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.

食事画像領域分割データセットの作成とその活用

岡本　開夢†　Cho　Jaehyeong†　會下　拓実†　柳井　啓司†

† 電気通信大学大学院情報理工学研究科　〒 182–8585 東京都調布市調布ヶ丘 1-5-1
E-mail: †{okamoto-ka,cho,ege-t}@mm.inf.uec.ac.jp, ††yanai@mm.inf.uec.ac.jp

あらまし　現在、領域分割の画像データセットは多数公開されているが、そのなかで食事におけるカテゴリは少数に限られている。また、食事画像データセットも多数公開されているが、画像毎に単独の食事名がアノテーションされたものが大部分である。また、画像に複数の食事がバウンディングボックス付きでアノテーションされているのは UECFood などごく少数に留まっており、さらに詳細に画素毎にアノテーションされたデータセットで大規模なものは存在していない。一方、食事画像における正確な食事量やそのカロリー量の推定には、画像内における食事領域の面積に対して関連しており、食事領域の領域分割が不可欠である．そこで、本研究では UECFood-100 のうち 1 万枚の画像に画素単位のアノテーションを付加しマスク画像を作成することで、領域分割モデルの学習データセットとして利用できるようにした。GrabCut を用いることでアノテーションコストを削減した。さらに、555 枚のラーメン画像だけスープやトッピングを考慮し詳細にアノテーションを行い、ラーメン画像に対して細かな領域分割及び領域マスクからの食事画像生成することで、食事マスク画像における活用例を示した。
キーワード　深層学習, 食事領域分割データセット, ラーメンデータ・セット, 画像生成

1. はじめに

現在、深層学習の発達により画像認識の精度が飛躍的に向上した他、画像生成や領域分割といったタスクにおいても優れた成果を残している。深層学習による教師あり領域分割においては、学習画像に画素ごとにアノテーションされたマスク画像データセットが必要となる。

領域分割に必要な大規模データセットとして PASCALVOC 2012 [1] や MS COCO [2] が挙げられるが、これらに含まれる画像は動物や乗り物のラベル付けされたものが大半を占めており、食事については少数のカテゴリに限られている。また単体の食事データセットにおいては単独の食事名がアノテーションされたものが大部分で、UECFood-100 [3] といった画像内の複数の食事に対してバウンディングボックスがアノテーションされているものは少数であり、画像ごとにアノテーションされたデータセットで大規模なものは存在していない。

一方、食事においては人々の生活に不可欠なものであり健康面を意識した食事量やカロリー量の推定が盛んにおこなわれている。それに伴い、現在食事画像からカロリー量を推定するようなアプリケーションが存在している。これら多くの場合には、食事画像から推定したカテゴリをもとに標準的なサイズにおけるカロリーや、ユーザーが入力したサイズをもとにカロリーを算出しており、直接画像から面積を算出しているものは限られている。しかし、カロリー量と食事量はともに密接に関連しており、食事画像の領域分割による食事量の推定なくしては正確なカロリー量を推定することは困難である。

また画像生成分野においては、実際の画像に近しい画像が生成することのできる Generative AdversarialNetworks(GAN) [4] の登場により大幅に改善されている。GAN はノイズから画像を生成するジェネレータとその画像の真偽を判定するディスクリミネータを敵対的に学習することで画像を生成を行っている他、その注目度の高さから様々な画像で使用されている。数文字や顔画像の生成が有名であるが、食事画像においてもマスク画像からの生成や、ある料理から別の料理の変換が行われている。

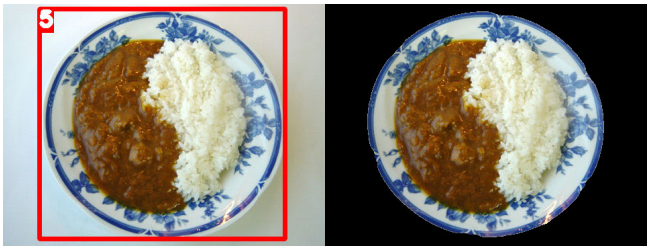

図 1　バウンディングボックスをもとに GrabCut を適用した例

本研究では既存の UECFood-100 のうちの 1 万枚の画像に画素単位でのアノテーションを付加し、領域分割モデルの学習用食事データセットとして作成を行う。まず UECFood-100 に既存のバウンディングボックスから新たに複数品目を考量したバウンディングボックスを付加する。次に、作学習成したバウンディングボックス内の領域に対して、GrabCut [5] を用いて初期マスク画像を作成する。GrabCut は領域の短形を与えることで前景と背景を分割する手法であり、これにより作成したマスク画像を精査することで領域分割用の食事データセットの作成を行う。図 1 は UECFood-100 に GrabCut を適用した例である。

また、別に食事マスク画像の活用例として画像生成を行う。

生成する画像は食事画像のなかで代表的なものとしてラーメンとし、作成する食事学習データとは別にアノテーションした555 枚のうち 500 枚を学習画像として用いる。ここではマスク画像から画像生成を行うタスクに注目し、実画像から領域分割によって得られたマスクを元に画像の生成を行う。用いる学習画像は食事領域に対してさらにスープやトッピングなどを考慮した 15 クラスのアノテーションを付加していることから、生成されるマスク画像は各クラスの領域を考量したマスク画像であり、このマスクを修正することで自由なラーメン画像も作成可能である。本研究で作成した画像データセットは近日中に http://foodcam.mobi/dataset/にて公開予定である。

2. 関 連 研 究

領域分割に用いられるデータセットとして PASCALVOC 2012 [1] や MS COCO [2] がベンチマークとして頻繁に用いられている。PASCALVOC は 2005 年から 2012 年に行われたコンペティションに用いられたデータセットで 2012 年版では飛行機や自転車を含む 22 クラス 9,993 枚を含む。MS COCO は Microsofot 社が提供するデータセットで 80 クラス 33 万枚の画像を含む。しかし食事クラスに着目した際に MS COCO に含まれるピザやホットドックといったわずか 10 クラスしか存在しておらず、食事領域分割モデルの学習画像としては限られたクラスにしか対応出来ていない。

一方、食事画像を用い食事量やカロリー量を考慮した研究がなされているが、岡元ら [6] の研究が挙げられる。これは、大きさが既知の基準物体とともに食品を上から撮影することでカロリー量を推定するシステムを考案している。まず、GrabCut [5] を使用し、画像内の基準物体と食事の領域を抽出し、それらを比較することで食事領域の実面積を推定する。その後、事前に作成したカロリーと面積の回帰式をもとにカロリー量の推定を行っている。本研究においてもマスク画像を作成する際に GrabCut [5] を用いる。次に、Myers らによる Im2Calories [7] が挙げられる。これはＣＮＮベースのカロリー量推定の研究であり、単一の画像からネットワークを用いて食事領域と深度情報を抽出し、それをもとにボクセル化した食事領域からカロリー量の推定を試みている。しかし 3 次元ベースの大規模なカロリー量付きデータセットが不足していたため、実際には 3 種類程度の食事のカロリー量の推定しか実現されなかった.

またマスク画像から画像を生成す研究として Park [8] らの研究が挙げられる。これは、conditional normalization 手法、Spatially-Adaptive Denormalization(SPADE) を提案した研究で、セグメンテーションマスク画像からの情報をネットワークの noramlization に与えることで、よりリアルな画像を生成することが可能となっている。マスク画像を用いる点で、本研究で用いるラーメン画像生成と類似な研究である。また、本研究で用いるラーメン画像生成はユーザーが自由に修正でき、その際に用いるマスク画像は輪郭があいまいなスケッチ画像ともとらえることができる。そこで、スケッチ画像から画像を生成した研究として、Chen らの SketchyGAN [9] が挙げられる。これは Masked Resifual Unit(MRU) と呼ばれる構造を提案しており、エッジ画像と実画像のペア画像を入力し学習したのちスケッチ画像と実画像のペア画像を用いてスケジューリングすることで、50 クラスの画像の生成を行っている。

3. データセット

本研究では既存の UECFOOD-100 [3] に対してインスタンスを考慮した新たなバウンディングボックスを付加した画像をもとにマスク画像 (UECFoodPIX) の生成を行う。マスク画像の生成には GrabCut [5] を用いる。また、SNS から収集したラーメン画像に対して、スープやトッピングを考慮したアノテーションを行うことで領域分割、画像生成用ラーメンデータ (UEC-Ramen555) の作成を行う。

3.1 UECFoodPIX

現在、多くの食事画像データセットが公開されており、食事分類タスクにおいては Food-101 [10]、UECFood-100 [3] や UECFood-256 [11] などが標準的なベンチマークとして用いられている。しかしその中でも、画像内の複数の食事に対してバウンディングボックスを持つものに関しては、UECFood-100/256 など少数に限られている。またこれらのバウンディングボックスは、UECFood-100/256 の食事カテゴリに含まれる 100 種類にとどまっており、この 100 カテゴリに含まれない食事部分にはバウンディングボックスが付けられていない。例としてパントーストとバターを含む画像の場合、バターは 100 カテゴリに含まれていないことから、パントーストの領域のみにバウンディングボックスが設けられており、バターの領域にはバウンディングボックスが設けらていない。またこのバウンディングボックスはインスタンスを考慮しておらず、図 2 のように画像に多数の寿司が含まれている場合には、多数の寿司を含む形で 1 個のバウンディングボックスのみが与えられているのが現状である。セグメンテーションマス付きの大規模な食事画像データセットに関しては、現状存在しておらず、UNIMIB2016 データセット [12] のみがセグメンテーションマスクにあたる食事領域情報をポリゴンとして提供している。

そこで本研究では、従来の UECFood-100 を拡張しインスタンスベースのバウンディングボックスとセグメンテーションマスクを付与した食事画像データセットの作成を行った。

まず初めに、UECFood-100 に含まれる 10,000 枚の画像に対して、新たに手動でインスタンスベースのバウンディングボックスを付与した。新たに作成したバウンディングボックスの数は従来に比べて 2 倍以上となったため、アノテーションコストを抑えるため手動での作業は、バウンディングボックスを付与するだけにとどまり食事カテゴリの割り当ては行わなかった。その代わりに、新しいバウンディングボックス内の領域と元の UECFood-100 データセットに含まれるバウンディングボックスの内の領域の比 a_o をとり、その比 a_o をもとに、新たなバウンディングボックスに対して自動的に食事カテゴリを付与した。しきい値を 0.5 に設定し、a_o の値がこの値よりも上回る場合には元 UECFood-100 の食事カテゴリを、下回る場合には' その他の食事' カテゴリを割り当てた。a_o の式は以下の通りである。

$$a_o = \frac{area(B_{uecfoodseg} \bigcap B_{uecfood100})}{min\{B_{uecfoodseg}, B_{uecfood100}\}} \quad (1)$$

図 2 に示すように、新しいバウンディングボックスは 100 カテゴリ以外の食品を含めた全ての食品に対して付与されており、複数個の食品に対して一個ずつに適用されている。クロワッサンの画像では、元の UECFOOD では手前のクロワッサンのみにバウンディングボックスが付けられているがが、新たに作成したバウンディングボックスでは奥のクロワッサンにもバウンディングボックスを設けられている。次に、GrabCut [5] を用いて学習用に 9,000 枚、評価用に 1,000 枚の画像に対してセグメンテーションマスクを付与した。全てのセグメンテーションマスクを手作業で付与するにはアノテーションコストがかかりすぎるため、新たに作成したバウンディングボックスをもとに GrabCut [5] を使用して食事領域を抽出することでマスク画像を自動的に作成した。図 3 では作成したセグメンテーションマスク画像の例を示す。自動生成したマスクにおいては、単品品目であるチャーハンの画像のように綺麗に食事領域を抽出されているものも多いが、カツオのたたきやとんかつの画像のように、つまみやキャベツといった領域に対しては画像分類タスクにおける食事のクラス (ここではカツオととんかつ) のマスクが別の領域に付いている場合がある。そのため、図 3 のように作成したマスクを人手で精査することで食事領域分割用データセットとした。

3.2 UEC-Ramen555

現在、公開されている多くの食事画像データセットは食事カテゴリによる分類されたデータセットがほとんどである。しかし、食事は同じカテゴリ内の食事でも用いられる材料が違う場合があり、それによって異なる形や色をしている食事が存在する。その中で、ラーメンは代表な的食事カテゴリであると共に多様なスープと具で構成されている食事で、その材料によって様々な種類に分類される。例えば、一口にラーメンといっても醤油、味噌やとんこつといったスープの種類が豊富である他、チャーシューや卵といった具材の個数も調理の仕方によって異なる。そこで、本研究ではラーメン画像の各要素のカテゴリごとにセグメンテーションマスクを作成して、ラーメン画像とラーメンの要素がピクセルレベルのラベルが付いているマスク画像に構成される画像データセットを作成した。これにより、スープや具材を考慮した領域分類や画像生成が可能となる。データセットは SNS によって収集した 555 枚のラーメン画像とマスク画像のペアであり、マスク画像では背景・5 種類のスープ・レンゲ・箸・チャーシュー・卵・切れた卵・海苔・メンマ・ナルトの 15 クラスで構成されている。マスク画像は人で作成し、図 4 に作成したラーメンデータセットの例を示す。各ラーメン画像のスープはそれぞれ異なってマスクが付加している他、左の画像と右の画像における卵、切れた卵に関してもそれぞれ別々のマスクを付加している。555 枚のうち 500 枚を学習用画像に、残りを評価用画像として、領域分割と画像生成のモデルに用いる。

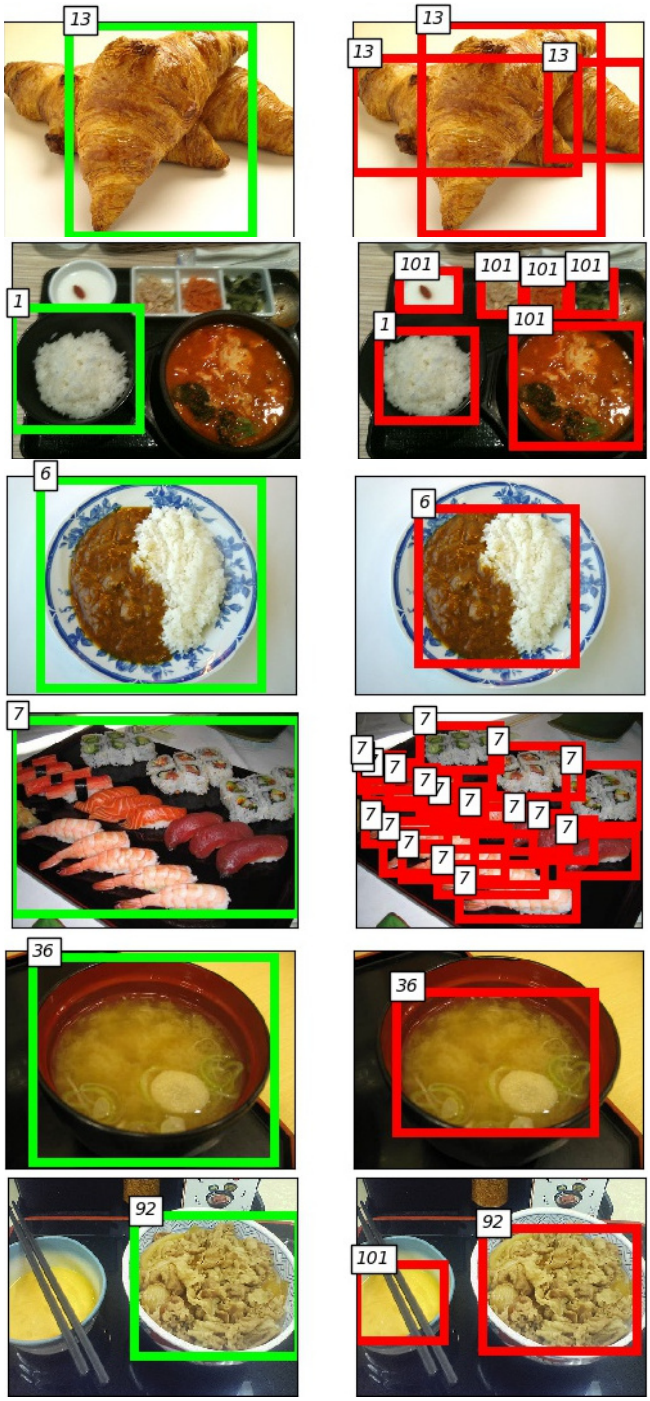

図 2 バウンディングボックスの例．左:UECFood-100 [3], 右:新たに作成したバウンディングボックス．番号は UECFood の食事カテゴリー番号．101 は新たに追加した「その他」カテゴリー．

4. 実　　験

まず、作成した食事領域分割のデータセット (UECFoodPix) を用いた領域分割を行う。領域分割用のモデルは Deeplab V3+ [13] とし Accuracy と mean Intersection over Union (mIoU) による定量評価を行い、評価用データを用いて分類された食事領域を示す。次に、ラーメンマスク画像を用いた領域分割と画像生成を行う。領域分割には同じく DeeplabV3+ を、画像生成には pix2pix [14] 使用し抽出された食事領域と生

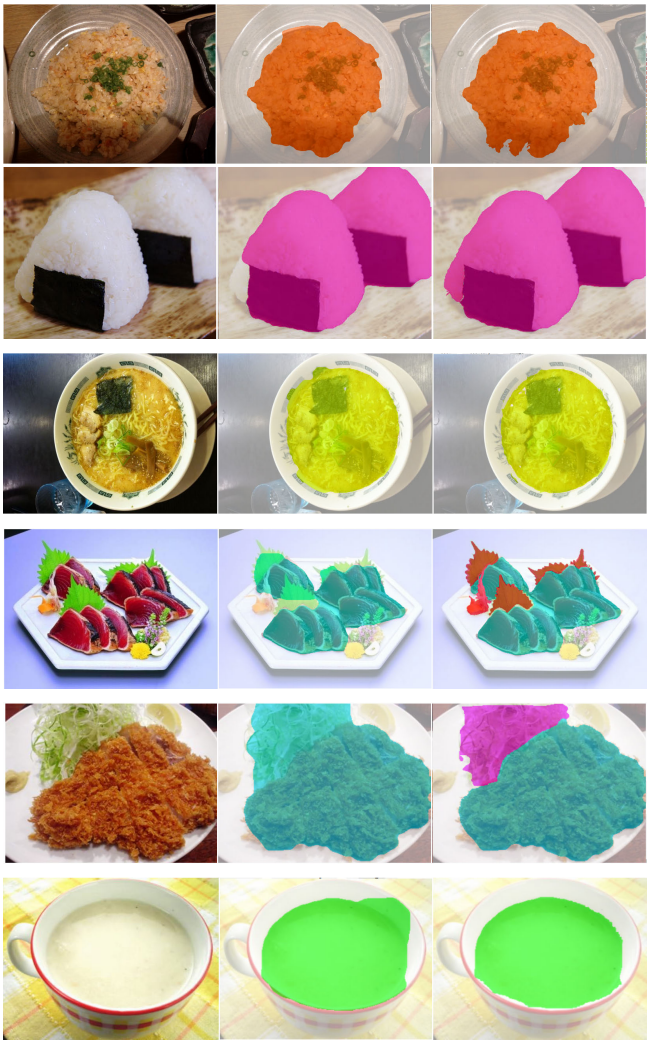

図 3　バウディングボックスの例　(左:実画像, 中:自動生成したマスク画像, 右:作成したマスク画像

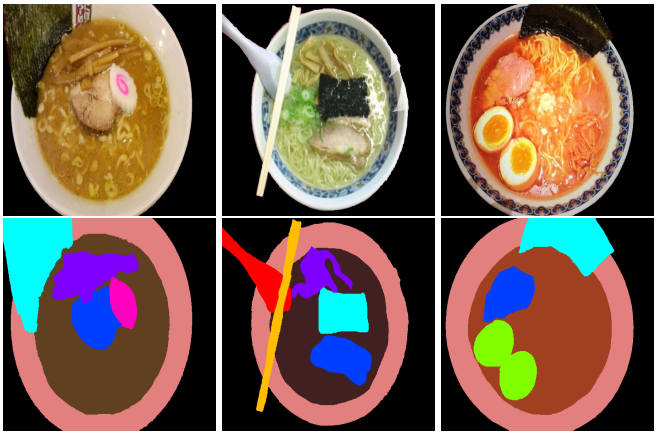

図 4　ラーメン画像のデータセットの例 (上：元画像、下：各要素を領域分割したマスク画像)

成画像を示す。

4.1　領域分割

作成した食事画像データセットのベンチマークテストとして、Chen らによって提案された Deeplab V3+ [13] による食事領域抽出を行った。Deeplab V3+は複数のスケールで格子状に分割しそれぞれの画像に対して畳み込みを行うピラミッド構造とエンコーダデコーダモデルを組み合わせたセマンティックセグメンテーションモデルであり、領域分割手法として頻繁に使われる手法である。今回のこの Deeplab V3+の画像特徴量を抽出するモデルとして ResNet-101 [15] を用いる。学習画像には、自動生成したマスク画像を使用した場合と、9.000 枚のうち 2,000 枚のマスク画像を人手で精査した場合を用いており、評価画像には自動生成したマスク画像 1,000 枚を精査したものを用いた。評価には各クラスにおける Accuracy と mIoU を用いた。結果として、自動生成マスク画像のみの場合には Accuracy が 0.560、mIoU が 0.416、人手で 2000 枚精査した場合には Accuracy が 0.597、mIoU が 0.436 となった。また、図 5 は評価画像における食事領域抽出の結果の一例であり、右が入力画像、中が精査したマスク画像、右がモデルが推定したマスク画像である。たこ焼きや酢豚の画像のように単品品目やまた複数品目の領域分割は行うことができたが、定食の野菜の部分やカツオのたたきのレモンの部分を誤ったクラスに分割していることが分かる。これは、自動生成したマスク画像と精査したマスク画像における違いから起きるものと考えられる。図 3 のように画像におい自動生成したマスクには食事領域以外の領域や別の食事領域に対してもマスクが設けられている場合があることから、モデルが各クラスの領域を学習する際に誤った分類がなされたと考えられる。表 1 から自動生成マスク画像を精査することで精度が向上することが示されたことから、9,000 枚すべての画像に対して精査を行うことで、ある程度の改善が期待される。このデータセットについてはすべての学習画像に対して精査したマスク画像を付加をしたものを学習画像として公開予定である。

表 1　領域分割の精度

学習画像	Acc	mIoU
自動生成マスク画像のみ (9000 枚)	0.560	0.416
自動生成マスク画像 (7000 枚) + マスク画像 (2000 枚)	0.597	0.436

4.2　ラーメン画像の領域分割とスケッチ画像をもとにした画像生成

まずはじめに、UECFoodPix と同様に作成したマスク画像と実画像を学習画像として Deeplab V3+を用いた領域抽出と pix2pix [14] を用いた画像生成を行った。各タスクには、作成したラーメン画像データセットの中で 500 枚のラーメン画像とマスク画像のペアを用いて各タスクのモデル学習を行い、残りペア画像はテストに使用した。pix2pix は conditional GAN [16] を拡張したネットワークで 2 つの画像ドメインのペアになった画像からドメイン間の関係を学習し、ドメイン間の画像変換をする手法である。入力には、画像から取得した 15 チャンネルのバイナリマスクを用いて学習を行い、画像生成に使用したネットワークを図 6 に示す。また実験結果は図 7 の通りとなり、左は入力した画像、中は入力から領域分割されたマスク画像、右は中のマスク画像を元にして生成した画像の結果である。入力画像のレンゲやトッピングが、生成画像では同じ位置に生成で

図 5　食事領域の抽出結果　(左:入力画像, 中:正解マスク画像, 左:推定したマスク画像)

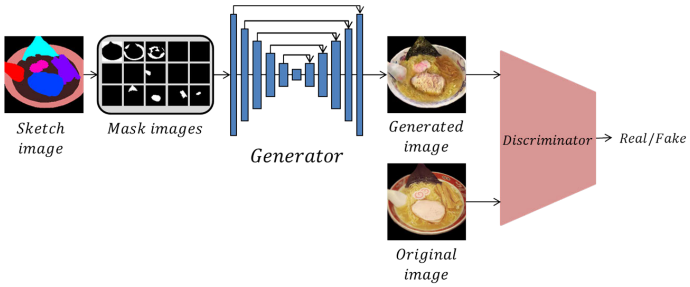

図 6　画像生成に使用した生成ネットワーク

きていることが分かる。また、領域分割されたマスク画像を修正することで任意のラーメン画像を生成することも可能あり、ラーメンのトッピングや箸の追加、除去や変更することが出来た。図 8 と図 9 にその結果を示す。しかしラーメンのどんぶりの模様やレンゲの色が入力画像と異なってことなっておりこれらのスタイルを考慮して画像を生成することは出来なかった。今回学習したモデルを用いて、ユーザーがラーメン画像を修正できる Web ベースシステムである “RamenAsYouLike” [17] の実装も行った。

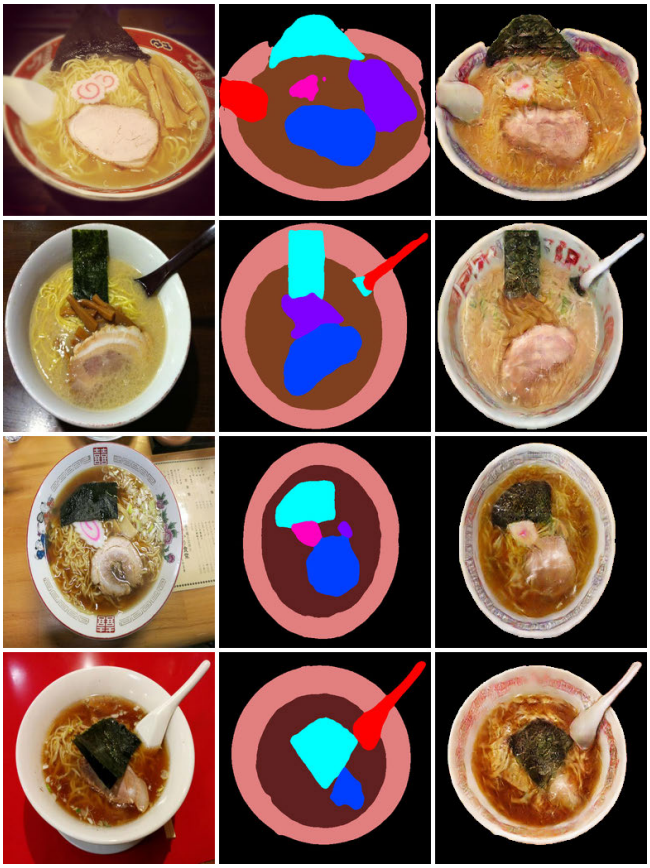

図 7　入力画像の領域分割と領域分割したマスク画像から画像生成を行った結果 (左：入力画像、中：領域分割結果、右：マスク画像から生成結果)

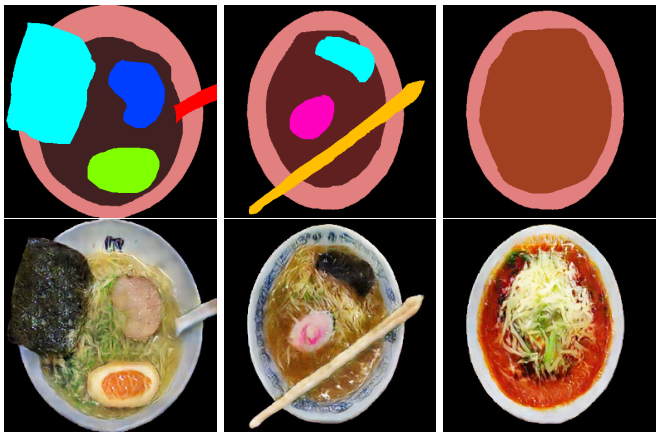

図 8　ユーザーがスケッチしたマスク画像からリアルな画像生成の結果 (上：入力したスケッチ画像、下：スケッチ画像を基にして生成した結果)

5. おわりに

本研究では、既存の UECFOOD-100 [3] に新たにインスタンスを考量したバウンディングボックスを付加し、そのバウンディングボックスに対して GrabCut を用いることで領域分割用の食事データセットを作成した。また、食事領域分割用データセットの活用例として領域分割、画像生成やユーザーによる修正からトッピングや食器の追加、変更などが簡単に可能であ

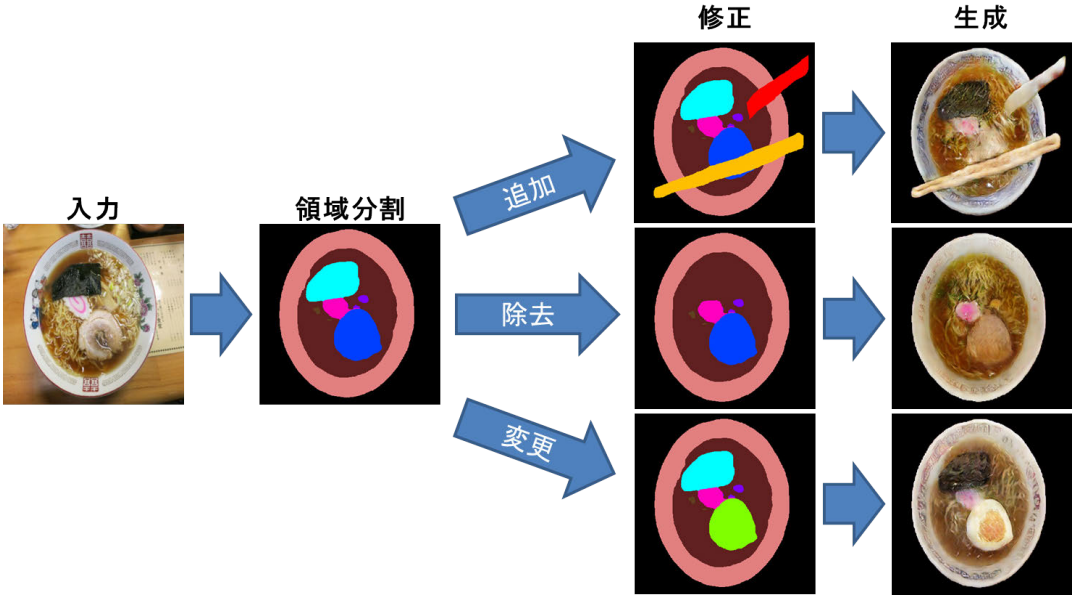

図 9　入力画像のマスク画像をユーザーが修正した後、画像生成を行った結果

ることを示した。

今後の課題として食事領域分割用のデータセットを今回作成したが、食事量やカロリー量推定におけるベンチマークとなるデータセットが未だ存在していないことから、これらの「量」データに対する対応が必要である。またラーメン画像生成においては、今回入力画像のレンゲ、模様や色といったスタイルを維持した画像生成を行うことが出来なかったことから、入力画像の各要素のスタイルを維持するようなスタイル抽出とスタイルを考慮した画像生成が課題として挙げられる。

謝辞: 本研究は, JSPS 科研費 15H05915, 17H01745, 19H04929, 17H06100 の助成を受けたものです。

文　　献

[1] M. Everingham, S. Eslami, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, “The PASCAL visual object classes challenge: A retrospective,” International Journal of Computer Vision, vol.111, no.1, pp.98–136, 2015.

[2] T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick, “Microsoft coco: Common objects in context,” Proc. of European Conference on Computer Vision, 2014.

[3] Y. Matsuda, H. Hajime, and K. Yanai, “Recognition of multiple-food images by detecting candidate regions,” Proc. of IEEE International Conference on Multimedia and Expo, pp.25–30, 2012.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” Advances in neural information processing systems, 2014.

[5] C. Rother, V. Kolmogorov, and A. Blake, “GrabCut: Interactive foreground extraction using iterated graph cuts,” ACM Trans. Graph., vol.23, no.3, pp.309–314, 2004.

[6] K. Okamoto and K. Yanai, “An automatic calorie estimation system of food images on a smartphone,” Proc. of ACM MM Workshop on Multimedia Assisted Dietary Management, 2016.

[7] A. Myers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and P.K. Murphy, “Im2Calories: towards an automated mobile vision food diary,” Proc. of IEEE International Conference on Computer Vision, pp.1233–1241, 2015.

[8] T. Park, M. Liu, and J. Zhu, “Semantic image synthesis with spatiallyadaptive normalization,” Proc. of IEEE Computer Vision and Pattern Recognition, 2019.

[9] W. Chen and J. Hays, “Towards diverse and realistic sketch to image synthesis,” Proc. of IEEE Computer Vision and Pattern Recognition, 2018.

[10] L. Bossard, M. Guillaumin, and L. Van Gool, “Food-101 – mining discriminative components with random forests,” Proc. of European Conference on Computer Vision, 2014.

[11] Y. Kawano and K. Yanai, “Automatic expansion of a food image dataset leveraging existing categories with domain adaptation,” Proc. of ECCV Workshop on Transferring and Adapting Source Knowledge in Computer Vision (TASK-CV), 2014.

[12] G. Ciocca, P. Napoletano, and R. Schettini, “Food recognition: a new dataset, experiments and results,” IEEE Journal of Biomedical and Health Informatics, vol.21, no.3, pp.588–598, 2017.

[13] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” Proc. of European Conference on Computer Vision, 2018.

[14] P. Isola, T. Zhu, and A. Efros, “Image-to-image translation with conditional adversarial networks,” Proc. of IEEE Computer Vision and Pattern Recognition, 2017.

[15] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Proc. of IEEE Computer Vision and Pattern Recognition, 2016.

[16] M. Mirza and S. Osindero, “Conditional generative adversarial nets,” arXiv preprint arXiv:1411.1784, 2014.

[17] J. Cho and K. Yanai, “Ramen as You Like: Sketch-based food image generation and editing,” Proc. of ACM International Conference Multimedia, 2019.