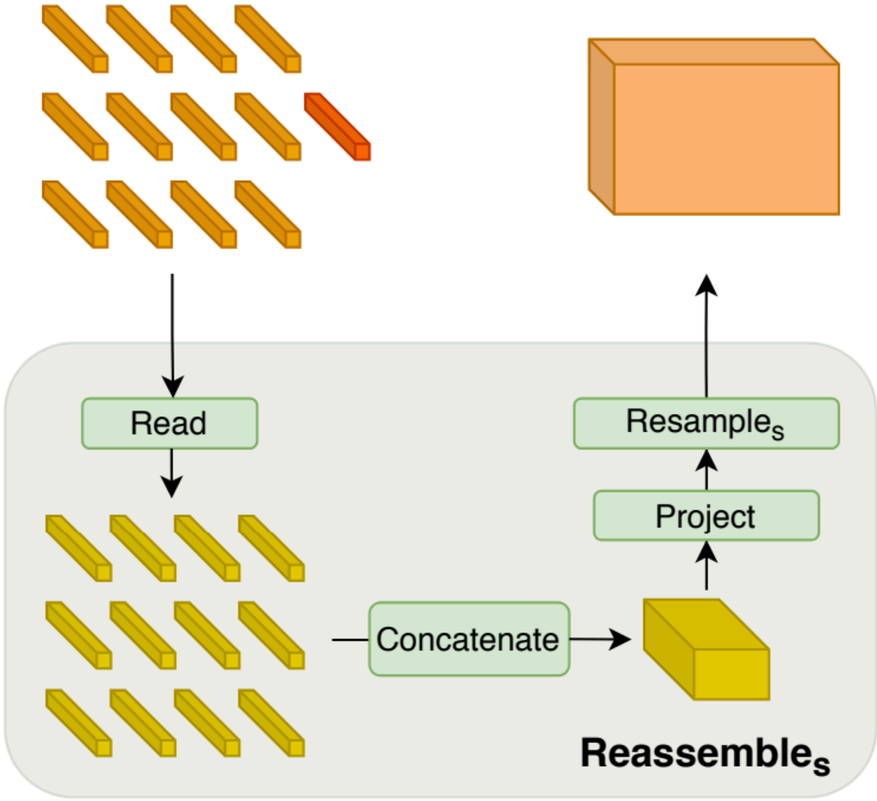
Vision Transformers for Dense Prediction, Ranftl et al. 2021



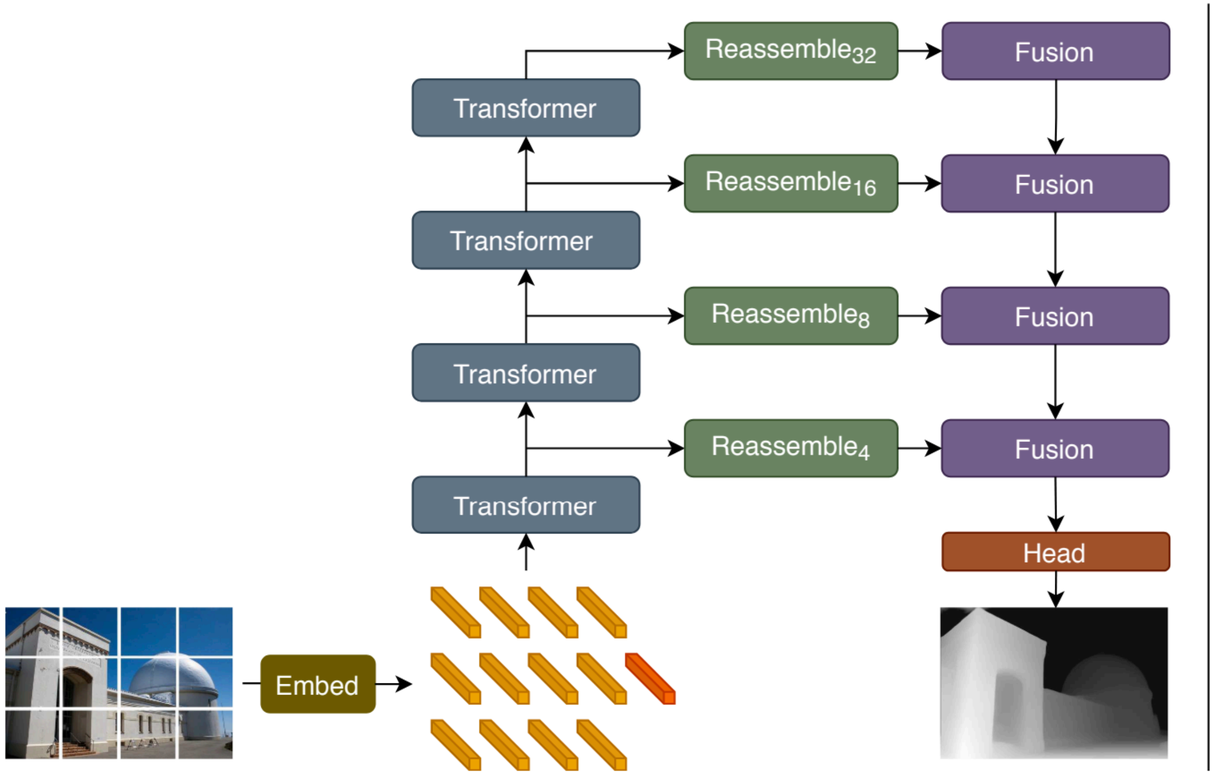
For MonoDepth:
Directly Regress Depth

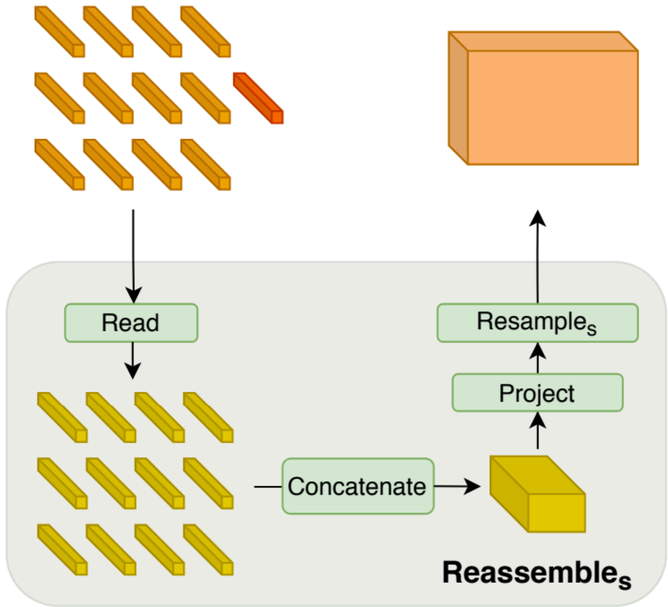
Vision Transformers for Dense Prediction, Ranftl et al. 2021

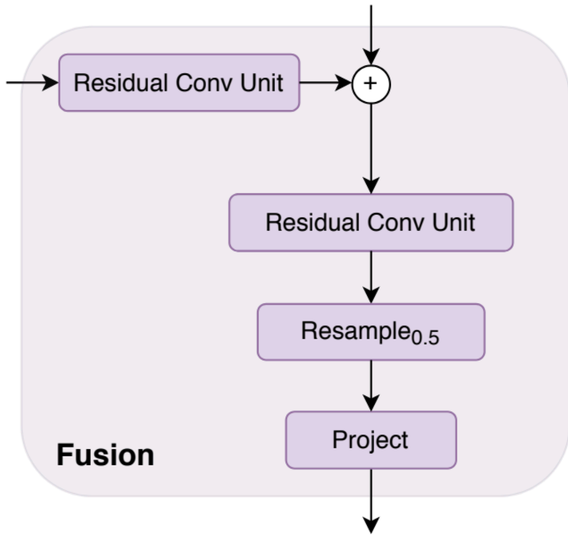


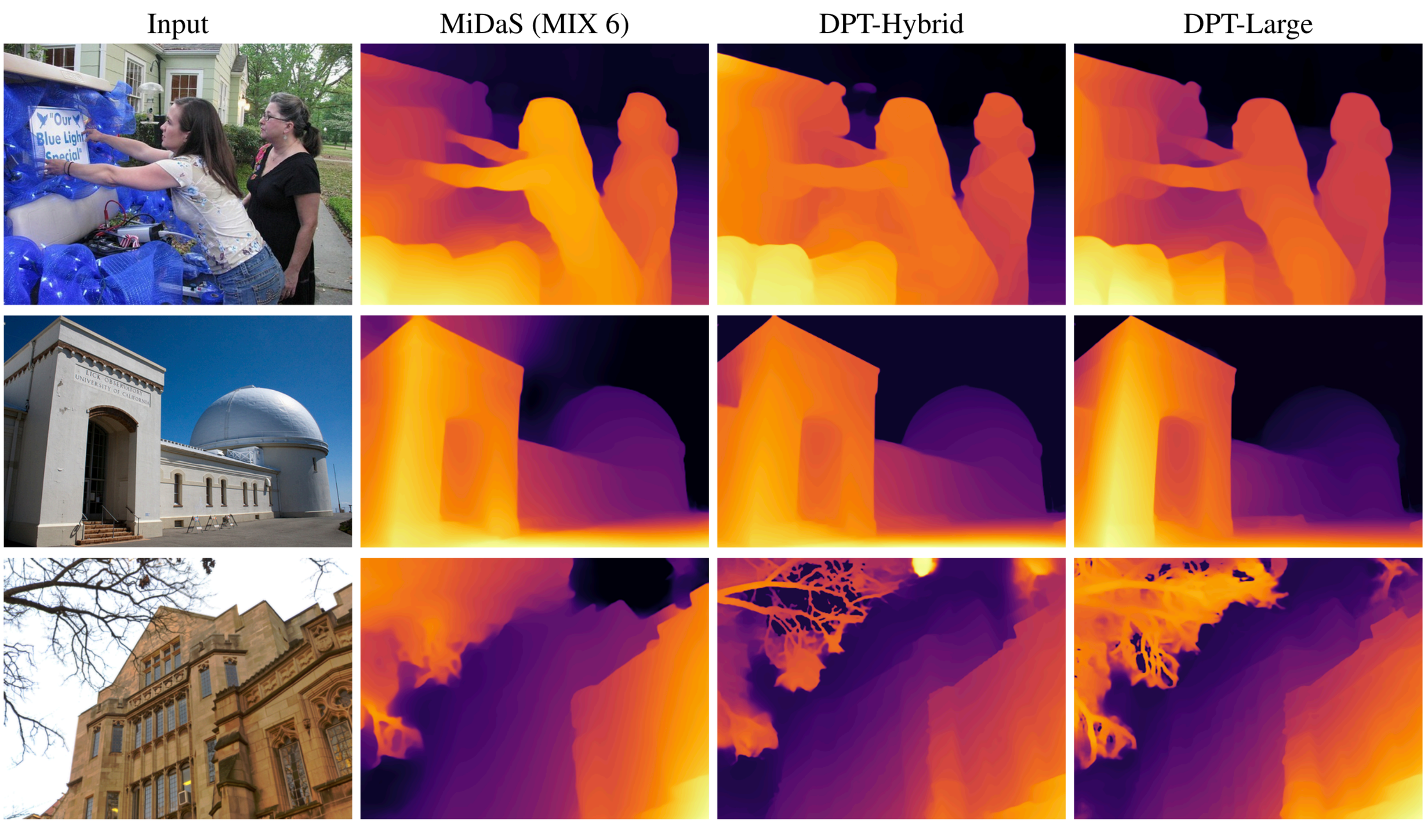


Vision Transformers for Dense Prediction, Ranftl et al. 2021

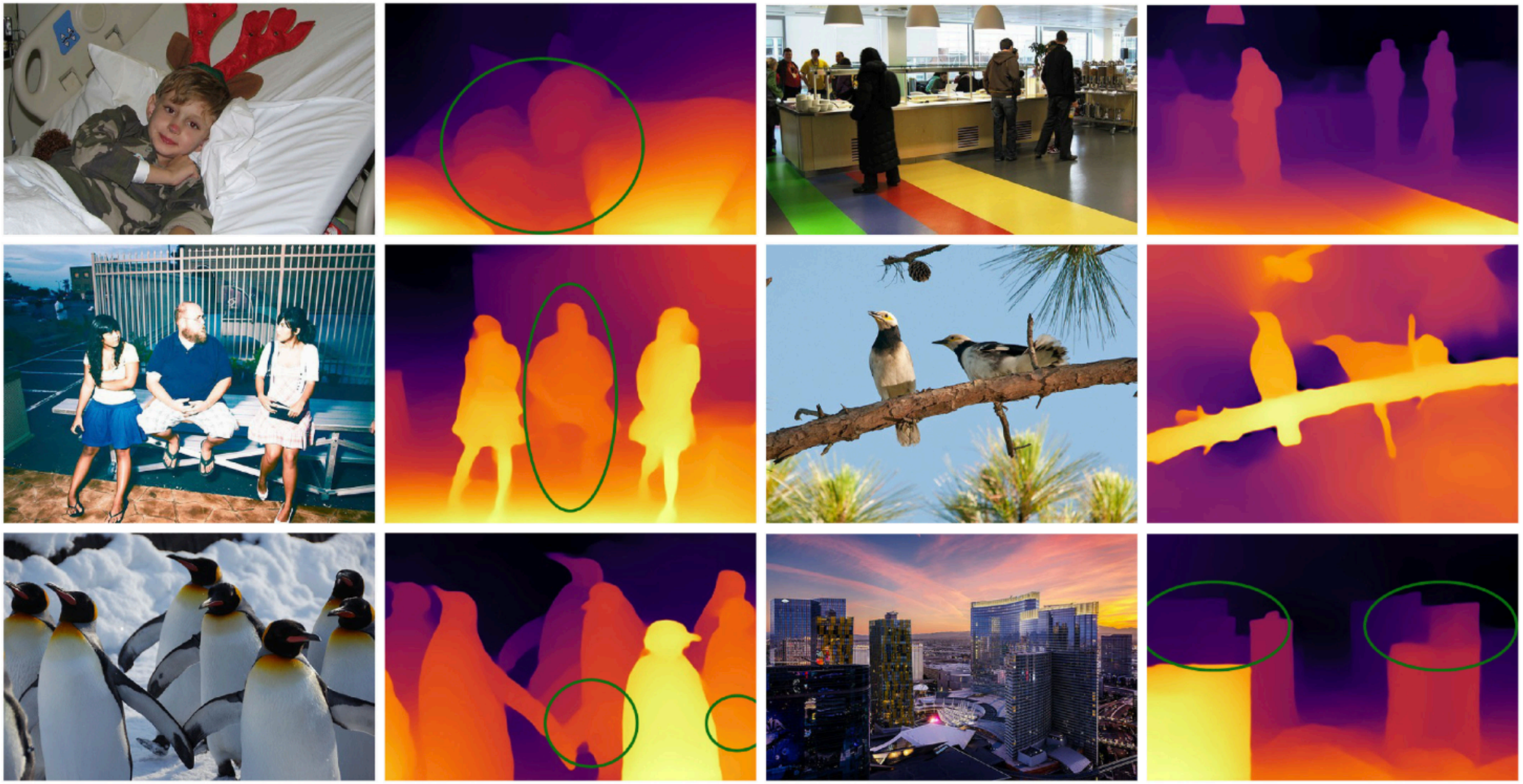







Input
MiDaS (MIX 6)
DPT-Hybrid
DPT-Large
"Our Blue Light Special"
LICK OBSERVATORY
UNIVERSITY OF CALIFORNIA

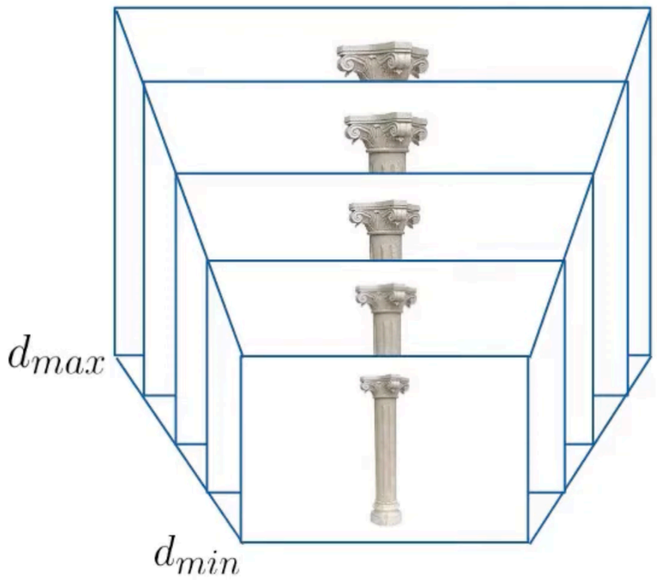
Failure Cases of Ranftl et al. 2022

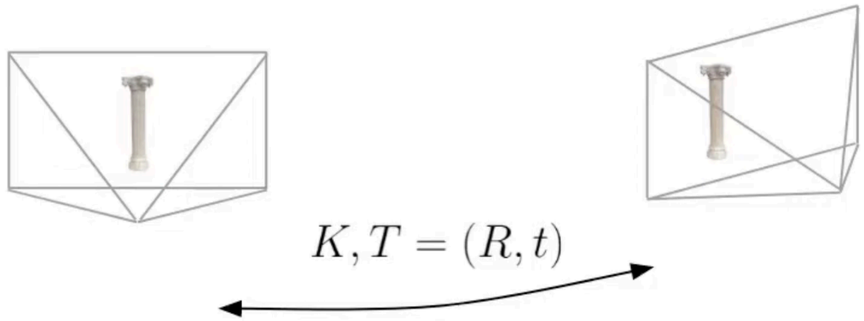


Stereo Depth Estimation

Source: https://www.youtube.com/watch?v=alw5wKyFHmo

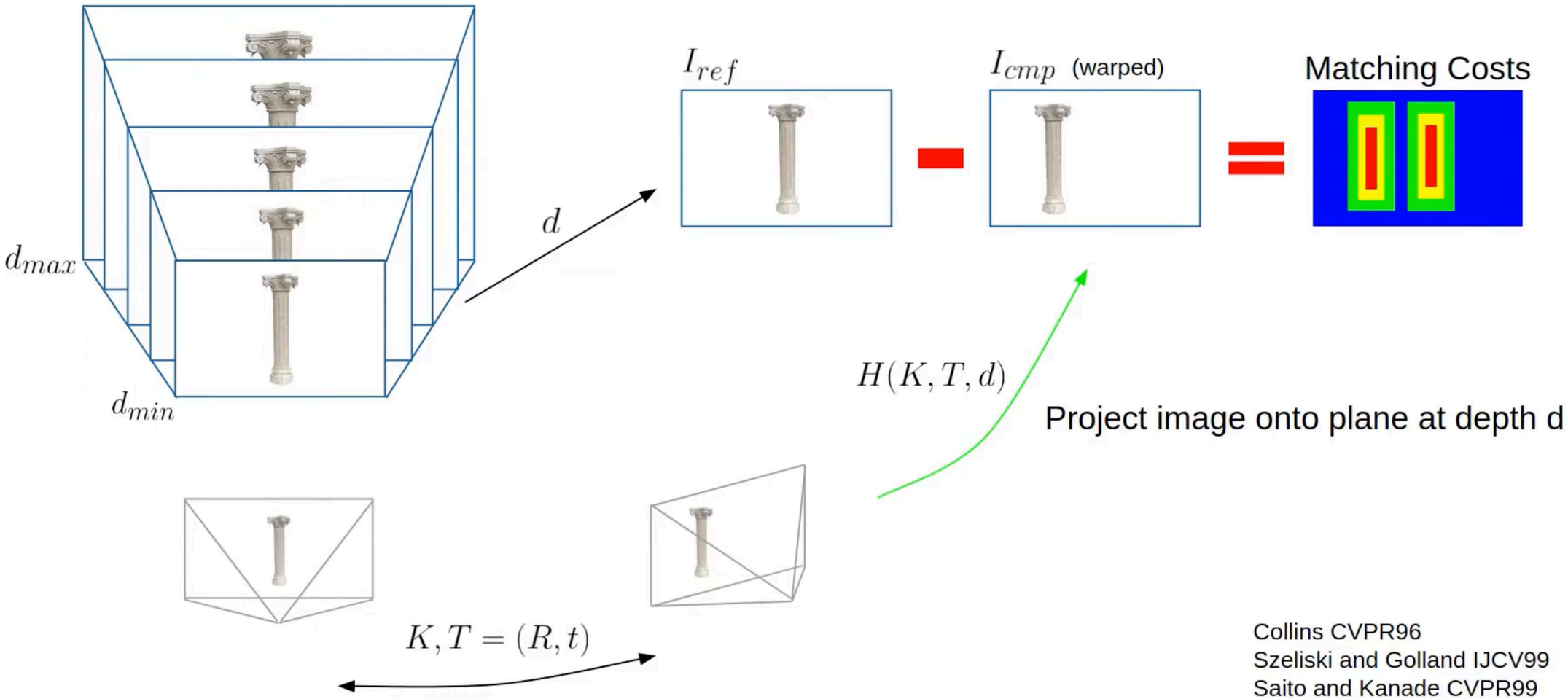
Plane Sweep Volumes





Source: https://www.youtube.com/watch?v=alw5wKyFHmo

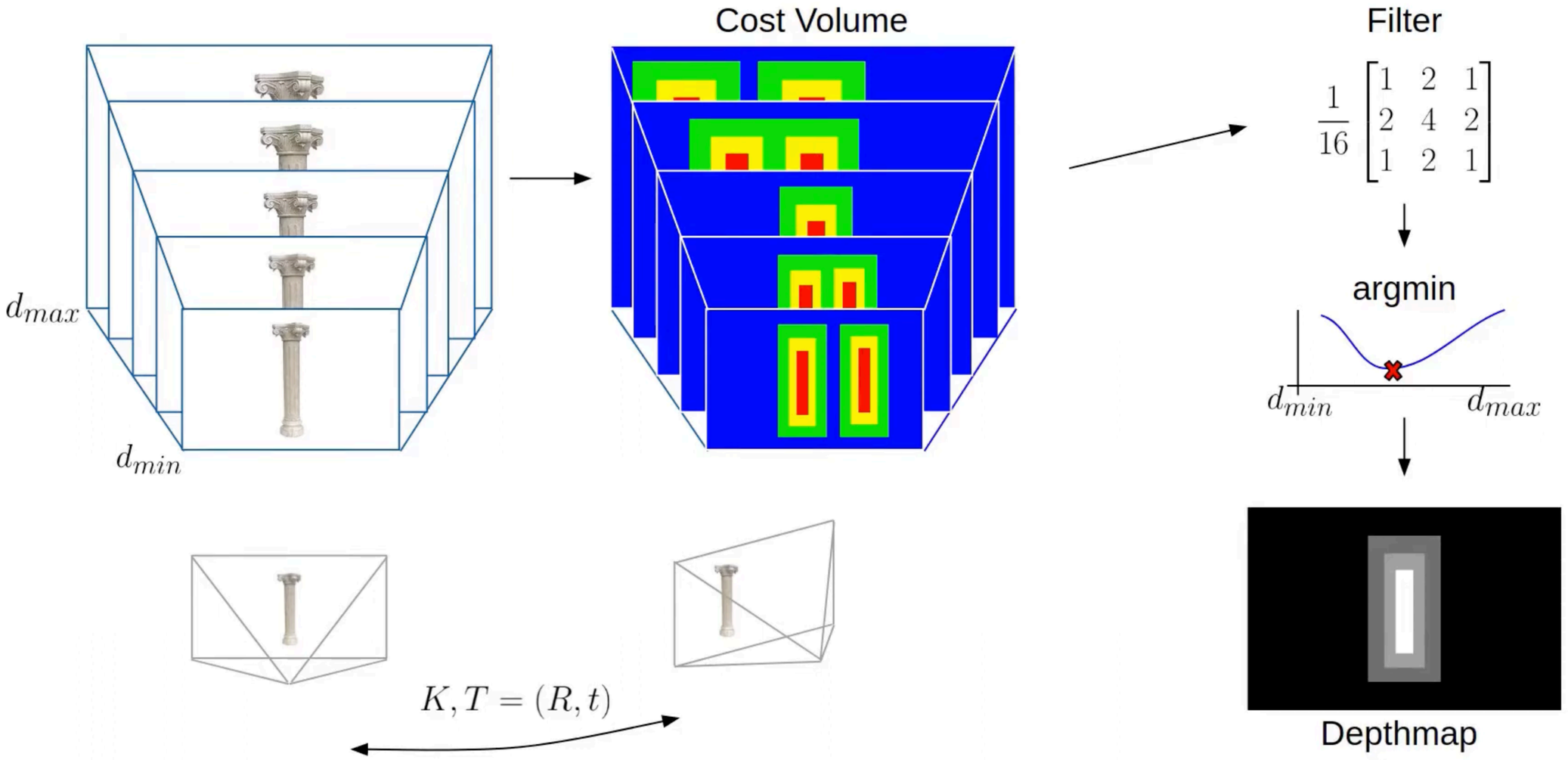
Plane Sweep Volumes



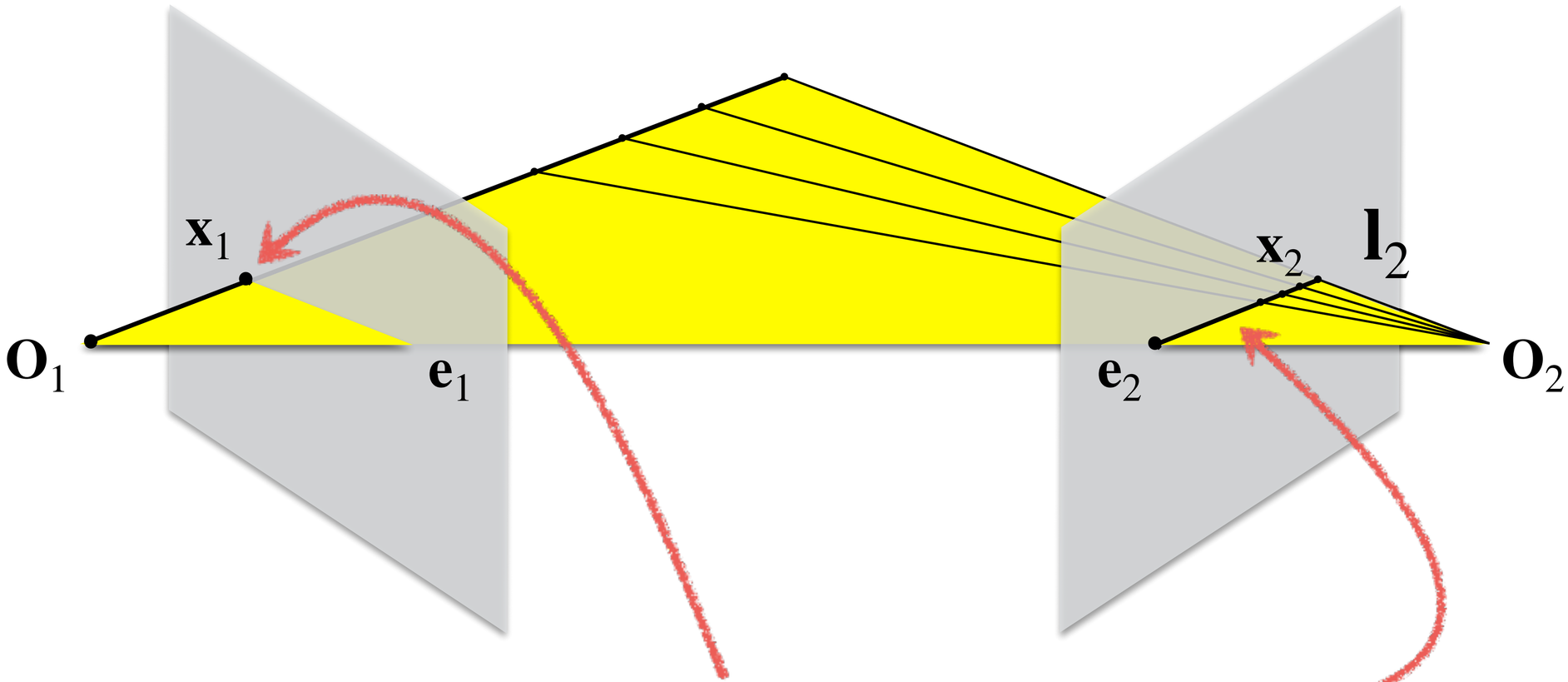
Collins CVPR96
Szeliski and Golland IJCV99
Saito and Kanade CVPR99

Source: https://www.youtube.com/watch?v=alw5wKyFHmo

Plane Sweep Volumes



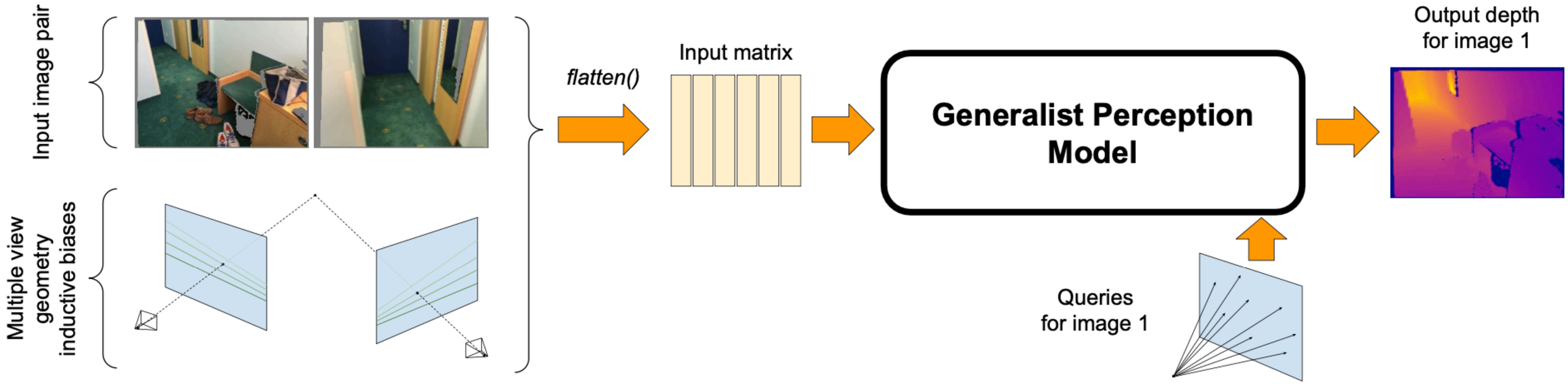
Epipolar constraint

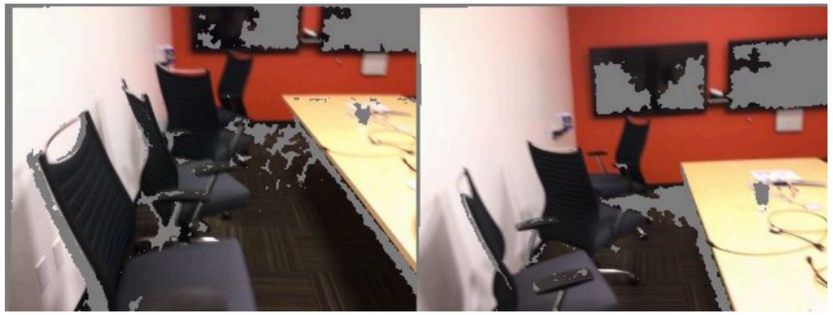


Potential matches for $\mathbf{x}_1$ lie on the epipolar line $\mathbf{l}_2$

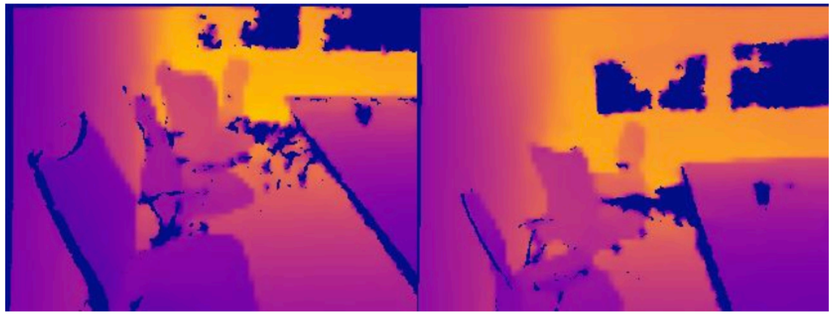
Adapted from: CMU 16-385 (Yannis, Kris)

Supervised **Stereo** Depth Estimation: Input-Level Inductive Biases for 3D Reconstruction Yifan et al. 2021

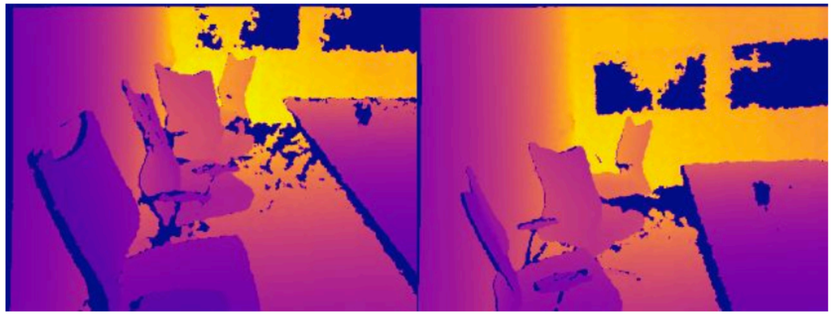




Input image pairs

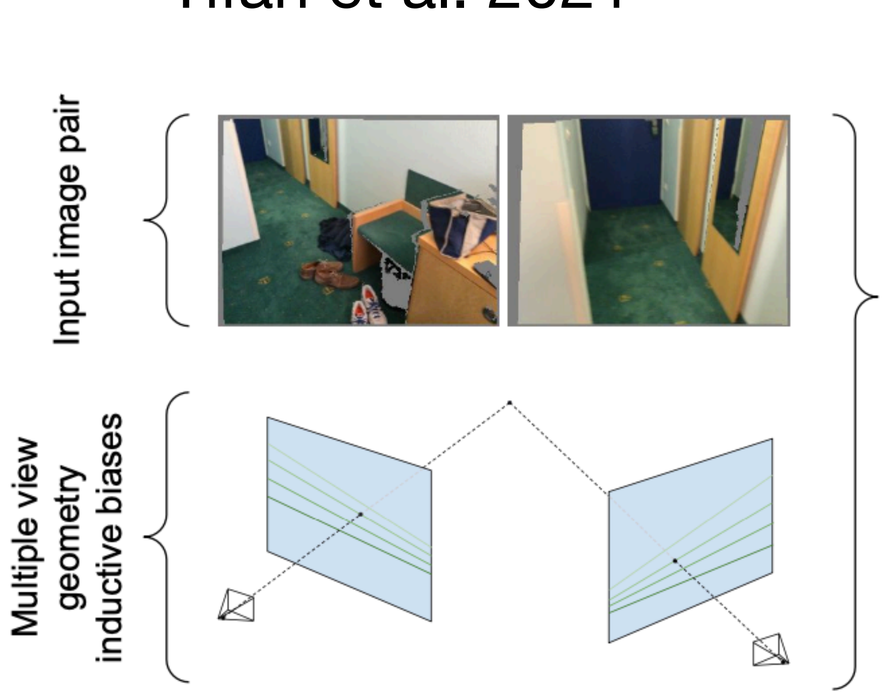


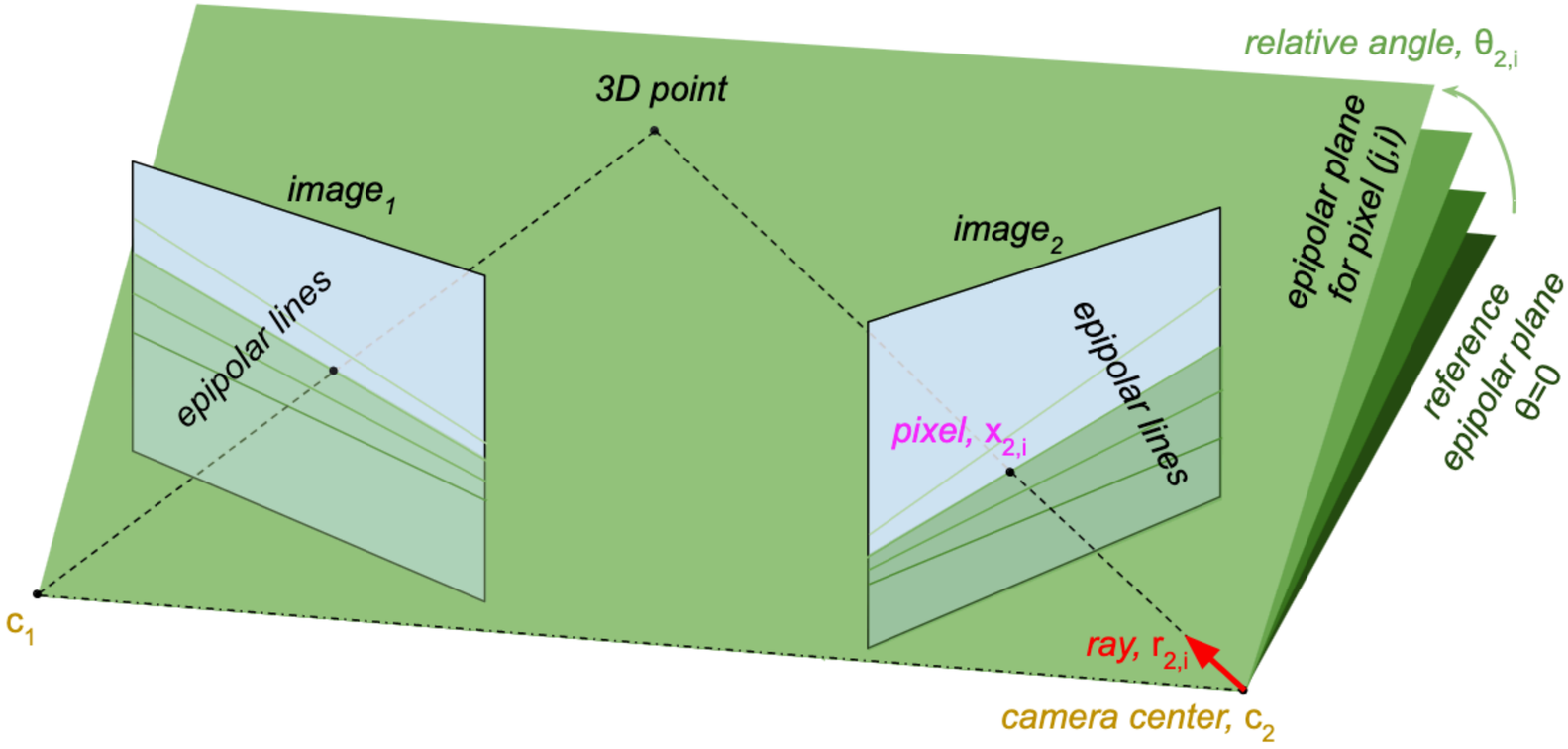
Model predictions

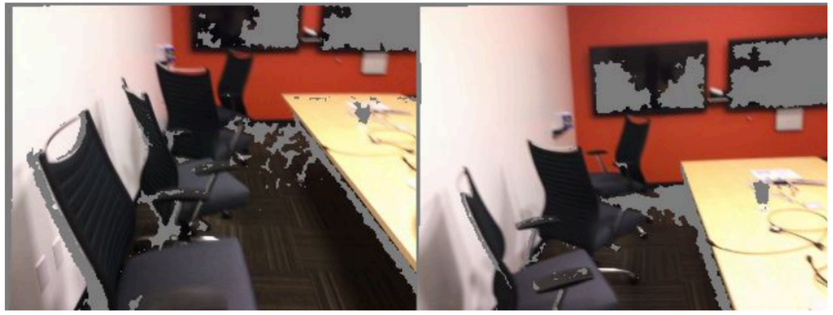


Ground truth depth maps

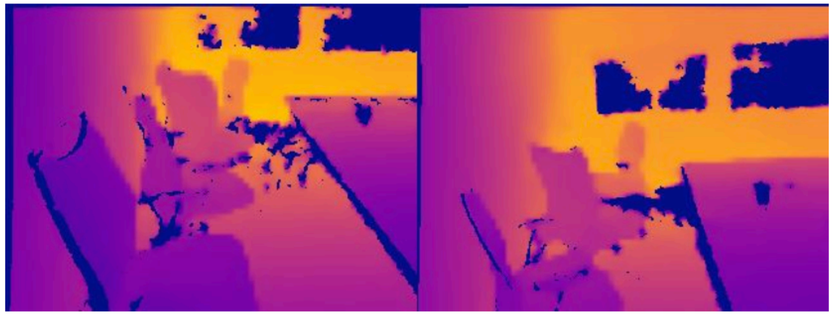
Supervised **Stereo** Depth Estimation:
Input-Level Inductive Biases for 3D Reconstruction
Yifan et al. 2021



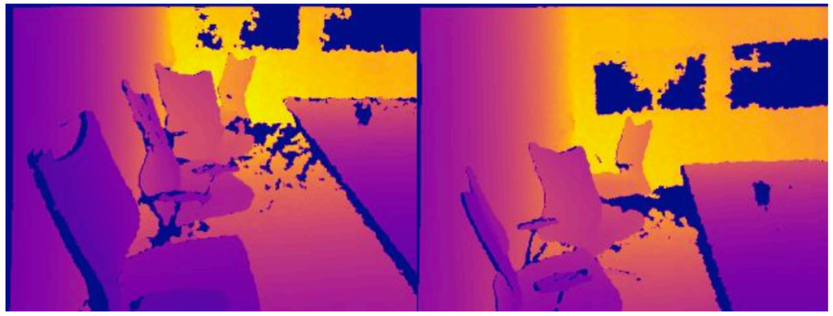




Input image pairs

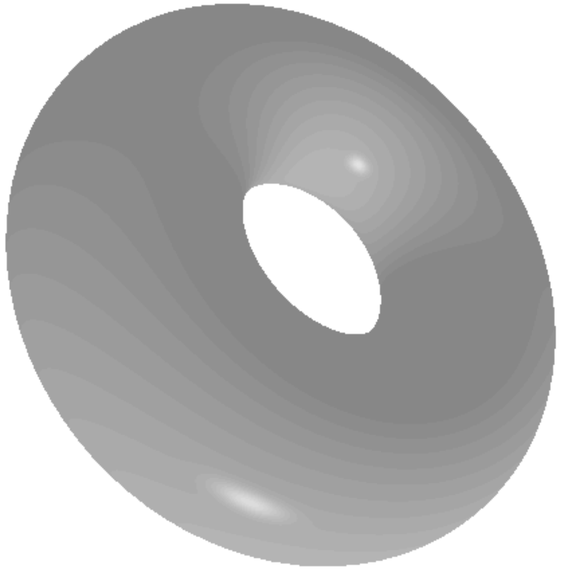


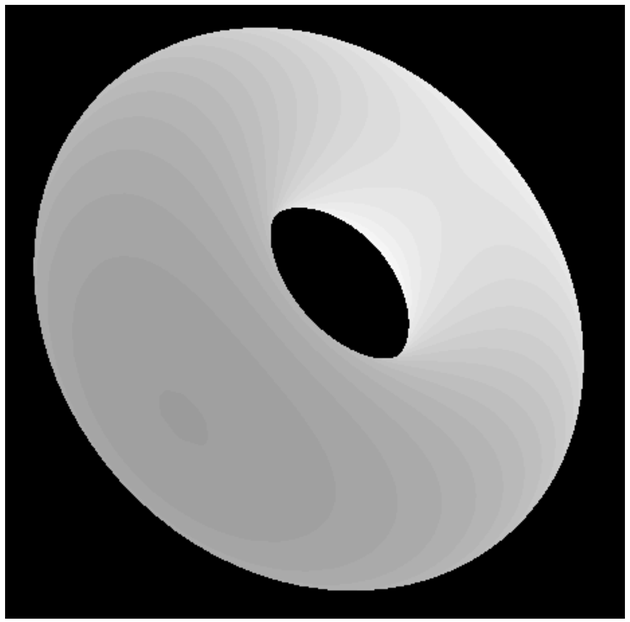
Model predictions



Ground truth depth maps

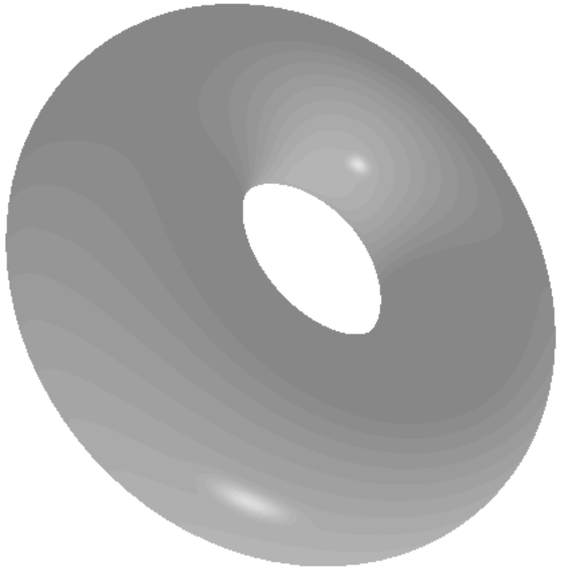
Depth Maps and Normal Maps

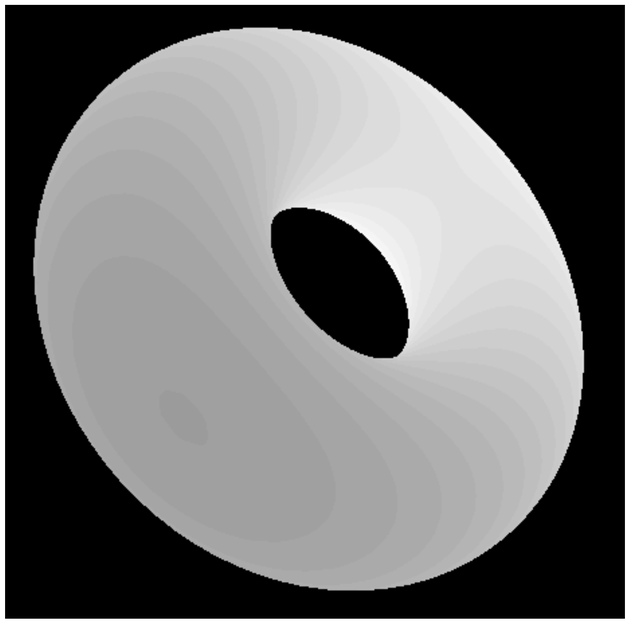


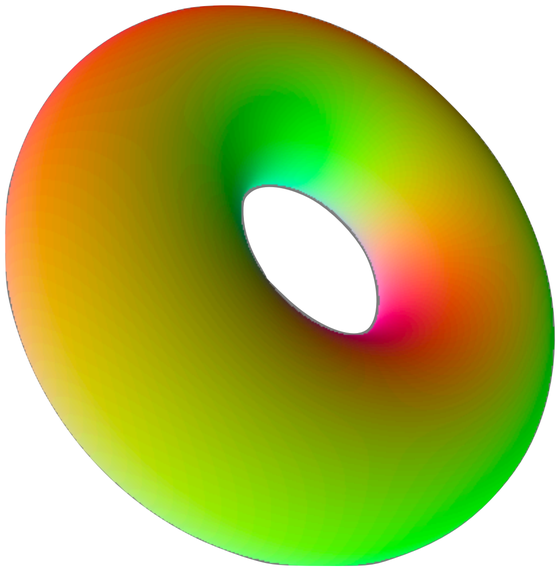


Slide credit: CMU 16-889: Learning for 3D Vision, Prof. Shubham Tulsiani

Depth Maps and Normal Maps



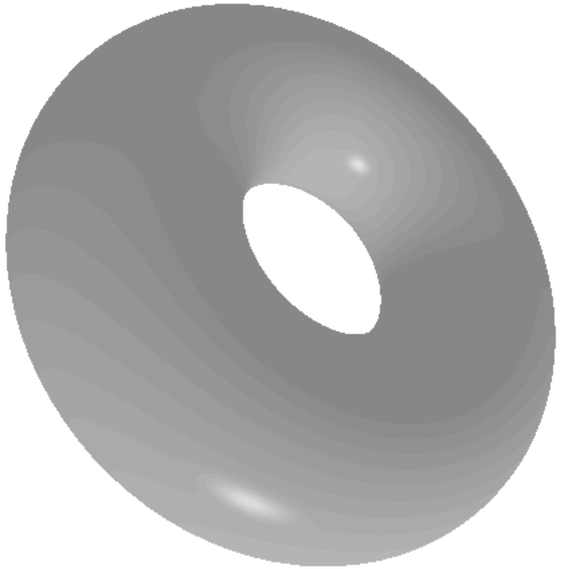


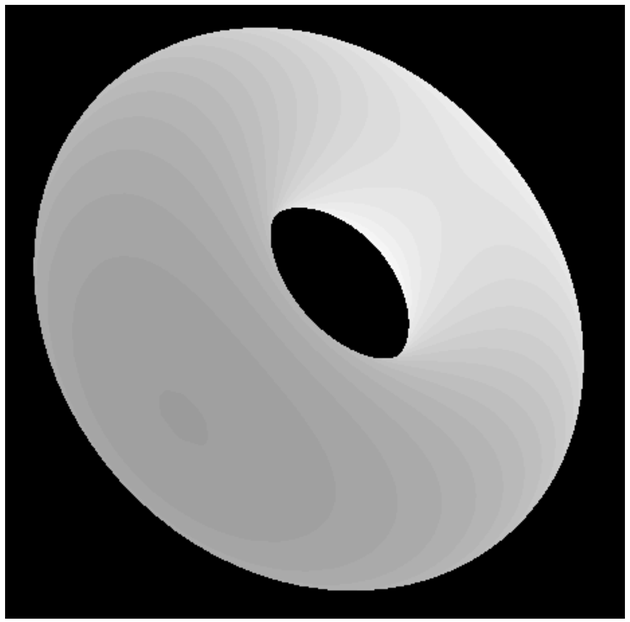


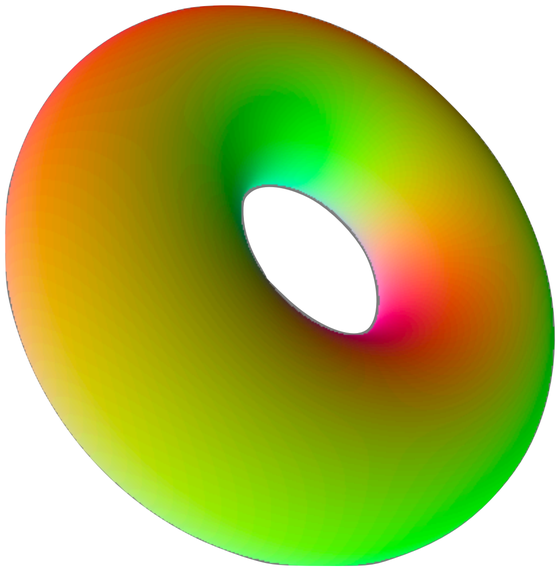
Can extend to capture other surface properties e.g. surface normals

Slide credit: CMU 16-889: Learning for 3D Vision, Prof. Shubham Tulsiani

Depth Maps and Normal Maps





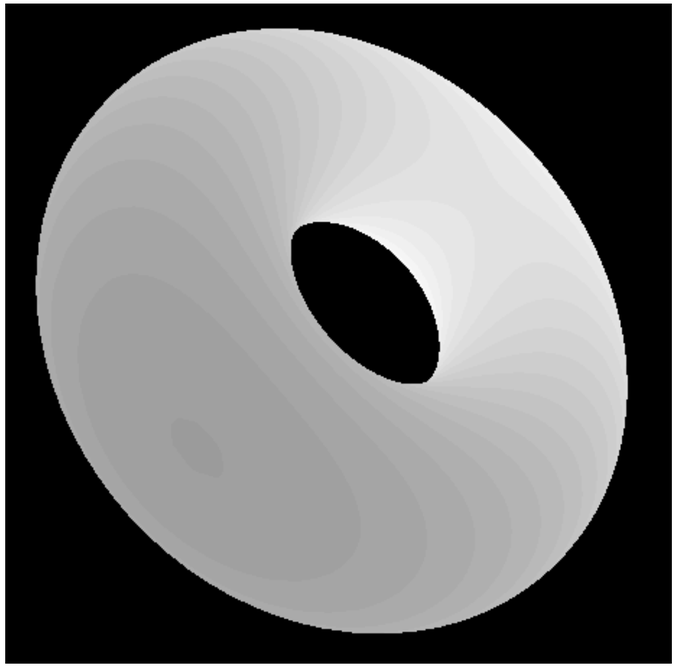


$$N[\mathbf{p}] \in \mathbb{S}^2$$

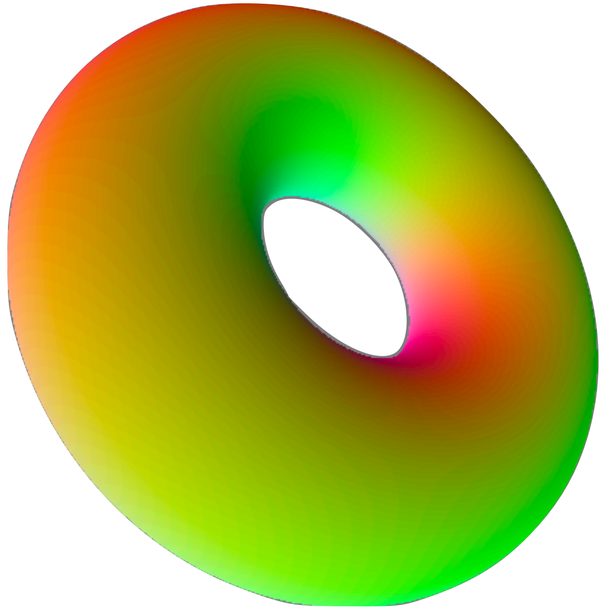
Can extend to capture other surface properties e.g. surface normals

Slide credit: CMU 16-889: Learning for 3D Vision, Prof. Shubham Tulsiani

Normal map is spatial derivative of depth map



∇_x



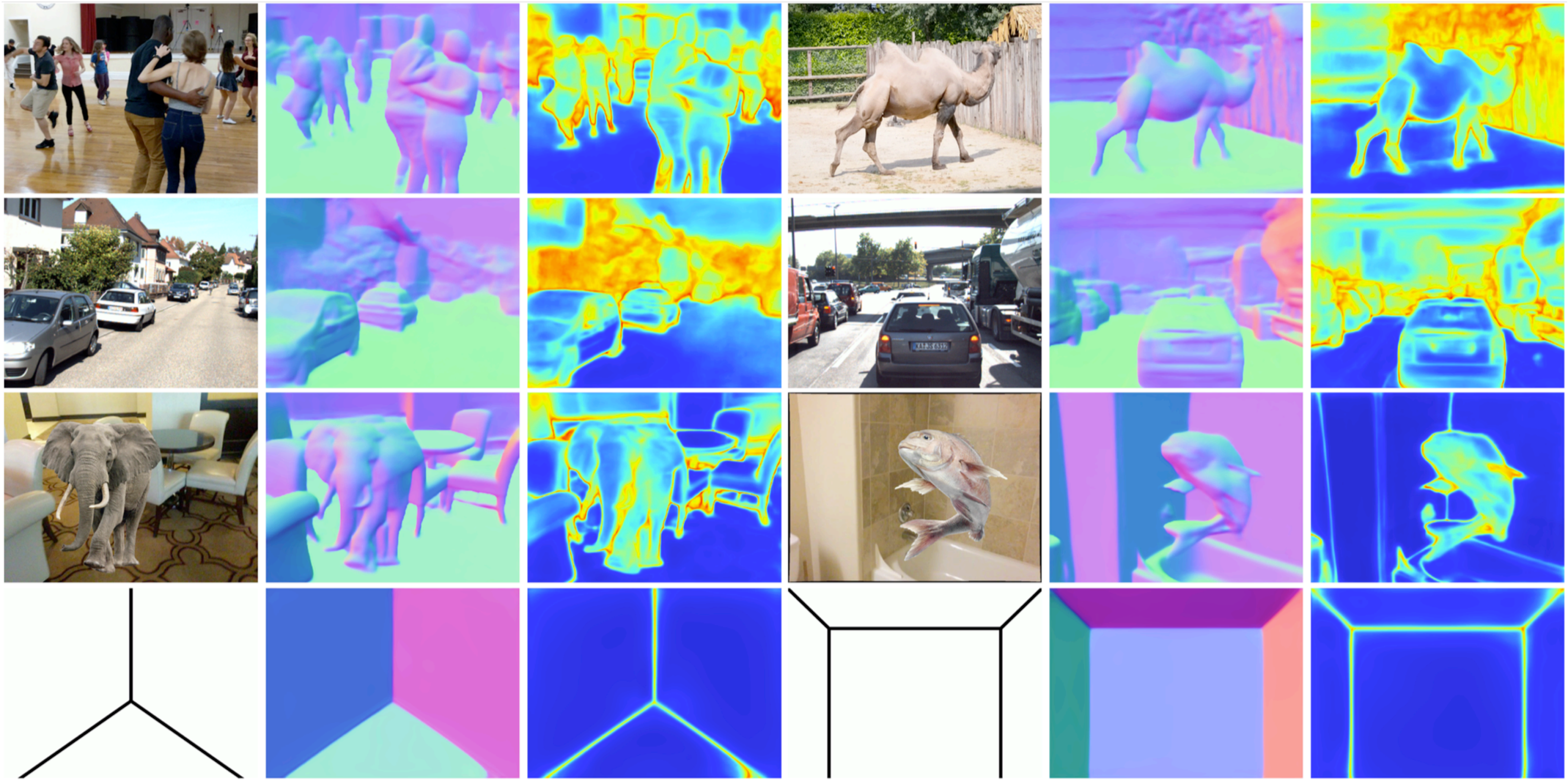
$$D : \mathbb{R}^2 \rightarrow \mathbb{R}^+$$

$$D(\mathbf{x}_{pix}) = d$$

$$N : \mathbb{R}^2 \rightarrow \mathbb{R}^3$$

$$N(\mathbf{x}_{pix}) = \nabla_{\mathbf{x}} D(\mathbf{x}_{pix}) = \mathbf{n}$$

It’s easier to predict normals from pixels than depth



Estimating and Exploiting the Aleatoric Uncertainty in Surface Normal Estimation, Bae et al., ICCV 2021

Problem: Depth not *metric*

ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth

Shariq Farooq Bhat
KAUST

Reiner Birkl
Intel

Diana Wofk
Intel

Peter Wonka
KAUST

Matthias Müller
Intel

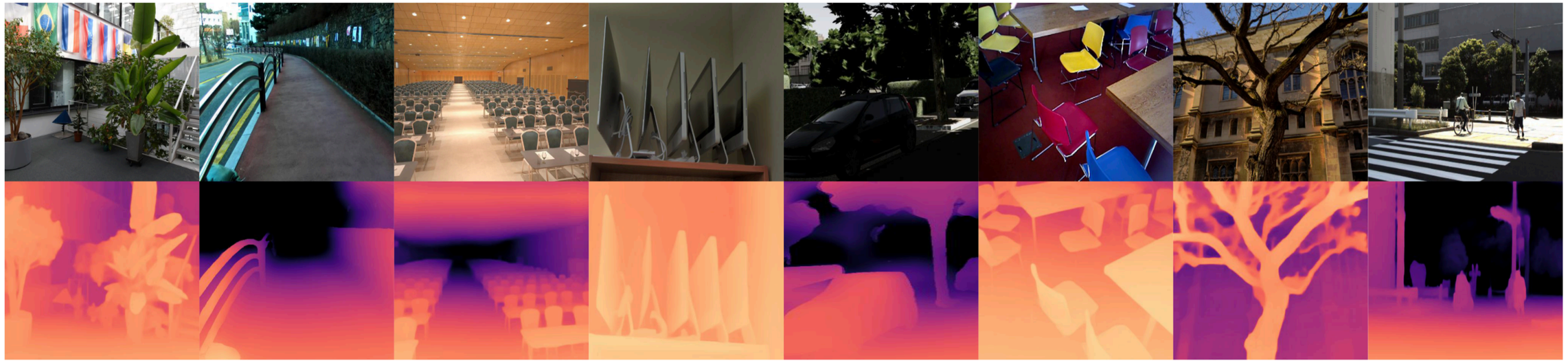


Figure 1. **Zero-shot transfer.** Our single multi-domain metric depth estimation model can be applied across domains, indoor or outdoor, simulated or real. **Top:** Input RGB. **Bottom:** Predicted depth. **From left to right:** iBims-1, DIML Outdoor, Hypersim, DIODE Indoor, vKITTI2, SUN-RGBD, DIODE Outdoor and DDAD.

Learning the Depths of Moving People by Watching Frozen People

Zhengqi Li Tali Dekel Forrester Cole Richard Tucker
Noah Snavely Ce Liu William T. Freeman

Google Research

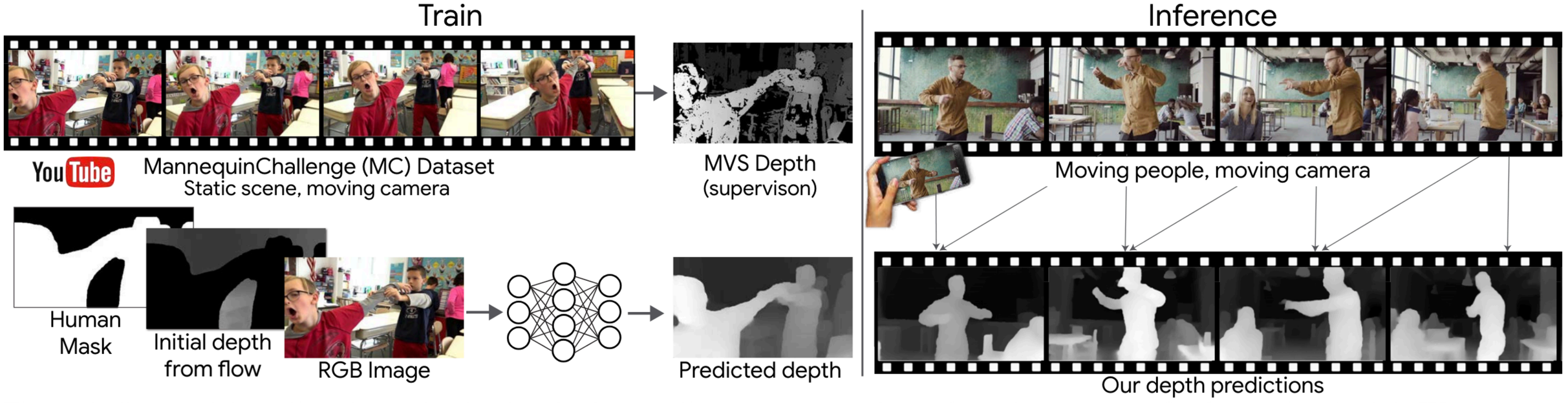
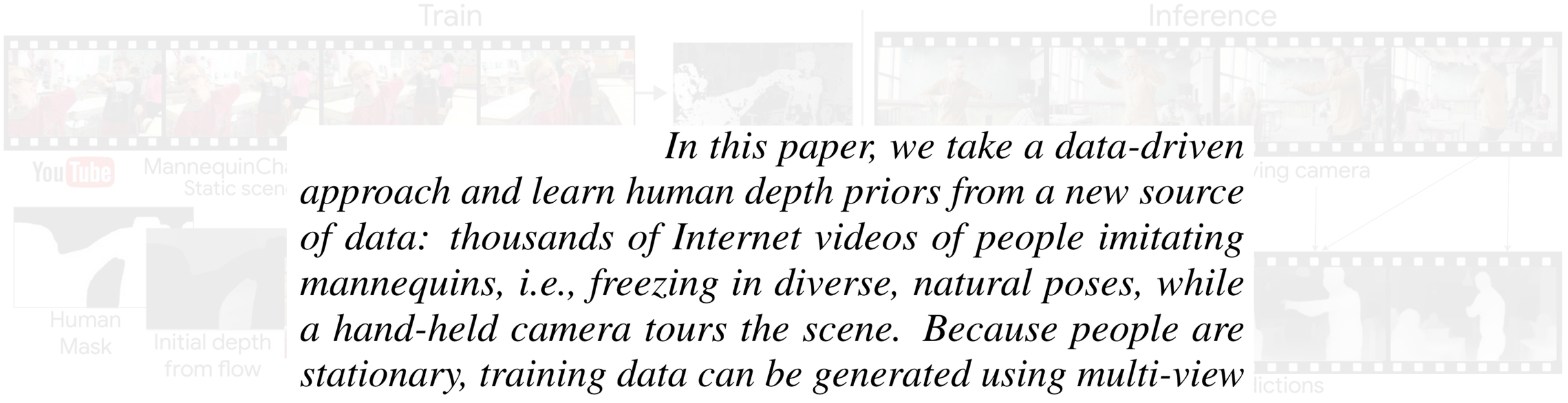


Figure 1. Our model predicts dense depth when both an ordinary camera and people in the scene are freely moving (right). We train our model on our new *MannequinChallenge* dataset—a collection of Internet videos of people imitating mannequins, i.e., freezing in diverse, natural poses, while a camera tours the scene (left). Because people are *stationary*, geometric constraints hold; this allows us to use multi-view stereo to estimate depth which serves as supervision during training.[2]

Learning the Depths of Moving People by Watching Frozen People

Zhengqi Li Tali Dekel Forrester Cole Richard Tucker
Noah Snavely Ce Liu William T. Freeman
Google Research

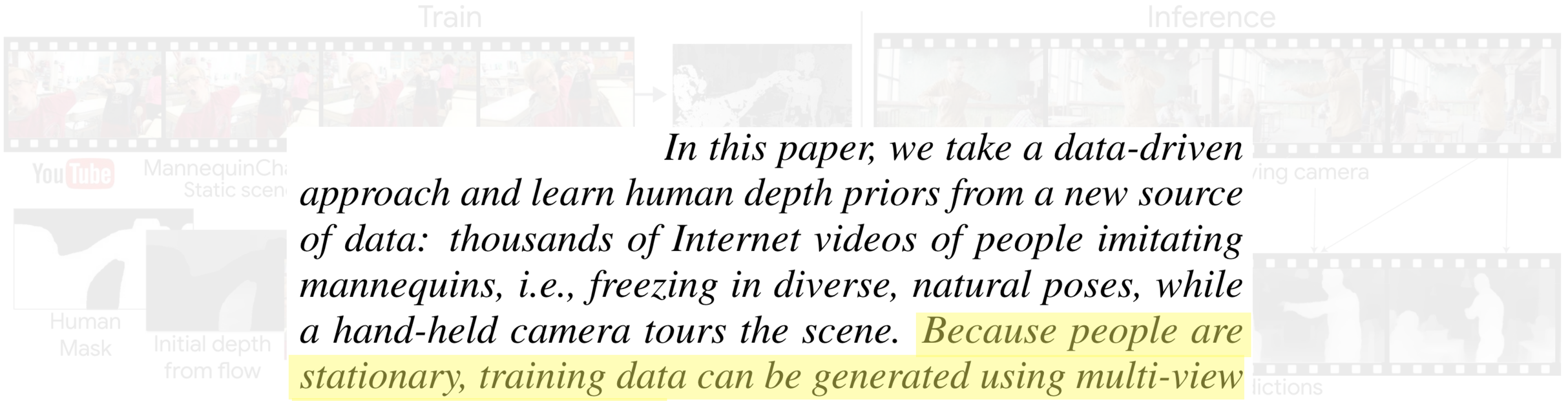


In this paper, we take a data-driven approach and learn human depth priors from a new source of data: thousands of Internet videos of people imitating mannequins, i.e., freezing in diverse, natural poses, while a hand-held camera tours the scene. Because people are stationary, training data can be generated using multi-view stereo reconstruction.

Figure 1. Our model pred ioving (right). We train our model on our new *MannequinChallenge* dataset—a collection of Internet videos of people imitating mannequins, i.e., freezing in diverse, natural poses, while a camera tours the scene (left). Because people are *stationary*, geometric constraints hold; this allows us to use multi-view stereo to estimate depth which serves as supervision during training.[2]

Learning the Depths of Moving People by Watching Frozen People

Zhengqi Li Tali Dekel Forrester Cole Richard Tucker
Noah Snavely Ce Liu William T. Freeman
Google Research



In this paper, we take a data-driven approach and learn human depth priors from a new source of data: thousands of Internet videos of people imitating mannequins, i.e., freezing in diverse, natural poses, while a hand-held camera tours the scene. Because people are stationary, training data can be generated using multi-view stereo reconstruction.

Figure 1. Our model pred … oving (right). We train our model on our new *MannequinChallenge* dataset—a collection of Internet videos of people imitating mannequins, i.e., freezing in diverse, natural poses, while a camera tours the scene (left). Because people are *stationary*, geometric constraints hold; this allows us to use multi-view stereo to estimate depth which serves as supervision during training.[2]

If Structure-from-Motion were *solved*, video would be a great data source…

Unsupervised Depth Prediction!

Frame at time t_1

...

Frame at time t_2

Goal: Learn Depth and Ego-Motion (relative camera pose) just from video!

Unsupervised Depth Prediction!

Frame at time t_1

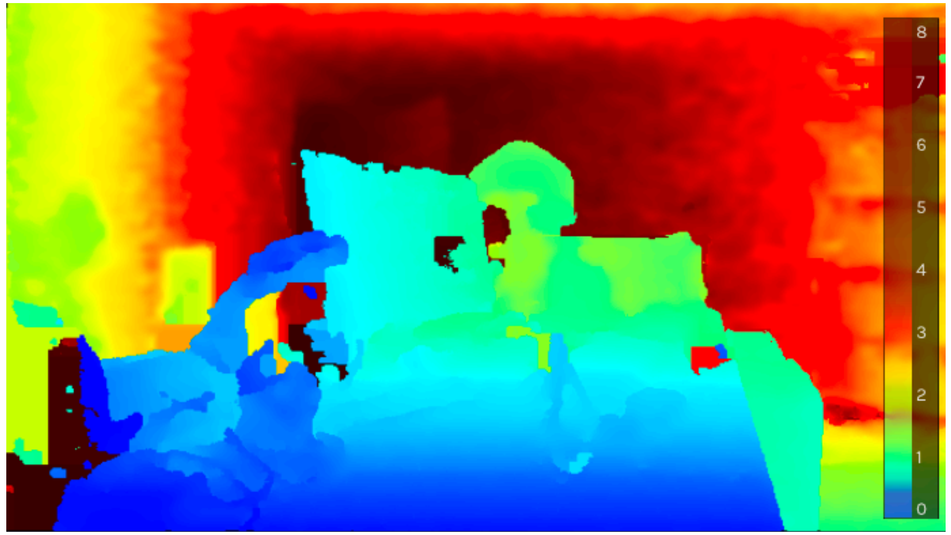
...

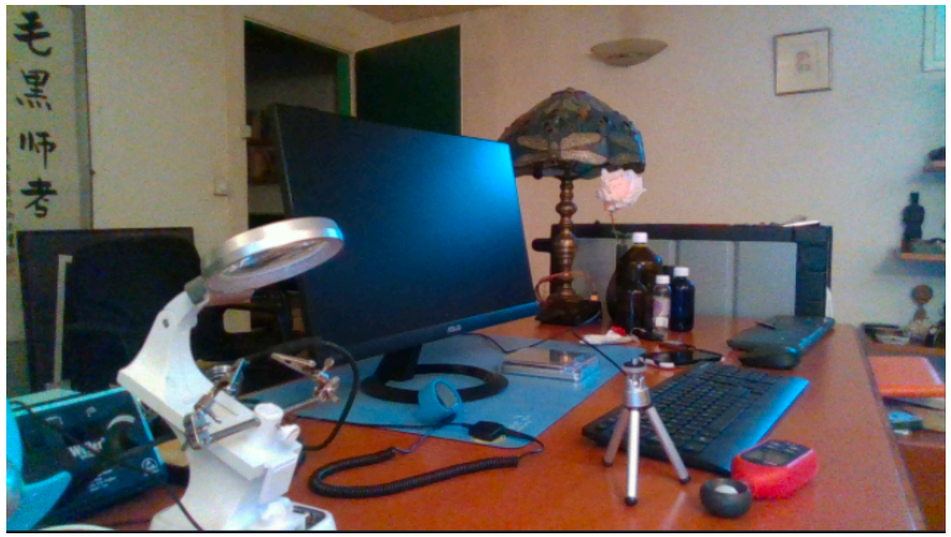
Frame at time t_2

Goal: Learn Depth and Ego-Motion (relative camera pose) just from video!

How?

“Unproject”: Go from pixel and depth to 3D point





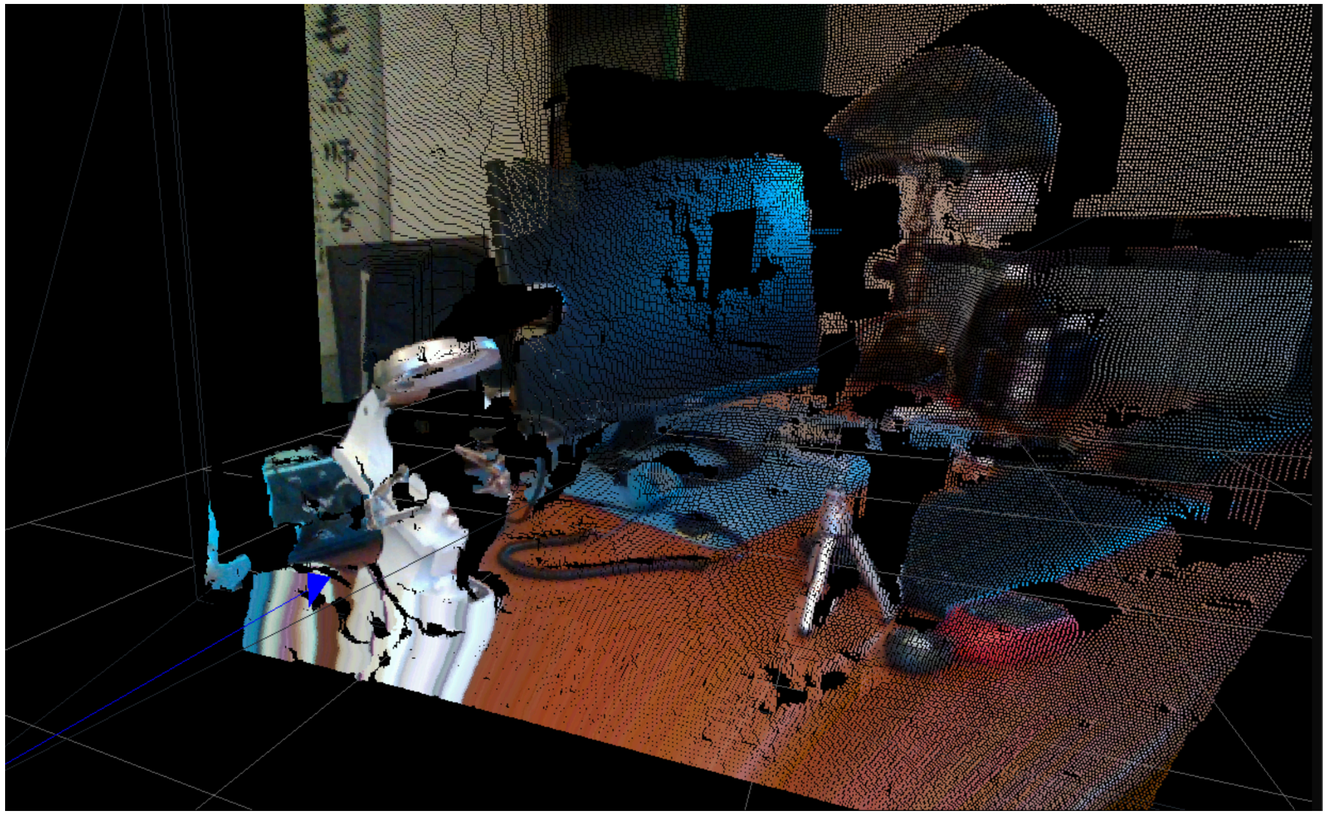


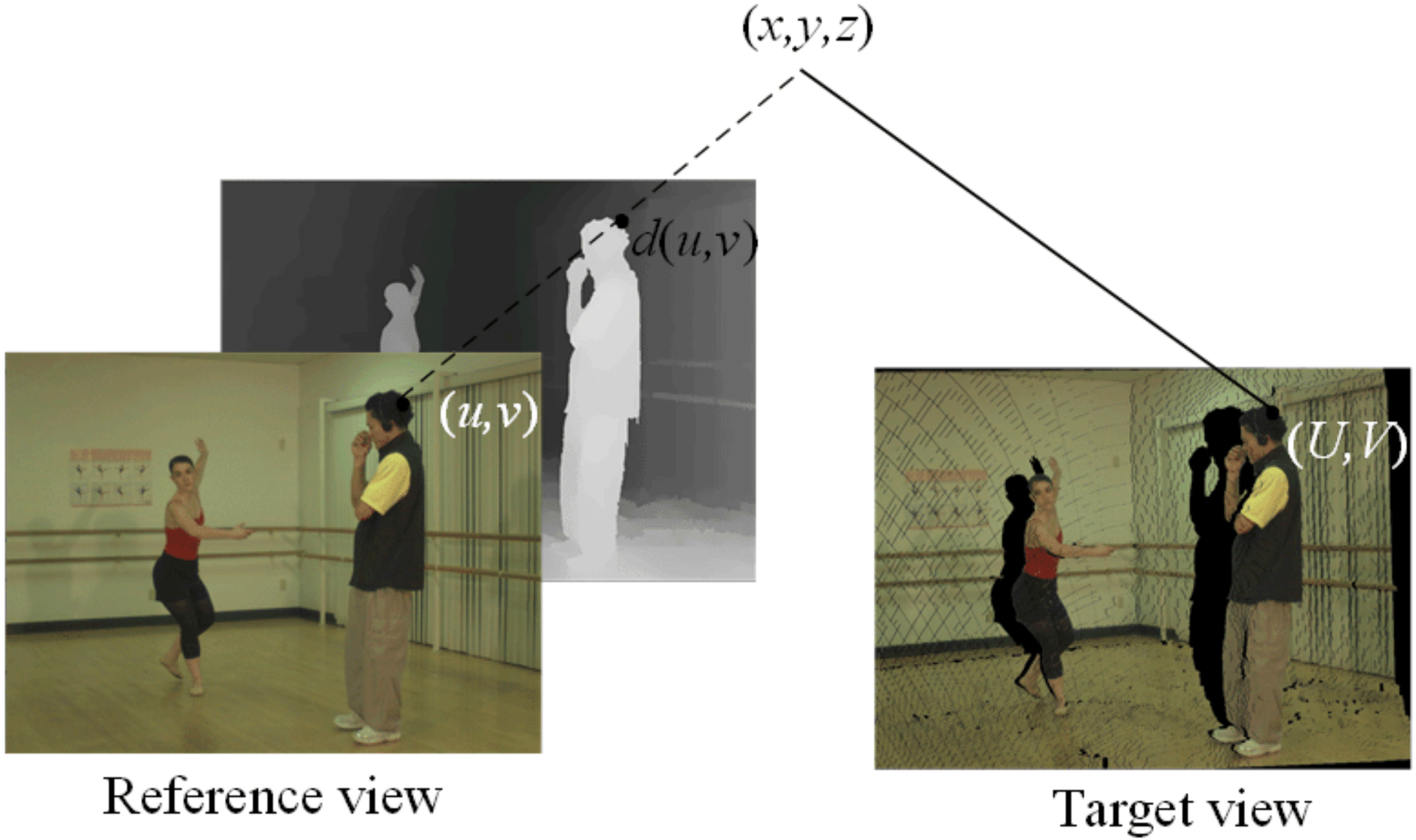
Image source: https://www.calvert.ch/maurice/2018/06/12/intel-realsense-d435-review/

“Unproject”: Go from pixel and depth to 3D point

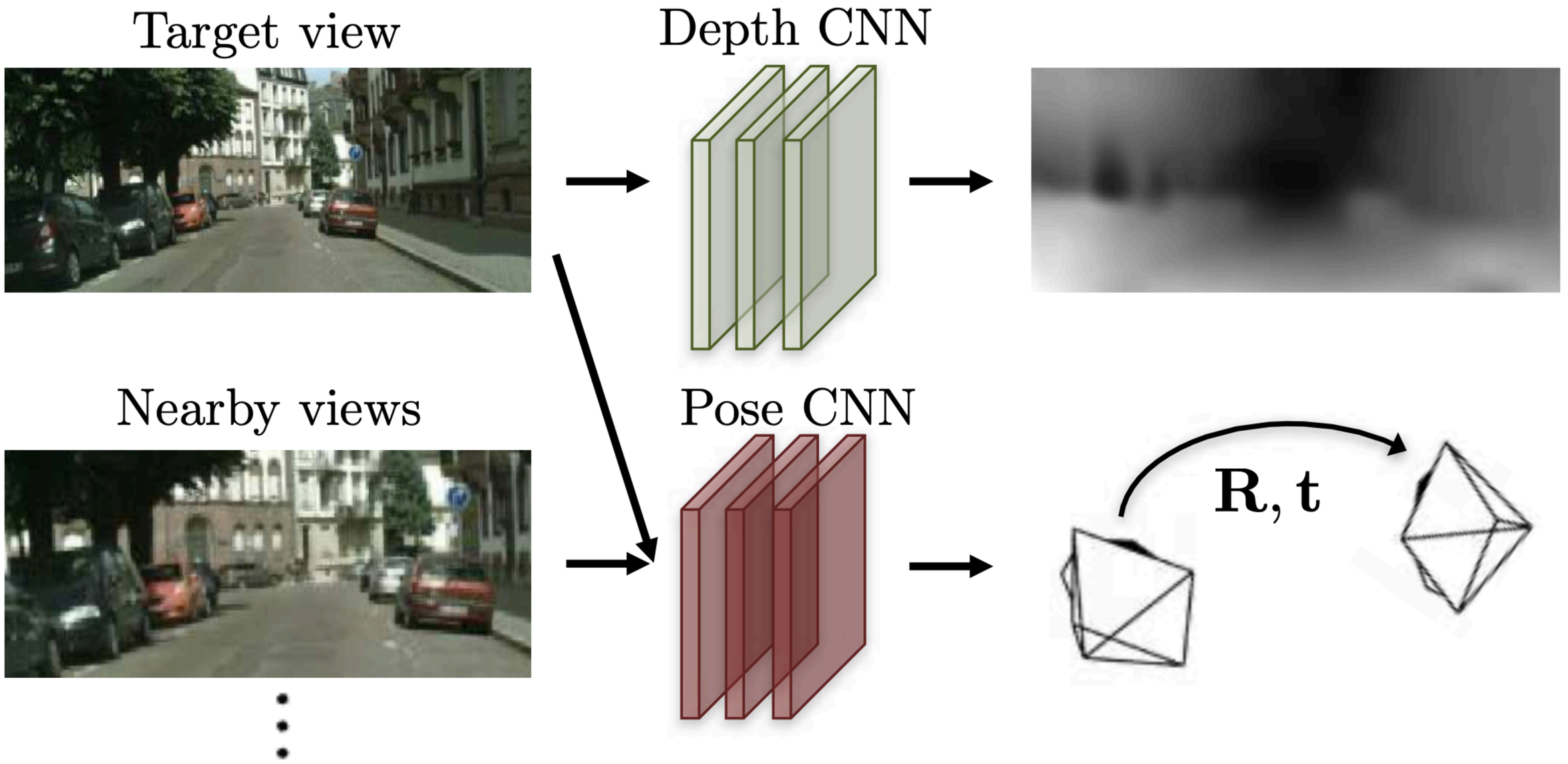
$$\mathbf{X} = Z \cdot \mathbf{K}^{-1} \begin{pmatrix} x_{pix} \\ y_{pix} \\ 1 \end{pmatrix}$$

$$\texttt{unproject}(\mathbf{x}, Z) = Z \cdot \mathbf{K}^{-1} \begin{pmatrix} x_{pix} \\ y_{pix} \\ 1 \end{pmatrix}$$

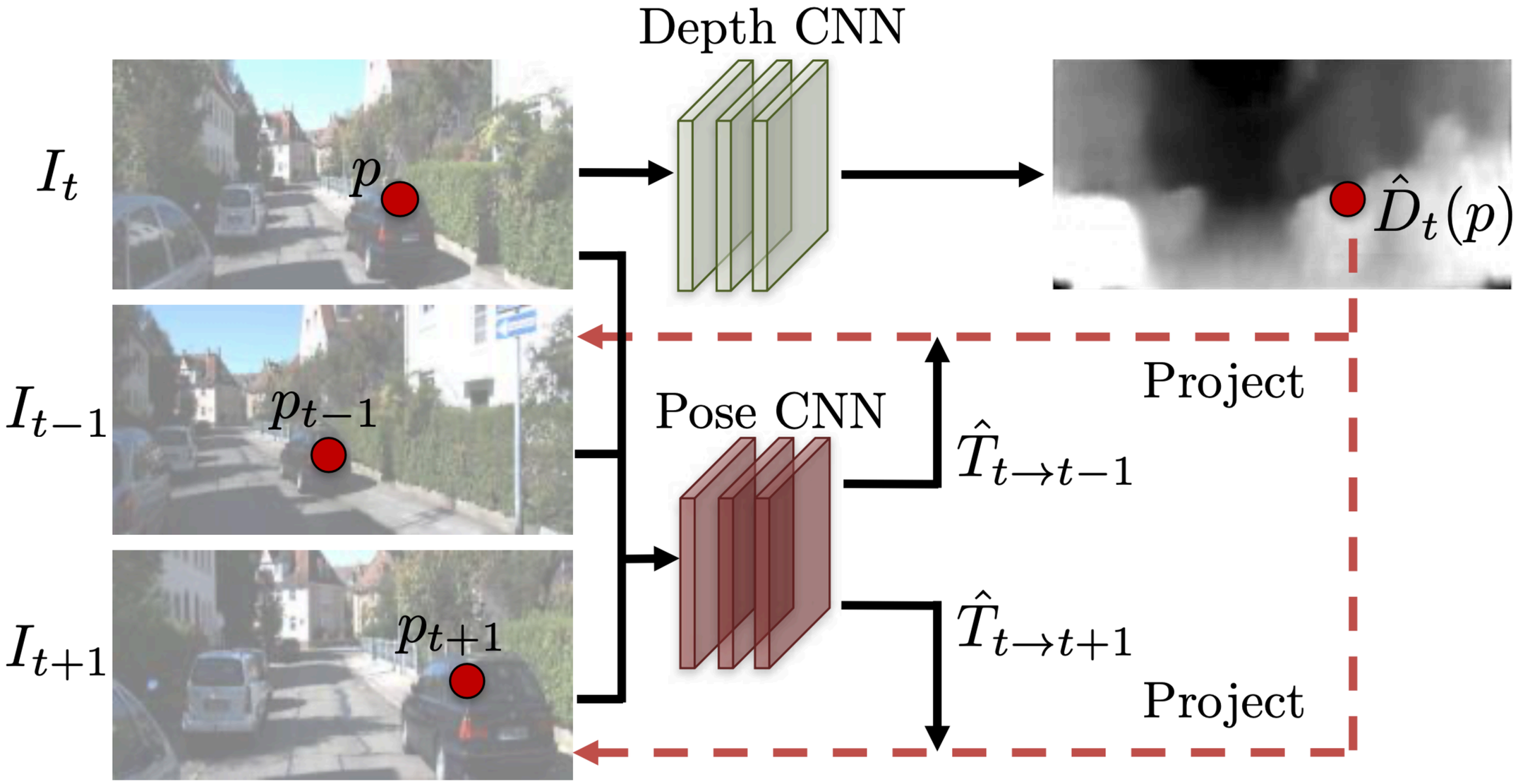
View Synthesis with Depth Maps (Depth Warping)

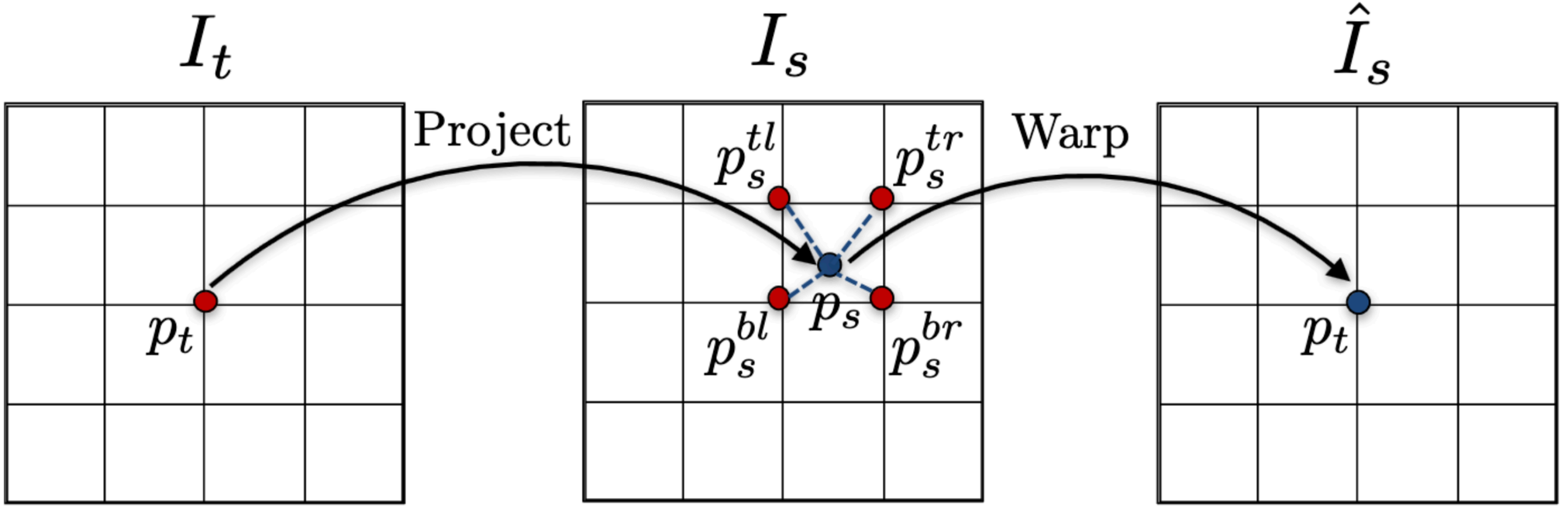


Unsupervised Depth and Ego-Motion from Video (Zhou et al. 2017)

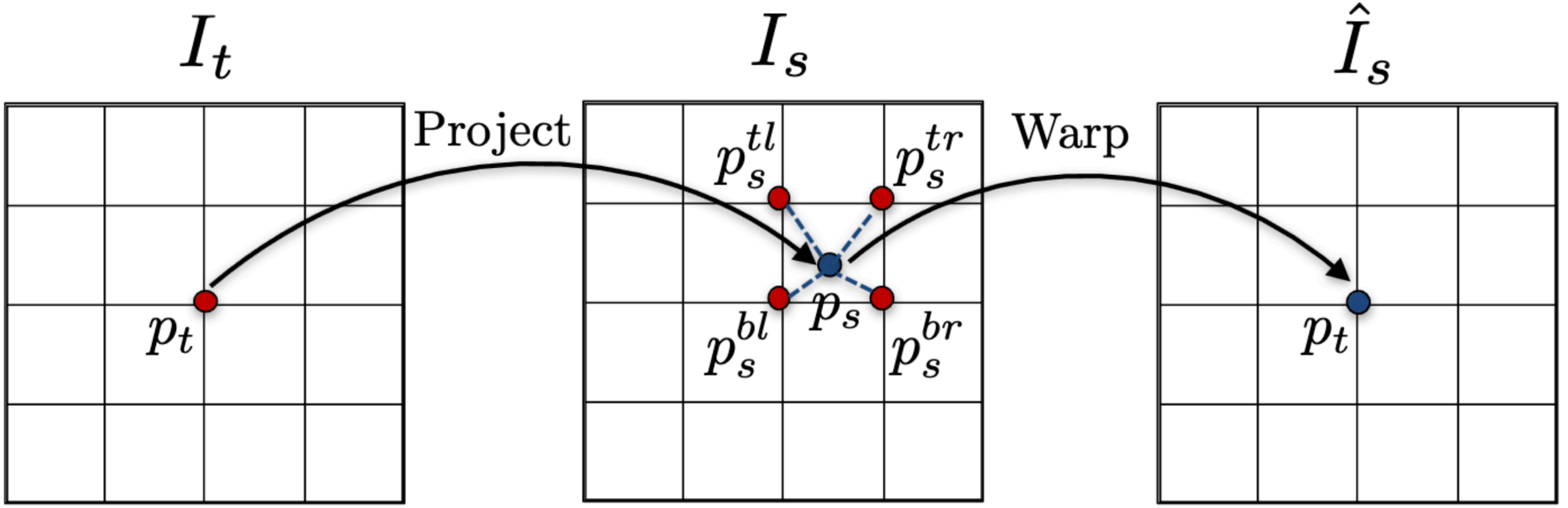


Unsupervised Depth and Ego-Motion from Video (Zhou et al. 2017)



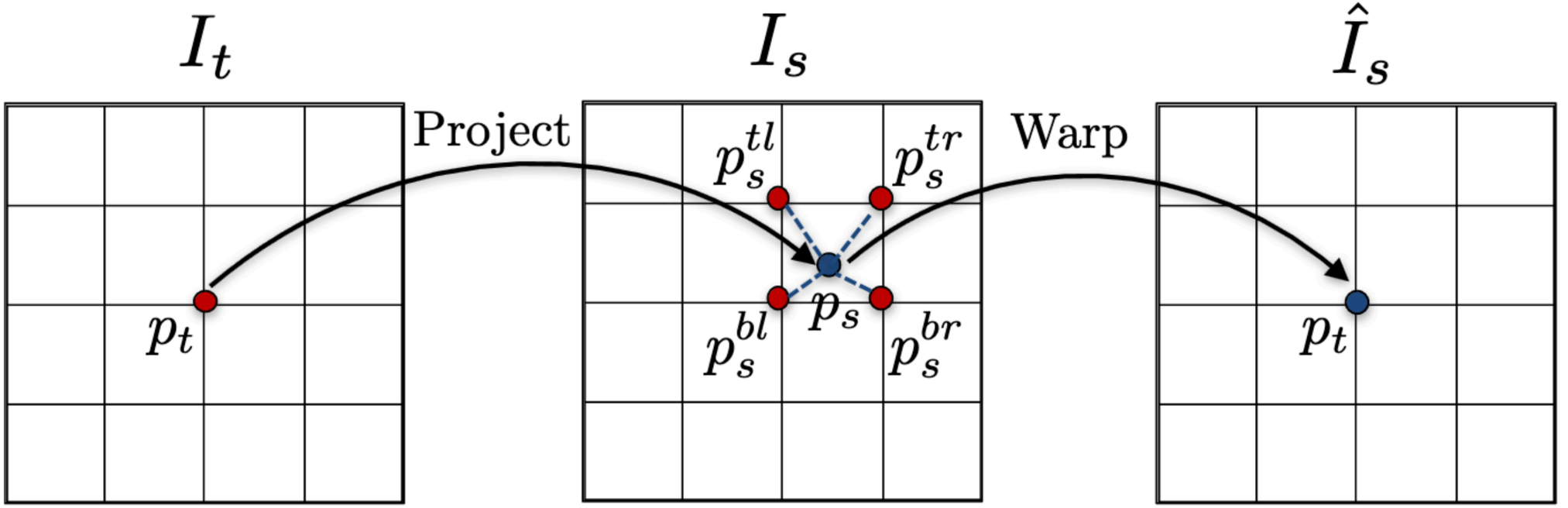


$$p_s \sim K\hat{T}_{t\to s}\hat{D}_t(p_t)K^{-1}p_t$$



$$p_s \sim K\hat{T}_{t\to s}\hat{D}_t(p_t)K^{-1}p_t$$

$$\hat{I}_s(p_t) = I_s(p_s) = \sum_{i\in\{t,b\},j\in\{l,r\}} w^{ij} I_s(p_s^{ij})$$

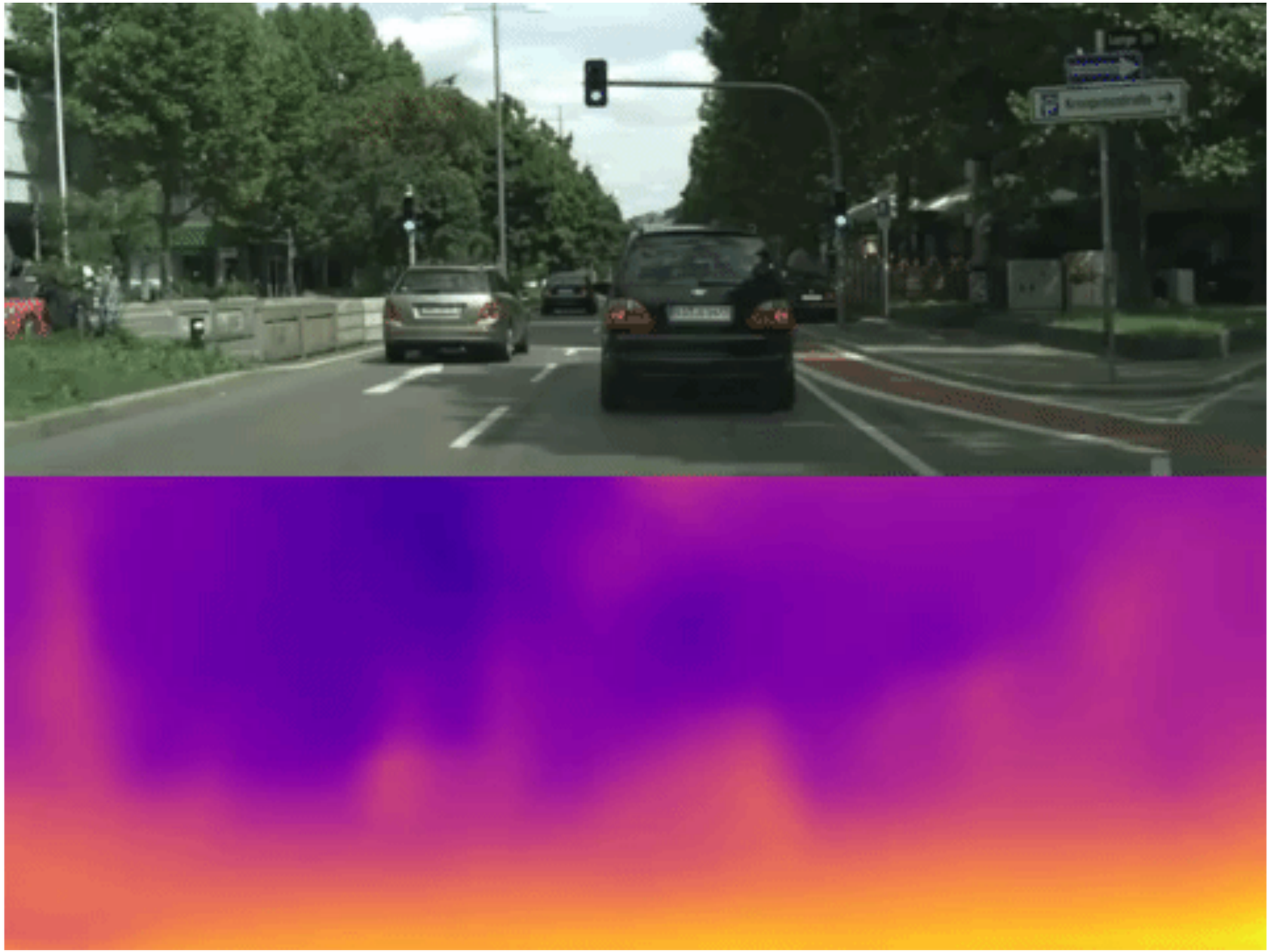


$$p_s \sim K\hat{T}_{t\to s}\hat{D}_t(p_t)K^{-1}p_t$$

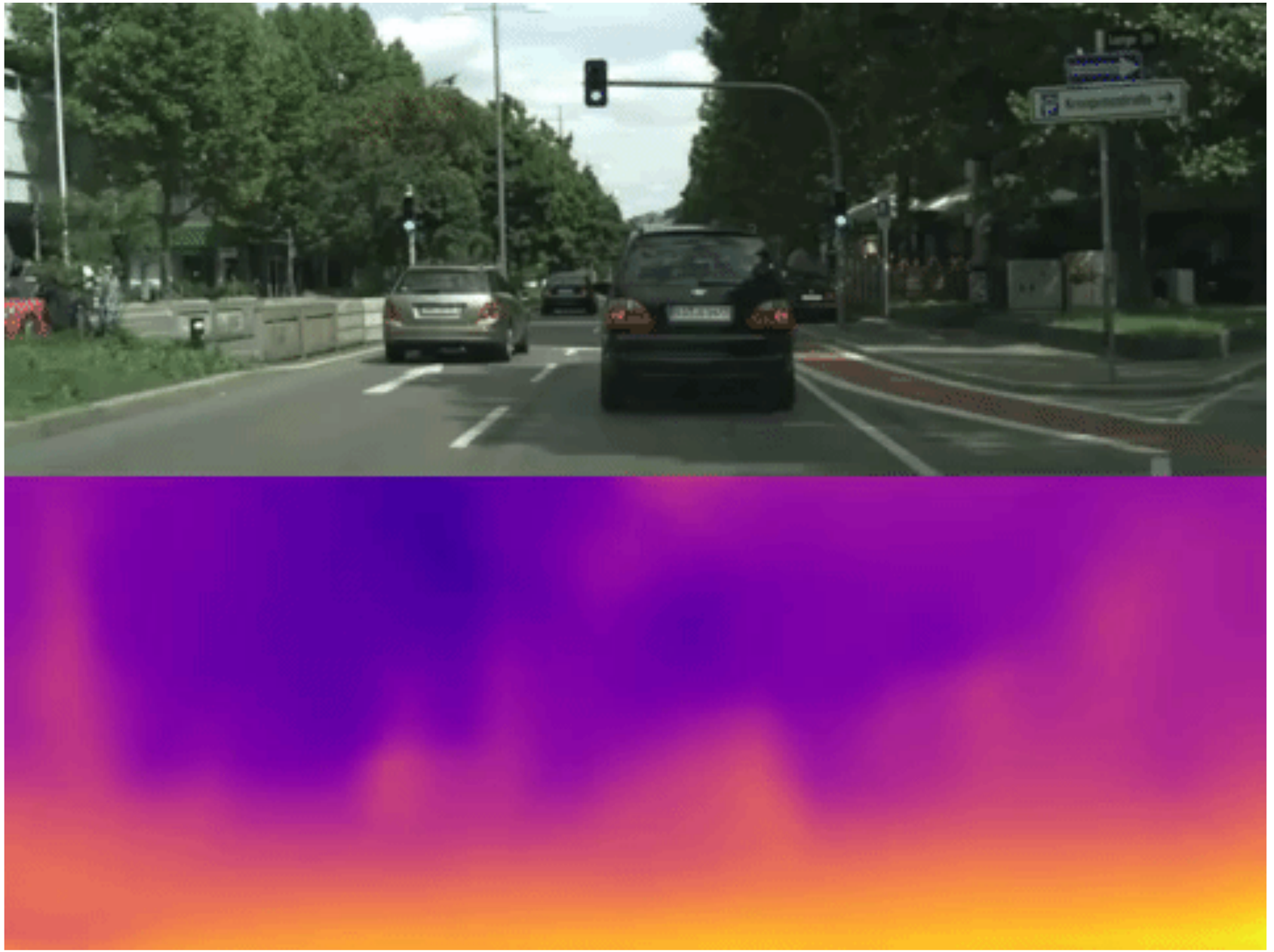
$$\hat{I}_s(p_t) = I_s(p_s) = \sum_{i\in\{t,b\},j\in\{l,r\}} w^{ij} I_s(p_s^{ij})$$

$$\mathcal{L}_{vs} = \sum_s \sum_p |I_t(p) - \hat{I}_s(p)|$$

The necessity of Multi-Scale Losses



Unsupervised Depth and Ego-Motion from Video, Zhou et al. 2017

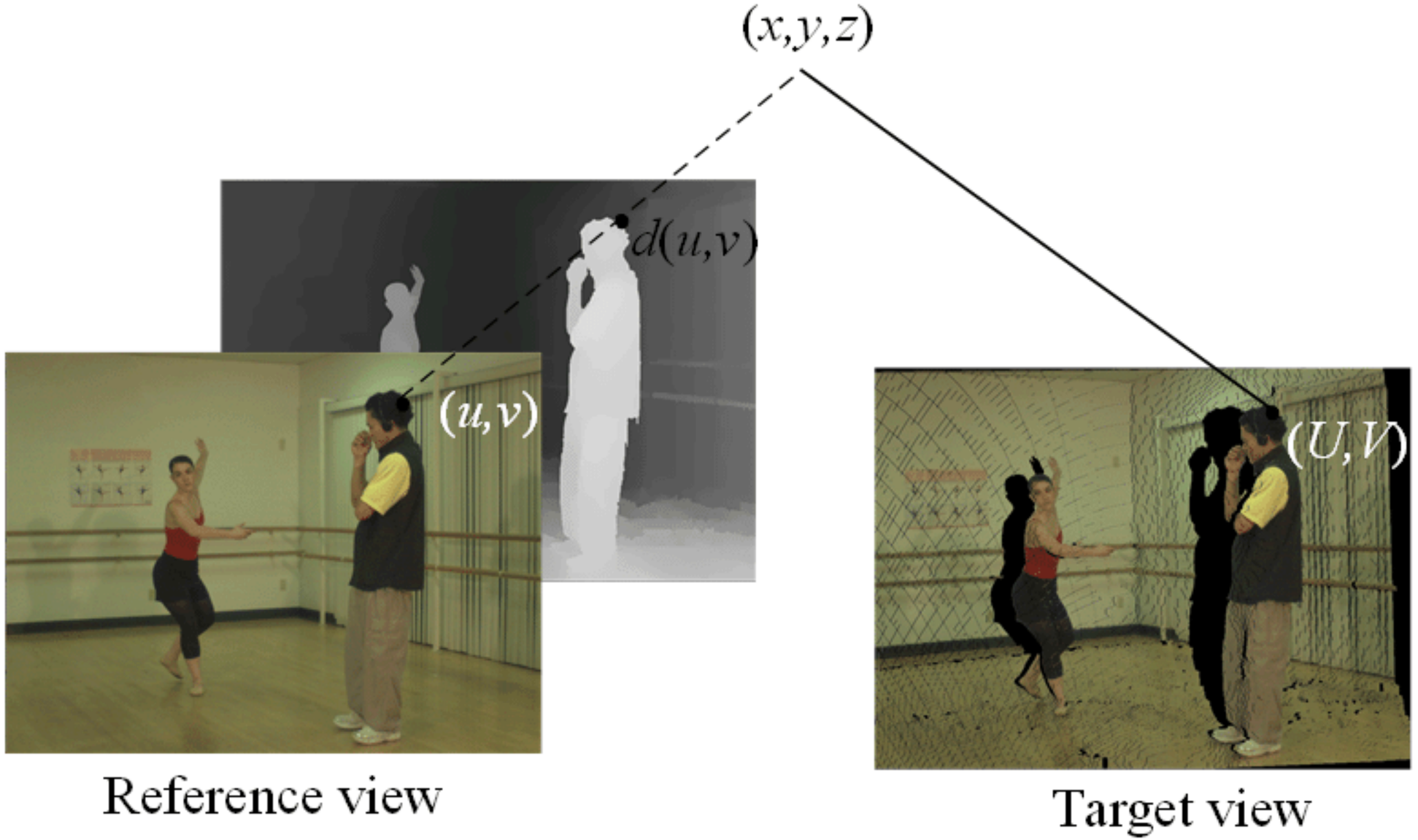


Unsupervised Depth and Ego-Motion from Video, Zhou et al. 2017

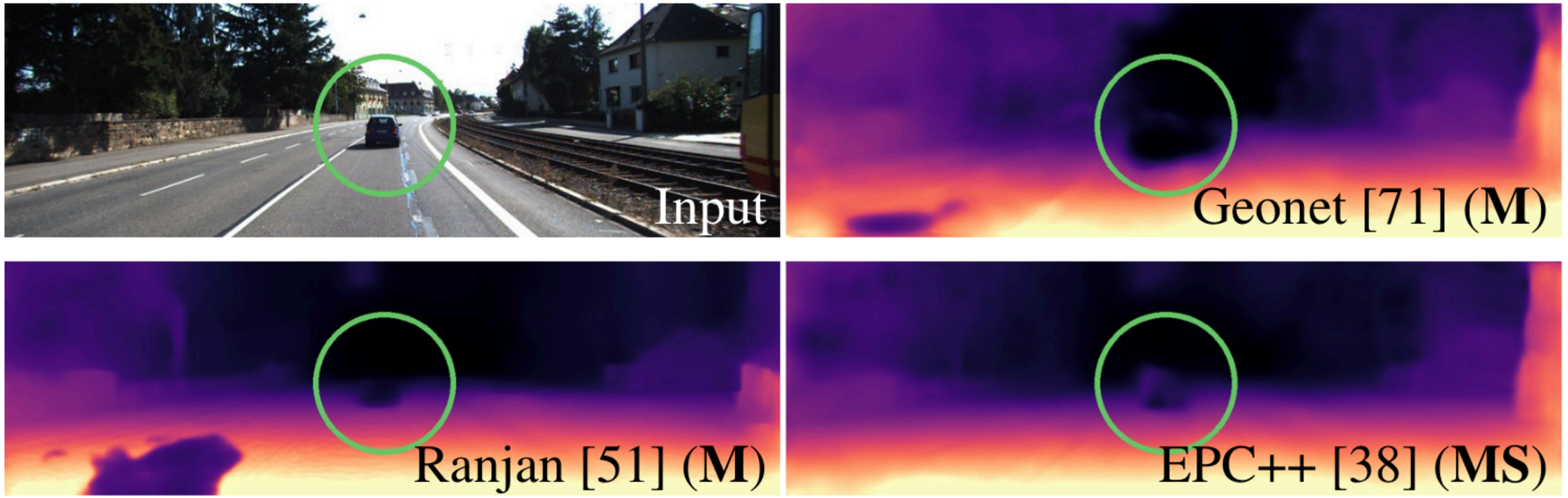
Self-supervised Learning of Depth and Pose from Video
Guizilini et al. 2021

Self-supervised Learning of Depth and Pose from Video
Guizilini et al. 2021

Can’t deal with occlusions / disocclusions



Assumes *static* scenes



Digging Into Self-Supervised Monocular Depth Estimation, Godard et al. 2018

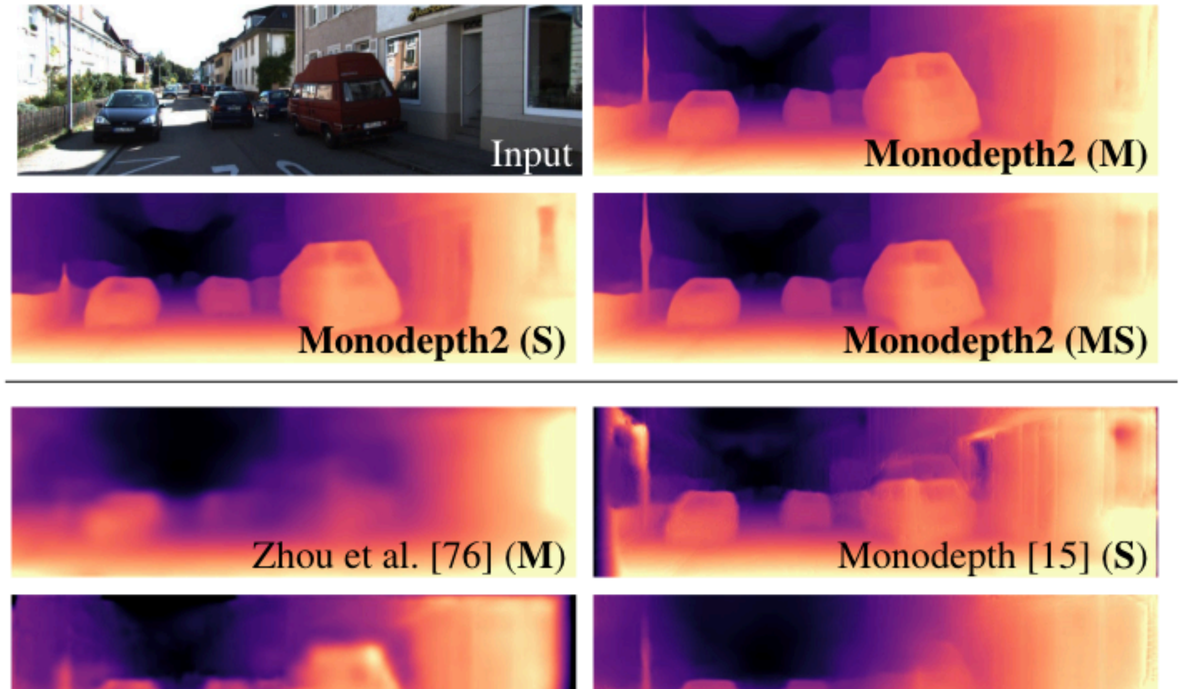
Digging Into Self-Supervised Monocular Depth Estimation

Clément Godard[1] Oisin Mac Aodha[2] Michael Firman[3] Gabriel Brostow[3,1]
[1]UCL [2]Caltech [3]Niantic
www.github.com/nianticlabs/monodepth2

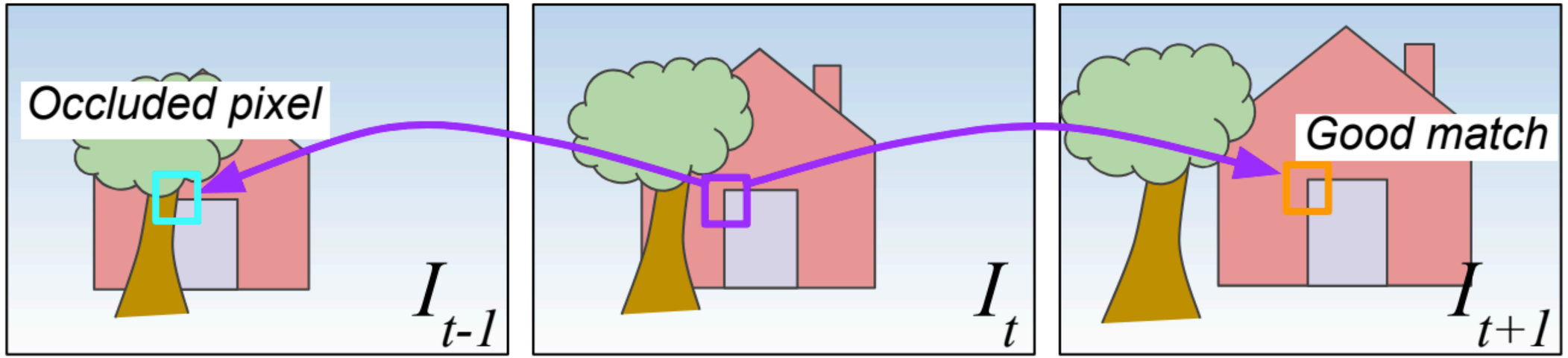
Abstract

Per-pixel ground-truth depth data is challenging to acquire at scale. To overcome this limitation, self-supervised learning has emerged as a promising alternative for training models to perform monocular depth estimation. In this paper, we propose a set of improvements, which together result in both quantitatively and qualitatively improved depth maps compared to competing self-supervised methods.

Research on self-supervised monocular training usually explores increasingly complex architectures, loss functions, and image formation models, all of which have recently



Dealing with Occlusions / Disocclusions



$pe($, $)$ = error

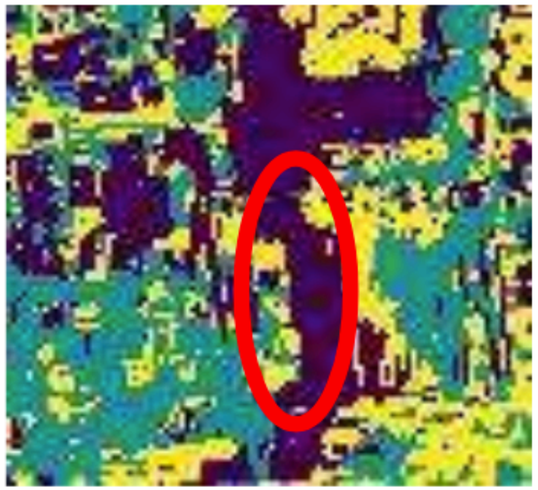
$pe($, $)$ = error

Baseline: $avg($, $)$ =

Ours: $min($, $)$ =

Digging into Self-Supervised Monocular Depth Estimation, Godard et al. 2018

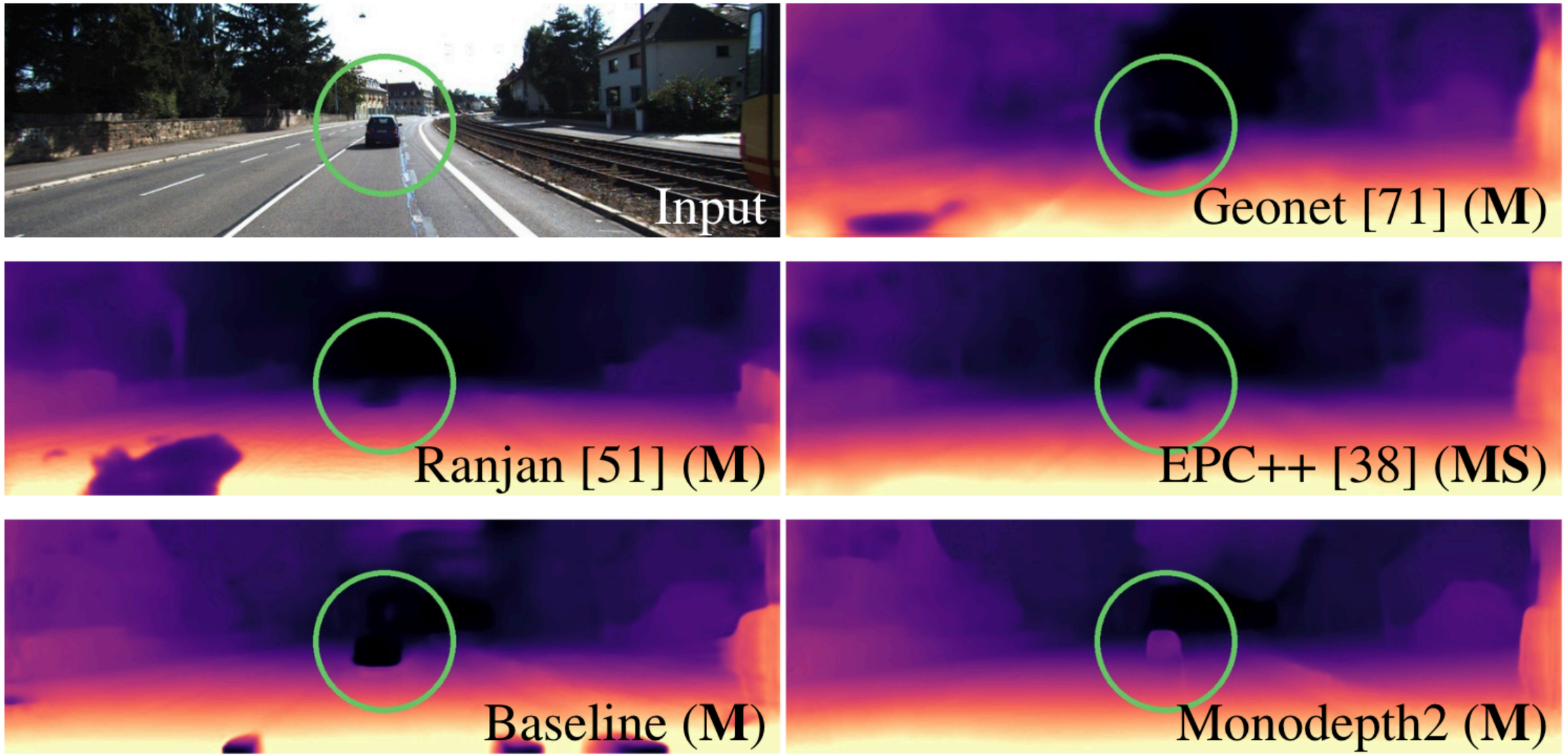
Dealing with Occlusions / Disocclusions



Colors here show which source frame each pixel in L is matched to.

Digging into Self-Supervised Monocular Depth Estimation, Godard et al. 2018

Dealing with moving objects



Dealing with moving objects: Predict per-pixel mask μ

We observe that pixels which remain the same between adjacent frames in the sequence often indicate a static camera, an object moving at equivalent relative translation to the camera, or a low texture region. We therefore set μ to only include the loss of pixels where the reprojection error of the warped image $I_{t' \to t}$ is lower than that of the original, unwarped source image I'_t, *i.e.*

$$\mu = \left[\min_{t'} pe(I_t, I_{t' \to t}) < \min_{t'} pe(I_t, I_{t'}) \right], \qquad (5)$$

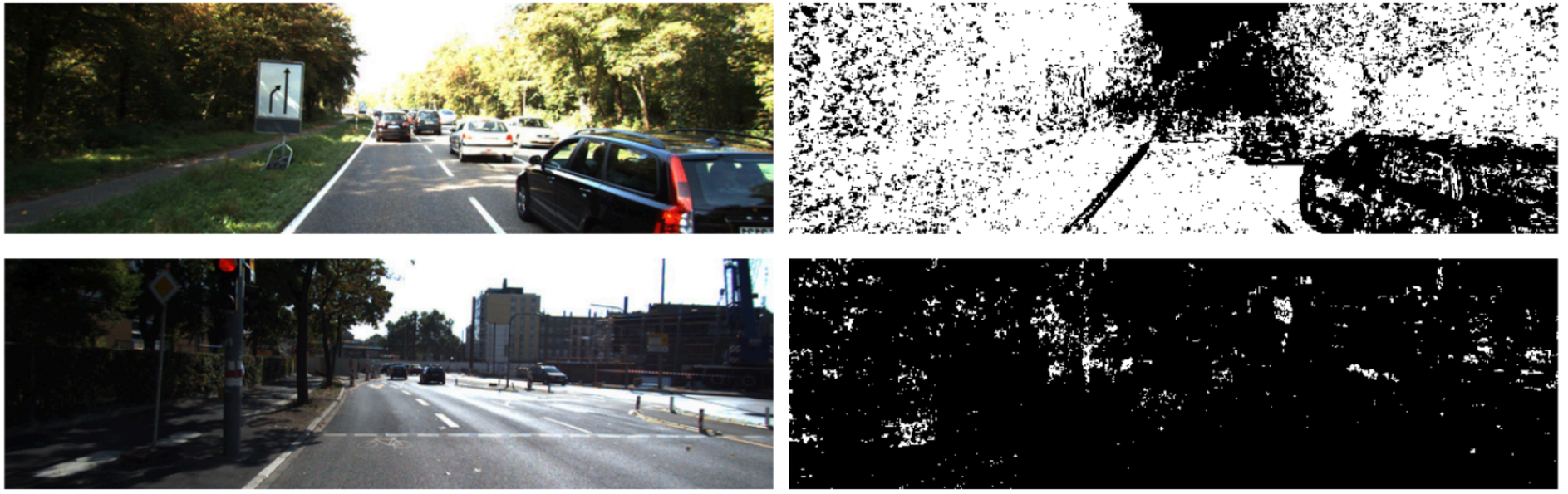
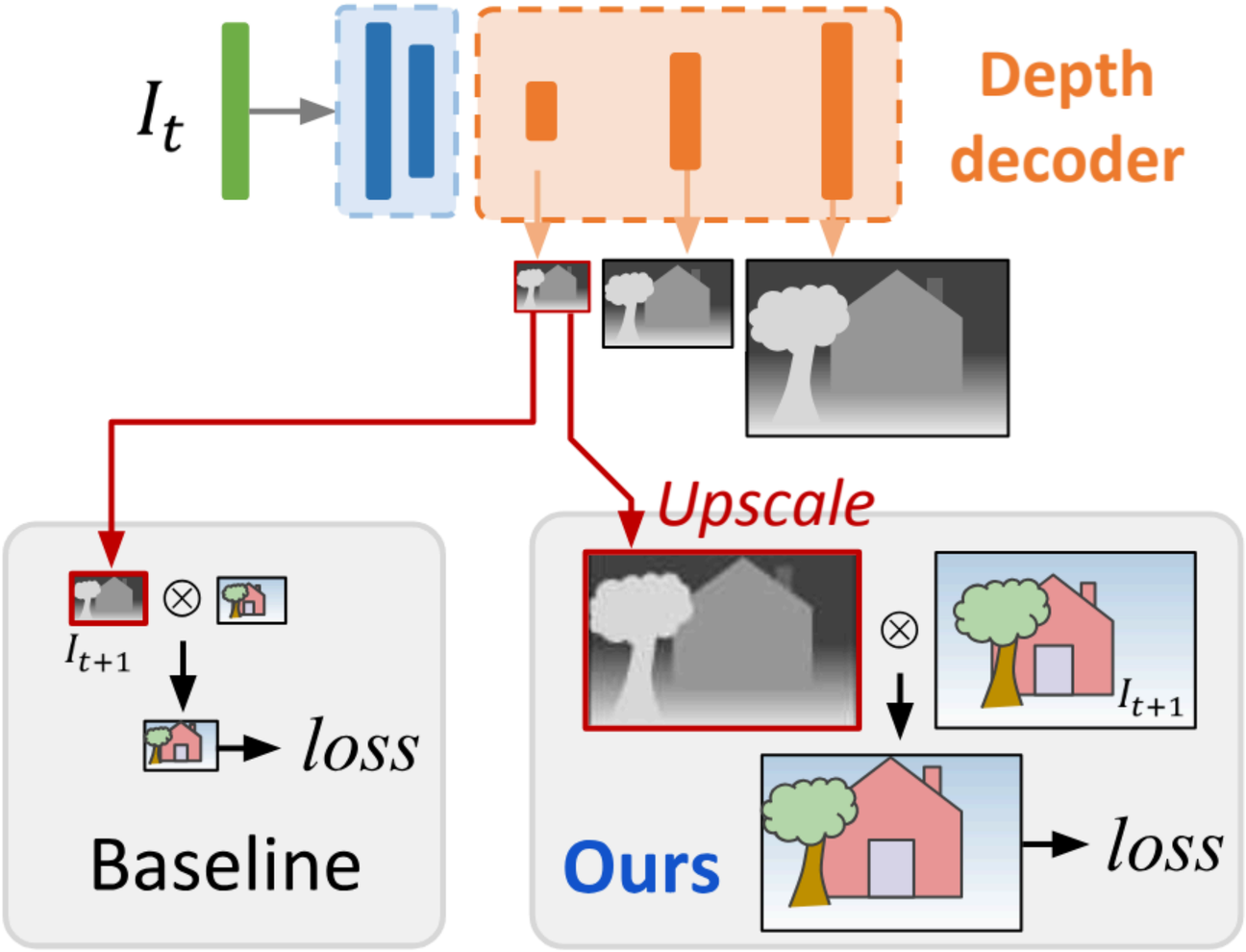


Figure 5. **Auto-masking.** We show auto-masks computed after one epoch, where black pixels are removed from the loss (*i.e.* $\mu = 0$). The mask prevents objects moving at similar speeds to the camera (top) and whole frames where the camera is static (bottom) from contaminating the loss. The mask is computed from the input frames and network predictions using Eqn. 5.

Better multi-scale loss



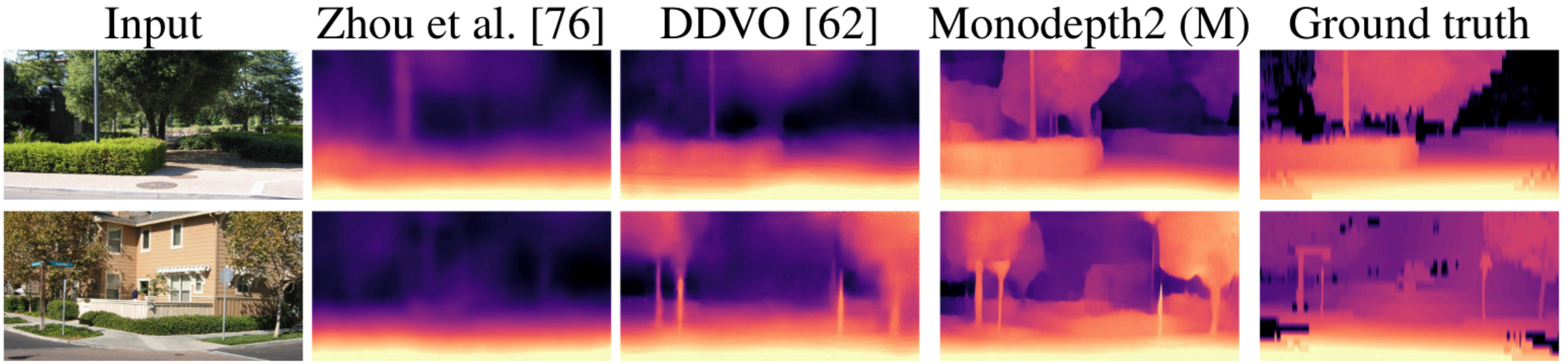


Figure 6. **Qualitative Make3D results.** All methods were trained on KITTI using monocular supervision.

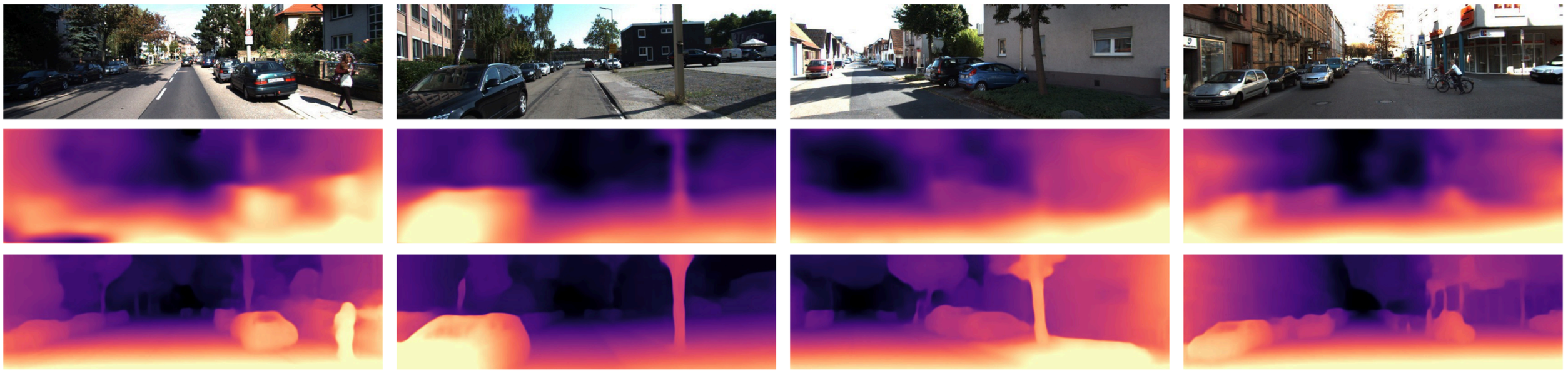
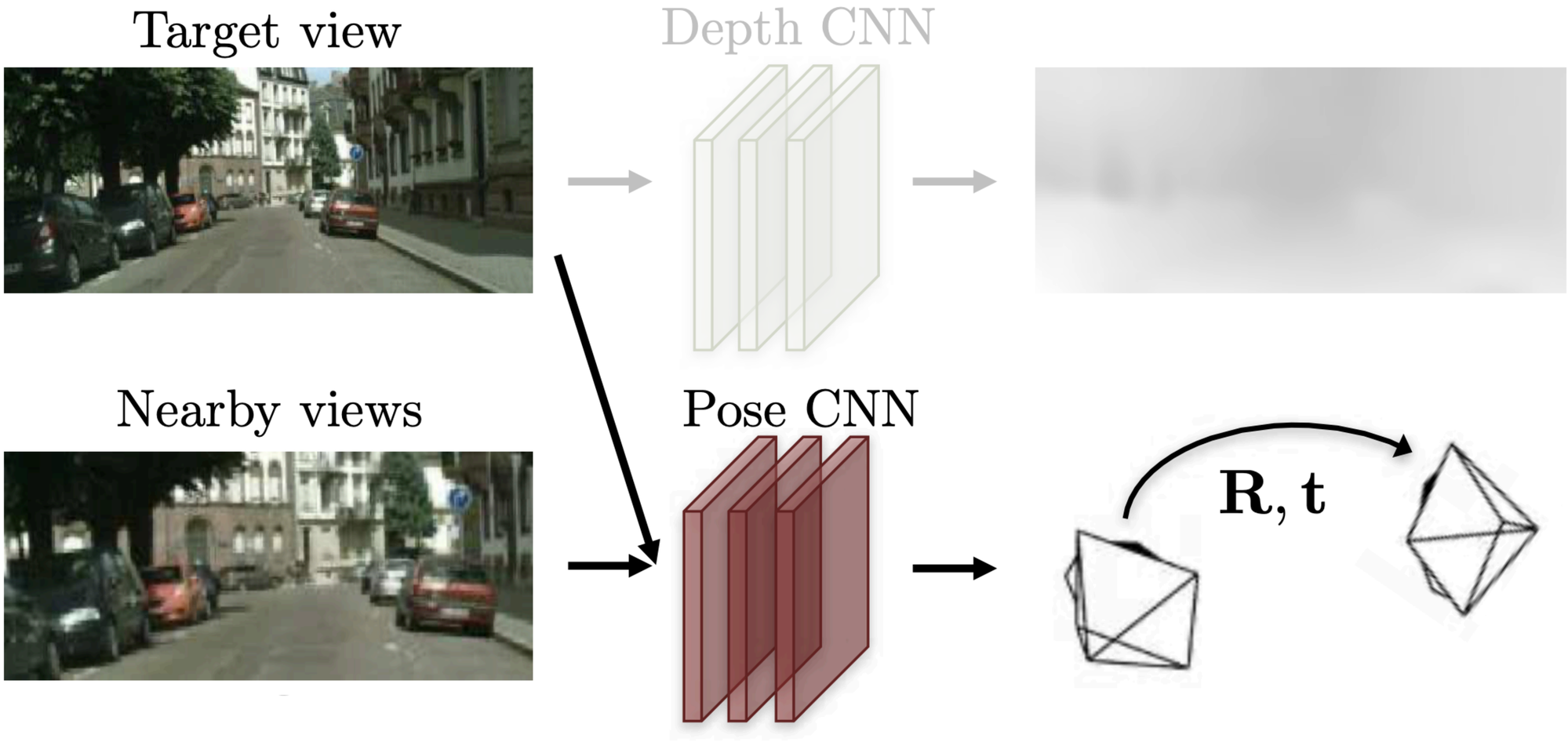

Input
Zhou et al. [76]
MD2 MS

Figure 8. **Failure cases. Top:** Our self-supervised loss fails to learn good depths for distorted, reflective and color-saturated regions. **Bottom:** We can fail to accurately delineate objects where boundaries are ambiguous (left) or shapes are intricate (right).

Problem: Only works for “simple” camera trajectories :/



CNNs are bad at predicting poses: no single pixel neighborhood informs completely about pose…