Mask-based Food Image Synthesis with Cross-Modal Recipe Embeddings

Zhongtao Chen
Univ. of Electro-Communications
Tokyo, Japan
chen-z@mm.inf.uec.ac.jp

Yuma Honbu
Univ. of Electro-Communications
Tokyo, Japan
honbu-y@mm.inf.uec.ac.jp

Keiji Yanai
Univ. of Electro-Communications
Tokyo, Japan
yanai@cs.uec.ac.jp

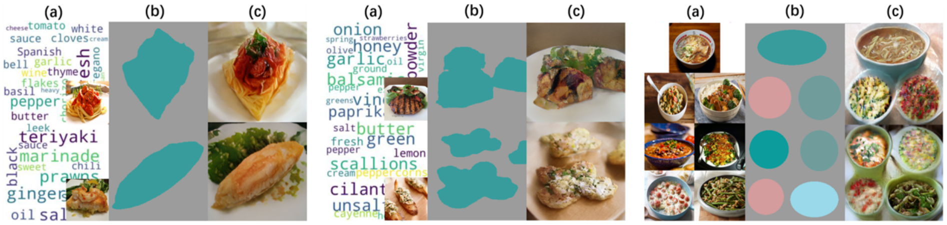

Figure 1: Example images synthesized by MRE-GAN from textual recipe embeddings with arbitrary mask shapes, (a) input recipe texts with the corresponding images, (b) input shape masks, and (c) generated images.

ABSTRACT

In this paper, we propose a Mask-based Recipe Embedding GAN (MRE-GAN), which enables us to generate a realistic food image based on a given mask image containing single or multiple food regions with cross-modal recipe embeddings for each food region. Thus, we can change meal shapes by modifying mask images, while by editing recipe text, we can change meal appearance. Our experimental findings confirmed that the proposed method could generate higher quality food images than the baselines, and we could change meal shapes and appearances by editing mask images and recipe texts as we liked.

CCS CONCEPTS

• **Computing methodologies** → **Computer vision**.

KEYWORDS

food image synthesis, mask-based image generation, cross-modal recipe embedding

ACM Reference Format:
Zhongtao Chen, Yuma Honbu, and Keiji Yanai. 2023. Mask-based Food Image Synthesis with Cross-Modal Recipe Embeddings. In *ACM Multimedia Asia 2023 (MMAsia '23), December 6–8, 2023, Tainan, Taiwan.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3595916.3626418

1 INTRODUCTION

With the development of deep learning technology, technologies of image synthesis and translation have been widely researched in recent years. The image synthesis technology based on GAN [6] is rapidly evolving, and the generation of a realistic image has been achieved. However, generating realistic food images remains a challenging task because the appearance, particularly the shape of food preparations significantly varies depending on the ingredients and instructions specified in a recipe. At the same time, with increasing interest in health and diet, research on dietary image recognition has received considerable attention. In particular, with the release of the large food dataset Recipe1M [13], cross-modal recipe retrieval between food images and cooking recipes has been actively studied. Cross-modal recipe retrieval realizes highly accurate recipe search by projecting image and text features in the same embedded space to learn modality-invariant representations.

In this paper, we propose a Mask-based Recipe Embedding GAN (MRE-GAN) that can generate a realistic food image based on a given mask image containing single or multiple food regions with cross-modal recipe embeddings for each food region. Thus, we can change food shapes by modifying mask images and food appearance by editing recipe texts or changing input food images. Figure 1 shows some generated images by the proposed method. We expect that our work helps and promotes new food-related applications such as an interactive multiple-dish image simulation system which enables us to obtain completed dish photos of newly-invented recipes before trying actual cooking.

The proposed GAN was inspired by the idea of Cross-Modal Recipe Embeddings by Disentangling Recipe Contents and Dish Styles (RDE-GAN) [15] in which recipe embeddings and shape features were disentangled when extracting image features. RDE-GAN could generate high-quality cross-modal recipe embeddings and food images by integrating recipe embeddings and food shape features. However, in the RDE-GAN, the disentangled food shape

MMAsia '23, December 6–8, 2023, Tainan, Taiwan

ACM ISBN 979-8-4007-0205-1/23/12…$15.00
https://doi.org/10.1145/3595916.3626418

features cannot be directly edited, and the shape of the generated images cannot be manually modified. Therefore, herein, we introduce region-based image synthesis into cross-modal embedding-based food image synthesis to generate food images based on arbitrary shape masks with cross-modal recipe embedding. No recipe-based image synthesis methods have been able to generate multiple-dish images thus far. To the best of our knowledge, we believe this is the first framework that generates food images using a shape mask image and cross-modal recipe embedding. As cross-modal embeddings can be extracted from either images or texts, we can generate food images from either recipe text embeddings or recipe image embeddings with a food region mask. We confirmed through comprehensive experiments, that our proposed method could generate higher quality food images even with multiple food items from recipe texts or given food images with arbitrary shape masks, as well as higher retrieval accuracy.

To summarize, the contributions of this work are as follows:

(1) We propose an MRE-GAN (Mask-based Recipe Embedding GAN), which is the first work on mask-based food image synthesis using cross-modal recipe embeddings.

(2) Through the extensive experiments, we confirmed that the proposed MRE-GAN could synthesize not only single dish images but also multiple-dish images, and outperformed the baselines regarding both the FID scores and the IS scores.

2 RELATED WORKS

2.1 Cross-Modal Recipe Retrieval

Cross-modal recipe retrieval is a mutual search across modalities from an image to recipe texts and from a recipe text to images. In the early study, Salvador *et al.* created a large dataset, Recipe1M. They proposed a collaborative, embedded learning approach, Joint Embedding (JE) [13], combining pairwise cosine loss and semantic regularization constraints. When performing general cross-modal joint embedding learning, a CNN is commonly used for encoding an image into a semantic vector of the image. Conversely, for recipe texts, an ingredient list and cooking instructions are encoded by the bidirectional LSTMs.

As an improved approach to JE [13], a Stack Attention Network (SAN) [2] was proposed to identify food areas in an image to learn joint embedding features. Later, some improvements have enhanced the performance by replacing the sine loss optimizer. AMSR [3] used the recipe's hierarchical attention with a simple triplet loss. AdaMine [1] batched all triplet losses in both joint latent space recipes and image embeddings, leveraging class-guided features. Additionally, the recent Modality-Consistent Embedding Network (MCEN) [5] has simplified the training and inference steps and introduced a task-specific encoder for text recipes based on hierarchical attention.

2.2 Recipe-to-Image Synthesis

Generating images from recipe texts is an inherently difficult task. Some recent works on cross-modal recipe retrieval [7, 15, 17, 20] integrated cross-modal recipe embeddings with food image synthesis, enabling us to generate food images from either text recipe embeddings or image embeddings.

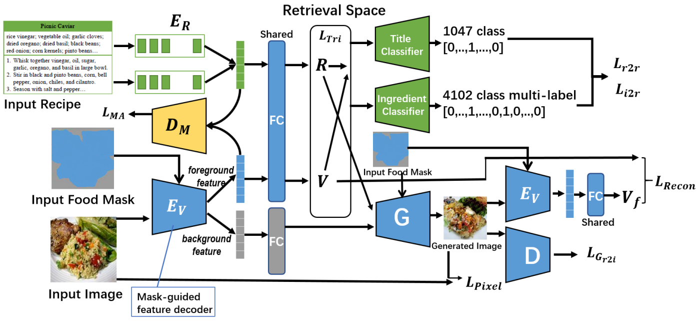
Picnic Caviar
rice vinegar; vegetable oil; garlic cloves; dried oregano; dried basil; black beans; red onion; corn kernels; pinto beans…
1. Whisk together vinegar, oil, sugar, garlic, oregano, and basil in large bowl.
2. Stir in black and pinto beans, corn, bell pepper, onion, chiles, and cilantro.
3. Season with salt and pepper…
Input Recipe
E_R
Retrieval Space
Shared
FC
L_{Tri}
R
V
Title Classifier
Ingredient Classifier
1047 class [0,..,1,...,0]
4102 class multi-label [0,..,1,...,0,1,0,..,0]
L_{r2r}
L_{i2r}
L_{MA}
D_M
Input Food Mask
foreground feature
E_V
background feature
FC
Input Image
Mask-guided feature decoder
Input Food Mask
G
Generated Image
E_V
FC
Shared
V_f
L_{Recon}
D
$L_{G_{r2i}}$
L_{Pixel}

Figure 2: The architecture of MRE-GAN.

R2GAN [20] and ACME [17] introduced GAN-based image generation in addition to triplet-based joint embedding. R2GAN proposed a two-level triplet ranking loss. The triplet loss was used in the generate image space as well as the shared embedding space. A reconstruct loss was also introduced in the generated image space to make generated images and input dish images identical. Consequently, cross-modal search performance was improved. However, R2GAN intended to improve search performance by introducing image generation rather than generating high-quality dish images. They generated only 64 × 64 images.

ACME [17] performed a reconstruction of ingredients and title category from a visual embedding and a recipe image from a textual embedding. ACME adopted adversarial cross-modal training [16]. So a visual embedding made from a recipe image and a textual embedding made from a recipe text cannot be distinguished from each other. The other basic parts are the same as those of AdaMine [1] and R2GAN [20]. Unlike R2GAN, ACME generated 128 × 128 images.

CookGAN [19] exclusively focused on food image synthesis from recipe texts that evolved R2GAN with the step-by-step up-sampling. Hence, it generated high-resolution food images based on the interaction of ingredients and cooking methods. In addition, CookGAN enabled the retaining of more detailed visual effects in the generated images by exploring the causal relationships contained in the recipe text information. X-MRS [7] further attempted to map the generated images into the joint embedding space, and used this embedding representation for recipe retrieval, thereby improving the accuracy of recipe retrieval.

The most relevant work to ours is RDE-GAN [15], which aimed to improve the visual quality of image generation by disentangling image features into recipe style features and dish shape features. With image feature disentanglement, RDE-GAN improved search accuracy and quality of the generated food images. However, in the RDE-GAN, the disentangled shape features were not directly editable by hand. Therefore, in this paper, we propose introducing mask-based image synthesis into cross-modal embedding-based food image synthesis to generate food images based on arbitrary shape masks with cross-modal recipe embedding.

3 MASK-BASED RECIPE EMBEDDING GAN

3.1 Overview of MRE-GAN

In this study, we propose a Mask-based Recipe Embedding GAN (MRE-GAN) that generates a food image based on a given region mask with recipe embeddings. With MRE-GAN, we can generate various food images including multiple dishes in any layouts as shown in Figure 1.

Figure 2 shows the archtecture of MRE-GAN, in which the proposed model has text and image encoders on the left side, and the parts of text prediction and image reconstruction on the right side. This architecture is based on the idea that embedding only the image feature of a food region into the shared space is expected to improve the quality of the generated image. We extract image features of food regions and a background region separately based on a given shape mask. This is because the feature extracted from the whole food image includes both the food regions and non-food backgrounds such as tableware and a part of a table; hence by providing a mask for the food part, only the food features can be extracted for the training of cross-modal embeddings.

3.2 Training of Cross-modal Embedding

A pair of a food image and the corresponding mask are provided to the mask-based image style encoder, E_V. A foreground food image feature and a background image feature are extracted from both the foreground food masks and background region separately. Each feature is a 1024-dimensional vector. The recipe text encoder, E_R, encodes an ingredient list via a bidirectional LSTM and a cooking instruction via a hierarchical LSTM into a recipe text vector, the same as that done in JE [13].

Next, the food image feature and the recipe text vector are passed through the fully connected layer with the shared weights to correlate the representation of both modalities with each other and embedded in the joint space. The triplet loss [14] which is a typical distance learning loss is used to embed the recipe embedding, R, and the image embedding, V, in the jointly shared space with the hard sample mining trick [8]. When calculating triplet losses, Hermans *et al.* [8] improved performance by choosing the hardest positive and negative samples for each anchor point in each batch. The triplet sample (x_a, x_p, x_n) means that x_a is an anchor point for one modality and is used as ground truth for evaluating the embedding of the corresponding modality. In contrast, x_p and x_n indicate the embedding of positive and negative features from another modality. The triplet loss ensures that a positive instance of one modality is closer to the anchor point of another modality and a negative instance of one modality is away from the anchor point of another modality. The triplet loss is represented by the following equation:

$$\begin{aligned} L_{Tri} &= \sum_V [d(V_a, R_p) - d(V_a, R_n) + \alpha]_+ \\ &\quad + \sum_R [d(R_a, V_p) - d(R_a, V_n) + \alpha]_+ \\ &\quad \text{where } [z]_+ = \max(z, 0). \end{aligned} \tag{1}$$

In addition to triplet loss, to adjust the distribution of the encoded features more, a modality discriminator, D_M, is also adopted such that it cannot be distinguished whether the feature vector is obtained from the image or the text. This idea was originally proposed in [16]. The Modality Alignment loss, L_{MA}, is represented as follows:

$$\begin{aligned} L_{MA} &= E_{i \sim p_{image}}[\log(D_M(E_V(i)))] \\ &\quad + E_{r \sim p_{recipe}}[\log(1 - D_M(E_R(r)))] \end{aligned} \tag{2}$$

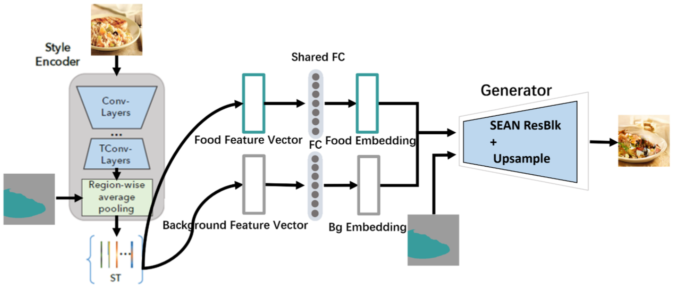

Figure 3: The part of the image encoder and generator.

3.3 Cross-Modal Translation

The previous studies [17, 20] have shown that the learned embedding of one modality allows the corresponding information of the other modality to be recovered, and improves semantic alignment. This has proven to enhance the consistency of the cross-modal transformation and it improves the representation of the learned embedding. Specifically, the recipe embedding, R, generates food images, and the visual embedding, V, is used to predict recipe ingredients. Therefore, our study aims to generate high-quality food images based on a food shape mask and cross-modal embeddings by adding a conditional GAN to the proposed network. We adopt a mask-based image encoder network that simultaneously extracts the corresponding style code from each semantic area of a given image. In our case, the features of semantic areas are extracted from foreground food areas and the background area independently according to the food shape mask corresponding to an input image.

As shown in Figure 3, the input of the input encoder is a set of food image and mask image, and the output of the image encoder is a set of 1024-dimensional feature vectors. Unlike a standard encoder built on a simple down-sampling convolutional neural network, our per-region style encoder employs a "bottleneck" structure to remove the information irrelevant to styles from the input image and obtain high resolution feature maps. In addition, considering that the style does not depend on the shape of the semantic region, the intermediate feature map generated by the network block, Transposed Convolutional Layers (TConv-Layers), is passed to the region-wise average pooling layer and reduced to a 1024-dimensional vector for each of the foreground and background regions.

The cross-modal translation consistency losss, L_{Cross}, is defined as the following formula:

$$L_{Cross} = L_{G_{r2i}} + L_{i2r}, \tag{3}$$

where $L_{G_{r2i}}$ and L_{i2r} represent recipe-to-image and image-to-recipe consistency losses, respectively.

3.3.1 Image Generation from Recipe Embedding and Food Shape Mask.

To generate a food image from the cross-modal recipe embeddings, first, we need to provide a food shape mask, recipe embeddings, and a background image feature to a generator. When the mask contains multiple types of food regions, we need to prepare a recipe embedding vector for each food region. As an architecture of the generator, we adopt a conditional generator employing Semantic REgion-Adaptive Normalization (SEAN) [21]. The modulation parameters are controlled by a region mask, recipe embedding and a background image feature. A conditional normalization technique, SEAN, is used to establish fine control over the style of the image in each semantic region such that we can add the different textures corresponding to recipe embeddings or a background image feature to each of the food regions and a background image feature to the background region.

For training of the image synthesis part, the image encoder is trained to extract region-by-region style codes from the input image according to the corresponding segmentation mask. The generator is trained to reconstruct the corresponding food image by providing the recipe embeddings for food regions and a background feature vector for a background region based on the corresponding semantic mask. Following SPADE [12] and SEAN [21], the input and reconstructed images are evaluated by the loss function, $L_{G_{r2i}}$, consisting of three loss terms: Adversarial loss, Feature matching loss, and Perceptual loss.

The loss function $L_{G_{r2i}}$ is given as follows:

$$L_{G_{r2i}} = \min_{E,G}((\max_{D1,D2} \sum_{k=1,2} L_{GAN}) + \gamma_1 \sum_{k=1,2} L_{FM} + \gamma_2 L_{percept}) \quad (4)$$

We set $\gamma_1 = \gamma_2 = 10$ following SPADE and SEAN.

For Adversarial loss, the formulation of conditional adversarial learning is expressed as follows:

$$\begin{aligned}\min_{E,G} \max_{D_1,D_2} \sum_{k=1,2} L_{GAN}(E,G,D_k) &= E[\max(0, 1 - D_k(R,M))] \\ &+ E[\max(0, 1 + D_k(G(ST,M),M))]\end{aligned} \quad (5)$$

where E is the image encoder, G is the generator employing SEAN, D_1 and D_2 are the two classifiers of different scales, R is the given recipe's image, M is the corresponding segmentation mask of R, and ST is the feature matrix combined with recipe embeddings and background feature vector.

For Feature matching loss, let T be the total number of layers in the discriminator D_k, and let $D_k^{(i)}$ and N_i is the output feature map and the number of elements of the i−th layer of D_k, respectively. Feature Matching loss, L_{FM}, is expressed as follows:

$$L_{FM} = E \sum_{i=1}^{T} \frac{1}{N_i} [\| D_k^{(i)}(R,M) - D_k^{(i)}(G(ST,M),M) \|_1] \quad (6)$$

For Perceptual loss, let N be the total number of layers used to calculate perceptual loss, $F^{(i)}$ be the output feature map of the i-th layer of the VGG network, and M_i be the number of elements of $F^{(i)}$. The perceptual loss, $L_{percept}$, is expressed as follows:

$$L_{percept} = E \sum_{i=1}^{N} \frac{1}{M_i} [\| F^{(i)}(R) - F^{(i)}(G(ST,M)) \|_1] \quad (7)$$

3.3.2 **Text Prediction from Image Embedding.** By applying the multi-label classifier to the visual feature, V, which predicts the components of the food image, image embedding can be classified into the correct food category, thereby maintaining the consistency of the translation. Although adversarial loss can generate a realistic image, the conversion might be inconsistent. Therefore, the generator is appropriate by embedding the generated image in a shared space and predicting the components in that embedding, encouraging generation of a food image in the corresponding food category.

When classifying image embeddings, V, we classify 1047 classes by title and 4102 classes using the multi-label ingredient list. The recipe prediction loss function, L_{i2r}, is expressed as follows:

$$L_{i2r} = L_{Title}(V, GT_{title}) + L_{Ing}(V, GT_{ing}) \quad (8)$$

where cross entropy loss by title classifier and ingredient classifier is L_{Title}, L_{Ing}, the ground-truth label for title classification is GT_{title}, and the ground-truth label for ingredient classification is GT_{ing}.

3.4 Single Modal Translation Consistency

In ACME [17], only cross-modal consistency was considered. In contrast, in our work, we take account of translation consistency within the single modal as well. The single modal translation consistency loss can be defined as follow:

$$L_{Single} = L_{r2r} + L_{i2i} \quad (9)$$

In the same way as L_{i2r}, text embeddings are classified by the title and the multi-label ingredient list. The recipe-to-recipe consistency loss, L_{r2r}, is represented as follows:

$$L_{r2r} = L_{Title}(T, GT_{title}) + L_{Ing}(T, GT_{ing}) \quad (10)$$

The image-to-image consistency loss, L_{i2i}, enforces that the image embedding of a generate image, V_f, is identical to the image embedding of the original image, V, and the generated image, I_f, looks the similar as the original image, I. The loss can be defined as follows:

$$L_{i2i} = L_{Recon}(V_f, V) + L_{Pixel}(I_f, I), \quad (11)$$

where L_{Recon} and L_{Pixel} are represented by the L1 loss and the L2 loss (MSE loss), respectively.

3.5 Total Loss

As mentioned above, we use the four losses, Triplet loss, L_{Tri}, Modality Adversalial loss, L_{MA}, Cross-Modal loss, L_{Cross}, and Single Modal loss, L_{Single}. The total objective function can be defined as follows:

$$L_{Total} = \lambda_1 L_{Tri} + \lambda_2 L_{MA} + \lambda_3 L_{Cross} + \lambda_4 L_{Single}, \quad (12)$$

where $\lambda_{1,\dots,4}$ represents the loss weight of each loss function.

4 EXPERIMENTS

4.1 Datasets, Metrics and Implementation

Dataset: All the experiments were conducted using the standard large-scale recipe dataset, Recipe1M [13], containing over 1 million recipes and images. We adopt the original data splits using 238,999 image-recipe pairs for training, 51,119 pairs for validation, and 51,303 pairs for testing. As the original Recipe1M has no region mask data, the food shape masks were automatically generated for all the images of Recipe1M using the method of "unseen food image segmentation" [10] which employed zero-shot segmentation [9] with the food image segmentation dataset, UEC-FoodPix Complete [11] as the base set. The detail is explained in the supplementary material. In addition, DeepLabV3+ [4] trained with UEC-FoodPix Complete was also used to generate food masks for comparison. The food masks generated by the unseen food method

Table 1: Comparison of image quality by the FID score (↓) and the IS score (↑).

Method	Text2Img(FID↓)	Img2Img(FID↓)	Img2Img(IS↑)
ACME[17]	390.52	391.29	2.19±.09
RDE-GAN[15]	83.82	84.31	6.99±.07
CookGAN[19]	–	–	5.41±.11
X-MRS[7]	28.60	27.90	–
Ours ($\text{Mask}_{\text{DeepLabV3+}}$)	56.72	56.11	–
Ours ($\text{Mask}_{\text{unseen}}$)	**27.44**	**27.12**	**8.27±.05**

achieved 73.0% MIoU (evaluated by 691 ground-truth masks included in the FoodSeg103 dataset [18]), while the food masks generated by the trained DeepLabV3+ achieved 54.1% MIoU. This indicated that the food masks generated by the unseen food method were much better.

Metrics: The Fréchet inception distance (FID) is a metric used to assess the quality of images created by a generative model. A lower value of FID indicates better visual diversity and quality. We randomly sample 1,000 recipes from the test set for image generation.

Implementation Details: In the experiments, we used the Adam optimizer for training the whole network with an initial learning rate of 10^{-4}, which was decreased after 50 epochs to 10^{-5}. The proposed model was trained with a total of 100 epochs. As the loss weights, we set $\lambda_1 = 1.0, \lambda_2 = 0.005, \lambda_3 = 0.002$, and $\lambda_4 = 0.002$, respectively. We empirically decided the value of each of the loss weights by referring the weights used in RDE-GAN [15] and ACME [17]. Basically, we adjusted the weighting constants so that each of the losses affected to the total loss equally. The resolution of the synthesized images is 256 × 256.

4.2 Evaluation of image generation

4.2.1 **Quantitative Results.** We show the comparison of MRE-GAN with the baselines on the quality of generated images in Table 1. As the baselines, we used ACME [17], RDE-GAN [15], CookGAN [19] and X-MRS [7]. In order to quantitatively evaluate the quality of generated image, we use FID to measure how close the distribution of original image and generated image is. In addition, we used Inception Score (IS) for comparison to CookGAN since the results of CookGAN were evaluated by only IS in the paper [19].

We trained two models, for training Ours ($\text{Mask}_{\text{DeepLabV3+}}$) of the proposed model, the shape mask of Recipe1M was calculated by the DeepLabV3+ model of pre-trained. For training Ours ($\text{Mask}_{\text{unseen}}$) of proposed model, the shape mask of Recipe1M was calculated by "unseen food image segmentation" [10].

From these results, our method clearly achieved the best quality over all the baselines including X-MRS which was the current SOTA. In addition, the results with the masks generated by DeepLabV3+ were much degraded from Ours ($\text{Mask}_{\text{unseen}}$), which reflected the segmentation accuracy (Unseen 73.0 vs DeepLabV3+ 54.1 in MIoU). Using more accurate masks is important for generating the higher-quality images.

4.2.2 **Ablation Studies.** We made the ablation studies on the losses related to the single modal translation consistency, since the effectiveness of the cross-modal consistency was confirmed in the existing works [15, 17]. Note that for all the ablation studies, we used the masks generated by the unseen food image segmentation method [10].

Table 2: Ablation studies.

Method	Text2Img(FID↓)	Img2Img(FID↓)
Without L_{i2i}, L_{r2r}	40.92	42.66
Without L_{i2i}	42.63	45.25
Without L_{r2r}	43.53	43.61
Without L_{pixel}	31.34	34.51
ALL($\text{Mask}_{\text{unseen}}$)	27.44	27.12

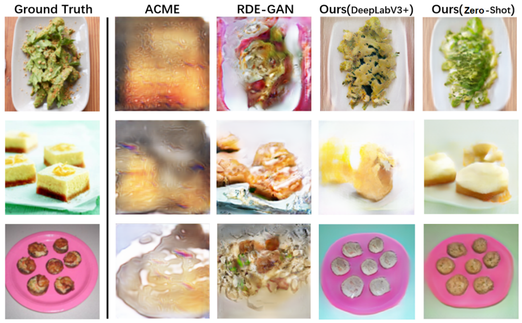

Figure 4: Comparison of synthesized images.

Table 2 shows the results, which indicates that the results without both or either L_{i2i} or L_{r2r} were degraded from the results with the full loss. Both the losses including L_{Pixel} were proven to be helpful for improvement of the image quality.

4.2.3 **Qualitative Results.** A comparative experiment, as shown in Figure 4, was performed to evaluate the visual quality of the generated image with the baseline methods quantitatively. The first column is ground truth corresponding to the recipe embedding used for image generation, and the second and the third column are the images generated by the baseline method ACME [17] and RDE-GAN [15]. The two columns from the right are images generated by the proposed model trained on different shape masks data. Compared with the baselines, it is shown that our proposed method can generate a realistic food image while maintaining the shape of the food. Moreover, the generated images with zero-shot food masks look more similar to the ground truth images than those with the standard segmentation, DeepLabV3+, masks.

The proposed model can generate images from either recipe text embeddings or visual image embeddings. To verify how different the images generated from the corresponding recipe and image embeddings, we generated food images from both embeddings as shown in Figure 5. Comparing the generated images in the third and fourth columns, it is considered that the images generated by text embeddings and the corresponding image embeddings are almost the same. This means that the proposed method can embed the corresponding recipe text and food image into almost the identical point in the shared space.

4.2.4 **Food Image Generation Manipulability.** The results of generating food images with multiple dishes are shown in Figure 6. They are generated with the embeddings from recipe style 1 as the style of green region and the embedding from recipe style 2 as the style of pink region. Note that tag cloud images in the row of style 1 and style2 means that recipe text embeddings were used as food embeddings for mask-based food image synthesis.

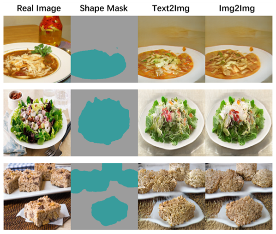

Figure 5: Food image generation from recipe text embeddings and image embeddings.

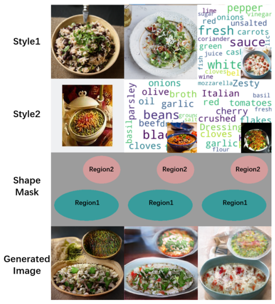

Figure 6: Multiple-dish food image generation.

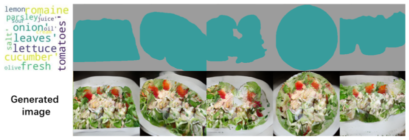

Figure 7: Food image generation with different shape masks.

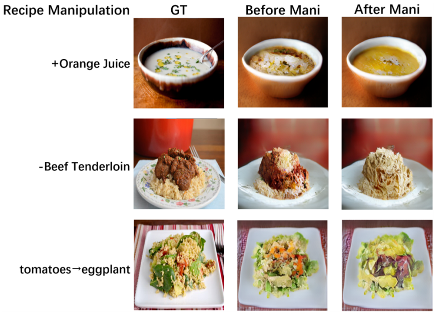

Figure 8: Changes in the generated image due to the operation of the input text.

As shown in the rightmost row of Figure 1, combining three or four dishes is also possible. As shown in the second column of Figure 6, we can use both recipe embedding and image embedding for different regions. To our best knowledge, this is the first work to generate multiple dish images from either recipe text embedding or food image embedding.

As the next results, we show the images generated from different food shape masks and the fixed recipe embedding in Figure 7. From this result, our proposed model can generate foods of any shapes by changing the shape of the input mask. However, we have a limitation on controlling the shape of the plates. Since food masks do not contain the region of plates, the plates are generated so as to surround the food mask, and the generated plates are sometime distorted.

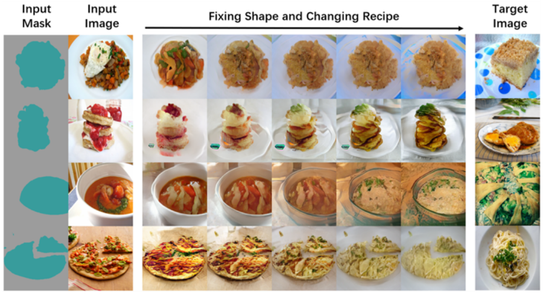

Figure 9: Images generated by gradually changing the embeddings between two target samples via linear interpolation.

Next, we conduct an experiment to verify what type of effect the change in input ingredients has on the generated image. According to Figure 8, when adding, deleting or replacing some ingredients in the recipe texts, the generated food image accordingly changed slightly. For example, in the first column, comparing the generated image after adding orange juice with the generated image before adding the juice, we found that the generated image also changed correspondingly after adding orange juice to the input ingredients. The same holds true for the other two cases where beef tenderloin was removed and tomatoes was replaced with eggplants.

Finally, we verified how the generated image changes when the recipe embedding continuously changes. If the space of the recipe embedding can be continuously expressed, the generated complementary image changes smoothly. This result is shown in Figure 9. It is a shape mask corresponding to the recipe embedding as an input, and is an image shown in the right column as a target. In the third line, it can be observed that the generated image changed smoothly, continuously and simultaneously with the change in the semantic embedding of the elements such as the color and texture of the pizza while maintaining the input shape mask. Thus, the proposed model also provides the expressive power of high-level features required for image generation.

5 CONCLUSIONS

In this study, we have proposed a new framework, Mask-based Recipe Embedding GAN (MRE-GAN), which synthesized food images from cross-modal recipe embeddings based on a given food segmentation mask. To the best of our knowledge, this is the first work on the combination of cross-modal embeddings and mask-based image synthesis, the major advantage of which is that we can generate images by combining a mask drawn by hand and text embeddings for each of the regions. Although MRE-GAN requires food region masks for training, we added food region masks to all the Recipe1M images using "unseen food image segmentation" [10] which enabled us to annotate food region masks to all the training images automatically with high accuracy. Finally, we confirmed through comprehensive experiments, that our proposed method, MRE-GAN, could generate high quality food images even with multiple food items from recipe texts or given food images with arbitrary shape masks, as well as high retrieval accuracy. In addition, we also showed that food image manipulation was clearly possible by changing the shape of a food mask or editing recipe texts.

Acknowledgments: This work was supported by JSPS KAKENHI Grant Numbers, 21H05812, 22H00540, 22H00548, and 22K19808.

REFERENCES

[1] Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. 2018. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In *Proc. of International ACM SIGIR Conference on Research Development in Information Retrieval*. 35–44.

[2] Jingjing Chen, Lei Pang, and Chong-Wah Ngo. 2017. Cross-modal recipe retrieval: How to cook this dish?. In *Proc. of International Conference on Multimedia Modeling*. 588–600.

[3] Jing-Jing Chen, Chong-Wah Ngo, Fu-Li Feng, and Tat-Seng Chua. 2018. Deep understanding of cooking procedure for cross-modal recipe retrieval. In *Proc. of ACM International Conference on Multimedia*. 1020–1028.

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. of European Conference on Computer Vision*. 801–818.

[5] Han Fu, Rui Wu, Chenghao Liu, and Jianling Sun. 2020. MCEN: Bridging Cross-Modal Gap between Cooking Recipes and Dish Images with Latent Variable Model. In *Proc. of IEEE/CVF Computer Vision and Pattern Recognition*.

[6] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, Vol. 27.

[7] Ricardo Guerrero, Hai X Pham, and Vladimir Pavlovic. 2021. Cross-modal Retrieval and Synthesis (X-MRS): Closing the Modality Gap in Shared Subspace Learning. In *Proc. of ACM International Conference on Multimedia*. 3192–3201.

[8] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737* (2017).

[9] Yuma Honbu and Keiji Yanai. 2021. Few-shot and zero-shot semantic segmentation for food images. In *Proc. of International Workshop on Multimedia for Cooking and Eating Activities*. 25–28.

[10] Yuma Honbu and Keiji Yanai. 2022. Unseen Food Segmentation. In *Proc. ACM International Conference on Multimedia Retrieval*. 19–23.

[11] Kaimu Okamoto and Keiji Yanai. 2021. UEC-FoodPix Complete: A Large-scale Food Image Segmentation Dataset. In *Proc. of ICPR Workshop on Multimedia Assisted Dietary Management (MADIMA)*.

[12] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic Image Synthesis With Spatially-Adaptive Normalization. In *Proc. of IEEE/CVF Computer Vision and Pattern Recognition*.

[13] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2017. Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. In *Proc. of IEEE/CVF Computer Vision and Pattern Recognition*.

[14] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A unified embedding for face recognition and clustering. In *Proc. of IEEE/CVF Computer Vision and Pattern Recognition*. 815–823.

[15] Yu Sugiyama and Keiji Yanai. 2021. Cross-Modal Recipe Embeddings by Disentangling Recipe Contents and Dish Styles. In *Proc. of ACM International Conference on Multimedia*.

[16] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial cross-modal retrieval. In *Proc. of ACM International Conference on Multimedia*. 154–162.

[17] Hao Wang, Doyen Sahoo, Chenghao Liu, Ee-peng Lim, and Steven C. H. Hoi. 2019. Learning Cross-Modal Embeddings With Adversarial Networks for Cooking Recipes and Food Images. In *Proc. of IEEE/CVF Computer Vision and Pattern Recognition*.

[18] W. Xiongwei, F. Xin, L. Ying, L. Ee-Peng, H. Steven, and S. Qianru. 2021. A Large-Scale Benchmark for Food Image Segmentation. In *Proc. of ACM International Conference on Multimedia*.

[19] Bin Zhu and Chong-Wah Ngo. 2020. CookGAN: Causality Based Text-to-Image Synthesis. In *Proc. of IEEE/CVF Computer Vision and Pattern Recognition*.

[20] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Yanbin Hao. 2019. R2GAN: Cross-Modal Recipe Retrieval With Generative Adversarial Network. In *Proc. of IEEE/CVF Computer Vision and Pattern Recognition*.

[21] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. 2020. SEAN: Image Synthesis With Semantic Region-Adaptive Normalization. In *Proc. of IEEE/CVF Computer Vision and Pattern Recognition*.