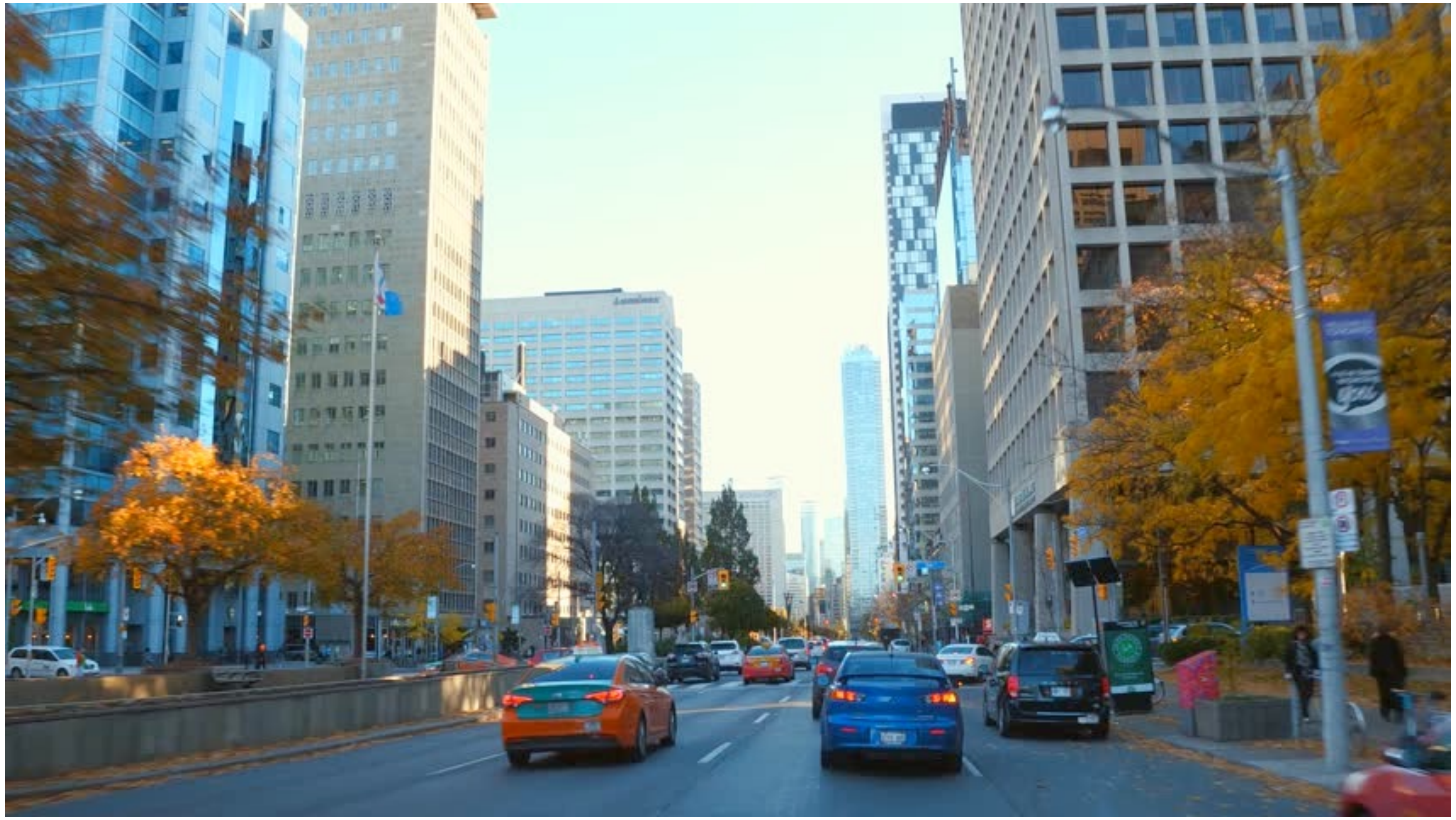
Learning 3D reconstruction in underconstrained settings

2.5D vs 3D

- 2.5D: Reconstruct only the visible pixels
- 3D: Reconstruct full 3D shapes

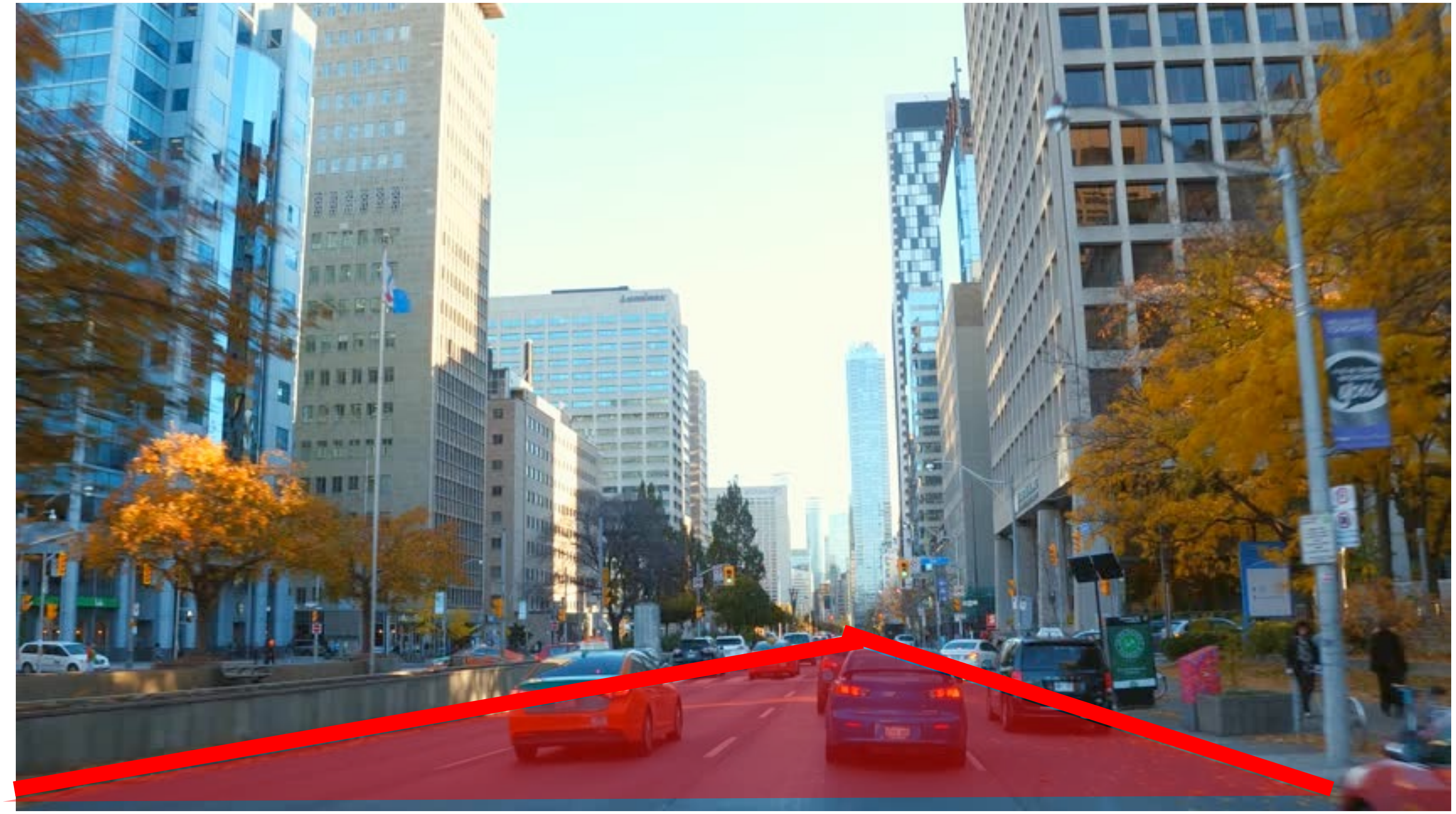
Estimating depth from a single image

- Why is this even possible?



Estimating depth from a single image

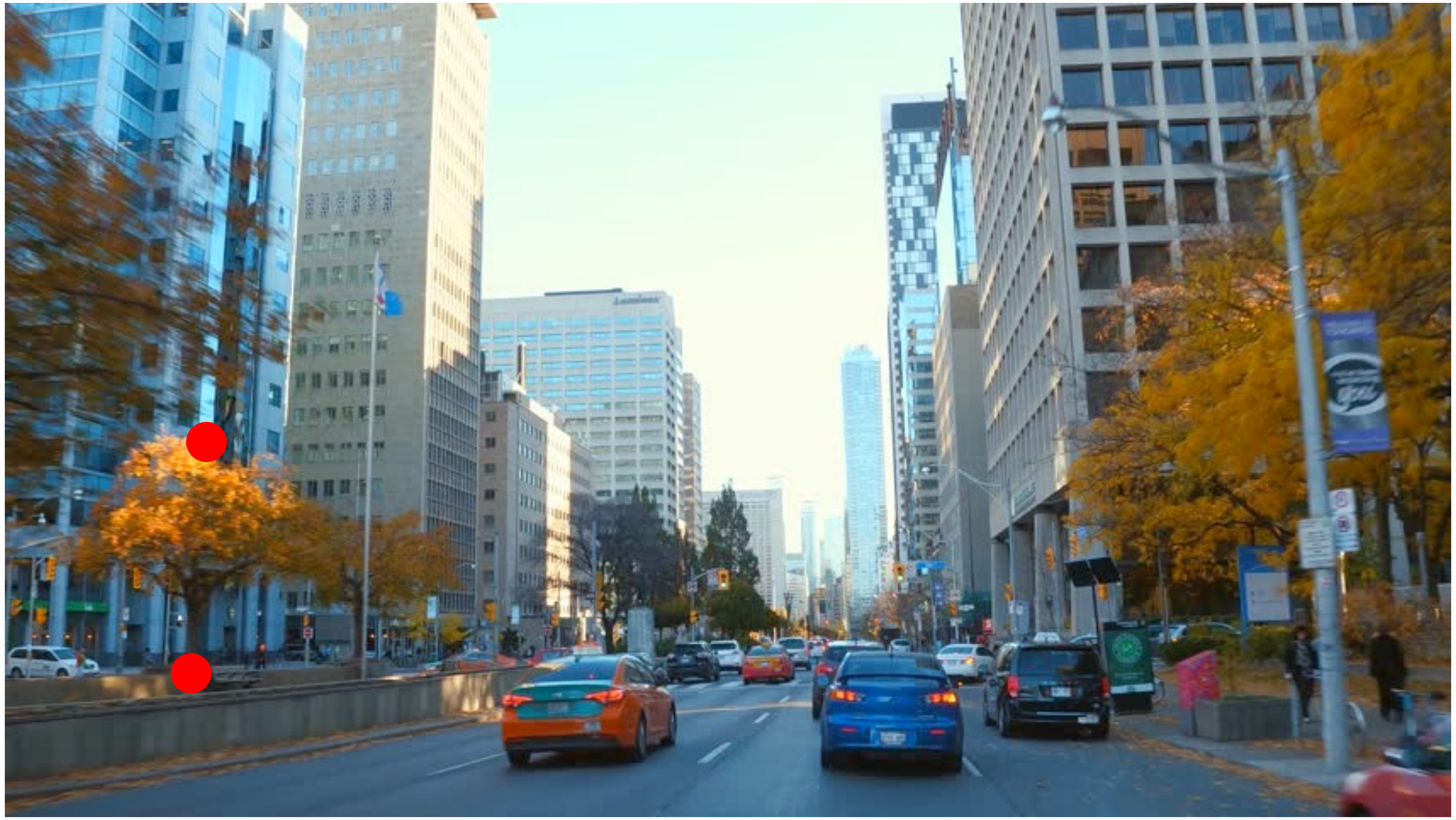
- Why is this even possible?



Vanishing lines indicate plane orientations

Estimating depth from a single image

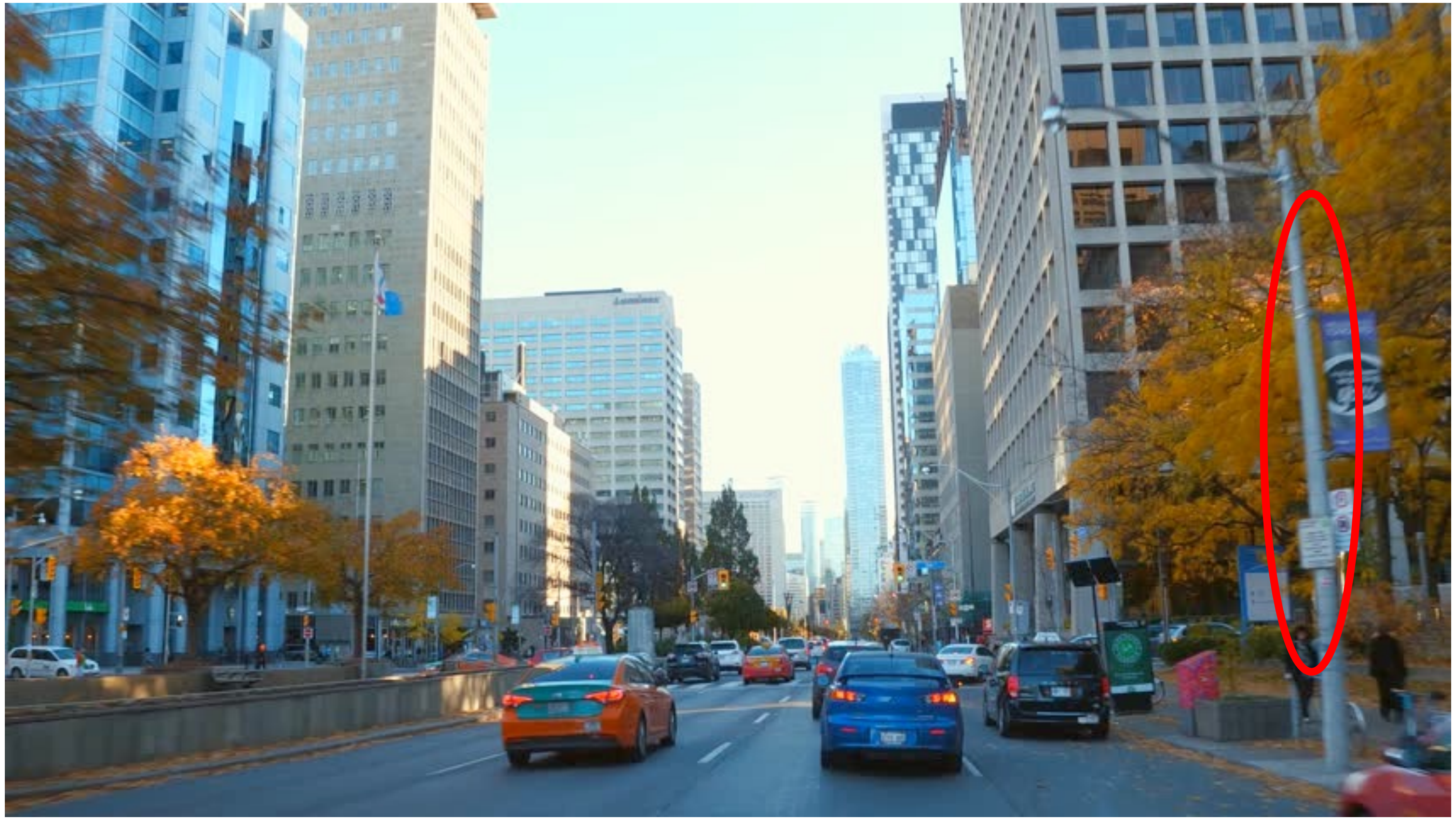
- Why is this even possible?



Apparent object height relative to true height indicates depth

Estimating depth from a single image

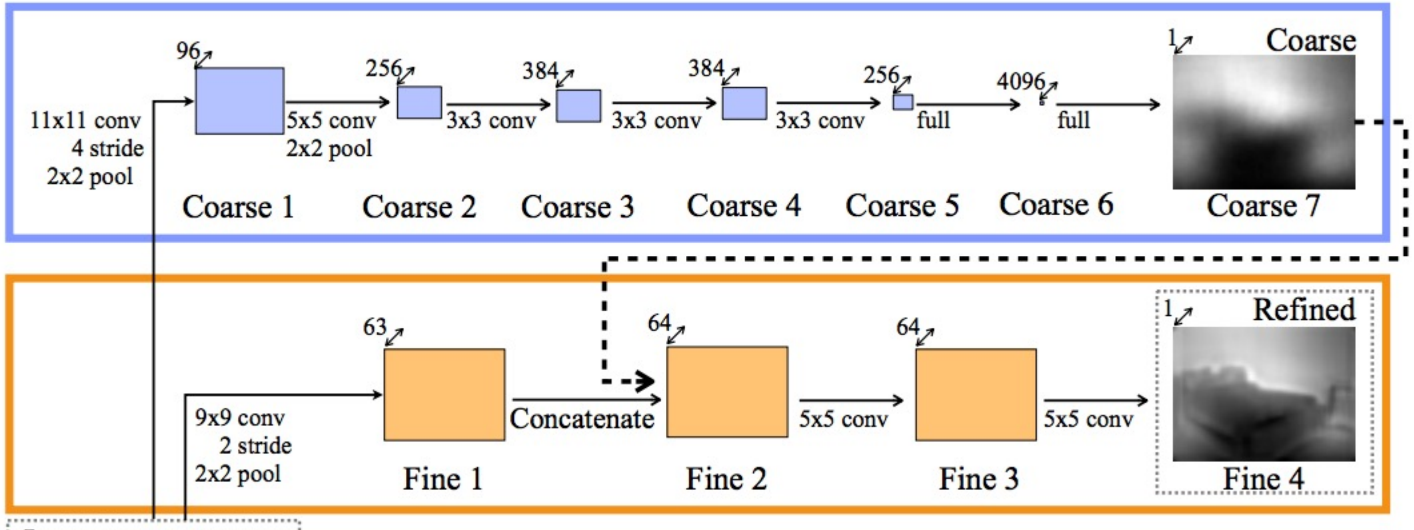
- Why is this even possible?



Occlusion indicates depth ordering

Estimating depth from a single image

- Image-in, image-out
- Similar to segmentation
- Again, resolution issues



		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

Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. David Eigen, Christian Puhrsch, Rob Fergus. In *NIPS,* 2014

Metric depth is a bad target

Metric depth is a bad target

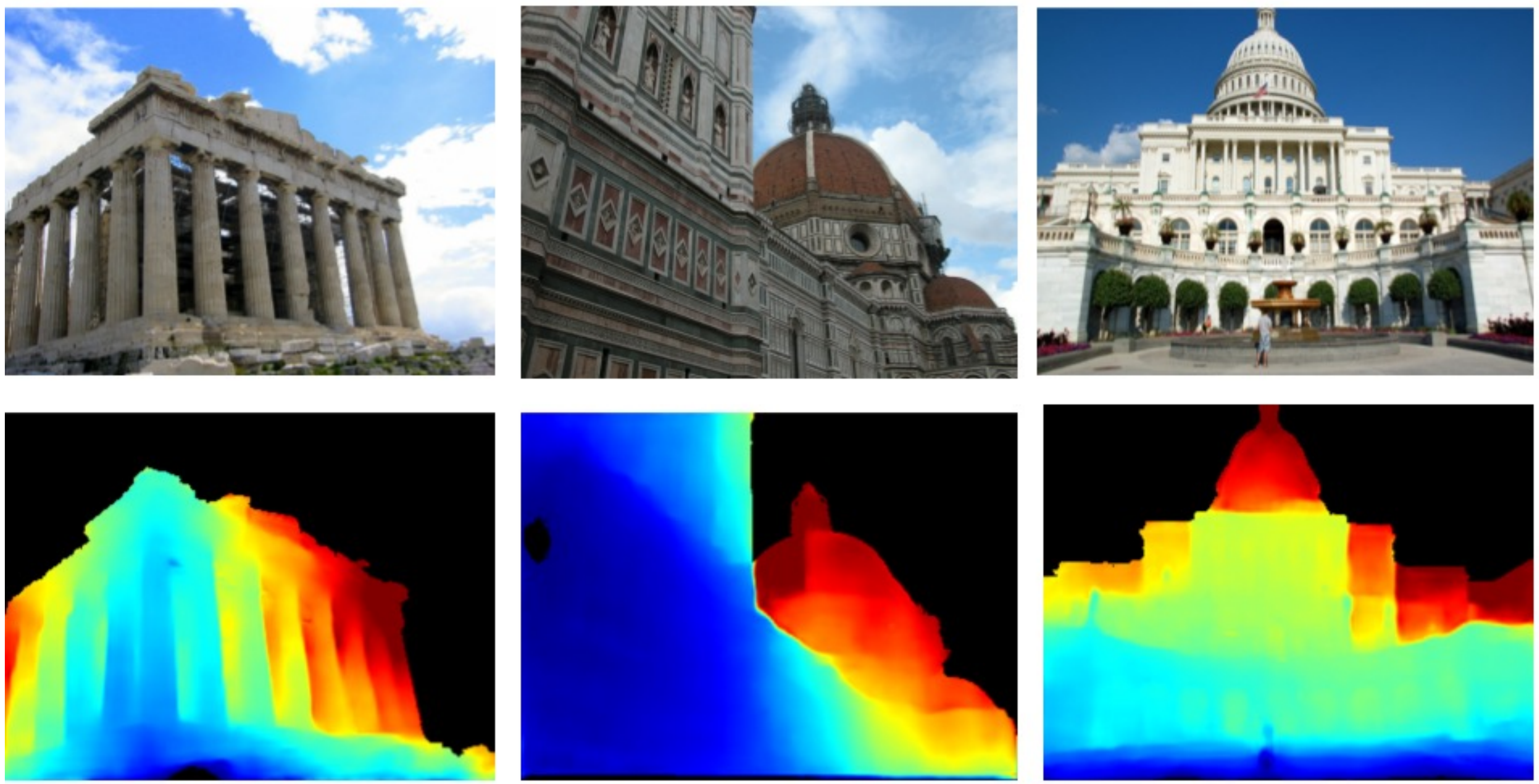
- Only relative depths matter
- Only logarithmic scales matter

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

Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. David Eigen, Christian Puhrsch, Rob Fergus. In *NIPS,* 2014

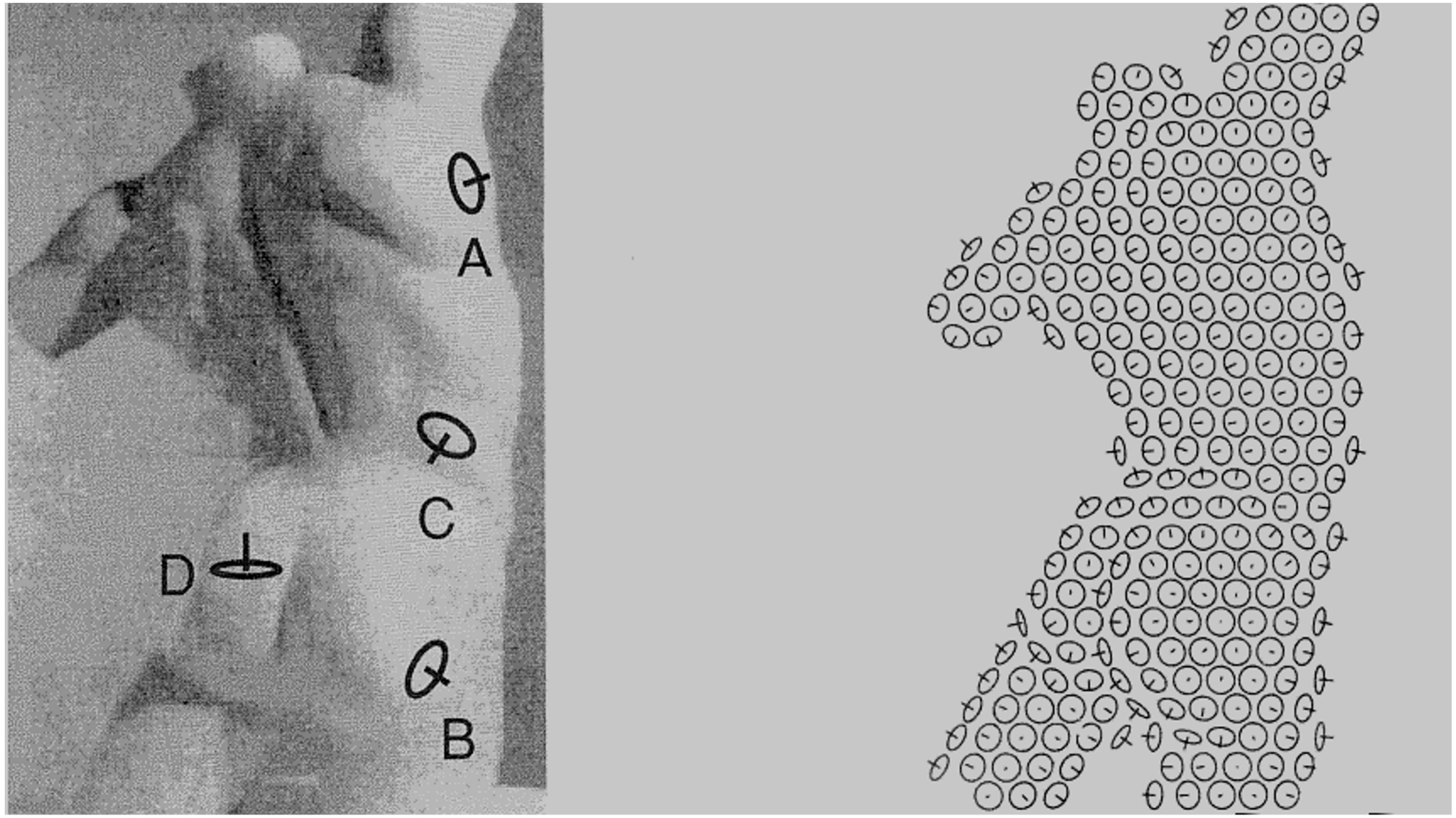
Depth estimation today

- MegaDepth, learnt from large SfM models



Parthenon, Athens | Florence Cathedral, Florence | United States Capitol, D.C.

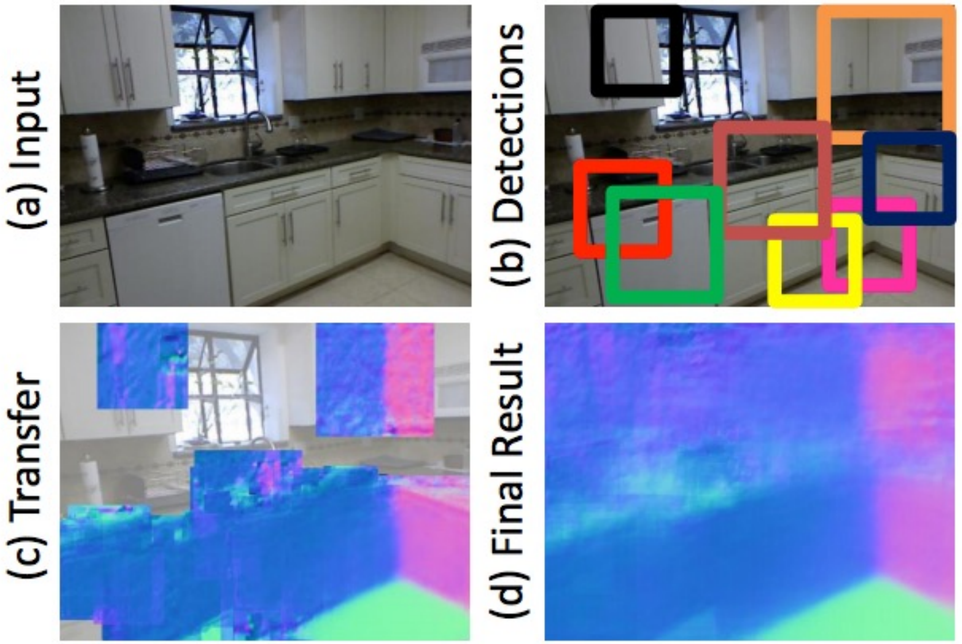
Humans perceive surface normals, not just depth, through a combination of various pictorial cues



Surface perception in pictures. Koenderink, van Doorn and Kappers, 1992

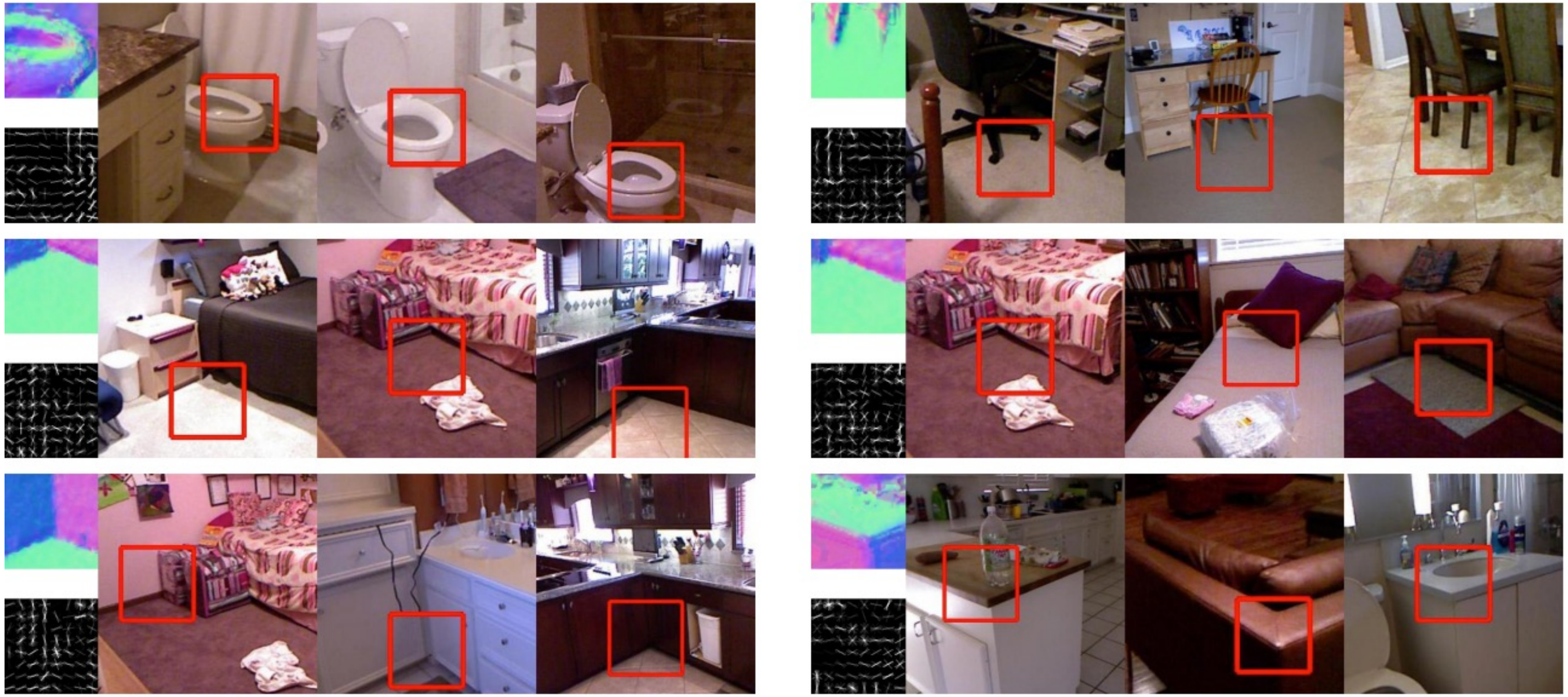
Slide credit: Jitendra Malik

Estimating normals from a single image

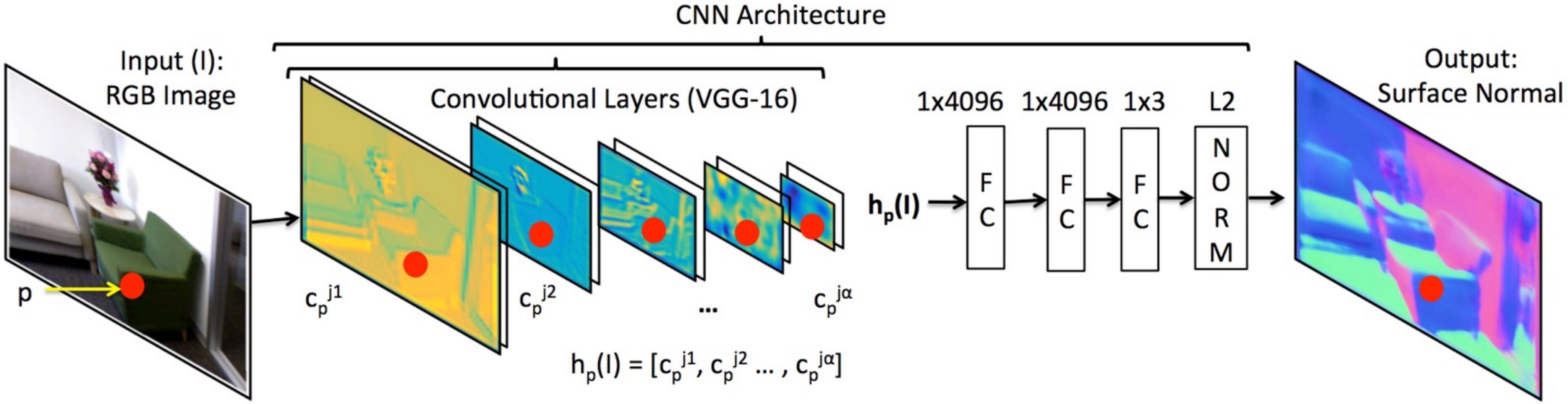


Data-Driven 3D Primitives for Single Image Understanding. David F. Fouhey, Abhinav Gupta, Martial Hebert. In *ICCV* 2013.

Estimating normals from a single image

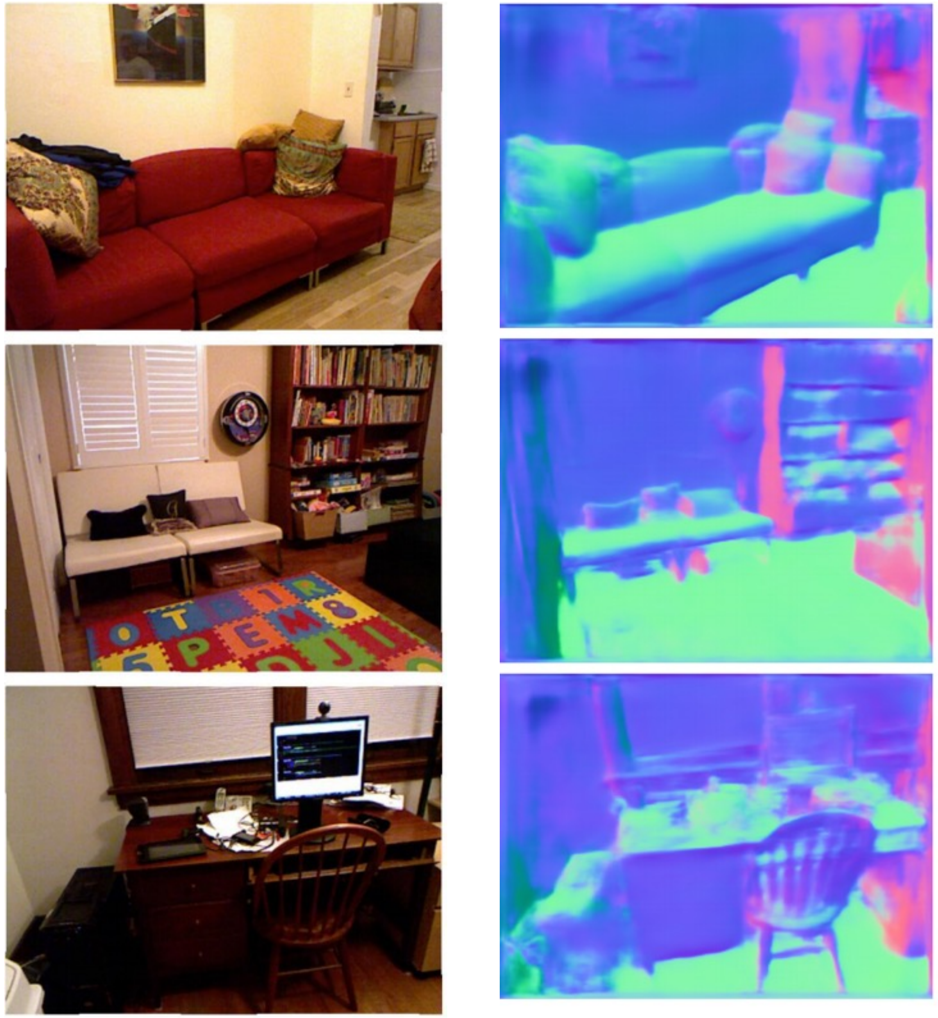


Estimating normals from a single image



Marr Revisited: 2D-3D Alignment via Surface Normal Prediction. Aayush Bansal, Bryan Russell, Abhinav Gupta. In *CVPR,* 2016

Estimating normals from a single image



2.5D vs 3D prediction

- Predicting depth / surface normals for every pixel is not full reconstruction
  - “2.5D reconstruction”
  - Does not contain parts of the scene that are hidden from view
- Can we do full 3D reconstruction?
- Simpler situation: can we do full 3D reconstruction of isolated objects?

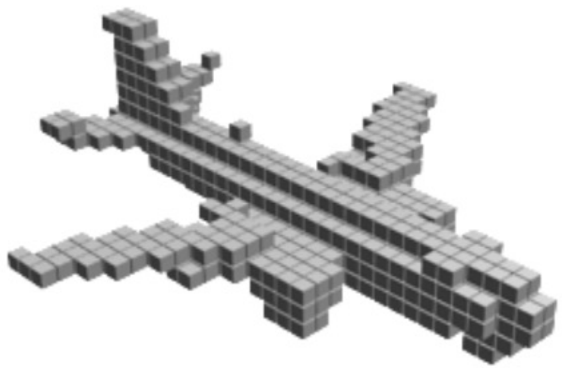
Shapenet

Reconstructing 3D shapes from images using machine learning

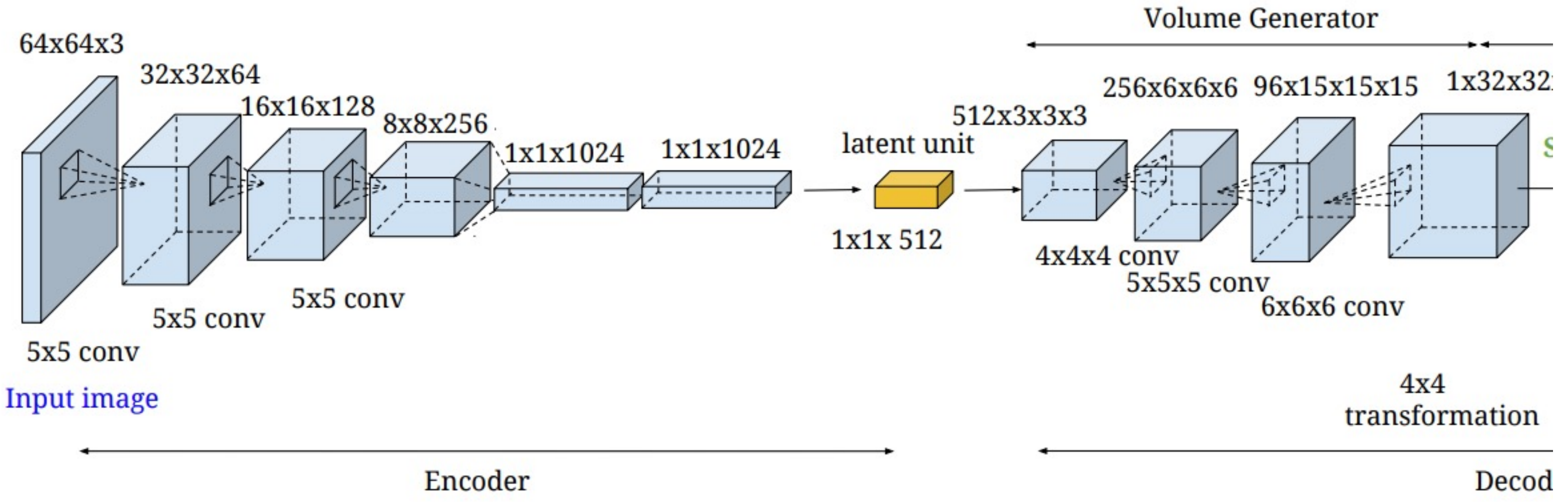
- Input:
  - Single image or multiple images of the same object
- Output:
  - 3D shape
- Representation?

Representation of 3D shapes

- Voxel grids
  - Discretize volume into grid cells
  - Identify cells that are occupied by object
- Advantages:
  - Easy representation for ML: analog of pixels
- Disadvantages:
  - Memory-inefficient
  - Difficult to capture surface



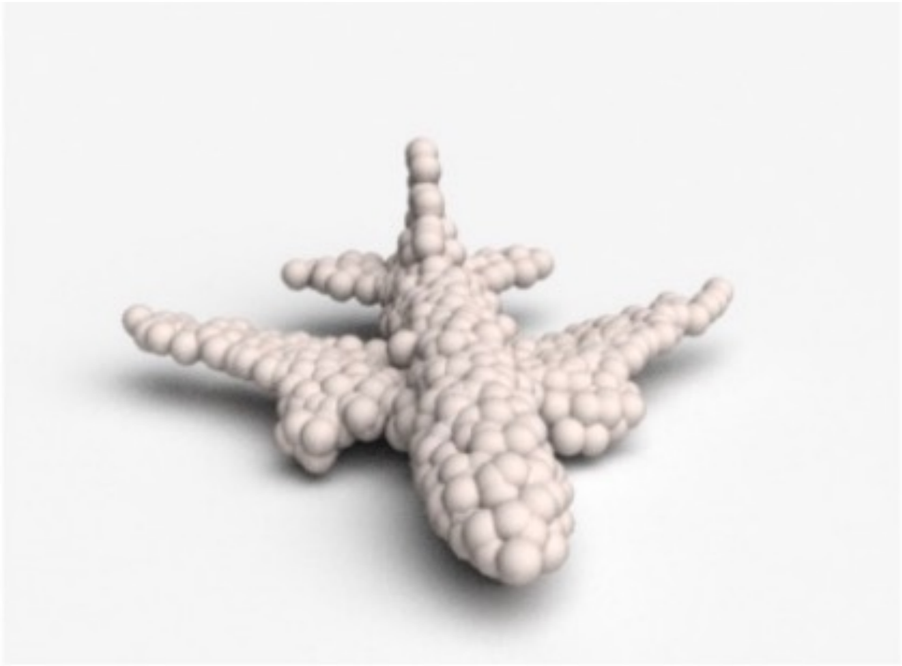
Architectures for generating voxel grids



1. Choy, Christopher B., et al. "3d-r2n2: A unified approach for single and multi-view 3d object reconstruction." *European conference on computer vision*. Springer, Cham, 2016.
2. Yan, Xinchen, et al. "Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision." *Advances in Neural Information Processing Systems*. 2016.

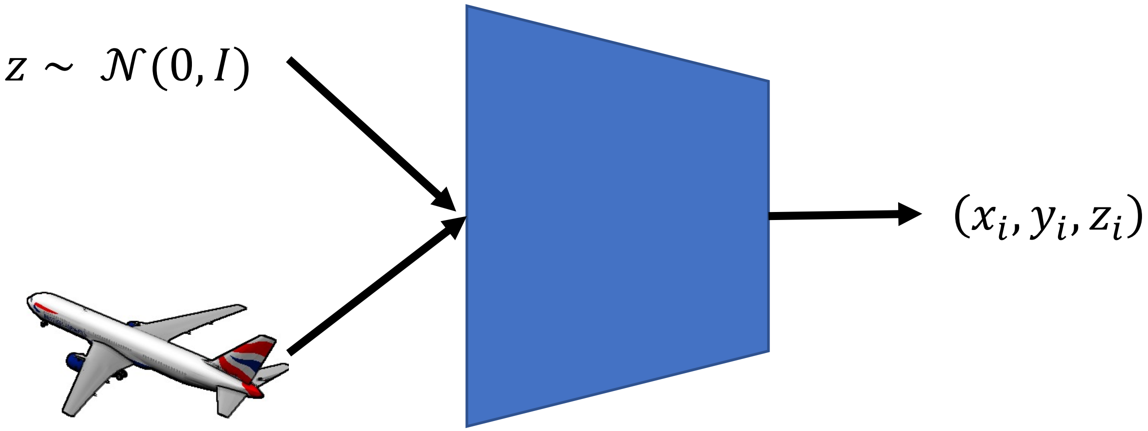
Representation of 3D shapes

- Point clouds
- Each point lies on surface
- Advantages:
  - Common representation produced by sensors (e.g. LiDAR)
  - Sparse, so memory efficient
- Disadvantages:
  - Difficult output to predict: sets
  - Difficult to extract surface



Architecture for generating point clouds

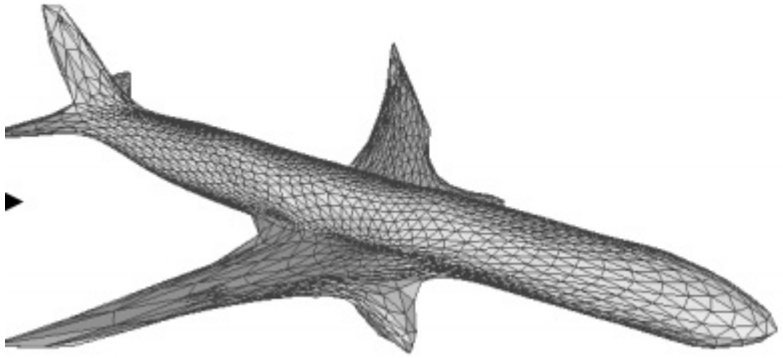
- Not an established answer
- One possibility: cloud of points = *samples* from an underlying distribution
- Generative modeling



Fan, Haoqiang, Hao Su, and Leonidas J. Guibas. "A point set generation network for 3d object reconstruction from a single image." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

Representation of 3D shapes

- Meshes
- Advantages
  - Common in graphics
  - Surfaces are triangles in the mesh
  - Sparse representation: memory efficient
  - Can easily encode color, texture, surface normals
- Disadvantages
  - Extremely difficult to predict: graph



Architecture for producing meshes

- Assume connectivity and faces are the same as that of a sphere
- Move only vertices
- Cannot change *topology* of objects

Wang, Nanyang, et al. "Pixel2mesh: Generating 3d mesh models from single rgb images." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

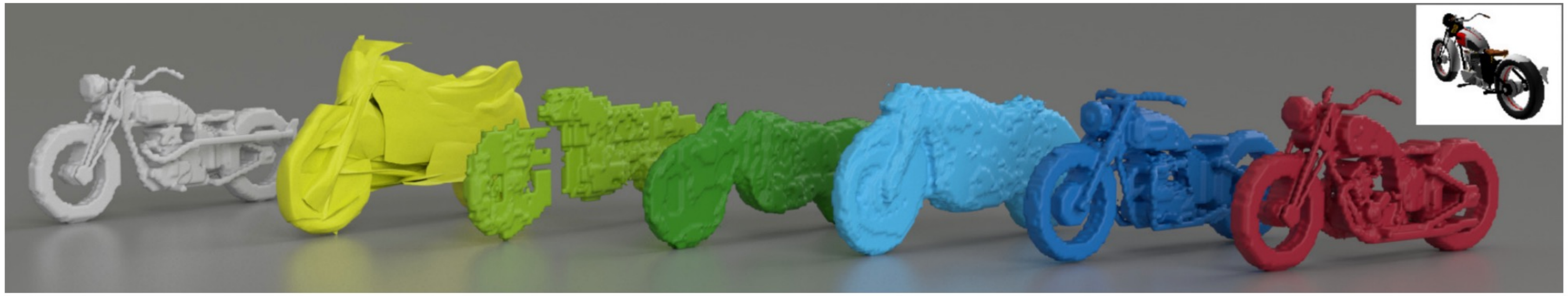
Where do we get ground truth?

	SHAPENET	IMAGENET
# Categories	55	**1000**
# Instances / class	50 – 8000	1300

- Models created by 3D artists
- Laser scans
- Structure-from-motion

Challenges with single view 3D reconstruction

- Clear evidence that SVR networks are mostly doing retrieval



AtlasNet (light green, 0.38) OGN (green, 0.46) Matryoshka Networks (dark green, 0.47) Clustering (light blue, 0.46) Retrieval (dark blue, 0.57))

GT Predicted

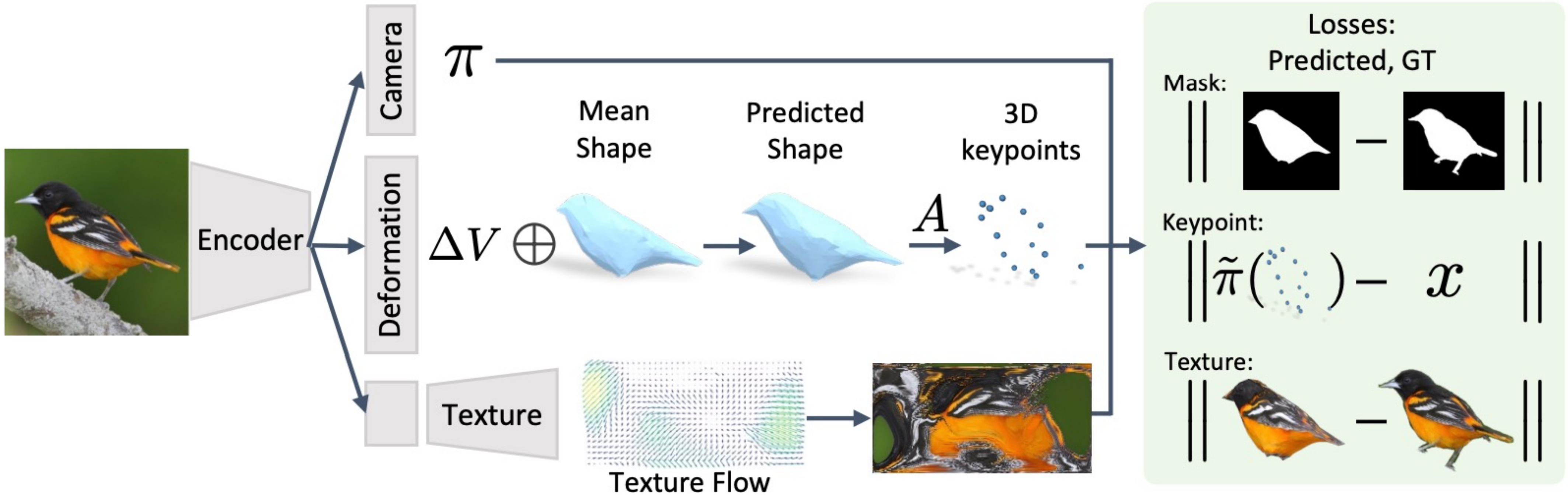
Tatarchenko, Maxim, et al. "What do single-view 3d reconstruction networks learn?." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

Supervision?

- Fully supervised [1]
- Supervised with multiple views from *known* cameras [2]
  - Predict shape from one image
  - Project it to other views
  - Ensure *photometric consistency*
- Supervised with multiple views from *unknown* cameras [3]
  - Also jointly learn to predict camera pose

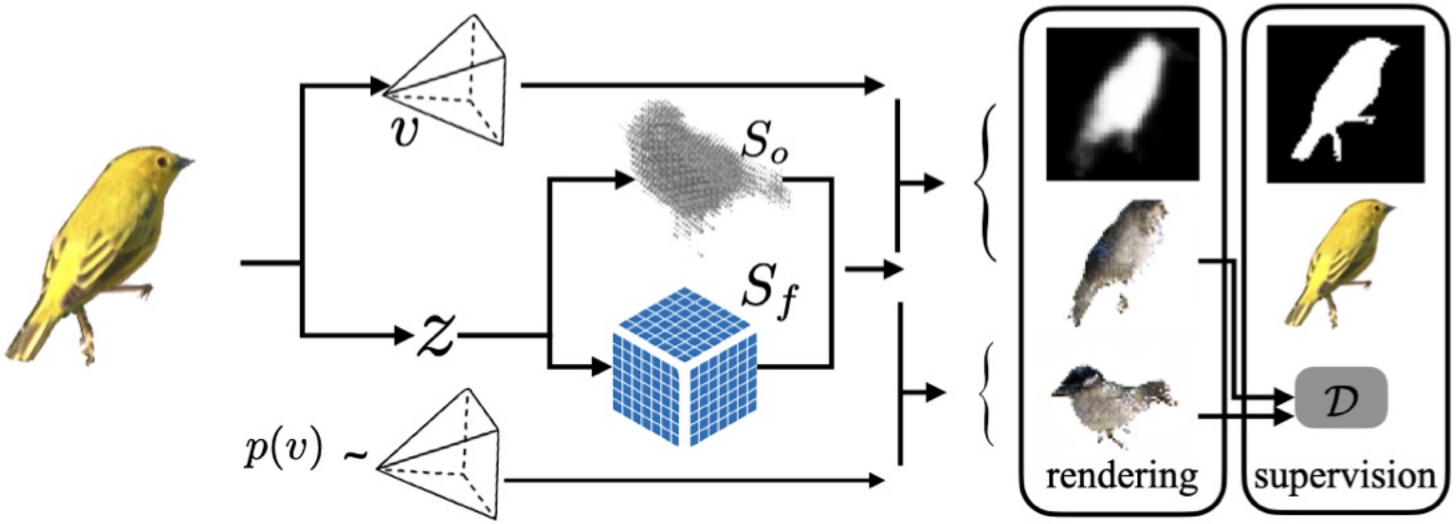
1. Choy, Christopher B., et al. "3d-r2n2: A unified approach for single and multi-view 3d object reconstruction." *European conference on computer vision*. Springer, Cham, 2016.
2. Yan, Xinchen, et al. "Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision." *Advances in Neural Information Processing Systems*. 2016.
3. Tulsiani, Shubham, et al. "Multi-view supervision for single-view reconstruction via differentiable ray consistency." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

3D reconstruction with limited ground truth



Kanazawa, Angjoo, et al. "Learning category-specific mesh reconstruction from image collections." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.

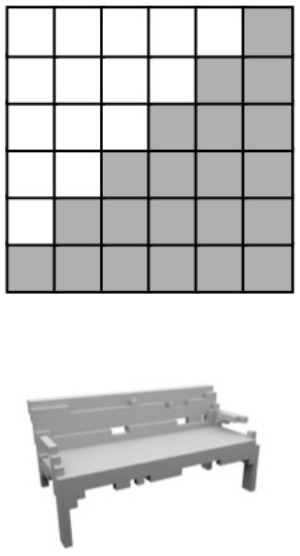
3D reconstruction with limited ground truth



Ye, Yufei, Shubham Tulsiani, and Abhinav Gupta. "Shelf-Supervised Mesh Prediction in the Wild." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

Neural representations of shape

Shape representations



(a) Voxel

- Easy to produce
- Very expensive to store
- Limited resolution

Implicit vs explicit equations

- Explicit representations of a curve
  - $y = f(x)$

- Implicit representation of a curve
  - $f(x, y) = 0$

Implicit representations of 3D shape

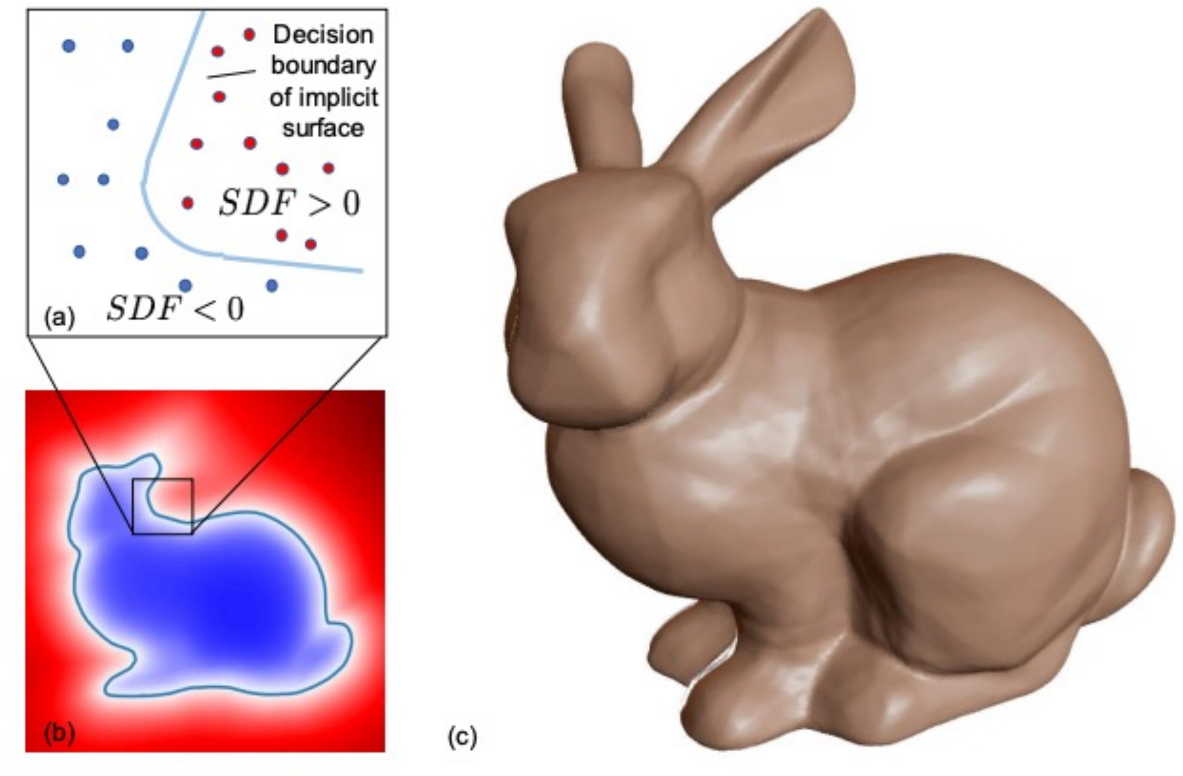
- Shape can be represented by the *level sets* of a function $f\colon \mathbb{R}^3 \to \mathbb{R}$

- Occupancy:
  - $f(x, y, z)$ is the probability (x, y, z) is inside the object
  - Surface is given by $f(x, y, z) = 0.5$

- Signed distance fields
  - $f(x, y, z)$ is the signed distance of (x, y, z) from the surface
  - Sign is positive for points inside, negative for points outside
  - Surface is given by $f(x, y, z) = 0$

Neural implicit representations

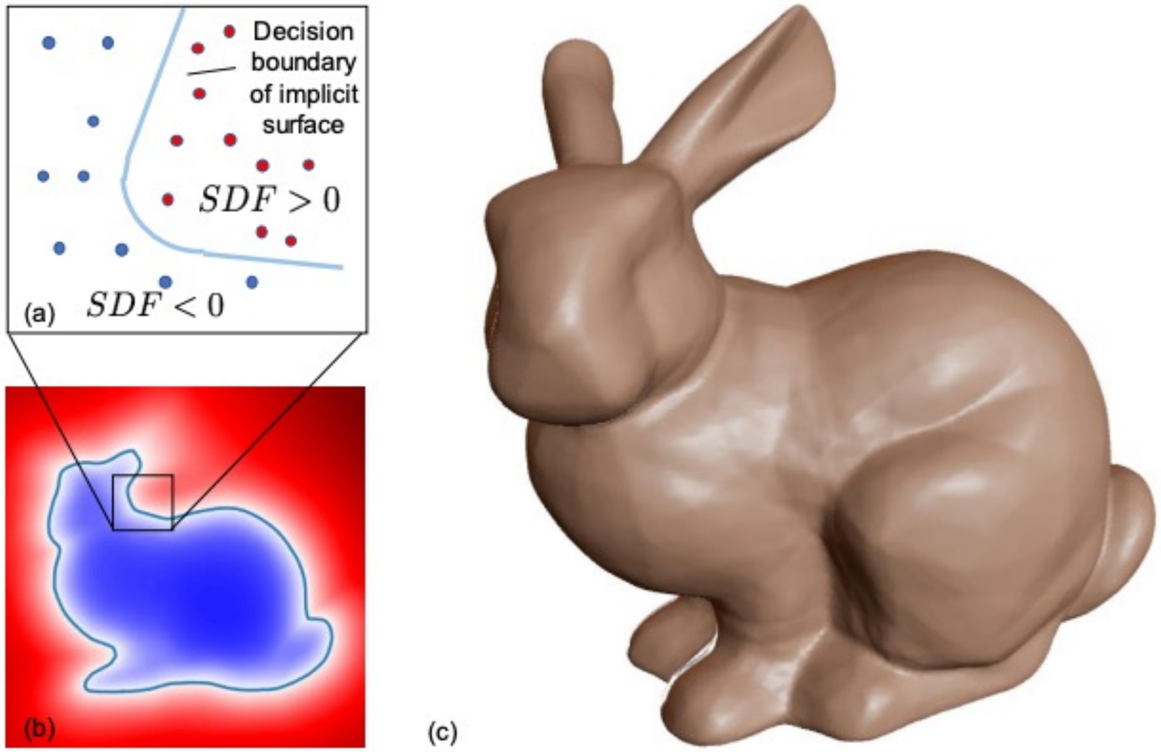
- Traditionally f is tabular array
- But can approximate with a neural network



Mescheder, Lars, et al. "Occupancy networks: Learning 3d reconstruction in function space." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

Park, Jeong Joon, et al. "Deepsdf: Learning continuous signed distance functions for shape representation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.

Neural Implicit shapes



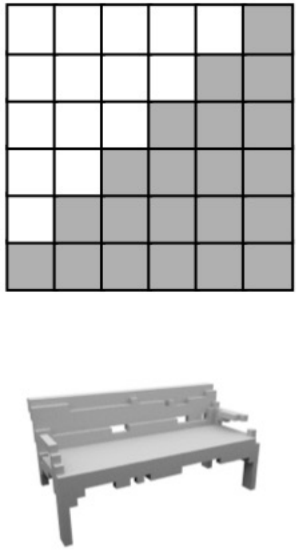
Park, Jeong Joon, et al. "Deepsdf: Learning continuous signed distance functions for shape representation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
Mescheder, Lars, et al. "Occupancy networks: Learning 3d reconstruction in function space." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

Representation of 3D shapes

- Implicit shapes
- A shape is a *function* that takes (x, y, z) as input and produces as output
  - Boolean on whether it is inside shape or not
  - Real value indicating distance from surface (”signed distance functions”)
- This *function* can be a *neural network*
- Thus each shape is a *neural network*
- Can additionally take e.g. feature vector as input

Park, Jeong Joon, et al. "Deepsdf: Learning continuous signed distance functions for shape representation." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.
Mescheder, Lars, et al. "Occupancy networks: Learning 3d reconstruction in function space." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2019.

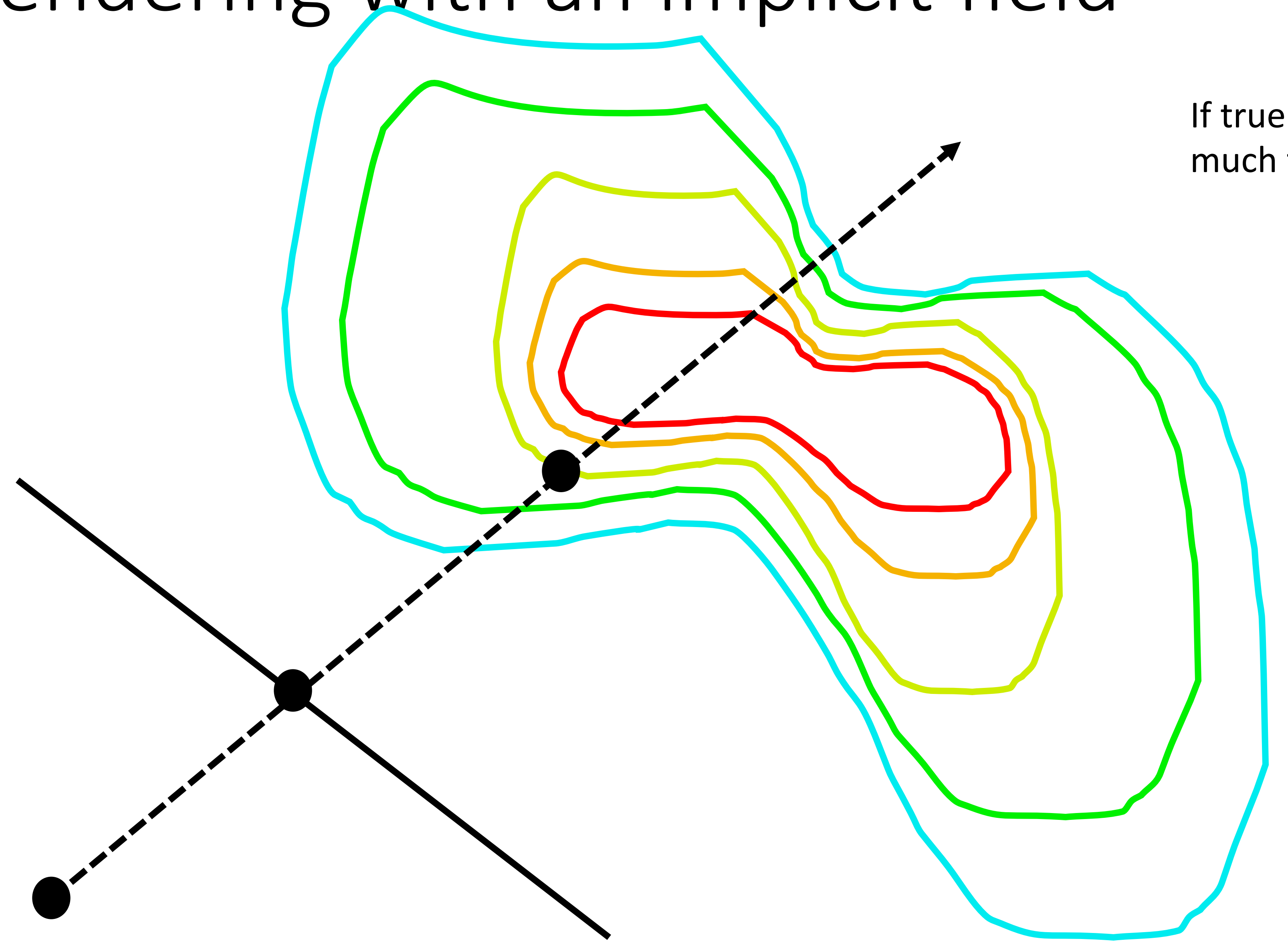
Shape representations



(a) Voxel

- Easy to produce
- Very expensive to store
- Limited resolution

Rendering with an implicit field

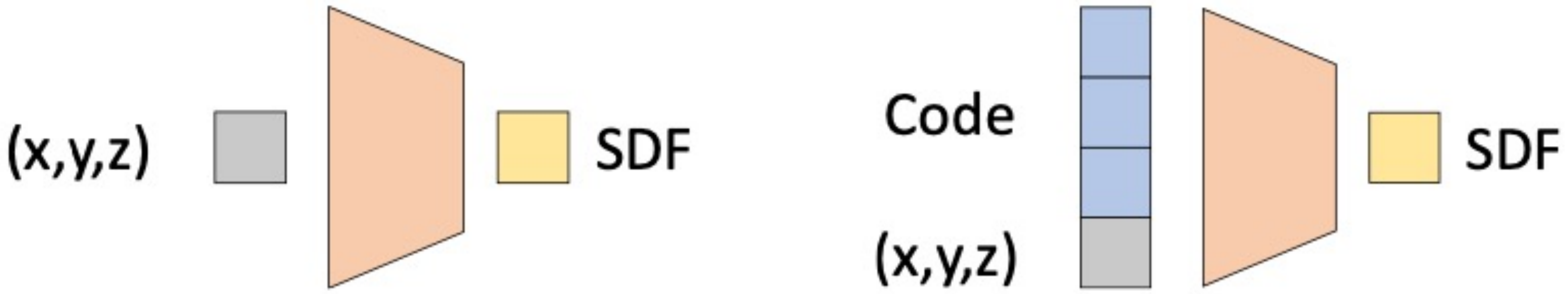


If true SDF, then can perform much faster – Sphere tracing

Generalization with neural fields

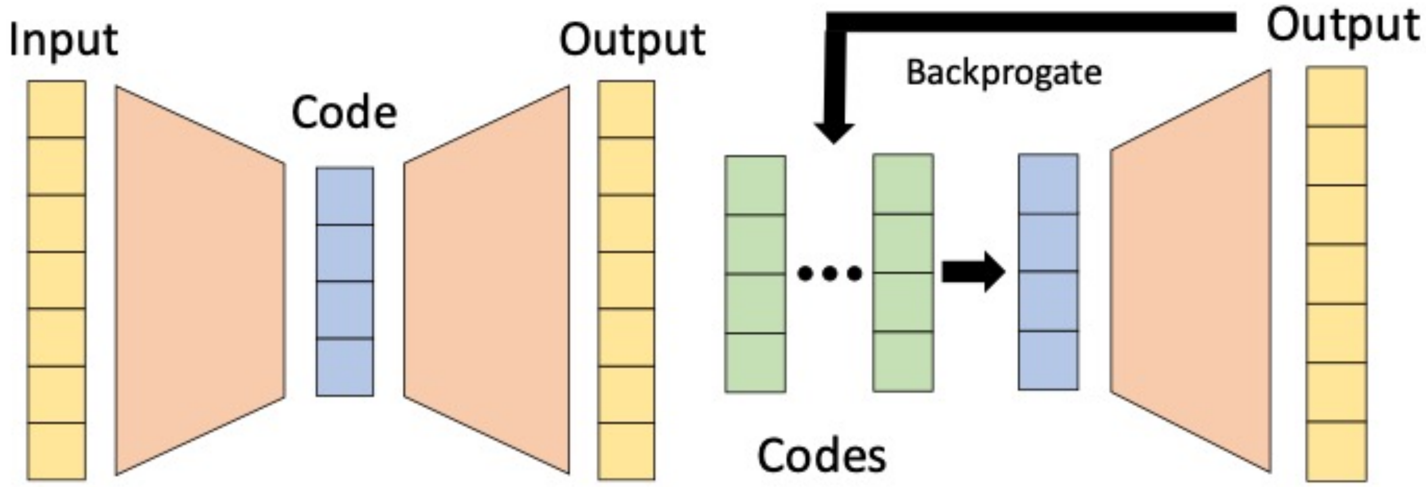
- Each neural field captures a particular shape
- Shape is encoded in the weights of the neural network
- How to generalize to new shapes?
  - Latent codes
  - Transfer learning

Implicit fields with latent codes



(a) Single Shape DeepSDF (b) Coded Shape DeepSDF

Producing latent codes for input shapes



(a) Auto-encoder **(b)** Auto-decoder

Fitting an implicit field

- Occupancy
  - Essentially a binary classification problem
  - Sample points, label them as inside or outside the surface

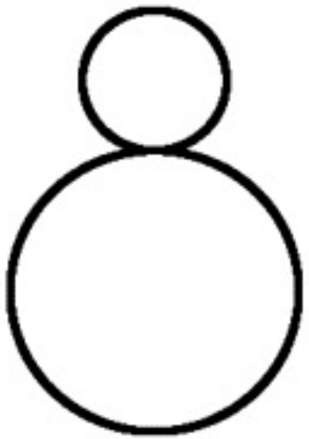
- SDF
  - Essentially a regression problem
  - Sample points, label them with true signed distance

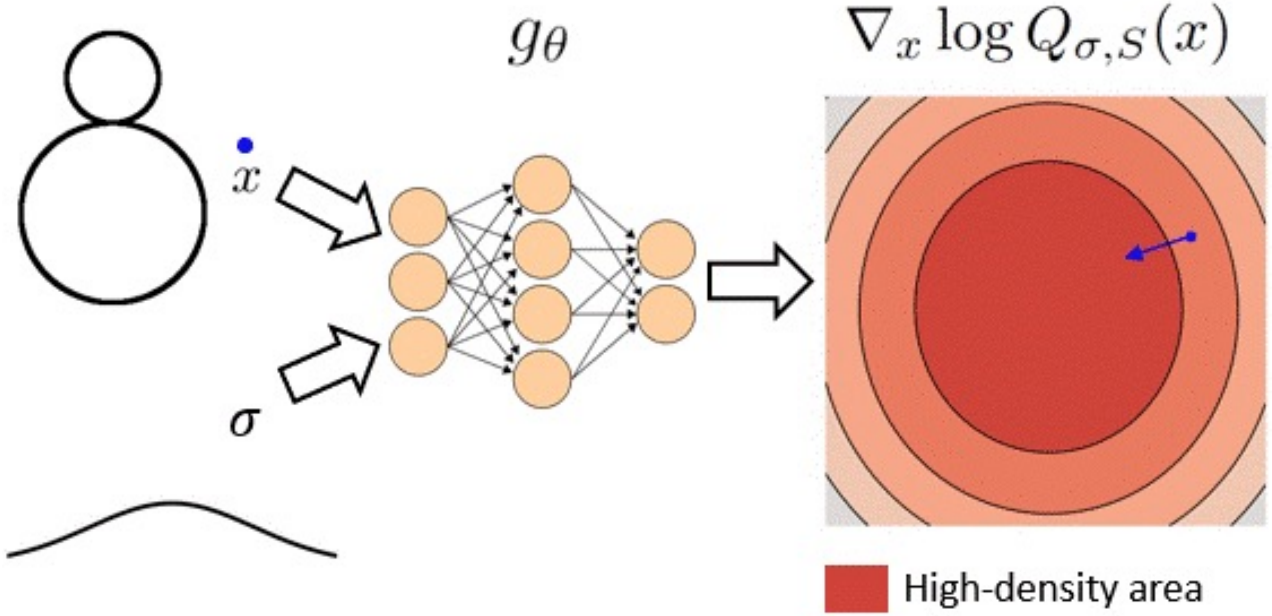
- In both cases, need watertight meshes to compute

Fitting implicit fields from point clouds

- Most 3D data comes in the form of point clouds
- Watertight meshes / ground truth SDFs generally hard to acquire
- How to train with point clouds?
- One approach: assume point clouds are sampled from underlying distribution
- Thus shape = generative model!

Fitting implicit fields from point clouds





Cai, Ruojin, et al. "Learning gradient fields for shape generation." *European Conference on Computer Vision*. Springer, Cham, 2020.

Representing high frequency details

- Standard neural networks use ReLU as activation
- So they approximate functions with piecewise linear functions
- Bad idea for high-frequency signals
  - Common in images, textured 3D surfaces etc
  - Need lots and lots of pieces!

Representing high frequency details – Fourier features

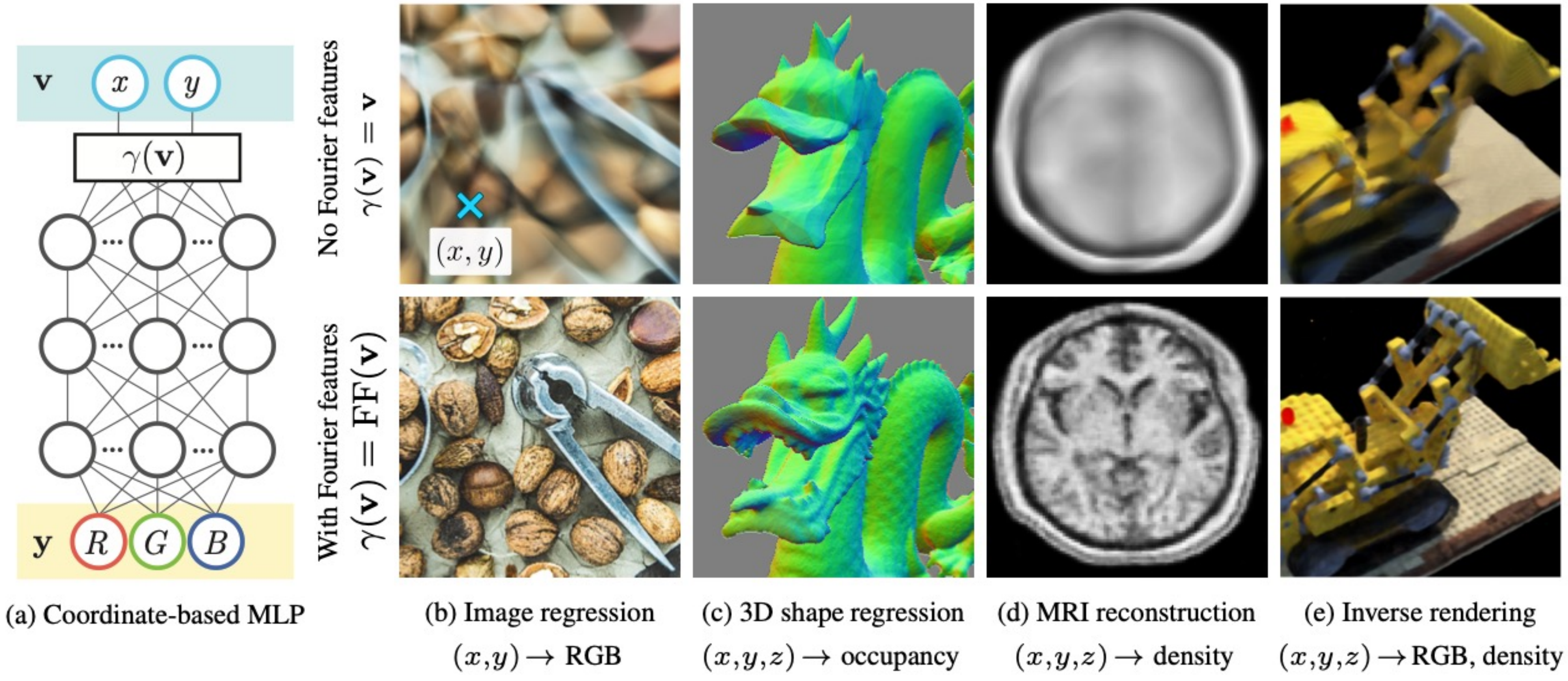
$$\mathbf{v} = (x, y, z)$$

$$\gamma(\mathbf{v}) = \left[a_1 \cos(2\pi \mathbf{b}_1^{\mathrm{T}} \mathbf{v}), a_1 \sin(2\pi \mathbf{b}_1^{\mathrm{T}} \mathbf{v}), \ldots, a_m \cos(2\pi \mathbf{b}_m^{\mathrm{T}} \mathbf{v}), a_m \sin(2\pi \mathbf{b}_m^{\mathrm{T}} \mathbf{v})\right]^{\mathrm{T}}$$

- Instead of $f(\mathbf{v})$, do $f(\gamma(\mathbf{v}))$

Tancik, Matthew, et al. "Fourier features let networks learn high frequency functions in low dimensional domains." *arXiv preprint arXiv:2006.10739* (2020).

Representing high frequency details – Fourier features

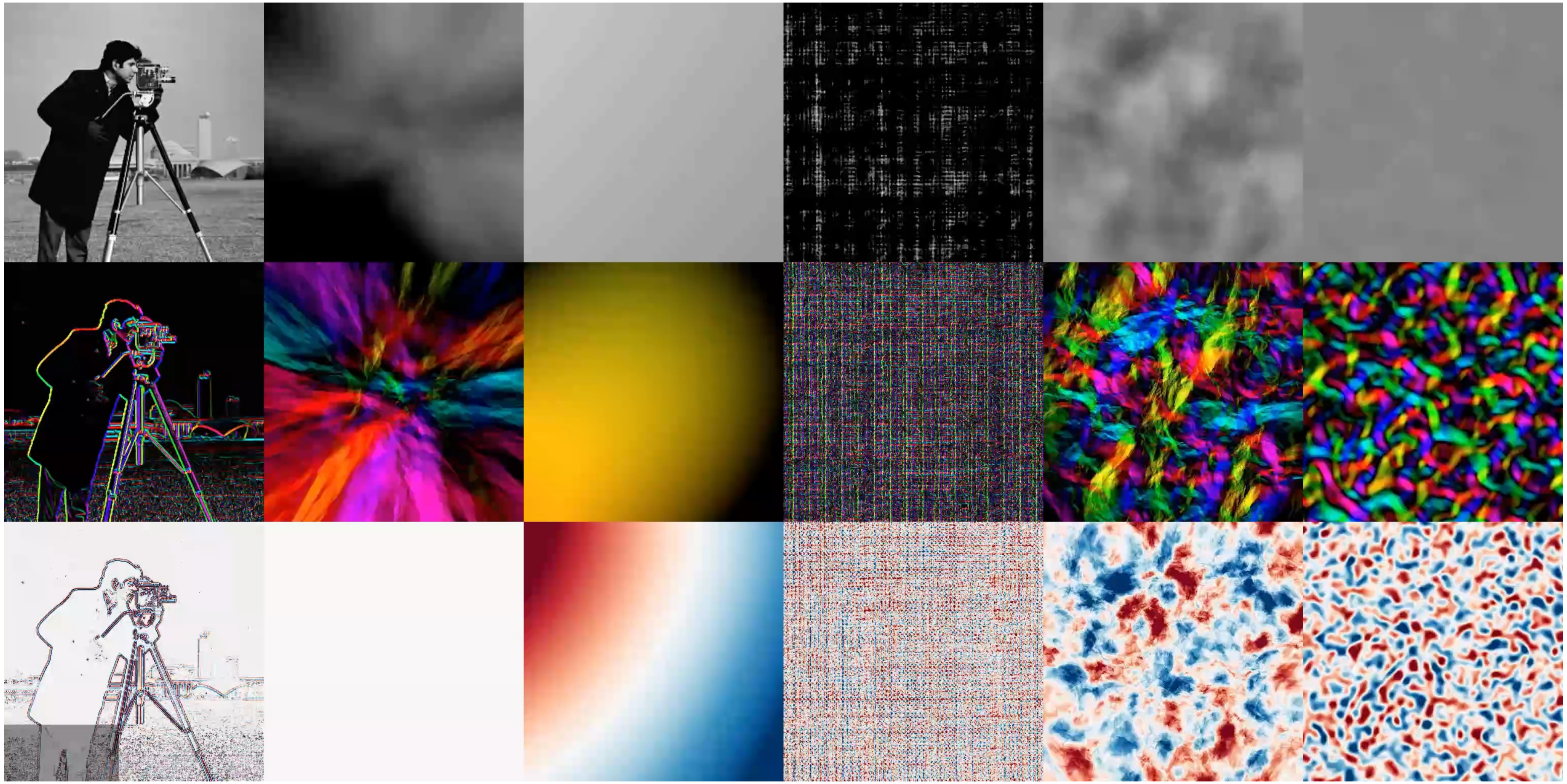


Tancik, Matthew, et al. "Fourier features let networks learn high frequency functions in low dimensional domains." *arXiv preprint arXiv:2006.10739* (2020).

Representing high frequency details - SIREN

- Instead of ReLU activations use sinusoidal activation
- Side-effect – all derivatives exist and are themselves SIREN models
  - Allows to model both signal and derivative

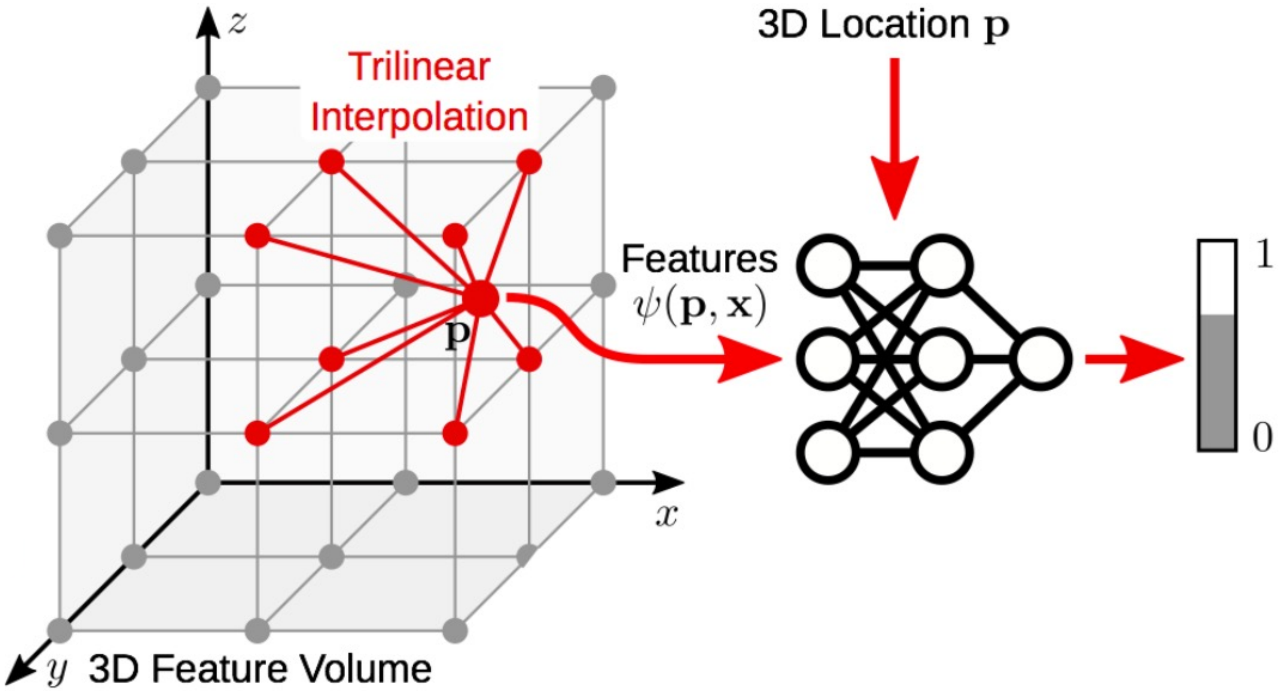
Representing high frequency details



Sitzmann, Vincent, et al. "Implicit neural representations with periodic activation functions." *Advances in Neural Information Processing Systems* 33 (2020).

Scene representations and detail – hybrid representations

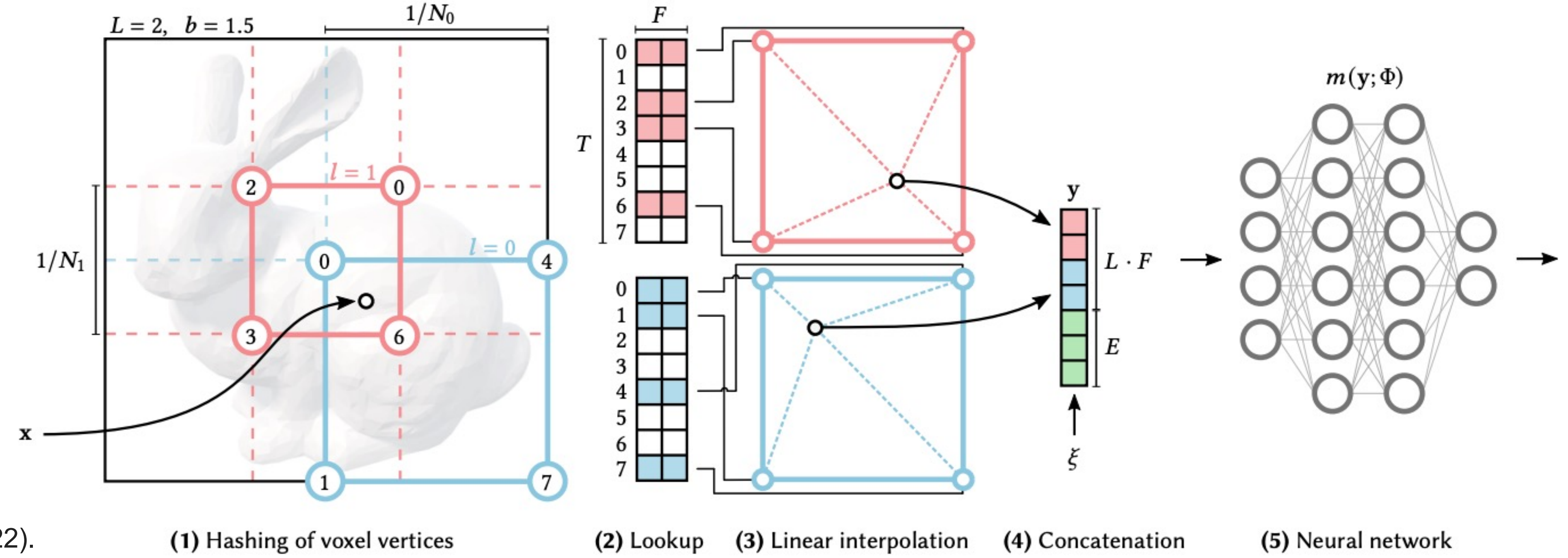
- Use a voxelized feature volume
- For each 3D point, index into feature volume with interpolation
  - Location-dependent "latent code"!
- Use MLP to decode latent code into occupancy



Peng, Songyou, et al. "Convolutional occupancy networks." *European Conference on Computer Vision*. Springer, Cham, 2020.

Scene representations and detail – hybrid representations

- Challenge: might need many many voxels
  - With multiple spatial resolutions
- Once again memory constraints
- Idea: maintain a smaller hash table of features
  - Hash voxel coordinates



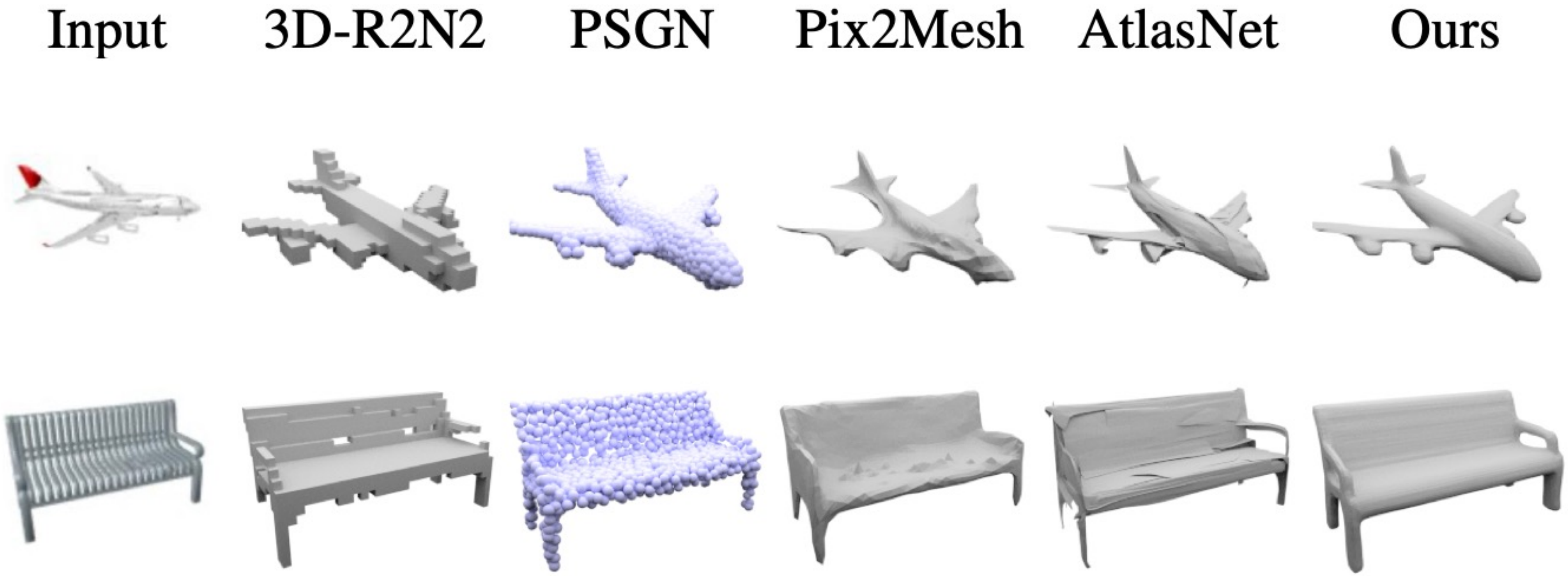
Müller, Thomas, et al. "Instant neural graphics primitives with a multiresolution hash encoding." *arXiv preprint arXiv:2201.05989* (2022).

Generalizing neural fields through transfer learning

- Use meta-learning framework
- Learn *initialization for network* θ_0
- In each training iteration
  - Sample a shape
  - Perform SGD steps to update parameters to $\theta_0 + \Delta\theta$
  - Backpropagate final loss to update θ_0
- Compared to latent code approach, allows greater fidelity/cheaper networks since new shapes can use different weights

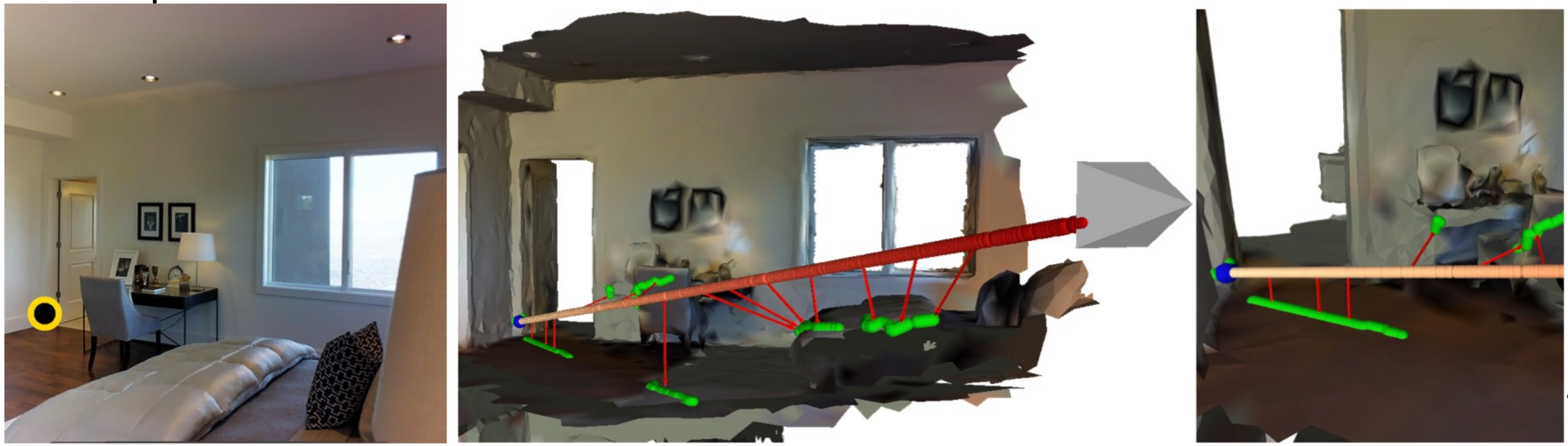
Sitzmann, Vincent, et al. "Metasdf: Meta-learning signed distance functions." *arXiv preprint arXiv:2006.09662* (2020).

Using implicit fields for 3D reconstruction



From single objects to scenes – problems with distance fields

- Signed distance fields no longer meaningful
- Unsigned distance fields meaningful but hard to analyze
- One approach: ray distance
  - For each point on the ray, distance to nearest intersection of the ray
  - But dependent on view



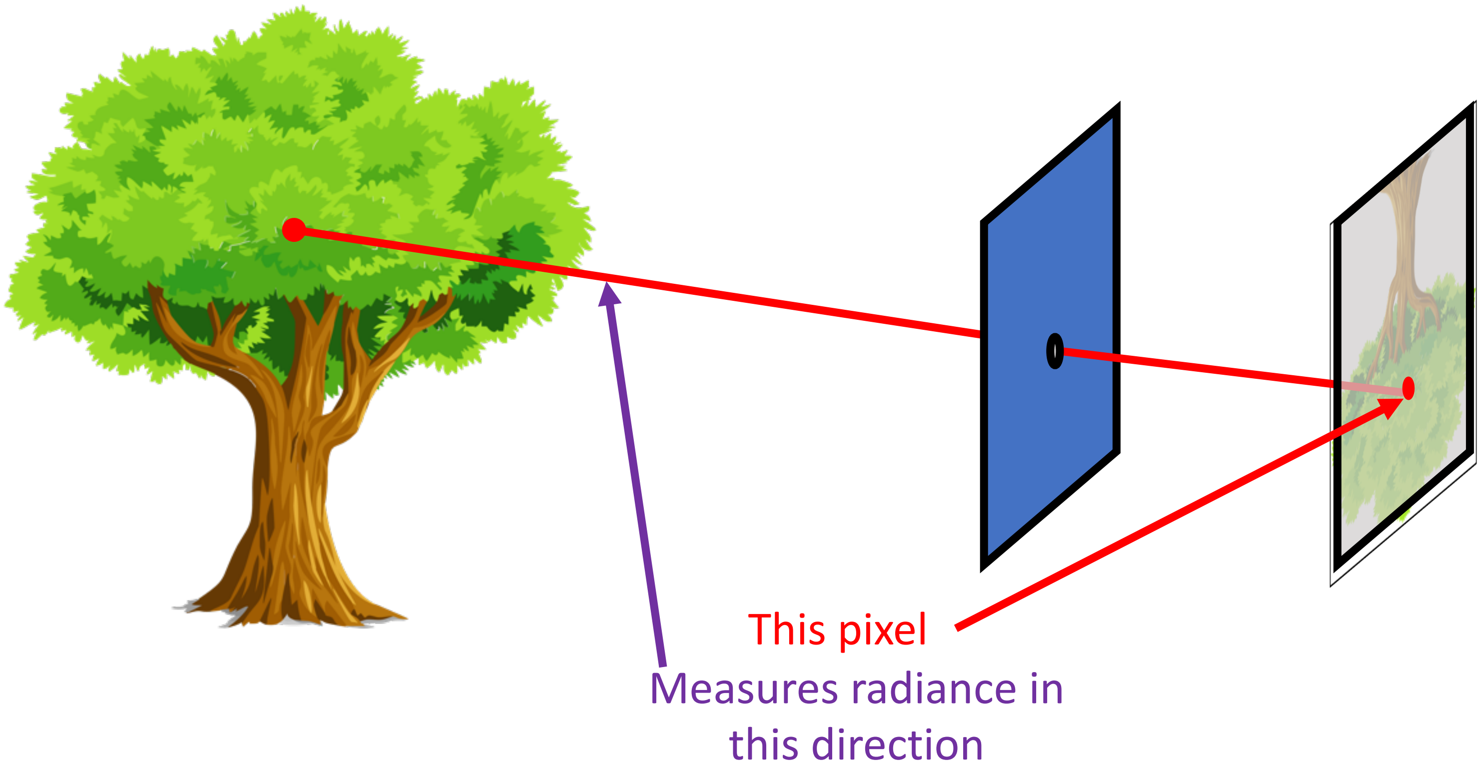
(a) Image with ray center (b) Third person 3D views with the red ray and nearest points

Kulkarni, Nilesh, Justin Johnson, and David F. Fouhey. "What's Behind the Couch? Directed Ray Distance Functions (DRDF) for 3D Scene Reconstruction." *arXiv preprint arXiv:2112.04481* (2021).

Neural fields of radiance

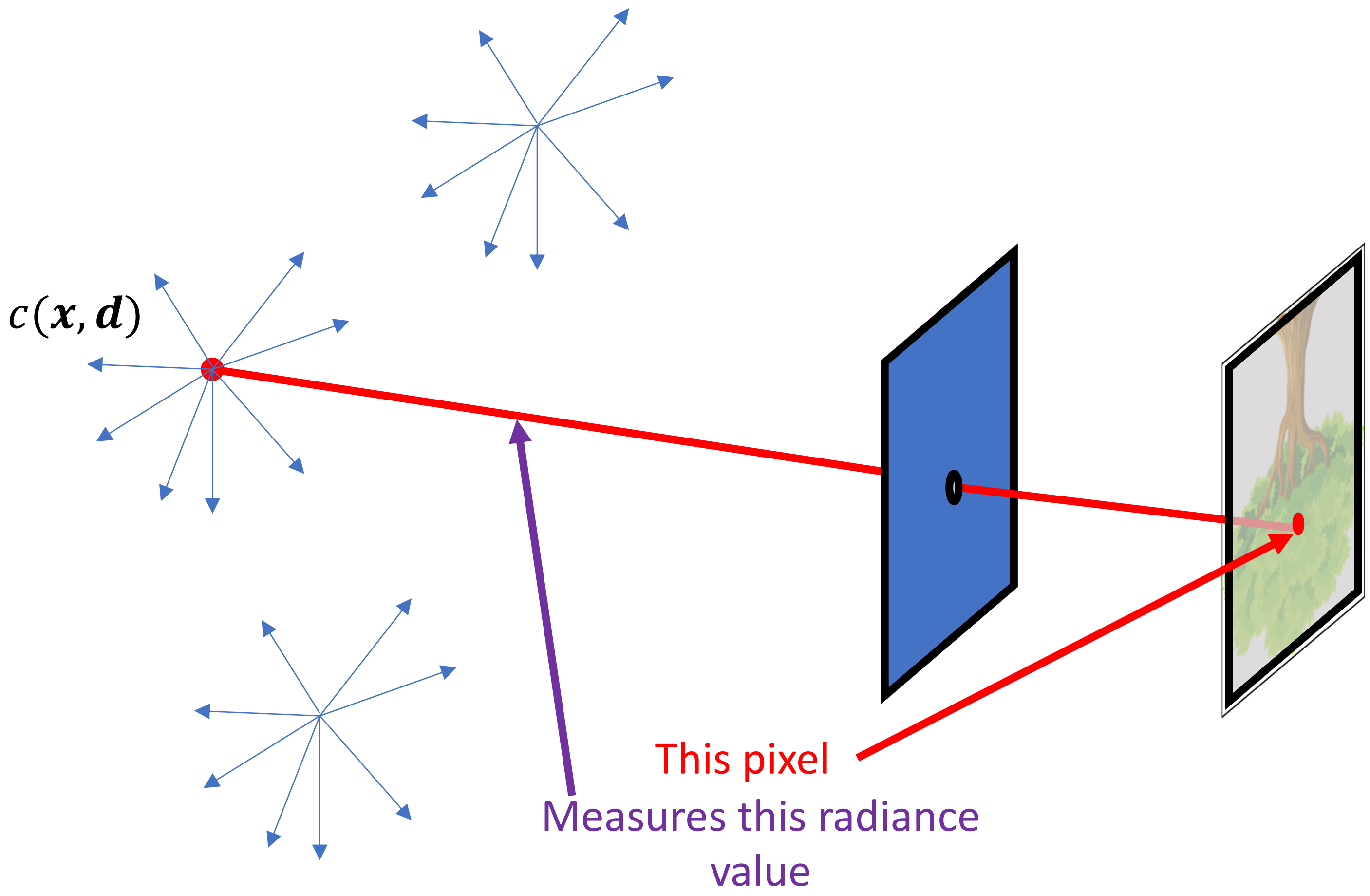
Radiance

- Pixels measure ***radiance***



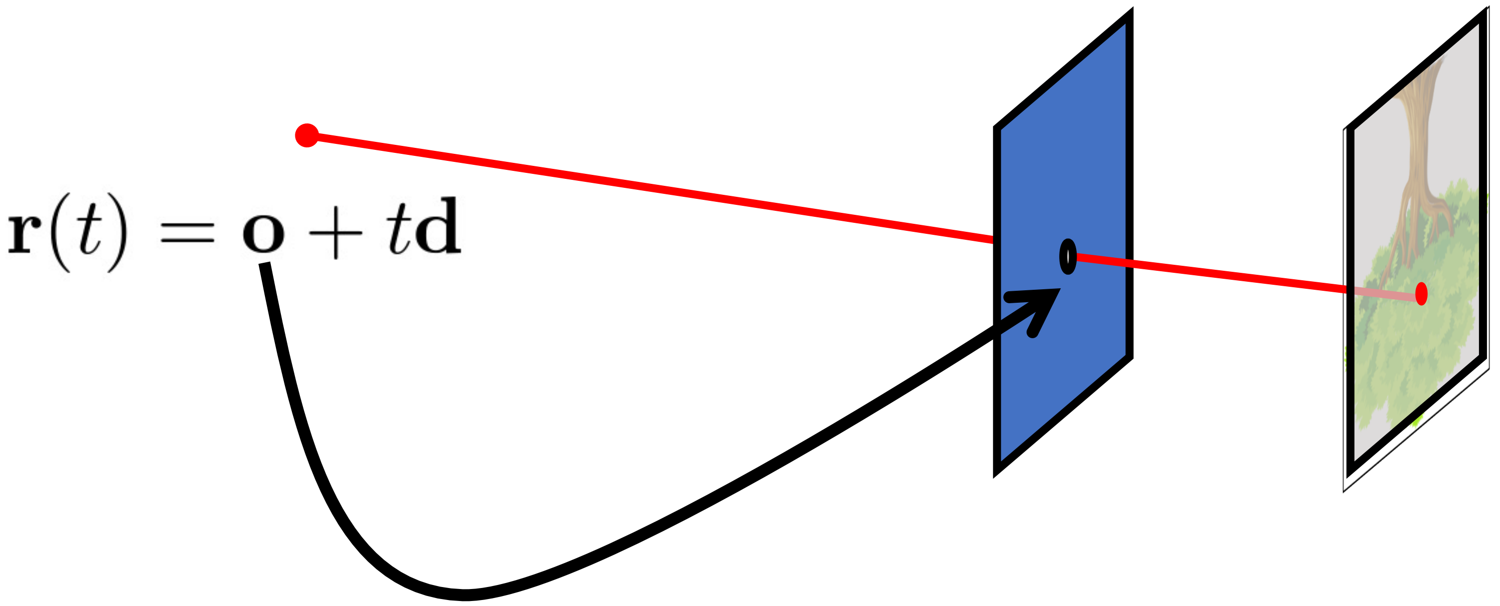
Radiance fields

- Radiance field $c(\boldsymbol{x}, \boldsymbol{d})$
- Also have density $\sigma(\boldsymbol{x})$: where are the surfaces?



Volume rendering with radiance fields

- Pixels measure ***radiance***



Volume rendering with radiance fields

Integral along ray

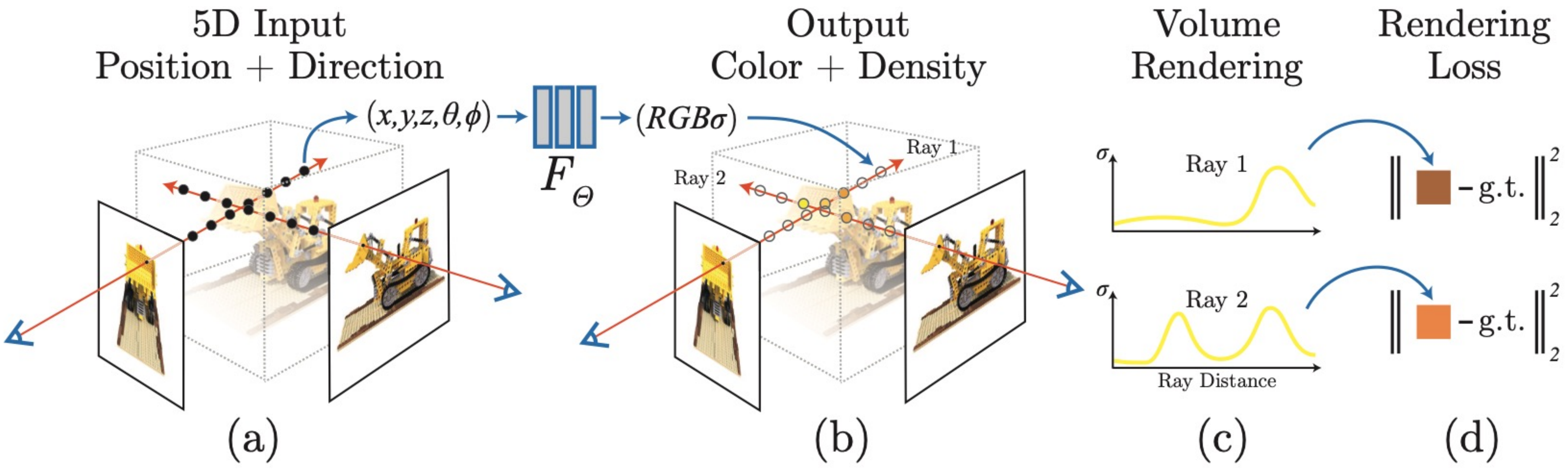
Observed color

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})dt$$

Soft visibility

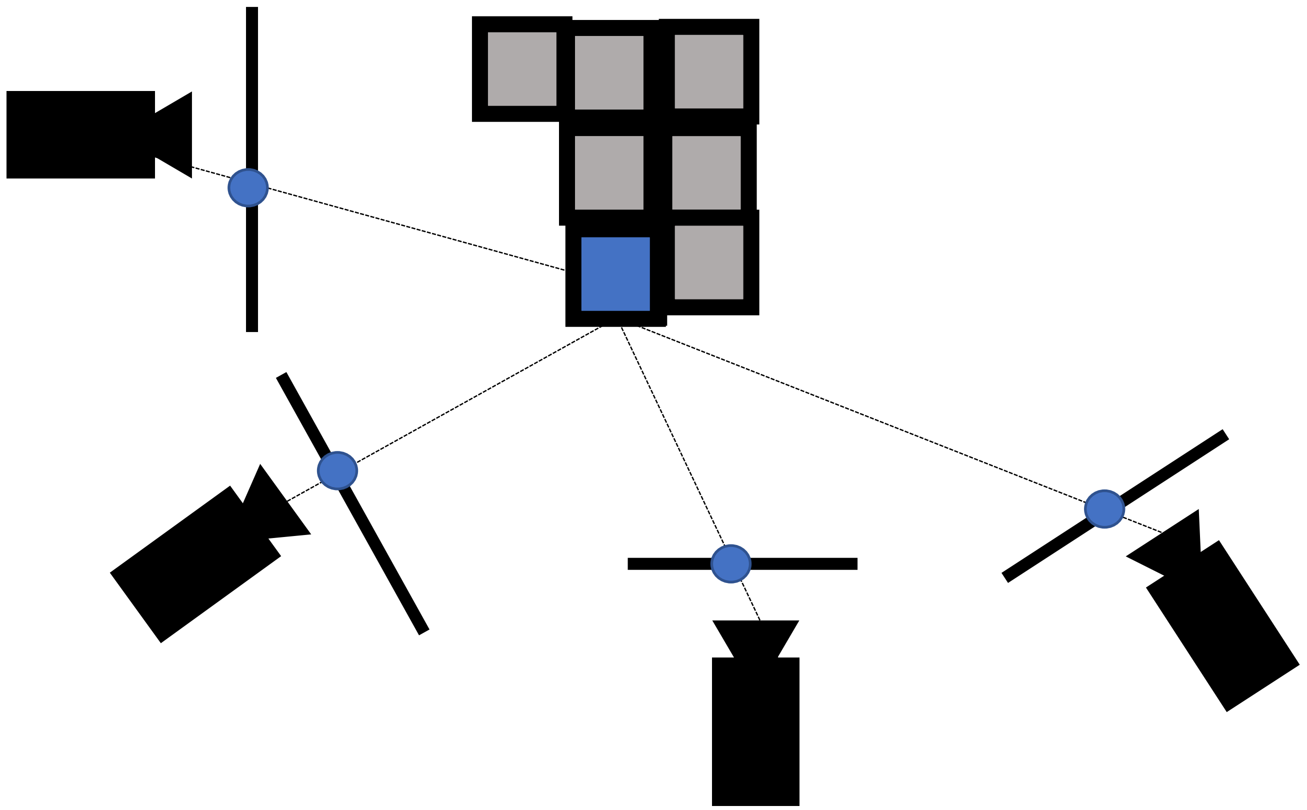
$$T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds\right)$$

Neural radiance fields

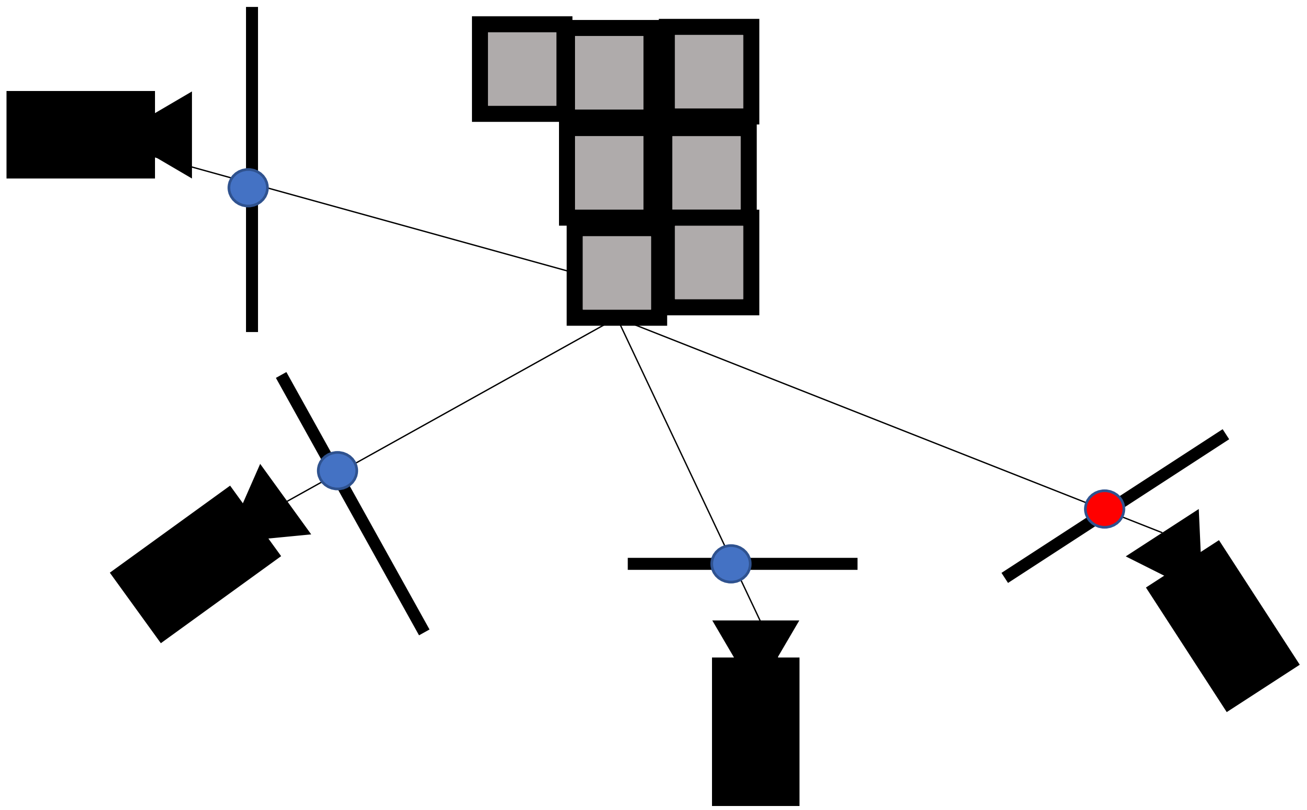


Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." *European conference on computer vision*. Springer, Cham, 2020.

Connections to classical algorithms: Space carving



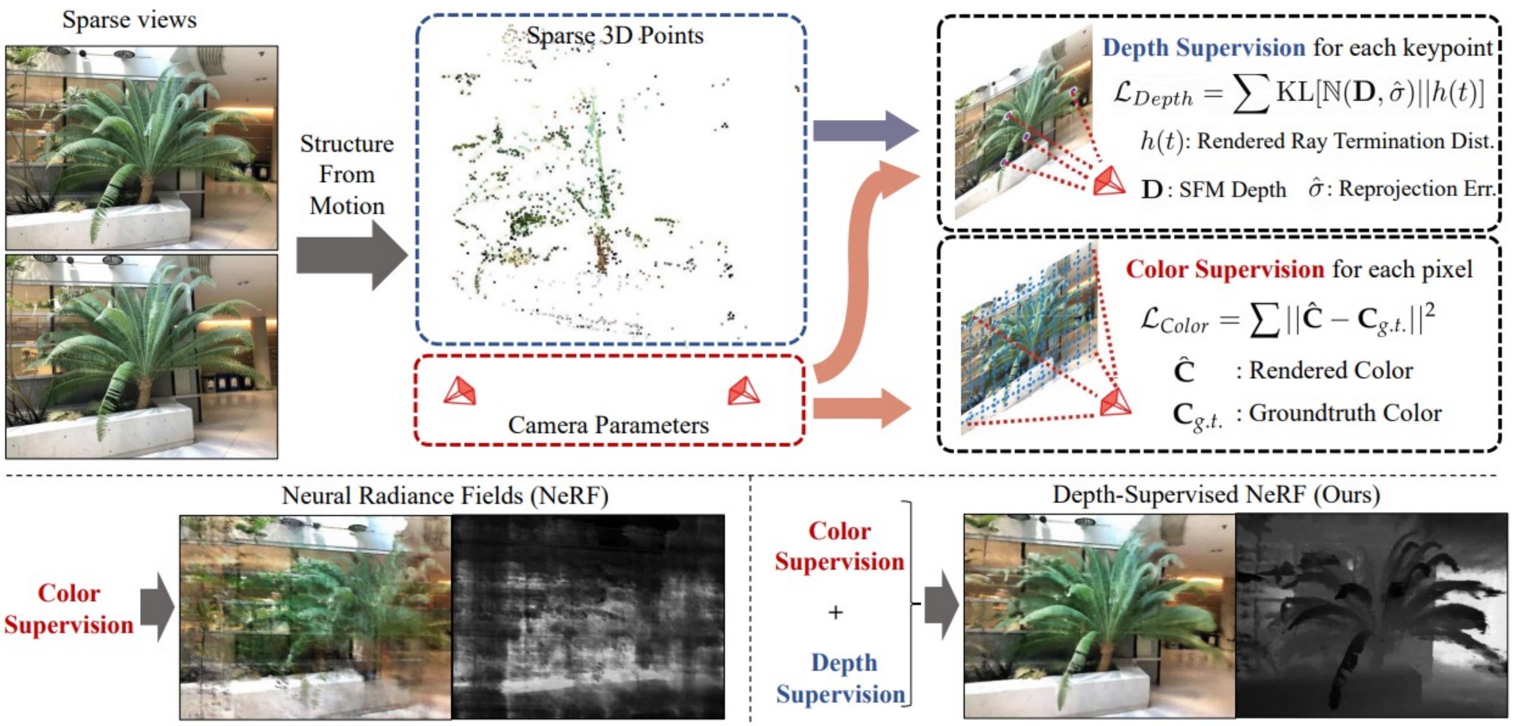
Connections to classical algorithms: Space carving



Leveraging classical 3D reconstruction

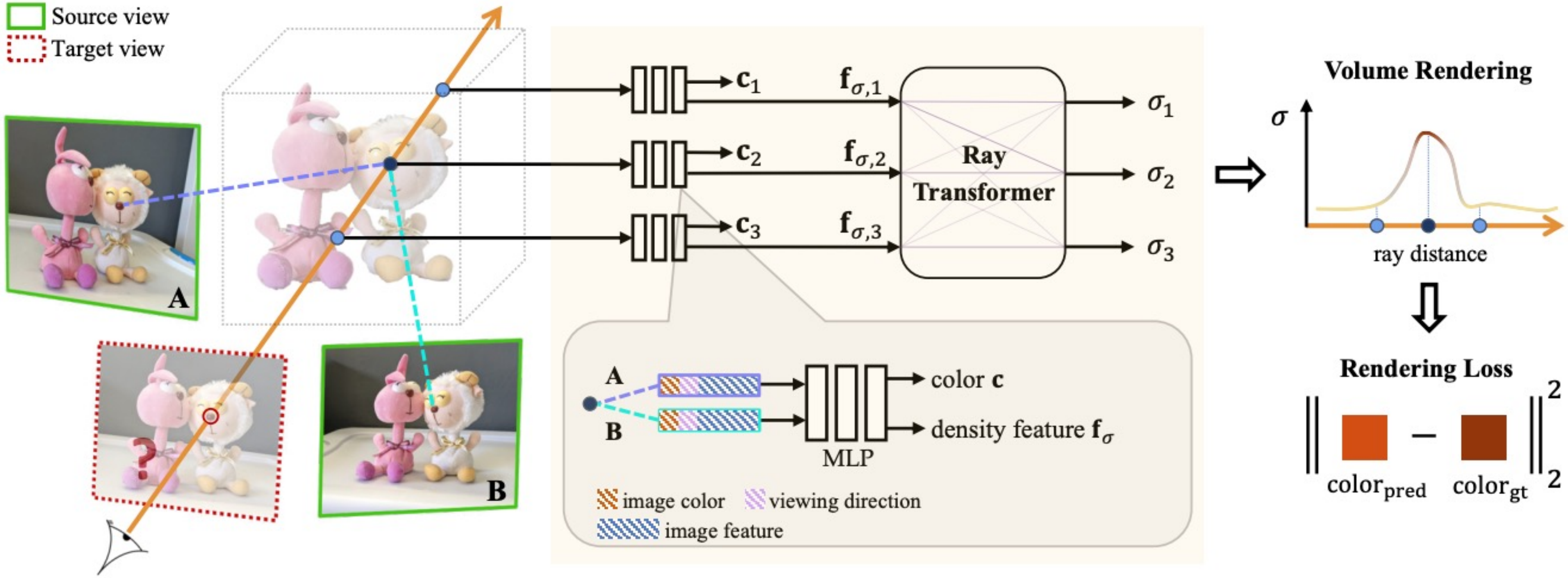
Depth-supervised NeRF: Fewer Views and Faster Training for Free

Kangle Deng[1] Andrew Liu[2] Jun-Yan Zhu[1] Deva Ramanan[1,3]
[1]Carnegie Mellon University [2]Google [3]Argo AI



Generalizing neural radiance fields across scenes

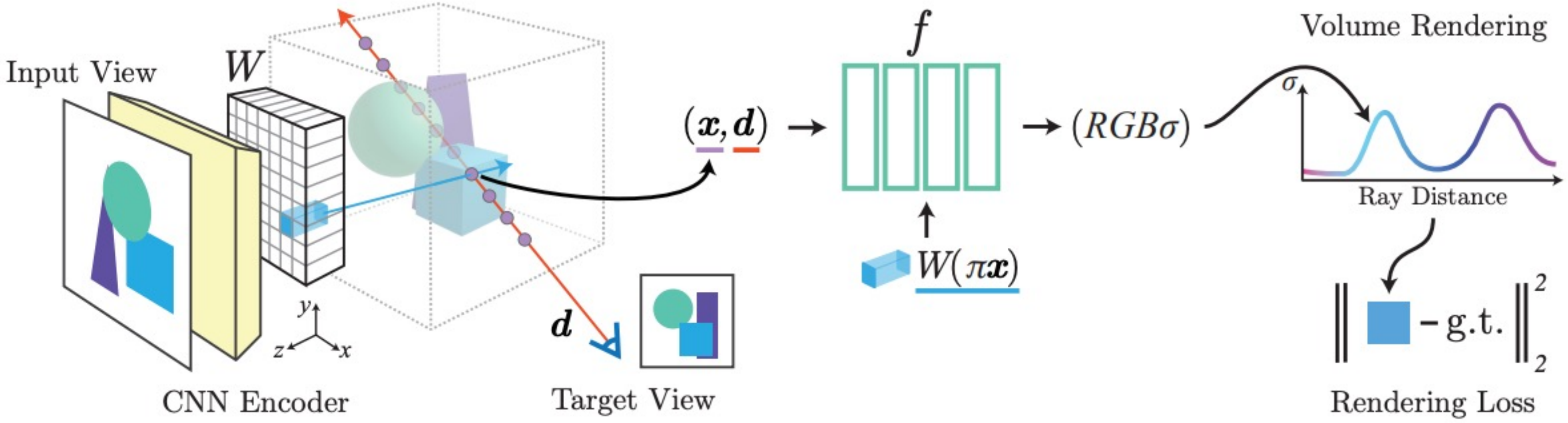
- Key idea: have neural network explicitly look up other views instead of storing radiance



Wang, Qianqian, et al. "Ibrnet: Learning multi-view image-based rendering." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

Generalizing neural radiance fields across scenes

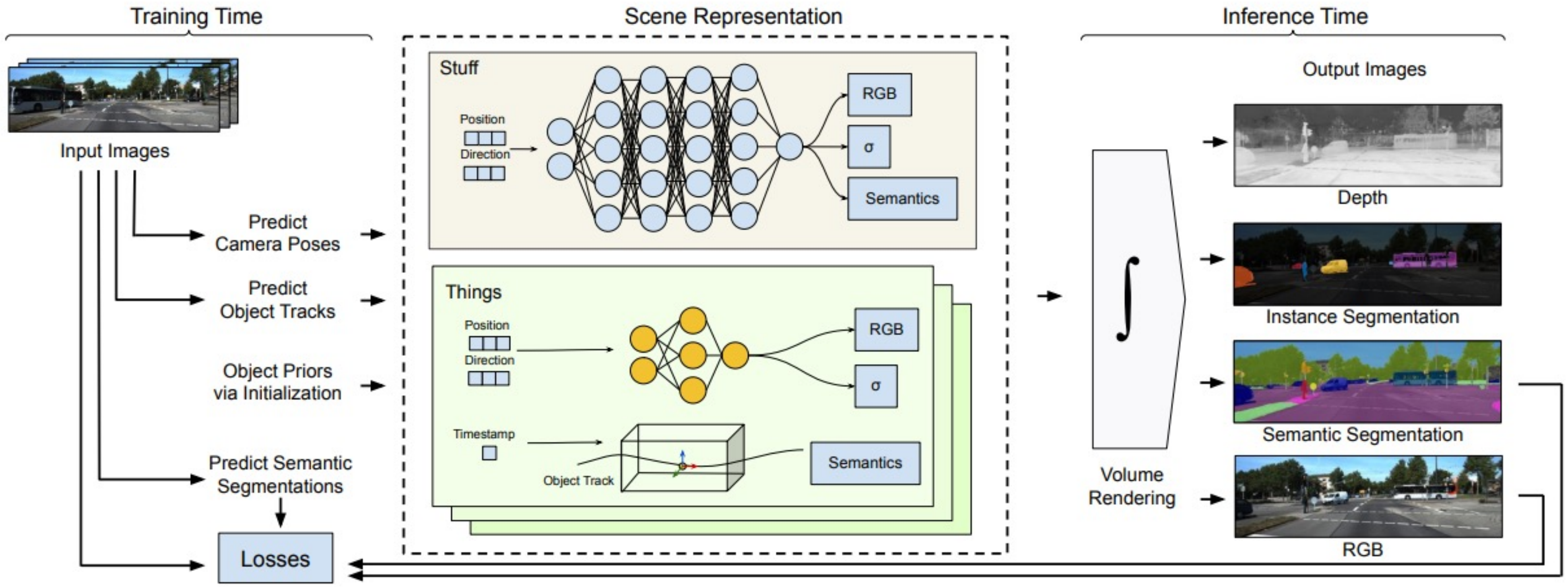
- Key idea: have neural network explicitly look up other views instead of storing radiance



Yu, Alex, et al. "pixelnerf: Neural radiance fields from one or few images." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.

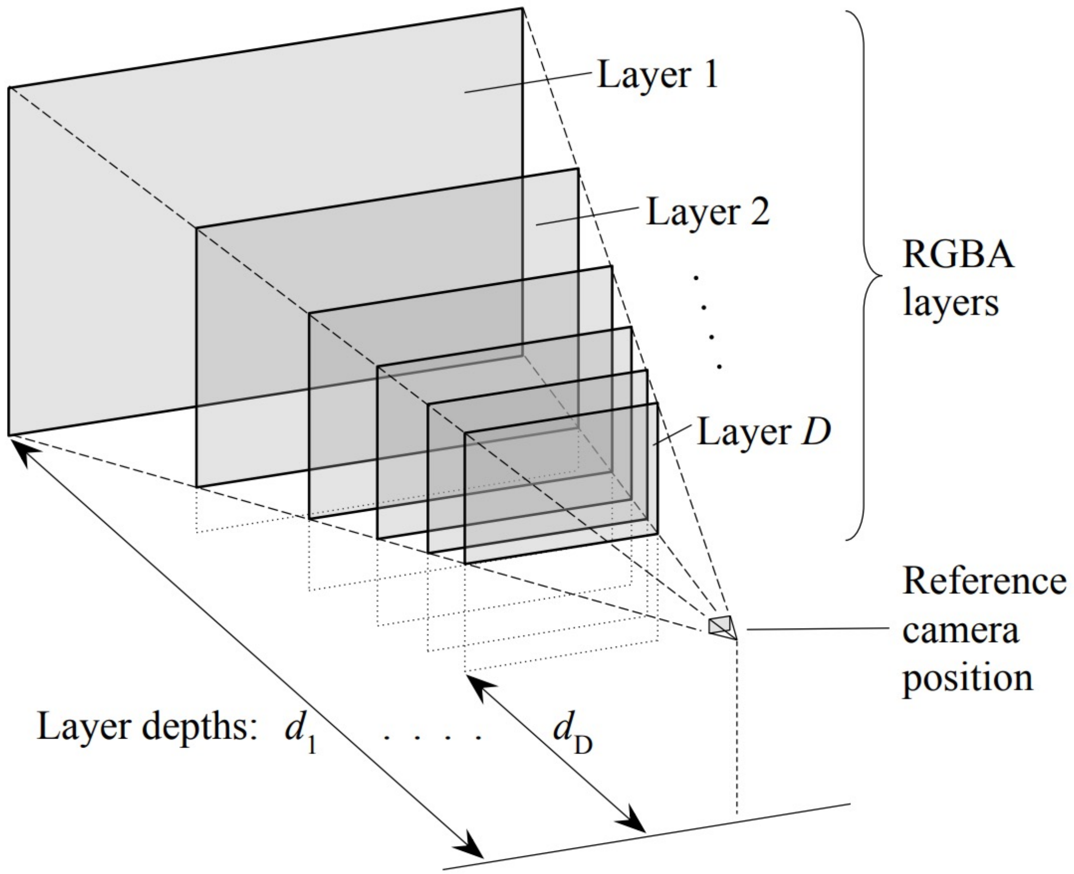
Neural fields of semantics

- Can use neural fields to store not just color but also semantics
- Useful way to encode cross-view consistency of recognition



Kundu, Abhijit, et al. "Panoptic Neural Fields: A Semantic Object-Aware Neural Scene Representation." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.

Other representations of 3D structure: Multiplane images



Tucker, Richard, and Noah Snavely. "Single-view view synthesis with multiplane images." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.

Challenges with neural fields

- Shape information is stored in neural network weights
    - Difficult to edit
- Appearance information entangled with shape and pose
- Generalization across complex scenes still work in progress