III: Depth Estimation

3D CV

Kirill Struminsky

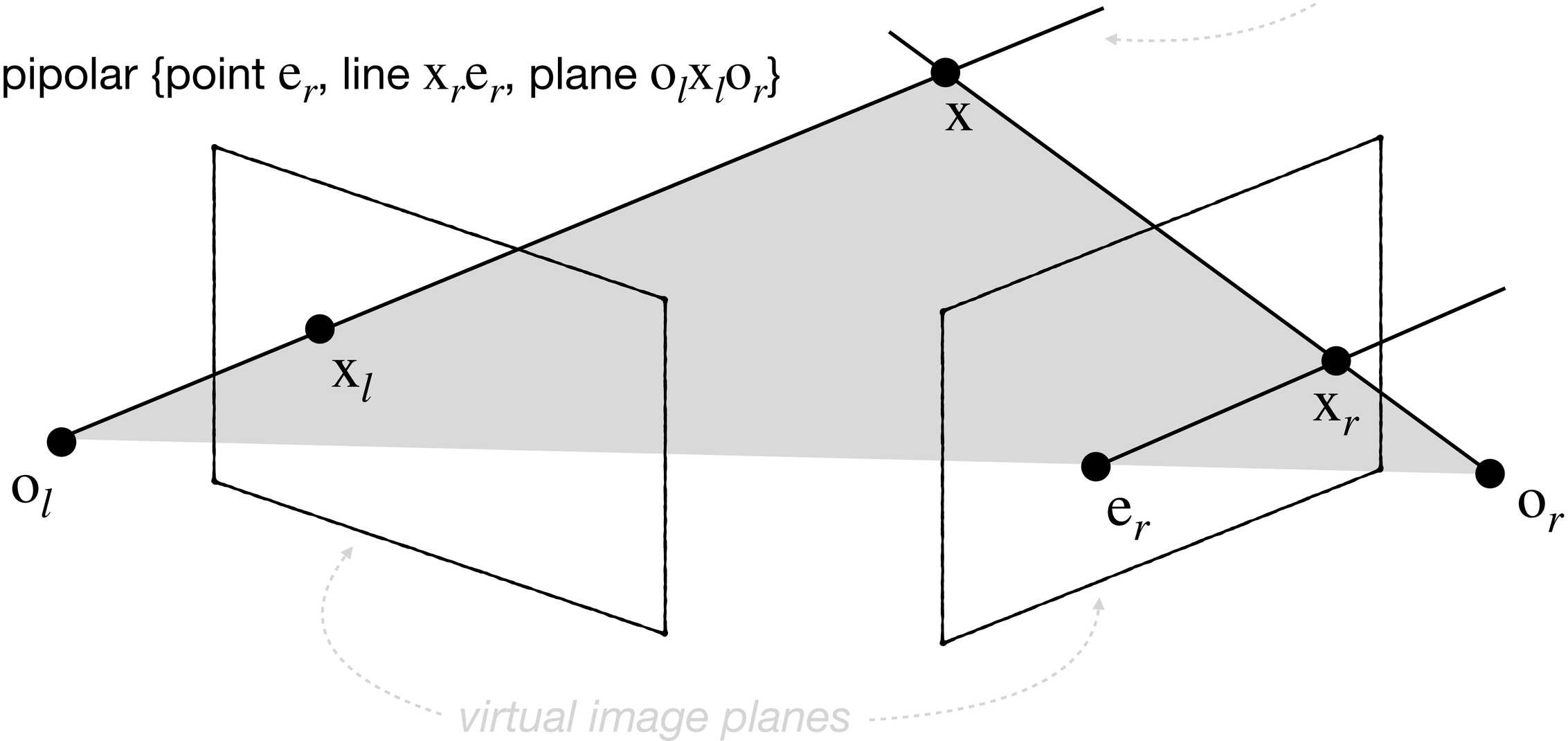
In the Previous Episode

- Two-view geometry
  - Epipolar geometry
  - Fundamental matrix
  - Triangulation
- Structure from motion

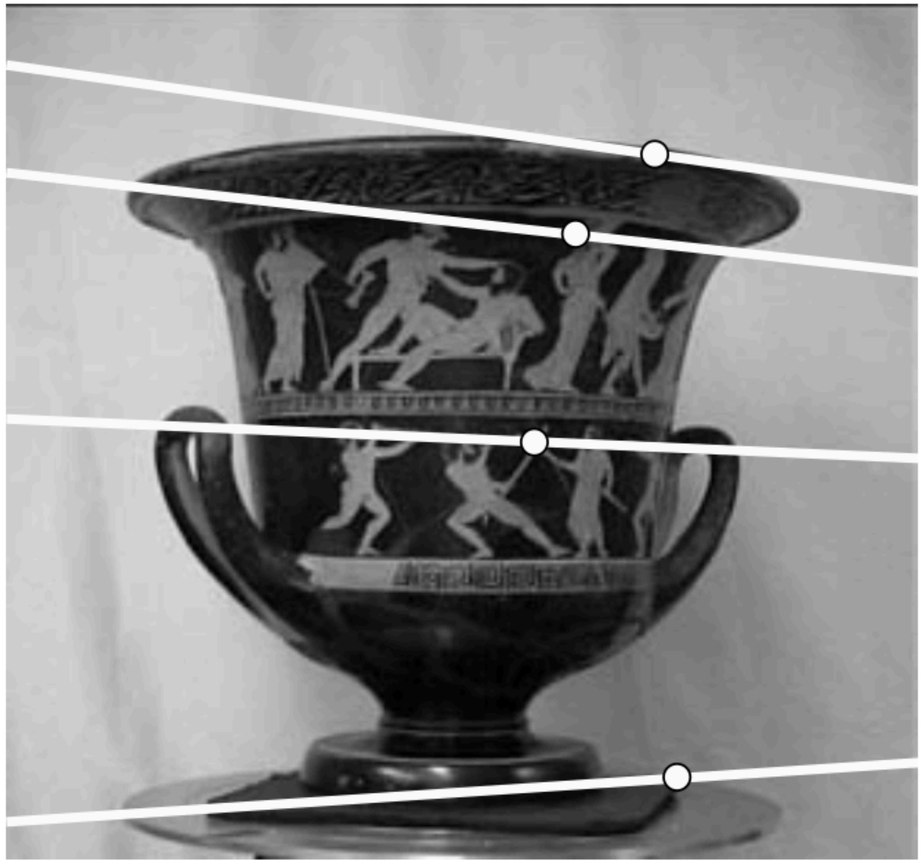
ComponentEpipolar Geometry

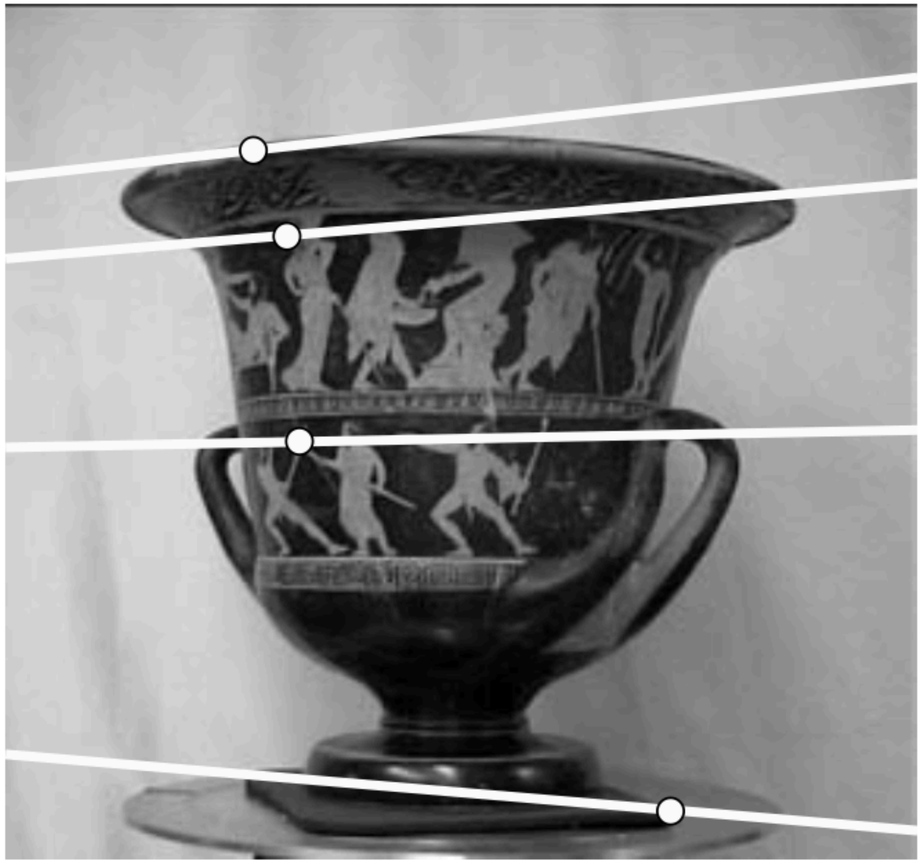
How does the image changes as we move?

epipolar {point e_r, line $x_r e_r$, plane $o_l x_l o_r$}



Epipolar Lines in Real Life

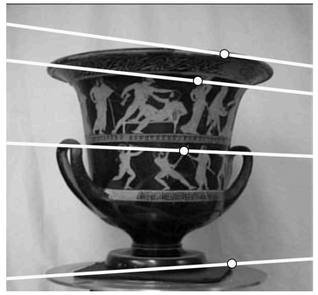


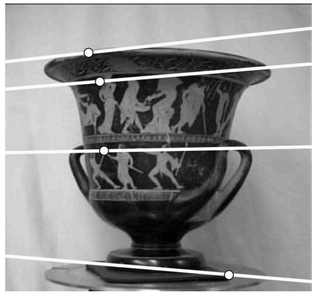


Hartley & Zisserman, Multiple View Geometry in CV, p. 241

Estimating Camera Positions

Using fundamental matrix F





- Estimate F using correspondences
- Consider $\mathrm{H}_l = \begin{bmatrix}\mathrm{K}_l | 0\end{bmatrix} \begin{bmatrix}\mathrm{I} & 0 \\ 0 & 1\end{bmatrix}$ and $\mathrm{H}_l = \begin{bmatrix}\mathrm{K}_r | 0\end{bmatrix} \begin{bmatrix}\mathrm{R} & \mathrm{t} \\ 0 & 1\end{bmatrix}$
- Fundamental matrix encodes R and t:

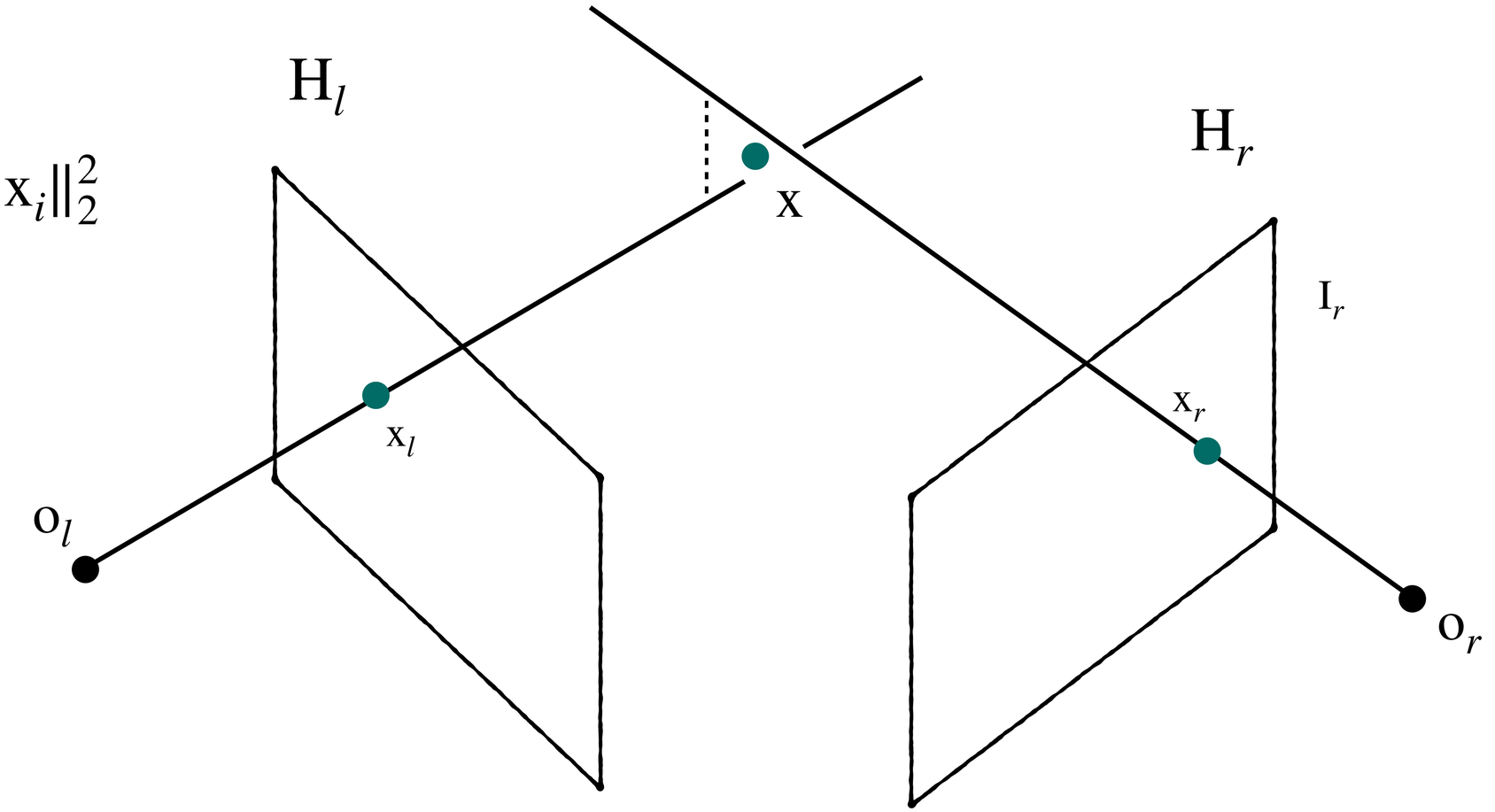
$$\mathrm{F} = \mathrm{K}_r^{-T}[\mathrm{t}]_\times \mathrm{R}\mathrm{K}_l^{-1}$$

- Assuming K_r and K_l are known we can recover R and t

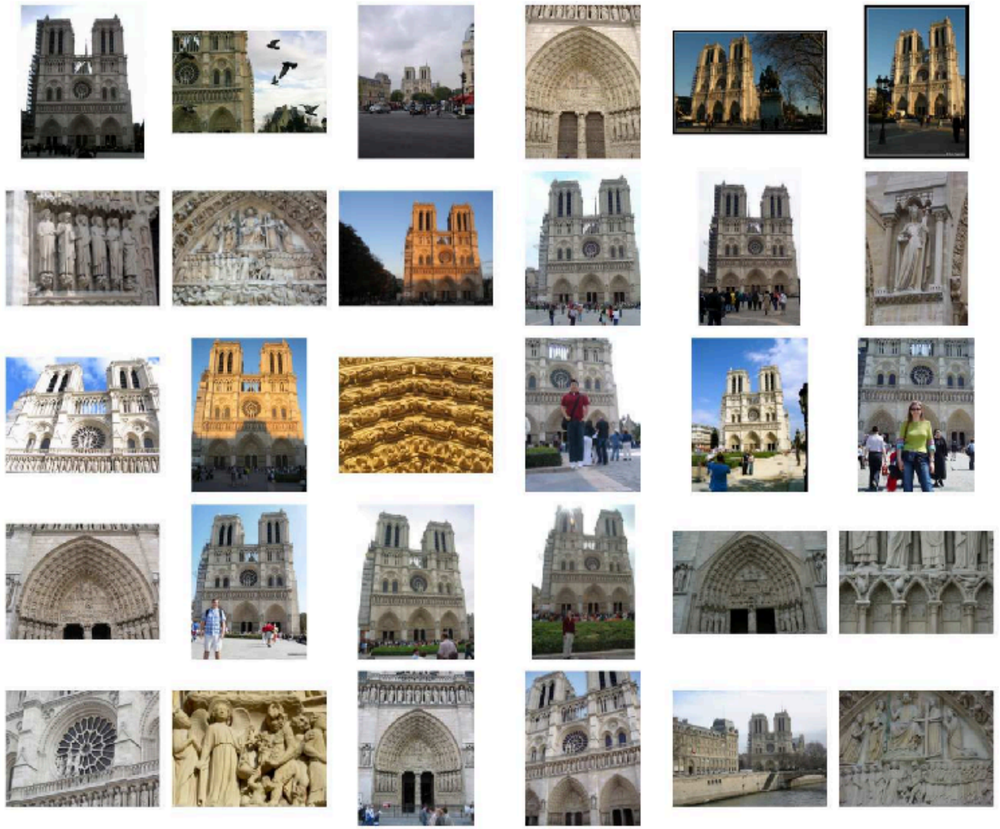
Structure Recovery

- Solve:

$$\mathrm{x} = \arg\min \sum_{i \in \{l,r\}} \|\mathrm{H}_i \mathrm{x} - \mathrm{x}_i\|_2^2$$

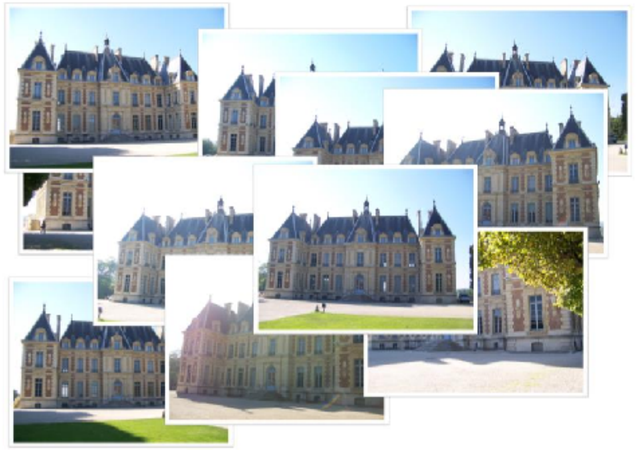


Structure from Motion

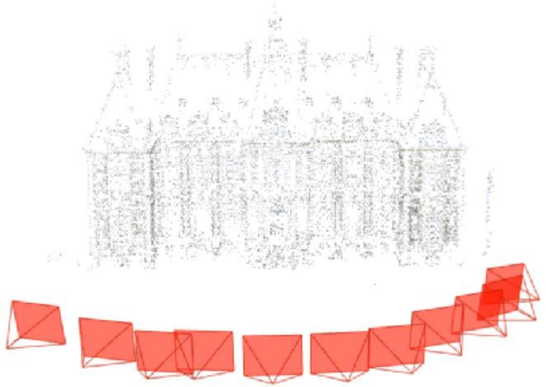


Snavely, Noah, Steven M. Seitz, and Richard Szeliski. "Photo tourism: exploring photo collections in 3D." ACM SIGGRAPH 2006

Today's Lecture: Depth Estimation



multi-view images



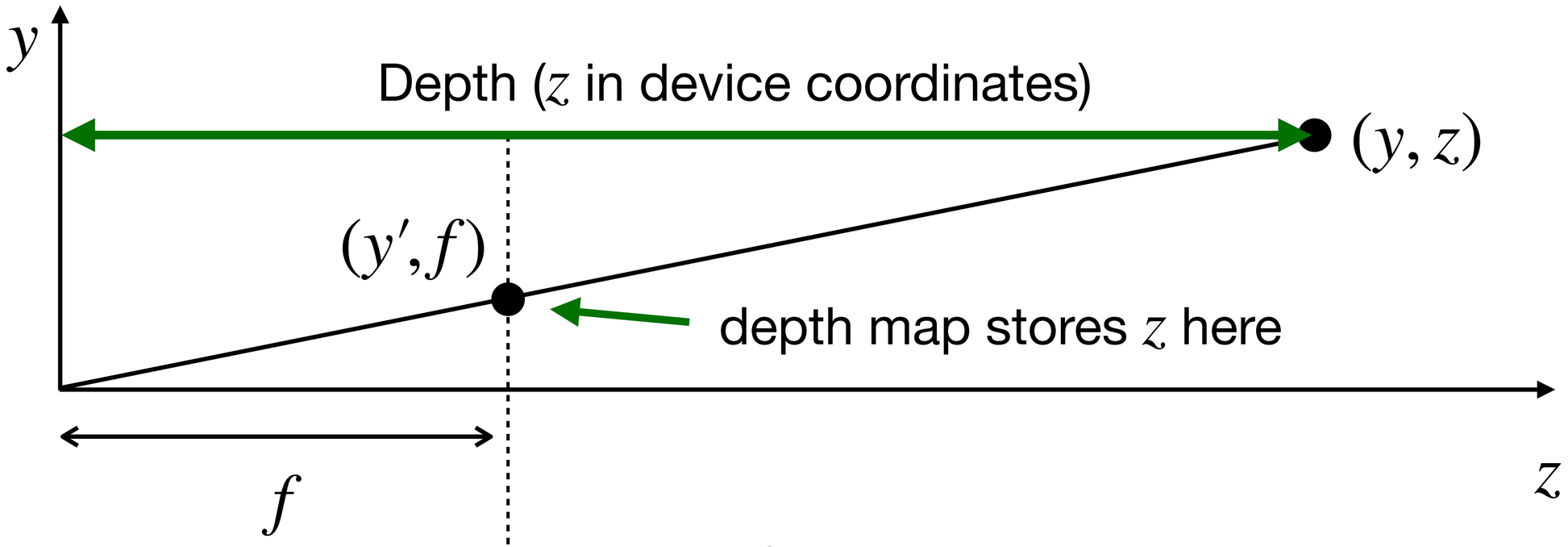
structure-from-motion

multi-view stereo

surface reconstruction

Depth Estimation Basics

Depth and Depth Maps



$$\frac{y'}{f} = \frac{y}{z} \quad \Rightarrow \quad y' = f\,\frac{y}{z}$$

https://github.com/isl-org/ZoeDepth

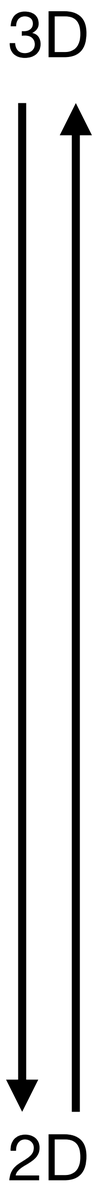
From 2D to 2.5D

- Consider $\mathrm{x} = (x, y, z, 1)$
- Project

$$[\mathrm{K}|0]\,\mathrm{x} = \begin{pmatrix} f_x & 0 & p_x & 0 \\ 0 & f_y & p_y & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \mathrm{x} = \begin{pmatrix} f_x x + p_x z \\ f_y y + p_y z \\ z \end{pmatrix}$$

- Find inhomogeneous coordinates

$$\begin{pmatrix} f_x x + p_x z \\ f_y y + p_y z \\ z \end{pmatrix} \rightarrow \begin{pmatrix} x_{pix} \\ y_{pix} \\ 1 \end{pmatrix}$$

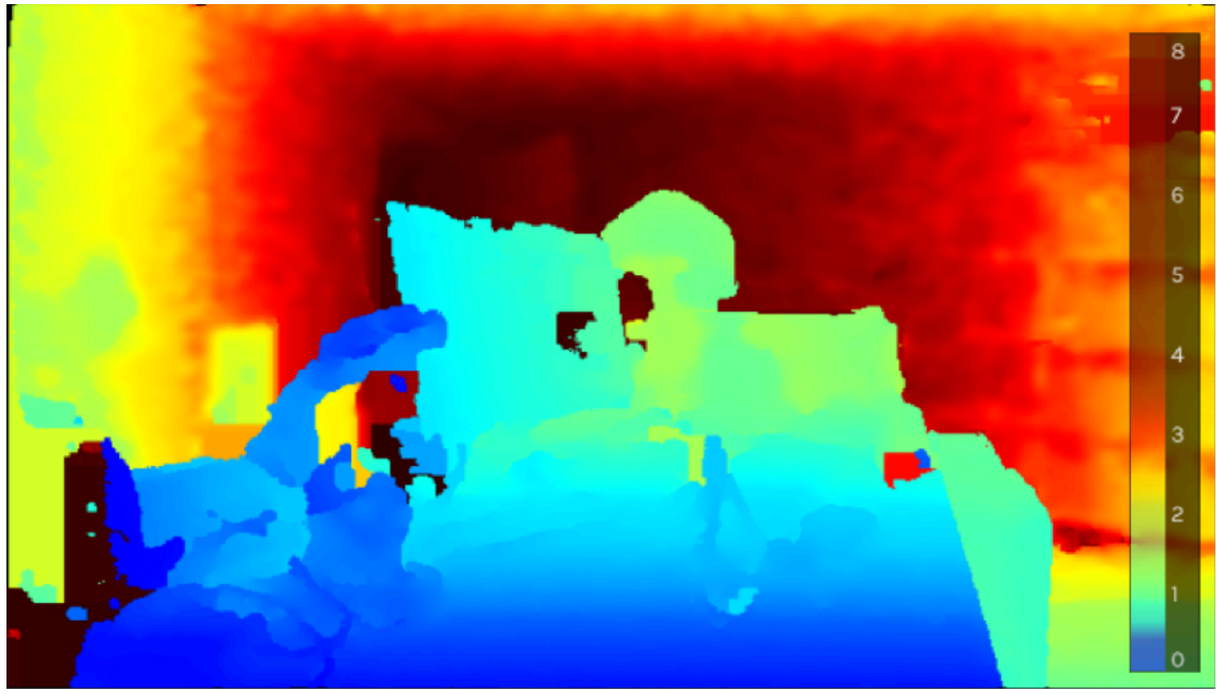


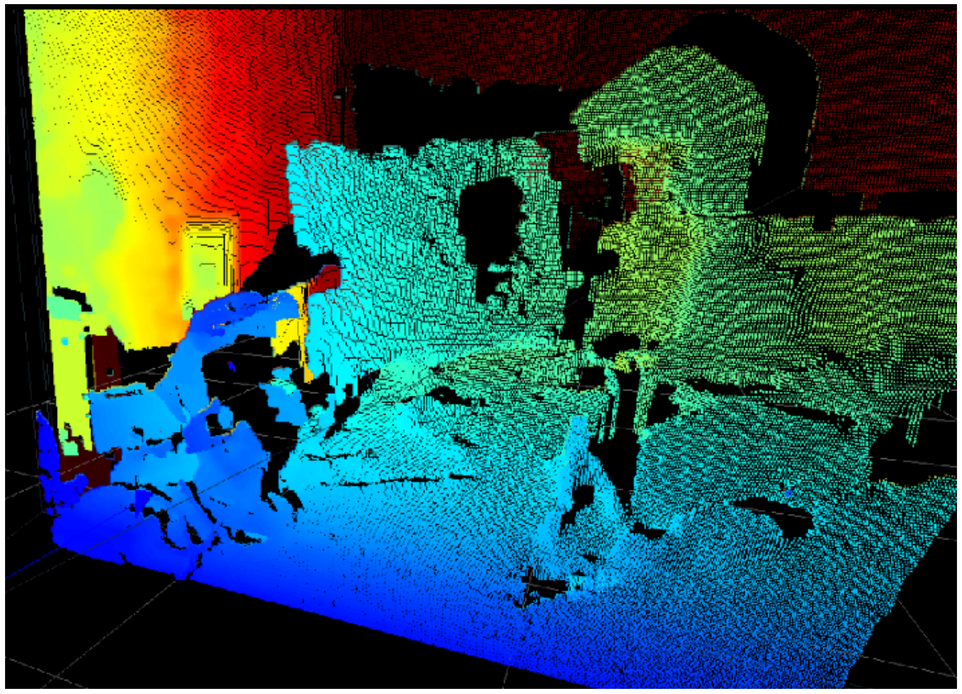
- Back projected point is $z\mathrm{K}^{-1} \begin{pmatrix} x_{pix} \\ y_{pix} \\ 1 \end{pmatrix}$
- Reverse projection

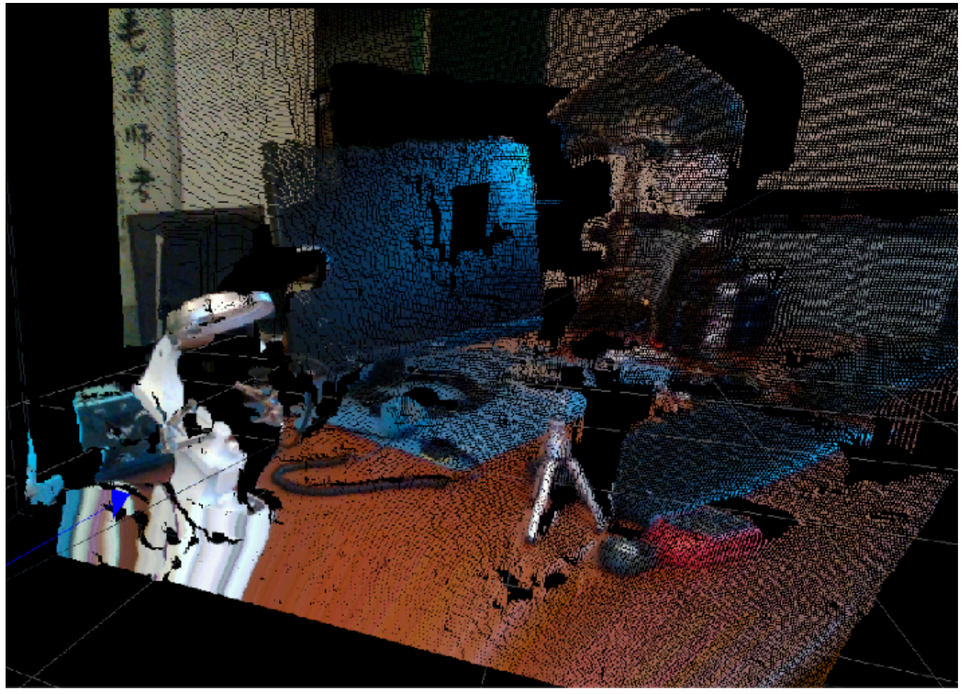
$$\mathrm{K}^{-1} \begin{pmatrix} z \cdot x_{pix} \\ z \cdot y_{pix} \\ z \end{pmatrix} = \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

- Get homogeneous coordinates

$$\begin{pmatrix} x_{pix} \\ y_{pix} \end{pmatrix} \rightarrow z \begin{pmatrix} x_{pix} \\ y_{pix} \\ 1 \end{pmatrix}$$







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

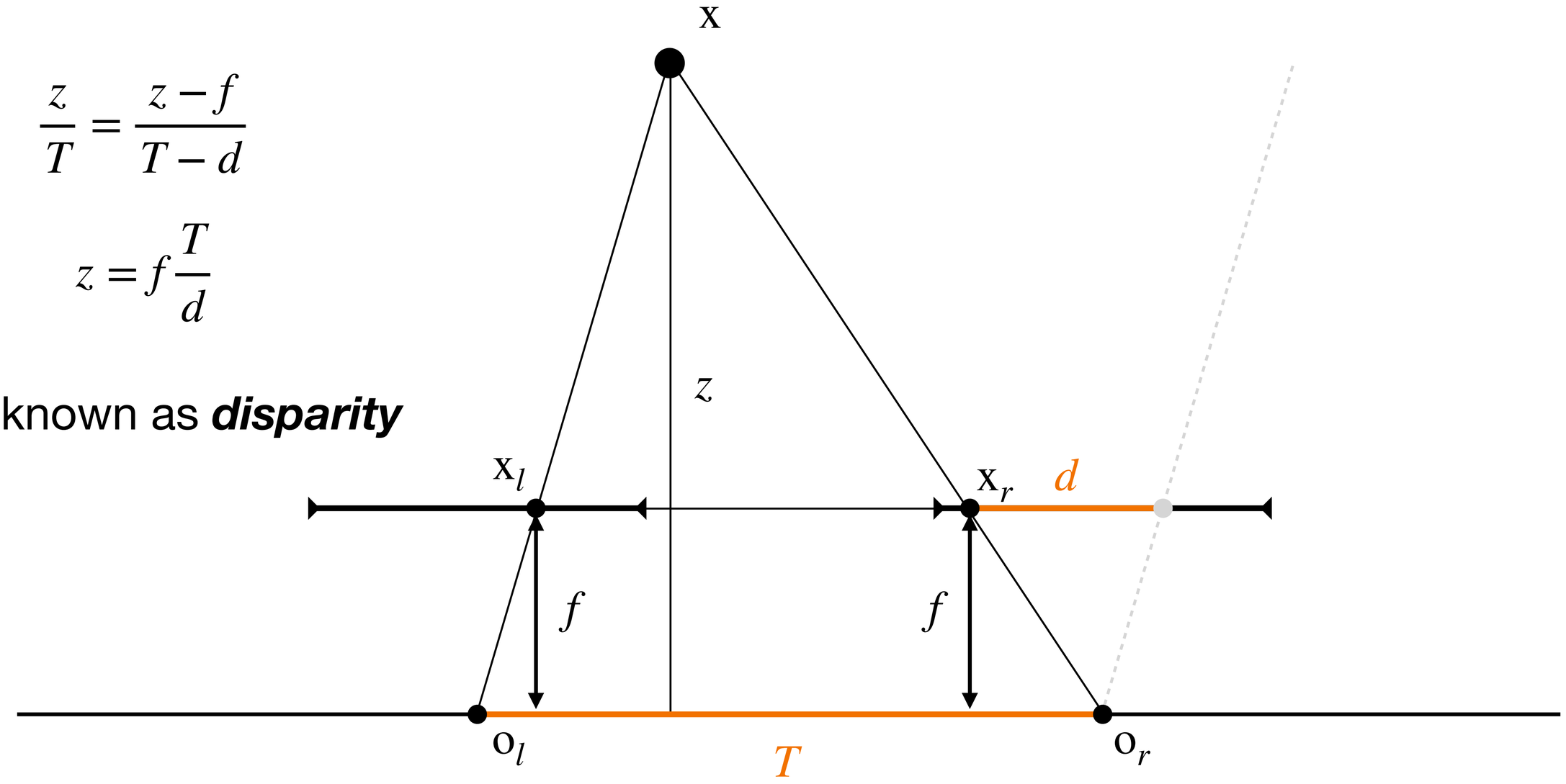
Binocular Vision

Depth Estimation in Two-View Geometry

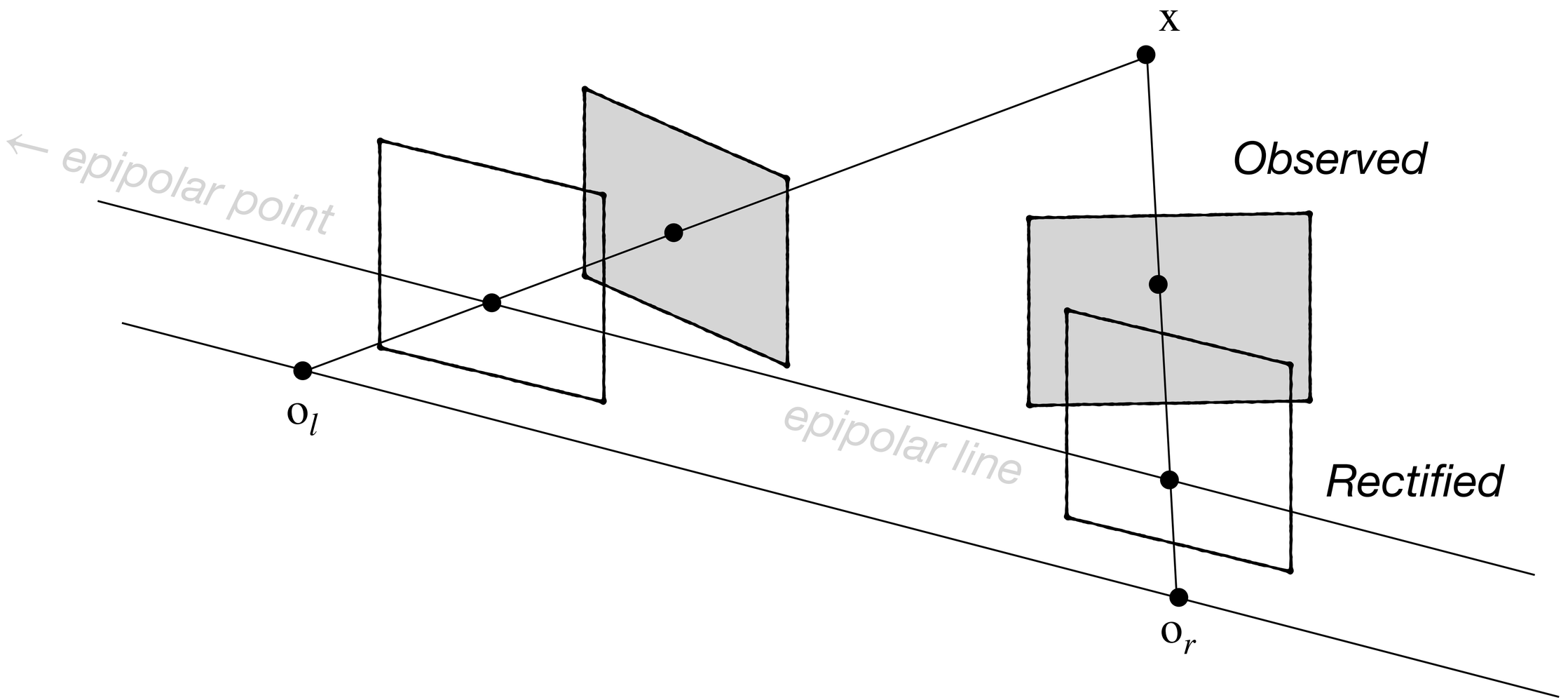
$$\frac{z}{T} = \frac{z-f}{T-d}$$

$$z = f\frac{T}{d}$$

d is known as ***disparity***



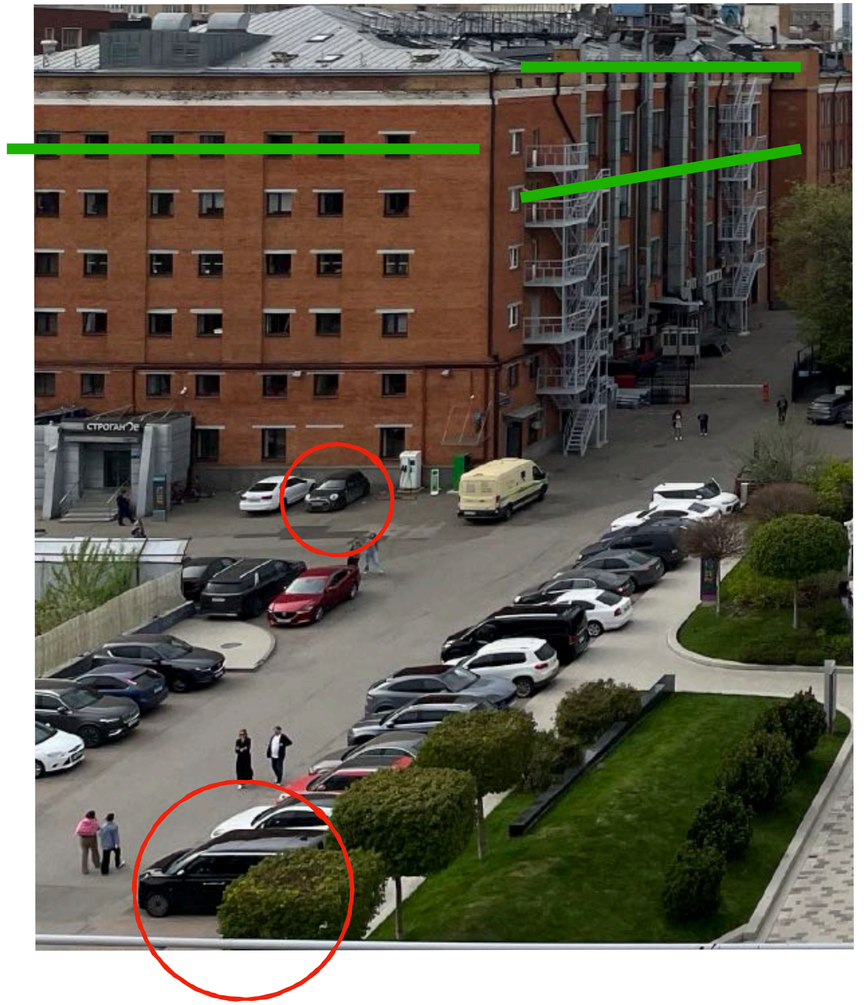
Rectification



Difficulties in Estimating Disparity

Learning-Based Approaches

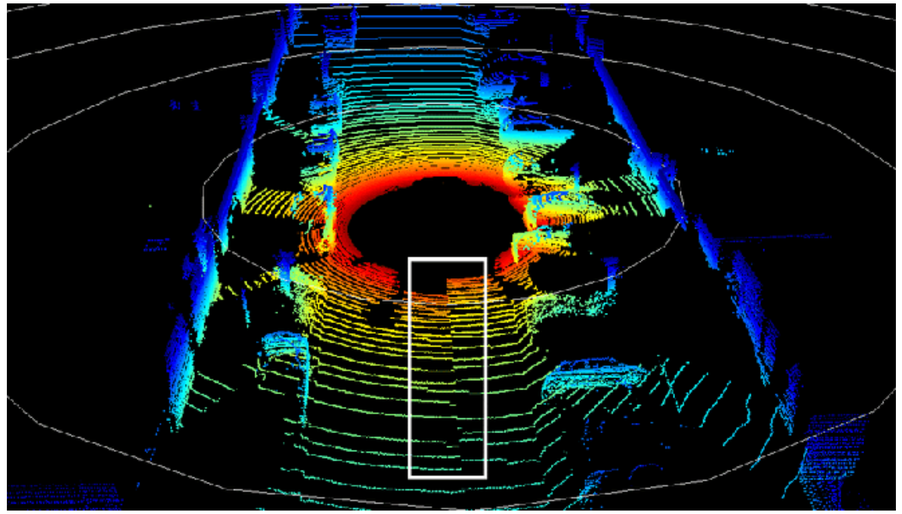
- Heuristics for disparity estimation have limited success
- We perceive depth even without binocular vision
  - There are many cues around us

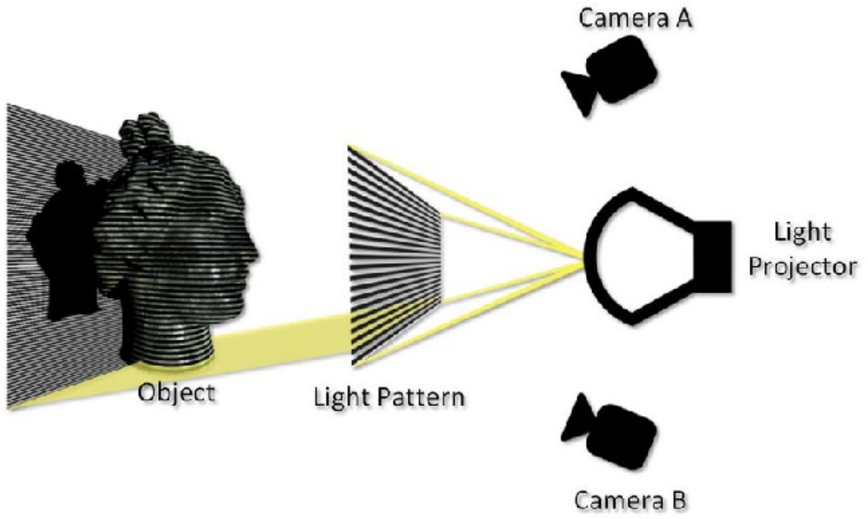


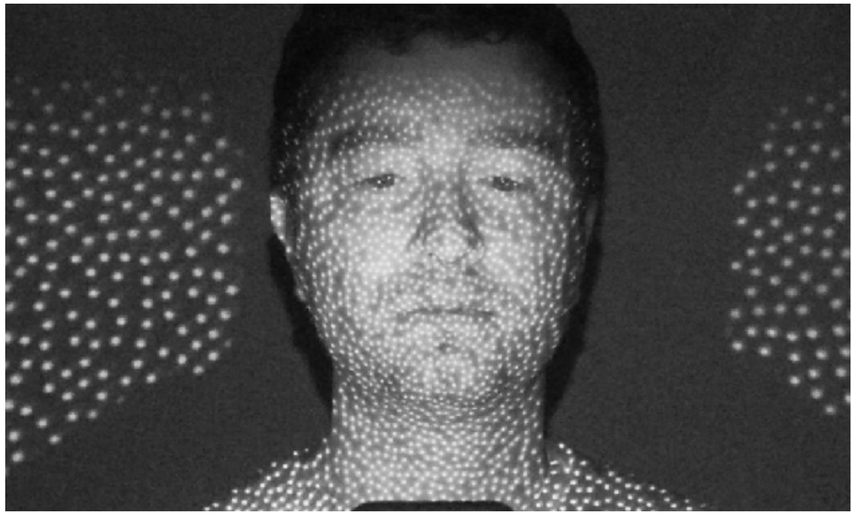
http://www.perceptionsense.com/2010/06/ames-room.html

Data

- Input: mono / stereo images
- Target: depth sensors
  - Structured light
  - Time of flight
  - Lidar
- Target: SfM
- Target: synthetic



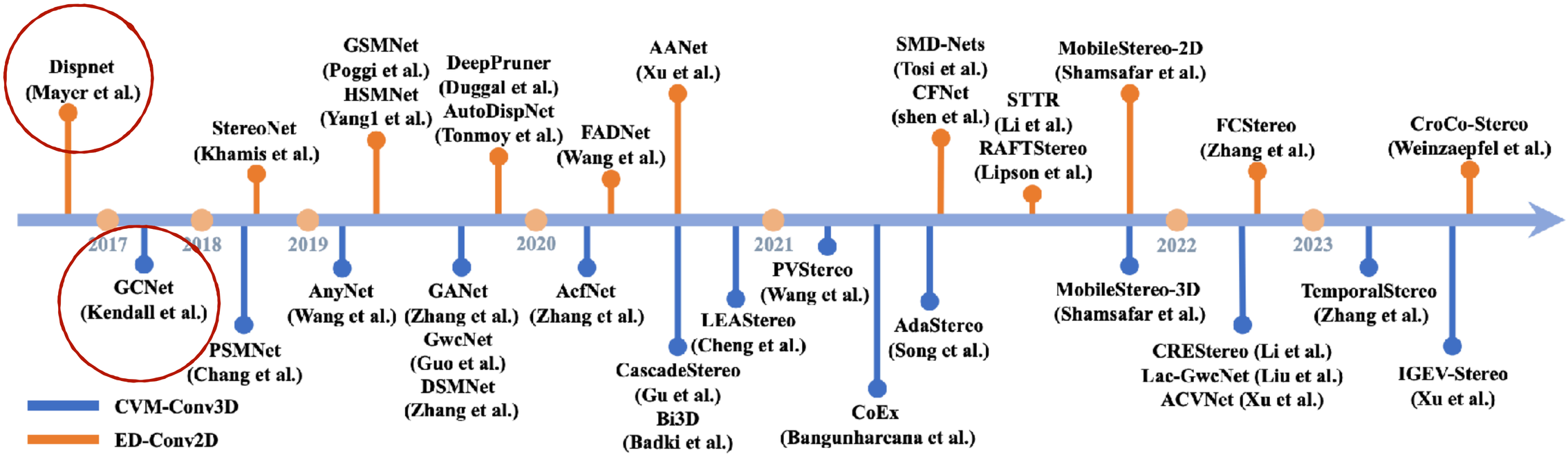




https://www.researchgate.net/figure/A-LiDAR-frame-in-the-KITTI-odometry-dataset-The-point-cloud-is-colored-by-the-laser-ID_fig2_351790098
https://www.idropnews.com/news/app-developer-access-to-truedepth-camera-worries-privacy-advocates/57116/

Stereo Depth: Disparity Estimation

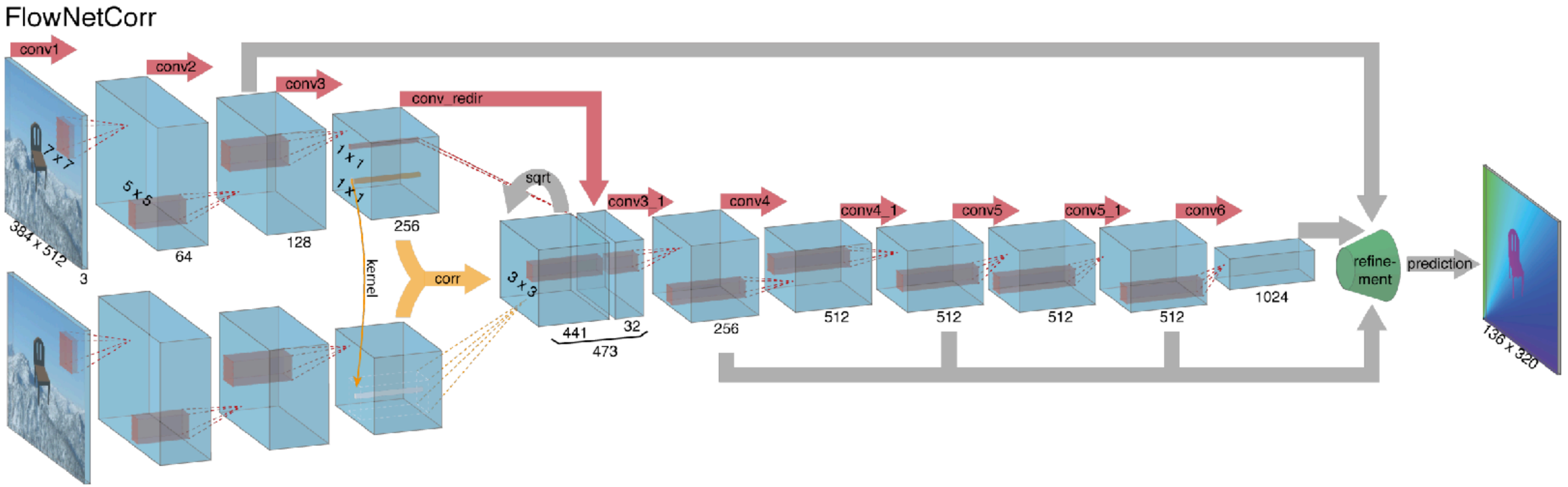
fast, based on 2D convolutions



high precision, uses 3D convolutions

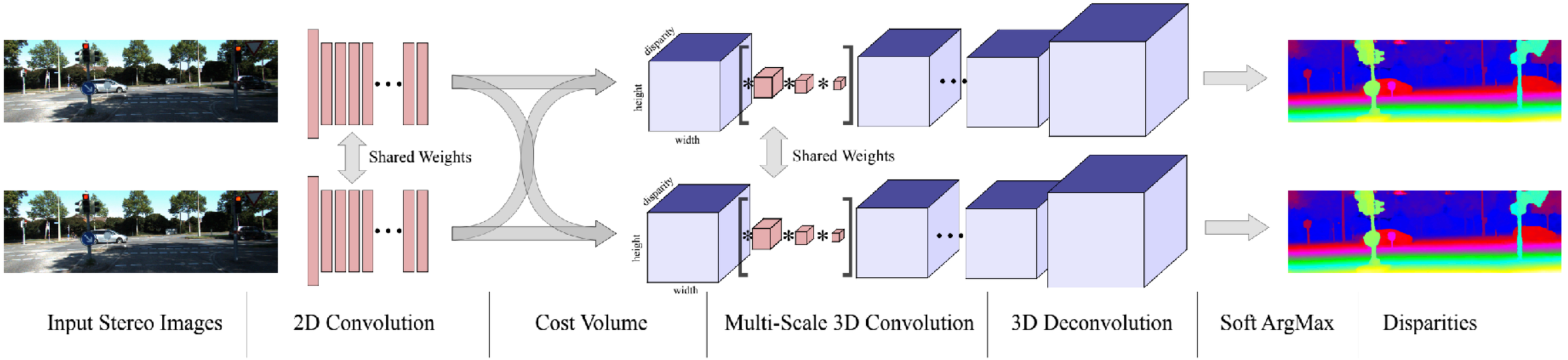
https://arxiv.org/abs/2312.00343

DispNet



Dosovitskiy A. et al. Flownet: Learning optical flow with convolutional networks
Mayer N. et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation

GCNet



Kendall A. et al. End-to-end learning of geometry and context for deep stereo regression

Monocular Depth Estimation

Depth Map Prediction

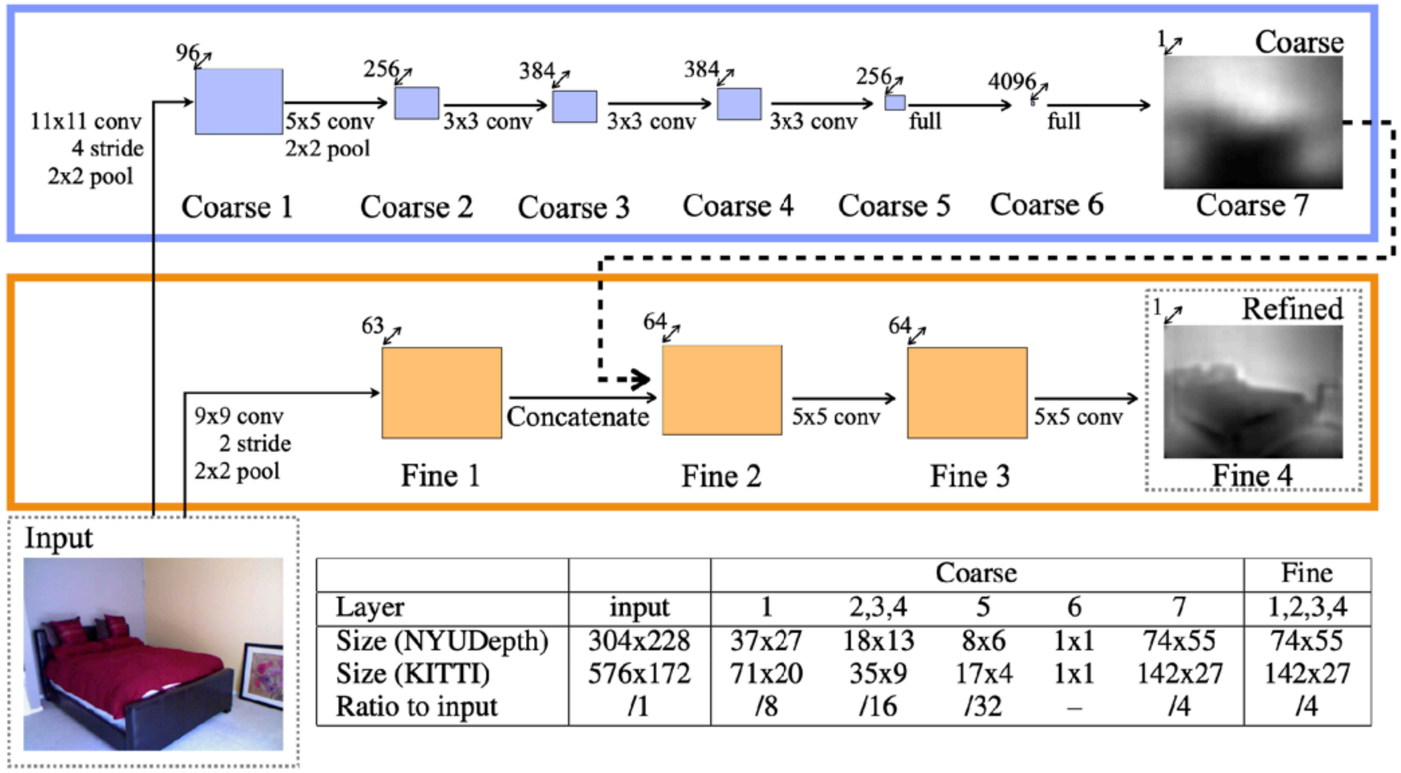
from a Single Image using a Multi-Scale Deep Network

- Scale invariant error for depths y and y^*
- Denote $d_i = \log y_i - \log y_i^*$

$$D(y, y^*) = \frac{1}{n^2} \sum_{i,j} \left((\log y_i - \log y_j) - (\log y_i^* - \log y_j^*) \right)^2$$

$$= \frac{1}{n} \sum_i d_i^2 - \frac{1}{n^2} \sum_{i,j} d_i d_j$$

$$= \frac{1}{n} \sum_i d_i^2 - \frac{\lambda}{n^2} \left(\sum_i d_i \right)^2$$



		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

https://arxiv.org/abs/1406.2283

Depth Map Prediction

from a Single Image using a Multi-Scale Deep Network

Towards Robust Monocular Depth Estimation

Mixing Datasets for Zero-Shot Cross-Dataset Transfer

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

https://arxiv.org/abs/1907.01341

Towards Robust Monocular Depth Estimation

Mixing Datasets for Zero-Shot Cross-Dataset Transfer

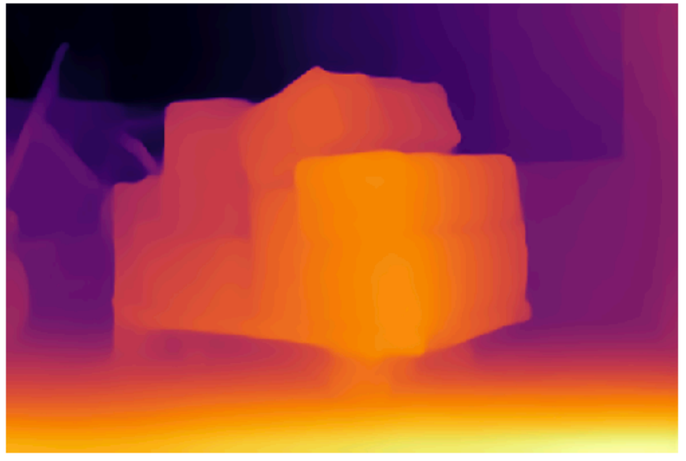
- In some datasets, intrinsic parameters of the camera are unknown
- As a result, disparity (target) is *shifted*
- MiDaS: predict disparity *up to scale and shift*

$$L_{ssi}(\hat{d}, \hat{d}^*) = \frac{1}{2M} \sum_{i=1}^{M} \rho(\hat{d}_i - \hat{d}_i^*)$$

- Where $\hat{d}_i = s d_i + t$ and $\hat{d}_i^* = d_i^*$ for $(s, t) = \arg\min_{s', t'} \sum_i (s' d_i + t' - d_i^*)^2$
- Architecture similar to U-Net

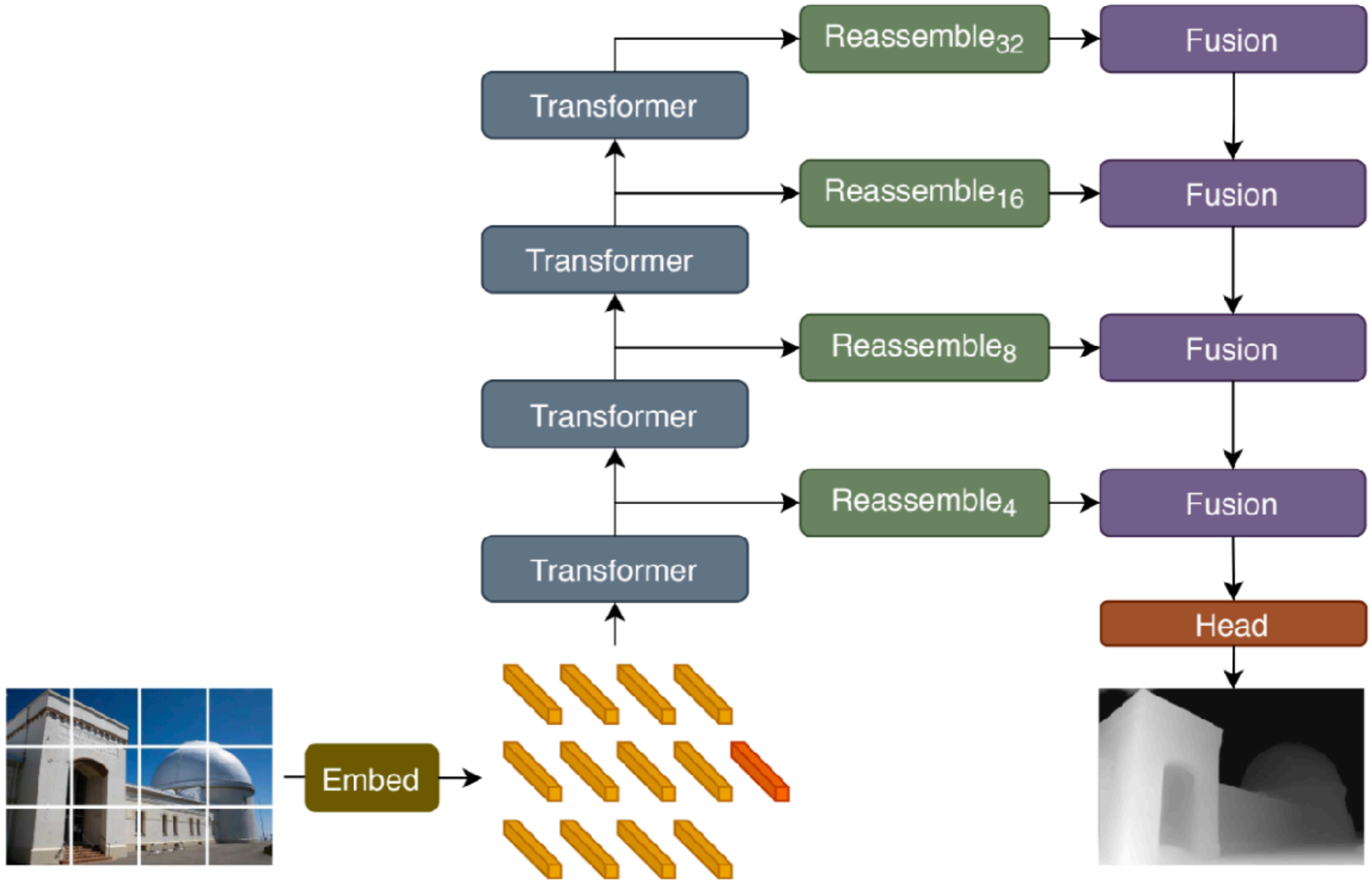




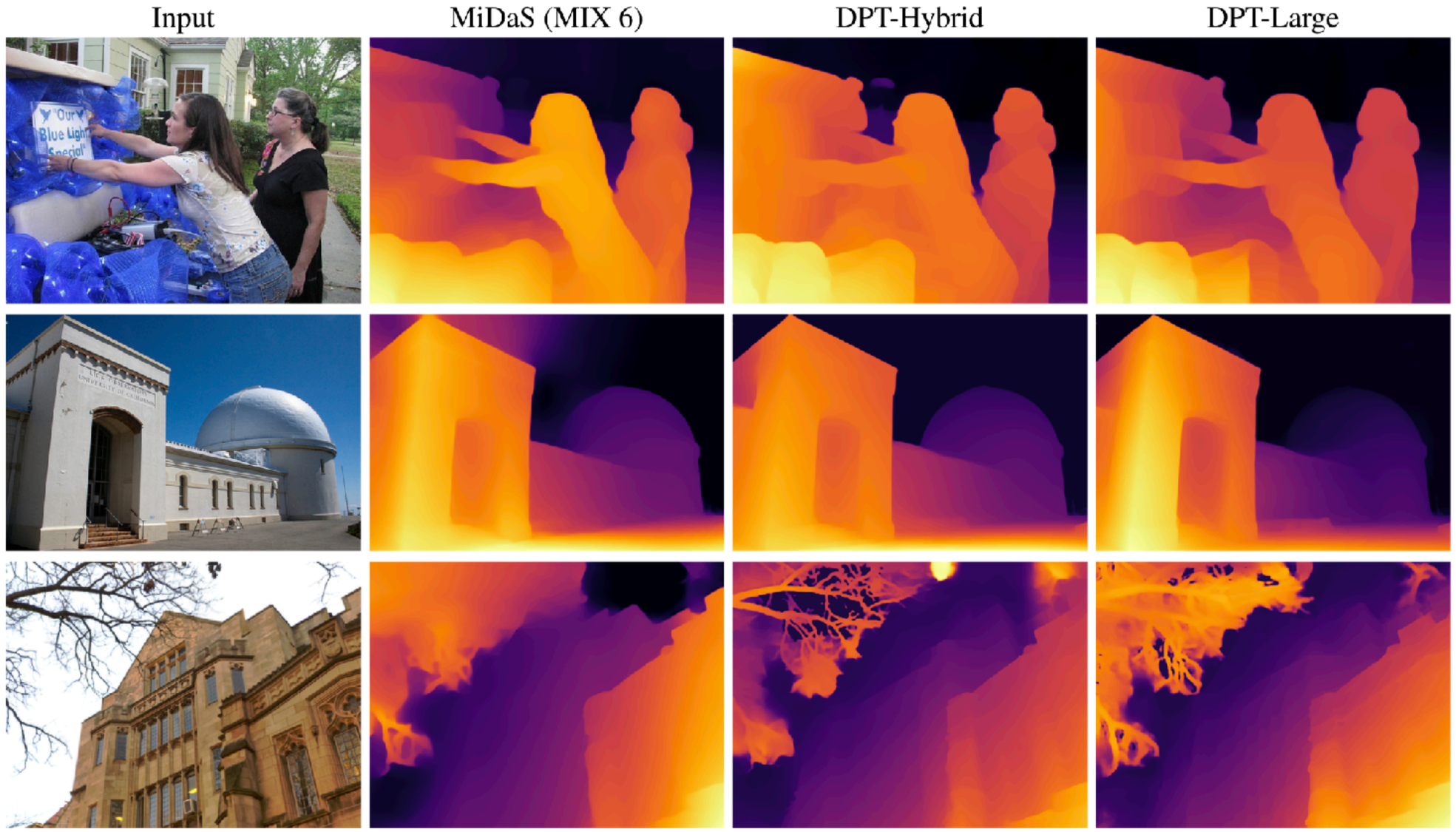




Vision Transformers for Dense Prediction



From Convolutions to Transformers

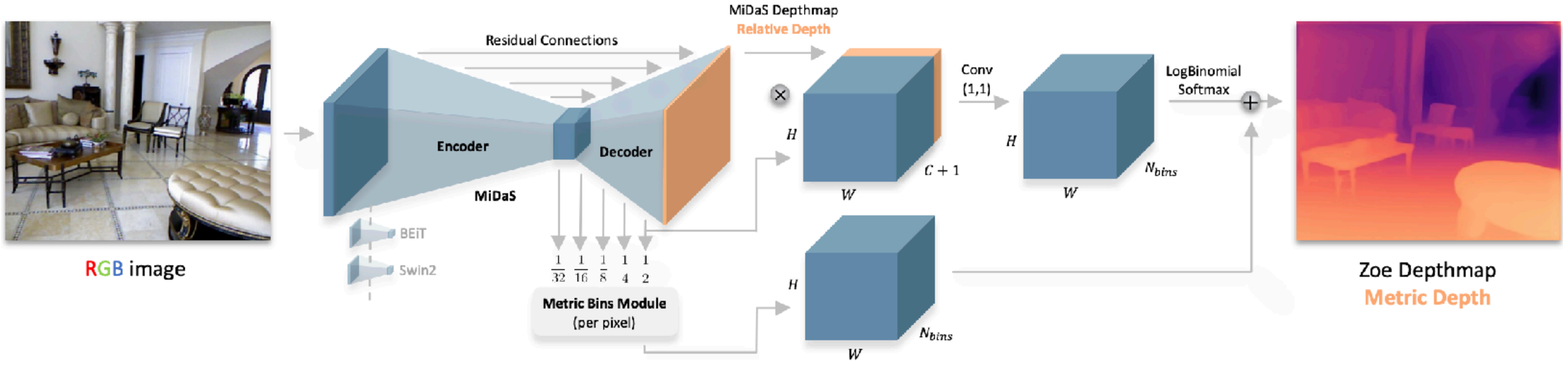


Failure Cases of MiDaS

ZoeDepth

Zero-shot Transfer by Combining Relative and Metric Depth

- Can we get metric depth?
- Pre-train on a mixture of datasets as in MiDAS
- Fine-tune a module that calibrates depth map



https://arxiv.org/abs/2302.12288, https://github.com/isl-org/ZoeDepth

Key Takeaways

- Depth maps allow approximating 3D structure
- Stereo depth estimation transforms into ***disparity*** estimation
- Scaling significantly improved monocular depth estimation

- Next time: point clouds