社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

クロスモーダルレシピエンベディングによるマスクに基づく食事画像生成

陳　仲涛†　本部　勇真†　柳井　啓司†

† 電気通信大学 大学院情報理工学研究科 情報学専攻
E-mail: †{chen-z,honbu-y}@mm.inf.uec.ac.jp, ††yanai@cs.uec.ac.jp

あらまし　近年、GAN をベースにした画像生成や変換などの技術は急速に進んでおり、本物らしい画像の生成を実現している。しかし、食事の場合は食材や作り方によって、見た目や形は大きく変わるため、リアルな食事画像を生成することは容易ではない。本研究では、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する Mask-based Recipe Embedding GAN (MRE-GAN) を提案する。MRE-GAN では、最先端のクロスモーダルレシピ検索 RDE-GAN をベースに、画像生成に SEAN 正規化を使用することで、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を行う。実験によって、提案手法はより高画質の画像生成を示し、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した。

キーワード　クロスモーダルレシピ検索、クロスモーダルエンベディング、食事画像生成

1. はじめに

近年、深層学習技術の発展により、画像生成や変換などの技術が盛んに研究されるようになってきている。その中でも、GAN (Generative Adversarial Networks) をベースにした画像生成の技術は急速に進んでおり、本物らしい画像の生成を実現している。しかし、食事の場合は食材や作り方によって、見た目や形は大きく変わるため、リアルな食事画像を生成することは容易ではない。

一方、健康とダイエットへの関心が高まり、食事画像認識に関する研究は大きな注目を集めている。特に、大規模な食事データセット Recipe1M の公開により、食事画像と料理レシピの間のクロスモーダルレシピ検索が盛んに研究されるようになってきている。クロスモーダルレシピ検索は画像とテキストを同じ埋め込みスペースに投影することで、モダリティ不変な表現を学習し、高精度なレシピ検索を実現した。

本研究では、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する Mask-based Recipe Embedding GAN (MRE-GAN) を提案する。MRE-GAN では、画像生成モデル GAN をベースにした正規化技術 SEAN [1] と最先端のクロスモーダルレシピ検索の技術 RDE-GAN [2] を融合することで、リアルな食事画像の生成を実現する。実験によって、レシピ情報と形状マスクを用いた高品質な食事画像の生成が可能であることを示す。

2. 関連研究

2.1 クロスモーダルレシピ検索

図 1 に示すように、クロスモーダルレシピ検索とは、画像からレシピへおよびレシピから画像へのモダリティを跨いだ相互の検索を行うことである。初期の研究 JE (Joint Embedding) [3] では大規模のデータセット Recipe1M を作成し、ペアワイズコサイン損失とセマンティック正則化制約を組み合わせることによる共同埋め込み学習アプローチを提案した。

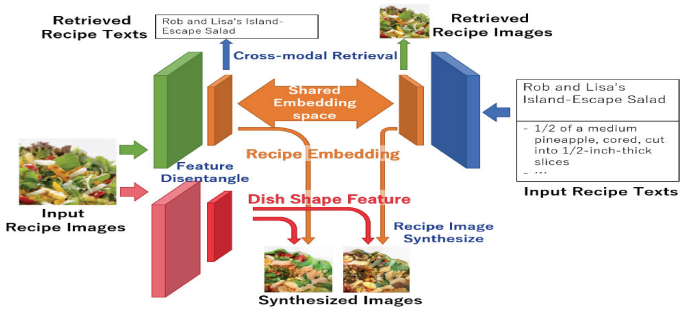

図 1　クロスモーダルレシピ検索の概要

一般的なクロスモダール共同埋め込み学習を行う際には画像を画像の意味的ベクトルにエンコードするための CNN を用意する。一方、レシピテキスト情報に対して、材料リストを双方向性 LSTM [4] で、調理手順を階層 LSTM でエンコードし、この二つを結合したものをテキストの意味的ベクトルとする。そして、画像の意味的ベクトルとテキストの意味的ベクトルをジョイントスペースにエンベディングするように距離学習を行うことにより、画像とテキストのマルチモーダル検索を実現した。JE [3] で実装した Word2Vec [5] でレシピのテキスト情報をエンコード済みのベクトルを本研究の実験で使用した。

上記の JE [3] を改良した手法として、SAN [6] は、画像内の食事領域を特定するためにスタックアテンションネットワークを適用し、共同埋め込み特徴を学習するようにした。その後、いくつかの改良手法では、正弦損失オプティマイザーを置き換えることでパフォーマンスの向上を示した。AMSR [7] は、レシピの階層的アテンションを単純なトリプレット損失で使用した。AdaMine [8] は、共同潜在空間のレシピと画像の埋め込みの両方ですべてのトリプレット損失をバッチ処理化し、クラスガイドな特徴を活用した。また、最近の Modality-Consistent Embedding Network (MCEN) [9] は学習と推論の手順を簡素化し、階層的なアテンションに基づいたテキストレシピ用のタスク固有のエンコーダーを導入した。

2.2 レシピテキストからの食事画像生成

レシピテキストから画像を生成することは、本質的に難しい問題であるが、Zhu らの R2GAN [10] はクロスモダールレシピ検索と GAN 画像生成を連携し、互換性のあるクロスモーダルエンベディング特徴を敵対的に学習することによって、レシピから食事画像への生成を実現した。

同年に Wang らの ACME [11](Adversarial Cross-Modal Embedding) でも、クロスモーダルレシピ検索の改良手法として、クロスモーダル埋め込みと敵対的学習を組み合わせたバッチハードトリプレット損失 [12] を提案し、レシピエンベッディングから画像の再構成を実現した。この手法は R2GAN [10] と同様に画像生成の敵対的学習によって検索精度を向上したが、２つの違いが存在している。一つ目、ACME [11] では学習時にテキストエンベディングのみを画像再構成に使用することであり、代わりに画像エンベディングは multi-label classifier によって、レシピのカテゴリや食材を予測することで、より精度を向上させた。二つ目、ハードサンプルマイニング [13] の戦略を利用したトリプレット損失では異なるモダリティのエンベディングから、正のペア或は負のペアを作って正のサンプル同士の距離を近く、負のサンプルの距離を遠くすることでベクトル表現がより意味を捉えたものになり、より効率的に距離学習を行なった。

これらの研究による画像生成の視覚的品質の改善を目標とした杉山らの RDE-GAN [2] は ACME のアイデアを発展させ、画像の特徴をスタイルとシェープに分離させるような項目を追加することで、さらなる検索精度の向上と直感的な可視化を実現した。しかし、画像の特徴を分離させるように、RDE-GAN [2] モデルの学習は two-stage に分けなければならないため、モデル全体の学習が不安定になる可能性がある。そこで、本研究は特徴分離の考え方を参考に、形状マスクに基づく一つの画像エンコーダを使用し、安定的な one-stage 学習と更なる可視化の品質と解像度を向上することを目指す。

同じくレシピテキストからの食事画像を目的とした CookGAN [14] は R2GAN を発展させ、食材と調理方法の相互作用に基づいて、段階的に料理の画像をアップサンプリングし、高解像度で生成することができた。更に、CookGAN [14] はレシピのテキスト情報に含まれた因果関係を探究することで、生成画像により詳細な視覚効果を保持できるようにした。

2.3 セマンテックマスクからの画像生成

本研究では、さまざまな画像から画像への変換の問題の中で、セマンティック画像合成は、入力セマンティックレイアウト画像を変更することでユーザーが簡単に制御できるため、特に有用な研究であると考えられる。

レイアウトから画像合成の代表的な研究として、Park らの SPADE [15] では、マスクを最初に埋め込みスペースに投影され、次に conv 層を通して変調パラメータ α と β を計算する。以前の条件付き正規化の方法とは異なり、α と β はベクトルではなく、空間次元のテンソルである。この計算された α と β はそれぞれ要素ごとに正規化されたアクティベーションに乗算、加算される。しかし、SPADE [15] は、画像の全体的なスタイルを制御するために 1 つのスタイルコードのみを使用する。

続いて Zhu らの SEAN [1] は SPADE [15] を改良し、領域ごとのスタイルコードを計算することを可能にし、更なる高品質な画像合成を示した。つまり、セマンティック領域適応正規化（SEAN）と呼ばれる新しい条件付き正規化手法はスタイル入力画像を使用してセマンティック領域ごとに空間的に異なる正規化パラメーターを作成できる。この空間的に異なる正規化パラメータがセマンティックレイアウトと入力スタイル画像に依存することである。本研究は、SEAN [1] のセマンティック領域ごとのエンコーディングを参考に、食事画像の食事領域と背景領域の画像特徴を分離できるようにした。異なる点として、画像生成の時に用いられるのは画像から直接的に抽出された特徴ではなく、レシピ検索用のエンベディングである。

3. 手　　法

3.1 概　　要

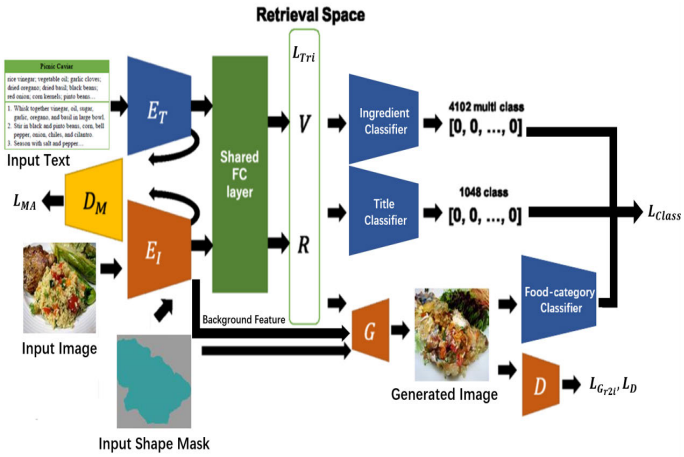

図 2　提案手法の概要図

本研究では、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する Mask-based Recipe Embedding GAN (MRE-GAN) を提案する。MRE-GAN では、杉山らの RDE-GAN [2] をベースに、画像生成に SEAN [1] 正規化を使用することで、レシピエンベッティングを用いて、形状マスクに基づく食事画像の生成を実現する。MRE-GAN では、複数品食事画像を自由なレイアウトで生成することが可能となる。

提案モデルは図 2 の示すように、左側のレシピ検索と右側のテキスト予測と画像再構成の構造になっている。食事領域の特徴のみを共有空間への埋め込みを行うことで、生成画像の品質の向上に繋がるという考えから、画像特徴を形状マスクに基づいて、食事領域と背景領域の２つに分離する手法を提案する。その理由は食事画像の特徴には料理の領域と、皿やフォークなどの無関係な背景が含まれていることが問題点であるため、料理部分のマスクを提供することで、料理の特徴のみをクロスモーダル検索の学習に使う。画像とテキストの共有空間への埋め込み表現を求め、その間の距離学習を通じてレシピ検索を行う。画像とテキストは共有空間へエンベッディングされた後、画像エンベディングからテキストの再構成にあたるタイトルと材料をそれぞれラベルのクラス分類器による予測と、テキストエンベディングから画像の再構成にあたる Generator に SEAN [1] 正規化で画像生成を行う。

先行研究より提案手法は一つの Per-Region Style 画像エンコーダ Ev を使用することで、one-stage で安定的な学習が可能にした。また、従来の RDE-GAN では画像のスタイルと形状の分離が不十分である問題に対し、形状マスクを学習データとして提供することで、マスクから Shape に関する情報を獲得する。

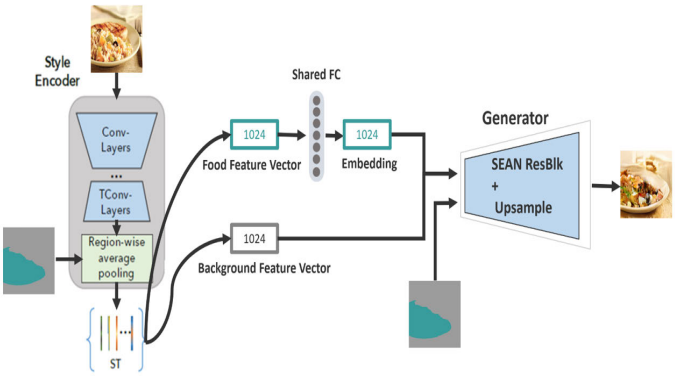

図 3　画像生成の流れ図

3.2　提案手法: Mask-based Recipe Embedding GAN

提案手法は Mask-based Recipe Embedding GAN(Mask-based RE-GAN) と命名する。Mask-based RE-GAN は画像とテキストの 2 つのモダリティの分布を調整してモダリティ不変な特徴表現かつ、画像生成タスクに適したエンベディング表現を取得し、クロスモーダル変換の一貫性を実現する。 4 つの目的関数で最適化され、全体の目的関数は下のように書き表せる。

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

本研究の実験では各損失にかかる重み λ はそれぞれ $\lambda_1 = 1.0, \lambda_2 = 0.005, \lambda_3 = 0.002, \lambda_4 = 0.002$ とする。検索学習コンポーネントは、画像の V とレシピの R の 2 つの高レベルの特徴ベクトルを受け取り、検索損失 $L_{Tri}(V, R)$ を計算する。モダリティアライメントコンポーネント $L_{MA}(V_m, R_m)$ は最後から 2 番目のレイヤーの特徴 V_m と R_m を入力として、敵対的な損失を使用してモダリティ不変性を実現し、2 つの分布をアラインメントさせることを目的としている。クロスモーダル translation consistency コンポーネント $L_{G_{r2i}}$ は、さらに image2recipe (V に基づいて成分を予測) と recipe2image (R からの画像生成) の 2 つのサブコンポーネントに分割される。それぞれ敵対的損失と分類損失を使用して最適化される。

3.3　画像とテキストの共有空間へのエンベディング

ペアとなった食事画像と形状マスクを Per-Region スタイルエンコーダに入力し、画像特徴は食事と背景の二つに分離する。この特徴量はそれぞれ 1024 次元のベクトルである。レシピテキストは RDE-GAN [2] と同様に材料リストを双方向性 LSTM、調理手順を階層 LSTM で意味ベクトルにエンコードする。その後、食事スタイルコードとテキストからの意味ベクトルを、重み共有の全結合層を通すことで各モダリティの表現を相互に関連付け、共有空間にエンベッディングされる。エンベッディング同士の距離学習は ACME [11] の提案した hard sample mining トリック [13] で triplet loss を計算する。これによって、レシピ検索用のレシピエンベディング R と画像エンベディング V を求める。次に、レシピのエンベディング表現 R を使用して食品画像を生成し、画像のエンベディング表現 V を使用して特定のインスタンスの材料を予測する。

最初のクロスモーダル最適化ステップでは、textual-visual の情報の早期アラインメント学習を実行することを目的とし、正のペアとなったレシピと画像から潜在的な特徴を抽出する。異なるモダリティからのエンコードされた特徴は分布が大きく異なる可能性があり、その結果、一般化が不十分になり、収束が遅くなってしまう。そこで、エンコードされた特徴の分布を調整するために、識別器 D_M を用意し、特徴表現が画像から取得されたかレシピから取得されたかを区別できないように、学習させる。この Modality Alignment loss は以下のように表す。

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

テキストエンコーダ E_T の出力, 画像エンコーダ E_I の出力を共有全結合層 FC に入力し, その出力のレシピエンベディング R と画像エンベディング V がクロスモーダル共有エンベディングとなるように学習する。ACME [11] と同様にハードサンプルのトリプレット損失を使う。トリプレット損失 [12] は、トレーニングサンプルのトリプレット (x_a, x_p, x_n) で計算される。ここで、x_a は一つのモダリティのアンカーポイントであり、対応するモダリティの埋め込みを評価するためのグラウンドトゥルースとして使用される。x_p と x_n は他のモダリティからの正と負の特徴の埋め込みを示す。トリプレット損失により、一方のモダリティの正のインスタンスがもう一方のモダリティのアンカーポイントに近くなり、一方のモダリティの負のインスタンスがもう一方のモダリティのアンカーポイントから離れていることが確保される。Hermans ら [13] は、トリプレット損失を計算するときに各バッチ内の各アンカーポイントに対して最もハードな正と負のサンプルを選択することにより、パフォーマンスの向上を示した。トリプレット損失は以下の式により表す。

$$\begin{aligned} L_{Tri} = {} & \sum_{V} [d(V_a, R_p) - d(V_a, R_n) + \alpha]_+ \\ & + \sum_{R} [d(R_a, V_p) - d(R_a, V_n) + \alpha]_+ \end{aligned} \qquad (3)$$

3.4　画像生成とテキスト予測

先行研究によって、一方のモダリティの学習された埋め込みが、もう一方のモダリティの対応する情報を回復できるようにし、セマンティックアラインメントを向上させることが示されている [11]。これにより、クロスモーダル変換の一貫性が強化され、学習されたエンベディングの表現を向上することが示された。具体的には、レシピエンベディング R を使用して食品画像を生成し、視覚的なエンベディング V を使用してレシピの成分を予測することである。そこで、本研究は高い表現力の持つレシピエンベディングを用いて、条件付き GAN を追加し、形状マスクに基づいた高品質な食事画像の生成を目指す。各意味的領域の特徴抽出元画像から食事領域と背景領域のそれぞれの特徴を抽出するために、入力画像の各セマンティック領域から対応するスタイルコードを同時に抽出する新しいスタイルエンコーダネットワークを使用する。

図 3 の示すように、スタイルエンコーダの入力は食事画像と対応するマスク画像であり、スタイルエンコーダの出力は、1024 次元のスタイル行列 ST である。ここで、単純なダウンスケーリング畳み込みニューラルネットワークで構築された標準のエンコーダーとは異なり、領域ごとのスタイルエンコーダーは、「ボトルネック」構造を採用し、入力画像からスタイルに関係のない情報を削除する。また、スタイルはセマンティック領域の形状に依存しないという事前の知識を取り入れてネットワークブロック TConv-Layers によって生成された中間特徴マップ（1024 チャネル）を領域ごとの平均プーリング層に渡し、1024 次元ベクトルに減される。

3.5 レシピエンベディングと形状マスクからの画像生成

クロスモーダルレシピ検索で獲得したレシピエンベディングからの画像生成はまず、画像生成に使用するスタイル特徴は食事スタイルベクトル（検索のレシピエンベディング）と背景スタイルベクトルを結合した行列に変更する。そして、既存の別の方法とは対照的に、SEAN によって学習される変調パラメータは、スタイルコードとセグメンテーションマスクの両方に依存するため、最終的にジェネレータは結合した行列と形状マスクを入力として、セマンティック領域適応正規化（SEAN）と呼ばれる条件付き正規化手法を使用することで、各セマンティック領域に画像のスタイルの詳細な制御を可能にする。

画像再構成のトレーニングではスタイルエンコーダは、対応するセグメンテーションマスクに従って、入力画像から領域ごとのスタイルコードを抽出するようにトレーニングされている。 ジェネレータは、結合した領域ごとのスタイルコードと対応するセマンティックマスクを入力として、対応する食事画像を再構成するようにトレーニングされている。 SPADE [15] と SEAN [1] に続いて、入力画像と再構成された画像は Adversarial loss、Feature matching loss、Perceptual loss の 3 つの損失項で構成される損失関数 $L_{G_{r2i}}$ によって定義される。

損失関数 $L_{G_{r2i}}$ の詳細は以下の式に示す。

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

先行研究の SPADE [15] と SEAN [1] によって、$\gamma_1 = \gamma_2 = 10$ と設置する。

$L_{G_{r2i}}$ を更に 3 つのサブロスに分けて説明する。

Adversarial loss について、E をスタイルエンコーダー、G を SEAN の生成器、D1 と D2 を異なるスケールの 2 つの識別器、R を与えられたスタイル画像、M を R の対応するセグメンテーションマスク、ST をレシピエンベディングを用いたスタイルコードとする。条件付き敵対学習を定式化すると、次のように表す。

$$\min_{E,G}\max_{D1,D2}\sum_{k=1,2} L_{GAN}(E,G,D_k) = E[\max(0, 1 - D_k(R,M))] + E[\max(0, 1 + D_k(G(ST,M),M))] \tag{5}$$

Feature matching loss について、T を弁別器、D_k を層の総数、$D_k^{(i)}$ を出力特徴マップ、N_i を i 番目の層の要素数としてする。特徴マッチング損失 L_{FM} を次のように表す。

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

Perceptual loss について、N を知覚損失の計算に使用される層の総数、$F^{(i)}$ を VGG ネットワークの i 番目の層の出力特徴マップ、M_i を $F^{(i)}$ の要素数とする。知覚損失 $L_{percept}$ を次のように表す。

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

3.6 クラス分類損失

敵対的損失はリアルな画像を生成できるが、変換の一貫性を確保することができないため、food-category classifier を使用する。これによって、ジェネレータが適切な対応するフードカテゴリの食事を生成できるように促す。

もう一方、食品画像の成分を予測する視覚的特徴 V に multi-label classifier を適用することによって、画像エンベディングを正しい食品カテゴリに分類できるようにすることで、変換の一貫性を維持する。

レシピのクラスは、タイトルで作成した 1048 クラスの分類と、材料リストで作成した 4102 クラスのマルチクラス分類を行う。タイトルの分類器と材料の分類器によるクロスエントロピーロスをそれぞれ L_{Title}、L_{Ingr}、クラス分類器を C、画像からエンコードされたエンベッディングを V 、テキストからエンコードされたエンベッディングを R、タイトル分類の正解ラベルを L_t、材料分類の正解ラベルを L_i として、損失関数を下のように表す

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

4. 実　　験

4.1 データセットと評価指標

実験では 100 万件以上のレシピとその画像からなる Recipe1M [3] データセットを使用する。形状マスクデータセットは 2 つの画像領域分割手法によって自動的に Recipe1M の全画像について作成する。 1 つ目は食事画像領域分割データセット UEC-FoodPix Complete [16] で学習した DeepLabV3+ モデル [17] による領域分割された Recipe1M の形状マスクである。 2 つ目は本部ら [18] の Few-Shot Segmentaion 手法による領域分割された Recipe1M の形状マスクである。こちらでも，UEC-FoodPix Complete を学習データとして利用している．

生成された画像の品質評価にはフレシェ開始距離（FID）を用いる。FID は実際の画像と生成された画像の間の特徴距離を測定するための最も一般的なメトリックの 1 つである。

4.2 実験設定とベースライン手法

本実験では提案モデルを学習して実験を行う。モデルの学習に関しては学習率を初期値 0.0001 とし、50 エポック経過後は 0.00001 とし、合計 100 エポック学習を行った。各 loss 関数のパラメータは $\lambda_1 = 1.0, \lambda_2 = 0.005, \lambda_3 = 0.002, \lambda_4 = 0.002$ とする。生成画像の解像度は 256x256x3 である。すべてのディープニューラルネットワークは Pytorch プラットフォームにおいて実装される。

生成画像の品質評価の実験では、ベースライン手法の ACME [11]、RDE-GAN [2] の画像生成と比較する。これらのベースライン手法は全て公式ホームページで公開した学習済みモデルを用いて、実装した。

4.3 画像生成の評価

生成画像の品質を定量的に評価するために、FID を用いて、元画像の分布と生成画像の分布がどれだけ近いかを測る尺度であり、ベースラインと比較した。また、提案モデル Ours($Mask_{DeepLabV3+}$) の学習には大量の食事データで学習した DeepLabV3+ モデルによる領域を分割された Recipe1M の形状マスクを使用した。提案モデル Ours($Mask_{Few-Shot}$) の学習には本部ら手法によって作成した Recipe1M の形状マスクを使用した。

既存手法と公平な比較をするために、提案手法による生成画像は 256x256 の解像度であるが、既存手法と同じの 128x128 の解像度にリサイズした後、FID を計算する。実験ではテキスト

表 1　本手法と既存手法の FID 比較

手法	FID ↓@画像から	FID ↓@テキストから
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

エンベディングからの画像再構成と画像エンベディングからの画像再構成の両方を行う。これらの結果を表 1 に示すように、Few-shot segmentation の手法による生成した形状マスクを用いた場合の提案手法は他の手法よりも FID が小さく、従来手法を上回る画像品質を達成したと言える。

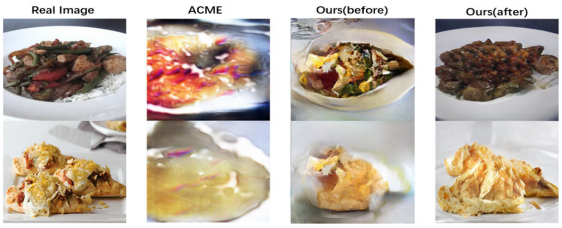

図 4　ベースライン手法による生成画像との比較

生成画像を定量的に評価するために、ベースライン手法と生成画像の視覚的品質の評価として、図 4 のような比較実験を行なった。一列目は画像生成に使用するレシピエンベディングに対応する現実の食事画像であり、二列目はベースライン手法 ACME [11] による生成画像である。右からのの二列は異なる形状マスクで学習した提案モデルによる生成画像であり、この形状マスクに関してはそれぞれ DeepLabV3+ [17] による Recipe1M [3] の形状マスクデータセットと本部ら [18] の作成した Recipe1M [3] の形状マスクデータセットである。二列目と四列目を比較すると、提案手法の方が食事の形状を保ちながら、リアルな食事画像を生成できることが示されている。

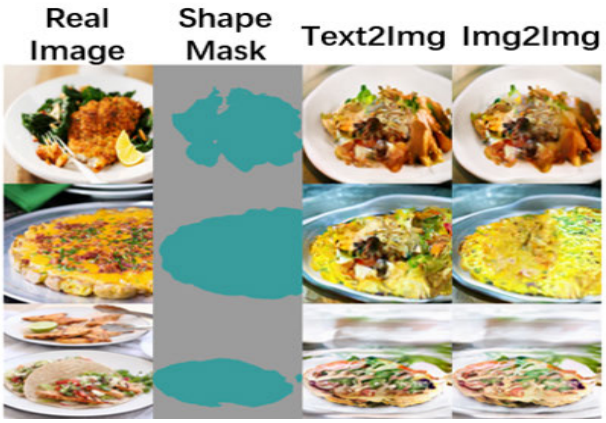

図 5　双方向からの食事画像生成

学習した提案モデルは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視覚的に検証するために、それぞれ意味的エンベディングと視覚的エンベディングを用いて、対応する形状マスクに基づいた食事画像の生成を行う。双方向からの食事画像を生成した結果を図 5 に示すように、ここでは画像のみを表示しているが、対応したテキストからの意味的エンベディングもあわせて行った。三列目と四列目の生成画像を比較すると、テキストエンベディングと対応する画像エンベディングによる生成された画像は同じ食事であると考えられる。また、生成された画像の豊富な視覚的良さから、高次元意味的エンベディングの表現力が反映されていると考えられる。

4.4　様々な食事画像生成

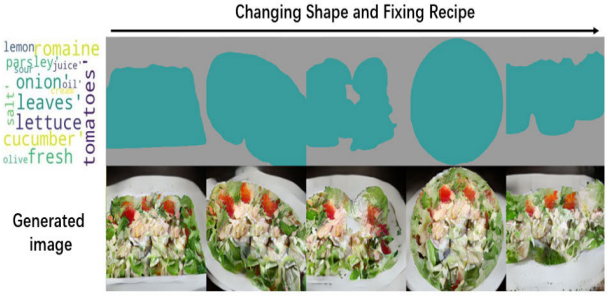

図 6　異なる形状マスクによる食事画像生成

異なる形状マスクから食事画像の生成を行う。その結果を図 6 に示す。この結果から、提案モデルによる画像生成の際は形状と意味的特徴を同時に考慮した生成結果を示している。しかし、食器の領域マスクは提供していないため、食器は食事マスクを囲むように生成され、生成された食器は歪んだケースが存在する。

図 7　入力テキストの操作による生成画像の変化

次に、入力の食材の変更による生成画像へどんな影響があるのかを検証するための実験を行う。図 7 の結果により、入力の食材のテキストからそれぞれ左の枝豆、或は、右のキャベツを削除した場合に、生成された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ことが確認できる。

そして、レシピエンベディングを連続的に変化させた際に生成画がどのように変化するのかを検証する。レシピエンベディングの空間を連続的に表現できていれば、生成した補完画像は滑らかに変化する。この結果を図 8 に示す。入力としてレシピエンベディングと対応する形状マスクであり、目標として右側の列目に示す画像である。三行目では、入力の形状マスクを保ちながら、ピザの色や質感などの要素が意味的エンベディングの変化と同時に、生成画像は滑らかで連続的に変化したことが分かる。むしろ画像生成に必要な高レベルな特徴の表現力も提案のモデルによって、得られていることが考えられる。

複数品の食事画像生成を行なった結果は図 9 に示す。事前に人工で作成した複数のセマンティック領域のあるマスク画像を

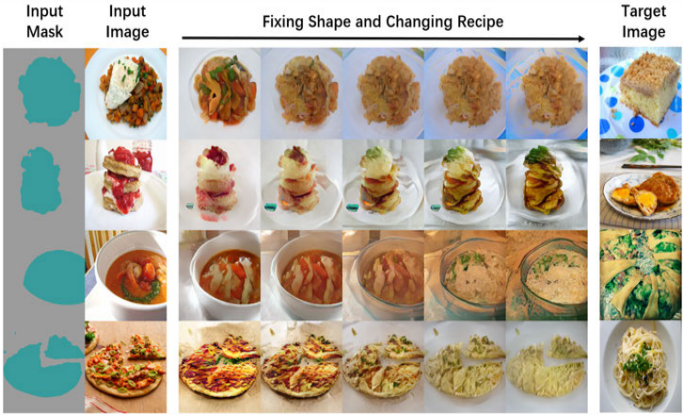

図 8　スタイルミックスによる生成画像

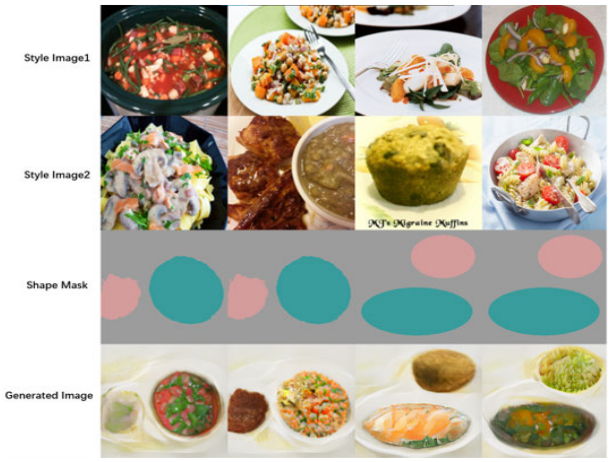

図 9　複数品の食事画像生成

用いて、それぞれレシピ 1 からのエンベディングを緑色領域のスタイルとして、レシピ 2 からのエンベディングをピンク色領域のスタイルとして、複数品の食事画像生成を行なった。この結果から、複数料理の画像生成が成功したことが示されている。

5. おわりに

本研究では、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する Mask-based Recipe Embedding GAN (MRE-GAN) を提案した。MRE-GAN では、杉山らの RDE-GAN [2] をベースに、画像生成に SEAN [1] 正規化を使用することで、クロスモーダルレシピエンベッティングによる、形状マスクに基づく料理の画像生成を実現した。既存研究に存在している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 で安定的な学習を可能にした。実験から、提案手法は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し、より高画質の画像生成を示した。

今後の課題として、複数品の生成画像には全体の一貫性を保持できないため、レシピ間の関係性を考慮することから検討していく必要がある。また、現在は食器の形状を指定することができないので、Recipe1M の画像に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ことを検討している。

文　献

[1] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[2] Yu Sugiyama and Keiji Yanai. Cross-modal recipe embeddings by disentangling recipe contents and dish styles. In *Proc. of ACM Multimedia*, 2021.

[3] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. Learning cross-modal embeddings for cooking recipes and food images. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[4] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, Vol. 18, No. 5-6, pp. 602–610, 2005.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013.

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

[7] Jing-Jing Chen, Chong-Wah Ngo, Fu-Li Feng, and Tat-Seng Chua. Deep understanding of cooking procedure for cross-modal recipe retrieval. In *Proc. of the 26th ACM international conference on Multimedia*, pp. 1020–1028, 2018.

[8] Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 35–44, 2018.

[9] Han Fu, Rui Wu, Chenghao Liu, and Jianling Sun. Mcen: Bridging cross-modal gap between cooking recipes and dish images with latent variable model. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020.

[10] Bin Zhu, Chong-Wah Ngo, Jingjing Chen, and Yanbin Hao. R2gan: Cross-modal recipe retrieval with generative adversarial network. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2019.

[11] Hao Wang, Doyen Sahoo, Chenghao Liu, Ee-peng Lim, and Steven C. H. Hoi. Learning cross-modal embeddings with adversarial networks for cooking recipes and food images. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[12] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015.

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

[14] Bin Zhu and Chong-Wah Ngo. Cookgan: Causality based text-to-image synthesis. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020.

[15] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

[16] Kaimu Okamoto and Keiji Yanai. UEC-FoodPix Complete: A large-scale food image segmentation dataset. In *Proc. of ICPR Workshop on Multimedia Assisted Dietary Management (MADIMA)*, 2021.

[17] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proc. of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

[18] Yuma Honbu and Keiji Yanai. Few-shot and zero-shot semantic segmentation for food images. In *Proc. of the 13th International Workshop on Multimedia for Cooking and Eating Activities*, 2021.