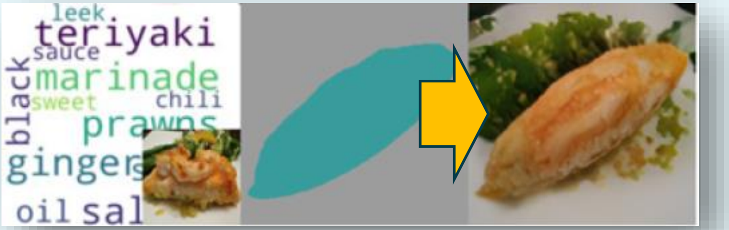
Mask-based Food Image Synthesis with Cross-Modal Recipe Embeddings

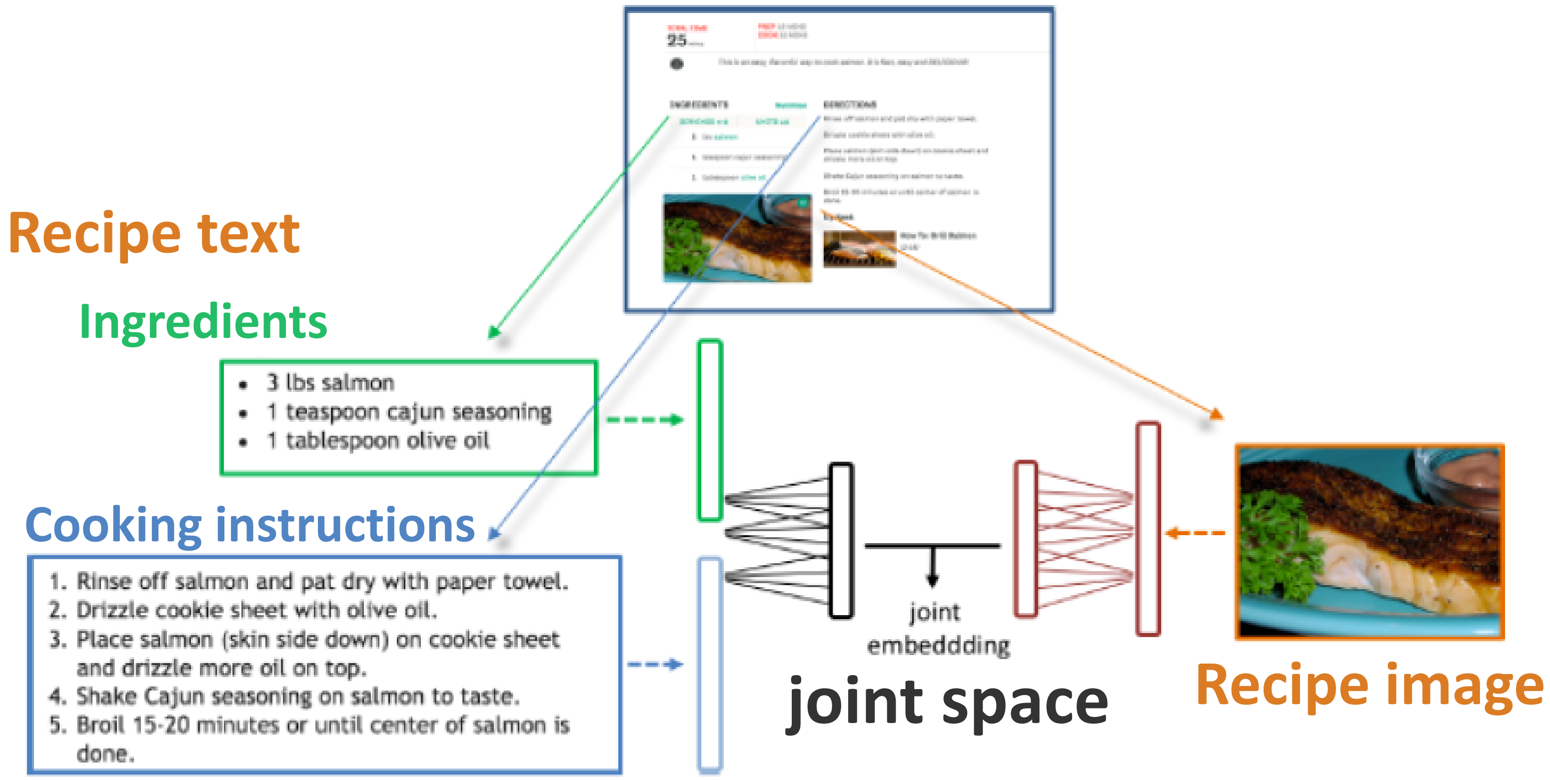


Zhongtao Chen, Yuma Honbu, Keiji Yanai

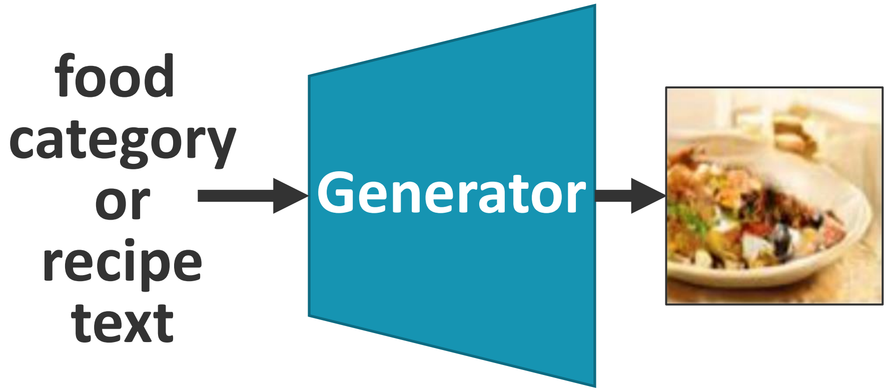
The University of Electro-Communications, Tokyo (UEC)

Background

- Cross-model recipe retrieval and food photo synthesis have drawn much attention in the food multimedia research community.



Cross-model recipe retrieval



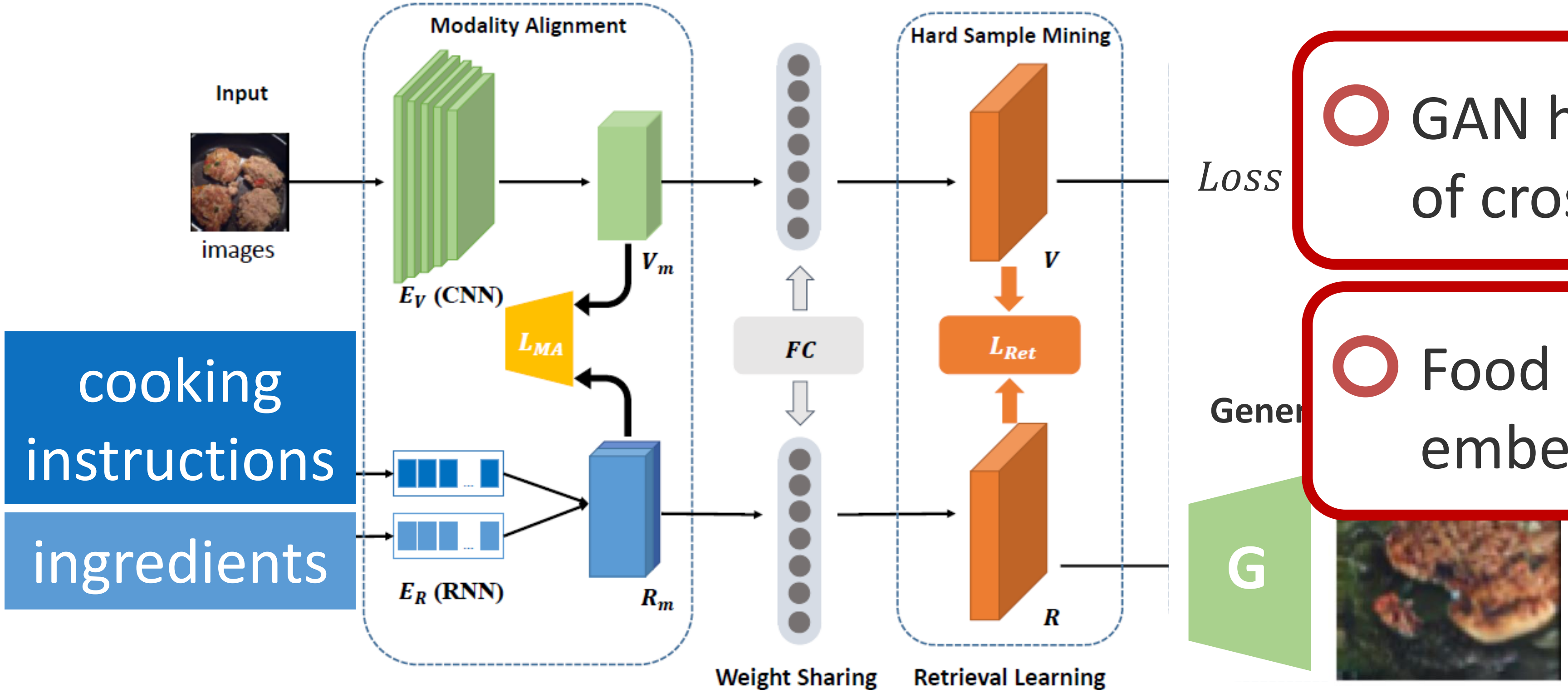
Food Image Synthesis

Cited from “Learning Cross-modal Embeddings for Cooking Recipes and Food Images”

Background: Cross-Modal Recipe Retrieval + GAN

Cross-modal recipe retrieval + GAN

Food Image Generation from cross-model embedding



- GAN has improved accuracy of cross-model recipe search.
- Food image generation from embedding became possible.

Hao Wang, Doyen Sahoo, Chenghao Liu, Ee-peng Lim, and StevenC. H. Hoi:
Learning cross-modal embeddings with adversarial net-works for cooking recipes and food images. CVPR2019

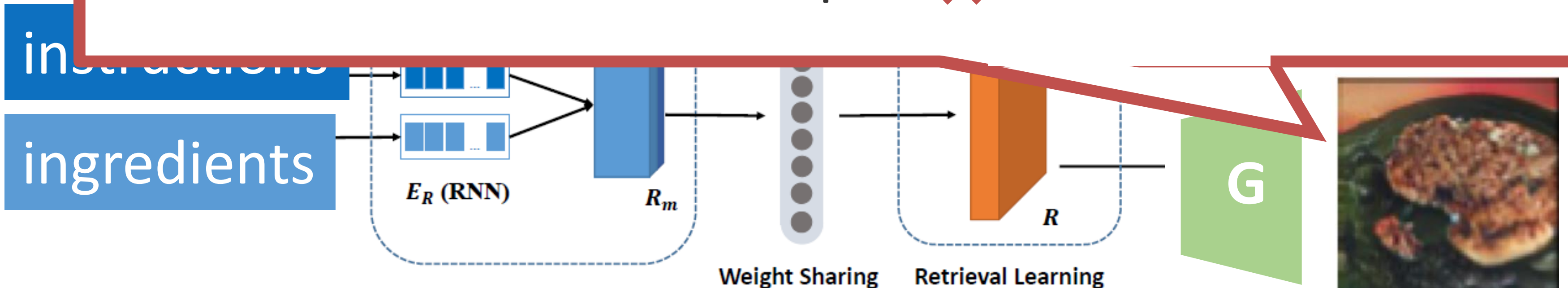
Motivation: We like to control the shape of generated foods img.

Cross-modal recipe retrieval + GAN

<u>Food Image Generation from cross-model embedding</u>

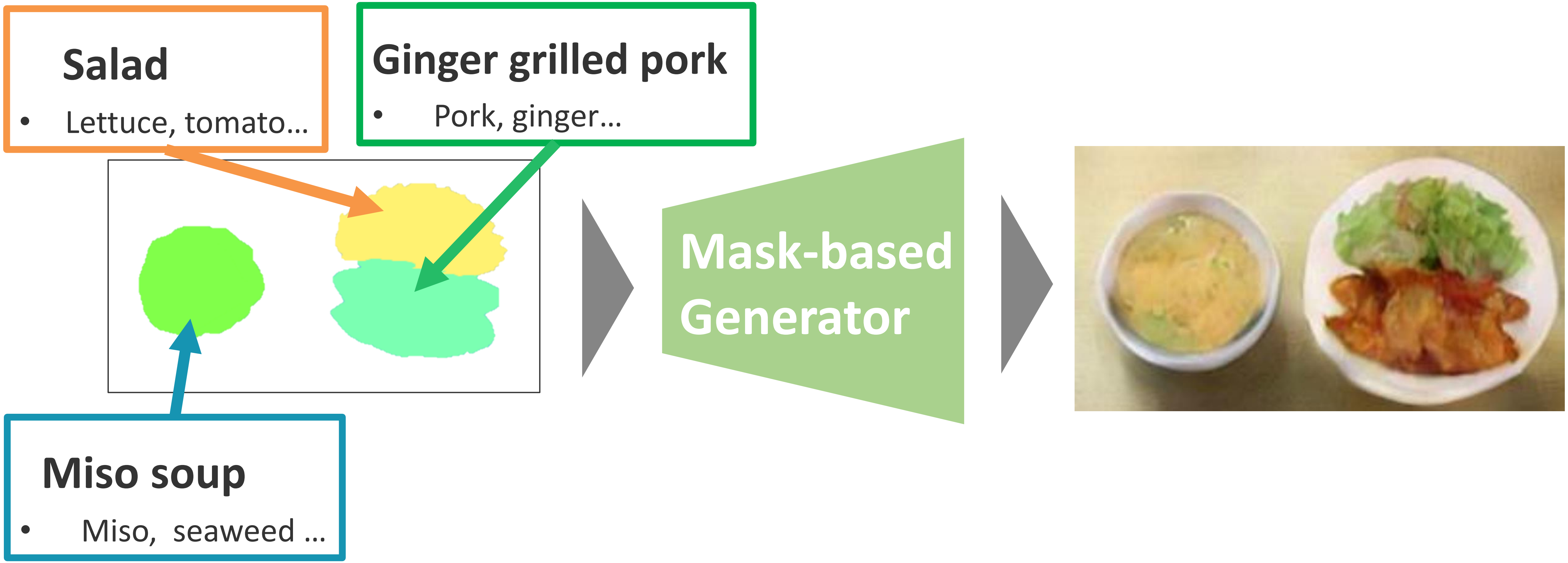
<u>The shape of foods is not controllable.</u>

✖ Preferred shape ✖ Preferred location



Hao Wang, Doyen Sahoo, Chenghao Liu, Ee-peng Lim, and StevenC. H. Hoi: Learning cross-modal embeddings with adversarial net-works for cooking recipes and food images. CVPR2019

Objective: mask-based food image synthesis



We aim to generate high-quality food images from shape masks and recipe information

Related work ① : Adversarial Cross-Modal Embedding (ACME)

ACME [Hao, CVPR2019]

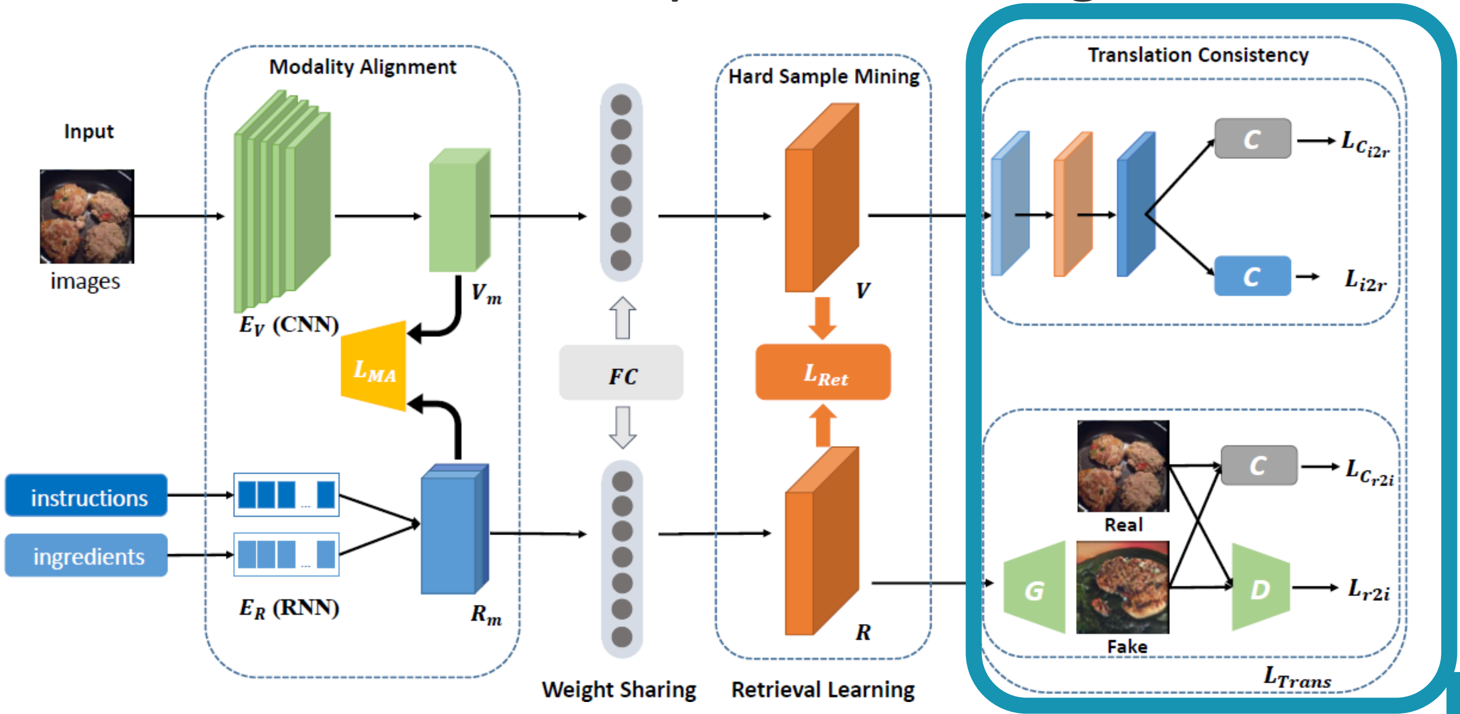
- Is the first work which added a food image generator (GAN) to cross-model recipe search model.

○ improvement of recipe search performance &
food image generation from cross-modal recipe embeddings

✖ low-quality images

Generated Images

Generated Images



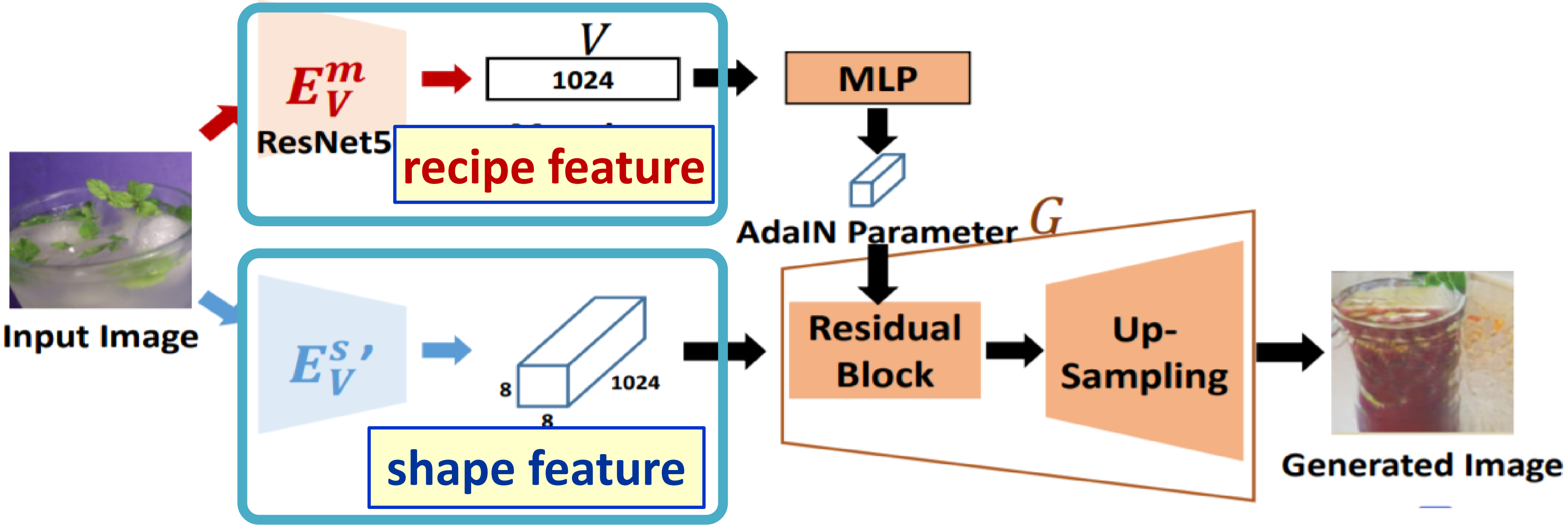
Hao Wang et.al : Learning Cross-Modal Embeddings with Adversarial Networks for Cooking Recipes and Food Images, CVPR 2019

RDE-GAN [Sugiyama and Yanai, ACM Multimedia2021]

Recipe Disentangling Embedding GAN (RDE-GAN) disentangles **recipe information** from **shape information** of recipe images.

○ high performance on cross-modal recipe retrieval and high-quality image generation

✖ instability of training process, and imperfect disentanglement on shape information



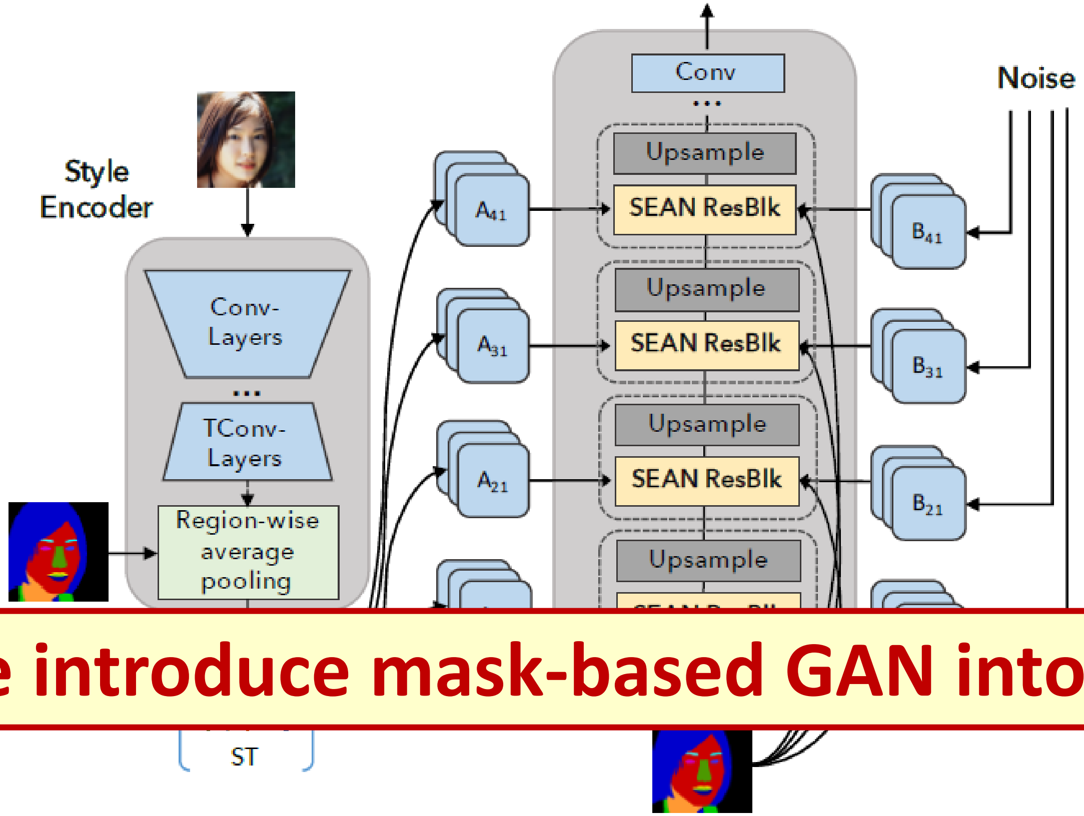
Y. Sugiyama and K. Yanai: Cross-Modal Recipe Embeddings by Disentangling Recipe Contents and Dish Styles, ACM Multimedia 2021

Related work ③: Mask-based image synthesis GAN

SEAN (Semantic region-Adaptive Normalization) [CVPR 2020]

- SEAN can control styles on each semantic mask independently.
  e.g. We can transfer the ramen style to semantic mask images.

style images



We introduce mask-based GAN into cross-modal food image synthesis.

masks

SEAN: Image Synthesis with Semantic Region-Adaptive Normalization, CVPR 2020.

Proposed method: **Overview**

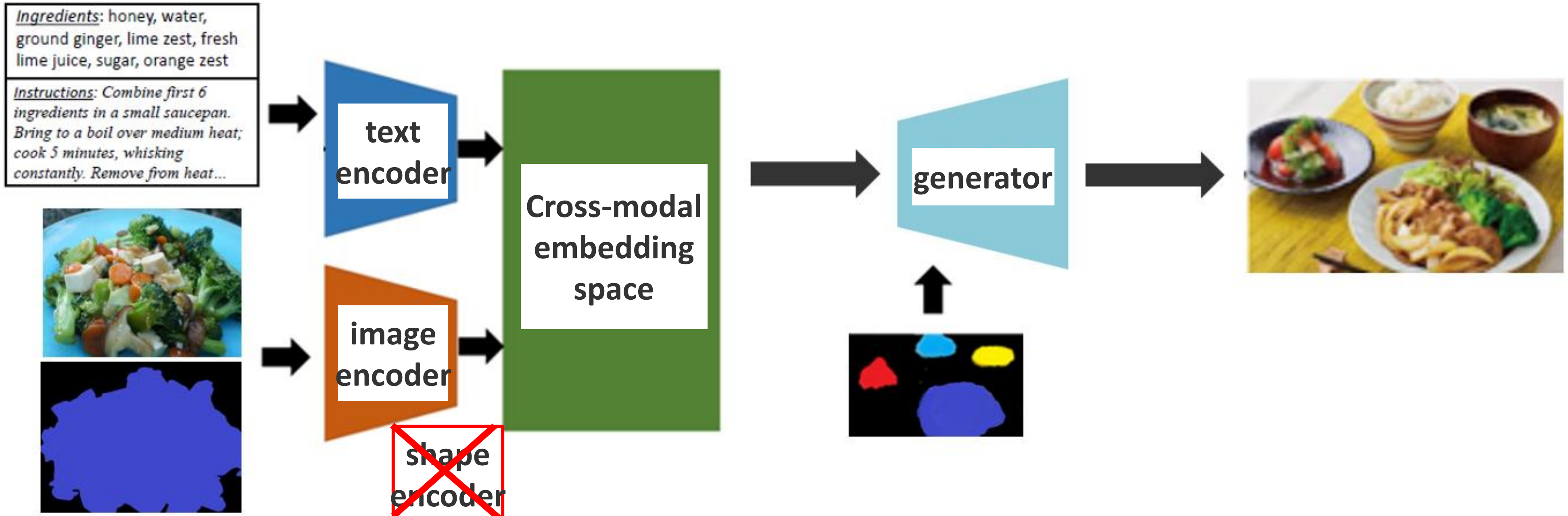
MRE-GAN (Mask-based Recipe Embedding GAN)

[generation time]

Given region masks and recipe embeddings, we can generate multiple-dish food images.

[training time]

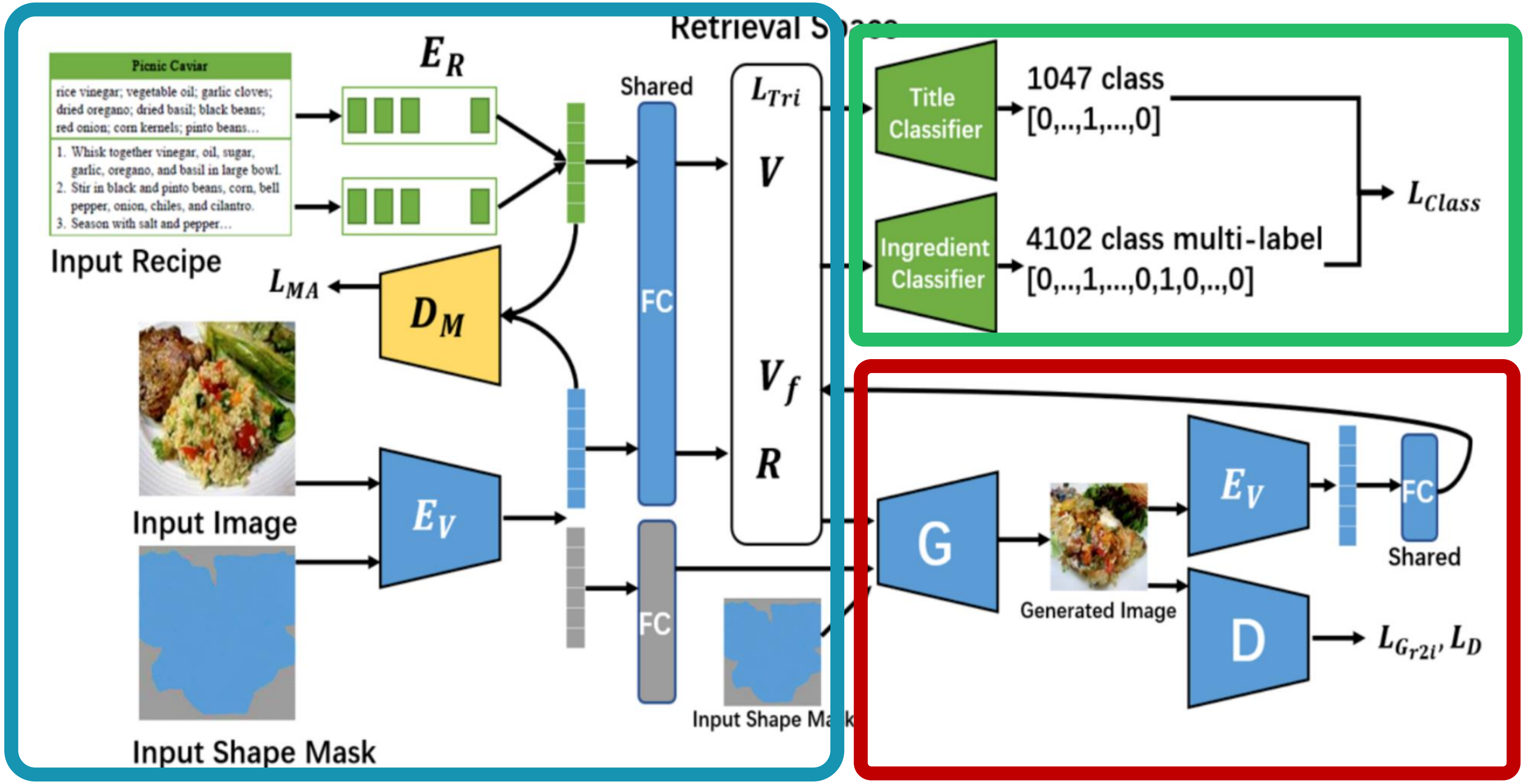
By providing masks for training images, a shape encoder is removed,
which makes training easier than RDE-GAN having both image and shape encoders.



Proposed method ： **architecture**

MRE-GAN (Mask-based Recipe Embedding GAN) consists three parts:

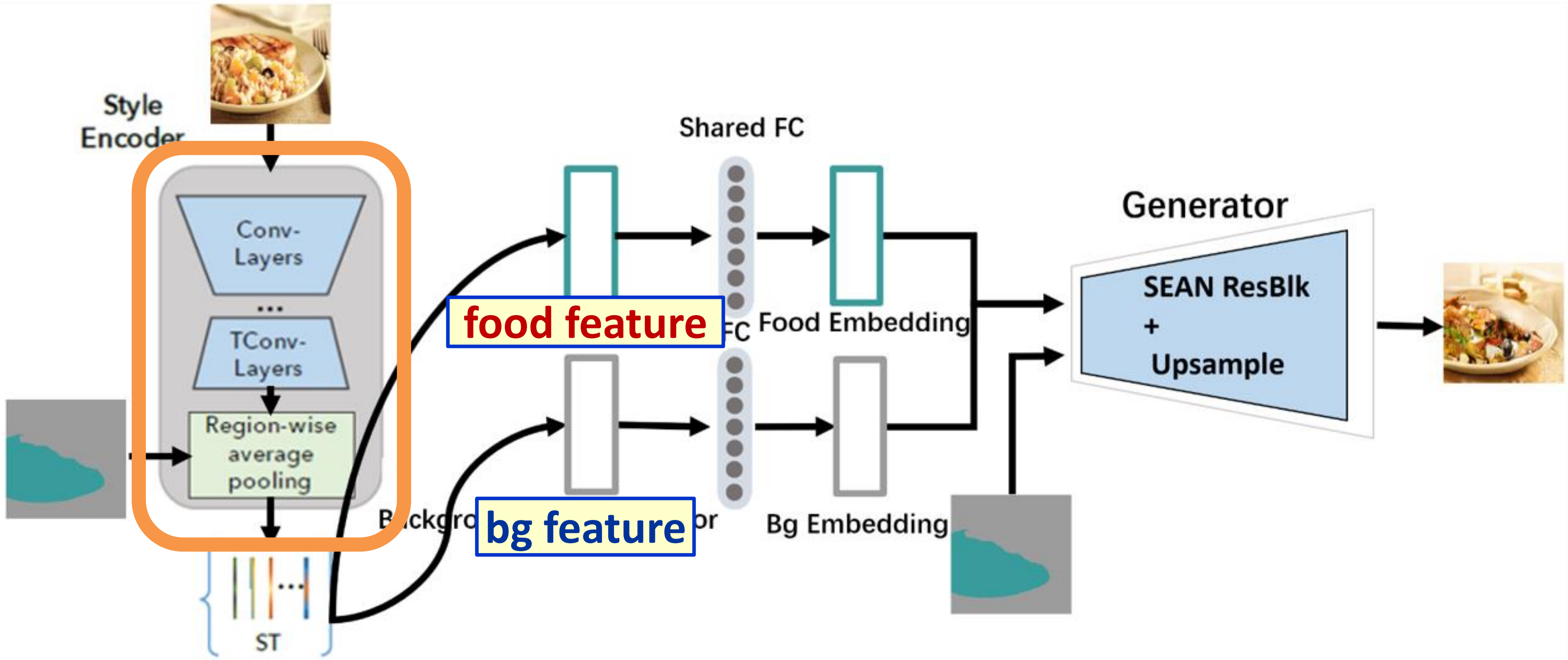
- **Mask-based GAN** + **Cross model recipe embedding** + recipe title/ingredient estimation
- Based on RDE-GAN, we added SEAN and removed a shape encoder to control food shapes.



Proposed method: **Improvement (1) from RDE-GAN**

Improvement ① : **Mask-based image encoder with masked average pooling**

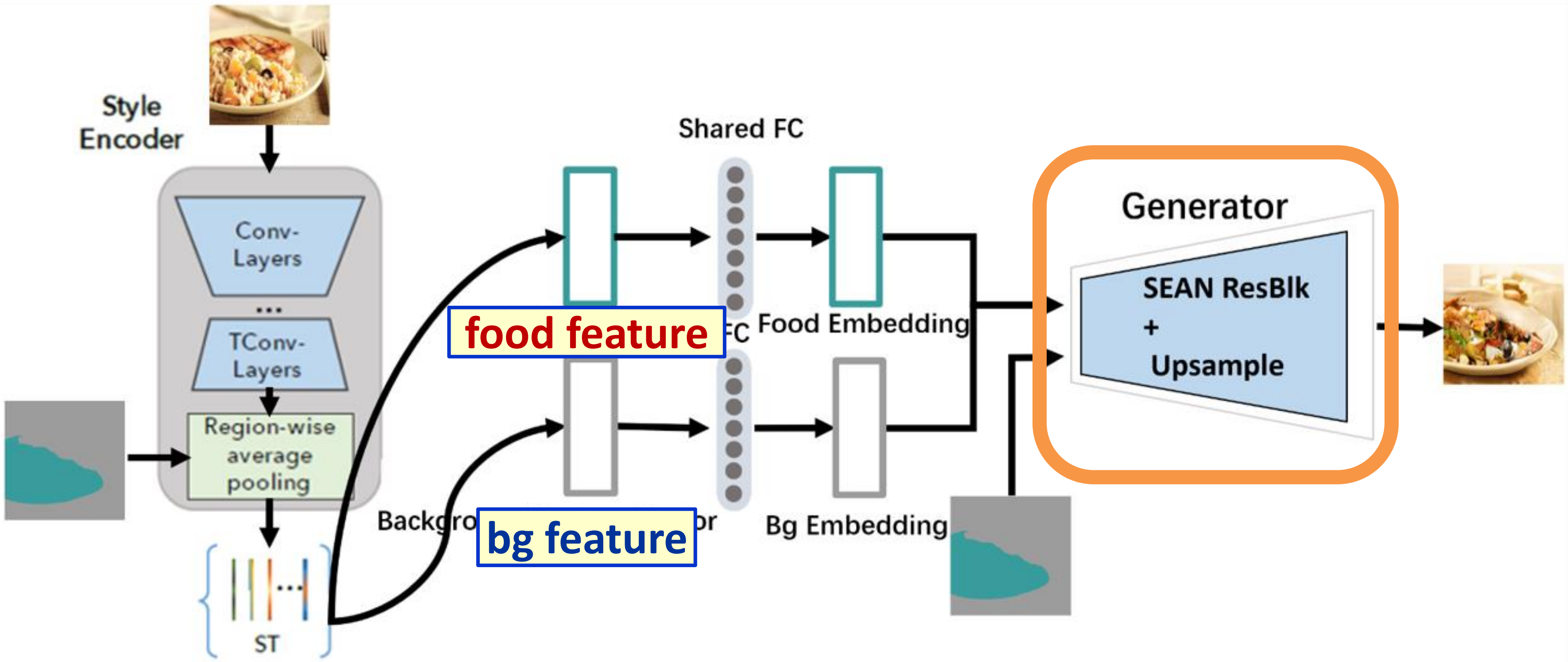
- Extract foreground food region features and background features separately from an input food image based on the corresponding food region mask
  - Use only food region features for training of cross-modal embeddings
  - Use triplet loss and cross-modal adversarial loss for training of cross-model embeddings



Proposed method: **Improvement (2) from RDE-GAN**

Improvement ② : **Masked-based image generator based on SEAN**

- Introduce a SEAN-based generator to control a style of each of the mask regions.
- Use adversarial loss, feature matching loss and Perceptual loss for training of a generator



Proposed method: **Loss functions (same as ACME and RDE-GAN)**

1) Adversarial training between text and image emb.

Modality Alignment Loss

$$L_{MA} = E_{i \sim p_{image}}[\log(D_M(E_V(i)))] + E_{r \sim p_{recipe}}[\log(1 - D_M(E_R(r)))]$$

2) Distance learning between texts and images

Triplet Loss

$$L_{Tri} = \sum_V [d(V_a, R_p) - d(V_a, R_n) + \alpha]_+ + \sum_R [d(R_a, V_p) - d(R_a, V_n) + \alpha]_+$$

3) Training image generator (GAN)

Adversarial, Feature matching, Perceptual Loss

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

4) Estimation of recipe title and ingredient from embeddings

Class Loss

$$L_{Class} = L_{Title}(V, L_t) + L_{Title}(R, L_t) + L_{Ingr}(V, L_i) + L_{Ingr}(R, L_i)$$

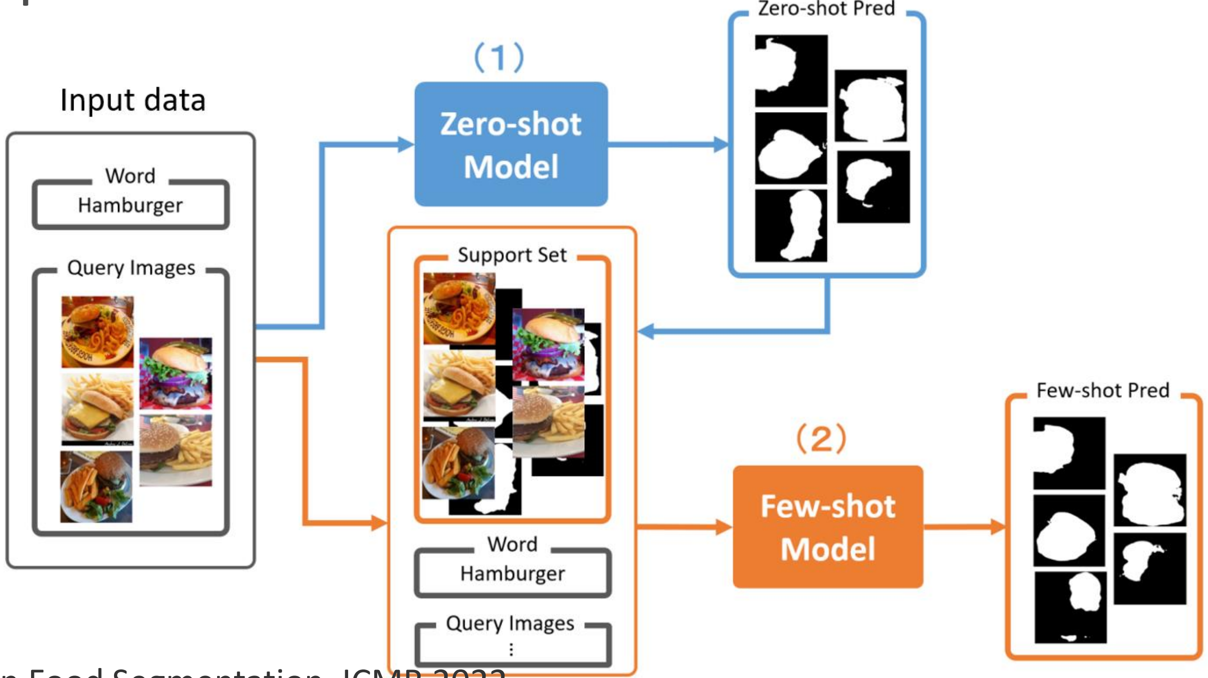
$$\Rightarrow \quad L_{Total} = \lambda_1 L_{Tri} + \lambda_2 L_{MA} + \lambda_3 L_{G_{r2i}} + \lambda_4 L_{Class}$$

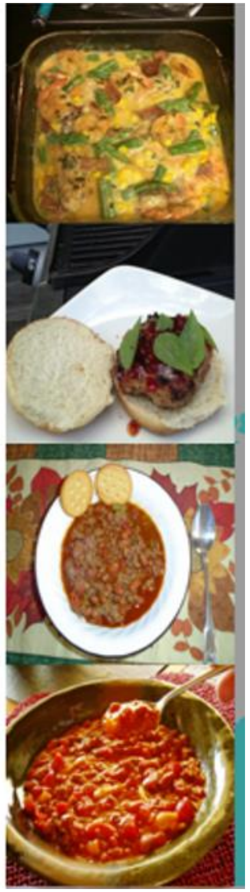
(λ_1=1.0 , λ_2 = 0.005 , λ_3 = 0.002, λ_4 = 0.002)

Preparing food masks for all the images of the Recipe1M dataset

To do that, we used "Unseen Food Segmentation[1]".

- We adapted a Zero/Few-shot segmentation method, PFENet[TPAMI 2021], for food domain.
- Using only recipe textual information, we created food masks for all the images in Recipe1M.
- MIoU of generated food masks: 73.0 %





Recipe1M images

Estimated food masks

[1] Yuma Honbu and Keiji Yanai: Unseen Food Segmentation, ICMR 2022.

Experiments

- **Quantitative/qualitative experiments for MRE-GAN**

- **Baselines**
  - Cross-modal GAN models: **ACME** [CVPR2019], **RDE-GAN** [ACMMM2021], **X-MRS**[ACMMM2021]   Food GAN models: **CookGAN** [CVPR2020]

- **Training data**

Recipe1M: We used 340,000 pairs of recipe texts and images.

(training: 238,999   val: 51,119   test: 51,303)

Food shape masks automatically generated by two methods: ① ②

① **DeepLabV3+ trained with food segmentation dataset, UECFoodPix Complete.**

② **"Unseen Food Segmentation" methods which is based on the combination of Zero-shot + Few-shot Segmentation**   ① **mIoU: 54.1%** ② **mIoU: 73.0%**

[1] Hao Wang et al. : Learning Cross-Modal Embeddings with Adversarial Networks for Cooking Recipes and Food Images. CVPR 2019.
[2] Yu Sugiyama and Keiji Yanai. Cross-Modal Recipe Embeddings by Disentangling Recipe Contents and Dish Styles. ACMMM2021
[3] Amaia Salvador, et al.: Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. CVPR2017.

Experiments (1): quantitative evaluation

Evaluation of the quality with FID and IS.

- **Compared with four baselines.**

(The smaller FID and the bigger IS means the higher quality.)

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 ($\mathrm{Mask}_{\mathrm{DeepLabV3+}}$)	56.72	56.11	–
Ours ($\mathrm{Mask}_{\mathrm{unseen}}$)	**27.44**	**27.12**	**8.27±.05**

MRE-GAN outperformed the baseline and MRE-GAN w/ low-quality masks.

Experiments (2): qualitative evaluation

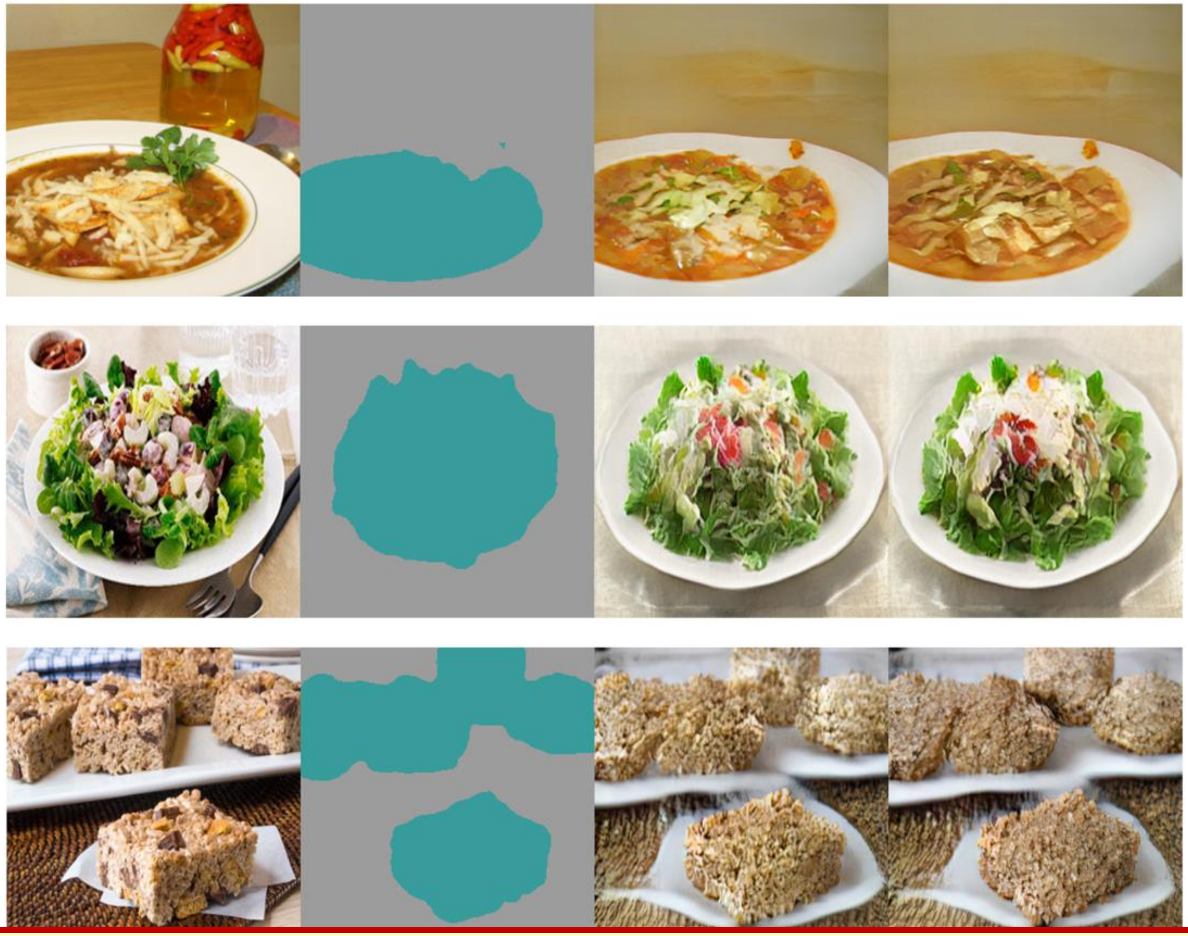
- Comparison on reconstruction ability (auto-encoder task)

Original ACME RDE-GAN **MRE-GAN**

MRE-GAN can synthesis images with the same style and shape as original.

Experiments (2): qualitative evaluation

- **Reconstruction from textual emb. (text2img) and image emb. (img2img)**



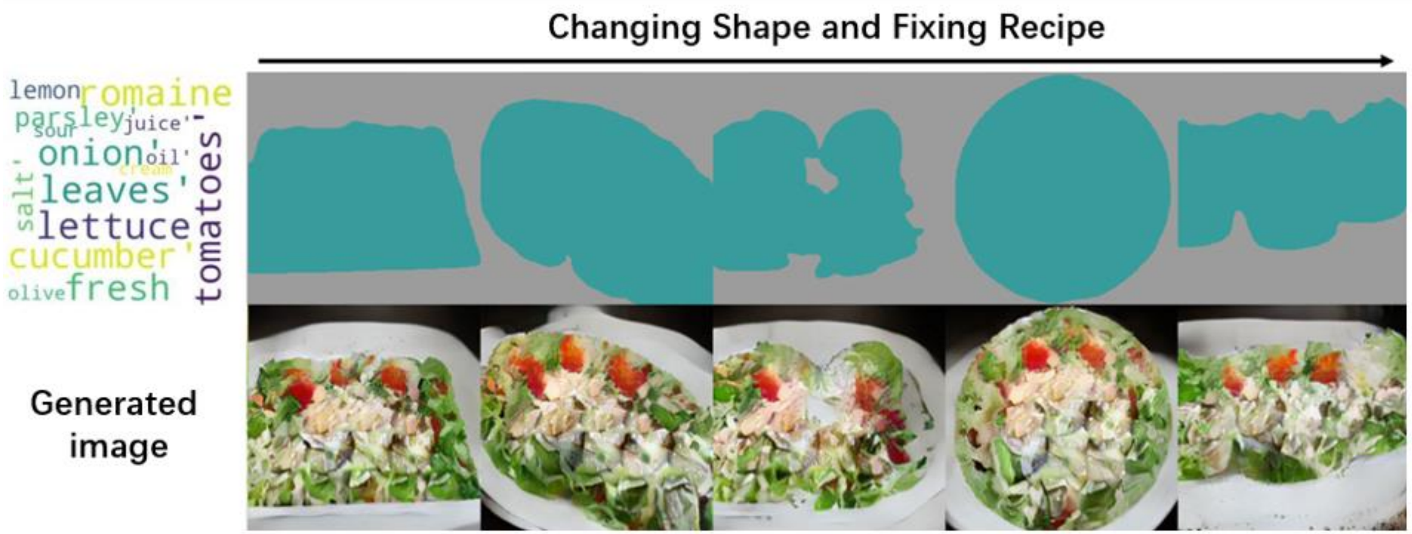
The image generated from both text and image emb. are almost identical .

Experiments (3) : multiple-dish images

Generating multiple-dish food images is possible.

Experiment (4): Modification of generated images

A) Changing shape masks with fixed recipe embeddings



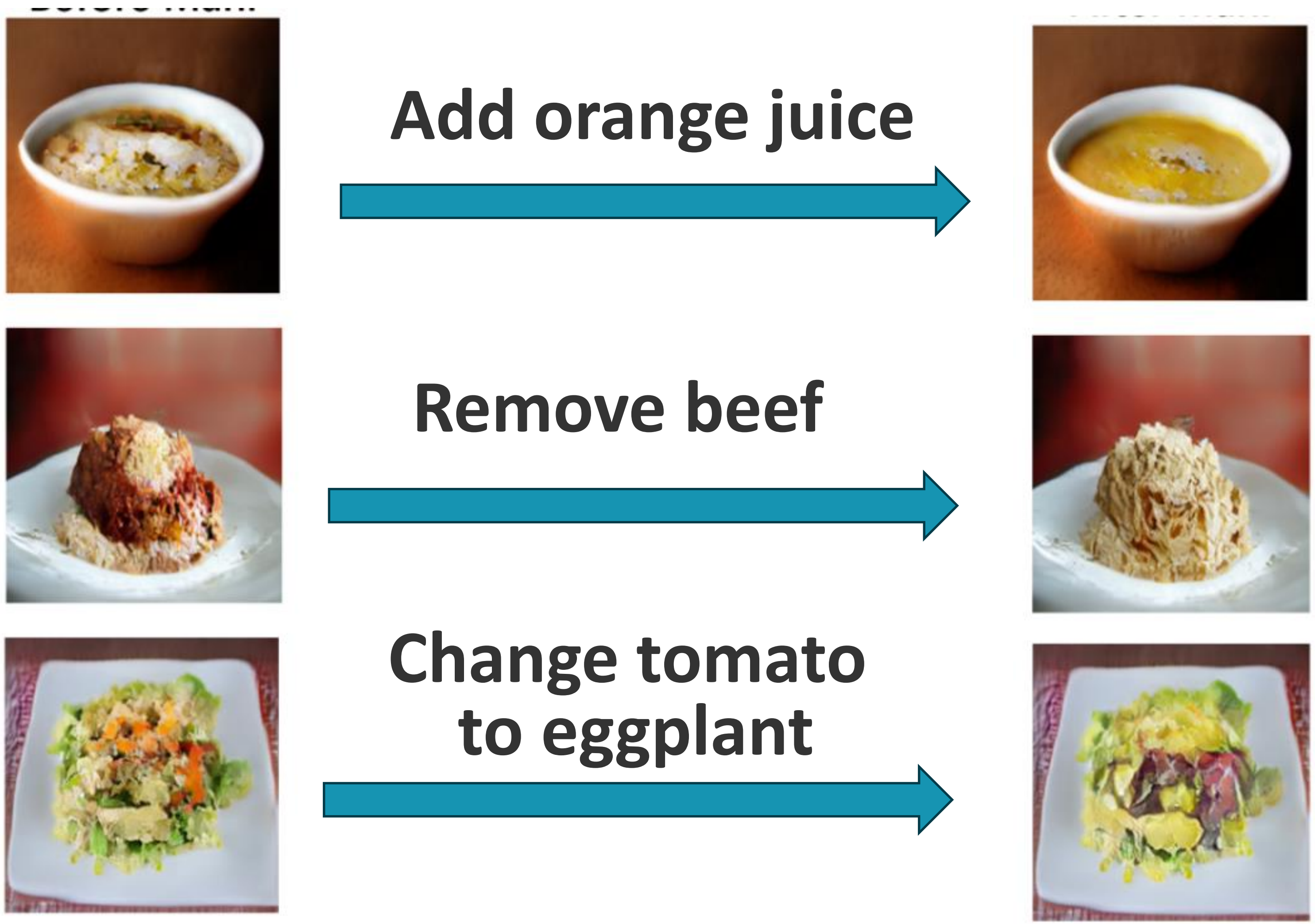
Experiment (4): Modification of generated images

B) Changing recipe embeddings gradually with fixed shape masks

(interpolating recipe embeddings between two recipe)

Experiment (4): Modification of generated images

C) Changing a part of the ingredient texts with fixed masks



Conclusions

- We proposed a Mask-based Recipe Embedding GAN (MRE-GAN) which generates food images from cross-modal recipe embeddings based on region mask images.

- We added food region masks to all the images in Recipe1M.

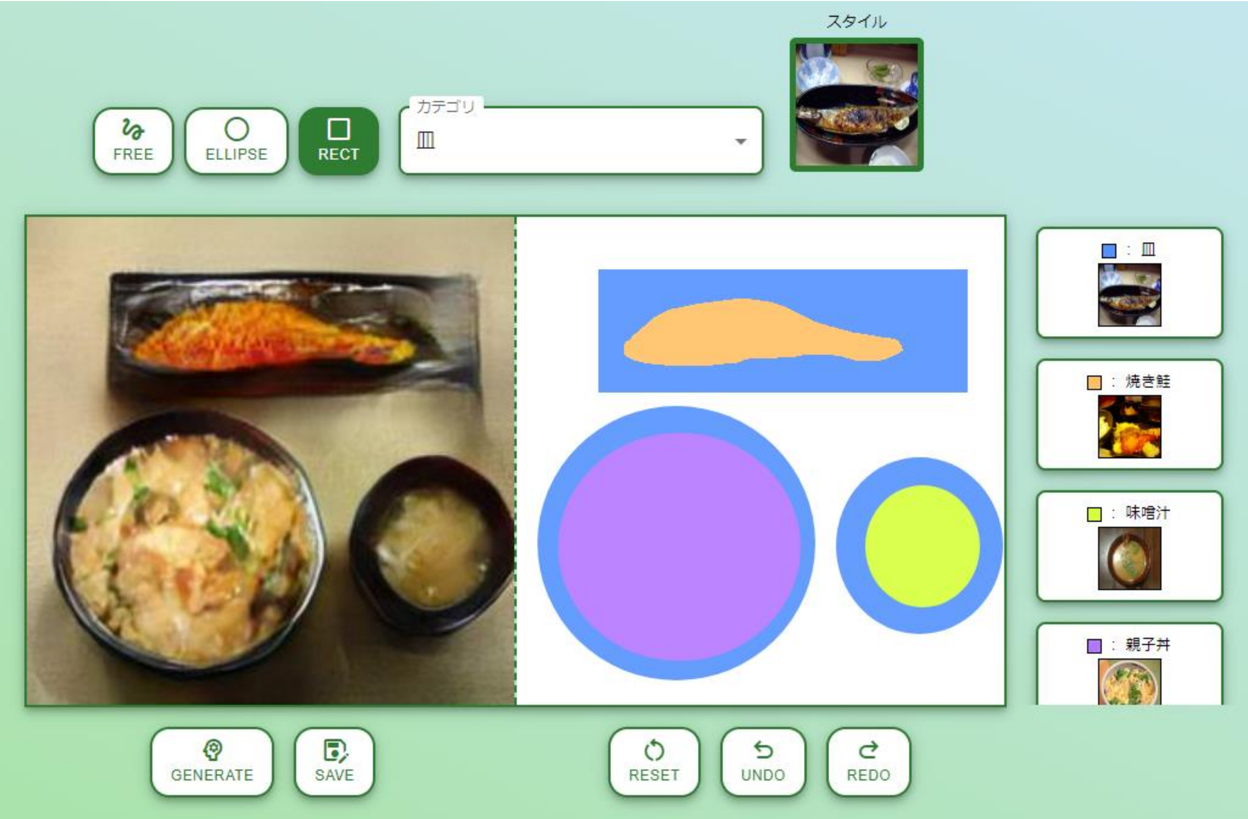
- We confirmed the effectiveness of MRE-GAN by the experiments.
  We successfully generated multiple-dish food images and arbitrary shape food images.

- ***Future works***
  - Currently, the shape of dish plates is not controllable. We like to add plate mask annotation to our food region mask dataset of Recipe1M.
  - We are working on Diffusion Model based food image generation with cross-model recipe embedding.

SetMealAsYouLike, ACM MM WS on MADiMa 2022.

- We added plate region masks to UEC-FoodPix Complete (100-kind food segmentation dataset) by Few-shot Segmentation methods.

- **We can generate set meal images from plate and food masks.**
  (not using cross-model embeddings)



Yuma Honbu and Keiji Yanai: **SetMealAsYouLike: Sketch-based Set Meal Image Synthesis with Plate Annotations,** ACM MM WS on MADiMA 2022.