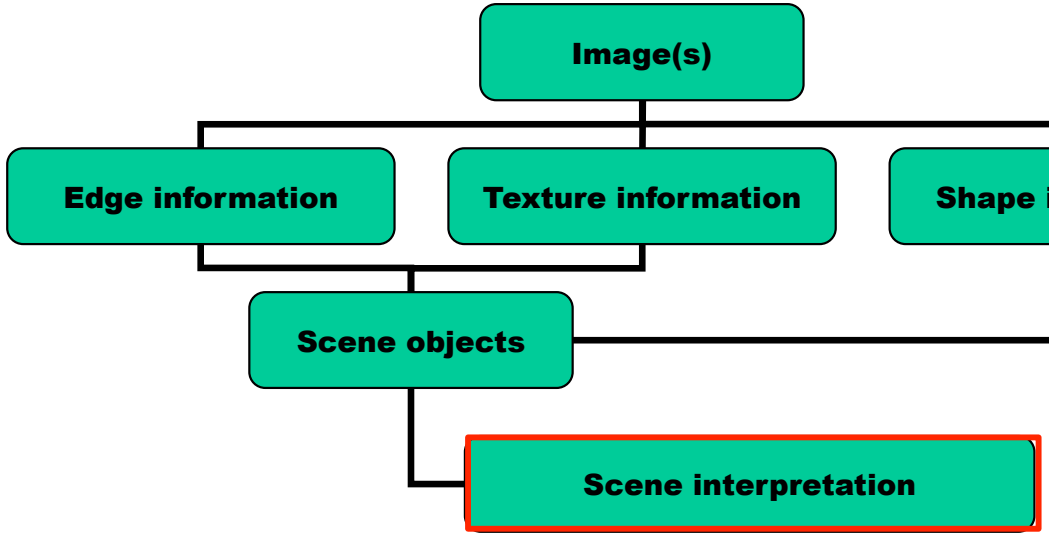
Course Outline

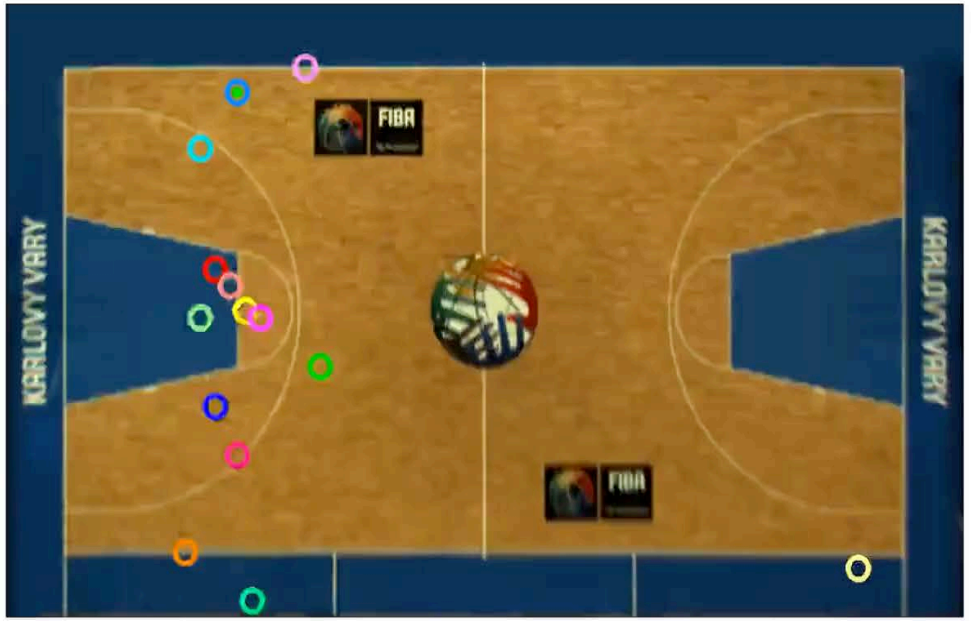


Introduction:
- Image formation

Extracting features:
- Contours
- Texture
- Regions

Shape recovery:
- From one image
- Using additional images

Tracking and Counting



From tracking individuals .....

Tracking and Counting

From tracking individuals ….. to tracking crowds.

Liu et al. , CVPR'22

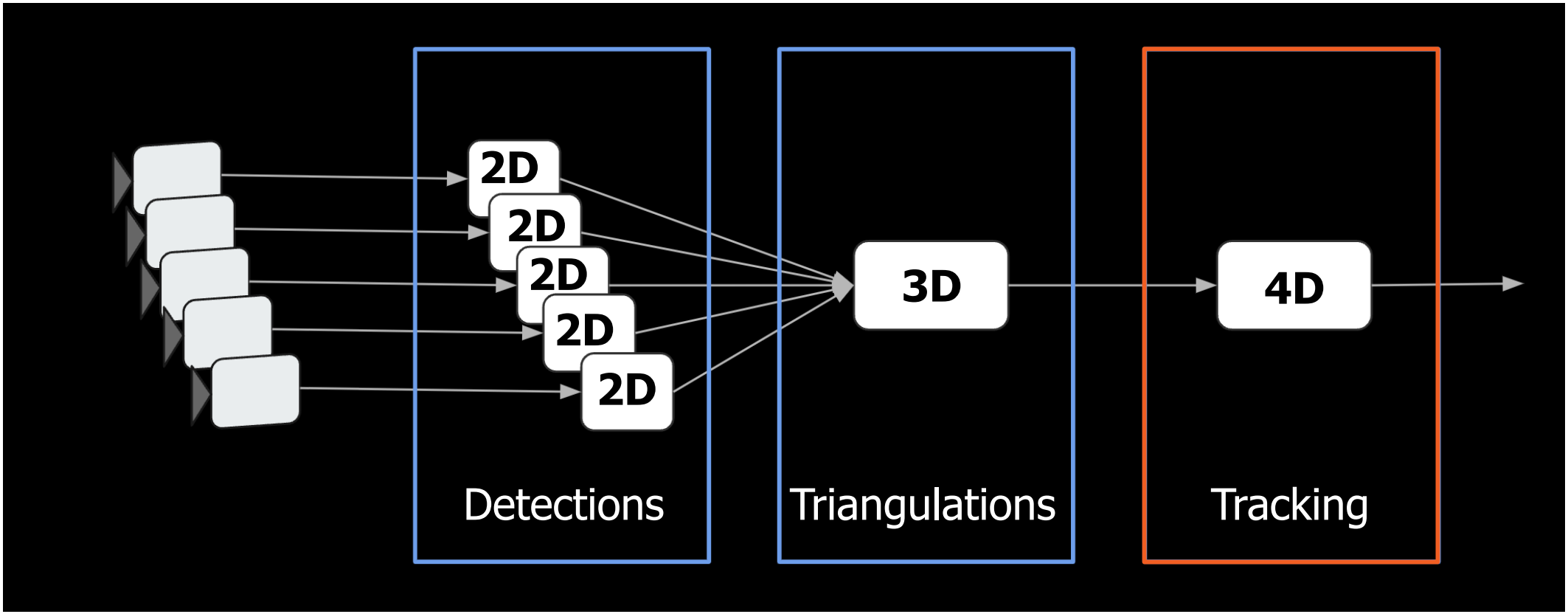
Tracking and Counting

From counting people … to counting stacked objects.

Multi-View Tracking of Soccer Players

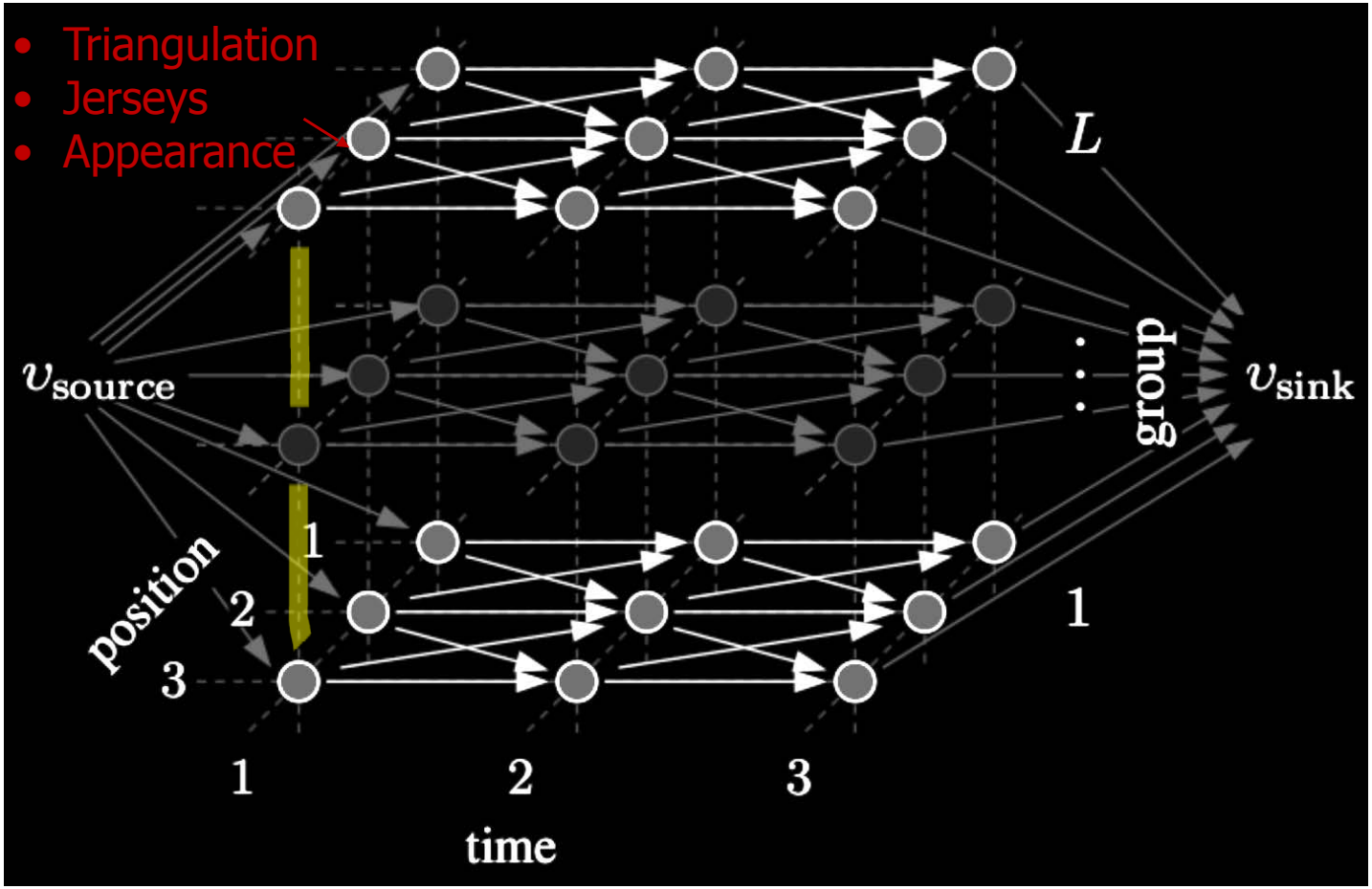
- Shallow CNNs for background subtraction.
- Deeper ones to detect the joints of players.
- Cellphones used as cameras.

Tracking Soccer Players



- 2D detections of joints in individual images.
- Triangulation across images to get 3D positions.
- Tracking over time.

Tracking over Time



Create a spatio temporal graph:

- The nodes are the detections.
- The edges are potential transitions from locations at time t to time t + dt.
- Find trajectories using linear programming.

Flow Based Formulation

Cost function and IP for people tracking

People Tracking Integer Program

$$\text{maximize} \sum_{(i,j)\in\mathcal{E}_p} p_i^j c_{pi}^j ,$$

$$\text{where} \quad c_{pi}^j = \frac{\log P_p(x_i|I^{t_i})}{1-\log P_p(x_i|I^{t_i})} ,$$

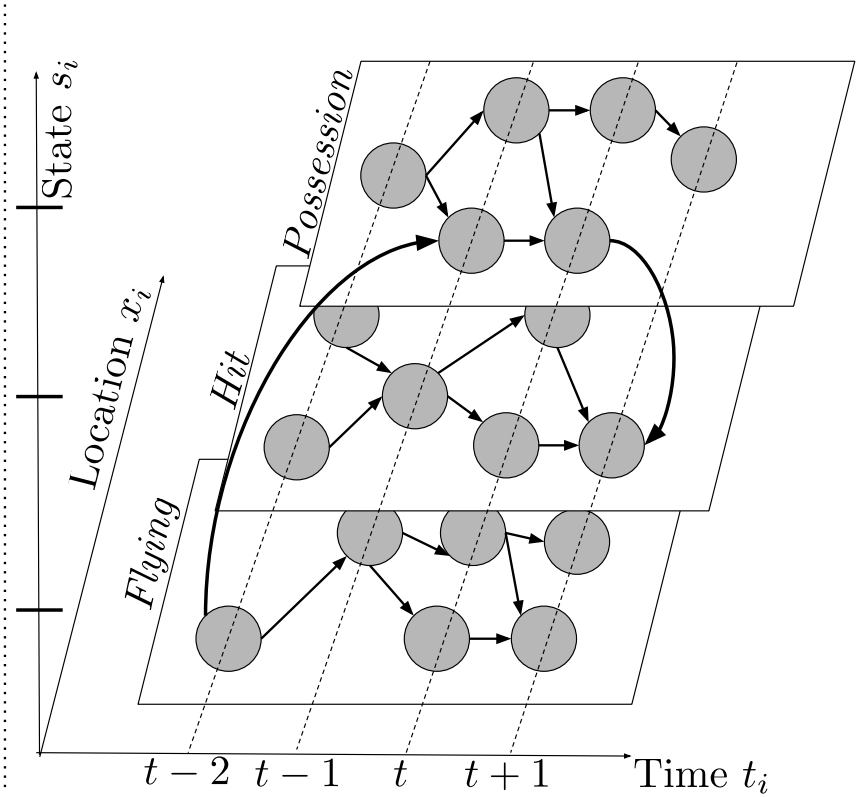
Joint Ball and People Tracking IP

$$\text{maximize} \sum_{(i,j)\in\mathcal{E}_b} f_i^j c_{bi}^j + \sum_{(i,j)\in\mathcal{E}_p} p_i^j c_{pi}^j$$

s.t.

$p_j^i \geq f_j^i$ **if in possession**

physical constraints otherwise



[Berclaz et al., 2011]

Tracking a Volleyball

Ball state indicated in top left

Red bounding box shows possession

The physics of the ball have to be modeled!

From Academia to the Real World

2005

2014

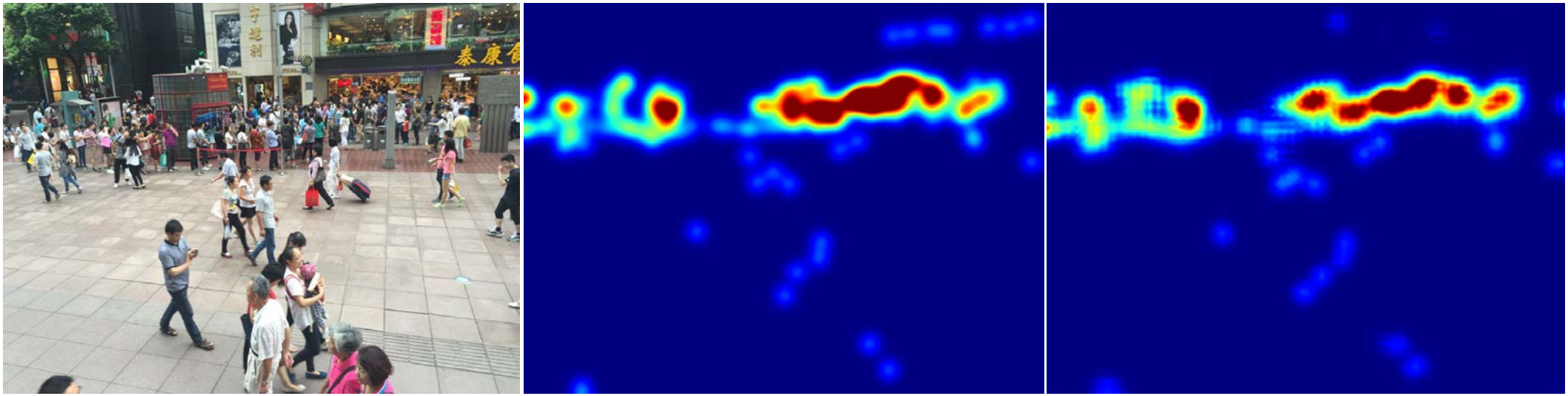
2015

2021

What about much Denser Crowds?

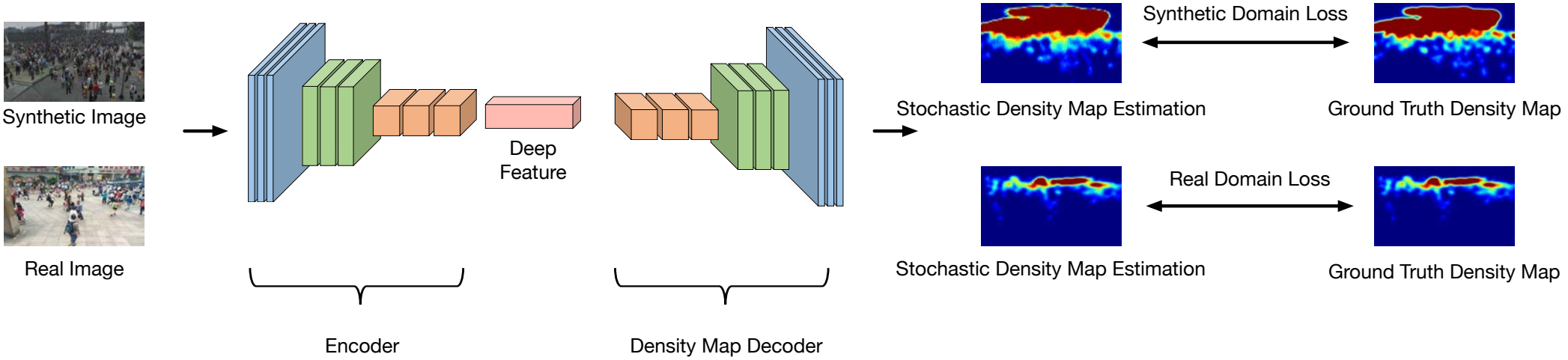
Do not track individuals but estimate density

Density and Perspective



- Density is heavily affected by perspective distortion
- Obtaining training data is time consuming
- Synthetic data can be used but it is not perfect

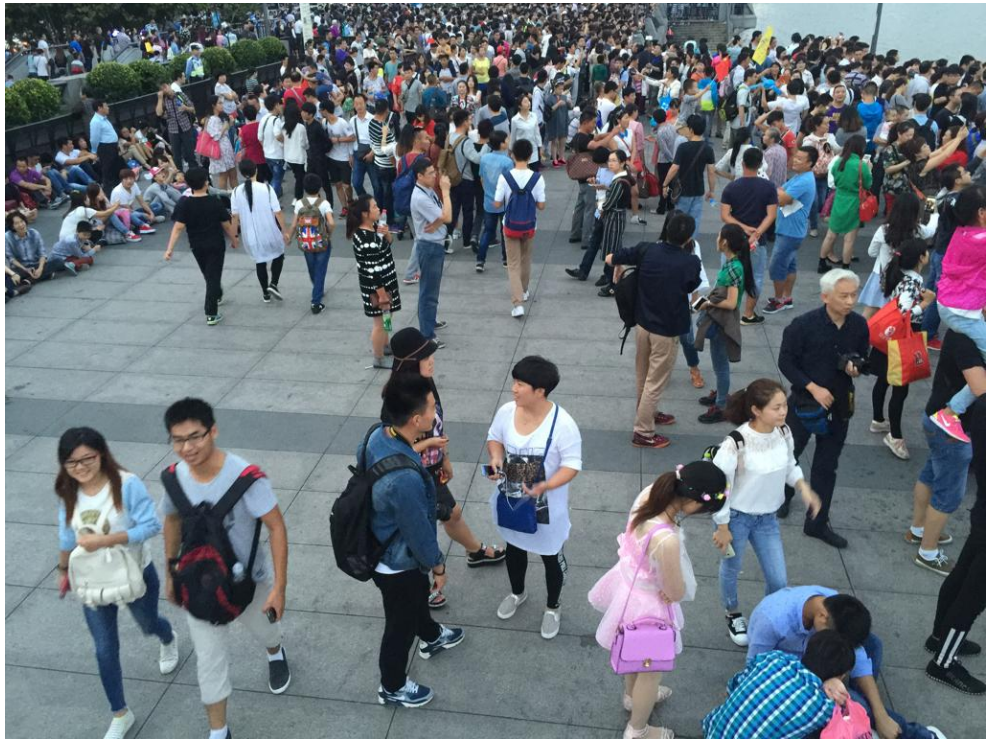
U-Net Counting



Train a U-Net on both real and synthetic data:

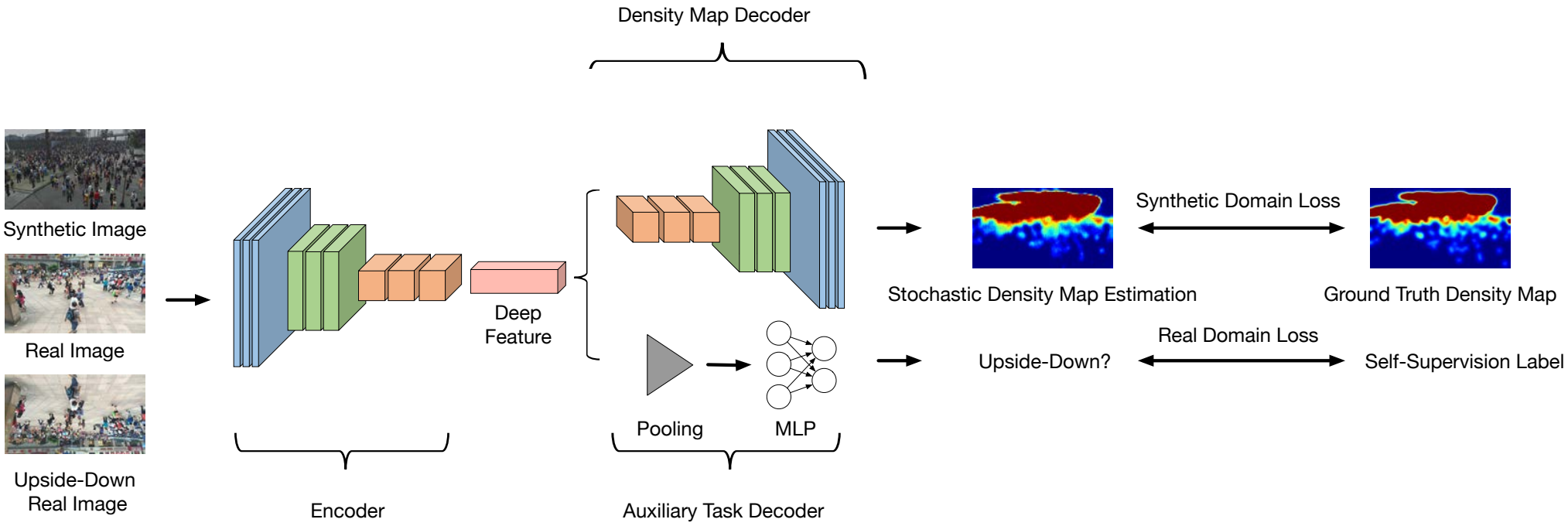
- How do you reduce the amount of necessary data?
- How do you label the real data?

Perspective Aware Features



- In normal images, the density seems higher near the top of the image.
- To learn features that account for perspective, train the network to recognize if an image is upside down.

Pre-Training

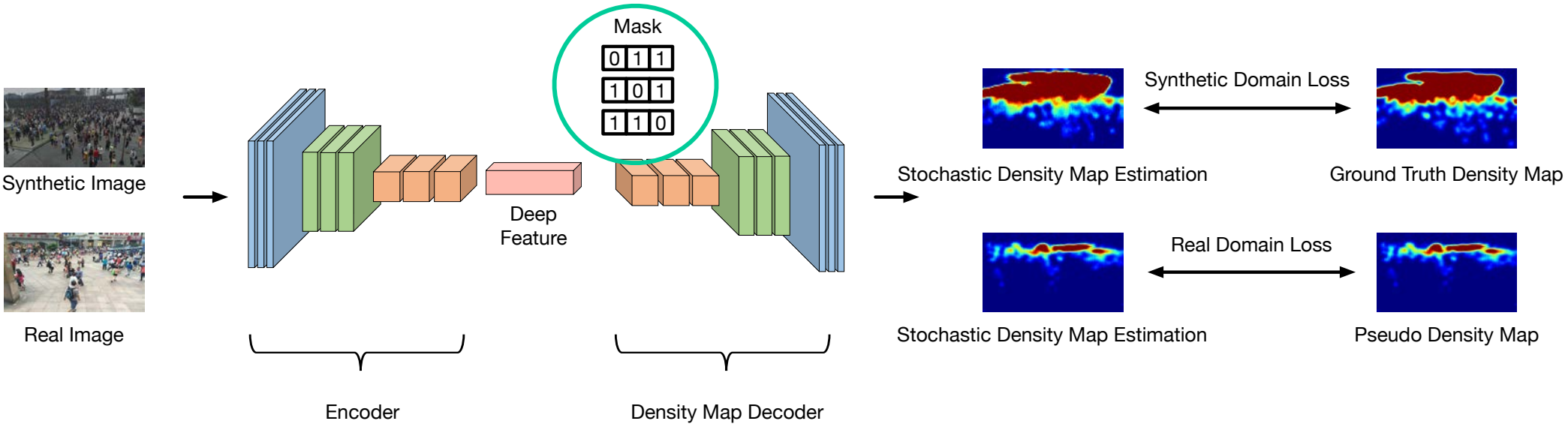


The network is trained to

- compute density on the synthetic images,
- to recognize if images are upside down.

➡ It leans perspective aware features.

Full-Training

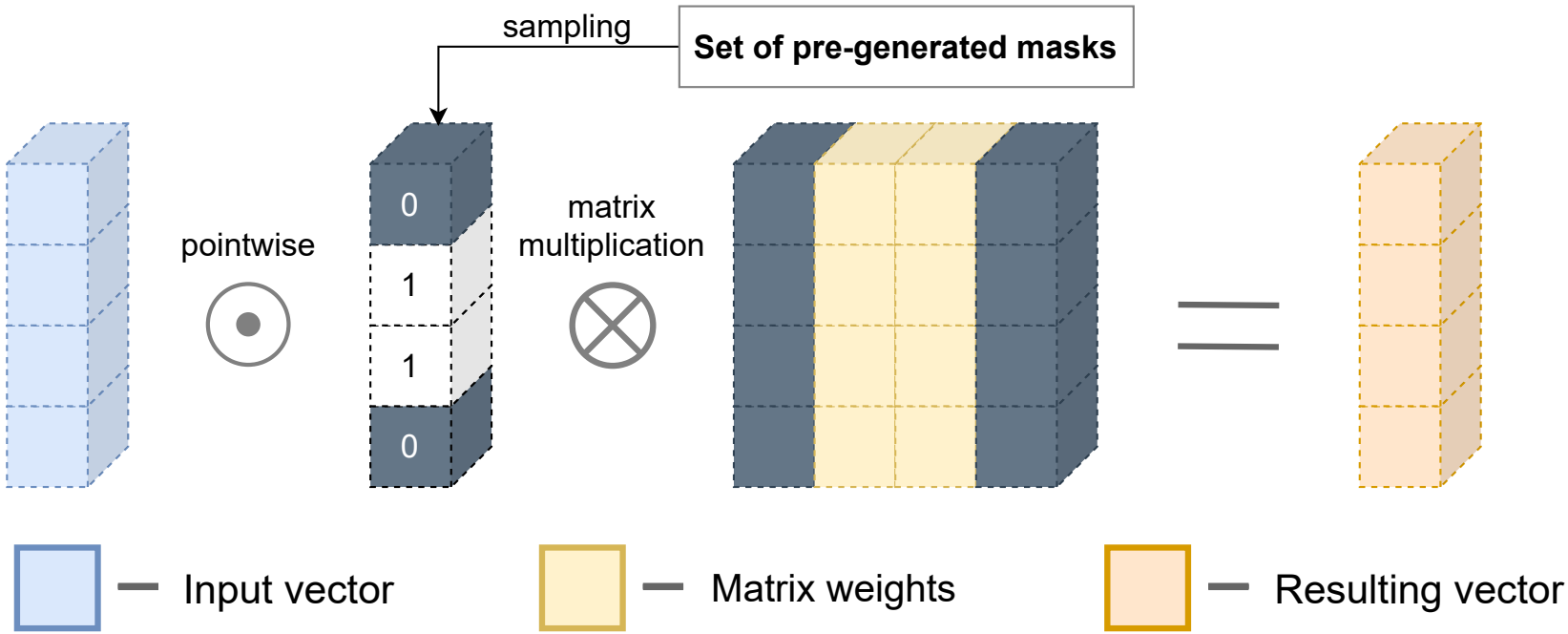


The network is fine-tuned on

- the synthetic images,
- the real images and automatically generated annotations.

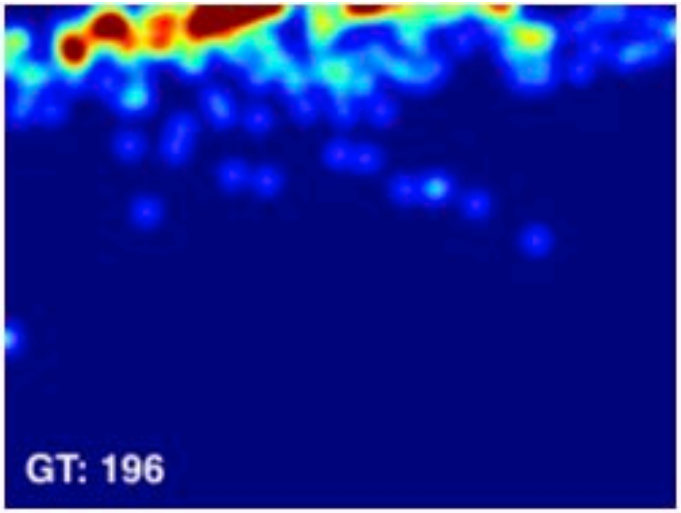
➡ It leans without manual annotations.

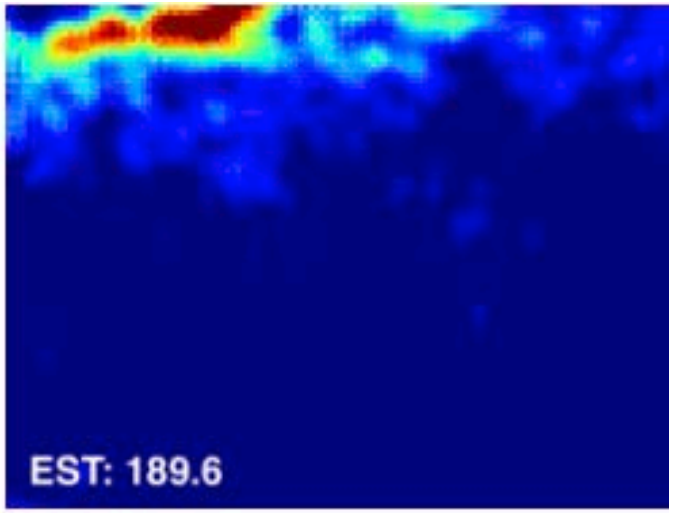
Estimating Uncertainty

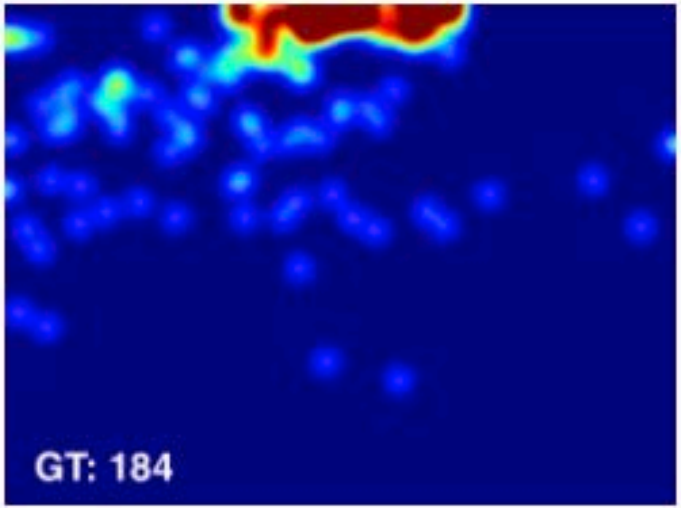


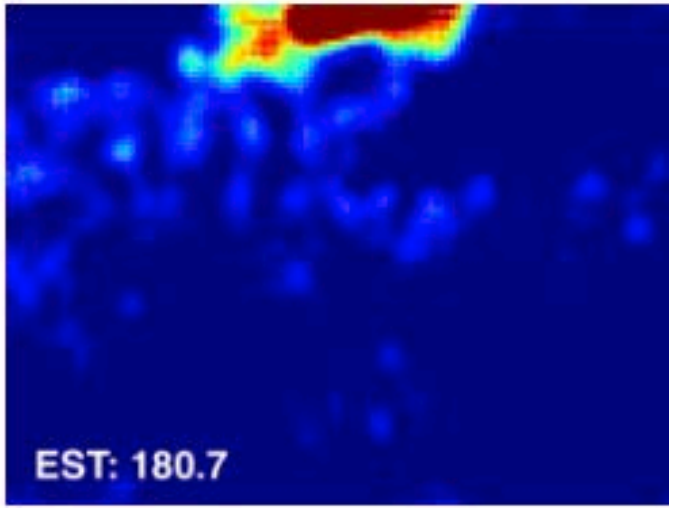
- During training, for every input vector, a binary mask is selected from a set of pre-generated masks and is used to zero out a corresponding set of features.
- By doing this several tines, one can assign a variance to the predictions.

Density Maps

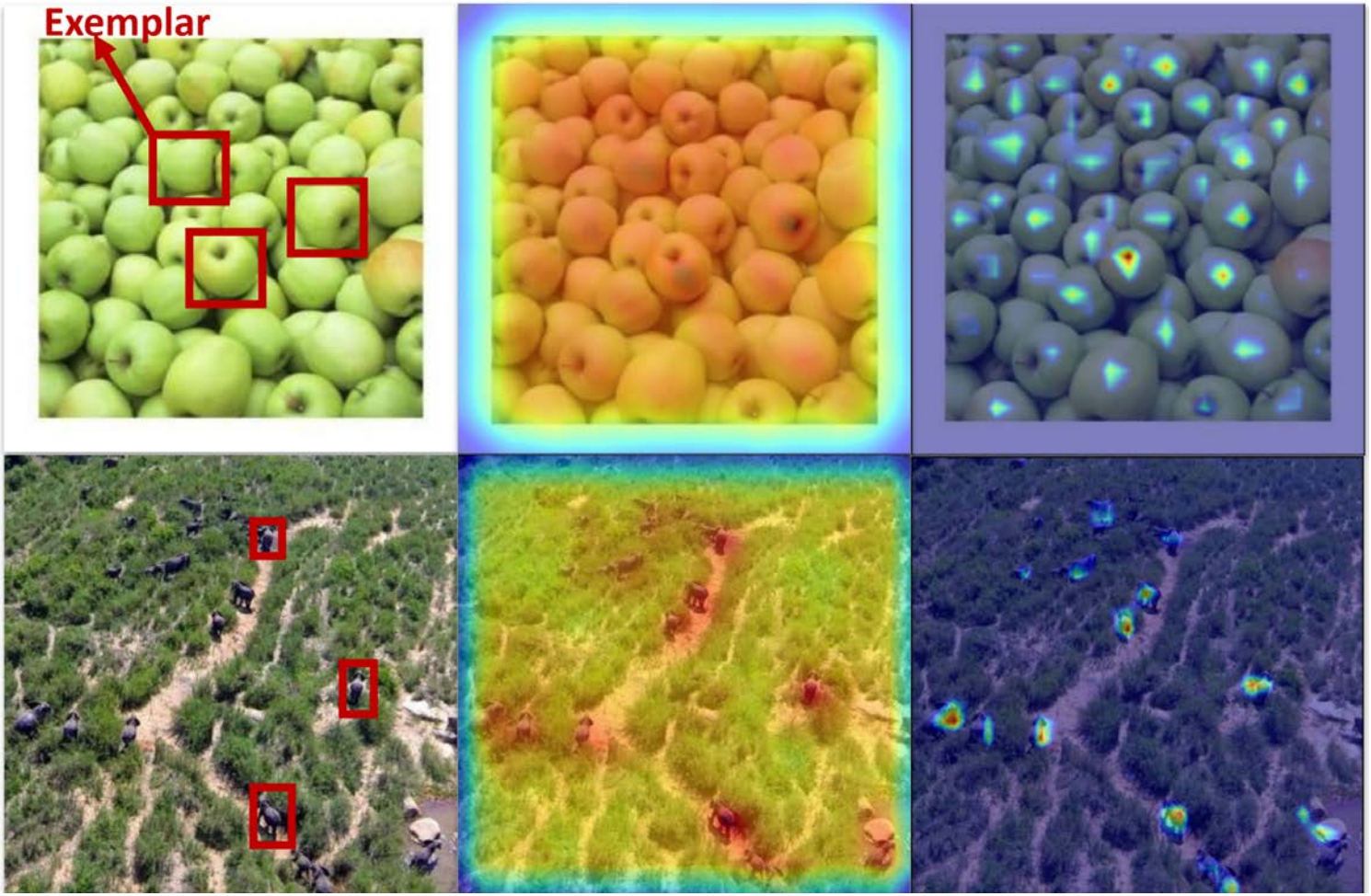








Other Kinds of Objects



Use an exemplar to make the counting class agnostic.

What about Stacked Objects?



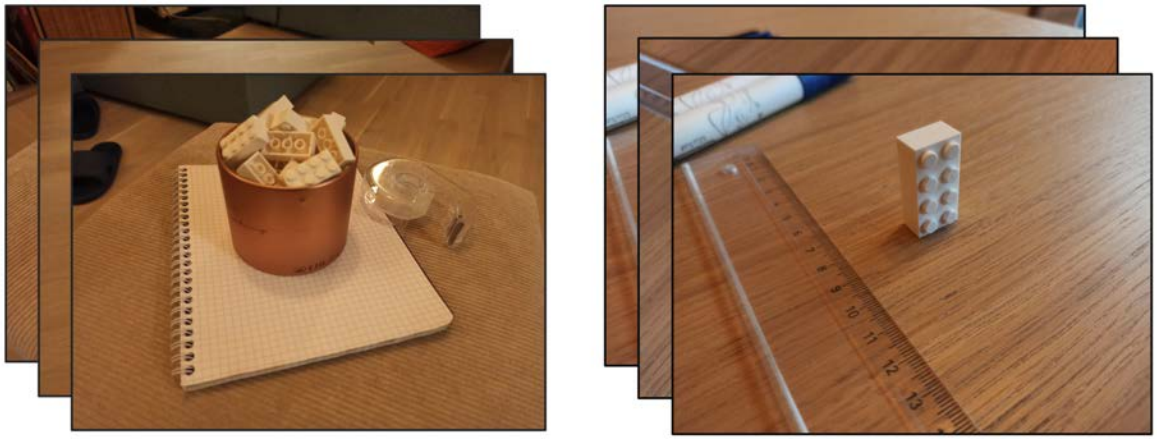


- We assumed that the target objects were not on top of each other.
- Here this is clearly not true anymore.

➡ Many of the objects to be counted are not even visible!

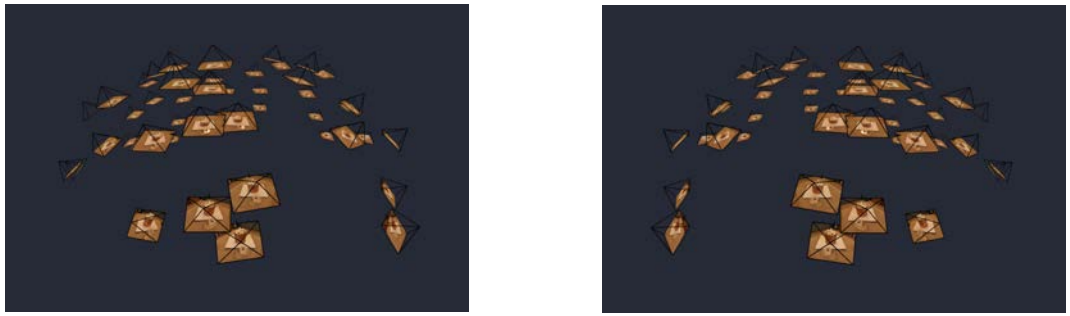
Problem Statement

Images



+

Calibrated cameras



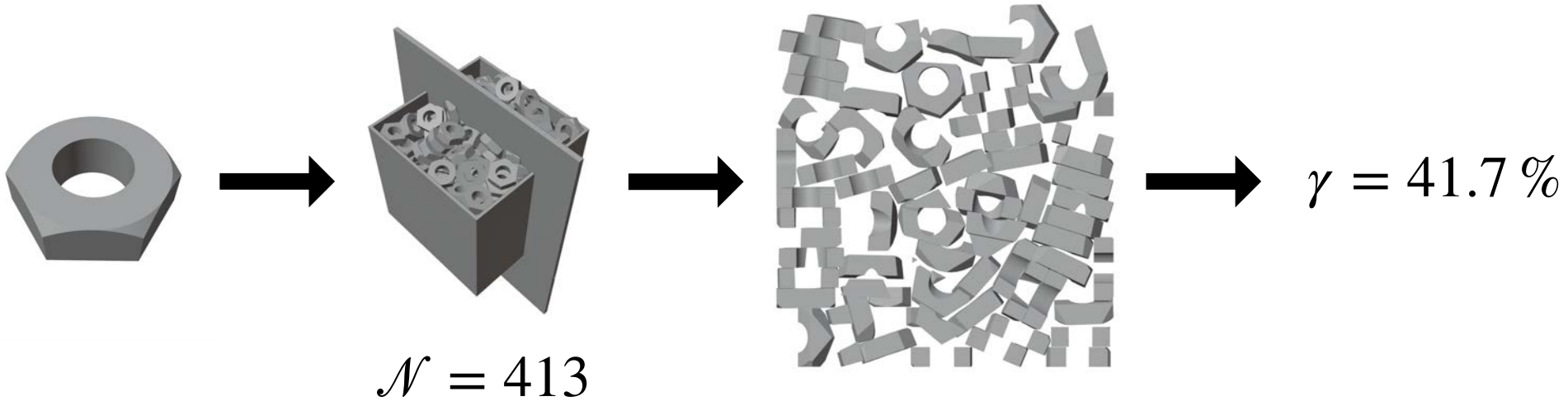
$\rightarrow \mathcal{N}$

Number or identical objects, including the occluded ones

Volume Occupancy

If we knew the volume of the stack V and of one single object V_C, we could get a rough count by computing:

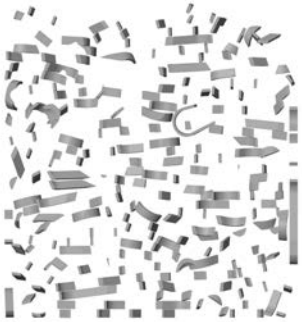
$$\mathcal{N} = \frac{V_C}{V}$$



$$\mathcal{N} = \gamma \frac{V_C}{V}$$

➡ γ is the volume occupancy we need to estimate, along with the total volume V.

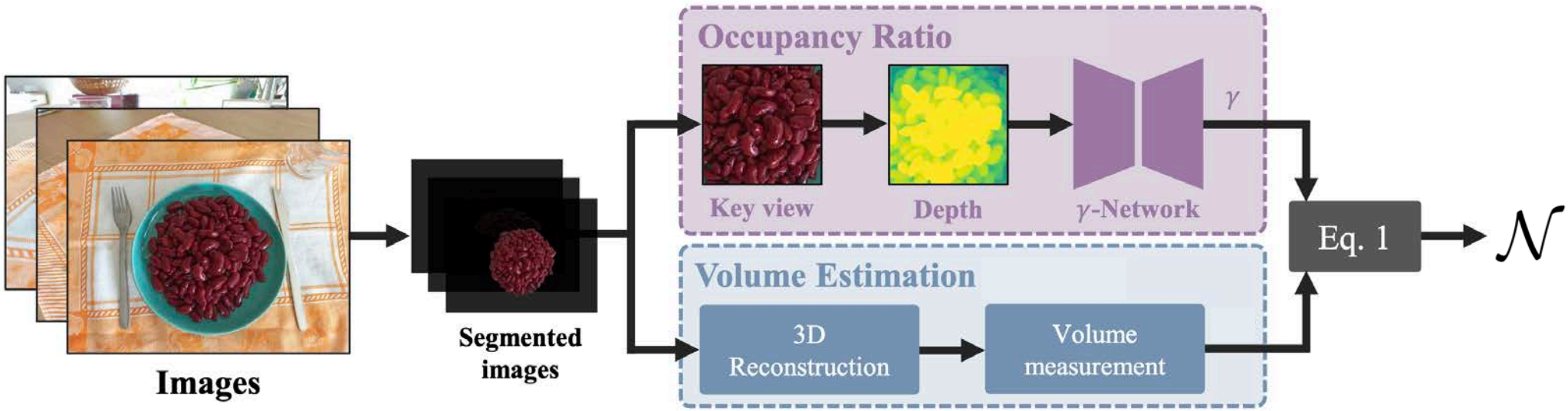
Occupancy Examples

Object	Container	Slice	Occupancy	Image
			$\gamma = 8.1\%$	
			$\gamma = 41.7\%$	
			$\gamma = 58.1\%$	

Pipeline

To estimate $\mathcal{N} = \gamma \frac{V_C}{V}$, we need:

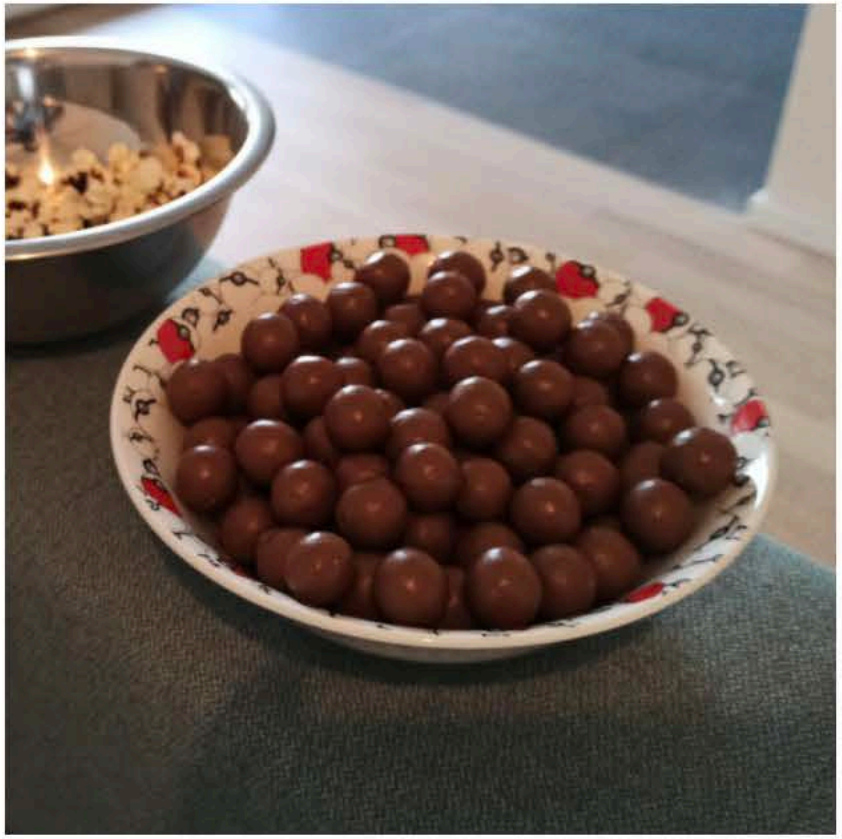
I. An occupancy estimation network
II. A volume estimation method
III. Large-scale labeled data for training purposes



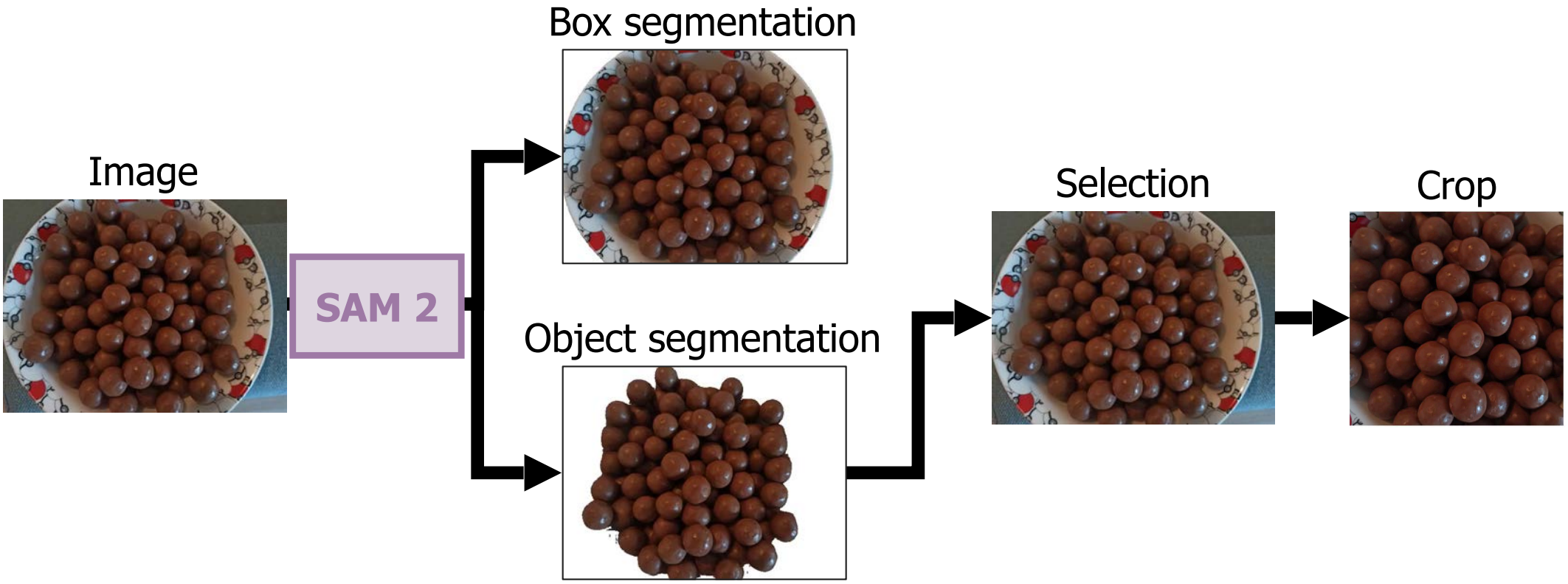
Learning to Estimate Occupancy

Synthetic training data

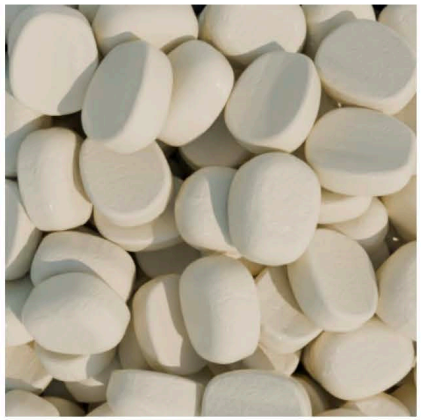
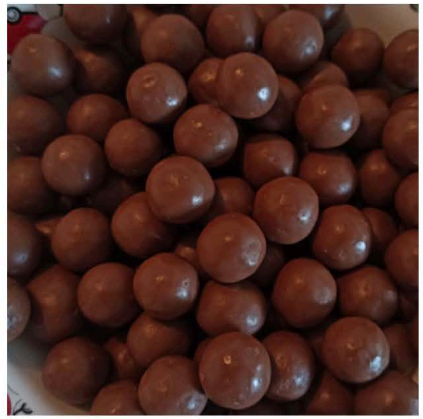
Real annotated data



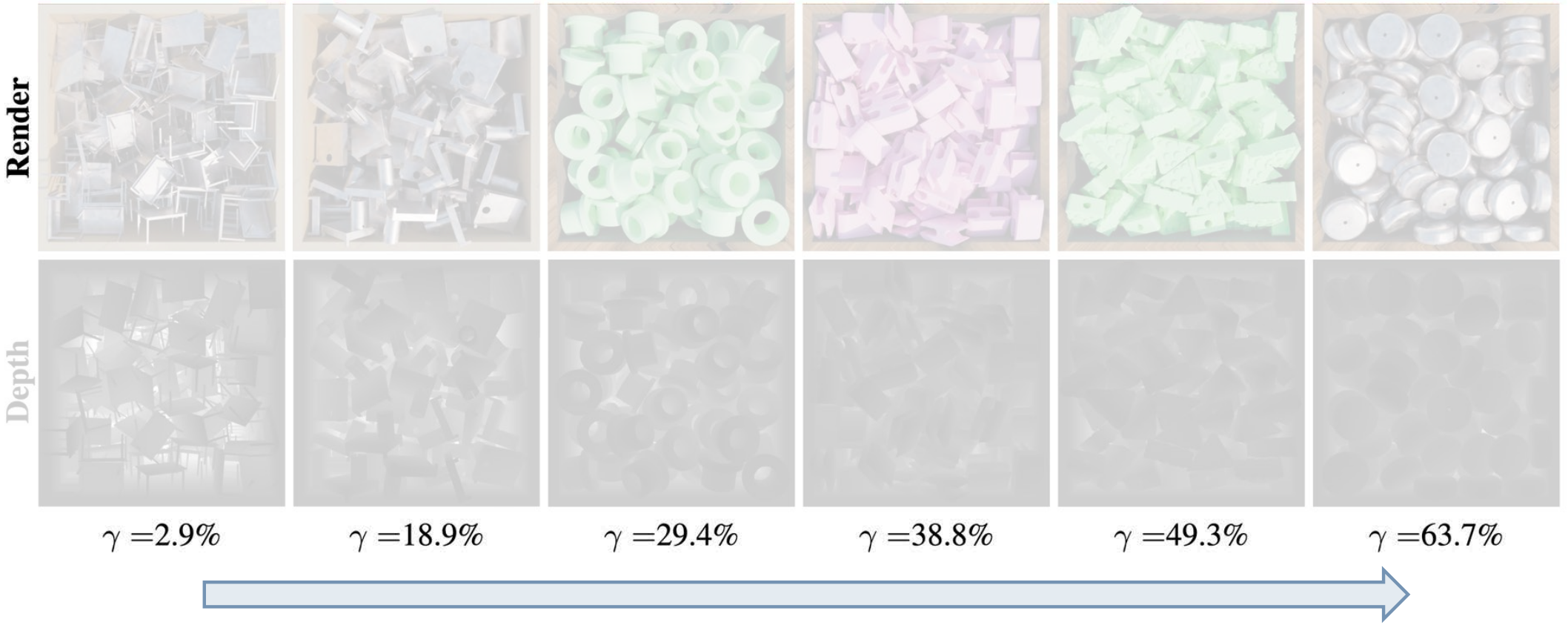
Selecting a Part of the Image



Volume Occupancy Pipeline (1)

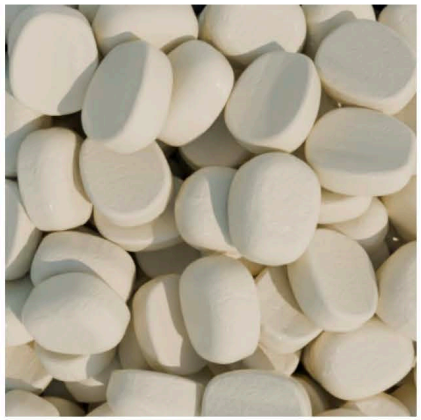
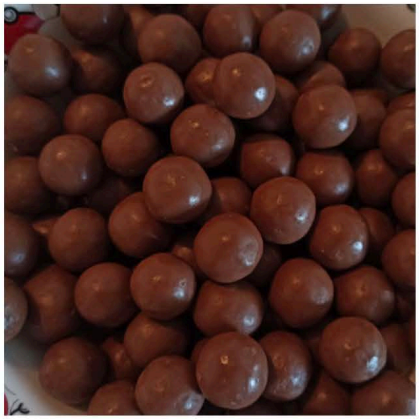
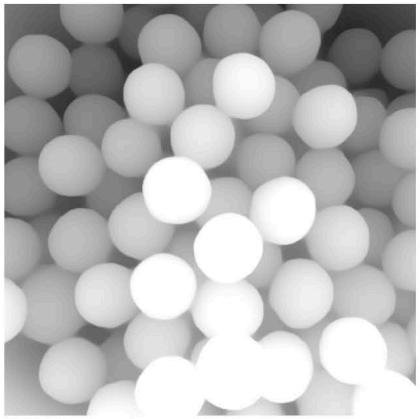
	Images	Selected crop
Synthetic data		
Real data		

RGB vs Depth



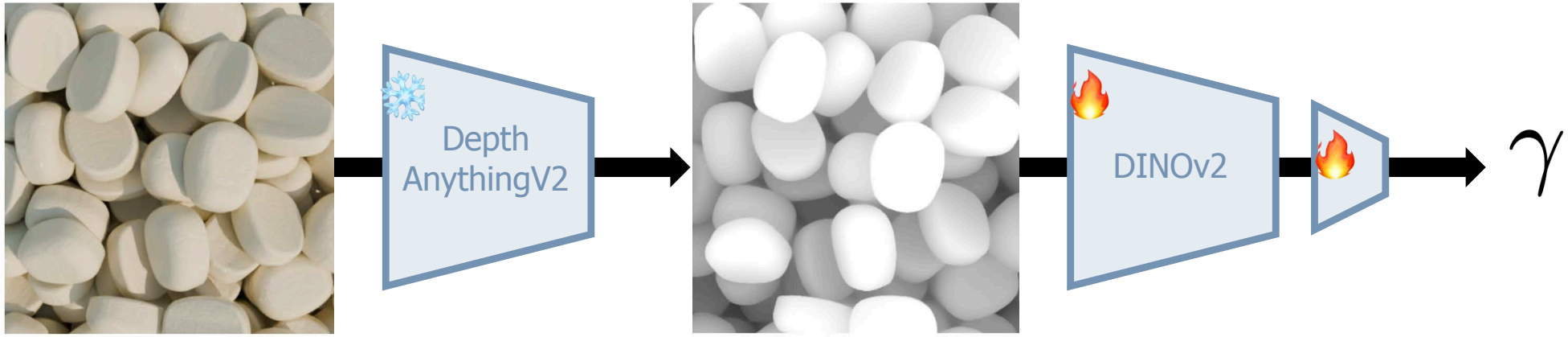
- Depth is relatively invariant to surface material.
- More useful than raw RGB to estimate depth.

Volume Occupancy Pipeline (2)

	Images	Selected crop	Depth
Synthetic data			
Real data			

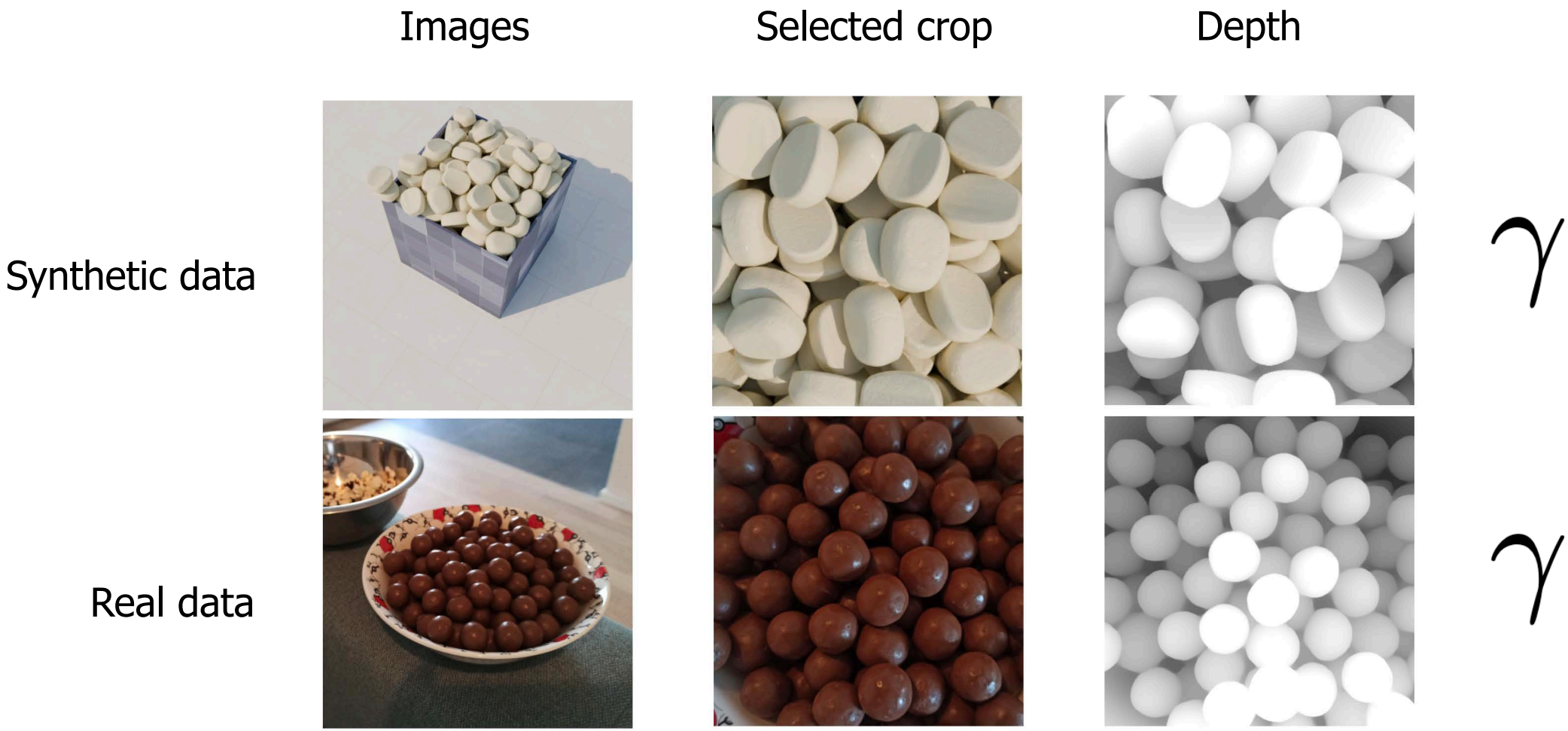
Depth Anything V2. *Yang et al*. NeurIPS'24

Volume Occupancy Network



- Fine-tune DINOv2 with a regression head to predict volume occupancy
- Even if we fine-tune only on synthetic data, DINOv2 is pre-trained on real data which further helps with generalization.

Volume Occupancy Pipeline (3)



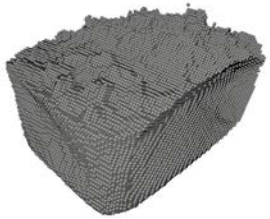
Global Volume Estimation

Images from calibrated cameras

Segments

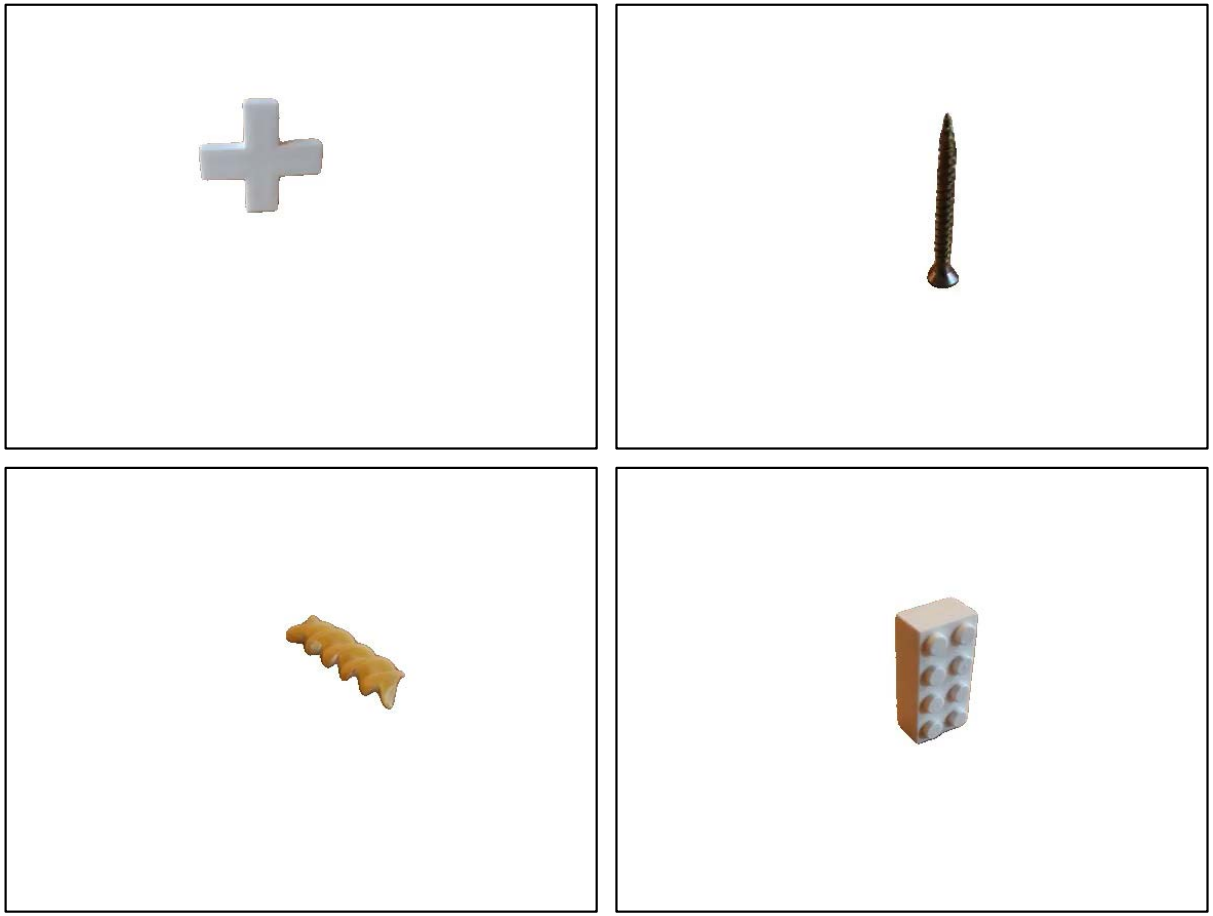
3DGS

Voxels



Local Volume Estimation

For new objects, perform volume estimation from a separate set of images to get V_C.



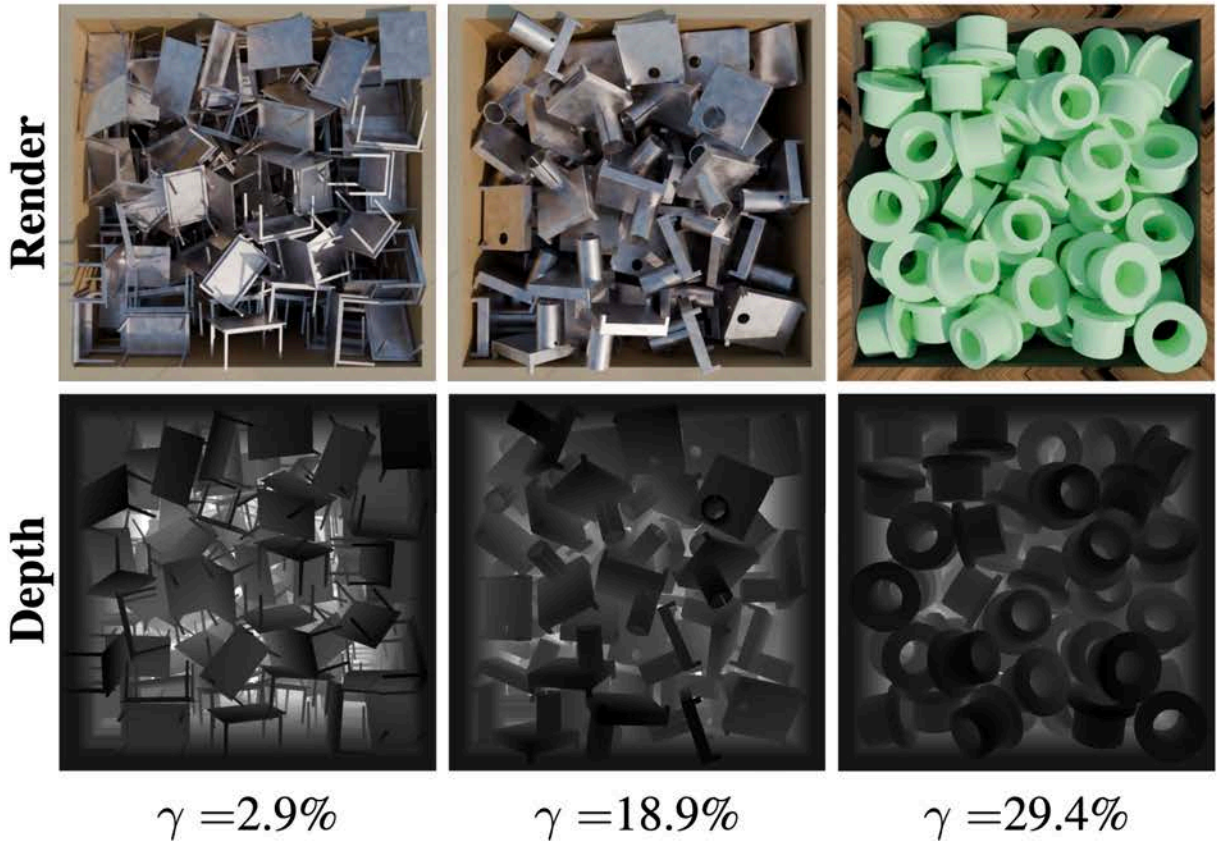
Creating a Training Dataset (1)

- Physically simulate 13000 scenes by dropping objects from ABC into parameterized containers

- Multi-view Physically-Based-Rendering

ABC: Matveev et al., CVPR'19

Creating a Training Dataset (2)



In the synthetic data, γ, V, and V_C are known !

Results

?

?

Comparison

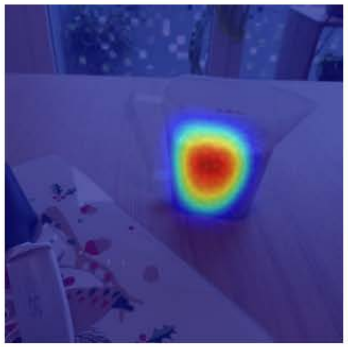
	NAE ↓	SRE ↓	MAE ↓	sMAPE ↓
BMNet+ [21]	0.93	0.98	966.76	131.44
SAM+CLIP [7, 16]	0.94	0.99	980.33	124.31
CNN	0.95	0.93	992.06	97.09
ViT+H	0.94	0.93	979.29	91.45
Human	0.79	0.84	823.23	76.85
Human-Vote	0.60	0.30	621.46	57.91
LlamaV 3.2	1.00	1.00	1037.5	190.48
Ours (Color)	0.57	0.27	607.98	74.33
Ours	**0.36**	**0.06**	**382.59**	**53.31**

- We outperform humans.
- We outperform a 2025 LLVM.
- Estimating from depth works better than from color.

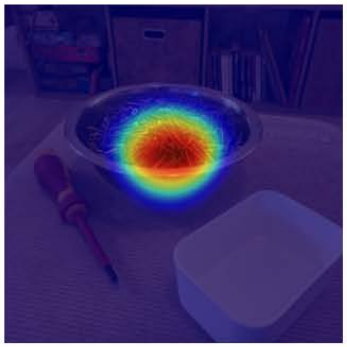
Future Work

- The current version assumes that the density is constant within the volume?
- Can we relax this assumption?


$N = 912$


$N_{\text{pred}} = 858$

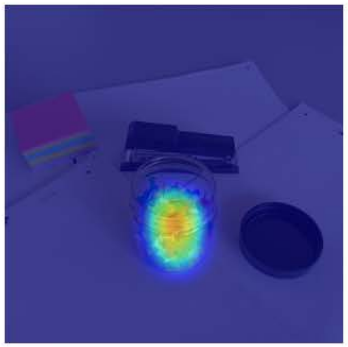

$N = 511$


$N_{\text{pred}} = 528$

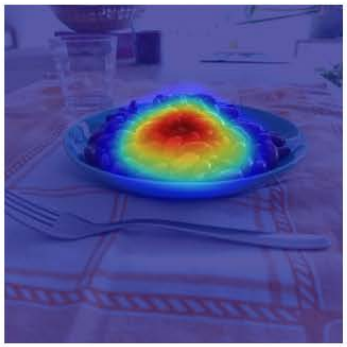

$N = 19$


$N_{\text{pred}} = 22$

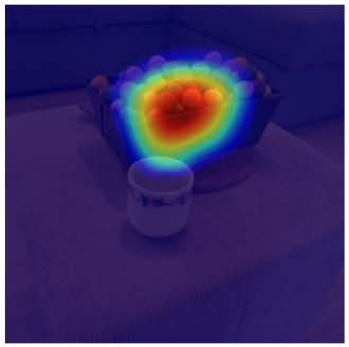

$N = 205$


$N_{\text{pred}} = 201$


$N = 207$


$N_{\text{pred}} = 173$


$N = 202$


$N_{\text{pred}} = 171$

Counting in Short

- Neural nets are key to regressing from images to numbers.
- However, older techniques are also required to deploy them effectively.
- Generating the required training data is a key part of implementing a working solution.