PRMU2022

クロスモーダル レシピエンベディングによる 形状マスクに基づく食事画像生成

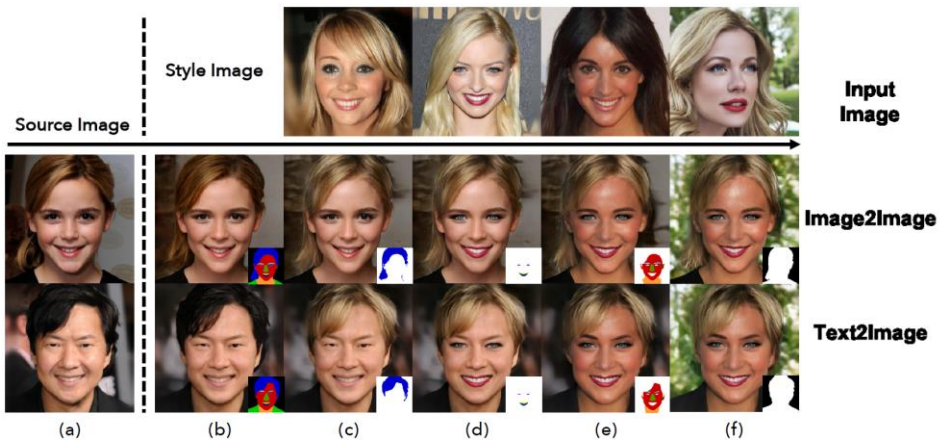
電気通信大学　大学院　情報学専攻

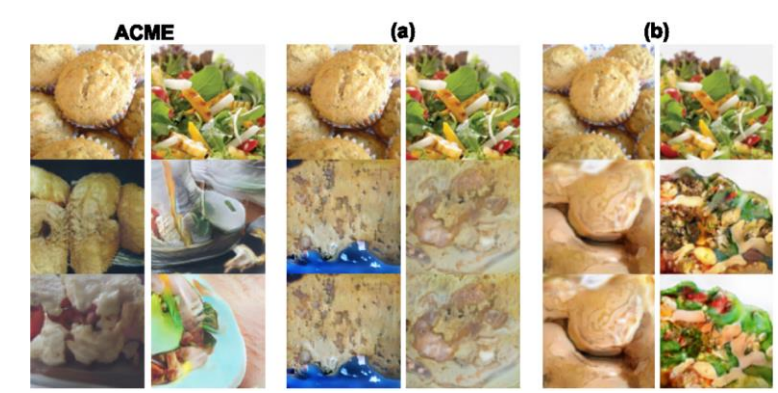
陳　仲涛　　本部勇真　　柳井啓司

はじめに

Image-to-Image

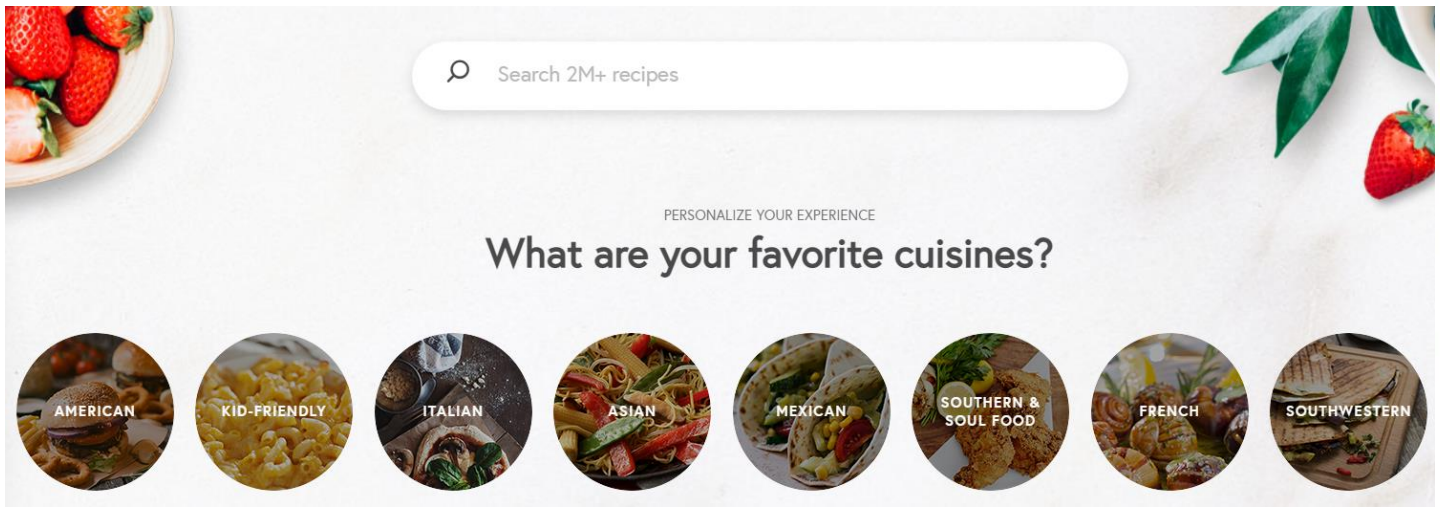
- 深層学習の発展により、画像合成や変換の技術が進歩
- Conditional GANは入力データを条件にリアルのような目標画像を生成
- CGANは食事画像の生成にも活用されているが、リアルな食事画像を生成することは容易ではない

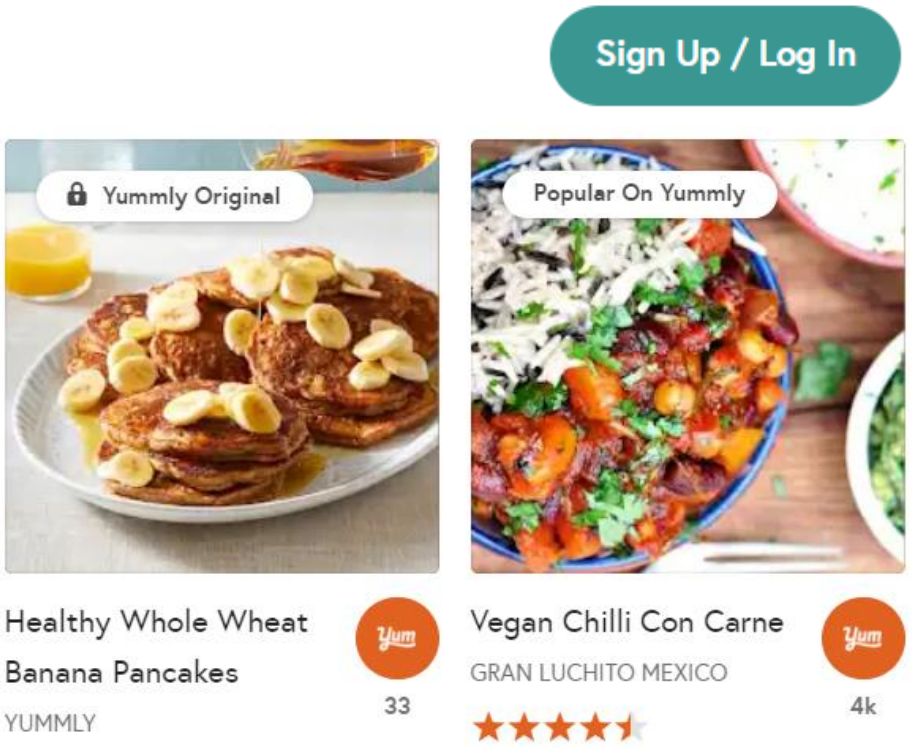




はじめに

- インターネットの普及に伴い、料理に関連する画像やレシピなどのデータが爆発的に増加している
- 食と健康への意識の向上から、料理に関する技術への関心が高まっている

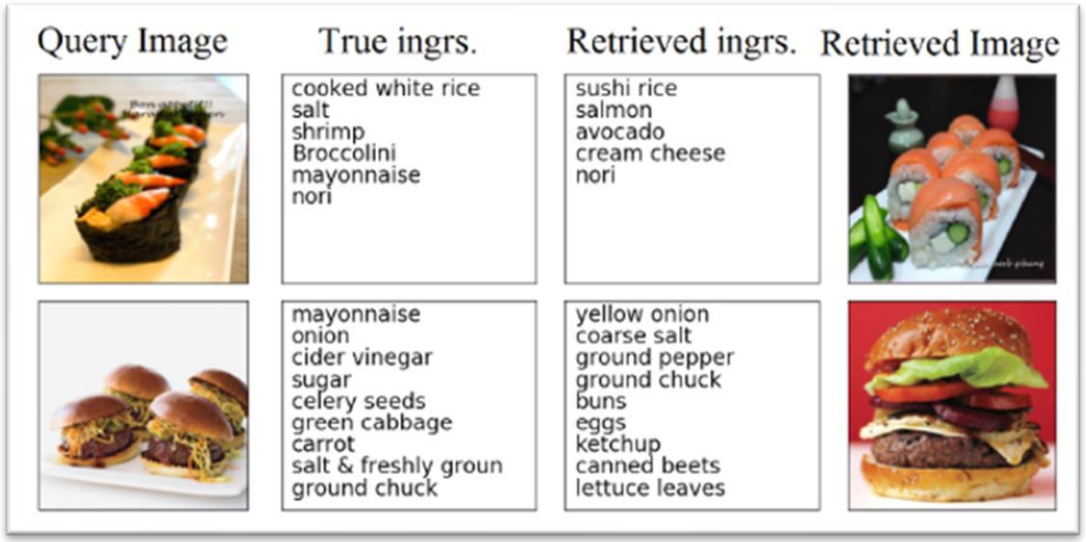


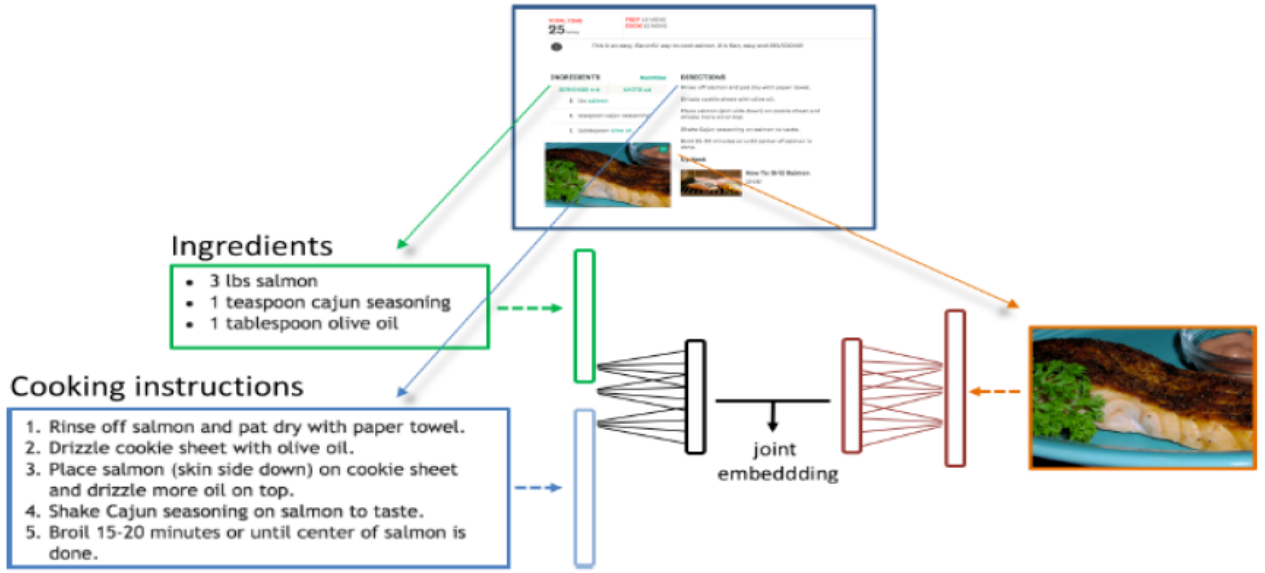


研究背景

クロスモダールレシピ検索

- **Recipe1M**が公開されてから、食事画像とテキストのクロスモーダルレシピ検索が多数研究されている
- **クロスモーダルレシピ検索**は、クロスモーダル検索の分野における重要なサブタスクであり、料理画像とレシピ（タイトル、材料、手順）の関連性を測定することに焦点を当てている
- 画像とテキストを同じ潜在空間にエンベディングすることで、異なるモダリティから関連のインスタンスを検索することである



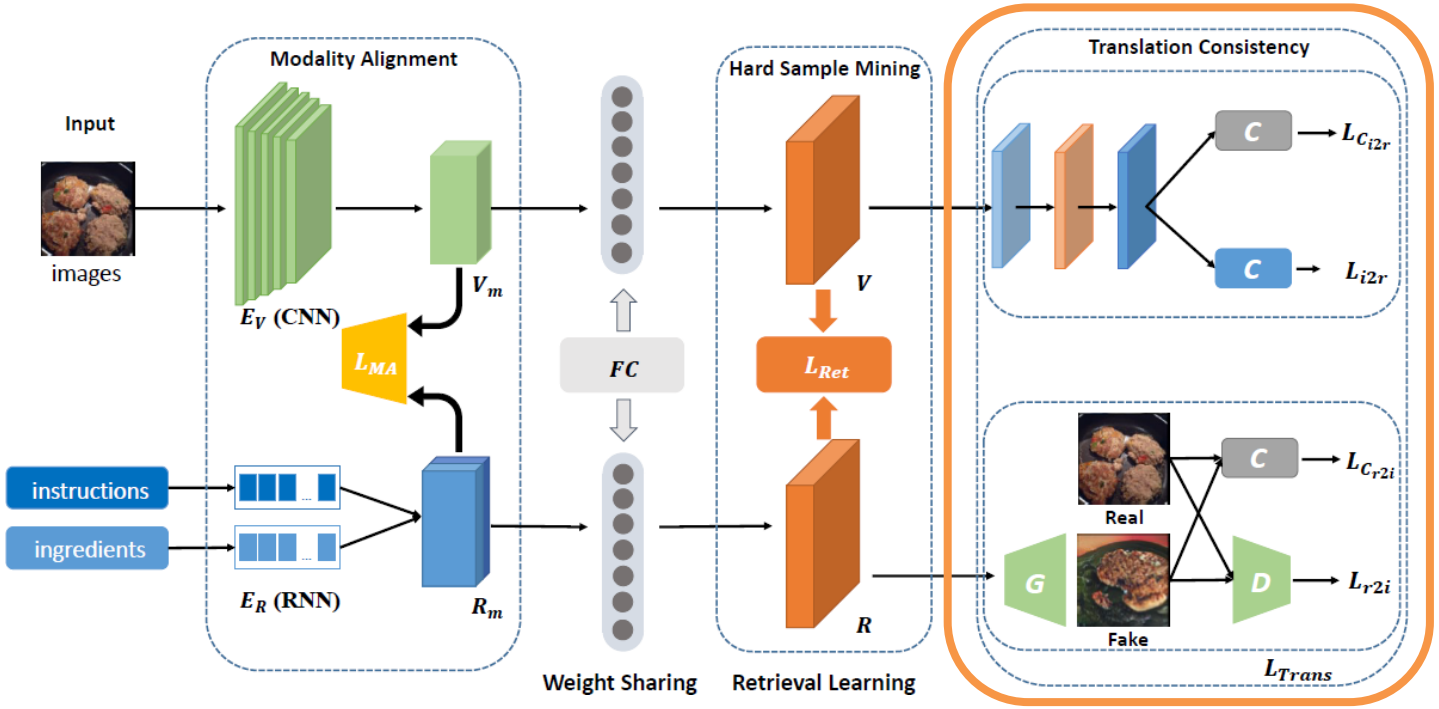


図はLearning Cross-modal Embeddings for Cooking Recipes and Food Imagesにより引用

関連研究①

ACME[Hao, CVPR2019]

- **ACME**はレシピ検索にテキスト予測と画像生成モデルを追加し、敵対的学習を行うことで、レシピ検索精度の向上を示し、レシピエンベディングを用いた食事画像の生成を実現
- この論文では、生成画像のクオリティが低い

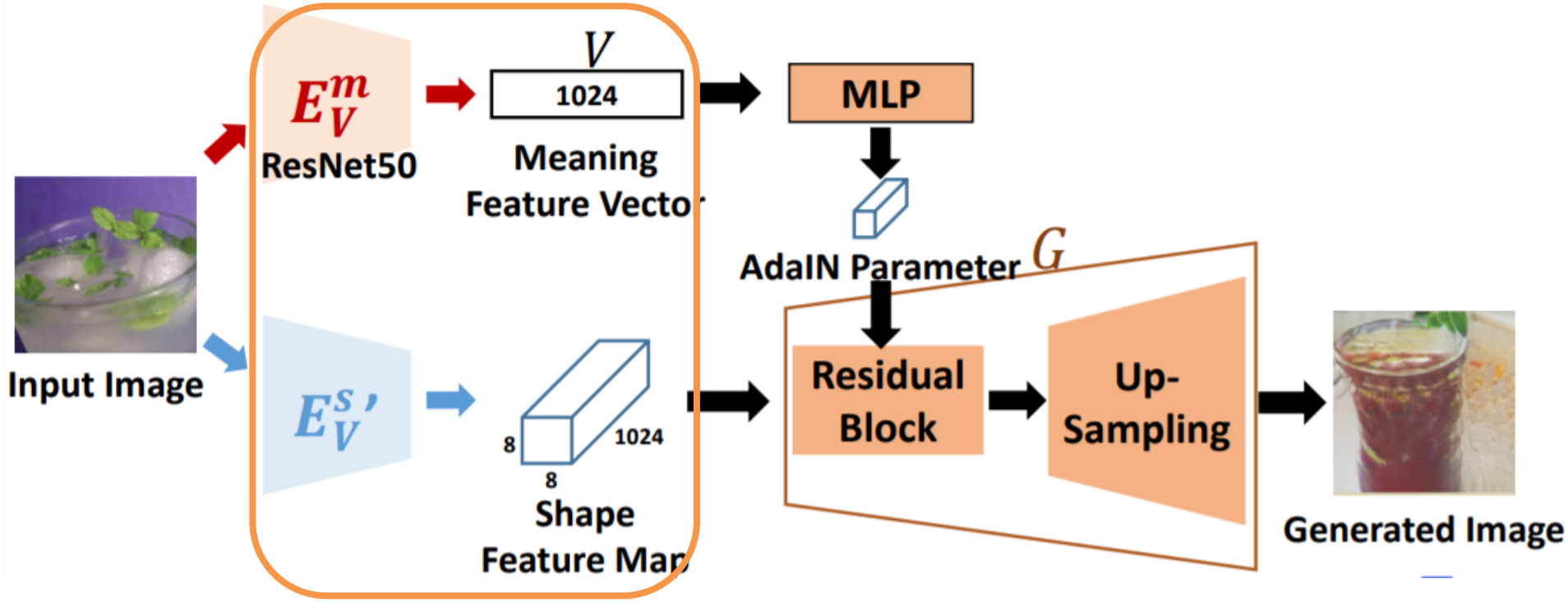


図はLearning Cross-Modal Embeddings with Adversarial Networks for Cooking Recipes and Food Images [CVPR 2019]により引用

関連研究②

RDE-GAN[Sugiyama, ACM Multimedia2021]

- **RDE-GAN**は画像特徴を意味と形状の二つに分離することで、より高い品質の画像生成を示した
- この論文では、二段階の学習で不安定な可能性があるため、
  画像のスタイルと形状の分離が不十分で生成画像の品質の改善が必要



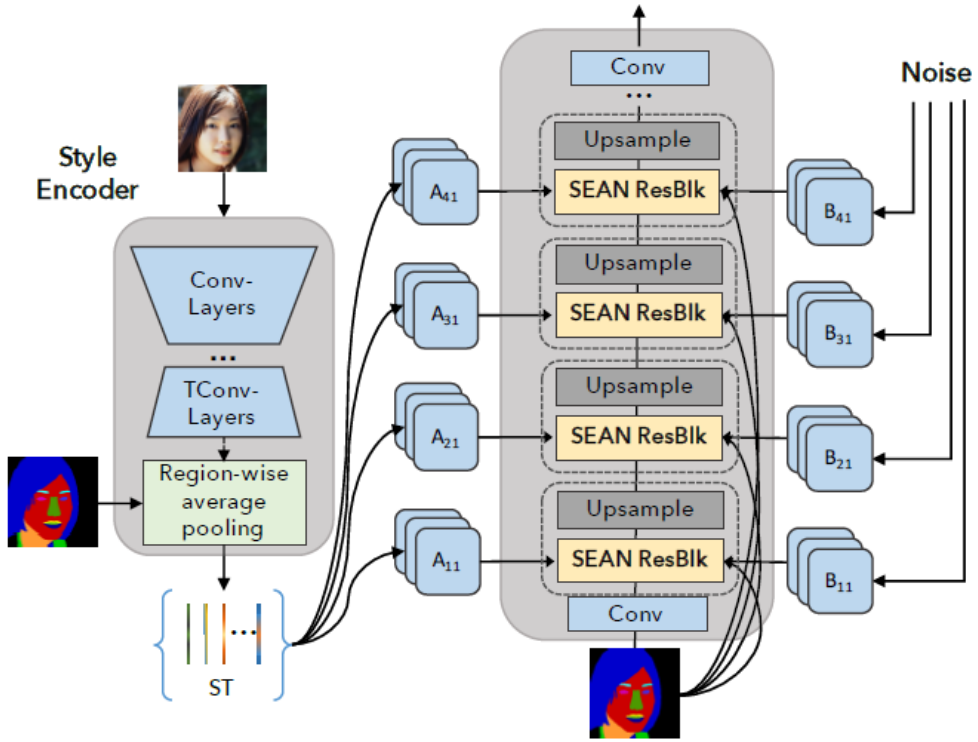
図はCross-Modal Recipe Embeddings by Disentangling Recipe Contents and Dish Styles [ACM Multimedia 2021]により引用

関連研究③

SEAN[CVPR 2020]

- **SEAN**はセマンティックマスク画像を用いて、パーツごとのスタイルを独立に計算

- レイアウトからの画像生成技術をRecipe-to-Imageタスクに適用



図はSEAN: Image Synthesis with Semantic Region-Adaptive Normalization[CVPR 2020]により引用

提案手法概要

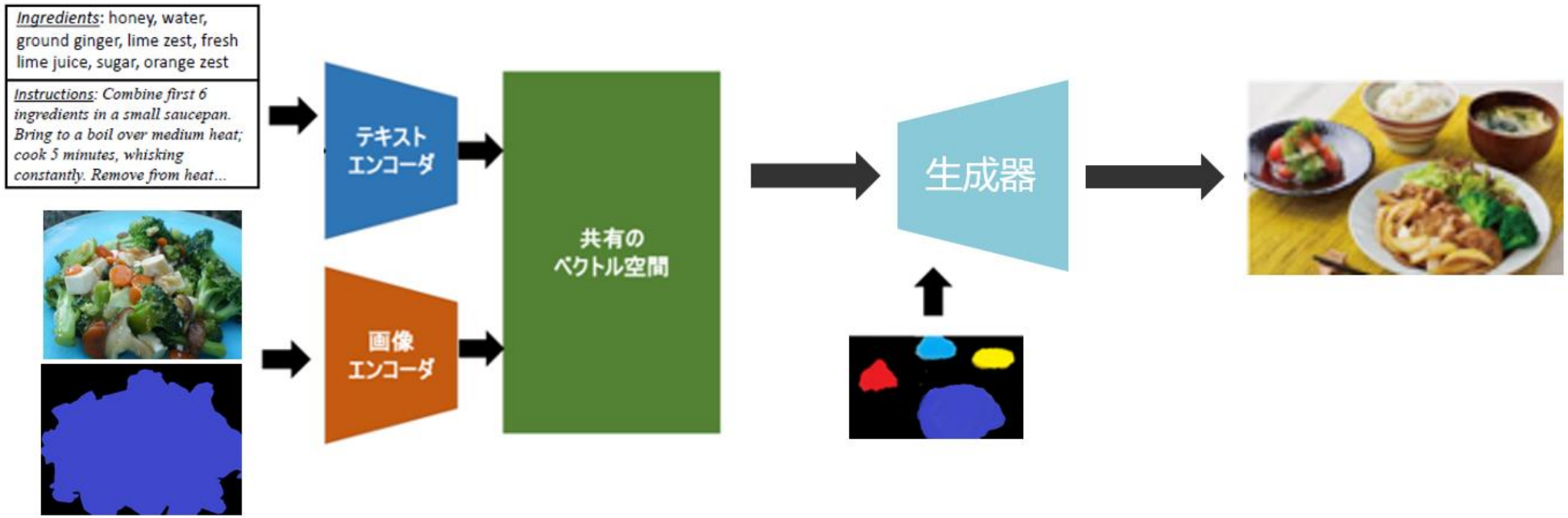
MRE-GAN(Mask-based Recipe Embedding GAN)

準備した食器に盛り付けをするように、複数の食事領域を含むマスク画像を用意し、
各領域にユーザが指定した料理画像を生成

[提案]

① 形状マスクを用意することで、画像生成にShapeに関する情報を提供

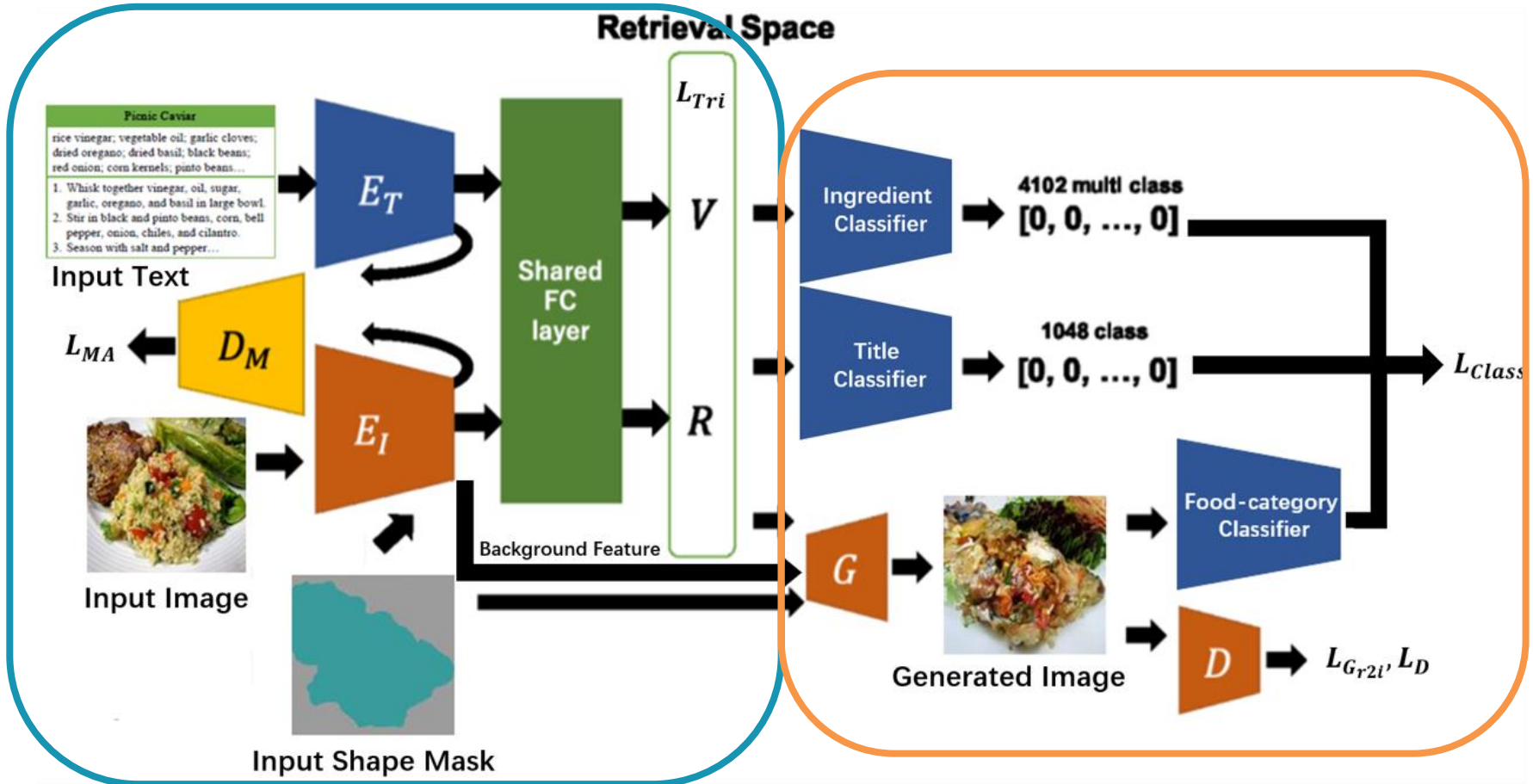
② 一つの画像エンコーダを用いて、One-Stageでモデルの学習を行い、高画質の画像を生成できるようにする



提案手法

詳細

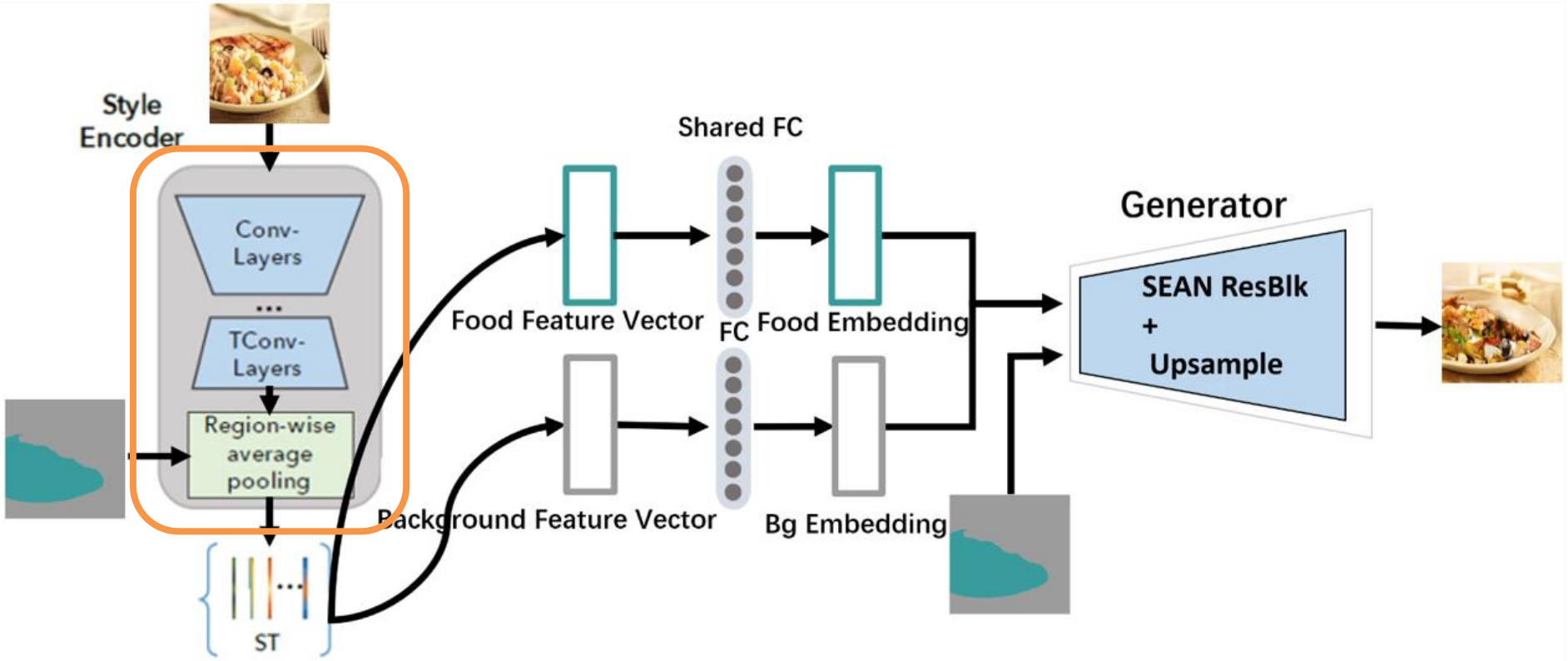
- モデル全体は主にクロスモーダルレシピ検索　と　画像生成とテキスト予測の部分で構成される
- 杉山らのRDE-GANをベースに、画像生成にSEAN正規化を使用することで、高品質な食事画像生成を行う



提案手法

改良①

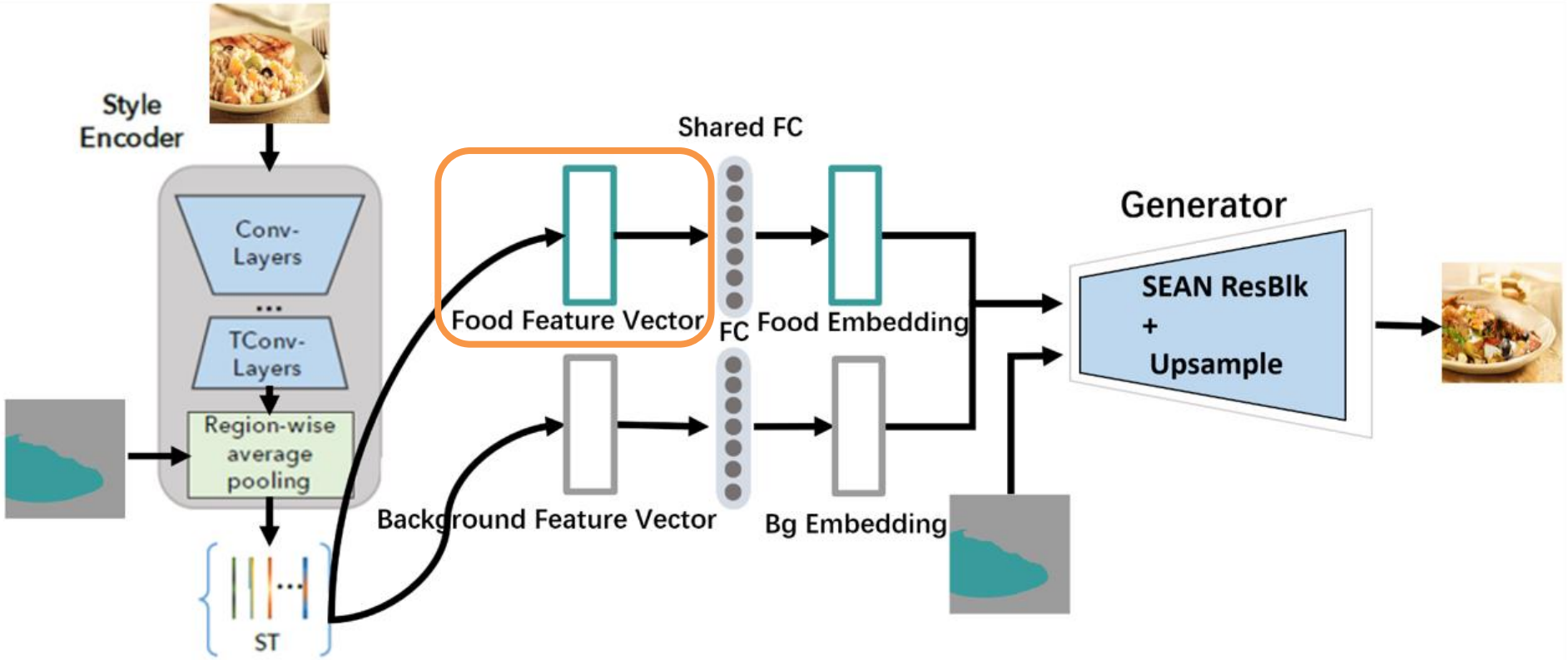
- 画像エンコーダの変更
  ーー一つのper-region style画像エンコーダを使用し、画像特徴を形状マスクに基づいた食事領域と背景領域の二つに分離
  ー食事領域の特徴のみをクロスモダールレシピ検索の学習に使用
  ーテキストと画像の早期アラインメントはModality Alignment Loss、エンベッディング同士の距離学習はTriplet Loss



提案手法

改良①

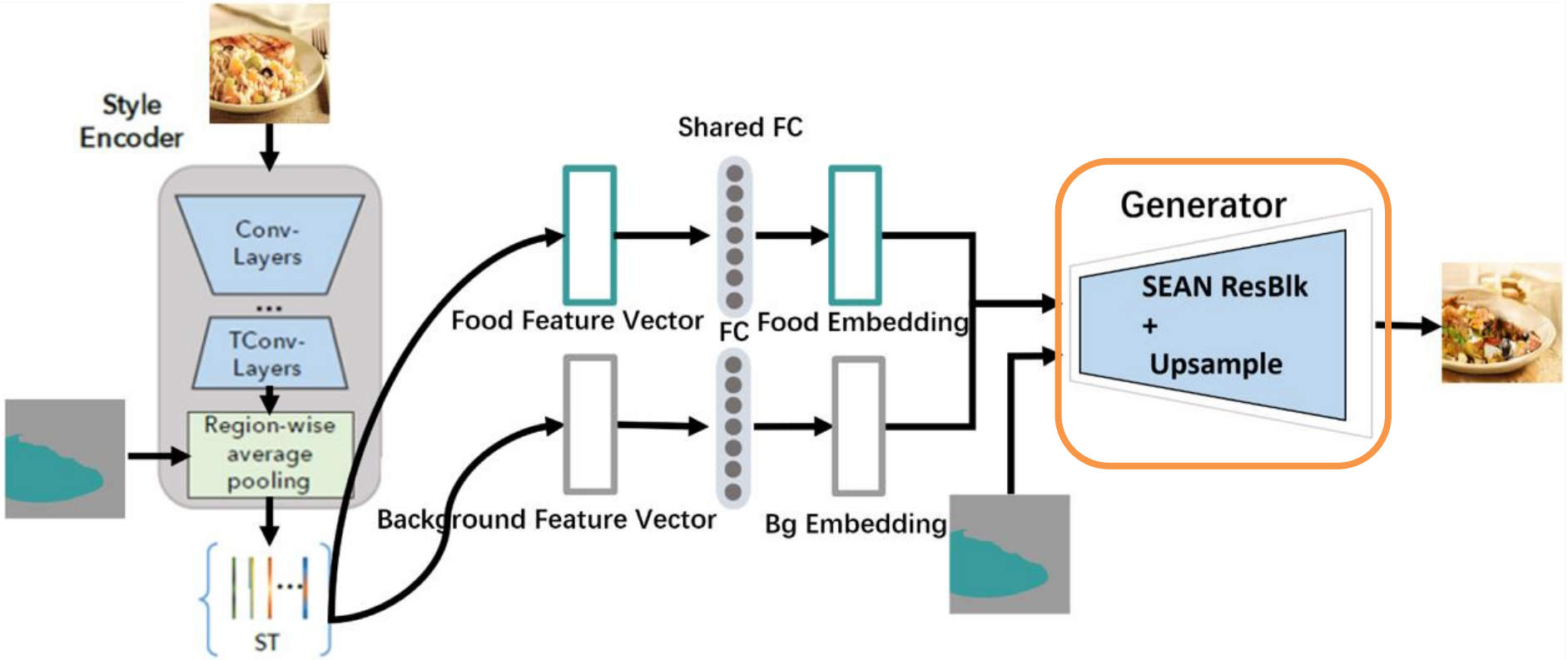
- 画像エンコーダの変更
  ーー一つのper-region style画像エンコーダを使用し、画像特徴を形状マスクに基づいた食事領域と背景領域の二つに分離
  ー食事領域の特徴のみをクロスモダールレシピ検索の学習に使用
  ーテキストと画像の早期アラインメントはModality Alignment Loss、エンベッディング同士の距離学習はTriplet Loss



提案手法

改良②

- 画像生成の変更
  ーSEANレイヤーの持つGeneratorを使用することで、生成画像のスタイルをより詳細に制御できる
  ーGANの学習：Adversarial loss、 Feature matching loss、 Perceptual loss



提案手法

Total loss

- Loss functions

ーテキストと画像の早期アラインメント:
Modality Alignment Loss

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

ーエンベッディング同士の距離学習:
Triplet Loss

$$L_{Tri} = \sum_V [d(V_a, R_p) - d(V_a, R_n) + \alpha]_+ + \sum_R [d(R_a, V_p) - d(R_a, V_n) + \alpha]_+ \quad (3)$$

ーGANの学習:
Adversarial Loss、 Feature matching Loss、 Perceptual Loss

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

ーエンベディング同士の分類:
Class Loss

$$L_{Class} = L_{Title}(V, L_t) + L_{Title}(R, L_t) + L_{Ingr}(V, L_i) + L_{Ingr}(R, L_i) \quad (8)$$

$$L_{Total} = \lambda_1 L_{Tri} + \lambda_2 L_{MA} + \lambda_3 L_{G_{r2i}} + \lambda_4 L_{Class} \quad (1)$$

各ハイパーパラメータはλ_1=1.0 , λ_2 = 0.005 , λ_3 = 0.002, λ_4 = 0.002とする

実験

- **MRE-GAN**の性能を検証するために、4つの実験を行った
- **ベースライン**
  ーACME
  ーRDE-GAN
- **学習データ**
  Recipe1Mデータセットに含む34万件のレシピと画像のペアを使用
  それぞれ学習：238,999　検証：51,119　テスト：51,303
  形状マスクはRecipe1Mの画像データに対して、領域分割を行うことで作成
  ①UECFoodPix Completeデータセットで学習済み**DeepLabV3+**による領域分割
  ②UECFoodPix Completeデータセットで学習済み**Zero-shot + Few-shot Segmentation**による領域分割

図4.2

実験

データセットの作成

- Zero-shot + Few-shot Segmentationは
  PFENet[TPAMI 2021] に食材の単語要素を加えることで、食事領域に特化させた**Few-shot**, **Zero-shot**モデルである **wPFE**, **zPFE** を提案した手法
- 二段階の推定でRecipe1Mのマスクを作成

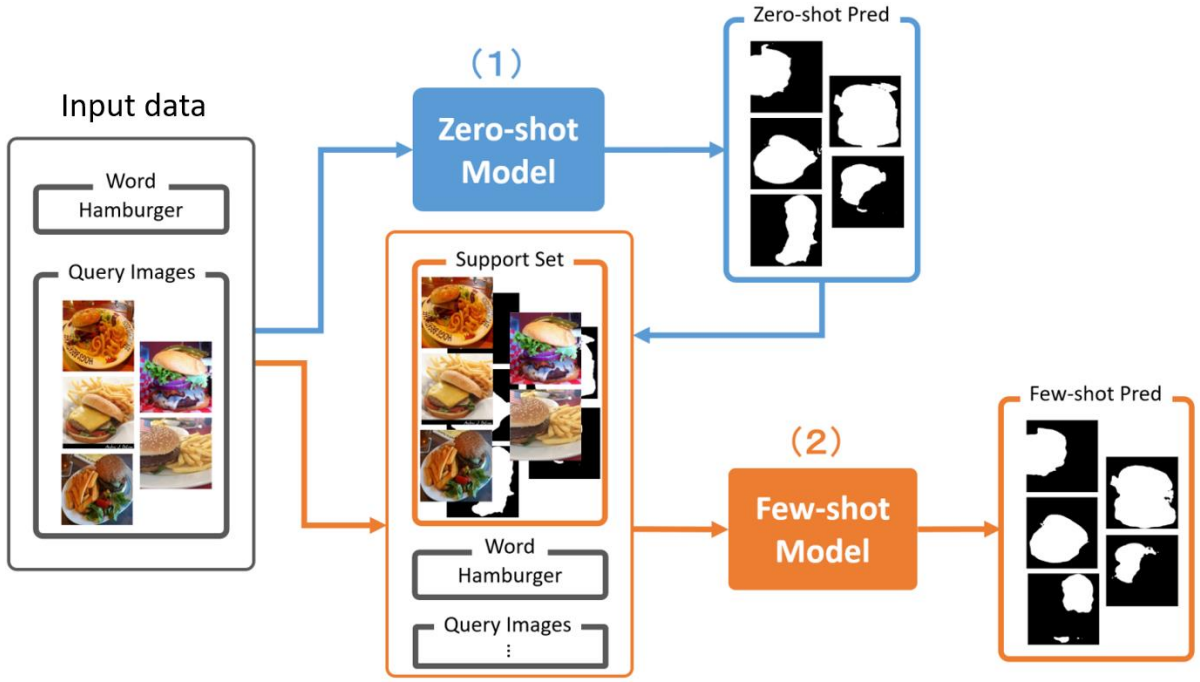


図4.1

図はYuma Honbu and Keiji Yanai: Few-Shot and Zero-Shot Semantic Segmentation for Food Images , Proc. of ICMR WS on Multimedia for Cooking and Eating Activities (CEA), (2021/08).により引用

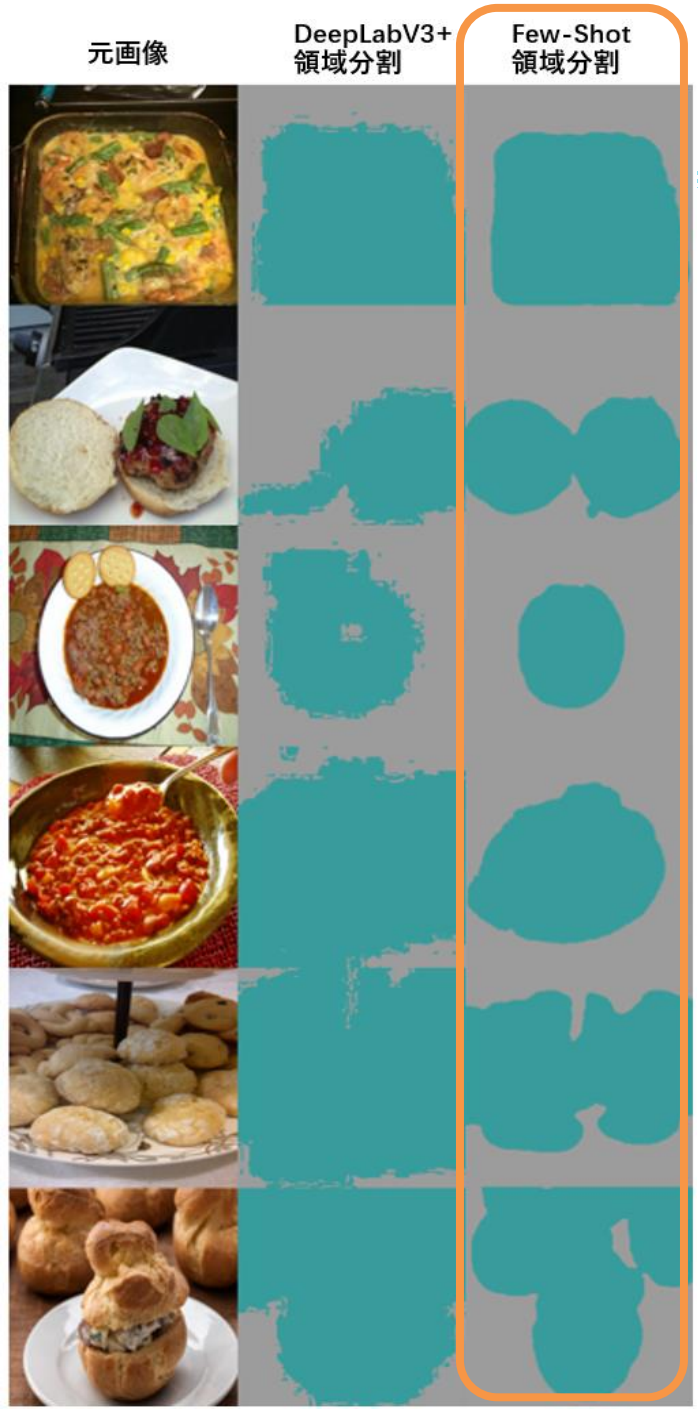


図4.2

実験１

画像生成の品質をFIDで評価

- **FID**は生成画像の品質を、元画像と生成画像の特徴量の分布間の距離で測る
  小さいほど実データに近い画像を生成できる
- 公平な比較をするために、提案手法による生成画像の解像度を256x256 ⇒ 128x128にリサイズ

表 4.1: 本手法と既存手法の FID 比較

手法	FID ↓@画像から再構成	FID ↓@テキストから再構成
ACME	183.8	182.9
RDE-GAN	158.9	158.6
Ours($Mask_{DeepLabV3+}$)	165.4	166.9
Ours($Mask_{Few-Shot}$)	126.2	127.8

実験１

画像生成の品質をFIDで評価

- **FID**は生成画像の品質を、元画像と生成画像の特徴量の分布間の距離で測る
  小さいほど実データに近い画像を生成できる
- 公平な比較をするために、提案手法による生成画像の解像度を256x256 ⇒ 128x128にリサイズ

既存研究より生成画像の品質が良い結果

表 4.1: 本手法と既存手法の FID 比較

手法	FID ↓@画像から再構成	FID ↓@テキストから再構成
ACME	183.8	182.9
RDE-GAN	158.9	158.6
Ours($Mask_{DeepLabV3+}$)	165.4	166.9
Ours($Mask_{Few-Shot}$)	126.2	127.8

実験１

画像生成の品質をFIDで評価

- **FID**は生成画像の品質を、元画像と生成画像の特徴量の分布間の距離で測る
  小さいほど実データに近い画像を生成できる
- 公平な比較をするために、提案手法による生成画像の解像度を256x256 ⇒ 128x128にリサイズ

既存研究より生成画像の品質が良い結果

表 4.1: 本手法と既存手法の FID 比較

手法	FID ↓@画像から再構成	FID ↓@テキストから再構成
ACME	183.8	182.9
RDE-GAN	158.9	158.6
Ours($Mask_{DeepLabV3+}$)	165.4	166.9
Ours($Mask_{Few-Shot}$)	126.2	127.8

実験２

画像生成の定性的評価

- 生成画像の視覚的品質はベースライン手法との比較を行う

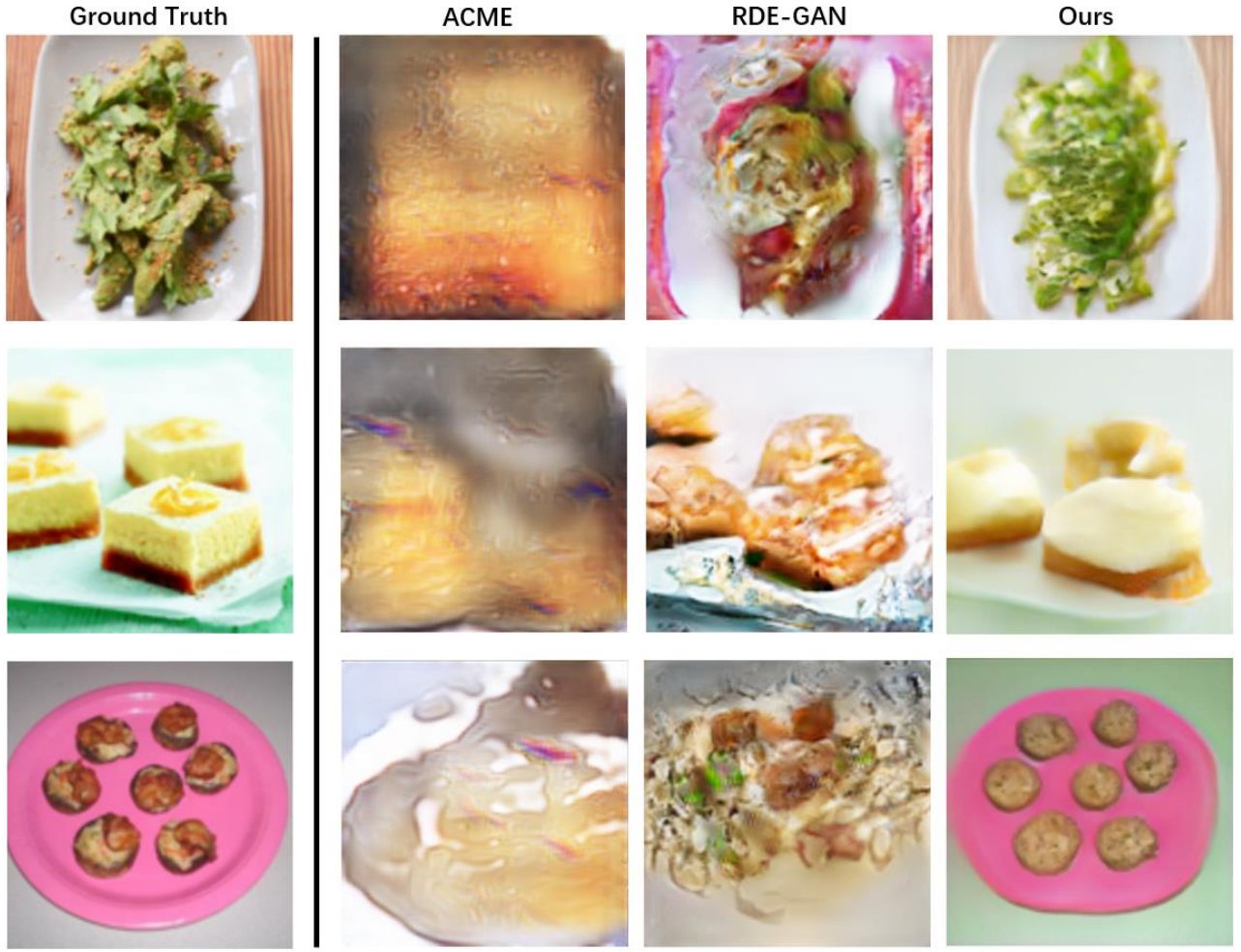


図4.3

実験２

**従来手法に比べ、提案手法の方が食事の形状を保ちつつ、
より本物に近い色やテクスチャの食事画像を生成できる**

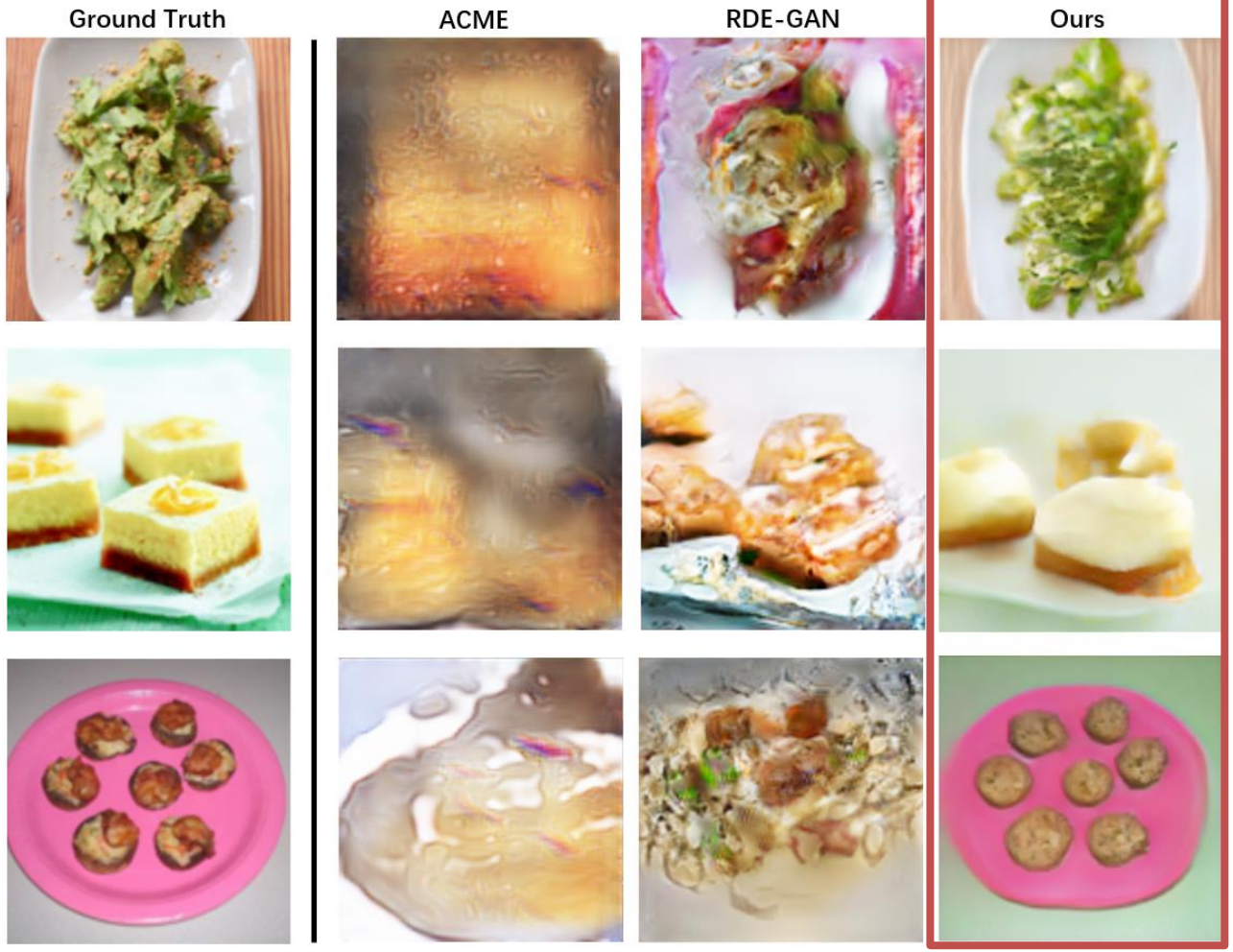


図4.3

実験２

エンベディング同士の距離学習

- 意味的エンベディングと視覚的エンベディングの間の距離学習はどの程度、進められたのかを検証

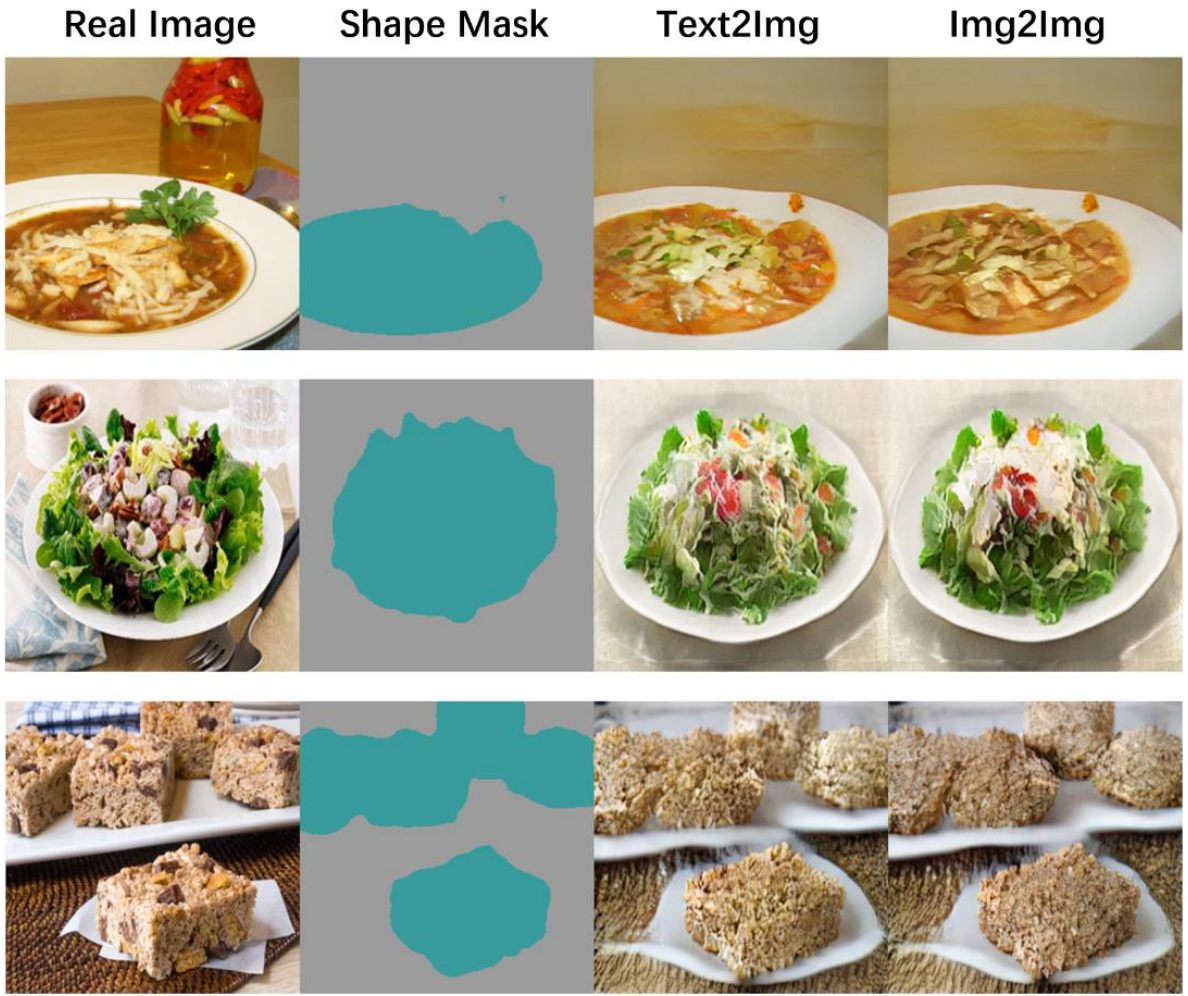
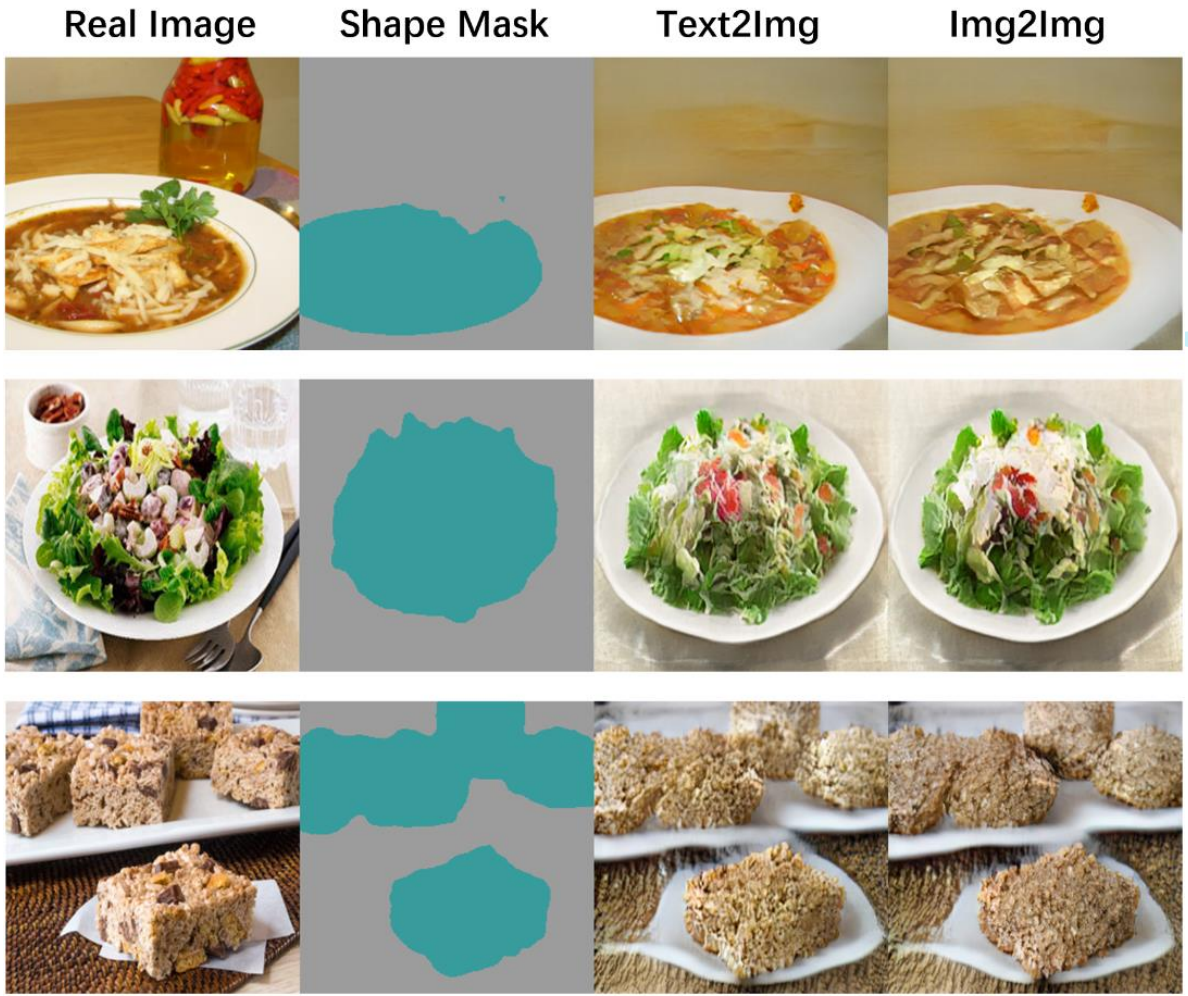


図4.4

実験２

エンベディング同士の距離学習

- 意味的エンベディングと視覚的エンベディングの間の距離学習はどの程度、進められたのかを検証



類似性が高い

図4.4

実験 3

複数品の画像生成

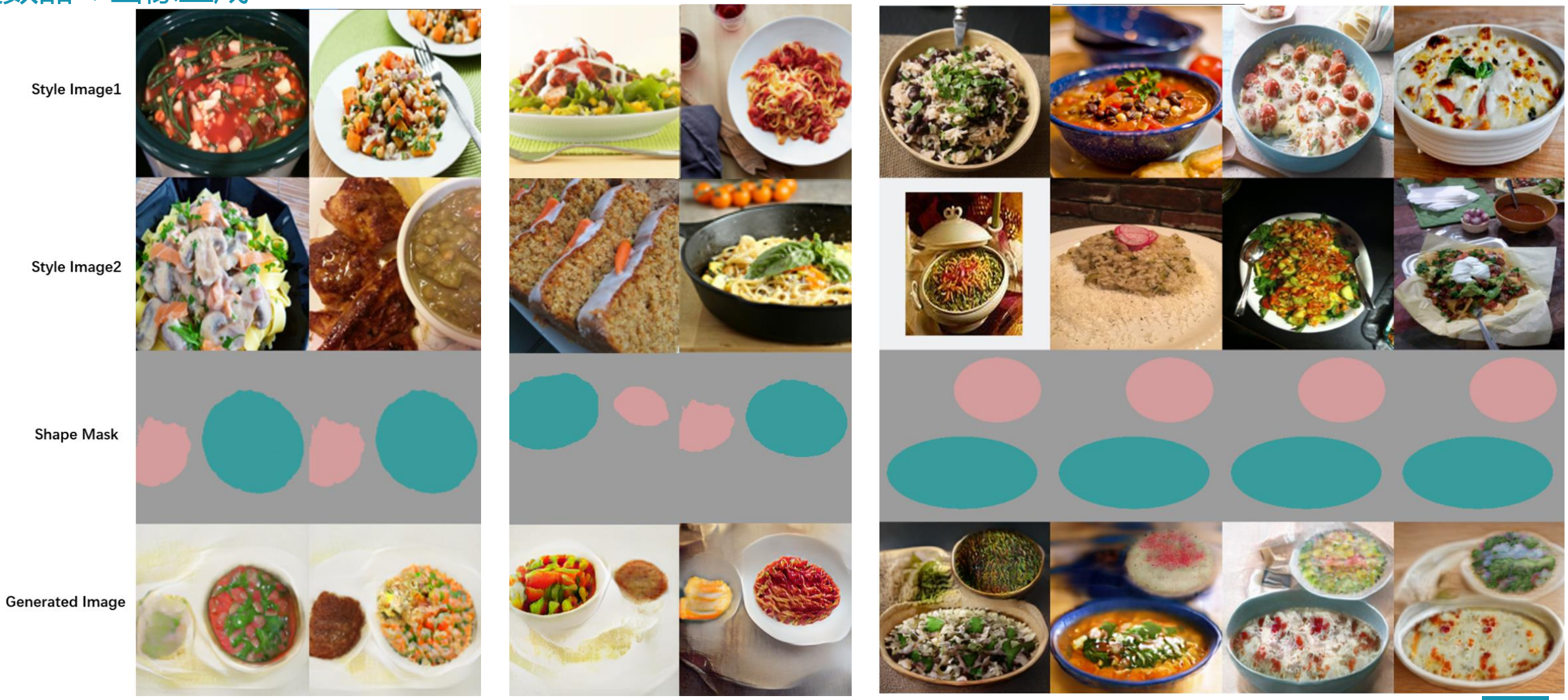


図4.5

実験４

生成画像の編集

- 同じレシピエンベディングを用いて、画像生成に形状マスクのみを変化させる
- 同じ形状マスクに基づいて、画像生成に使用するテキストエンベディングA → Bに変化させる
- 同じ形状マスクに基づいて、画像生成に入力食材のみを変化させる

実験４

生成画像の編集

- 同じレシピエンベディングを用いて、画像生成に形状マスクのみを変化させる
- 同じ形状マスクに基づいて、画像生成に使用するテキストエンベディングA → Bに変化させる
- 同じ形状マスクに基づいて、画像生成に入力食材のみを変化させる

実験４

画像の空間的再構成

- 同じレシピエンベディングを用いて、異なる形状マスクからの食事画像の生成
- 提案モデルによる画像生成は形状と意味的特徴を同時に考慮した生成結果を示す

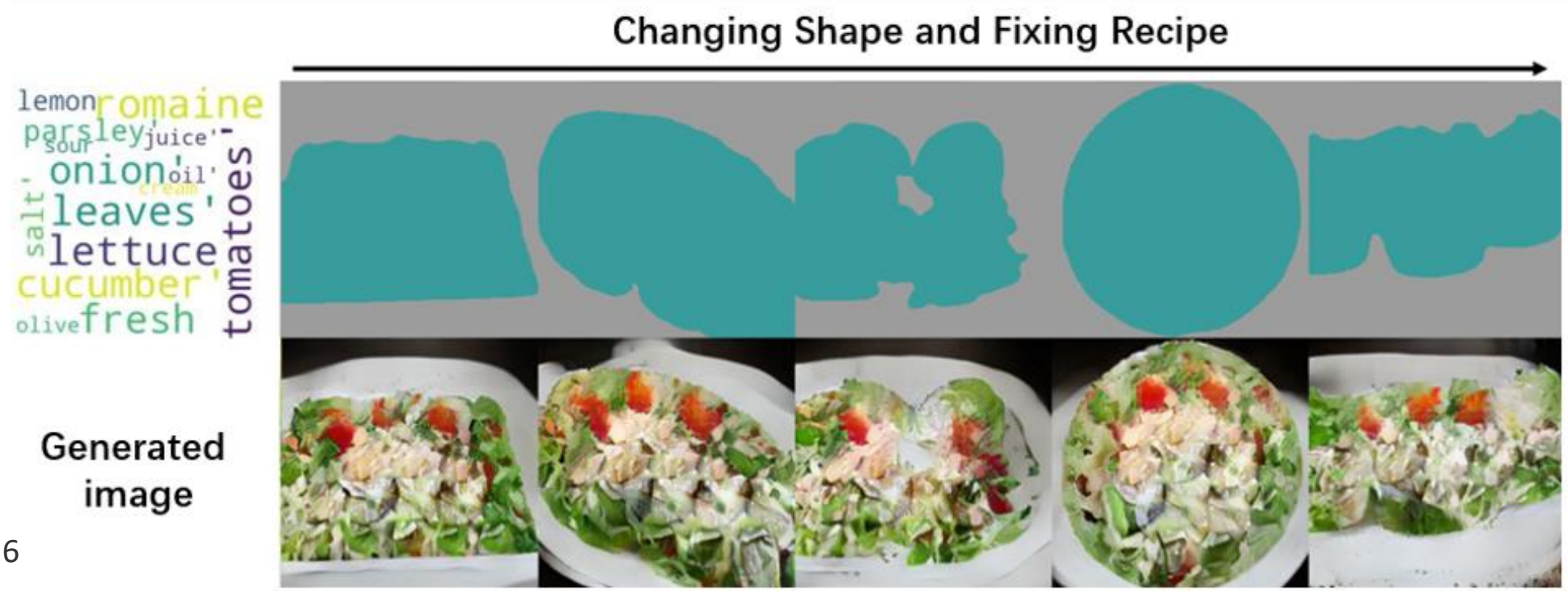


図4.6

実験４

画像の空間的再構成

- 提案モデルによる画像生成は形状と意味的特徴を同時に考慮した生成結果を示す

形状と意味的特徴を同時に考慮した結果を示す

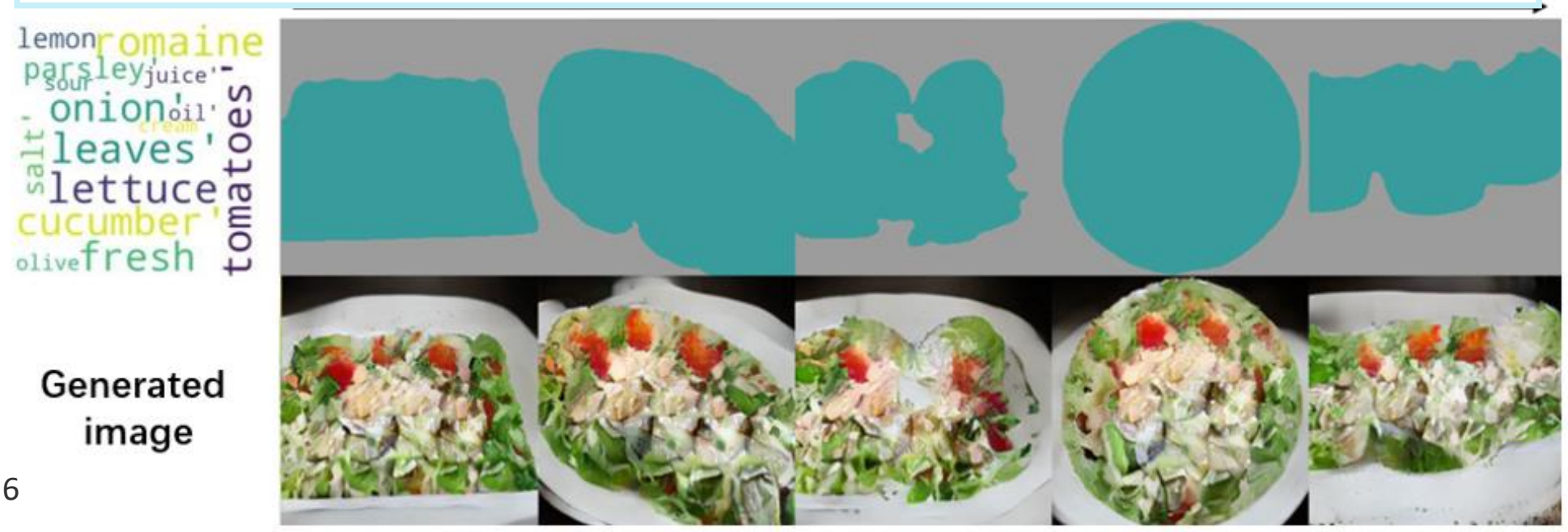


図4.6

実験４

生成画像の編集

- 同じレシピエンベディングを用いて、画像生成に形状マスクのみを変化させる
- 同じ形状マスクに基づいて、画像生成に使用するテキストエンベディングをA → Bに変化させる
- 同じ形状マスクに基づいて、画像生成に入力食材のみを変化させる

実験４

レシピエンベディング特徴空間の連続性

スタイルミックスによる食事画像の生成

図4.7

実験４

生成画像の編集

- 同じレシピエンベディングを用いて、画像生成に形状マスクのみを変化させる
- 同じ形状マスクに基づいて、画像生成に使用するテキストエンベディングA → Bに変化させる
- 同じ形状マスクに基づいて、画像生成に入力食材のみを変化させる

実験４

入力食材の変更による生成画像の編集

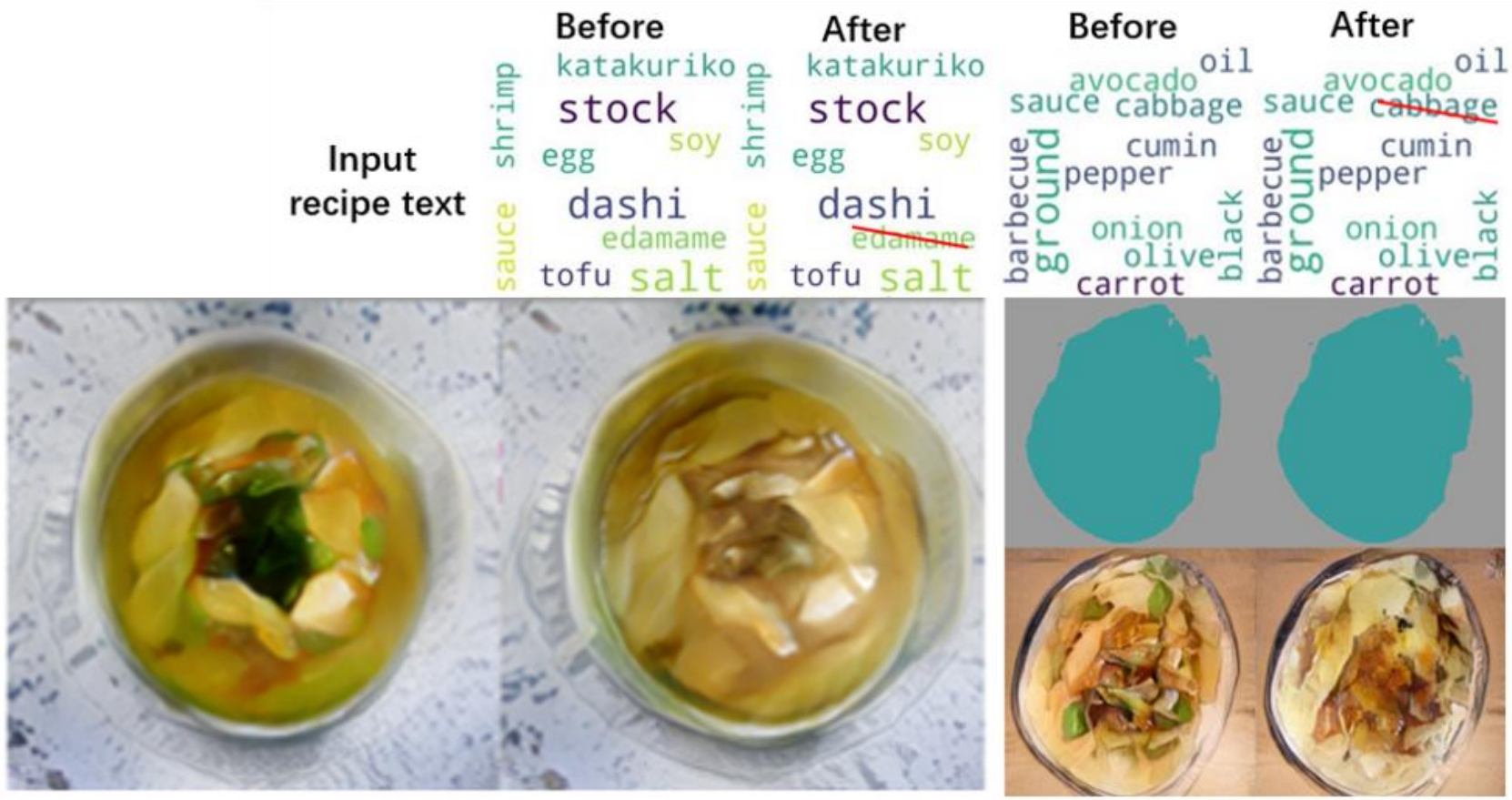
- 入力の食材のテキストからそれぞれ左の枝豆、或は、右のキャベツを取り除いた場合に、生成画像もこの変更に応じて変わった外観の食事画像を生成している

図4.8

実験４

入力食材の変更による生成画像の編集

- 左側の入力食材からえだまめを除いた場合に、生成画像に緑色の要素が消えた



実験４

入力食材の変更による生成画像の編集

- 右の入力食材からキャベツを取り除いた場合に、この変更に応じて生成画像の見た目も変わった
- この結果から、提案モデルは入力食材を通して、生成する料理の見た目を調整できる

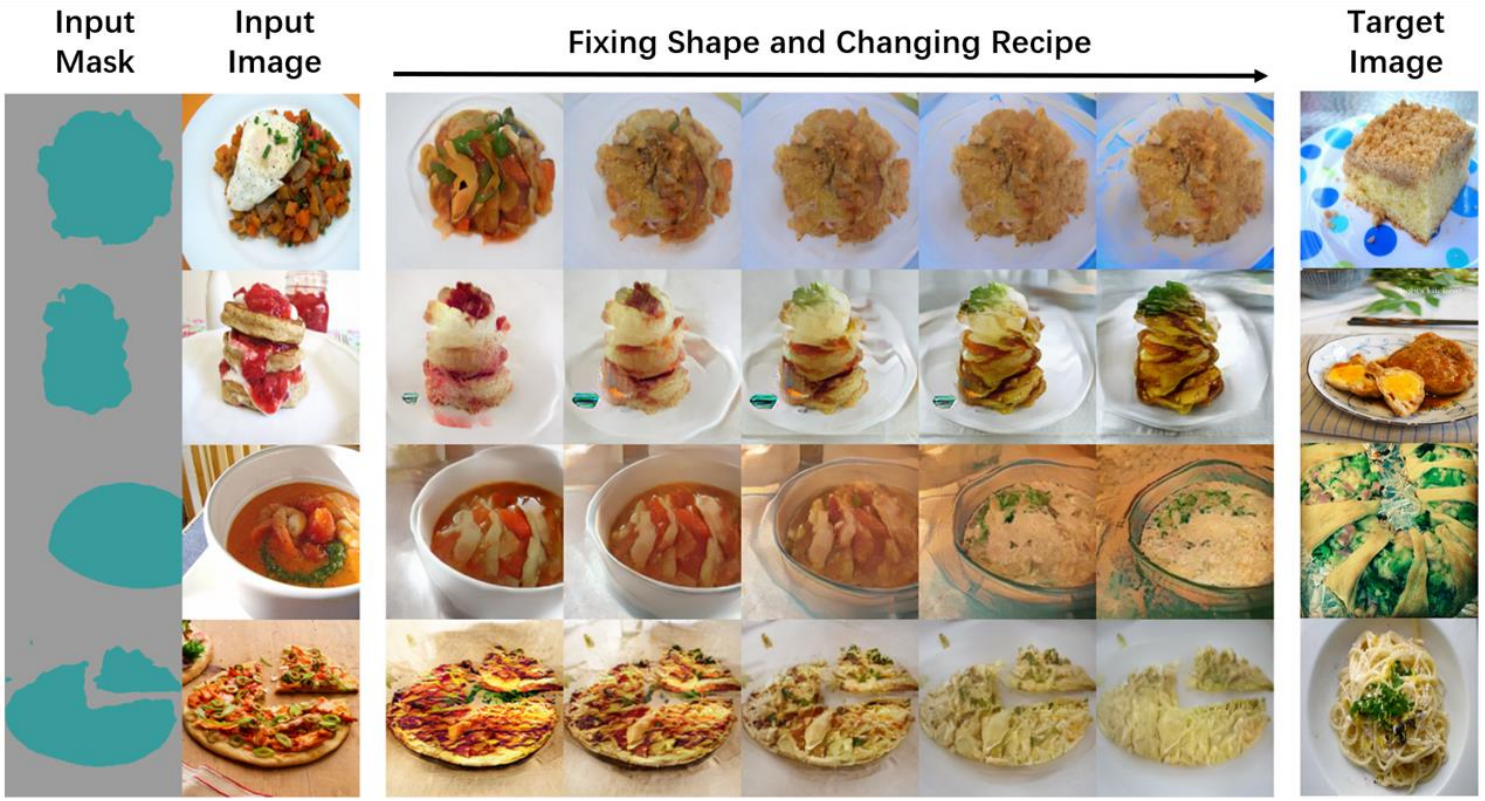
図4.8

まとめ

- 本研究は、準備した食器に盛り付けをするように、
  各領域にユーザが指定した料理画像を生成するMRE-GANを提案
- 既存研究に画像の意味と形状の分離が不十分である問題を解決し、
  One-Stageで安定的な学習を実現
- 実験から、提案手法はより高画質の生成画像を示し、
  複数品を含む任意の形状マスクに基づいたレシピからの画像生成を実現

- 今後の課題として、
  ①複数品の一貫性を保持できないため、レシピ間の関係性を考慮
  ②食器の形状を指定することができないので、
  　食器領域のアノテーションを付与することで、食事と食器の形状を両方指定可能とする

補足資料


Input Mask
Input Image
Fixing Shape and Changing Recipe
Target Image


Style Image1
Style Image2
Shape Mask
Generated Image