Calorie-Aware Food Image Editing with Image Generation Models

Kohei Yamamoto[0000−0002−1733−458X], Honghui Yuan[0009−0001−4334−9363], and Keiji Yanai[0000−0002−0431−183X]

The University of Electro-Communications, Tokyo, Japan
{yamamoto-k,yuan-h,yanai}@mm.inf.uec.ac.jp

Abstract. With the development of AI models such as ChatGPT, artificial intelligence has become deeply integrated into our daily lives. Additionally, the growing focus on health and wellness has accelerated the development of AI applications in healthcare. In the realm of dietary management, some smartphone applications offer automated calorie calculation and nutrient tracking features. However, these systems often rely on nutrition facts labels, making it challenging for users to visually comprehend the portion sizes corresponding to their desired caloric intake. To address this limitation, this paper proposes a novel food image editing model that incorporates image generation AI to adjust the caloric content of food images. Our model begins by extracting features, such as estimated current calories, food regions, and visual attributes, from input food images. Subsequently, a conditional edge image is generated based on the desired caloric value and food regions. By provding these features into an image generation model, our model produces a new food image that aligns with the specified caloric target. This approach enables accurate and visually intuitive dietary management.

Keywords: Generative AI · Food Image Editing · Food Image Generation · Food Calorie Estimation

1 Introduction

In the field of artificial intelligence research, language models and vision-language models trained on massive amounts of data using Transformers [22] have been rapidly evolving. Furthermore, generative AI, represented by ChatGPT[1], is becoming increasingly prevalent in society. In the field of image generative AI such as Stable Diffusion [17], midjourney[2], and DeepFloyd IF[3] are well-known. However, it is not possible to accurately generate images that match one's imagination, and it takes time and trial and error to generate images that are manipulated in size or that conform to physical laws. In particular, human hands,

[1] https://openai.com/chatgpt
[2] https://www.midjourney.com/
[3] https://www.deepfloyd.ai/deepfloyd-if

eating behavior, quantities, and characters are difficult for generative AI to accurately represent, and they are useful for distinguishing generated images from real ones. Food images remain a challenging domain for generative AI. Even with text prompts combining a food category and quantity, it is highly unlikely that the generated image will strictly adhere to the textual description. In fact, when an image is generated using Stable Diffusion v1.4 with the prompt "a photo of spaghetti for one person," the result is as shown in Figure 1. Although it is instructed in English to be for one person, the output image is more than one serving.

Fig. 1. Output images generated by Stable Diffusion v1.4 with the input "a photo of spaghetti for one person"

Moreover, as sports facilities improve, new health equipment is developed, and health promotion activities increase, health consciousness is rising across society. Today, it's easy to track exercise and manage diets using smartwatches and apps, making health management more accessible.

Diet-tracking apps, for instance, can automatically calculate calories and record nutrients, offering increasingly sophisticated features. However, these calculations often rely on standardized values from food labels, which can lead to inaccuracies when portion sizes vary. Additionally, it is not always easy to visually estimate the amounts specified on food labels.

In this study, we propose a food image editing model that considers calorie amounts by combining an image generation model and a food calorie estimation model. To generate images that consider calorie amounts, we need an original image and the desired calorie amount value as inputs.

The main contributions of this paper are as follows:

- We propose a method by combining an image synthesis model and a food calorie estimatino model to modify a given food image into a food image with the specified calorie amount, which enables us to visually grasp the change in calorie amount and to manage diet appropriately.
- By the comprehensive experiments, we confirmed the effectiveness of the proposed method.

2 Related Work

2.1 Calorie Estimation from Images

There are primarily two methods for estimating calorie content from images. The first method involves using a reference object while the second method directly estimates calorie content using deep learning.

In the method using a reference object, Smith *et al.* [8] conducted a study. Their study first places a reference paper in front of the food and detects the corners of the paper to recognize the three-dimensional space. Next, they manually construct a 2D mesh of the food and project it into 3D to calculate the volume. Finally, the calorie content is calculated based on the volume.

On the other hand, in the method using deep learning directly, studies by Ege *et al.* [3] and Maeta [13] exist.

Ege *et al.*proposed three methods for estimating calorie content from images. Among the three methods, it was found that accuracy improved by simultaneously learning calorie content, category classification, ingredient estimation, and cooking procedures.

Based on the study by Ege *et al.*, Maeta used the Swin Transformer V2, a state-of-the-art deep learning model, and an original output function, AutoBinning Softmax-Regression, to improve accuracy. Similar to Ege *et al.*'s study, this work performs simultaneous learning to estimate both calorie amounts and food category.

In this work, we like to estimate calorie content even in images without reference objects, and therefore adopt Maeta's method, which is the latest method for directly predicting calorie content using deep learning.

2.2 Food Image Generation

Many studies on food image synthesis used the Recipe1M citerecipe1M dataset. CookGAN [25] by Zhu *et al.*is a text image generation model that considers ingredients and procedures using a cooking simulator sub-network. CookGAN [4] by Han *et al.*generates meal images conditioning the generation network with an attention-based ingredients-image association model for associating food items with images. ChefGAN [16] by Pan *et al.*used an image-recipe embedding model to generate a similar representation of a recipe and an image and generate the image considering the recipe.

2.3 Deep Learning Applications for Food

Deep learning applications for food transformation mainly include food transformation and food recognition.

For food transformation research, there are mobile food transformation apps by Tanno *et al.* [20], food image transformation by Horita *et al.* [5], and Enchanting Your Noodles by Nakano *et al.* [14]. For food recognition research, there are

CaloriesCaptorGlass by Naritomi *et al.* [15] and CalorieCam360 by Terauchi *et al.* [21].

The research by Tanno *et al.*and Horita *et al.*on food transformation uses cCycleGAN, which adds a conditioning vectors to the generator and discriminator of CycleGAN [26] to enable conditioning. The mobile app of Tanno *et al.*is a product that allows you to convert meals on your smartphone and experience meal transformation closely.

Nakano *et al.*'s research, transforms simple ingredients such as somen noodles and rice into "attractive" dishes such as ramen, curry, and fried rice through a VR headset. This research is the first food editing research using VR in the food domain, and by changing the type of food in the video in real time, it was possible to obtain effects other than visual effects, such as changing the taste of the food being eaten.

Naritomi *et al.*'s research on food recognition calculates the volume of food using AR glasses, and displays the volume and calorie content in 3D. Previous studies often measure calorie content from two-dimensional images. However, by using AR glasses, it is possible to estimate calorie content using the depth of food, which is difficult to predict in 2D.

Terauchi *et al.*proposed a system that can recognize food, estimate calorie content, and measure food intake using commercially available 360-degree cameras, which have become widespread in recent years. By using 360-degree cameras, it has become possible for multiple people to simultaneously record food log.

This study differs from these applied studies in that it focuses on food image editing rather than food transformation, and it utilizes recent diffusion models for transforming food amounts.

3 Method

3.1 Overview of the Method

In this study, it is necessary to adjust the overall shape of food, which requires powerful image control.

ControlNet [24] is representative of the method using additional networks. ControlNet fixes the diffusion model to be used and constructs and learns a newly learnable encoder part of the diffusion network. This allows us to take advantage of the generative capabilities of the diffusion model while avoiding overfitting even with small datasets and promoting early convergence of learning. It can generate images with various conditions such as Canny images, OpenPose images, and depth images. Due to its high controllability, it is very often used in the creation of human images and manga generation.

Therefore, we decided to use ControlNet [24] for this study. In order to get SoftEdge images, which requires the detection of the food region. Therefore, we used Grounded-Segment-Anything (Grounded-SAM)[4] for food region detection.

[4] https://github.com/IDEA-Research/Grounded-Segment-Anything

Furthermore, a calorie estimation model is required to determine how much to expand or contract the food region based on the calorie amount. Finally, a mechanism is needed to adjust the appearance, and in this case, we decided to use LoRA [6] and Reference-only [5].

Considered these points, the proposed method has the structure shown in Figure 2. As inputs, the model takes an input image and the desired calorie

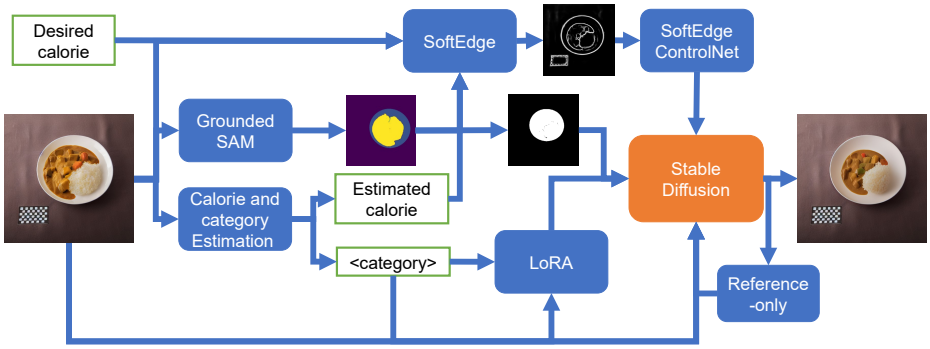

Fig. 2. Structure of the proposed method

amount after transformation. First, the input image is provided to the calorie estimation model to obtain the calorie amount and category. At the same time, by inputting the input image and prompt "a photo of food" to Grounded-SAM, the bounding box and the segmentation mask are obtained. Next, the segmentation mask obtained from Grounded-SAM is fine-tuned for SoftEdge image editing. Based on the fine-tuned segmentation mask and the ratio of the desired calorie amount to the measured calorie amount, the SoftEdge image of the input image is adjusted. Furthermore, LoRA is learned from a single image, and Reference-only is used to obtain appearance features. Finally, using the input image, segmentation mask, SoftEdge image, LoRA, and Reference-only obtained so far, a food image with the adjusted calorie amount is generated.

3.2 Calorie Estimation

In order to generate an image considering the calorie amount, it is necessary to estimate the current calorie amount of the input image. Therefore, in this study, we used Maeda's model [13] as the calorie estimation model. Maeda's model is trained on a food image dataset with calorie information created by Ege *et al.* [2]. Ege's dataset contains 15 food category and 4,877 images, from 6 recipe sites.

However, in this study, in order to obtain higher calorie detection accuracy, we collected a dataset for additional learning. Similar to Ege *et al.*, we collected

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

images of 14 food categories for which there were enough images from recipe sites with calorie information. By retraining Maeda's model using these images, we aimed to improve the accuracy of the calorie recognition model.

3.3 Grounded-SAM

In this study, we primarily adjust the quantity using ControlNet [24] with SoftEdge images. To edit the SoftEdge image, it is necessary to recognize which area contains food or a plate. Therefore, we used Grounded-SAM, which obtains a bounding box and segmentation mask of an object by inputting arbitrary text and an image. Figure 3 is shown the structure of Grounded-SAM. Grounded-

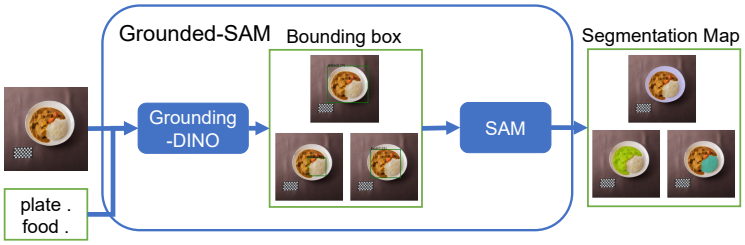

Fig. 3. Structure of Grounded-SAM.

SAM is an open segmentation model that uses Grounding-DINO [11] and Segment Anything (SAM) [9]. Grounded-SAM performs object detection with the zero-shot object detection model Grounding-DINO and performs segmentation with SAM based on the bounding boxes. In this study, to detect the plate and food areas, we input an image and "plate . food ." into Grounded-SAM to segment the plate and food areas. Note that Grounded-SAM text entry is to be separated by a period, and "plate . food ." means that plates and foods are recognized separately.

3.4 Fine-tuning the Segmentation Mask and Editing the SoftEdge

In this section, we obtain the SoftEdge image to transform the food quantity. In general, the segmentation mask obtained by Grounded-SAM is too narrow to edit the SoftEdge image. We expand the mask by 10% so that the line drawing can be edited easily.

By adjusting the segmentation mask, one food area is obtained for one plate area. Based on this, the food area in SoftEdge image is edited to the desired calorie amount as shown in Figure 4.

First, the input image is converted to a SoftEdge image. The SoftEdge image is generated by Holistically-nested Edge Detection (HED) [23], an image-to-image prediction deep learning model composed of a convolutional network.

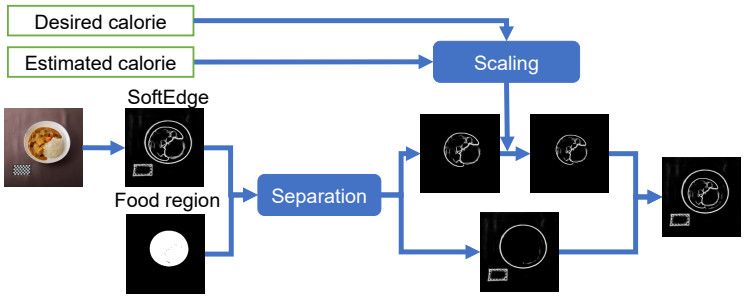

Fig. 4. Editing the SoftEdge image. The figure shows the case where the SoftEdge image is resized to $\frac{2}{3}^{\frac{1}{3}} \fallingdotseq 0.874$ times in the vertical and horizontal directions, assuming the desired calorie amount is 400 kcal and the measured calorie amount is 600 kcal.

Second, the food area and background are separated using the adjusted segmentation mask. Third, the food area of the SoftEdge image is adjusted. The editing is performed by considering the ratio of the desired calorie amount to the measured calorie amount as the volume ratio, and using the area ratio to perform a resizing process. Since the calorie amount is obtained by volume, if the vertical and horizontal (and depth) sizes are simply multiplied by n times, the volume ratio (calorie amount) becomes n^3 times larger. If we want to increase the calorie amount by n times, we need to multiply the vertical and horizontal sizes of the area by $n^{\frac{1}{3}}$. For example, if the desired calorie intake is 400 kcal and the measured calorie intake is 600 kcal, the volume ratio of the food area's SoftEdge image is $4 : 6 = \frac{2}{3} : 1$. Therefore, by resizing both the length and width by a factor of $\frac{2}{3}^{\frac{1}{3}} \fallingdotseq 0.874$, the desired volume (calorie amount) can be obtained. Finally, by combining the edited SoftEdge image with the plate area SoftEdge image, a SoftEdge image for generating images with the adjusted calorie amount is obtained. Note that the fine-tuning of the segmentation mask and the editing of the SoftEdge image work similarly even when there are multiple plates or foods. Furthermore, since the segmentation map is used as an inpainting mask for the diffusion model, only when increasing the calorie amount, the plate mask is also enlarged in proportion here.

3.5 Preserving Appearance

As described in the previous subsection, ControlNet [24] can adjust the shape of the food. However, ControlNet does not preserve the food appearance. Therefore, in this work, we use two mechanisms, LoRA [6] and Reference-only[6], to preserve the appearance.

LoRA is an additional network that learns only the differences in weights for new concepts while fixing the weights of the learned network. In Stable Diffusion, training of LoRA is performed by inserting it into the linear layer of the attention

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

layer. Since it is a very lightweight network, training is easy, and it is also used to easily express features such as style, characters, background, and pose.

In addition to LoRA, Reference-only is also used to preserve the appearance. Reference-only is a structure as shown in Figure 5, that can reflect the appearance features of the original image by saving the intermediate features of self-attention without conditioning at each layer and combining them with the calculation of attention keys and values with condition. Although Reference-only

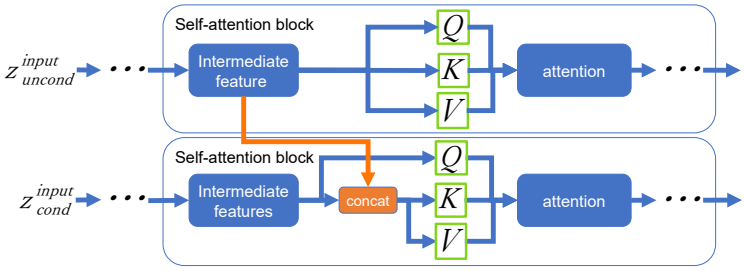

Fig. 5. Structure of Reference-only.

does not require training of networks unlike LoRA, it is necessary to have intermediate features without conditioning. Therefore, the image generation time increases. In this study, we use both LoRA and Reference-only to preserve the appearance of the original image.

Finally, as shown in Figure 6, an image with an adjusted calorie amount is generated using the input image, segmentation mask, edited SoftEdge image, and LoRA obtained so far.

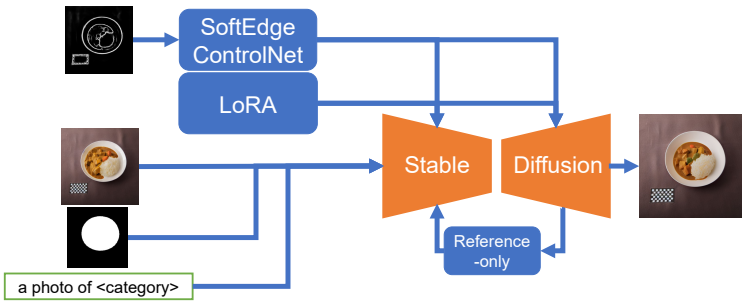

Fig. 6. The final output part of the proposed method.

To generate a calorie-edited image, we use Stable Diffusion's inpainting mechanism. For the prompt, we use "a photo of <category>" where <category> is the category name estimated by the calorie estimation model. As the inpaint-

ing mask, we use the combination of all the plate and food areas obtained so far. Furthermore, the pre-trained LoRA and Reference-only for preserving the appearance are also used to generate the final calorie-edited image.

4 Experiments

Since both the images in Ege's dataset [2] and the dataset collected in this experiment were collected from the internet, there may be inconsistencies between images and calorie amount information. Therefore, a part of the dataset, UEC-FoodCal (tentative), which was collected independently in collaboration with a company with chefs and registered dietitians, is used for generating calorie-edited images and comparing them with actual calorie amounts.

4.1 Calorie Estimation

Referencing Ege's model [2], we collected recipes and images from online recipe sites for 14 dishes. Using only the collected images this time, we retrained Maeta's model [13] and evaluated these model. Table 1 shows the results of calorie recognition for Maeta's model and the retrained model. The images used for evaluation are 1/5 of each dataset, approximately 1,000 images for Ege's dataset and approximately 7400 images for collected dataset.

Table 1. Calorie Recognition Results. Bold indicates the best accuracy among all data. Red indicates the best performance for each dataset.

		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

The results showed that the models achieved the higher calorie estimation accuracy for the datasets they were trained on. However, considering the larger number of training images, higher image resolution, and more evaluation images, the retrained model in this study is considered to be more effective. For category estimation, the original model seems to be more effective. However, this may be because the data used in Ege's dataset was well-curated from 7 recipe sites, resulting in similar recipes and photography styles. In this study, we decided to generate images with adjusted calorie amounts using the retrained model, which was trained on the larger dataset.

4.2 Calorie-Aware Edited Food Images

We present the results of the edited images when generating various sizes in Figure 7. For all the images, the categories are estimated correctly, and the

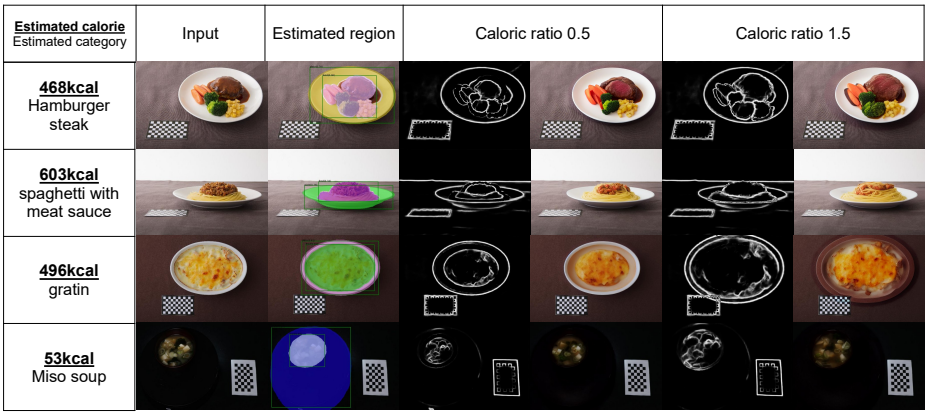

Fig. 7. Results of edited images. The calorie amount estimated by the calorie estimation model and the category are described to the left of the input image.

calorie values are also reasonably estimated. Moreover, the segmentation map is obtained, the SoftEdge image is edited, and it can be seen that the food area of the generated image is changed. In the top row, like the hamburger steak, it can be seen that if there is a word for another food, such as "steak", it may deviate from the original image. However, the side dishes relatively maintain their shape and color, and the amount of food itself is also changing. Additionally, miso soup images in the bottom row, a significant area is dyed with a deep black, and it can be seen that in some cases, the contour of the plate cannot be captured in the SoftEdge image. Even in such cases, changes in the amount of food can be seen.

Furthermore, Figure 8 shows the modified images when the size of the meal is changed to various sizes with the seed value of the Stable Diffusion set to 0.

Comparing the original image with the generated image, although it is successful in enlarging the food area of the generated image, it sometimes fails to reduce it. When increasing the size of the meal, the larger food area than we expect is generated. However, we can obtain some reduced amount of food images by generating multiple candidates with different seed values. (See the Appendix for details.) Furthermore, the ratio of the food area to the overall resolution is similar to the ideal ratio of the food area to the input image, except when the calorie amount is reduced to 0.3 times. Therefore, we can say that the food amount change is also successful.

4.3 Comparison with Real Image

Here, we show how much difference there is from the actual amount of food by comparing generated images with real amount-changed images. Figure 9 shows the real and generated mages. Note that when the original images are compared with each other, the size of the plate and the angle of view are different depending

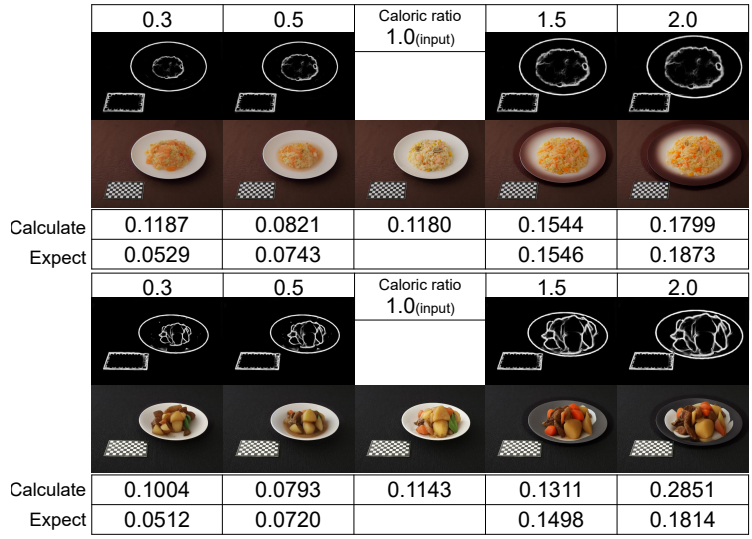

Fig. 8. Calorie-modified images when changing to various sizes. "Calculate" represents the ratio of the food area estimated by Grounded-SAM to the entire edited image. "Expect" indicates the ratio of the ideal food area to the entire input image.

on the size of the meal. In addition, when the ratio of the meal area detected by Grounded-SAM between the input image and the converted image is regarded as the area ratio, the volume ratio is regarded as the calorie ratio, and the calculated value of "(the food area ratio to the entire screen of the edited image / the food area ratio to the entire screen of the input image) $^{(\frac{3}{2})}$" is regarded as the calorie ratio by the food area.

Comparing the real image and the generated image, we can say that although the shapes are different, the visual amount of food is similar. It is not possible to simply compare the real images with the generated images because plates are different. However, by comparing the generated image with the input image, the size can be recognized, which can be said to be a strength of this method. Looking at the calorie ratio by food area, when it was edited with the calorie ratio of 0.5, the value was far from the designated calorie amount ratio, but when it was edited with the calorie ratio of 1.5, the value was close to the designated calorie amount ratio.

4.4 Quantitative Evaluation

In the food image modification considering calorie amounts, it is important to estimate the calorie amount and to actually reduce or enlarge the meal. Therefore, There is no fixed quantitative evaluation for this task.

In the image generation in consideration of the calorie amount, it is important how much the meal area is actually increased or decreased according to the

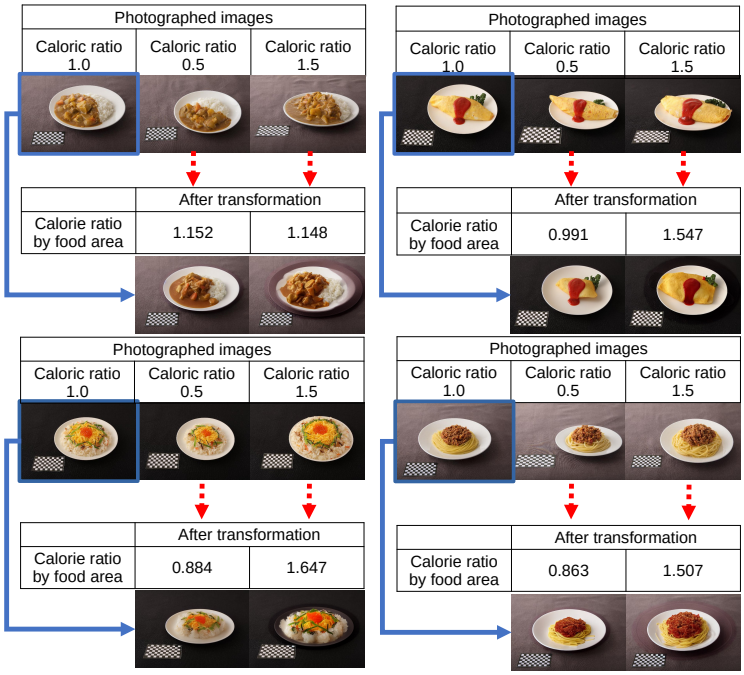

Fig. 9. Comparison of the real amount-changed image with the generated image. The lower part shows the output image when the calorie amount (0.5 times or 1.5 times the reference calorie amount) in the same column is input with the image in the upper left of each category as the input. Note that the size of the plate and the angle of view of the photographed image differ depending on the size of the meal. In addition taken Images at $\times 0.5$ and $\times 1.5$ are reference images for comparison and are not used for processing.

desired calorie amount. Therefore, in this subsection the evaluation is made by measuring the ratio of calories by meal area of the input image.

In case of x0.5, the calorie ratio should be 0.5 in the ideal case, while it should be 1.5 in case of x1.5. However, there are differences depending on the category, such as curry, stir-fried noodles, and pilaf, which showed appropriate values, and hamburger steak, miso soup, and omurice, which had excessively high calorie ratios. Regarding the average value over all the categories, we can see that it often takes a slightly larger calorie ratio than groundtruth value, 0.5 or 1.5. In addition, regarding the deviation, it can be seen that the values of hamburger steak and miso soup are scattered.

5 Conclusion

In this paper, we proposed a metho for food image editing considering calorie amounts. By using Maeta's model for calorie recognition and collecting a

Table 2. Calorie ratio of food regions for 100 generated images for each category and calorie ratio. Values represent mean ± standard deviation.

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

new dataset of images with calorie information, we were able to make the estimated calorie amount more accurate. Additionally, we used Grounded-SAM to detect the food area and adjusted the detected segmentation map. By editing the SoftEdge image, we generated food images considering arbitrary desired calorie amounts. Furthermore, by performing inpainting using these data, we were able to generate food images with changed calorie amounts. In quantitative evaluation, it was confirmed that the size change was appropriately performed based on the region.

Future challenges include discovering a powerful size changing mechanism, and developing a calorie estimator that considers the actual size or increases the number of food types. Although in this study we used ControlNet with a SoftEdge image as a size changing mechanism, we believe that intermediate features like Reference-only and ControlNet using a segmentation map of the food area may be effective. Furthermore, since each model in the mechanism of this research can be replaced, if a better model is available, it is possible to increase the number of estimated categories and improve the accuracy of calorie estimation. Therefore, a more versatile model can be created by constructing a upgraded calorie estimator applicable to a wider variety of foods.

We believe that calorie estimation and the generation of food images based on calorie amounts are effective for health management and adjusting food intake. We hope that our work stimulates new food computing researches on calorie-aware food image synthesis and modification.

Acknowledgments. This work was supported by JSPS KAKENHI Grant Numbers, 22H00540 and 22H00548.

References

1. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: arXiv preprint arXiv:2010.11929 (2020)
2. Ege, T., Yanai, K.: Image-based food calorie estimation using knowledge on food categories, ingredients and cooking directions. In: Proc.of ACM International Conference on Multimedia. pp. 367–375 (2017)
3. Ege, T., Yanai, K.: Simultaneous estimation of food categories and calories with multi-task cnn. In: Proc.of IAPR International Conference on Machine Vision Applications (MVA). pp. 198–201 (2017). https://doi.org/10.23919/MVA.2017.7986835
4. Han, F., Guerrero, R., Pavlovic, V.: CookGAN: Meal Image Synthesis from Ingredients. In: Proc. of IEEE/CFV Winter Conference on Applications of Computer Vision (2020)
5. Horita, D., Tanno, R., Shimoda, W., Yanai, K.: Food category transfer with conditional cyclegan and a large-scale food image dataset. In: Proc.of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management. pp. 67–70 (2018)
6. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: arXiv preprint arXiv:2106.09685 (2021)
7. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: Proc.of IEEE International Conference on Computer Vision. pp. 1780–1790 (2021)
8. Kim, J.h., Lee, D.s., Kwon, S.k.: Food classification and meal intake amount estimation through deep learning. Applied Sciences **13**, 5742 (05 2023). https://doi.org/10.3390/app13095742
9. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.B.: Segment Anything. In: Proc.of IEEE International Conference on Computer Vision. pp. 3992–4003 (2023), https://api.semanticscholar.org/CorpusID:257952310
10. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., et al.: Grounded language-image pre-training. In: Proc.of IEEE Computer Vision and Pattern Recognition. pp. 10965–10975 (2022)
11. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding DINO: Marrying dino with grounded pre-training for open-set object detection. In: arXiv preprint arXiv:2303.05499 (2023)
12. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proc.of IEEE International Conference on Computer Vision. pp. 10012–10022 (2021)
13. Maete, K.: Estimating calorie content from meal images using vision transformer. In: Master's Thesis, The University of Electro-Communications (2023)
14. Nakano, K., Horita, D., Sakata, N., Kiyokawa, K., Yanai, K., Narumi, T.: Enchanting your noodles: Gan-based real-time food-to-food translation and its impact on vision-induced gustatory manipulation. In: Proc.of IEEE Conference on Virtual Reality and 3D User Interfaces. pp. 1096–1097 (2019)
15. Naritomi, S., Yanai, K.: CalorieCaptorGlass: Food calorie estimation based on actual size using hololens and deep learning. In: Proc.of IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops. pp. 818–819 (2020)

16. Pan, S., Dai, L., Hou, X., Li, H., Sheng, B.: ChefGAN: Food image generation from recipes. In: Proc.of ACM International Conference on Multimedia. p. 4244–4252 (2020)
17. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proc.of IEEE Computer Vision and Pattern Recognition. pp. 10684–10695 (2022)
18. Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: Proc.of IEEE International Conference on Computer Vision. pp. 8429–8438 (2019). https://doi.org/10.1109/ICCV.2019.00852
19. Shi, Y., Xue, C., Pan, J., Zhang, W., Tan, V.Y., Bai, S.: DragDiffusion: Harnessing diffusion models for interactive point-based image editing. In: arXiv preprint arXiv:2306.14435 (2023)
20. Tanno, R., Horita, D., Shimoda, W., Yanai, K.: Magical rice bowl: A real-time food category changer. In: Proc.of ACM International Conference on Multimedia. pp. 1244–1246 (2018)
21. Terauchi, K., Yanai, K.: CalorieCam360: Simultaneous eating action recognition of multiple people using an omnidirectional camera. In: Proc.of ACM International Conference on Multimedia Retrieval. pp. 644–648 (2023)
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proc.of Neural Information Processing Systems. vol. 30 (2017)
23. Xie, S., Tu, Z.: Holistically-nested edge detection. In: Proc.of IEEE International Conference on Computer Vision. pp. 1395–1403 (2015)
24. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proc.of IEEE International Conference on Computer Vision (2023)
25. Zhu, B., Ngo, C.W.: CookGAN: Causality Based Text-to-Image Synthesis. In: Proc.of IEEE Computer Vision and Pattern Recognition (2020)
26. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proc.of IEEE Computer Vision and Pattern Recognition. pp. 2223–2232 (2017)