3D Mesh Reconstruction of Foods from a Single Image

Shu Naritomi, Keiji Yanai
The University of Electro-Communications, Tokyo, Japan

Our recent work

- **Hungry Networks**: 3D Mesh Reconstruction of a Dish and a Plate from a Single Dish Image for Estimating Food Volume. [1]
  - ACM Multimedia Asia 2020.

- **Pop'n Food**: 3D Food Model Estimation System from a Single Image. [2]
  - IEEE MIPR 2021

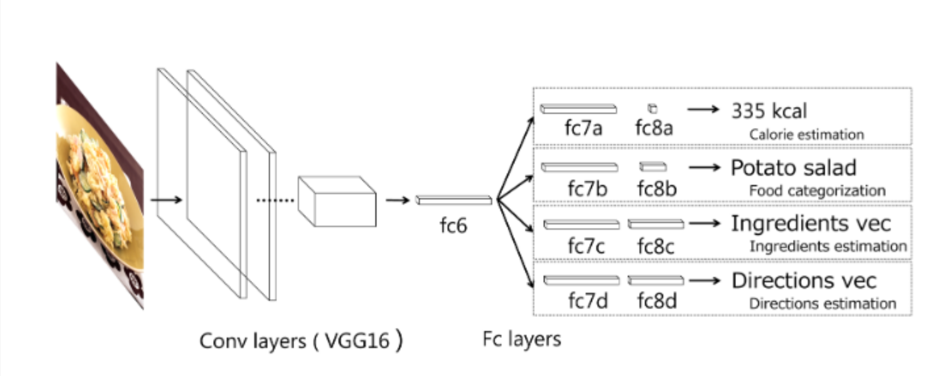
[1] S. Naritomi, and K. Yanai. **Hungry Networks**: 3D Mesh Reconstruction of a Dish and a Plate from a Single Dish Image for Estimating Food Volume. In Proc. of ACM Multimedia Asia 2020.

[2] S. Naritomi, and K. Yanai. **Pop'n Food**: 3D Food Model Estimation System from a Single Image.
In Proc. of IEEE 4th International Conference on Multimedia Information Processing and Retrieval 2021

Introduction

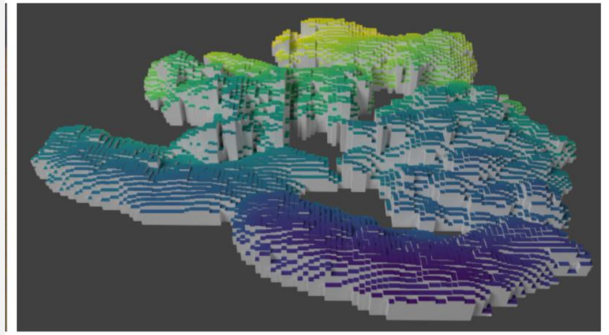
- Dietary calorie management has been an important topic.
- There is a lot of research on calorie estimation in the multimedia community.

2D based



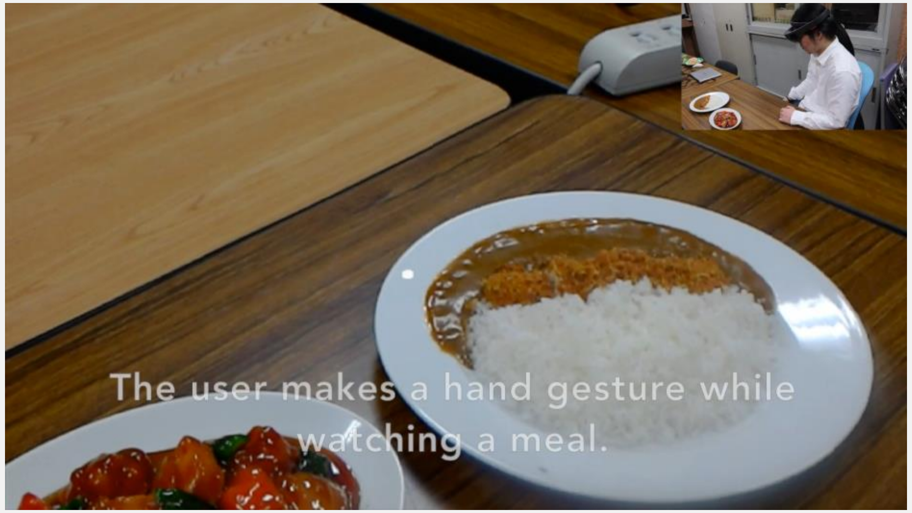
[Ege et al., IEICE2018]

Depth based



[Im2Calories, ICCV 2015]

Sensor based

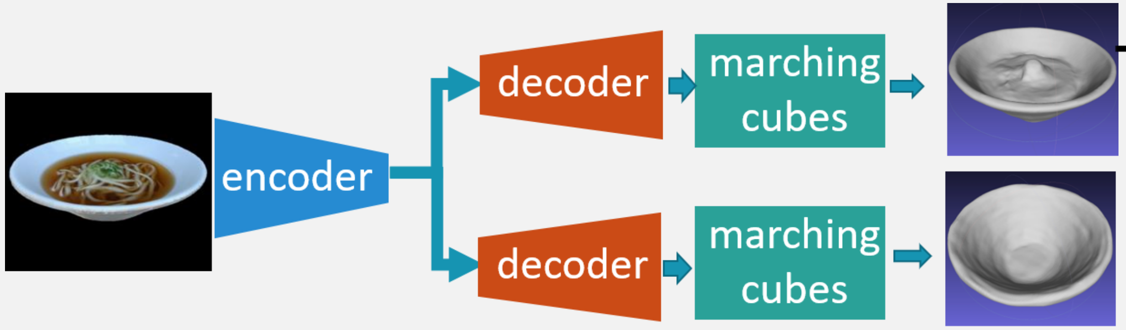


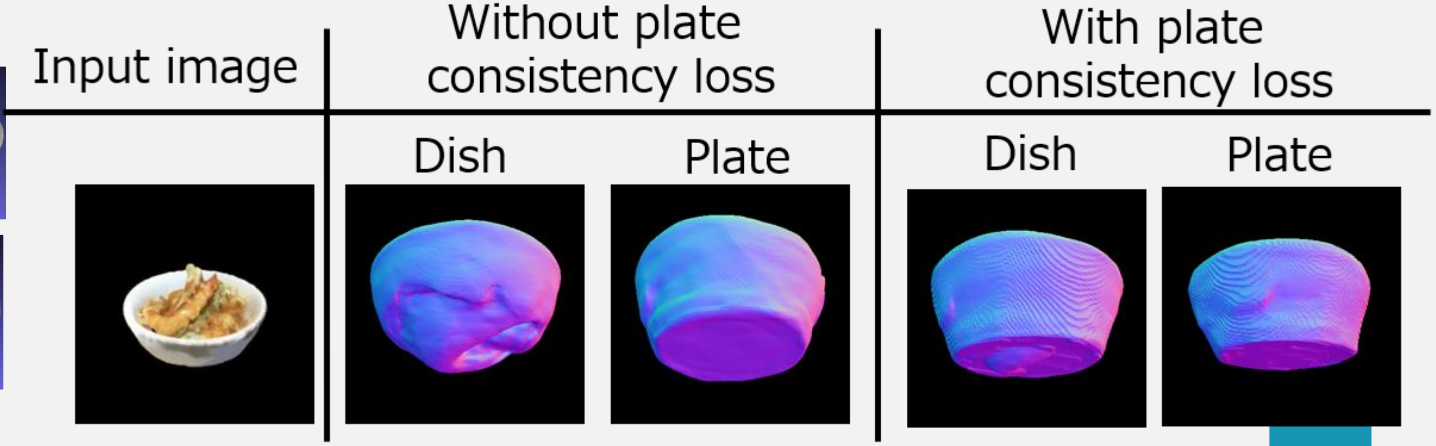
[CalorieCaptorGlass, IEEE VR 2020]

Introduction

- Reconstruct 3D dish (food + plate) volume and 3D plate volume from a single dish image

- Achieve consistency between the plate part of the two reconstructed volumes introducing plate consistency loss.

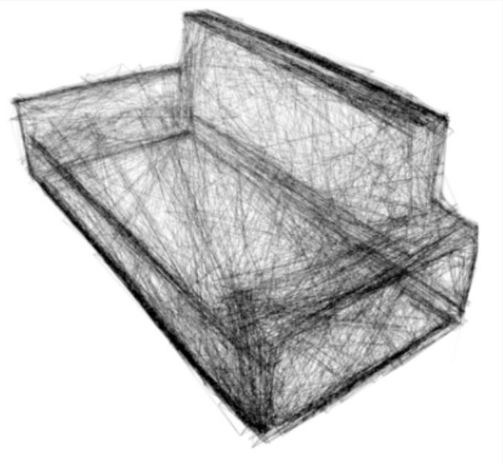




Appropriate 3D representation

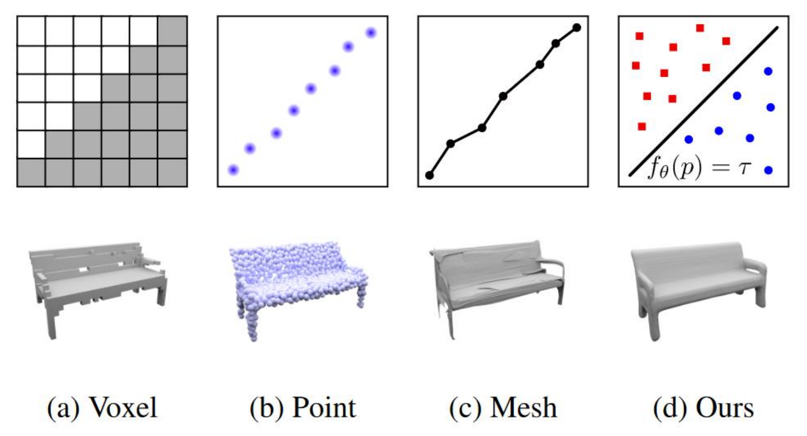
- we want to estimate the food volume.
  - Voxel : ✕ Not suitable for high resolution
  - Point cloud: ✕ The connection between points is unknown.
  - Mesh: ○ It is easy to achieve high resolution.
    the volume can be calculated easily if the conditions are met.

- Conditions for obtaining volume from Mesh
  - Watertight
  - no self-intersection



Self-intersection [Mesh R-CNN, ICCV2019]

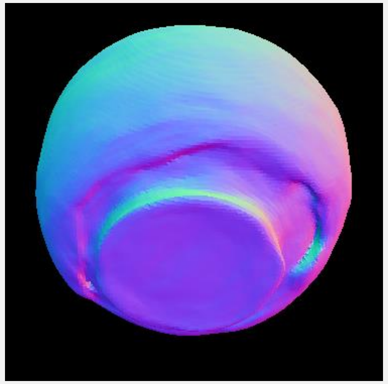
Appropriate 3D representation

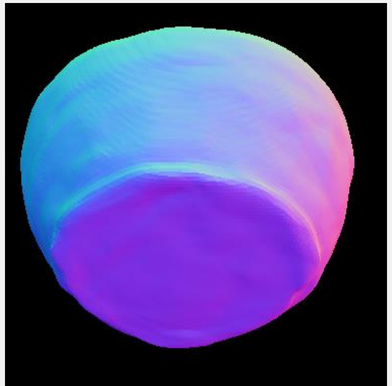


[Occupancy Networks, CVPR2019]

- Mesh Template : self-intersection occurs frequency.
- **Occupancy, SDF**: When a marching cube is used to extract the mesh, it is watertight and does not self-intersect.

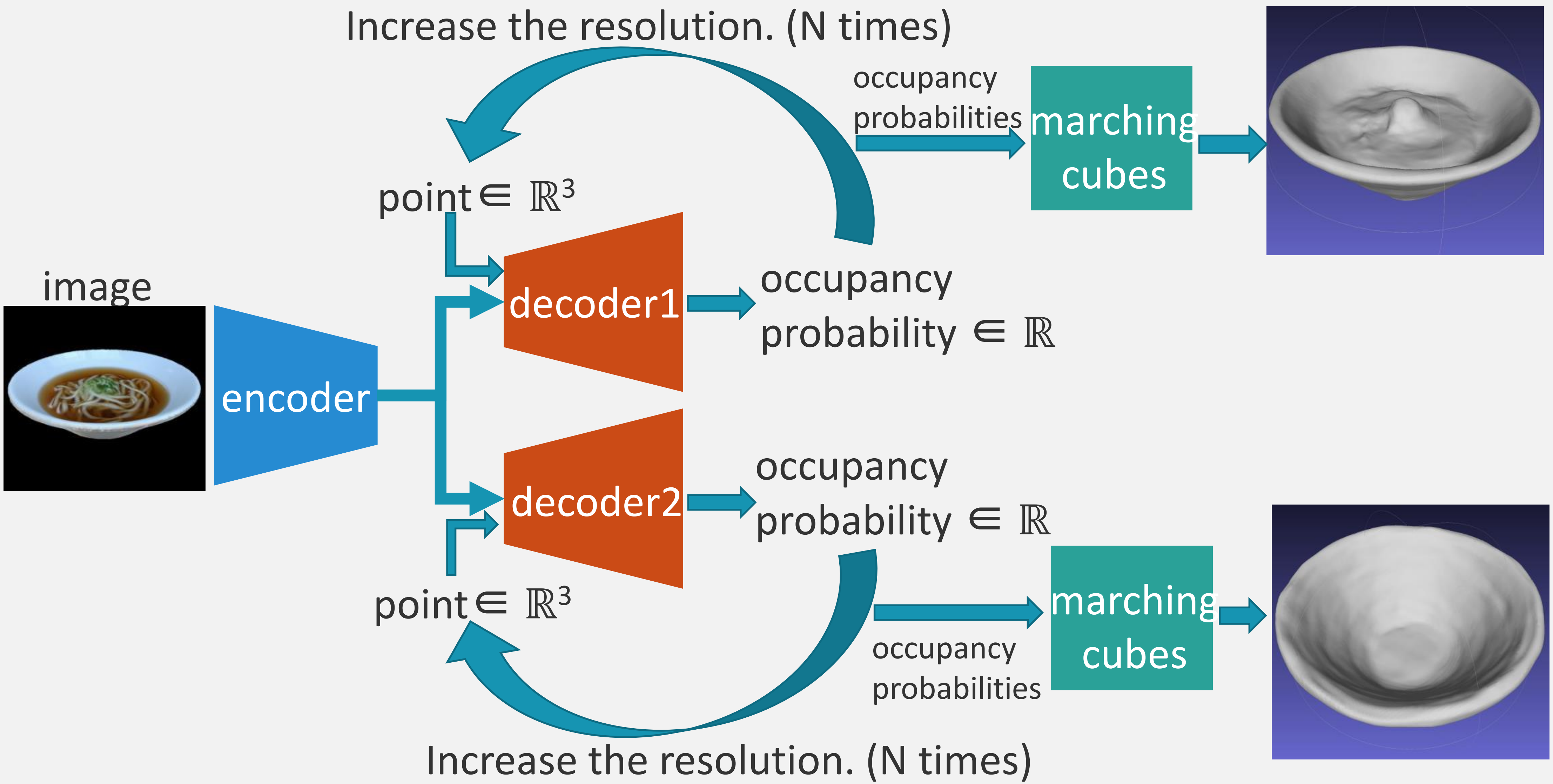
- The problem of situations where the shapes of the plate do not match.
  - The point $p \in R^3$ contained inside the plate is not contained inside the dish mesh.



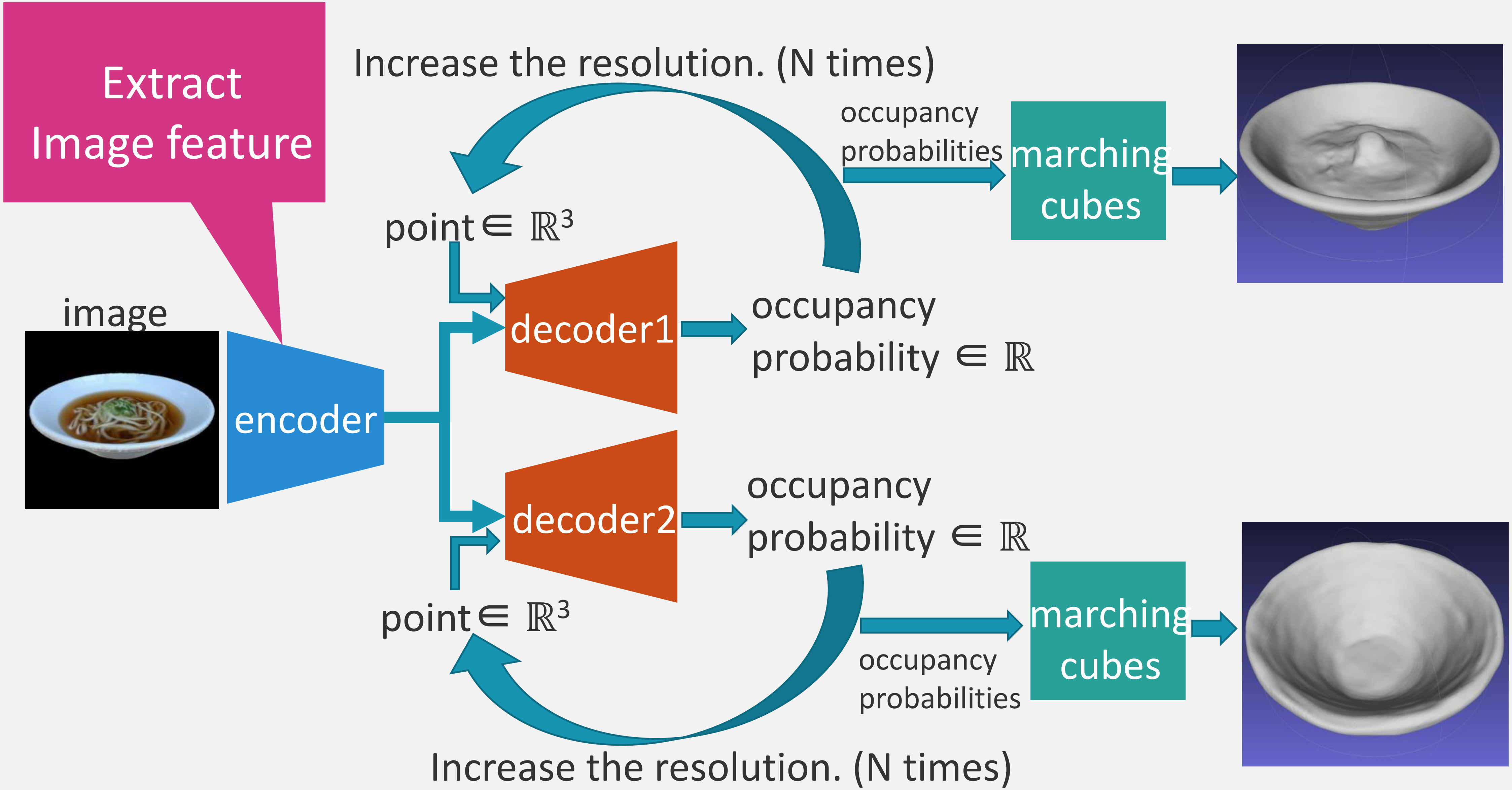


occupancy is reasonable

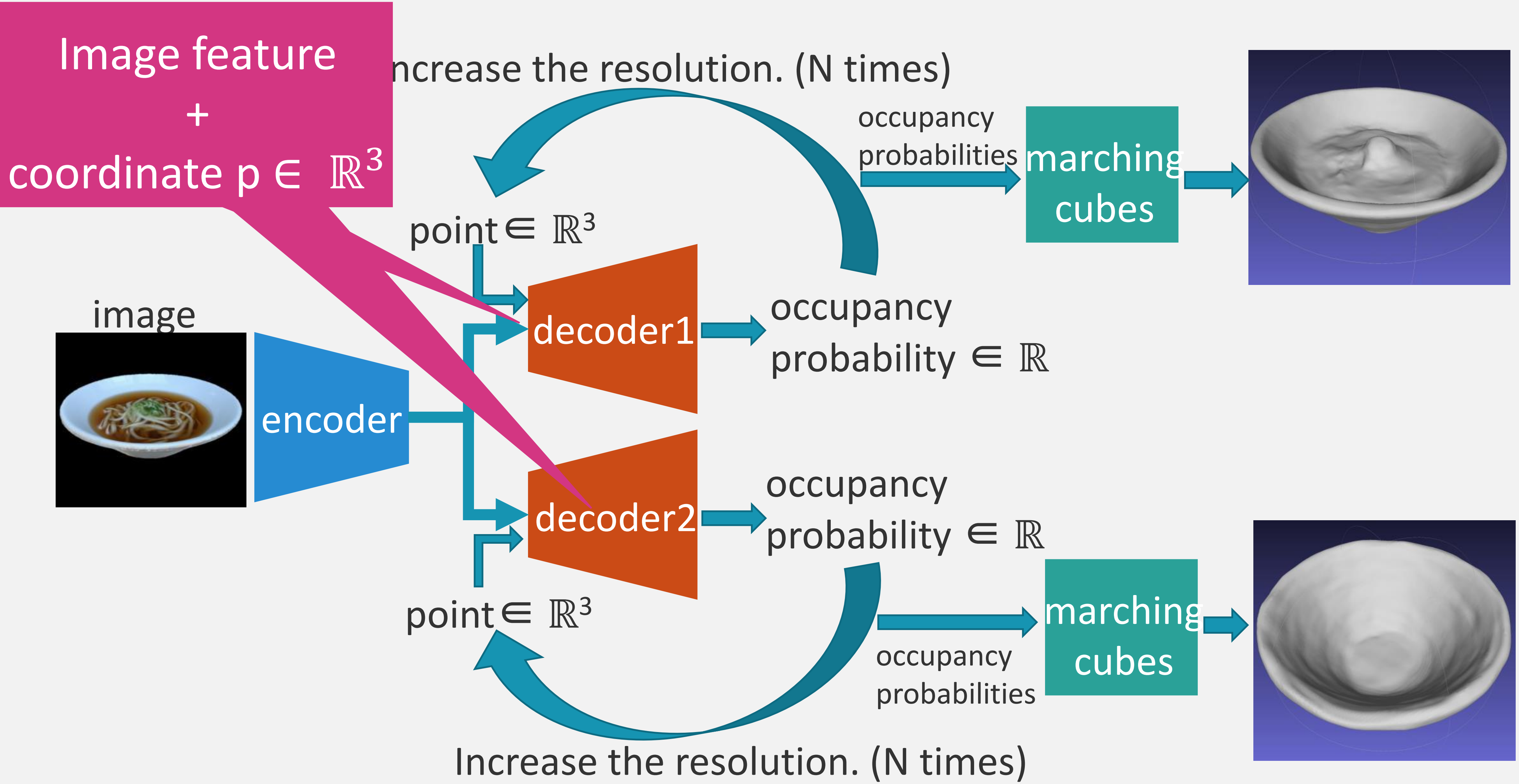
Hungry Networks : inference



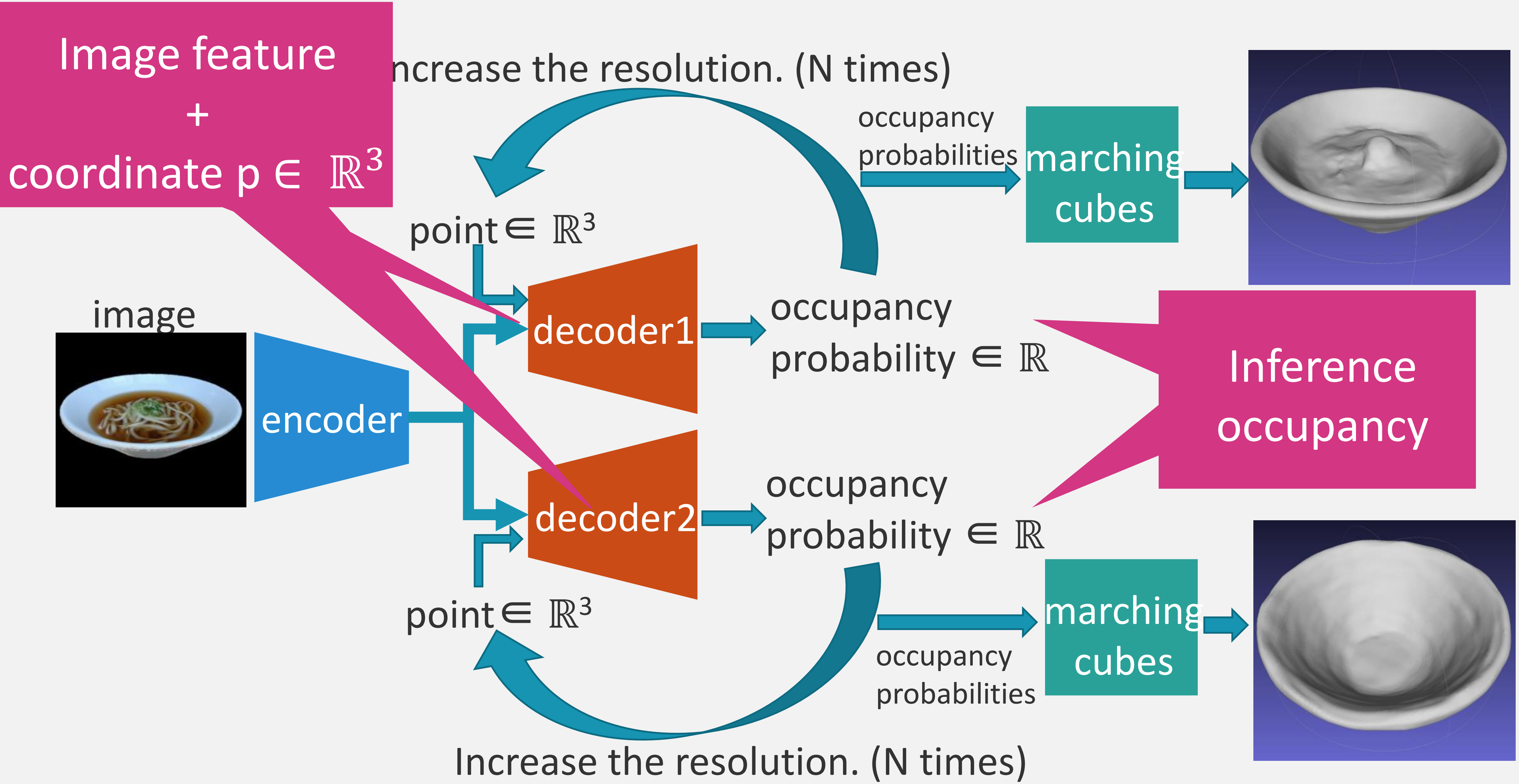
Hungry Networks : inference



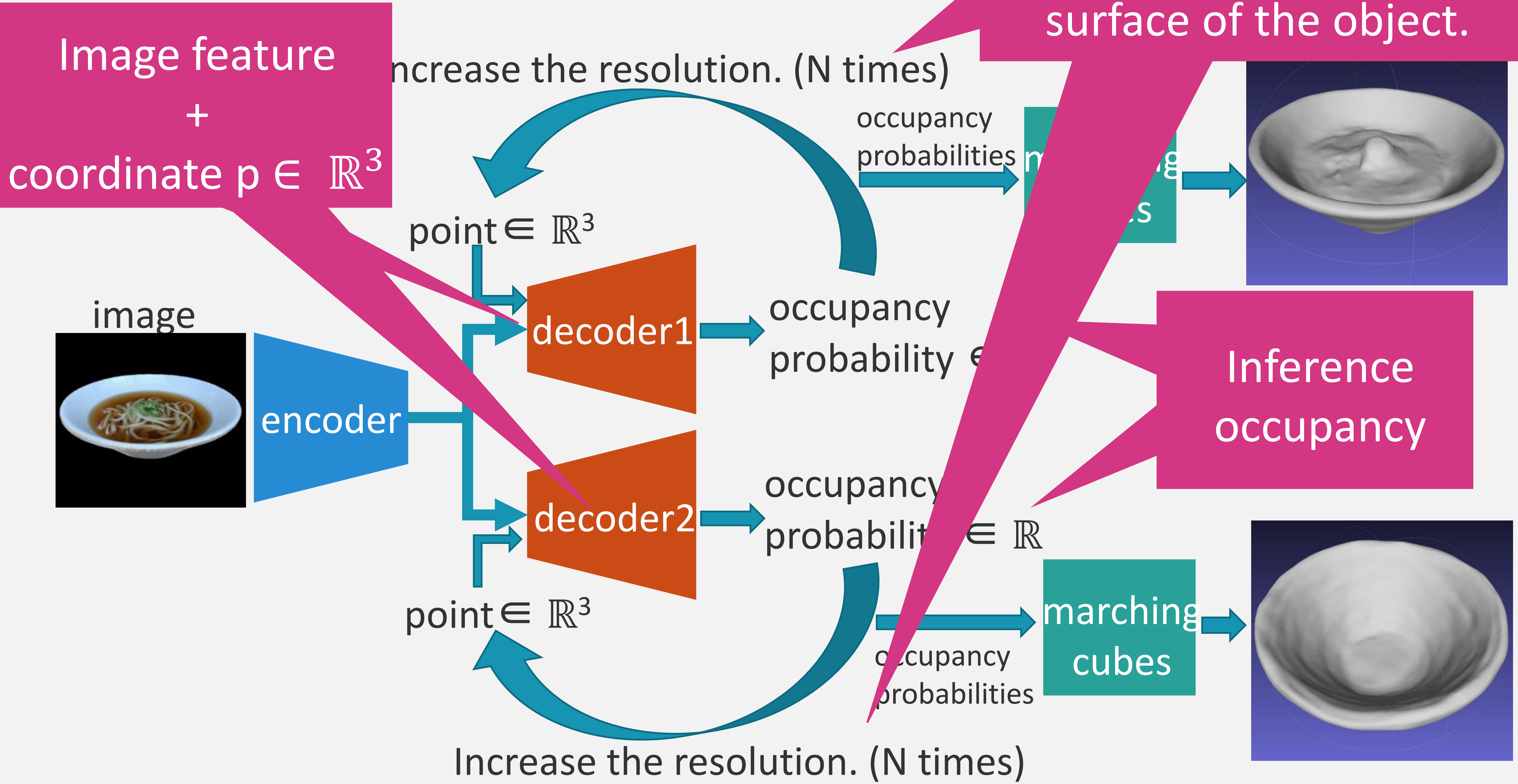
Hungry Networks : inference



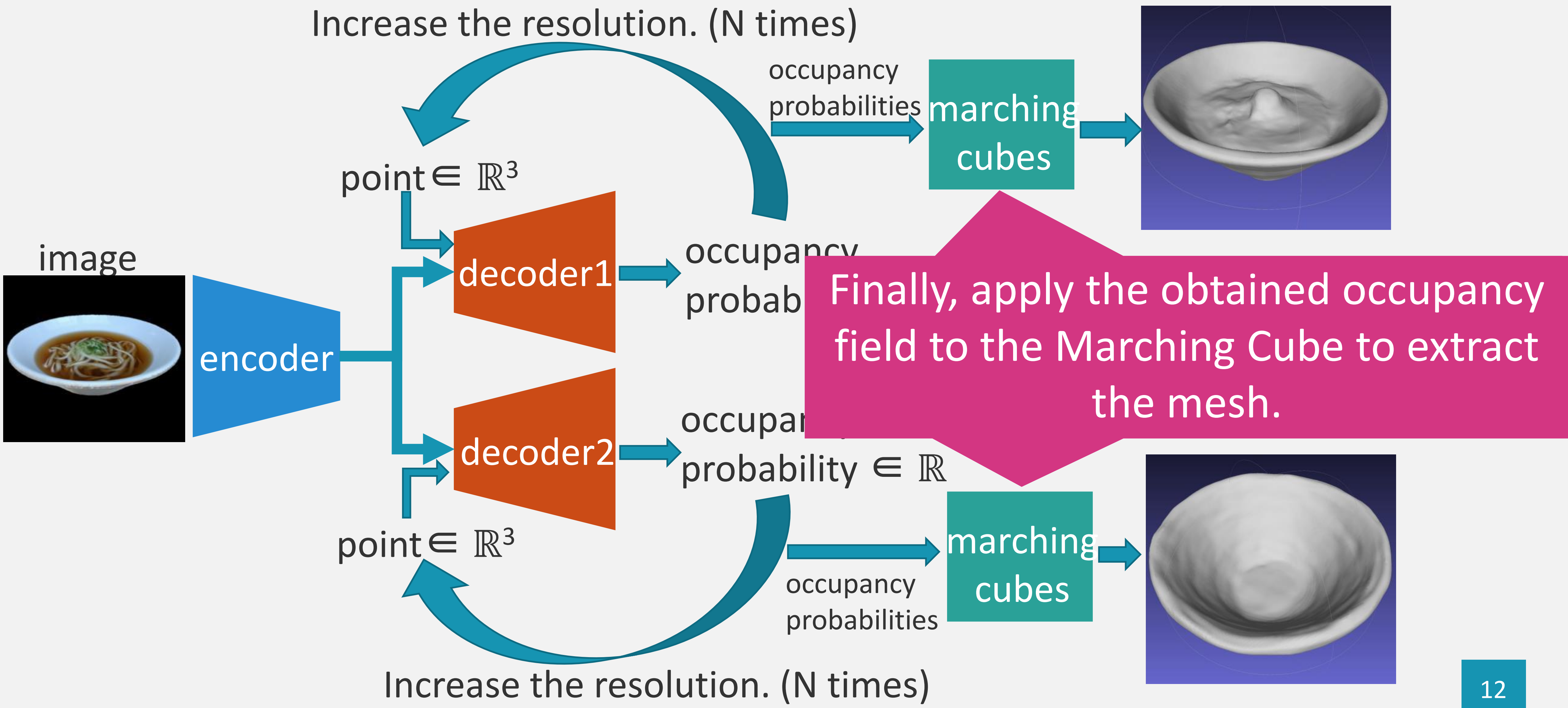
Hungry Networks : inference



Hungry Networks : inference



Hungry Networks : inference

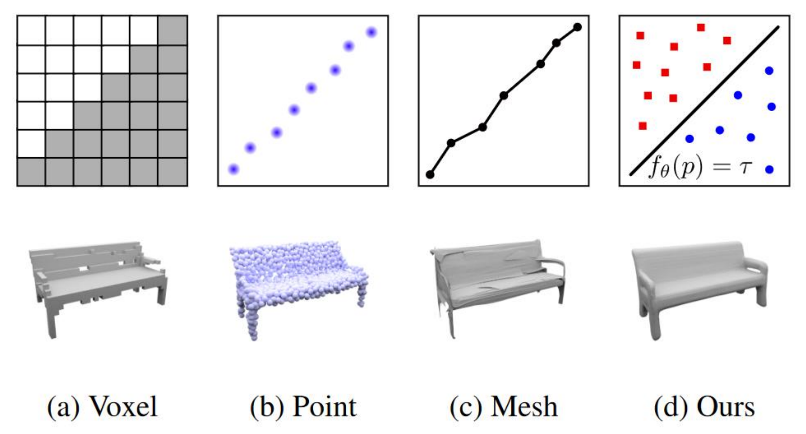


Finally, apply the obtained occupancy field to the Marching Cube to extract the mesh.

Hungry Networks : training

- Learning the occupancy is actually a binary classification.
    - Binary cross entropy loss

$$\mathcal{L}_{\mathcal{O}}(f_d(x,p),o(p)) = \mathcal{L}_{bce}(f_d(x,p),o(p))$$



[Occupancy Networks, CVPR2019]

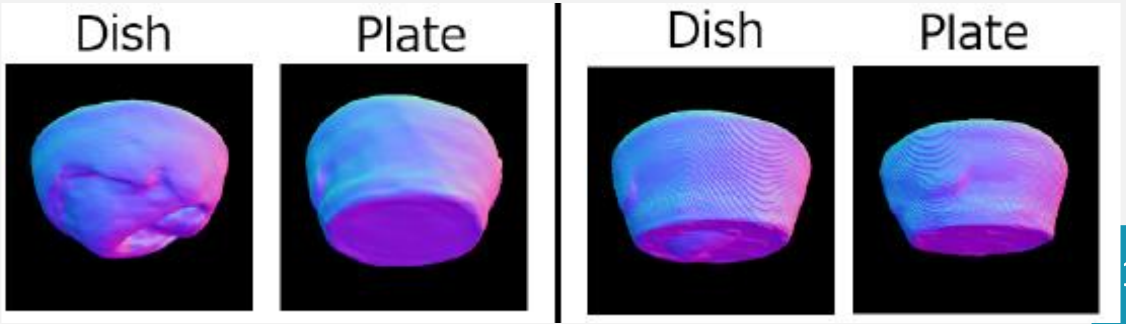
$p \in R^3$	: input point coordinate
x	: image feature vector
$o(p) \in R$	: occupancy of point p
$f_d(x,p) \in R$	: decoder that outputs occupancy

Hungry Networks : training

- **Plate consistency loss (proposal method)**
  - Loss function for matching plate parts of the 3D shape of dish and plat

Dish occupancy $f_{d1}(x,p)$	Plate occupancy $f_{d2}(x,p)$	$f_{d2}(x,p) - f_{d1}(x,p)$
0	**0**	**0**
1	**0**	**-1**
0	**1**	**1**
1	**1**	**0**

$$\mathcal{L}_C(f_{d1}(p), f_{d2}(p)) = \max(f_{d2}(p) - f_{d1}(p)\ ,0)$$



Hungry Networks : training

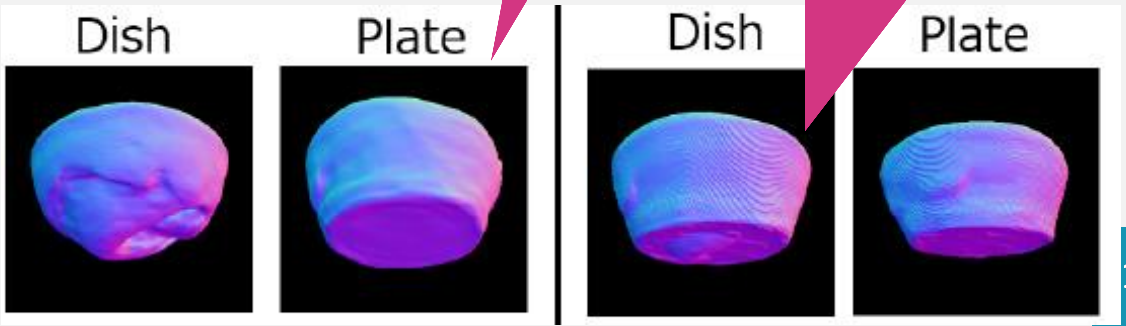
- **Plate consistency loss (proposal method)**
  - Loss function for matching plate parts of the 3D shape of dish and plat

Dish occupancy $f_{d1}(x,p)$	Plate occupancy $f_{d2}(x,p)$	$f_{d2}(x,p) - f_{d1}(x,p)$
0	**0**	**0**
1	**0**	**-1**
0	**1**	**1**
1	**1**	**0**

$$\mathcal{L}_C(f_{d1}(p), f_{d2}(p)) = \max(f_{d2}(p) - f_{d1}(p)\ ,0)$$

Without Plate consistency loss

With Plate consistency loss



Hungry Networks : training

- **Plate consistency loss (proposal method)**
  - Loss function for matching plate parts of the 3D shape of dish …at

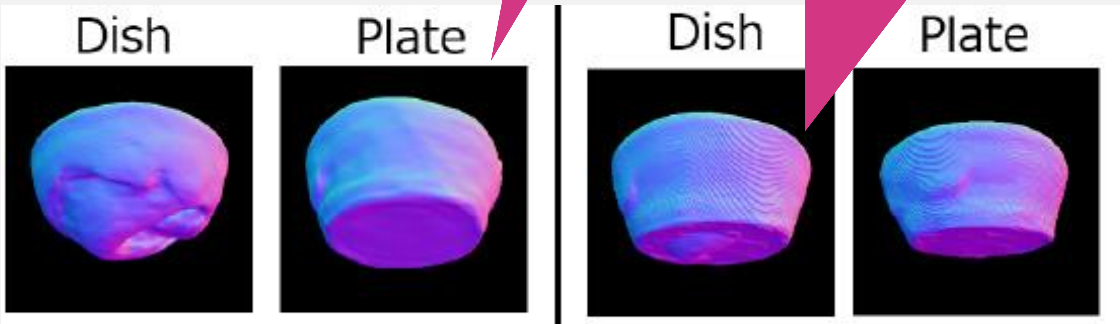
Dish occupancy $f_{d1}(x,p)$	Plate occupancy $f_{d2}(x,p)$	$f_{d2}(x,p) - f_{d1}(x,p)$
0	**0**	**0**
1	**0**	**-1**
0	**1**	**1**
1	**1**	**0**

There is a problem if the difference is **1**.

$$\mathcal{L}_C(f_{d1}(p), f_{d2}(p)) = \max(f_{d2}(p) - f_{d1}(p)\ ,0)$$

Without Plate consistency loss

With Plate consistency loss



Hungry Networks : training

- Mini batch loss

$$
\begin{aligned}
x_i &= f_e(I_i) \\
y1_{i,j} &= f_{d1}(x_i, p_{i,j}) \\
y2_{i,j} &= f_{d2}(x_i, p_{i,j})
\end{aligned}
$$

$$
\begin{aligned}
\mathcal{L}_{\mathcal{B}} = \frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \sum_{j=1}^{K} \Big(& \lambda_1 \mathcal{L}_{\mathcal{O}}(y1_{i,j}, o1_i(p_{i,j})) \\
& + \lambda_2 \mathcal{L}_{\mathcal{O}}(y2_{i,j}, o2_i(p_{i,j})) \\
& + \lambda_3 \mathcal{L}_{\mathcal{C}}(y1_{i,j}, y2_{i,j}) \Big)
\end{aligned}
$$

$f_e(I_i)$ Encoder that outputs image feature

I_i i-th image

$\mathcal{B}$ mini batch

Training dataset

- There is no dataset containing a 3D mesh of dish.
  - Build a new dataset
- 240 Dish 3D models、 38 plate 3D models.
  - Using a commercially available 3D scanner.




Training dataset

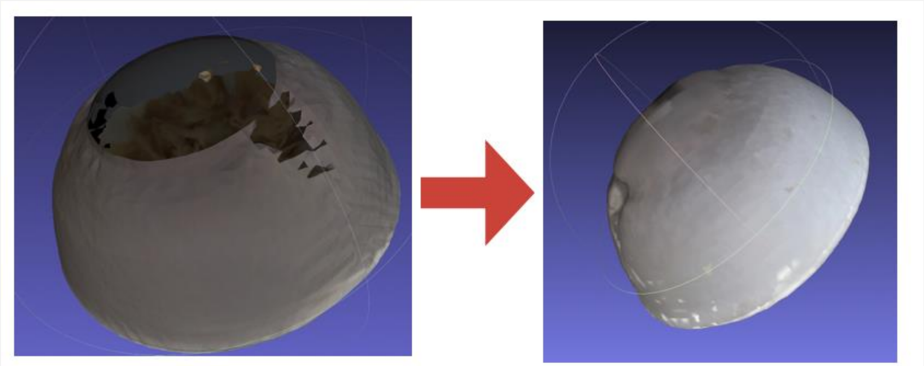
- The mesh output by the scanner cannot be learned as it is.
- problem
  - (1) The center of the model does not coincide with the origin.
  - (2) Not watertight.
  - (3) The size is not unified.
  - (4) Containing noise.
  - (5) The coordinates of the plate parts of a dish mesh and a corresponding plate mesh do not match to each other.

Training dataset

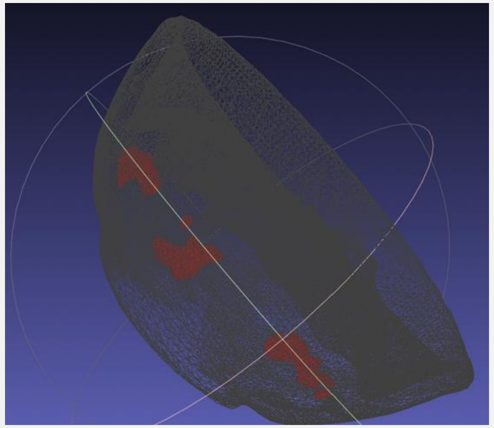
- The mesh output by the scanner cannot be learned as it is.
- problem
  - (1) The center of the model does not coincide with the origin.
  - (2) Not watertight.
  - (3) The size is not unified.
  - (4) Containing noise.
  - (5) The coordinates of the plate parts of a dish mesh and a corresponding plate mesh do not match to each other.

Training dataset : modify scanned mesh data.

- (2) Not watertight.
  - The 3D model taken by the scanner lacks the surface that was in contact with the floor.
  - Apply Poisson Surface Reconstruction

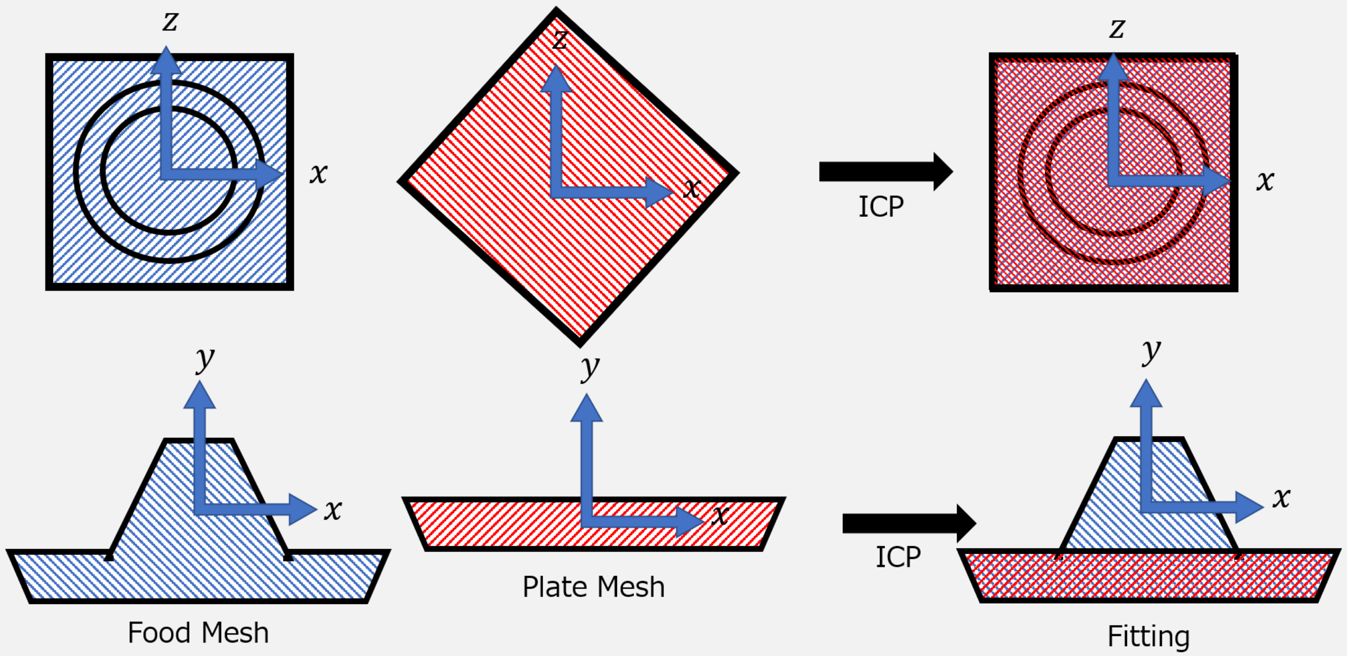


- (4) Contains noise
  - Eliminate using TSDF Fusion



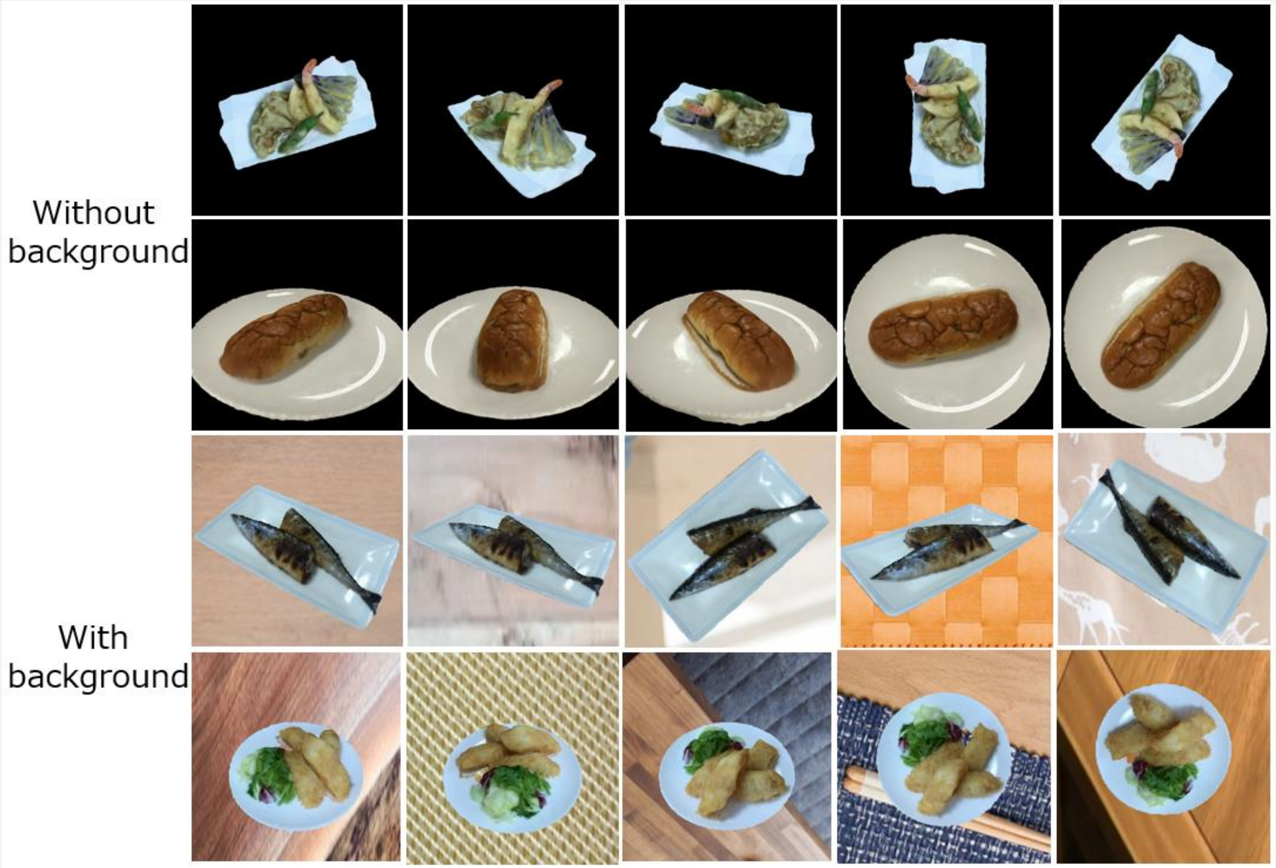
Training dataset : modify scanned mesh data.

- (5) The coordinates of the plate parts of a dish mesh and a corresponding plate mesh do not match to each other.
  - Align dish and plate meshes using ICP (Iterative closest point)



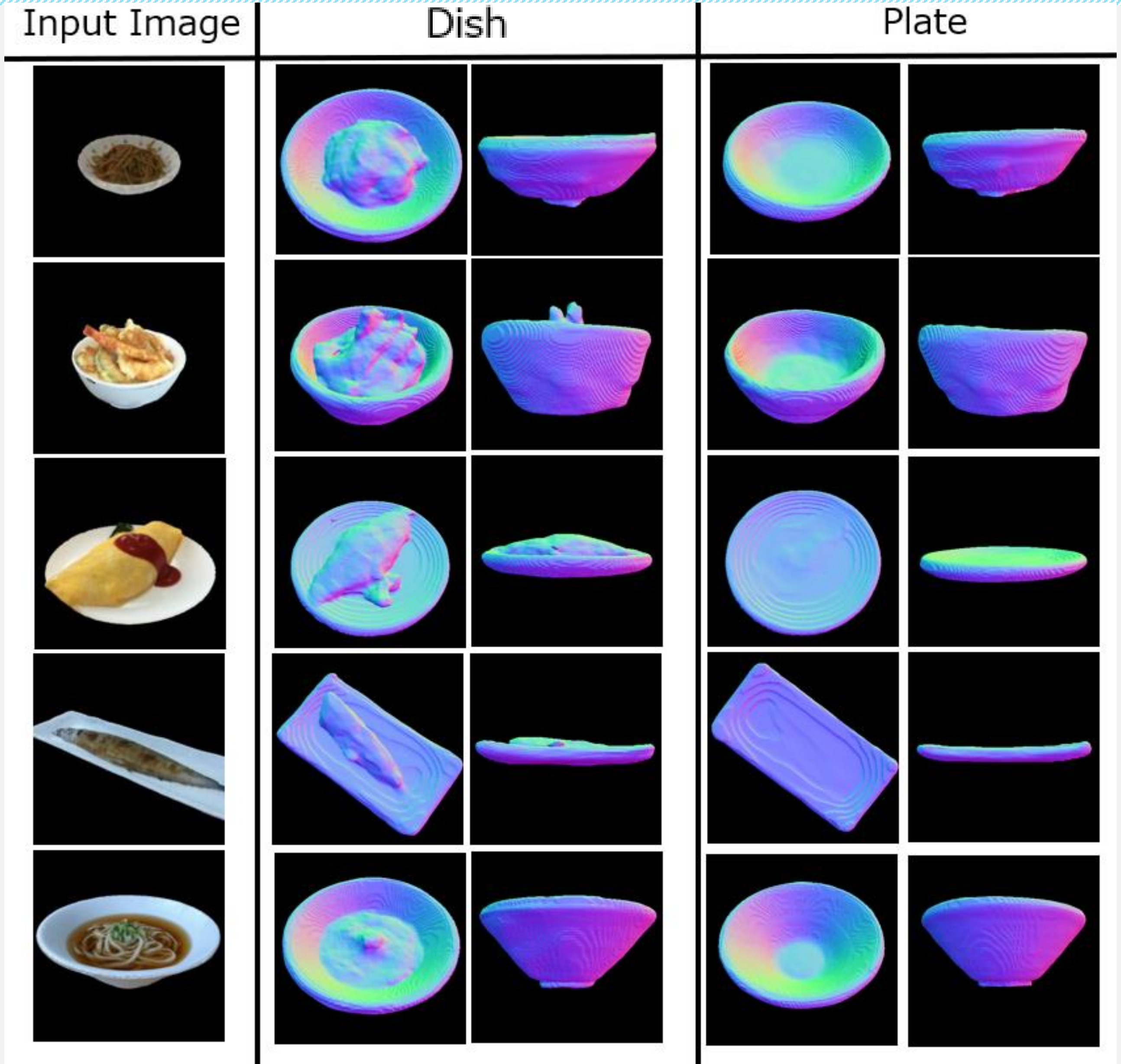
Training dataset : image

- Rendered using blender as well as 3D-R2N2 [13].
- Two patterns of images are available, with background or without



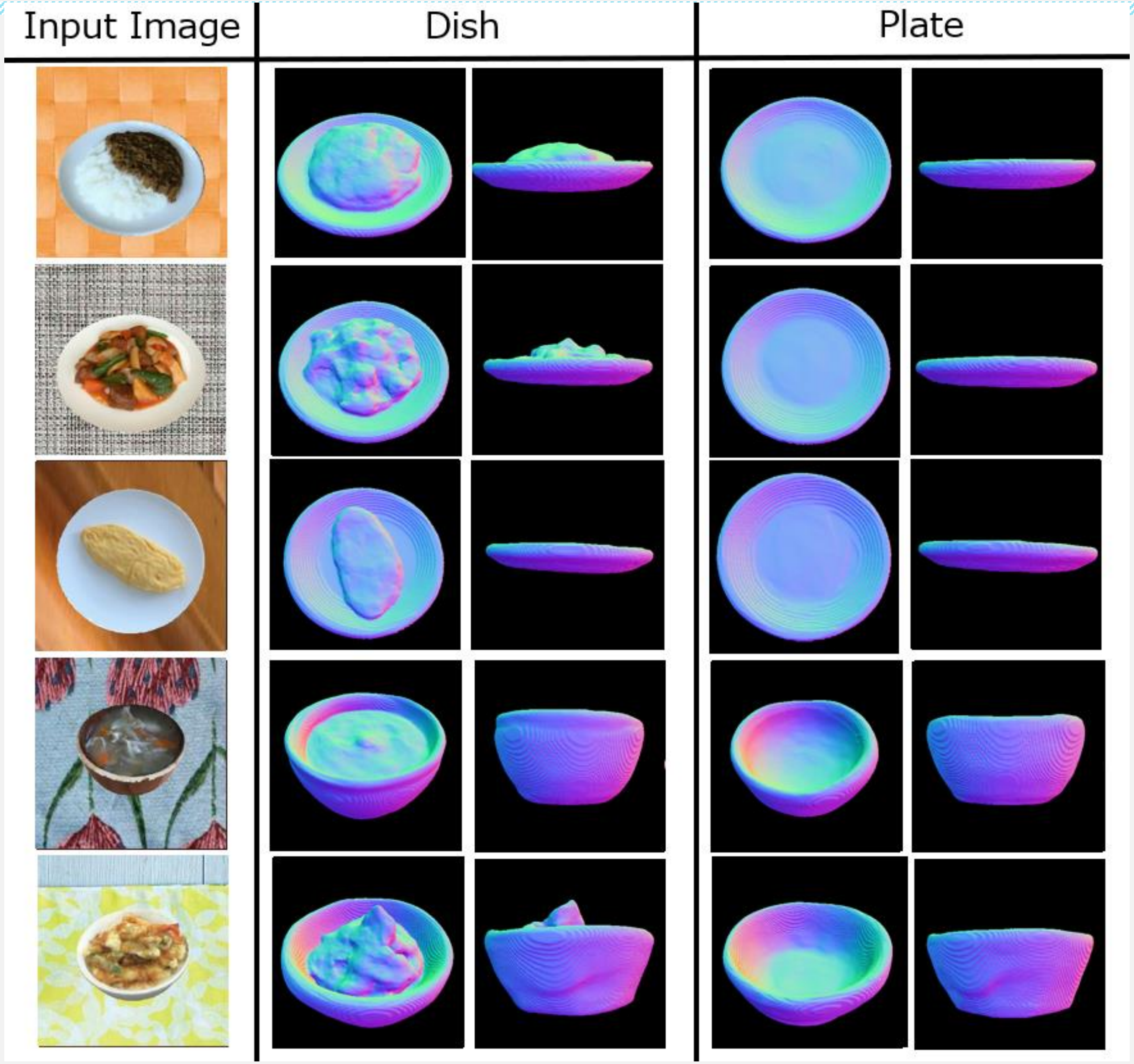
Experiment : Qualitative evaluation

- ResNet18
- λ_3=20
- Without background



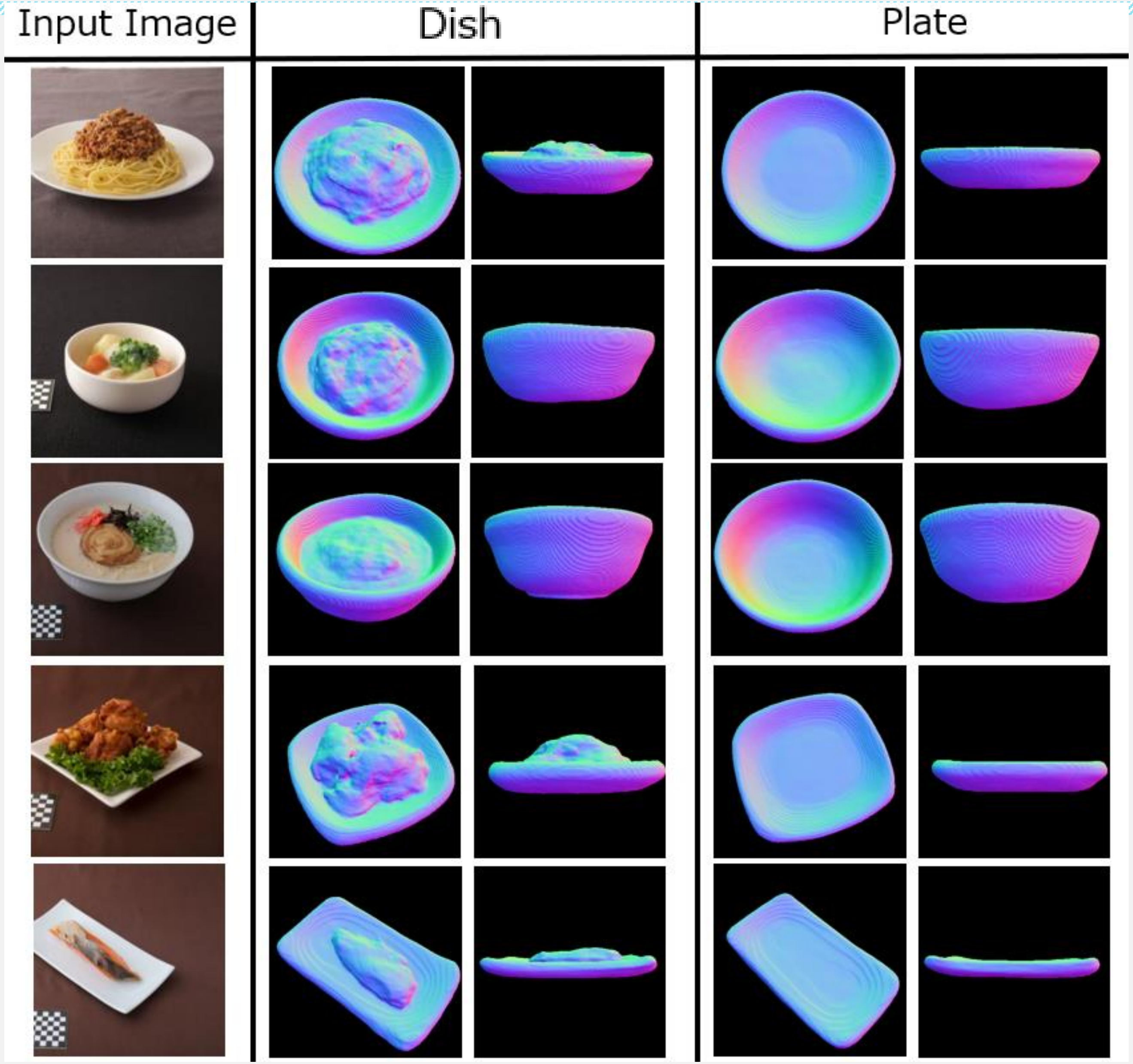
Experiment : Qualitative evaluation

- ResNet18
- λ_3=20
- With background



Experiment : Qualitative evaluation

- ResNet18
- λ_3=20
- With background



Experiment : Quantitative evaluation

- weighting plate consistency loss

λ_3	IoU (dish)	IoU (plate)	Chamfer L1 (dish)	Chamfer L1 (plate)	**plate consistency**	**Volume error**
0	**0.624**	**0.621**	**0.0189**	0.0186	0.0256	0.0252
20	0.550	0.607	0.0262	**0.0182**	0.0168	**0.0155**
50	0.542	0.610	0.0260	0.0209	**0.0152**	0.0161

$$\mathcal{L}_{\mathcal{B}} = \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{K}\Big(\lambda_1\mathcal{L}_{\mathcal{O}}(y1_{i,j}, o1_i(p_{i,j})) + \lambda_2\mathcal{L}_{\mathcal{O}}(y2_{i,j}, o2_i(p_{i,j})) + \underline{\lambda_3}\mathcal{L}_{\mathcal{C}}(y1_{i,j}, y2_{i,j})\Big)$$

Experiment : Quantitative evaluation

- weighting plate consistency loss

plate consistency loss contributes to reducing volume error.

λ_3	IoU (dish)	IoU (plate)	Chamfer L1 (dish)	Chamfer L1 (plate)	plate consistency	Volume error
0	**0.624**	**0.621**	**0.0189**	0.0186	0.0256	0.0252
20	0.550	0.607	0.0262	**0.0182**	0.0168	**0.0155**
50	0.542	0.610	0.0260	0.0209	**0.0152**	0.0161

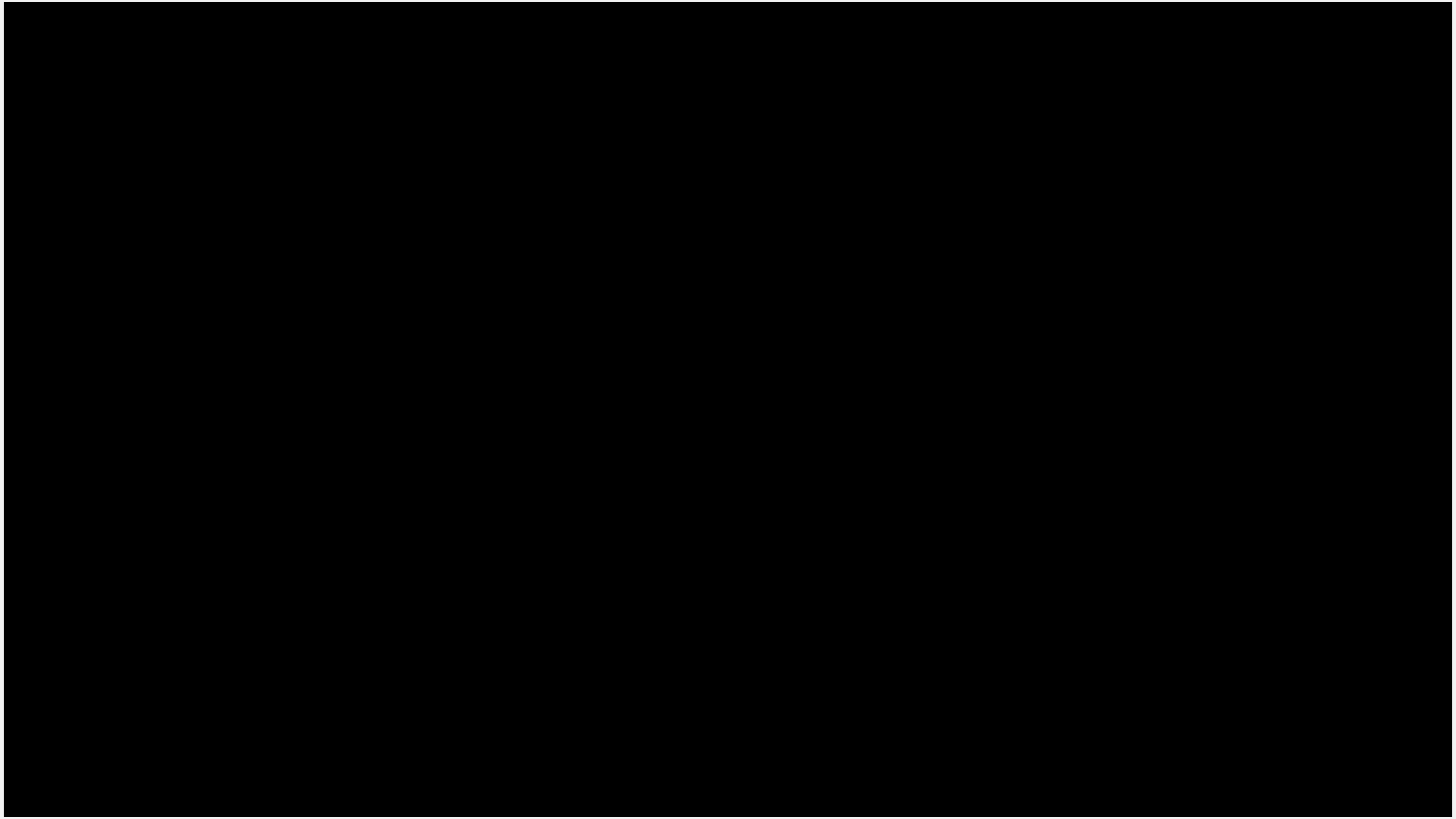
$$\mathcal{L}_{\mathcal{B}} = \frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\sum_{j=1}^{K}\Big(\lambda_1 \mathcal{L}_{\mathcal{O}}(y1_{i,j}, o1_i(p_{i,j})) + \lambda_2 \mathcal{L}_{\mathcal{O}}(y2_{i,j}, o2_i(p_{i,j})) + \underline{\lambda_3}\mathcal{L}_{\mathcal{C}}(y1_{i,j}, y2_{i,j})\Big)$$

Experiment : Quantitative evaluation

- 2 patterns of learning image with / without background
    - Image with background + ResNet18 + λ_3=20 Is the most accurate.

encoder	background	IoU (dish)	IoU (plate)	Chamfer L1 (dish)	Chamfer L1 (plate)	Plate consistency score	Volume error
ResNet 18	none	0.560	0.634	0.0265	0.0193	**0.0146**	0.0150
ResNet 50	none	0.564	0.617	**0.0251**	0.0186	0.0148	0.0147
ResNet 18	yes	**0.565**	**0.645**	0.0254	**0.0173**	**0.0146**	**0.0146**
ResNet 50	yes	0.558	0.628	0.0252	**0.0173**	0.0157	0.0157

Application



https://youtu.be/YyIu8bL65EE

Conclusion

- Hungry Networks
    - Reconstruct 3D dish (food + plate) volume and 3D plate volume from a single dish image
- Introducing plate consistency loss
    - Matching plate parts of the 3D shape of dish and plate
    - Contributes to the accuracy of volume estimation
- Creating a 3D meal dataset for training
    - We showed that it can correspond to the real dish image.

Method objective

Method objective

Method objective

Appropriate 3D representation

Proposed networks

- Hungry Networks
    - Reconstruct two meshes of dish and plate from a single dish image
    - Extend Occupancy Networks [17], an occupancy-based method

- Introducing plate consistency loss
    - Loss function for matching plate parts of the 3D shape of dish and plate