Calorie-Aware Food Image Editing with Image Generation Models

Kohei Yamamoto, Honghui Yuan, Keiji Yanai
The University of Electro-Communications, Tokyo, Japan

Background

- Generative AI is undergoing a rapid evolution.
- Image generation AI is permeating society with the trend of Chat-GPT.

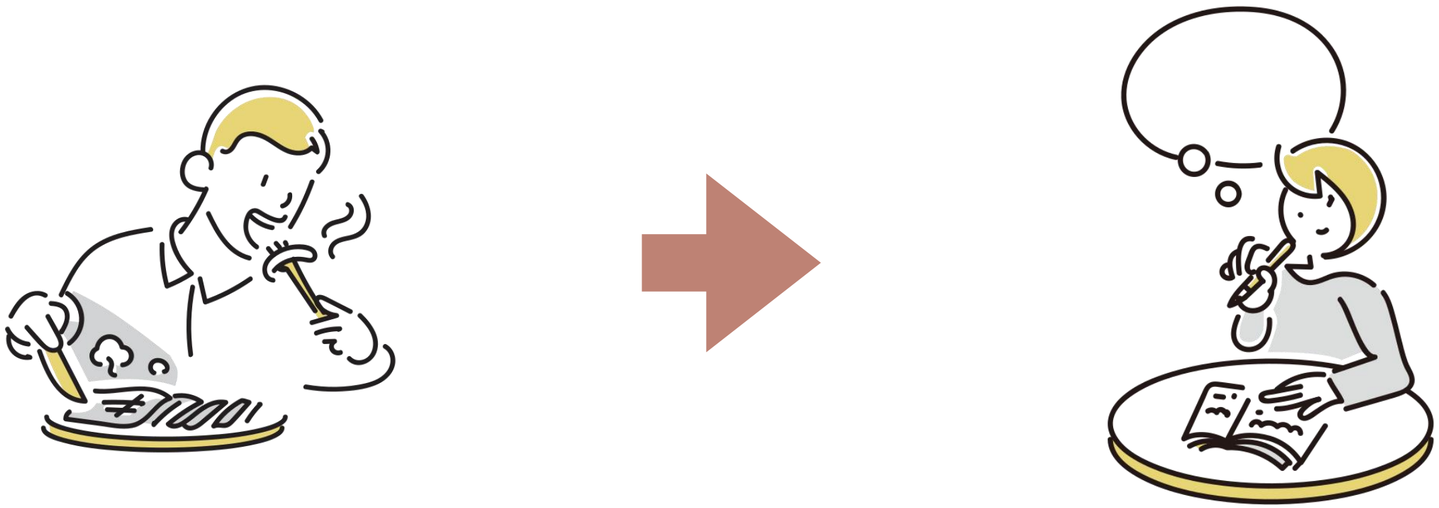
Stable Diffusion

Midjourney

DeepFloyd IF

Background

- Society is becoming more health conscious.
- Improving eating habits can help prevent obesity and lifestyle-related diseases.
- More and more people are using apps to keep track of their diet.



Problem

- Nutritional labeling is uniform and visual quantities are unknown.
- Image generation based on size and quantity is difficult.

NUTRITION INFORMATION		
Servings per can: 2		
Serving size: 210g		
	Average Quantity Per serving	Average Quantity Per 100g
ENERGY	895kJ	425kJ
PROTEIN	10.8g	5.1g
FAT: TOTAL	1.2g	0.6g
-SATURATED	0.2g	0.1g
CARBOHYDRATE	33.7g	16.1g
-SUGARS	15.5g	7.4g
DIETARY FIBRE	11.9g	5.7g
SODIUM	1300mg	620mg
POTASSIUM	650mg	310mg
IRON	2.7mg	1.3mg

Nutritional Information Label

cited from https://www.kateiveyfitness.com/blog/how-to-read-nutrition-labels/

Pasta for one person generated by SD1-4 model

Purpose

- Edit food images based on calorie amount.

687kcal

1031kcal

344kcal

No studies have been conducted on changing amounts of food

Related Work

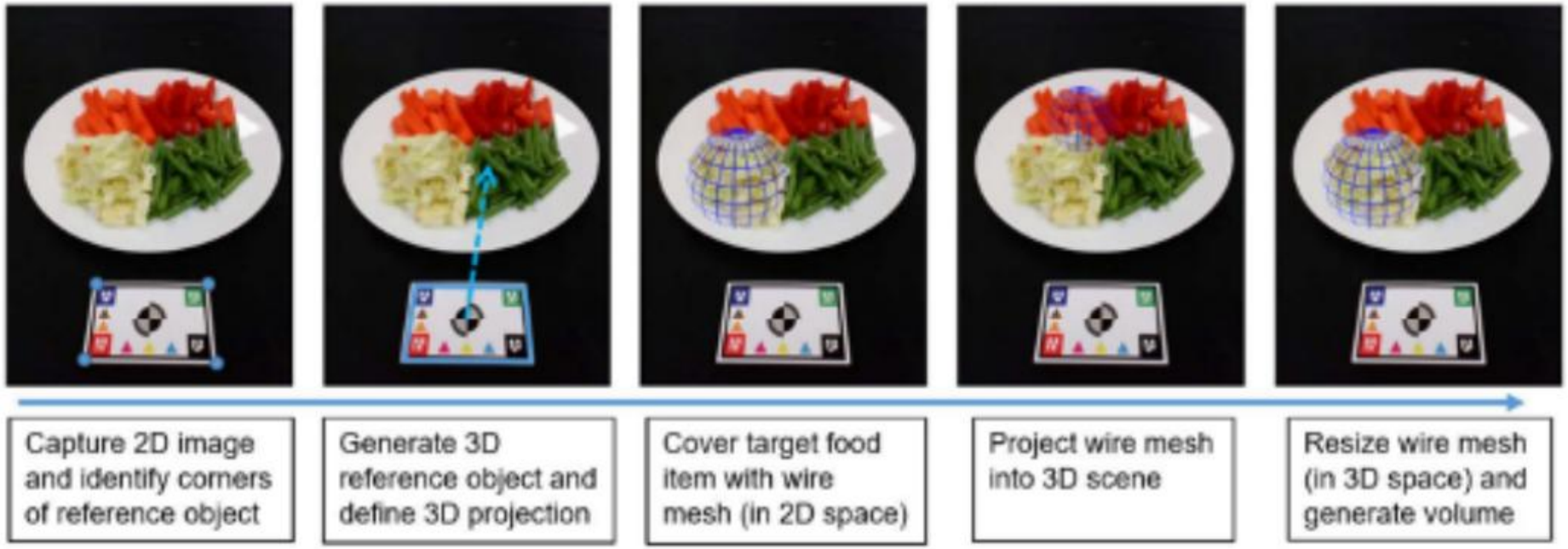
Two main methods exist for calorie measurement with food images.

① Method using the reference object

② Direct estimation using deep learning

Related Work ① Method using the reference object

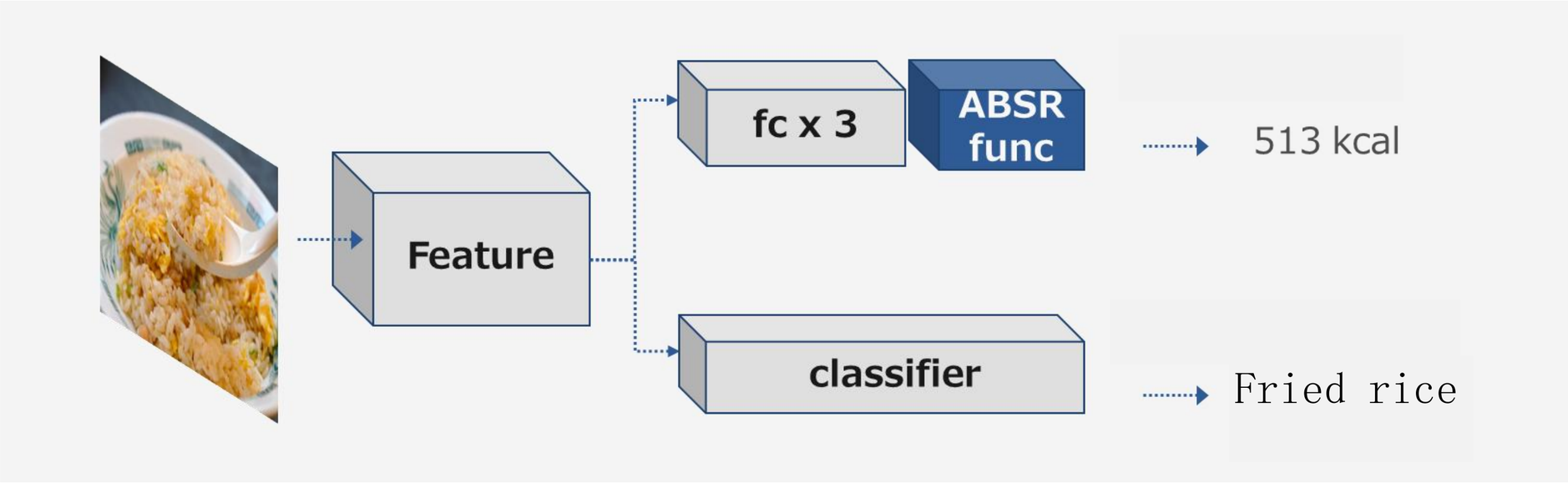
A reference paper is placed in front of the food[1] to recognize three-dimensional space.



[1] Ji-hwan Kim, Dong-seok Lee, Soon-kak Kwon, "Food Classification and Meal Intake Amount Estimation through Deep Learning", Applied Sciences, vol.13, no.9, pp.5742, 2023.

Related Work ② Direct estimation method

Using the latest deep learning model, Swin Transformer V2, and by defining a unique output function[2] to improve accuracy.

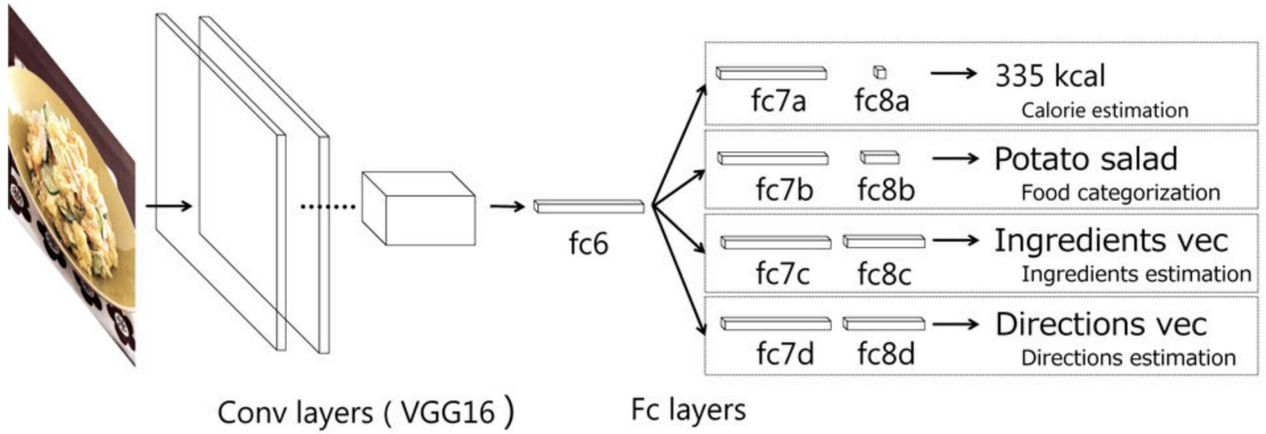


[2] Katsutoshi Maeda. “Estimation of Calorie Amount from Food Images Using Vision Transformer” (in Japanese), 2023.

Related Work ② Direct estimation method

The accuracy was improved by simultaneously learning[3]

- calorie content,
- category classification,
- ingredient estimation,
- cooking procedure.

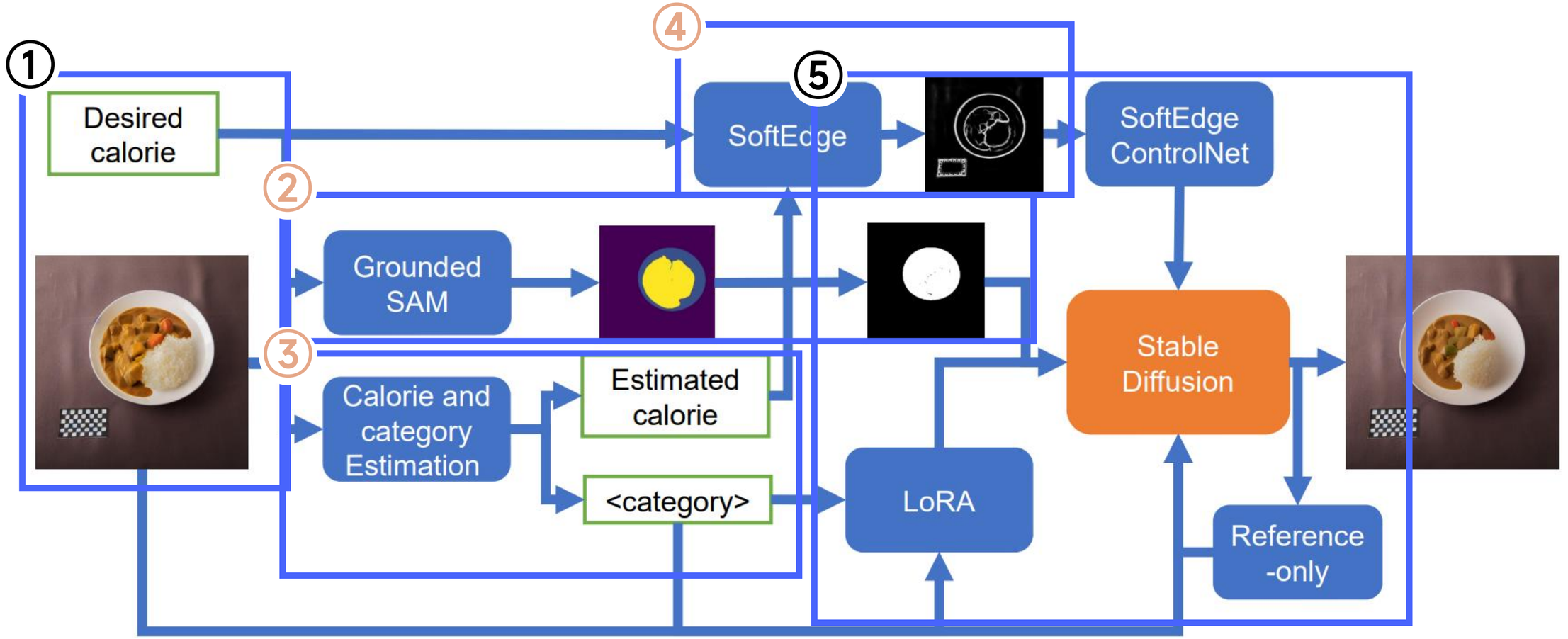


Using direct regression techniques.

[3] Takumi Ege ar ries, ingredients and cooking directions

Proposed Method

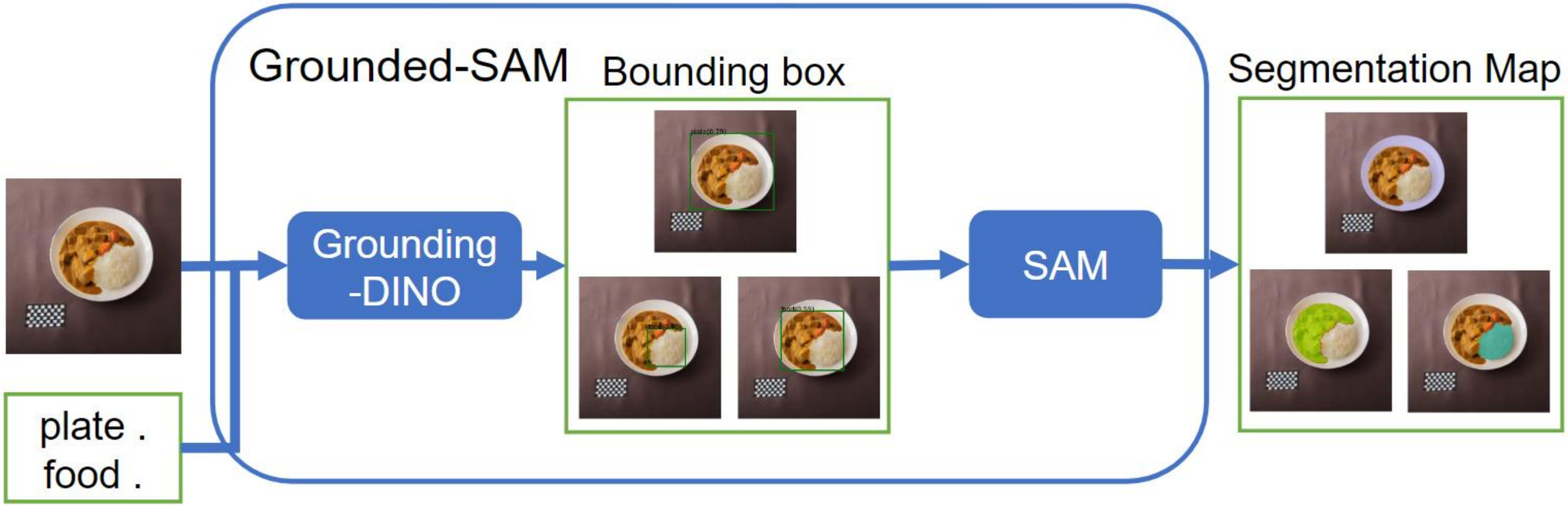
Overview of the proposed method



Proposed Method

Grounded-SAM[4]

- Segment the plate and food areas.

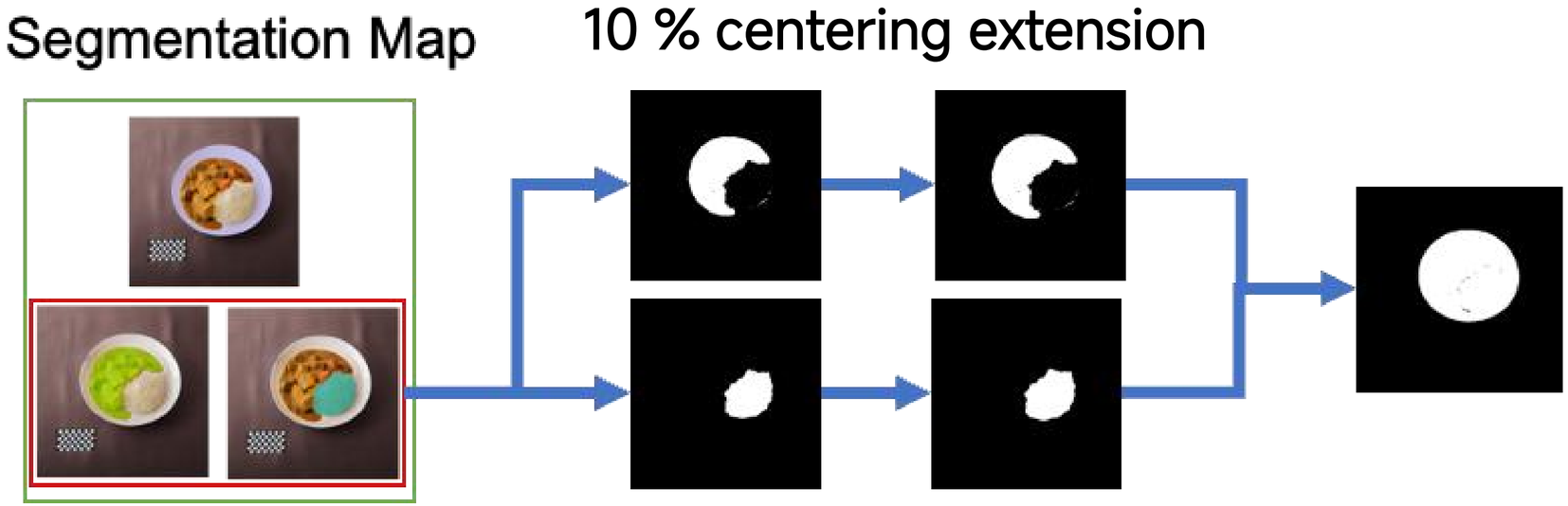


[4] https://github.com/IDEA-Research/Grounded-Segment-Anything
[5] Liu, Shilong, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." arXiv: 2303.05499 (2023).
[6] Kirillov, Alexander, et al. "Segment anything." arXiv:2304.02643 (2023).

Proposed Method

Mask Adjustment

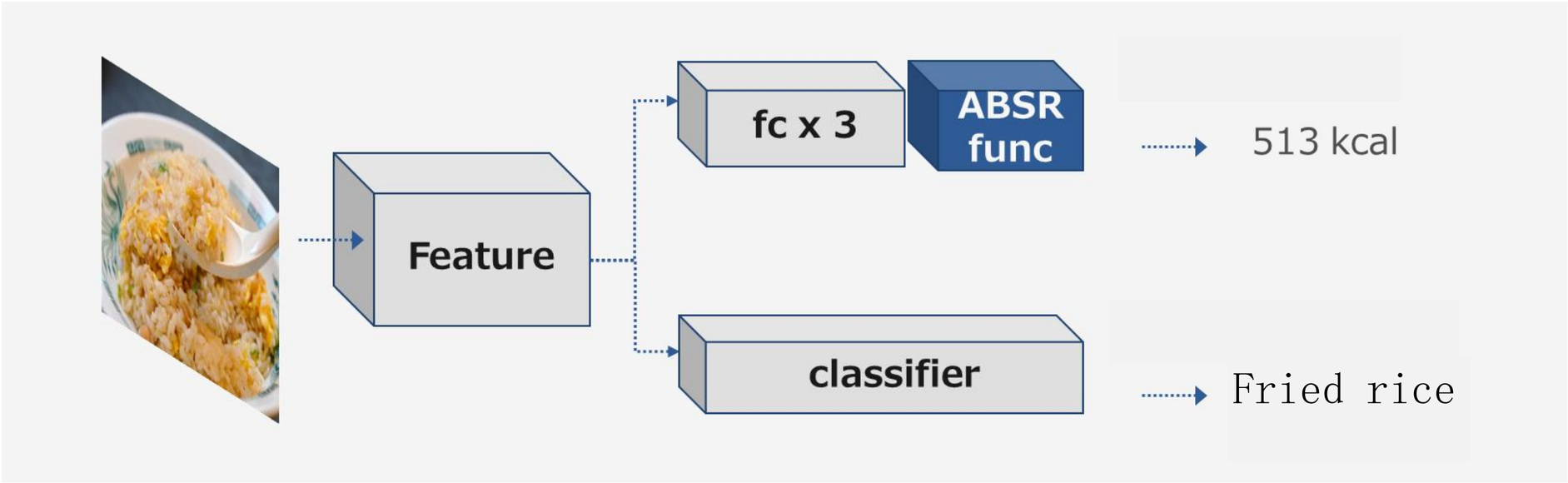
- The detection area of Grounded-SAM is too small to cover the entire food area.
- Therefore, “fine expansion of the food area and merging of the food area” is conducted.



Proposed Method

Calorie Estimation

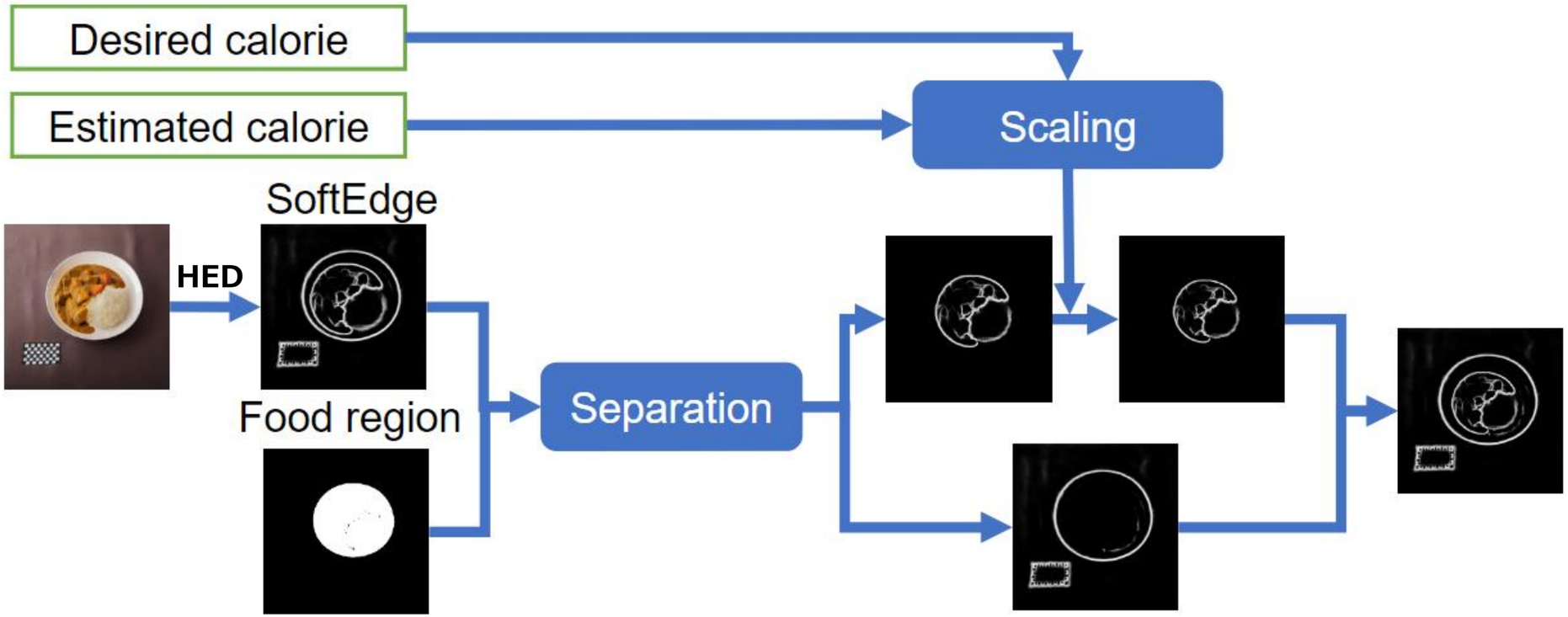
- Caloric content and category estimation model was retrained.
- Increased number of images to improve its accuracy.



Proposed Method

SoftEdge(HED[7])

- Extracts the editable food portion
- Edits the SoftEdge image using the cubic root of the calorie.

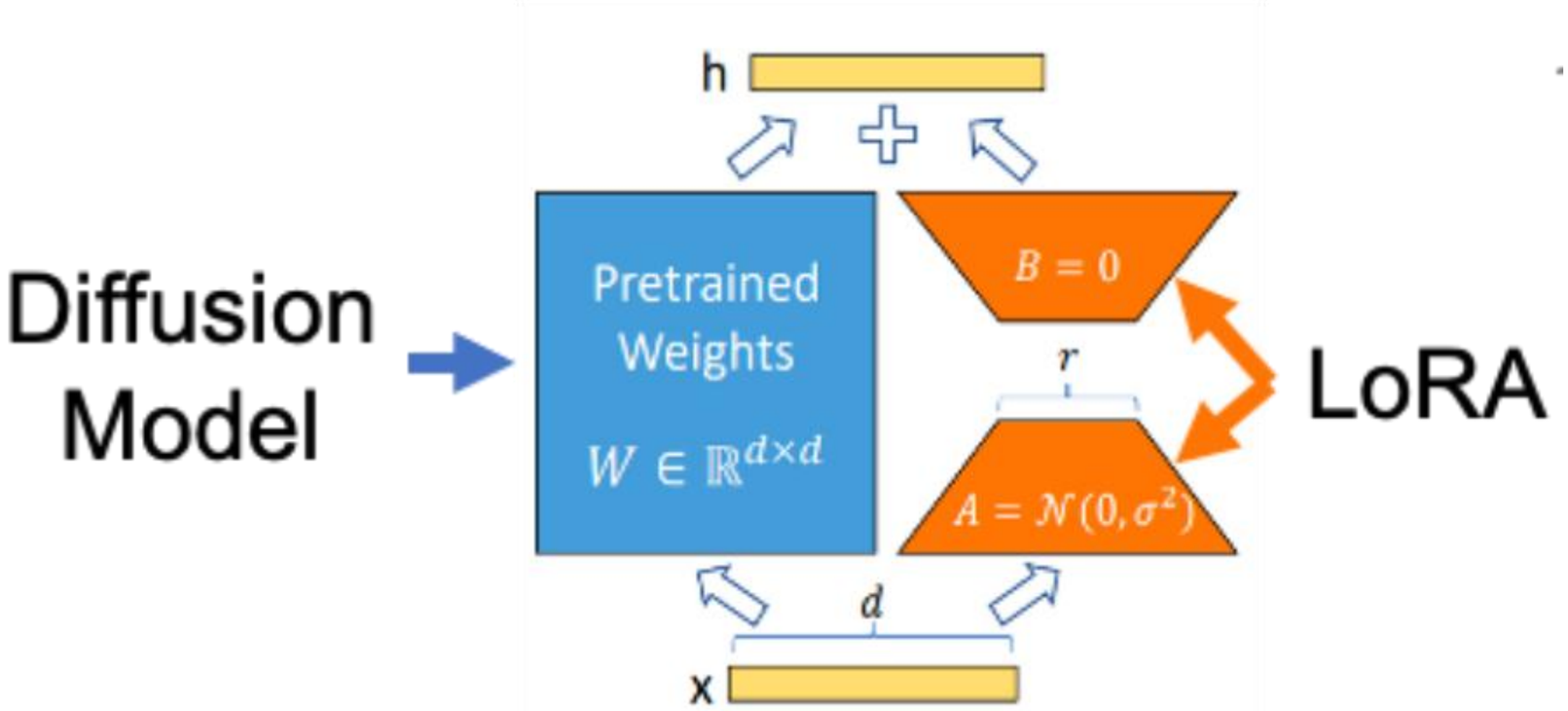


[7] Xie, Saining, and Zhuowen Tu. "Holistically-nested edge detection." CVPR. 2015.

Proposed Method

Appearance Preserving

(1): LoRA study[8].
Learning the appearance of a single image with DragDiffusion[9] settings

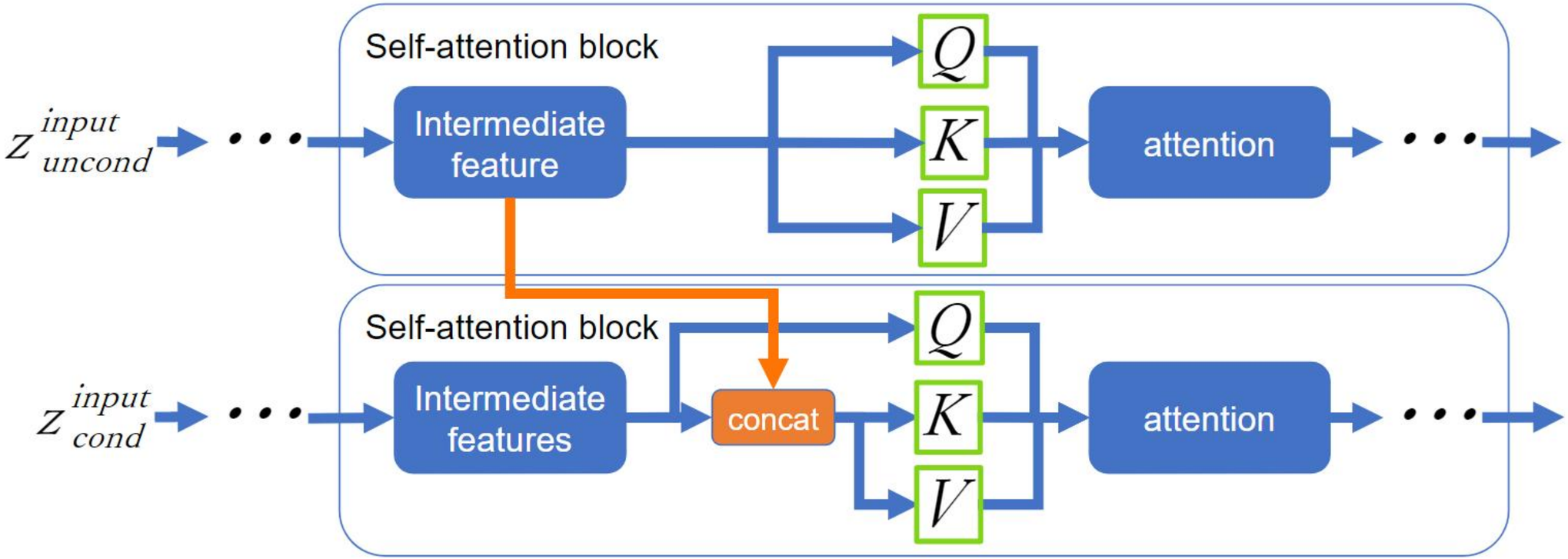


[8] Hu, J. Edward et al. “LoRA: Low-Rank Adaptation of Large Language Models.” *arXiv:2106.09685* (2021)
[9] Shi, Yujun, et al. "DragDiffusion: Harnessing Diffusion Models for Interactive Point-based Image Editing." *arXiv:2306.14435* (2023).

Proposed Method

Appearance Preserving

(2): Reference-only[10].
Preserves intermediate features and combines them during inference.

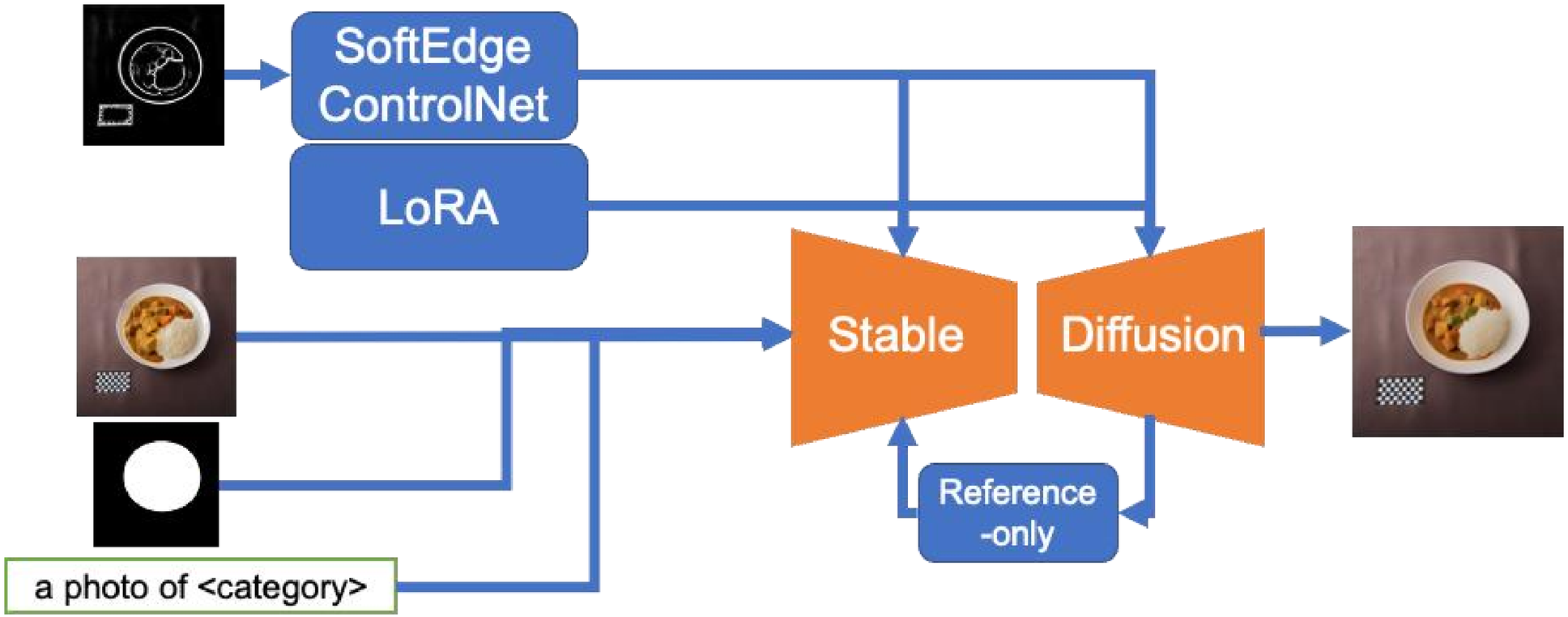


[10] https://github.com/Mikubill/sd-webui-controlnet/discussions/1236

Proposed Method

Generation

Output edited image as Inpainting model using StableDiffusion



Experiments

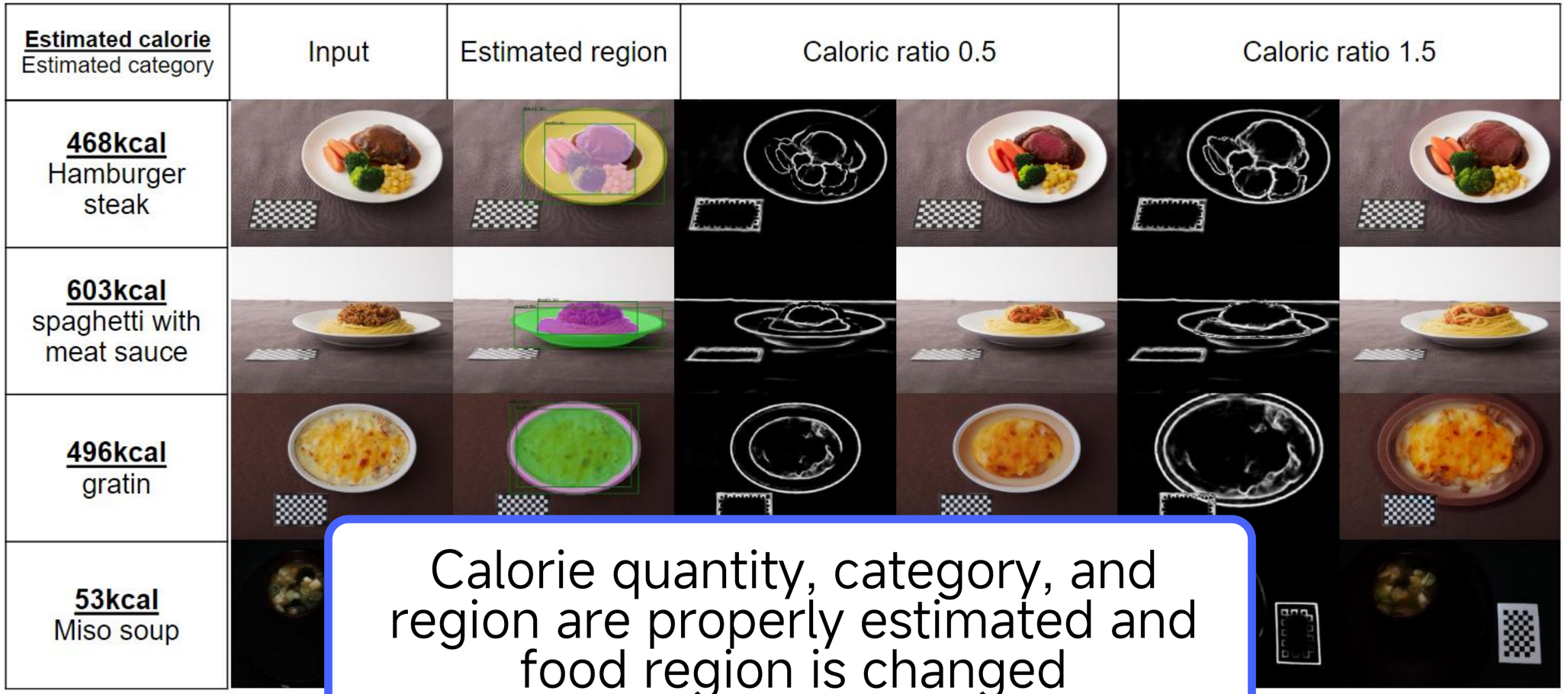
Calorie Estimation

- The model was re-trained with approximately 7.6 times more images than the original model
- evaluated with 7,407 food images with caloric content collected from the Internet.

		Original		Re-trained	
		Ege *et al.*[2]	new dataset	Ege *et al.*[2]	new dataset
Calorie	Absolute error [kcal]↓	87.5	161.0	166.4	**80.0**
	Relative error [%] ↓	27.8	68.1	61.7	**25.5**
	Ratio within 20% error ↑	0.536	0.301	0.240	**0.623**
Category	Top-1 accuracy [%] ↑	**89.2**	27.9	46.8	73.0

Experiments

Results of edited images

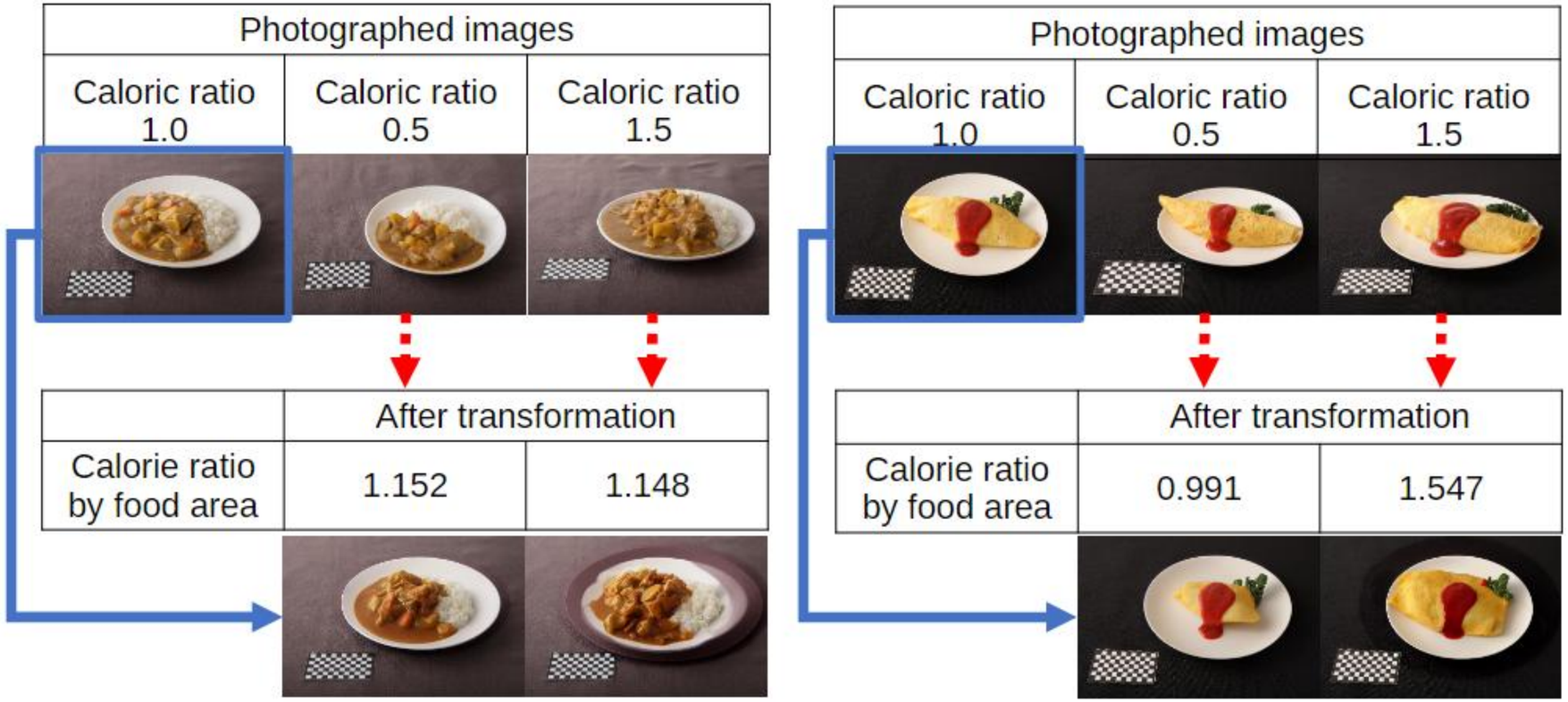


Calorie quantity, category, and region are properly estimated and food region is changed

Experiments

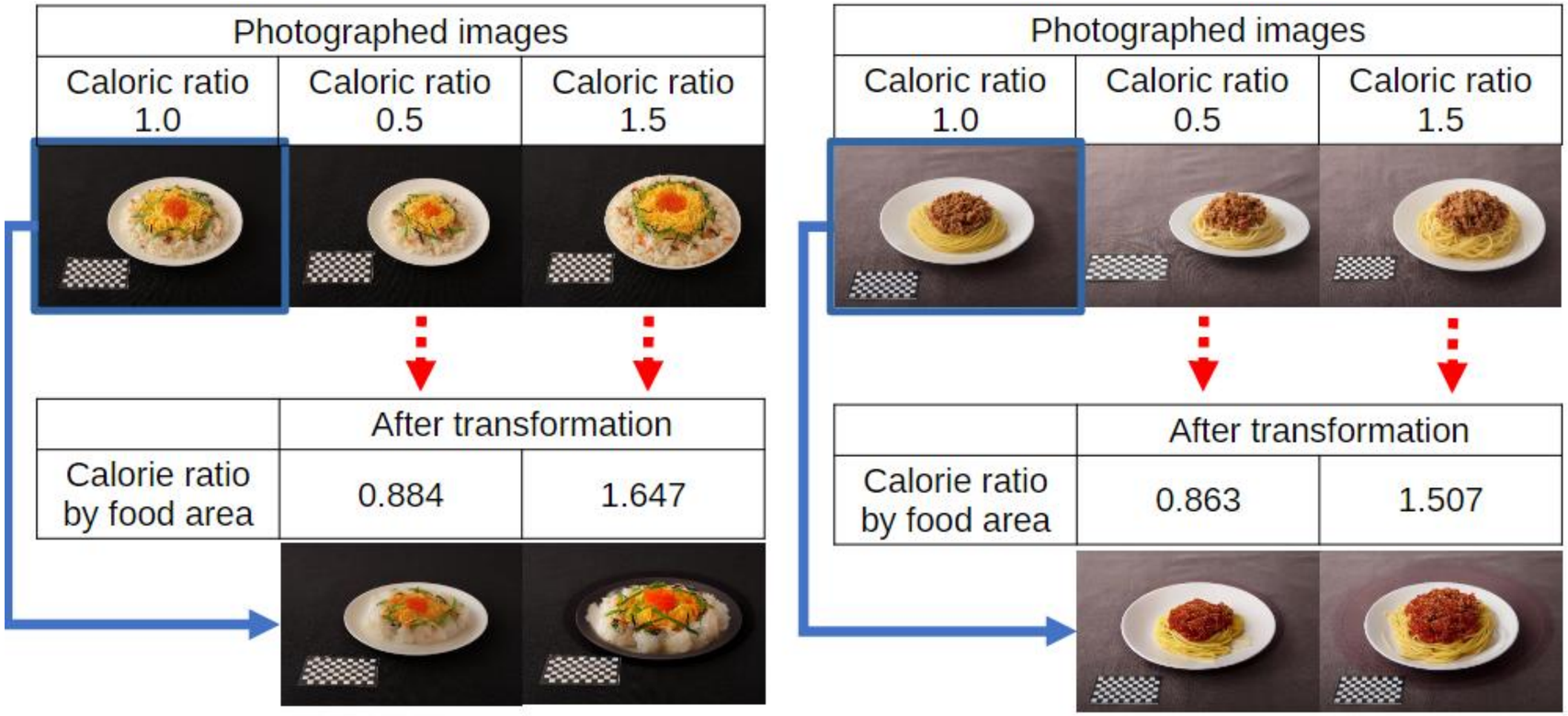
Comparison with photographed image

No inconsistency with the photographed real image.



Experiments

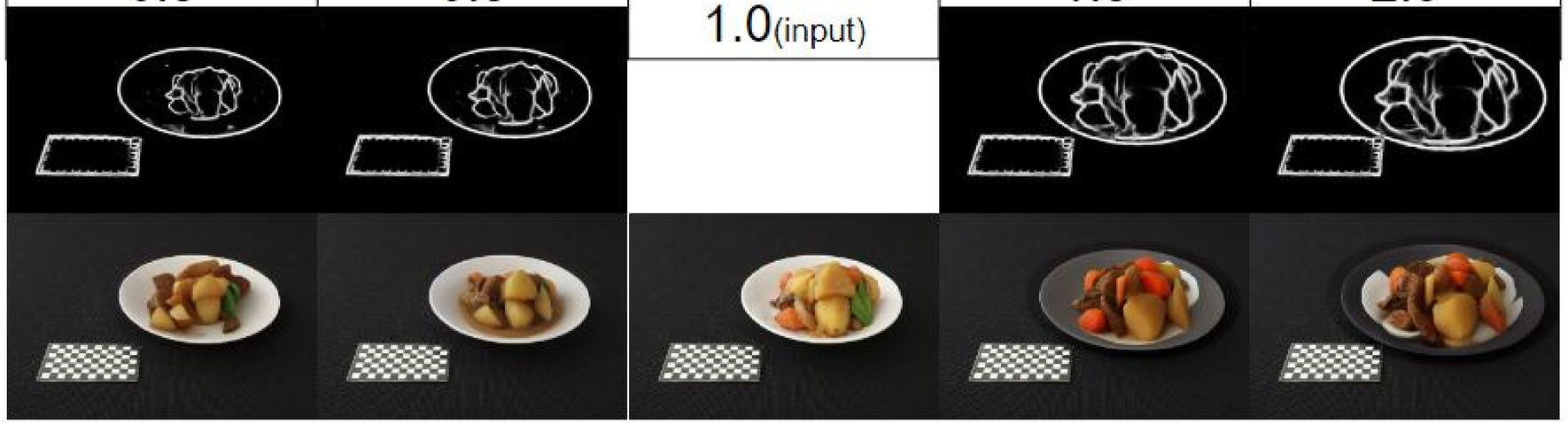
Comparison with photographed image



Experiments

Change size (with seed value of 0)

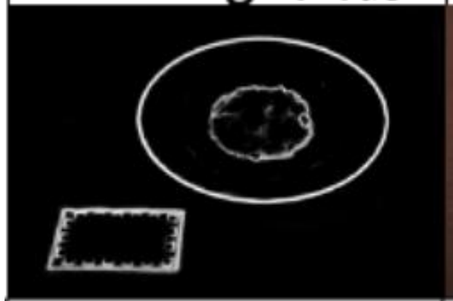
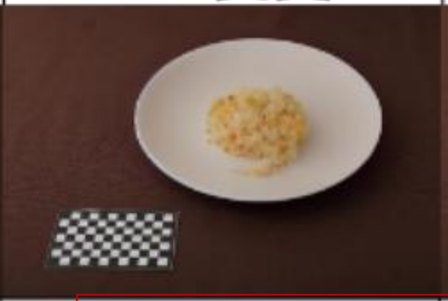
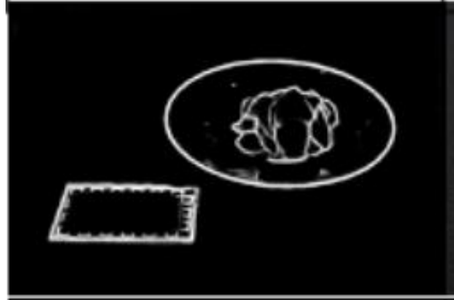
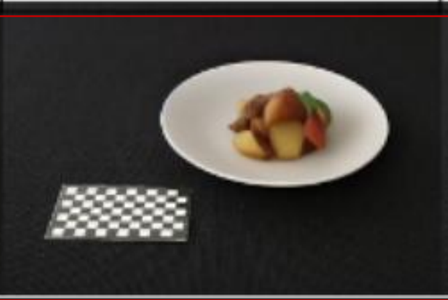
Except for the calorie ratio of 0.3, the values change according to the SoftEdge image.

	0.3	0.5	Caloric ratio 1.0(input)	1.5	2.0
					
Calculate	0.1004	0.0793	0.1143	0.1311	0.2851
Expect	0.0512	0.0720		0.1498	0.1814

Experiments

Seed change (for a calorie ratio of 0.3)

Improvement is seen when Seed values are changed.

SoftEdge画像	seed:0	seed変更1	seed変更2	seed変更3
				
理想　0.0529	0.1187	0.0530	0.0594	0.0517
				
理想 0.0512	0.1004	0.0621	0.0526	0.0511

Experiments

Quantitative evaluation

Calorie ratio of food regions for 100 generated images

Food Category	Estimated Calorie Ratio at x0.5	Estimated Calorie Ratio at x1.5
Nikujaga	0.522 ± 0.0417	2.485 ± 1.3268
Fried rice	0.507 ± 0.0378	1.905 ± 1.0061
Chirashi-sushi	0.428 ± 0.1687	1.257 ± 0.5095
Curry	0.559 ± 0.0714	1.558 ± 0.6483
Stie-fried nooodles	0.511 ± 0.0112	1.530 ± 0.2707
Gratin	0.498 ± 0.0544	1.397 ± 0.2116
Hamburg steak	1.184 ± 0.1941	2.683 ± 1.4070
Miso soup	1.321 ± 1.6865	3.576 ± 1.5958
Mixed rice	1.505 ± 0.9103	1.716 ± 0.2736
Omelet rice	0.818 ± 0.1661	1.932 ± 0.9982
Pilaf	0.511 ± 0.0104	1.561 ± 0.5165
Potato salad	0.730 ± 0.2682	1.523 ± 0.3880
Spaghetti with meat sauce	0.374 ± 0.1403	1.260 ± 0.6171
Cream stew	0.572 ± 0.1448	1.381 ± 0.4155
All categories average	0.717 ± 0.2790	1.840 ± 0.7275

Conclusion

Image editing based on calorie content

- Calorie Content Estimator Training
- Results of Image Editing
- Quantity changes and quantitative evaluation

Remaining challenges

- Increase the types and the accuracy of meals that can be recognized by the caloric content estimator
- Address cases of failure to segment regions well.