Predicting 2.5D / 3D

CS 280 2025

Angjoo Kanazawa

With many useful slides from Shubham Tulsiani

Monocular Depth Estimation

Depth from a Single Image

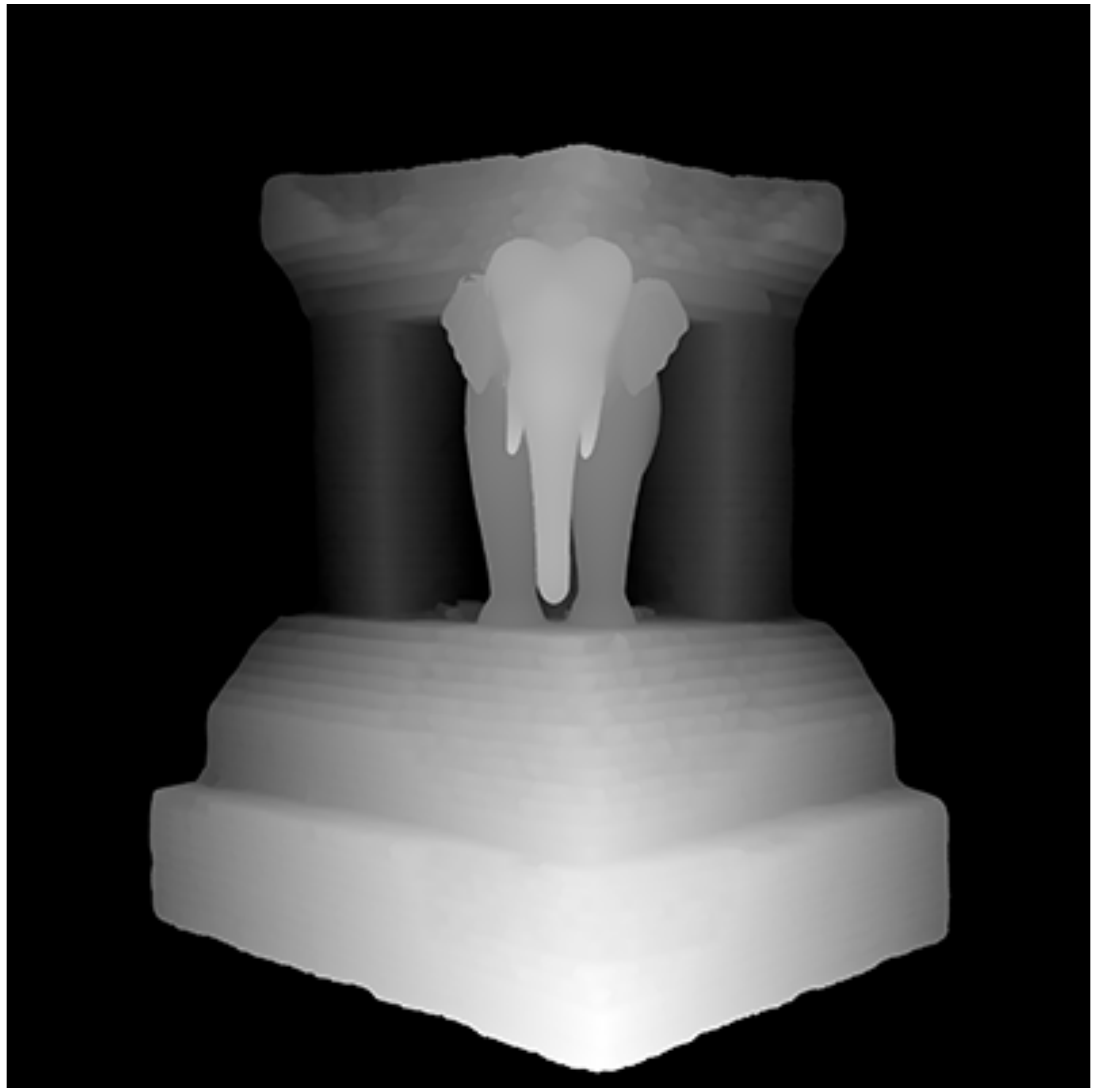
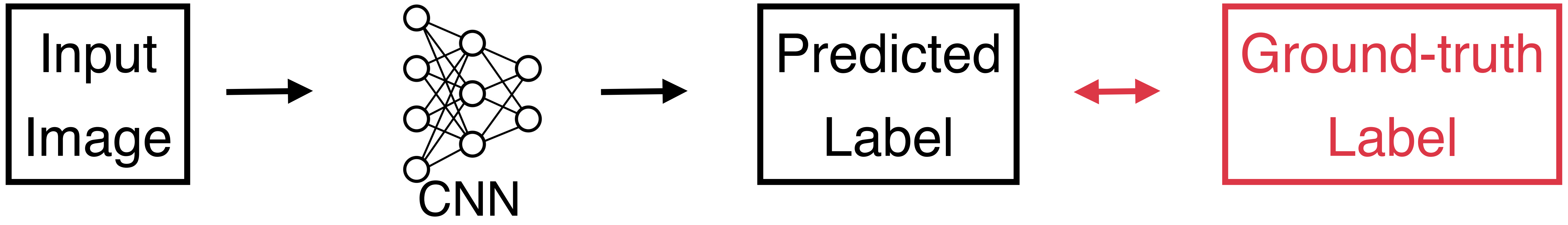


Image credits: Paul Bourke

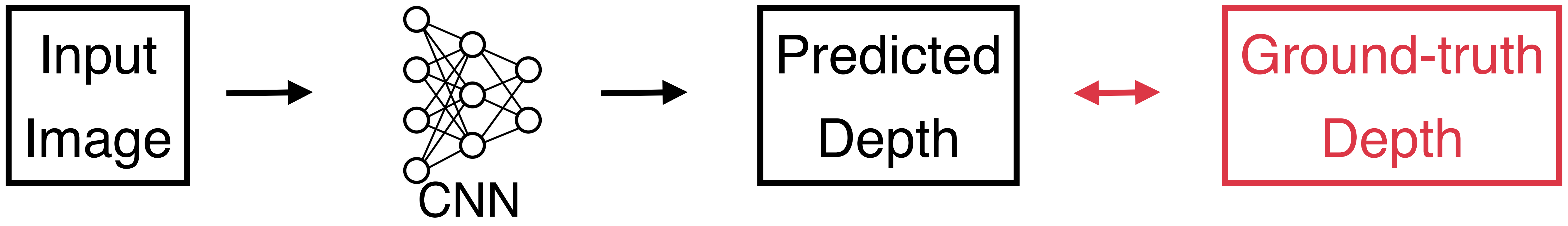
Learning from Direct Supervision



IMAGENET

COCO
Common Objects in Context

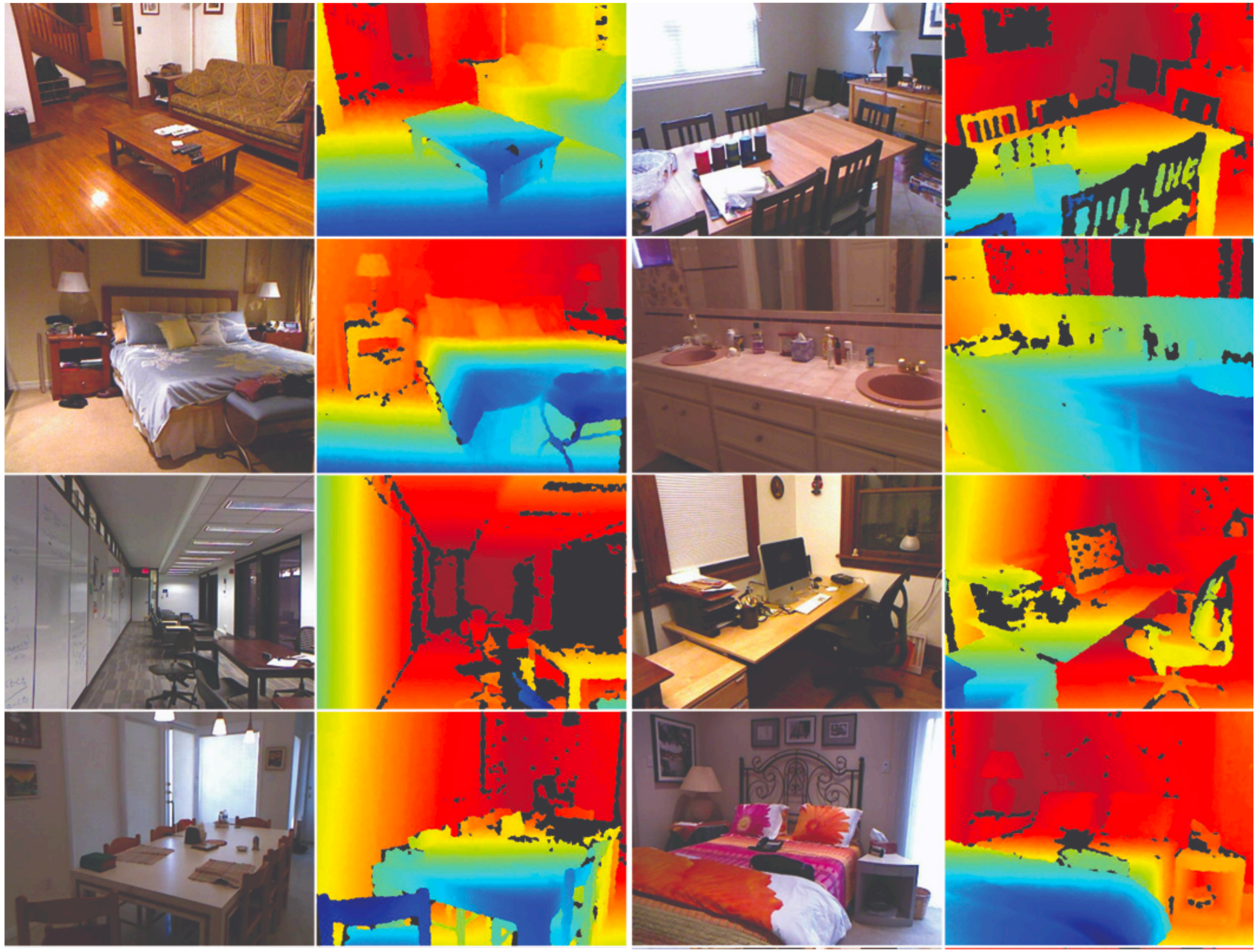
Learning from Direct Supervision



A caricature recipe for learning:

- Step 0: Decide on model and objectives
- Step 1: Collect training data (lots of [image, depth] pairs)
- Step 2: Learn a predictor
  - Step 2a: Wait a few days, drink coffee and watch training curves
- Use the predictor!

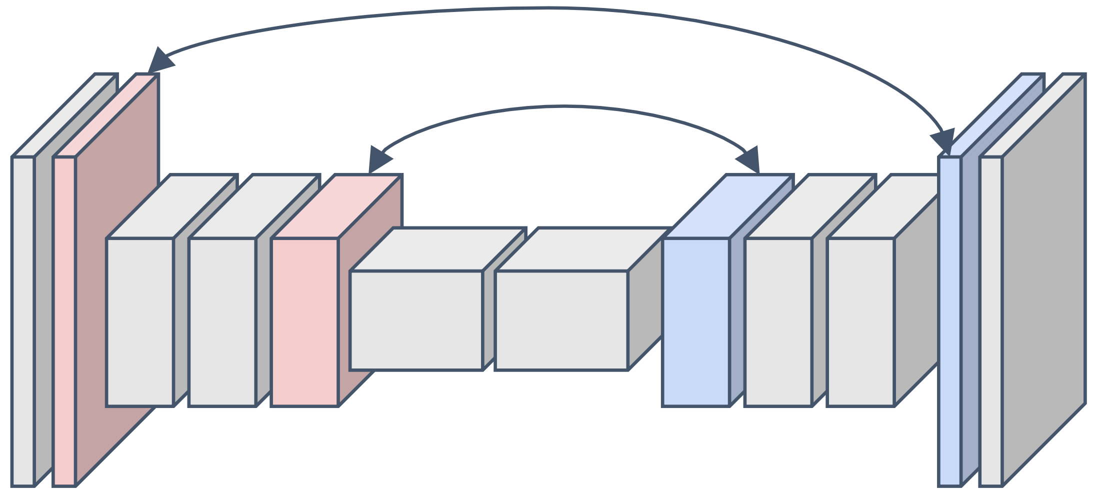
Capturing Depth



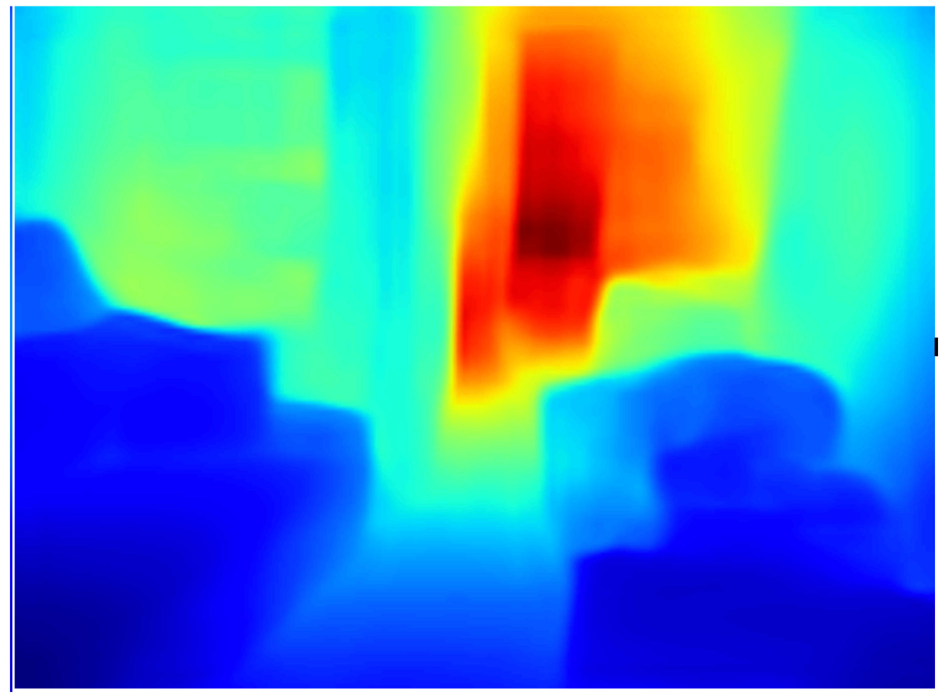
NYU Dataset. Silberman et. al.

Depth Prediction: An initial Approach

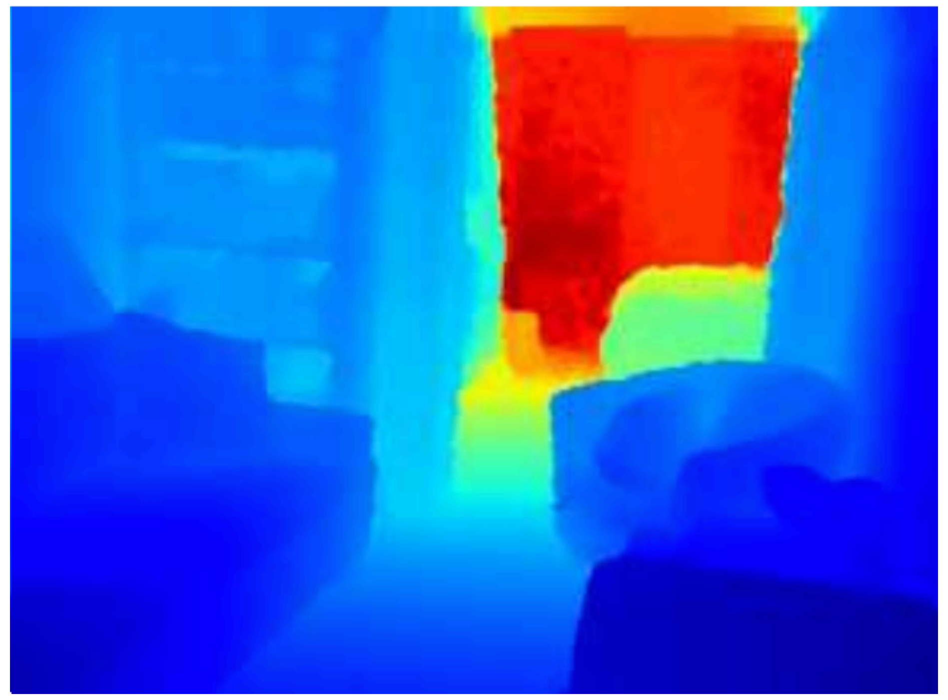
Input Image



Convolutional Network



Predicted Depth



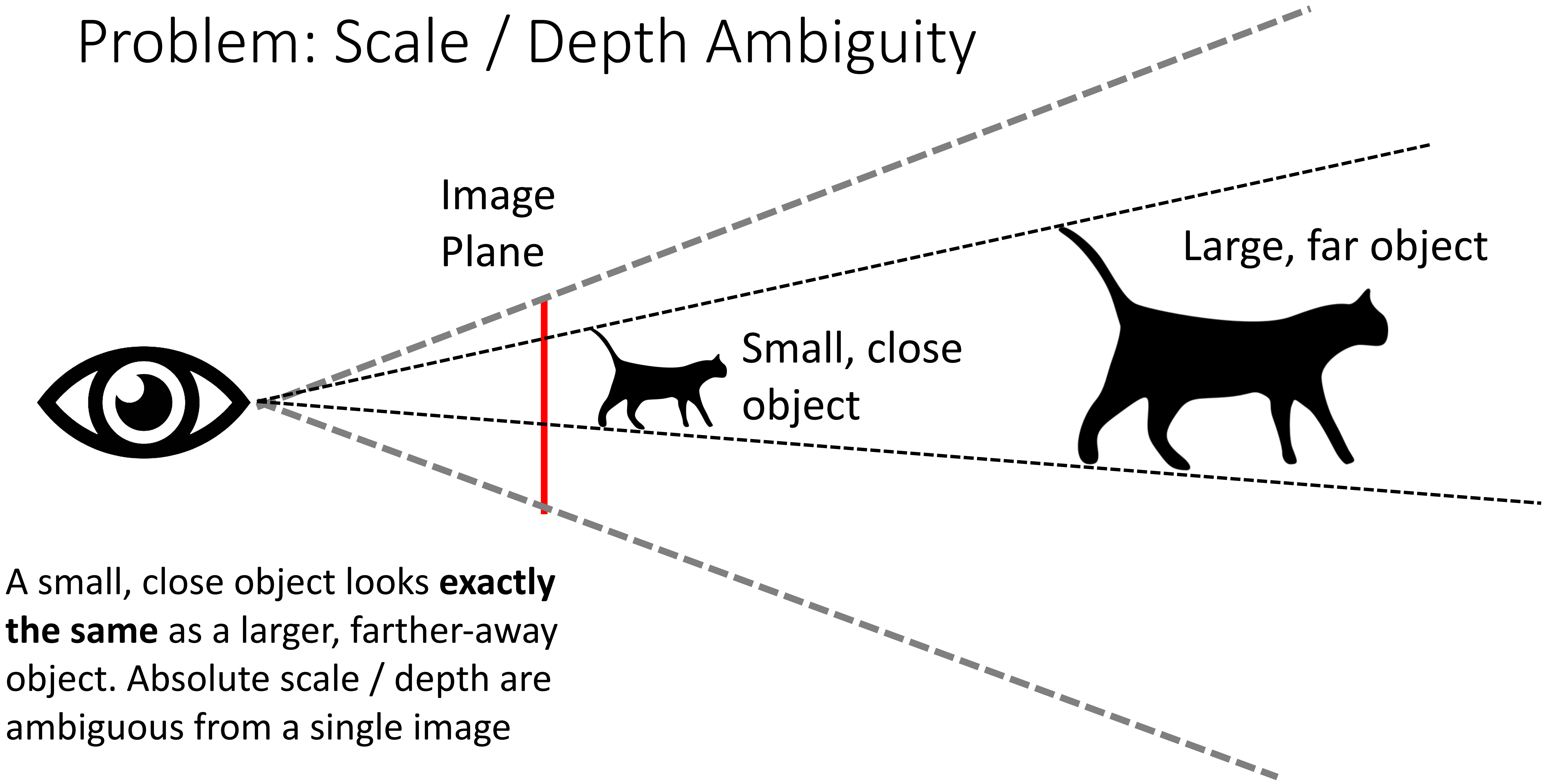
GT Depth

$$L(y, y^*) = \sum_i \|y_i - y_i^*\|^2$$

Slide credit: Justin Johnson

Depth Prediction

Problem: Scale / Depth Ambiguity



A small, close object looks **exactly the same** as a larger, farther-away object. Absolute scale / depth are ambiguous from a single image

Need a scale-***invariant*** learning objective:

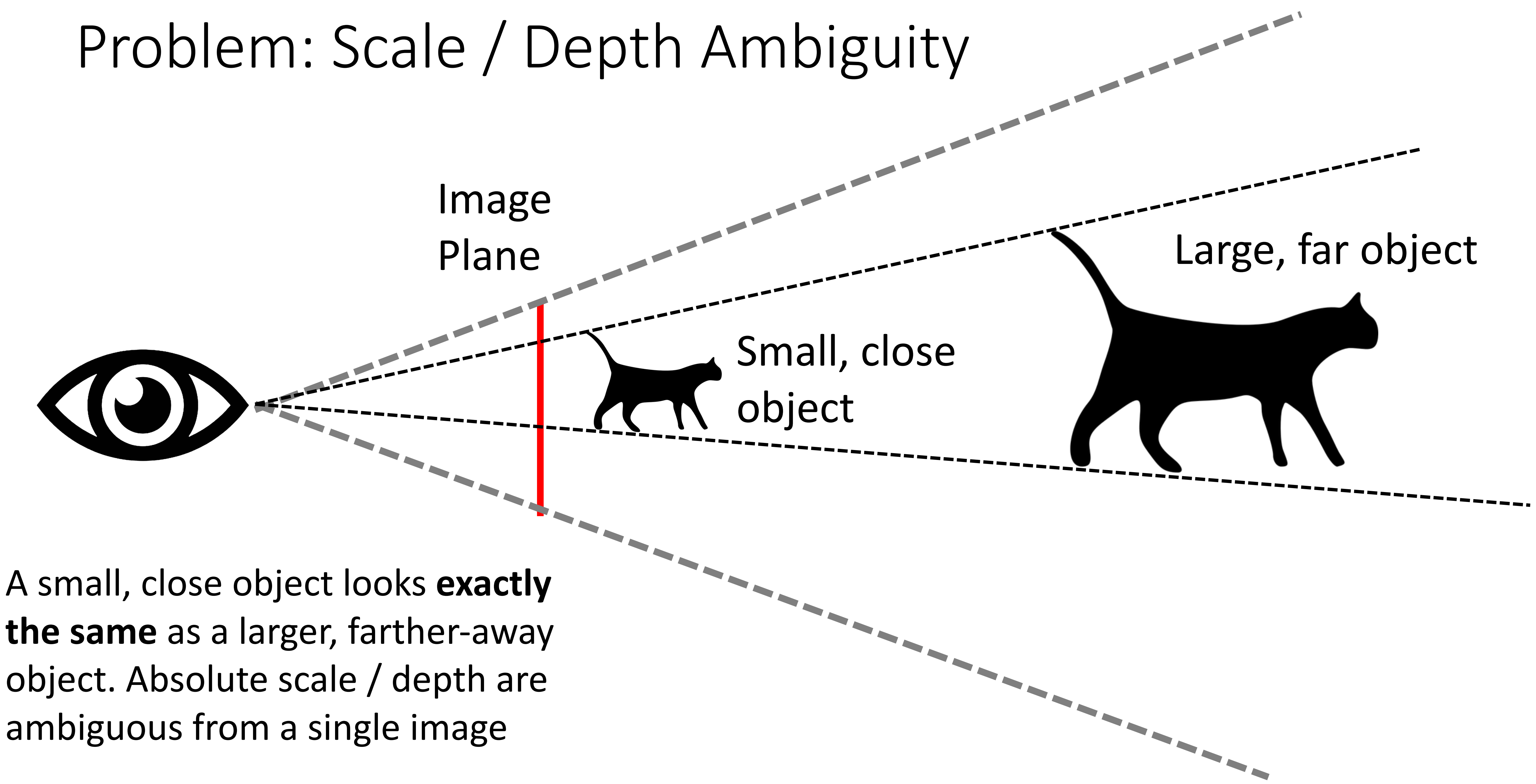
$$L(y, y^*) = L(\alpha y, y^*)$$

(for any scalar)

$$L(y, y^*) = \sum_i \|y_i - y_i^*\|^2$$

Image credit: Justin Johnson

Depth Prediction



A small, close object looks **exactly the same** as a larger, farther-away object. Absolute scale / depth are ambiguous from a single image

Use a scale-***invariant*** learning objective:

$$L(y, y^*) = L(\alpha y, y^*)$$

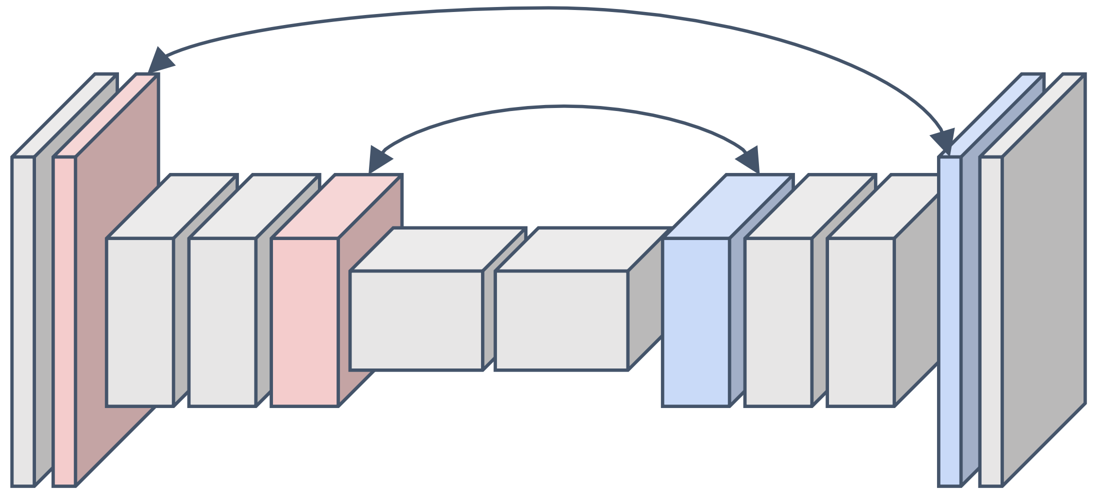
(for any scalar)

$$\min_{\alpha} L(\alpha y, y^*)$$

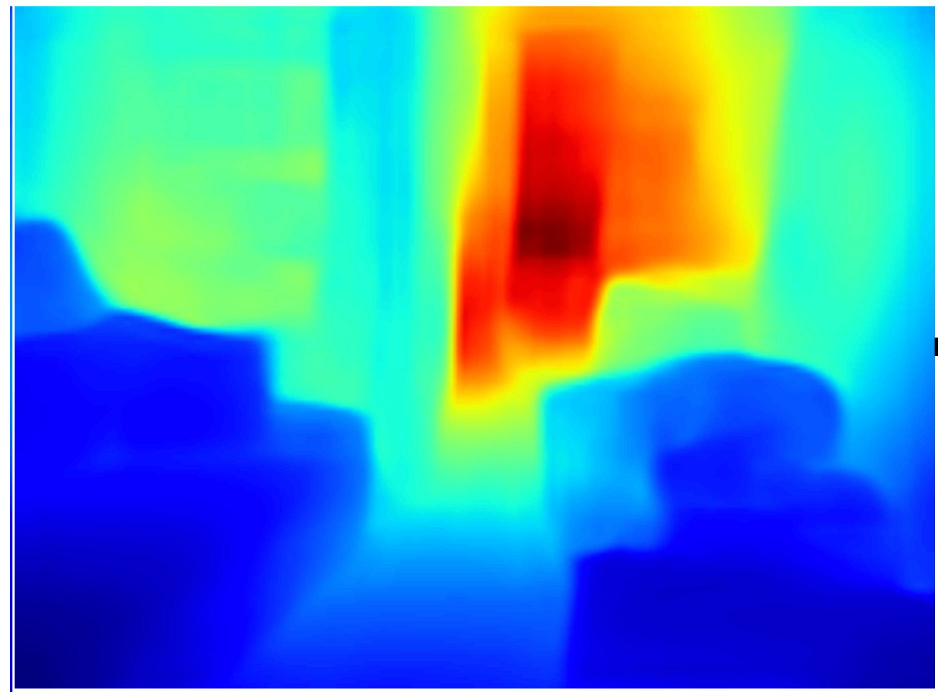
Solve for the alpha that minimizes the loss the most, and minimize that the loss

Depth Prediction via a Scale-invariant loss

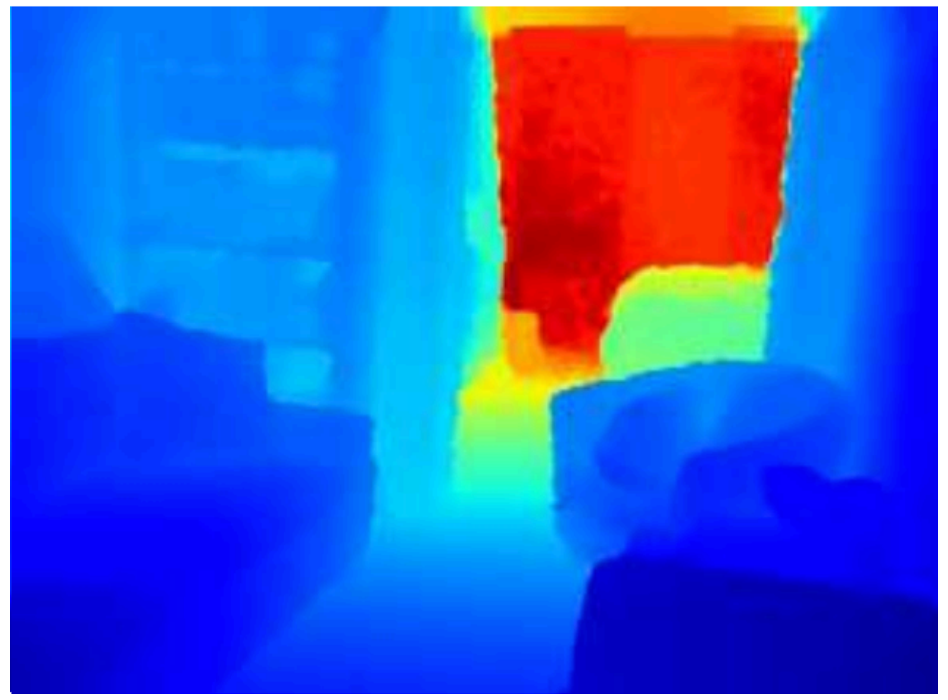
Input Image



Convolutional Network



Predicted Depth



GT Depth

$$L(y, y^*) = \sum_i \| \log y_i - \log y_i^* + \alpha(y, y^*) \|^2$$

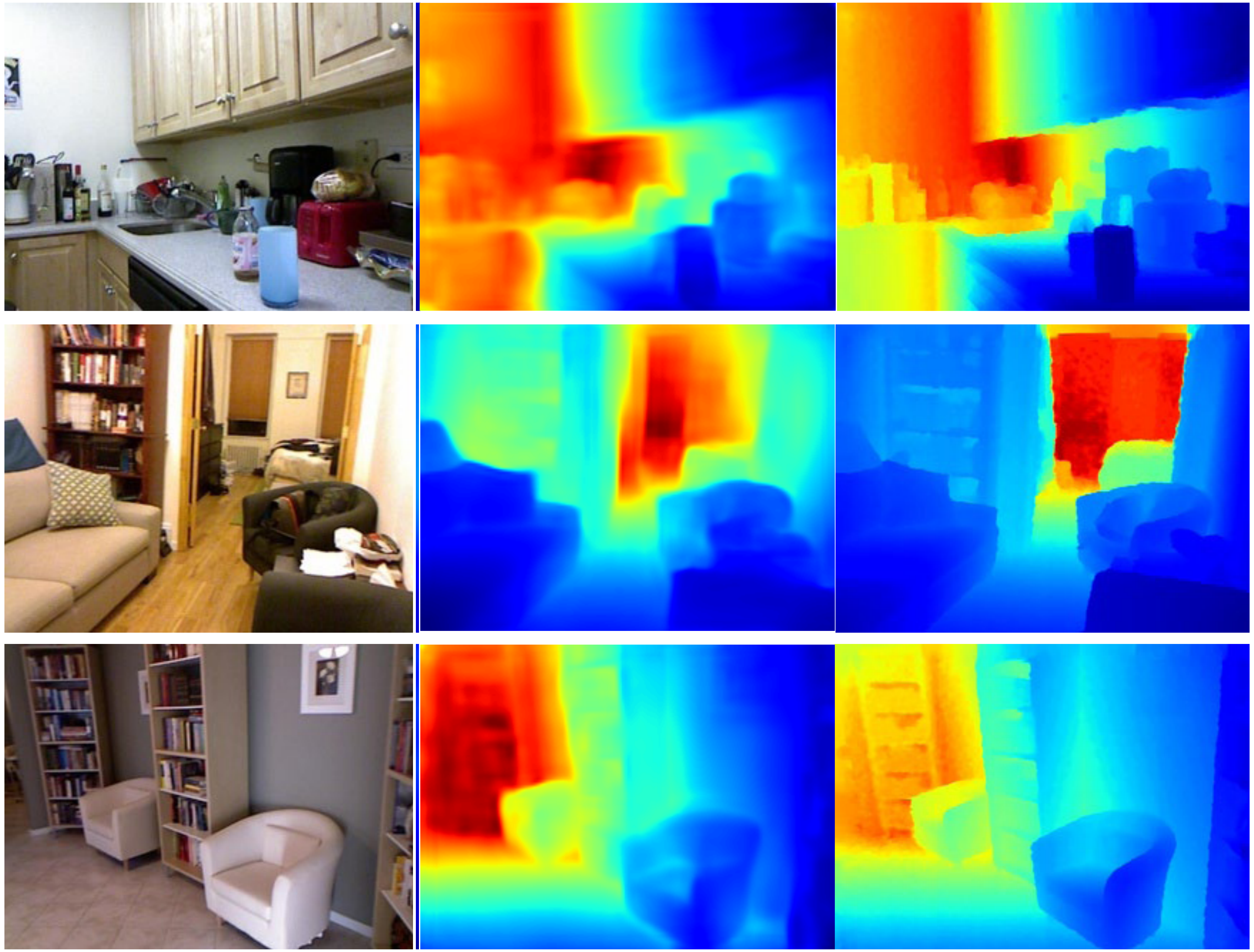
$$\alpha(y, y^*) = \frac{1}{n} \sum_i (\log y_i^* - \log y_i)$$

Solution to $\alpha y - y*$ in log-space

Depth Map Prediction from a Single Image using a Multi-scale Deep Network. Eigen, Puhrsch, and Fergus. NeurIPS 2014

Slide credit: Justin Johnson

Depth Prediction: Sample Results



- Accurate coarse estimates

- Inaccurate around boundaries

Depth Map Prediction from a Single Image using a Multi-scale Deep Network. Eigen, Puhrsch, and Fergus. NeurIPS 2014

Improving Depth Prediction

1. More data!

2. Training Objectives
   - Alternate scale-invariant losses?
   - Better regularizers?

3. Improved Architectures

$$\alpha(y, y^*) = \frac{1}{n}\sum_i (\log y_i^* - \log y_i)$$

3D movies dataset

Fig. 2. Sample images from the 3D movies dataset. We show images from some of the films in the training set together with their inverse depth maps. Sky regions and invalid pixels are masked out. Each image is taken from a different film. 3D movies provide a massive source of diverse data.

Depth Datasets

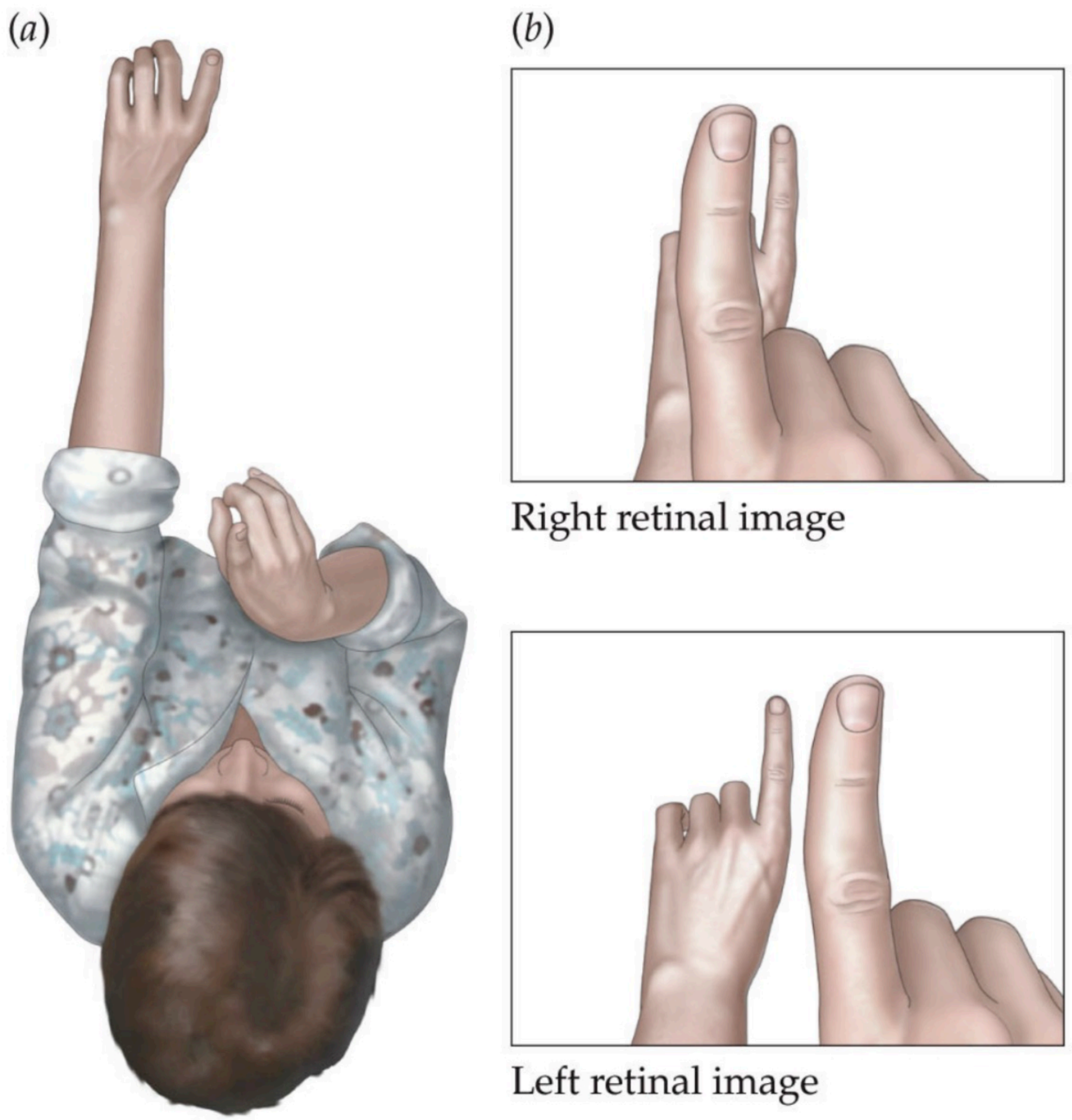
Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer

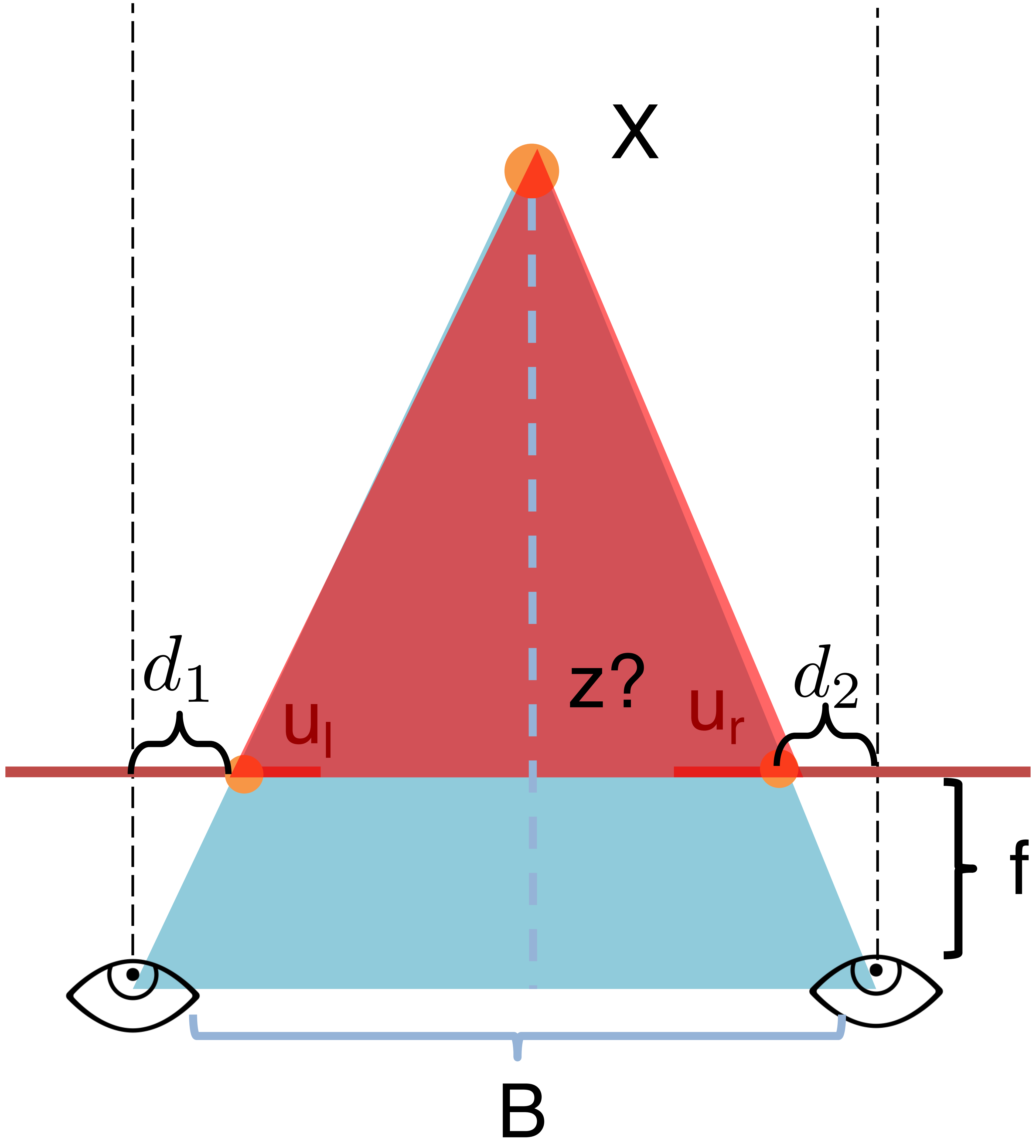
René Ranftl*, Katrin Lasinger*, David Hafner, Konrad Schindler, and Vladlen Koltun

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	$\boxed{\text{No scale \& shift}}$	3600
WSVD [33]	✓	✓	✓	✓	✓	Medium	**High**	Stereo	$\boxed{\text{No scale \& shift}}$	1.5M
3D Movies	✓	✓	✓	✓	✓	Medium	**High**	Stereo	$\boxed{\text{No scale \& shift}}$	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

Disparity

Recall…





Depth from Disparity

$$\text{disparity} = u_{\text{left}} - u_{\text{right}}$$

Lots of disparity = Near by

Small disparity = Far away

At infinity 0 movement

$$disparity = \frac{fb}{depth} \sim \frac{1}{depth}$$

Why is disparity a nice space to predict??

- Easy to bound [0, 1] with d_{max}
- Linear in inverse depth

Web Stereo Video Supervision for Depth Prediction from Dynamic Scenes. Wang et. al.

Scale and shift ambiguity still exists

$$disparity = \frac{fb}{depth} \sim \frac{1}{depth}$$

- Scale: Focal length and baselines are unknown!
- Shift: Values depend on d_{max}, which is image dependent
  - Principle points can also vary

$$disparity - (c_R - c_L) = \frac{fb}{depth}$$

c^{R}, c^{L}: principal point in left, right

- So also trained with scale & shift invariant loss!

Scale and Shift-invariant Depth Prediction

Towards Robust Monocular Depth Estimation:
Mixing Datasets for
Zero-shot Cross-dataset Transfer

René Ranftl*, Katrin Lasinger*, David Hafner, Konrad Schindler, and Vladlen Koltun

Trained jointly across many datasets

Scale and shift-invariant loss on disparity (inverse-depth):

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

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

Also use additional regularizers (e.g. gradients should match)

What this means

You still need to solve for scale and shift at test time!

- At test time you have to scale and shift it to get depth:

$$\hat{z} = \frac{1}{a \cdot \hat{d}_{\text{pred}} + b}$$

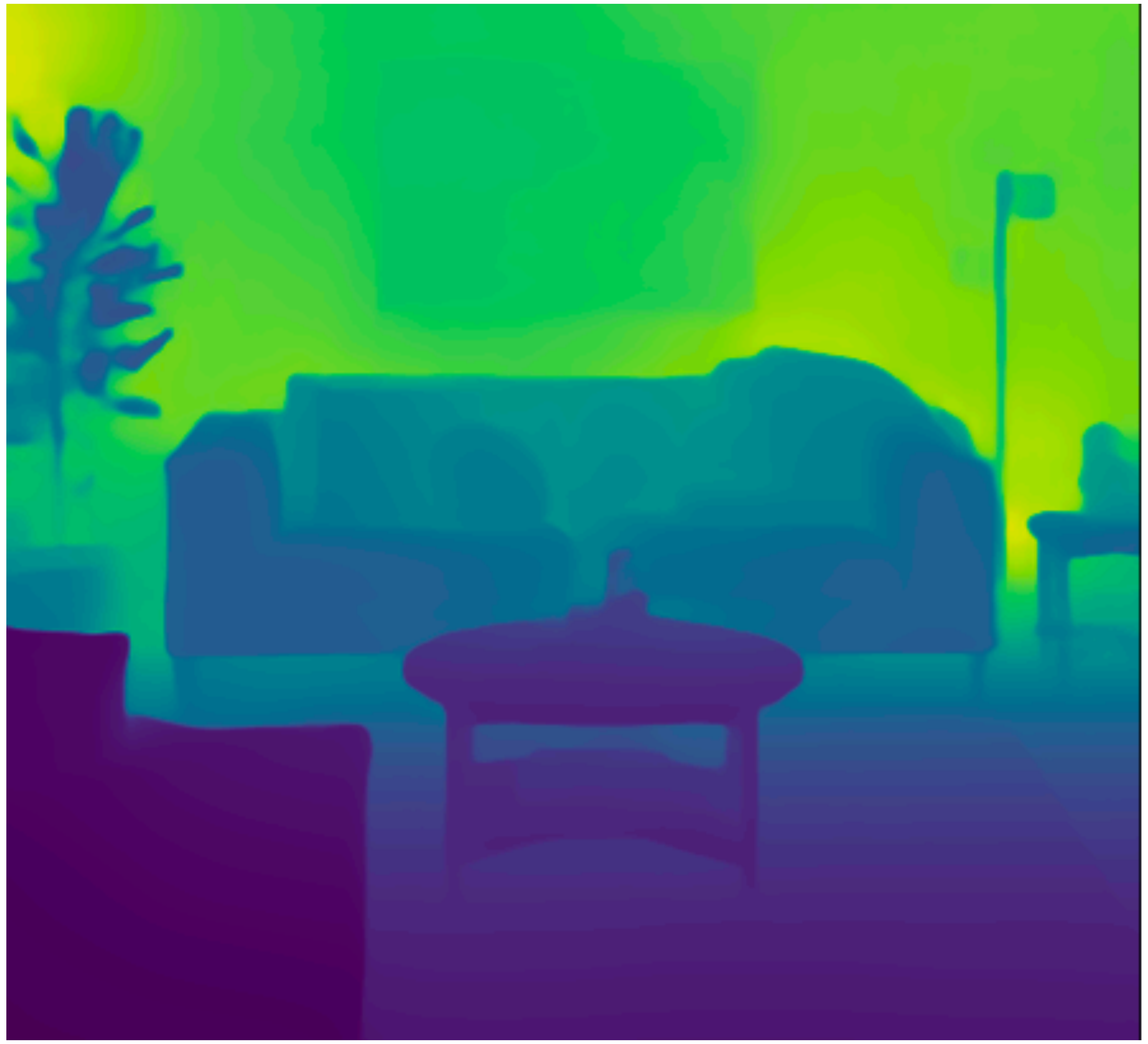
- Scale (a) : global stretch factor, for focal length * baseline
- Shift (b) : one global offset where to places the center of disparity
- Finicky…

Depth Prediction: Sample Results

Predictions of inverse depth (upto a scale and shift)

Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-dataset Transfer. Ranftl et. al.

Depth Prediction: Sample Results



Depth Prediction: Sample Results

Don’t judge a depth by its color — see prediction in 3D!

Sensitive to scale and shift

Learning to Recover 3D Scene Shape from a Single Image

CVPR 2021

Wei Yin†, Jianming Zhang‡, Oliver Wang‡, Simon Niklaus‡, Long Mai‡, Simon Chen‡, Chunhua Shen†*

† The University of Adelaide, Australia ‡ Adobe Research

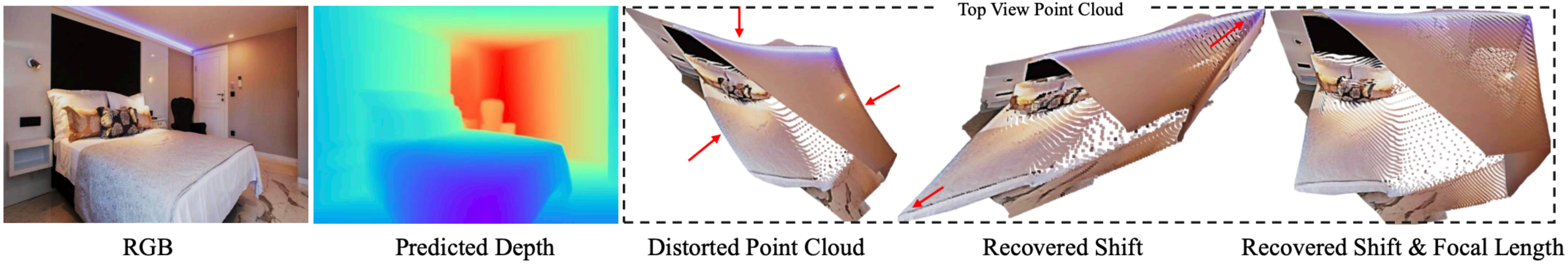
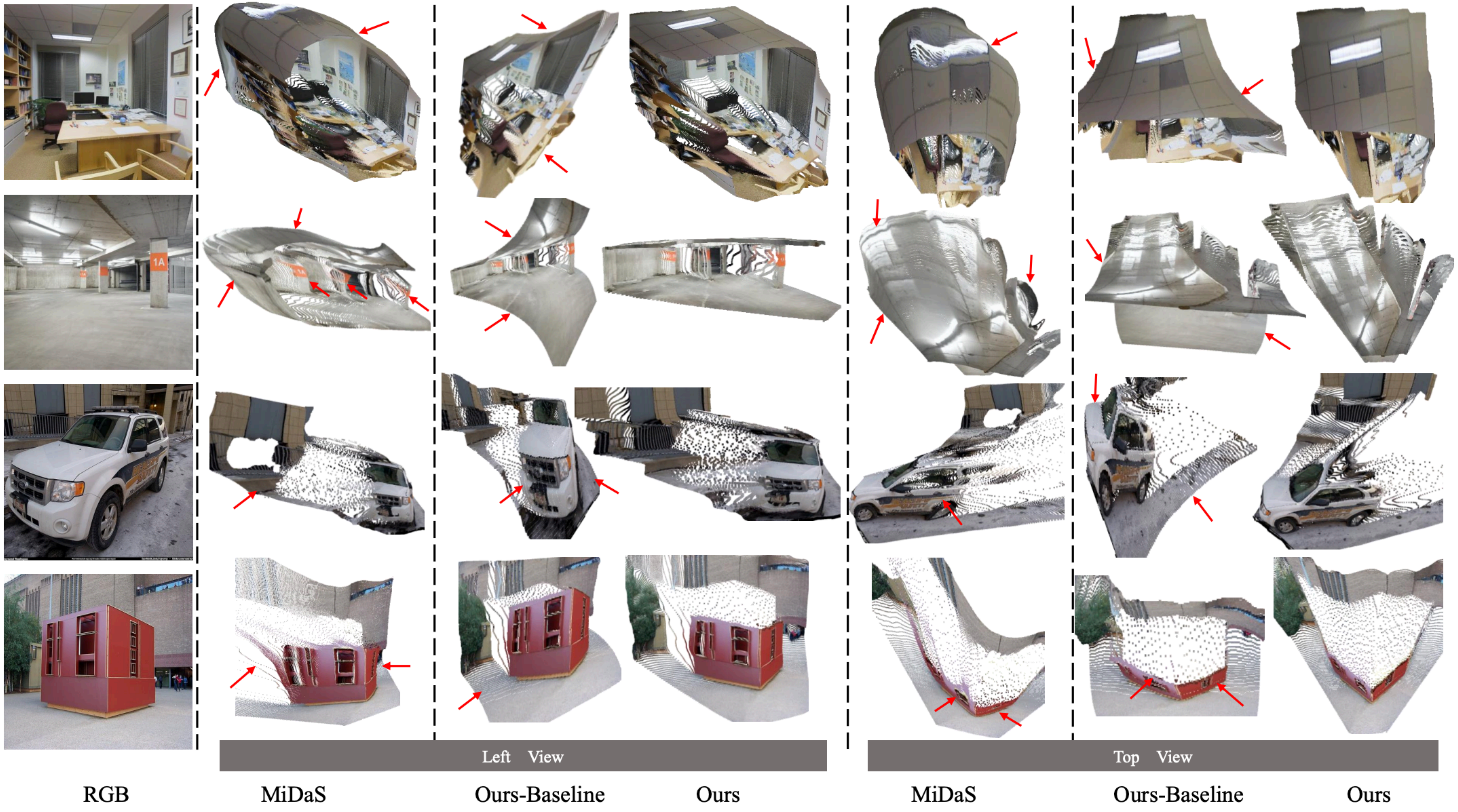


Figure 1: 3D scene structure distortion of projected point clouds. While the predicted depth map is correct, the 3D scene shape of the point cloud suffers from noticeable distortions due to an unknown depth shift and focal length (third column). Our method recovers these parameters using 3D point cloud networks. With recovered depth shift, the walls and bed edges become straight, but the overall scene is stretched (fourth column). Finally, with recovered focal length, an accurate 3D scene can be reconstructed (fifth column).

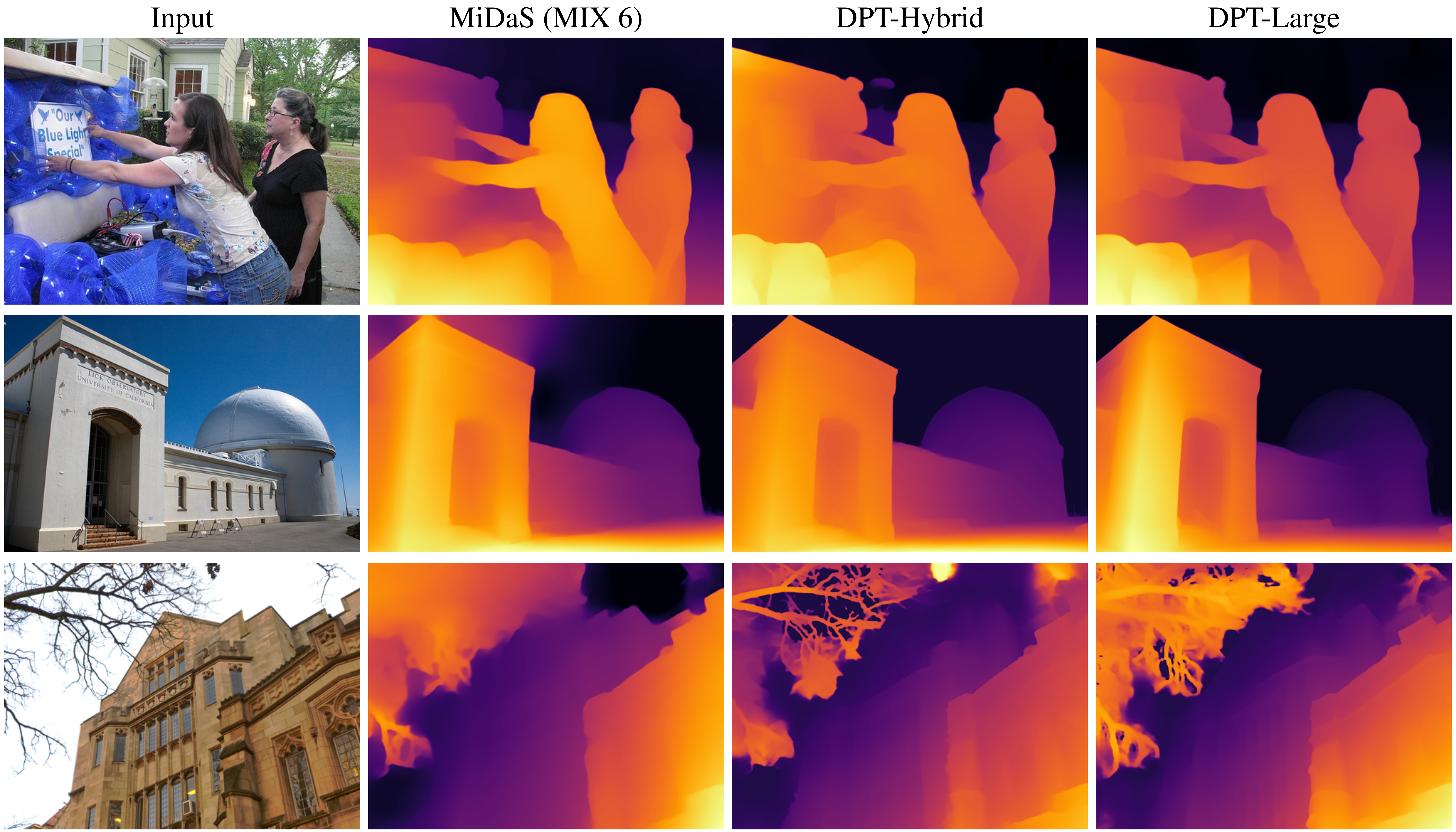
Same disparity output! Very different depth



Depth Prediction: An Active Research Area

Vision Transformers for Dense Prediction

René Ranftl Alexey Bochkovskiy Vladlen Koltun



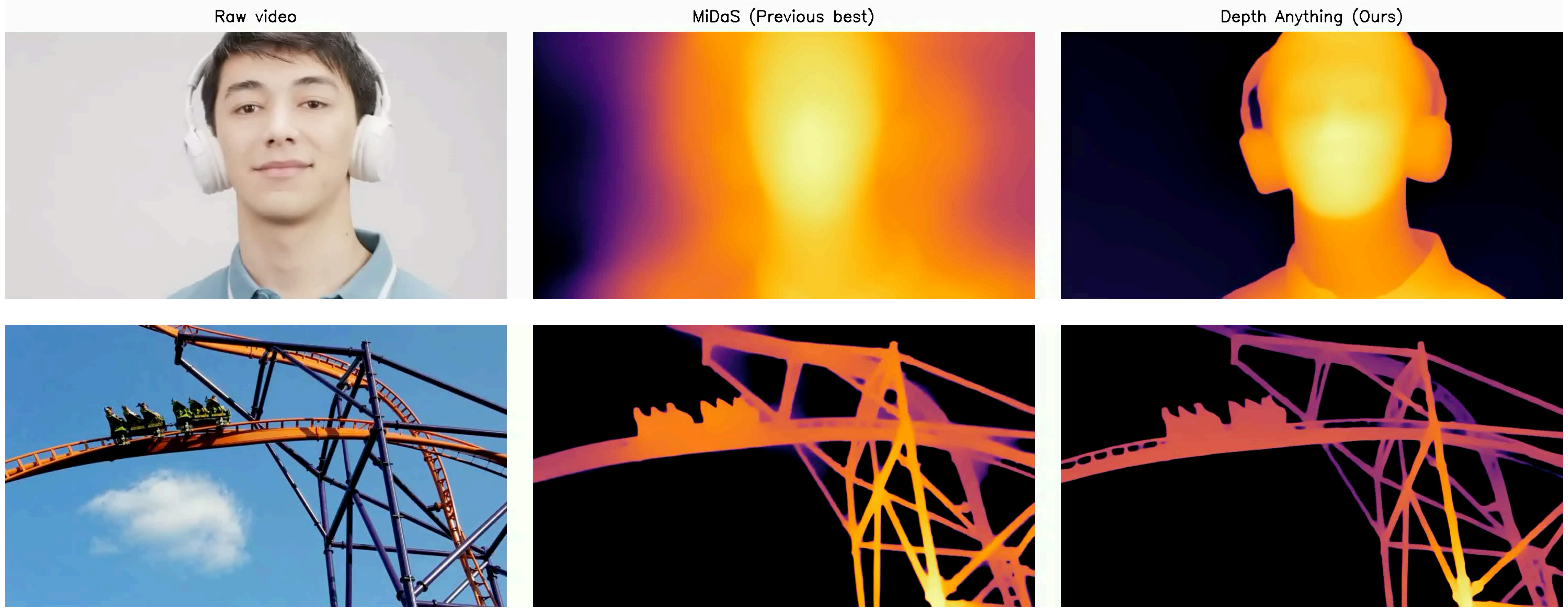
DPT, arXiv 2020

Using Transformers instead of convolutional predictors

Depth Prediction: An Active Research Area

Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data

Lihe Yang[1] Bingyi Kang[2†] Zilong Huang[2] Xiaogang Xu[3,4] Jiashi Feng[2] Hengshuang Zhao[1†]
[1]The University of Hong Kong [2]TikTok [3]Zhejiang Lab [4]Zhejiang University
† corresponding authors
https://depth-anything.github.io



trained on 1.5M labeled images and **62M+ unlabeled images** jointly

CVPR 2024

Depth Prediction: An Active Research Area

Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation

Bingxin Ke Anton Obukhov Shengyu Huang Nando Metzger
Rodrigo Caye Daudt Konrad Schindler
Photogrammetry and Remote Sensing, ETH Zürich

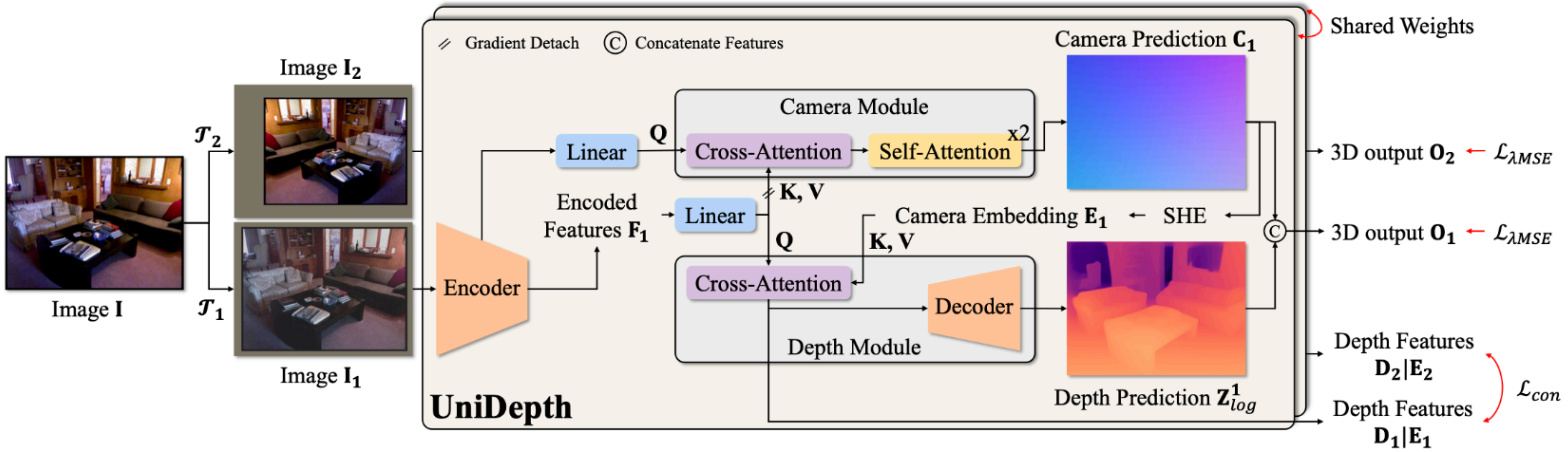
Marigold, CVPR 2024

Adapt SOTA diffusion models for depth prediction

Depth Prediction: An Active Research Area

UniDepth: Universal Monocular Metric Depth Estimation

KITTI Benchmark | 1st (at submission time) | custom badge | inaccessible | custom badge | inaccessible



Just directly predict metric depth with some consistency loss

[UniDepth: Universal Monocular Metric Depth Estimation](),
Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, Fisher Yu,
CVPR 2024,
Paper at [arXiv 2403.18913]()

Depth Prediction: An Active Research Area

MoGe: Unlocking Accurate Monocular [redacted]eometry Estimation for Open-Domain [redacted]ages with Optimal Training Supervision

[redacted]g Wang[1,2], Sicheng Xu[2], Cassie Dai[3,2], Jianfeng Xiang[4,2], Yu Deng[2], Xin Tong[2], Jiaolong Yang[2]
[1]USTC, [2]Microsoft Research, [3]Harvard, [4]Tsinghua University

CVPR 2025 Oral

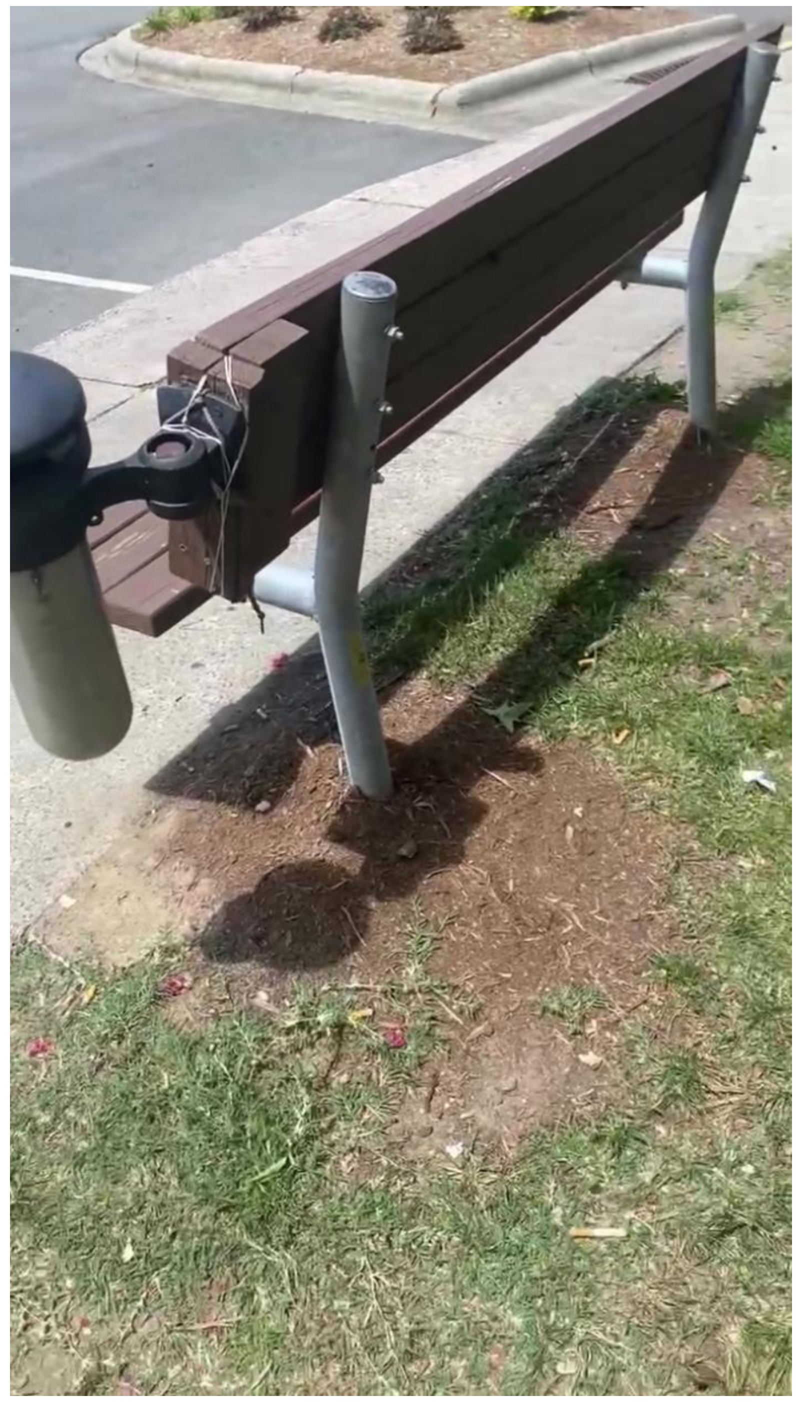
Paper | arXiv | Code | HF Demo

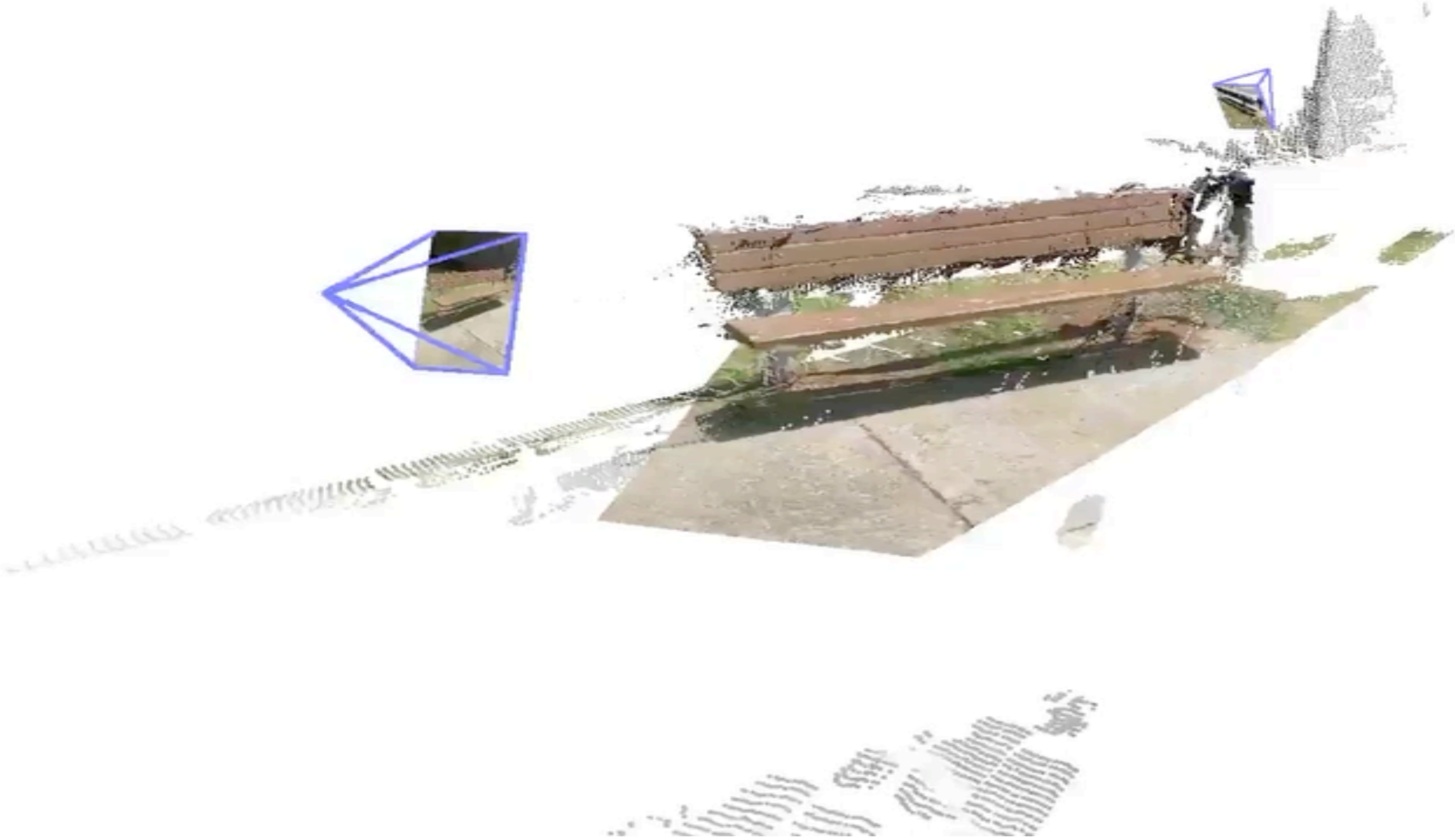
Predict per-pixel xyz points in a canonical coordinate frame instead of depth

But.. mono depth is 2.5D!
What about actual 3D?

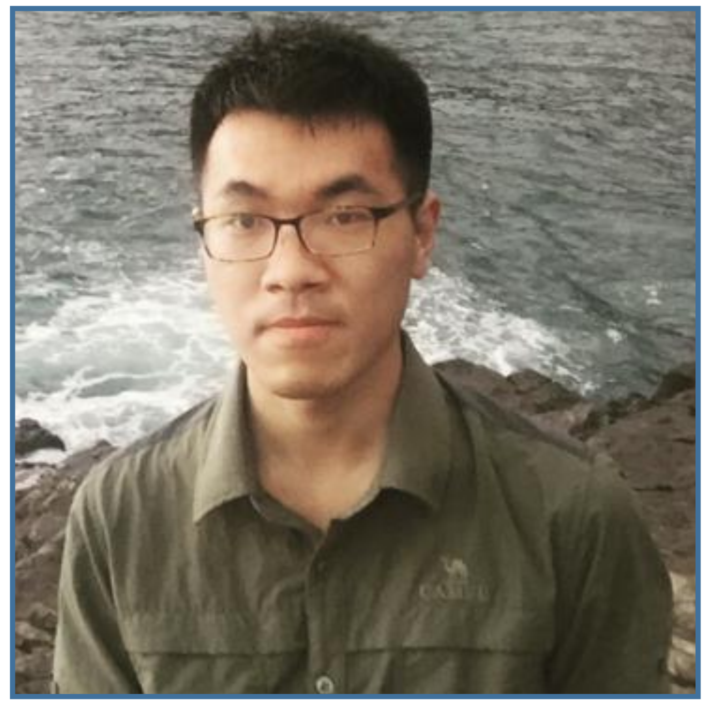
DUST3R

DUSt3R [Wang et al CVPR 2024]





DUSt3R: Dense Unconstrained Stereo 3D Reconstruction





Shuzhe Wang
Aalto University

Vincent Leroy
Naverlabs Europe

Yohann Cabon
Naverlabs Europe

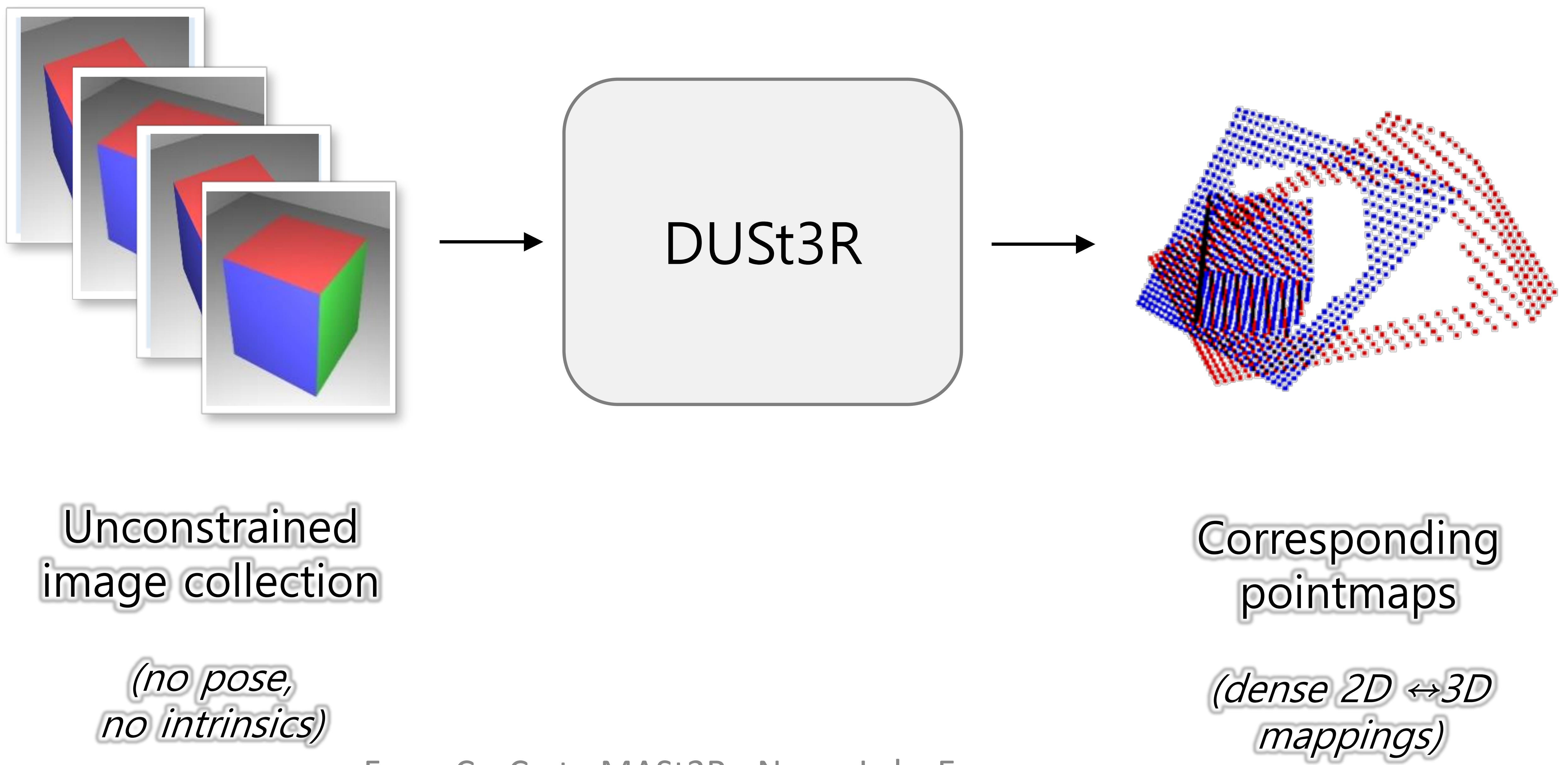
Boris Chidlovskii
Naverlabs Europe

Jérome Revaud
Naverlabs Europe

Next slides from this talk!

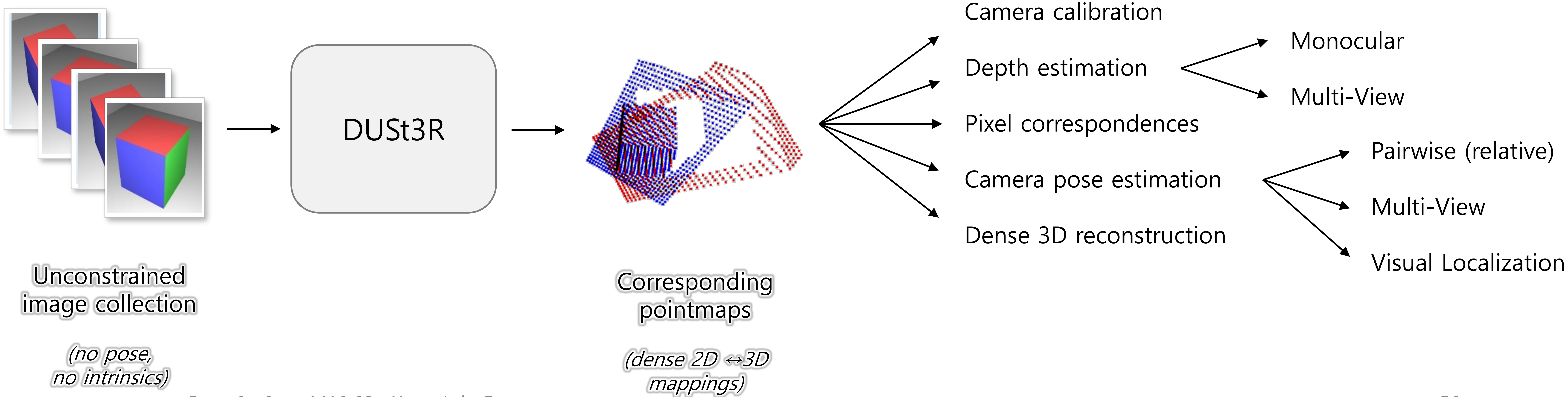
DUSt3R:
Dense Unconstrained Stereo 3D Reconstruction

- Pointmaps as a proxy output that:
    - *capture 3D scene geometry (point-cloud)*
    - *connect pixels ↔ 3D points*
    - *spatially relate 2 viewpoints (relative pose)*

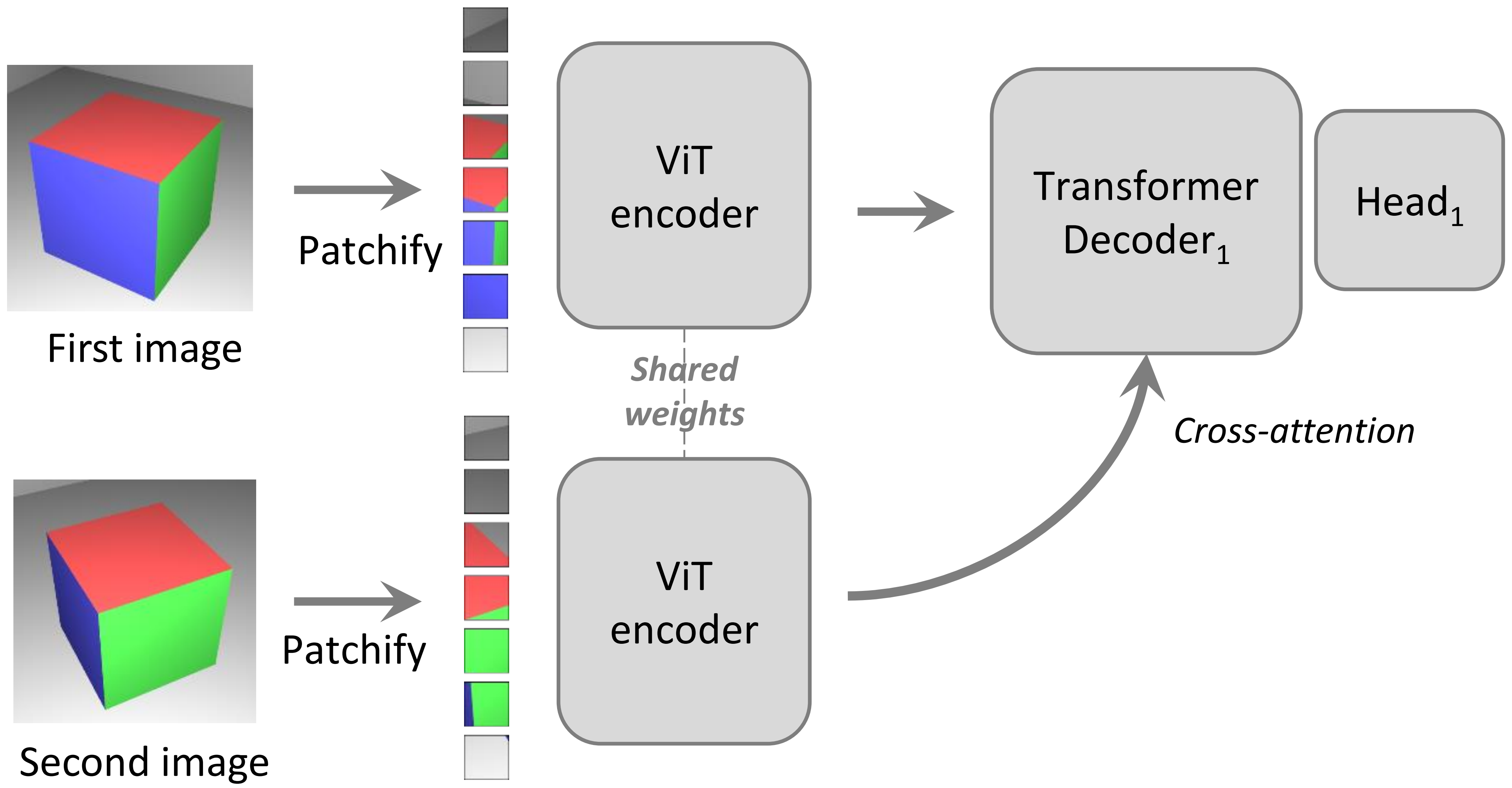


DUSt3R: Dense Unconstrained Stereo 3D Reconstruction

- Pointmaps as a proxy output that:
  - *capture 3D scene geometry (point-cloud)*
  - *connect pixels ↔ 3D points*
  - *spatially relate 2 viewpoints (relative pose)*

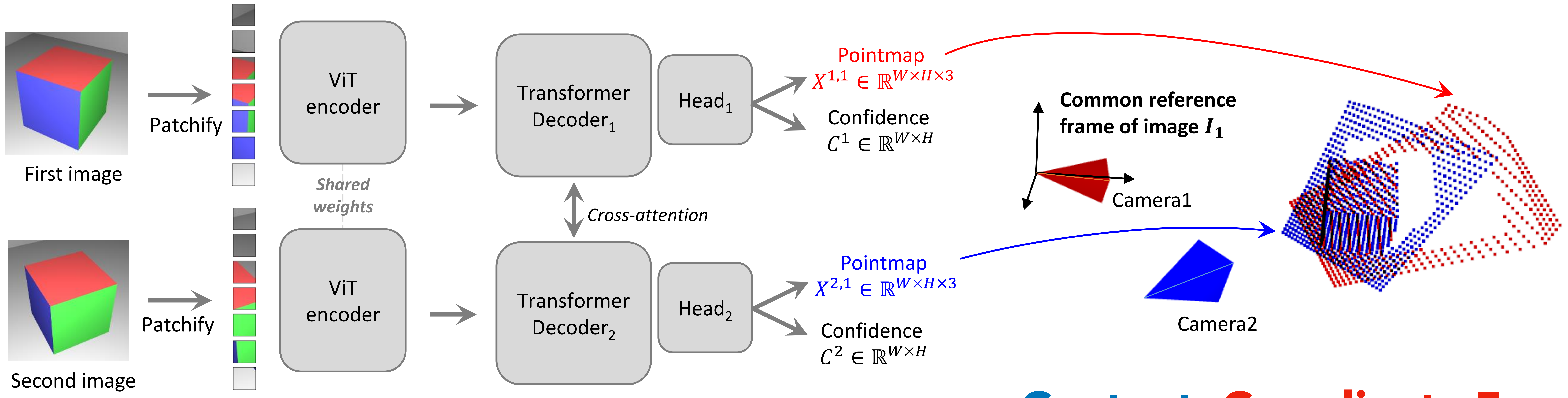


DUSt3R:
Dense Unconstrained Stereo 3D Reconstruction



Start from CroCo …

DUSt3R:
Dense Unconstrained Stereo 3D Reconstruction

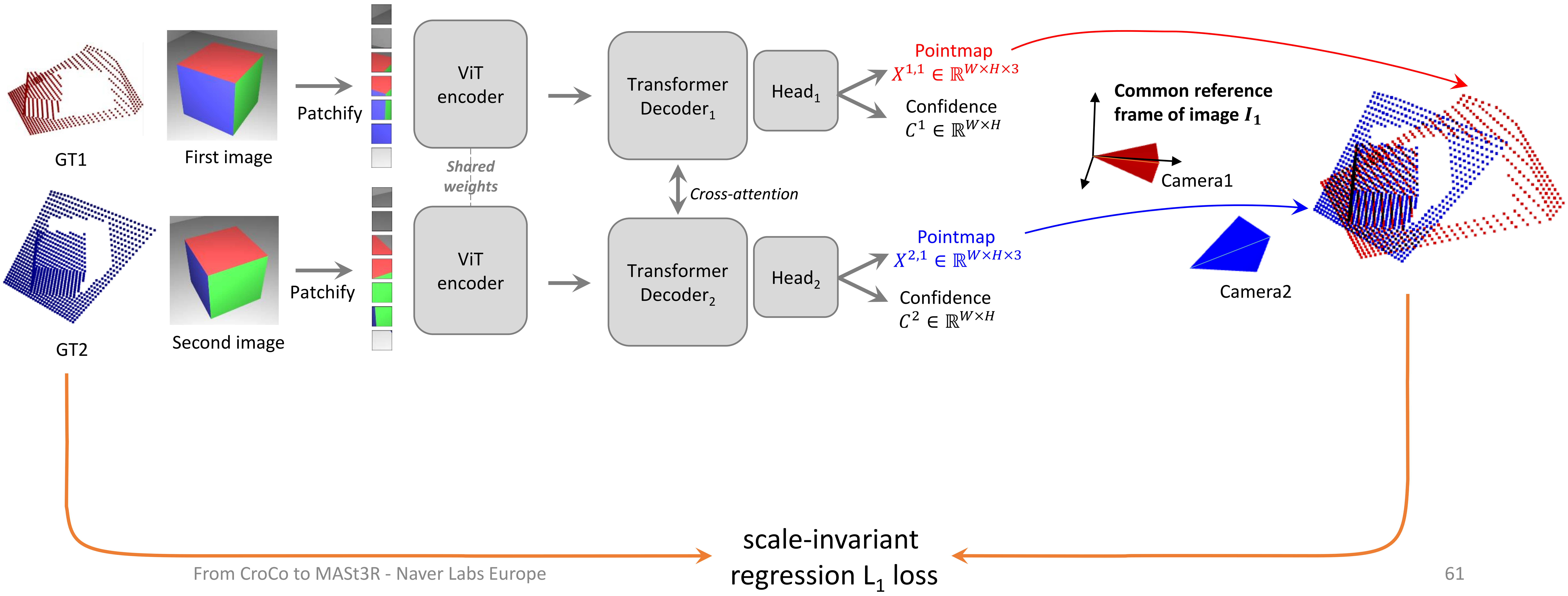


Content **Coordinate Frame**

$$X^{\boxed{n},\boxed{m}}$$

Start from CroCo and add a 2nd decoder

DUSt3R:
Dense Unconstrained Stereo 3D Reconstruction



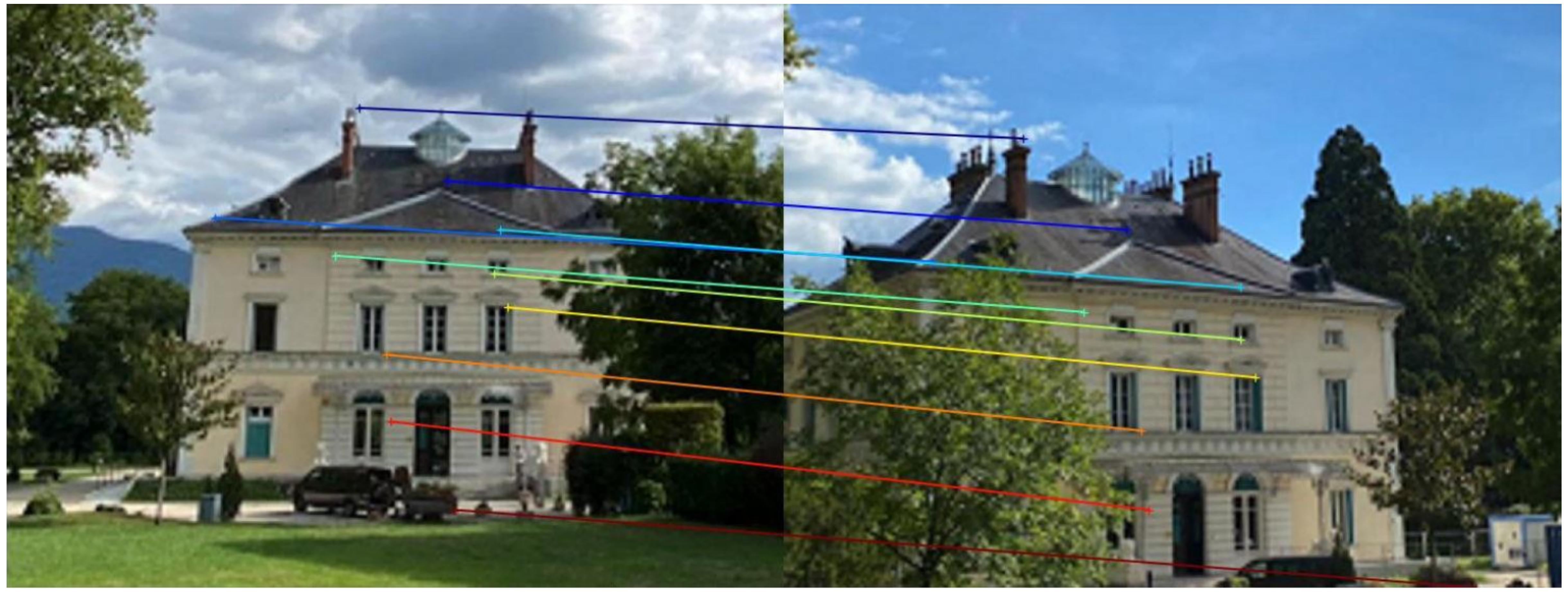
DUSt3R: Dense Unconstrained Stereo 3D Reconstruction

Train it on lots of data!!

- Training data

Datasets	Type	N Pairs
Habitat [103]	Indoor / Synthetic	1000k
CO3Dv2 [93]	Object-centric	941k
ScanNet++ [165]	Indoor / Real	224k
ArkitScenes [25]	Indoor / Real	2040k
Static Thing 3D [68]	Object / Synthetic	337k
MegaDepth [55]	Outdoor / Real	1761k
BlendedMVS [161]	Outdoor / Synthetic	1062k
Waymo [121]	Outdoor / Real	1100k

Many things you can do with Dust3r



- Point matching: NN in 3D space
- Recovering focal length
  - Assume principal point is at the center
  - Solve for $(u,v) - f\dfrac{(X,Y)}{z}$ across all pixels weighted by confidence:

$$f_1^* = \arg\min_{f_1} \sum_{i=0}^{W} \sum_{j=0}^{H} C_{i,j}^{1,1} \left\| (i', j') - f_1 \frac{(X_{i,j,0}^{1,1}, X_{i,j,1}^{1,1})}{X_{i,j,2}^{1,1}} \right\|,$$

Many things you can do with Dust3r

- Relative Pose Estimation (between img 1 and 2)
  - Option 1: Use the focal length & 2D correspondence to get Essential matrix
  - Option 2: Solve Procrustes alignment between $X^{1,1}$ and $X^{1,2}$ by running the network twice by flipping the inputs
  - Option 3: PnP with RANSAC

Dust3r for multiple views

Global Alignment Optimization

- Run DUST3R on all pairs, then solve for world point maps with cameras

$$\left\| X^{B,w} - \sigma^{(B,A)} P^{(B,A)\mapsto w} X^{B,A} \right\|$$

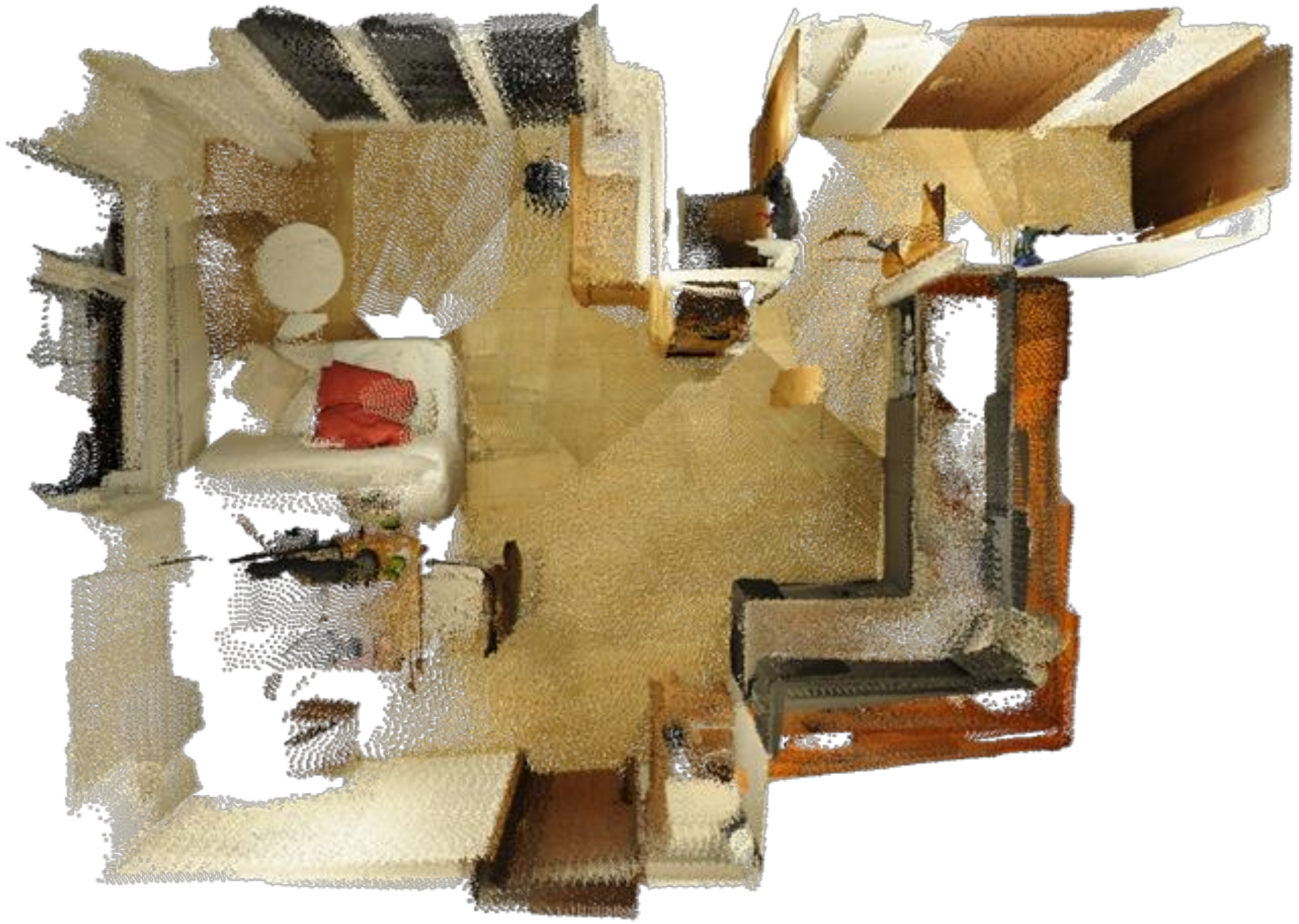
$X^{B,w}$: Content of View B in World frame

$\sigma^{(B,A)}$: Scale for the pair

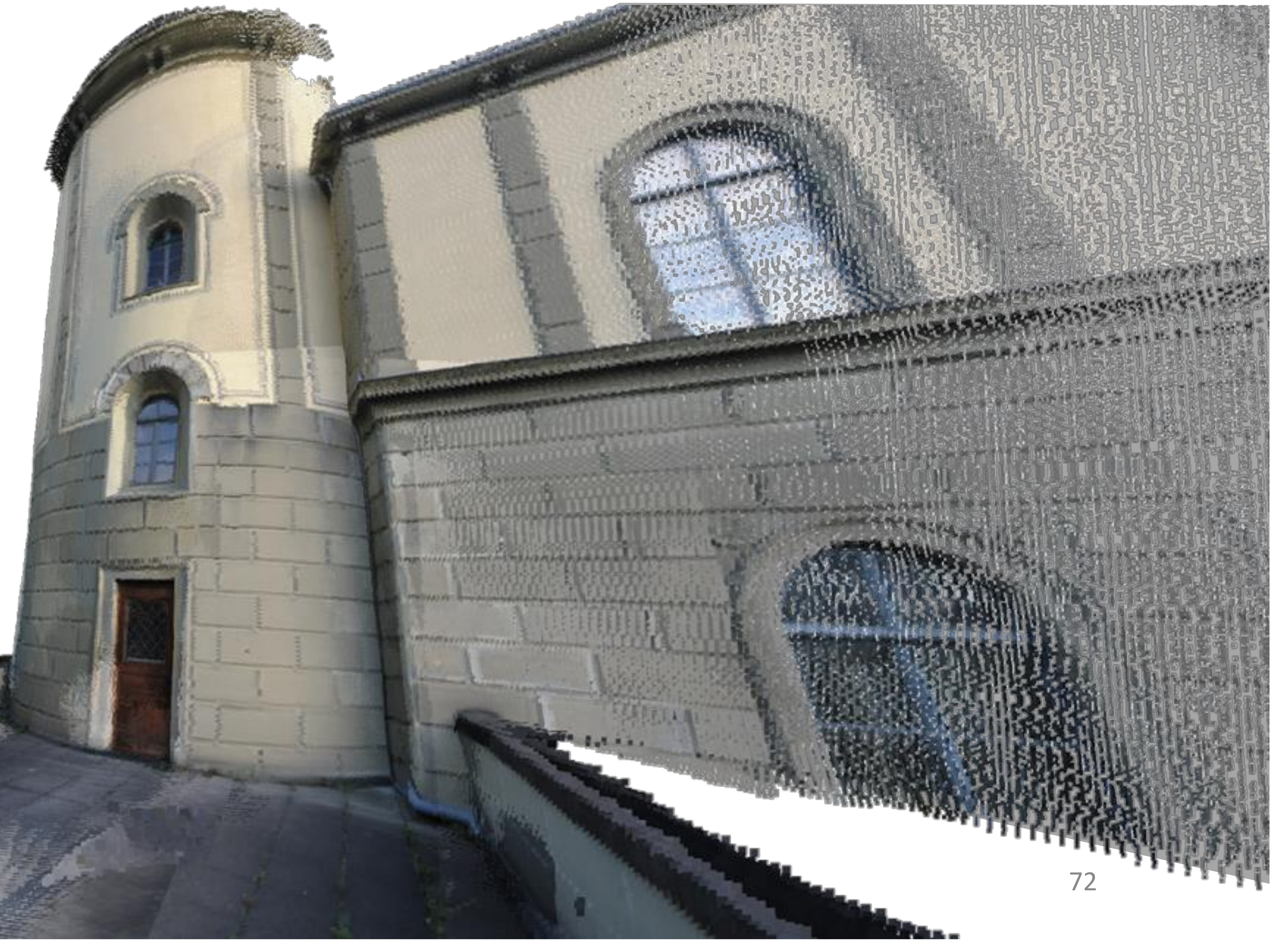
$P^{(B,A)\mapsto w}$: Transformation for that pair, i.e from this A to world

$X^{B,A}$: Content of View B in A's coordinate system

The same model works indoor …



… and outdoor



Opposite View reconstructions

Img 1

Img 2