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.

皿領域の推論を活用した食事の弱教師あり領域分割

下田　和†　　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: †shimoda-k@mm.inf.uec.ac.jp, ††yanai@mm.inf.uec.ac.jp

あらまし　画像から食事の領域を求めるうえで皿領域と食事領域を区別することは重要である。本論文では食事画像における皿領域をアノテーションを用いずに推論する手法を提案する。具体的には、食事クラス識別器と食事/非食事画像識別器における可視化結果の差分から皿領域を求める。食事クラス識別器と食事/非食事画像識別器における識別においては、食事画像と共起の強い物体における認識の貢献度合いが変化する。皿領域は食事画像と強い共起があることが想定される。そこで、食事クラス識別器と食事/非食事画像識別器の差分を求めることで、皿領域を疑似的に推論することができるはずである。本論文では、提案手法により、アノテーションを用いずに皿領域が推論可能であることを示した。また、皿領域の推論結果を用いた弱教師あり食事領域分割手法を提案し、これにより既存手法を上回る精度を達成した。

キーワード　深層学習, 弱教師あり領域分割, 皿領域の推論

1. はじめに

深層学習を用いた領域分割手法は広く研究されており、精度は飛躍的に向上している。しかしながら、領域分割モデルを学習するには、領域レベルのアノテーションが必要であり、これを用意するには大きなコストが必要である。近年、領域分割のアノテーションコストの課題を解決するために、弱教師あり領域分割が広く研究されている。弱教師あり領域分割は画像ラベルのみを学習に用いて、テスト時においては領域分割を行うタスクである。画像ラベルは、領域ラベルと比較してアノテーションコストが安価であるため、弱教師あり領域分割が代用可能な精度を達成できれば、大幅なアノテーションコストの削減が期待できる。

食事画像において領域分割は非常に重要なタスクの一つである。物体の領域を正確に推定することができれば食事の量の推定や、カロリー量の推定などへの活用が期待される。しかしながら、食事画像においては大規模な領域分割データセットは存在しない。一方で、UECFOOD100 や Food101 など大規模なクラスラベルが付与されたデータセットは多く存在する。また、食事画像は人々の生活に密接に結びついており、大量の食事画像が SNS にアップロードされている。これらの SNS の画像の多くはテキストやタグなどの情報が付与されており弱教師情報を保持していると考えることができる。このようなクラス分類データセットや膨大な Web 画像を用いて学習し、領域分割を行うことができれば大きな利益になるはずである。本研究においては、食事画像データセットにおける弱教師あり領域分割の精度向上させる手法を提案する。

本研究においては、一般物体の弱教師あり領域分割で有効性が示されている変化領域の検出を用いた弱教師あり領域分割手法 [1] を用いる。この手法は Class Activation Map(CAM) を用いて、クラス分類結果を可視化し、CRF と変化領域の検出によりクラス分類の可視化結果を実際の物体の領域に近づけていく手法である。この手法は、一般物体の弱教師あり領域分割のベンチマークで高い精度を達成しているが、食事画像に適用する上では一般物体と食事画像の違いを考える必要がある。特に、食事画像においては常に皿が存在し、皿領域を食事の領域分割結果に含めるかという問題がある。領域分割において皿領域を領域分割結果に含めたほうがよいかどうかは、応用方法次第で変わることが想定される。例えば、カロリー推定であれば皿領域が食事領域に含まれることは望ましくなく、食事領域をインペインティングする必要があれば、皿領域も領域分割結果に含まれていたほうが望ましいはずである。もし、皿領域を推論することができればどちらの場合にも対応することができる。そこで、本研究では領域レベルのアノテーションを用いずに皿領域についても弱教師ありで推論をする。図に提案手法の意図を示した。

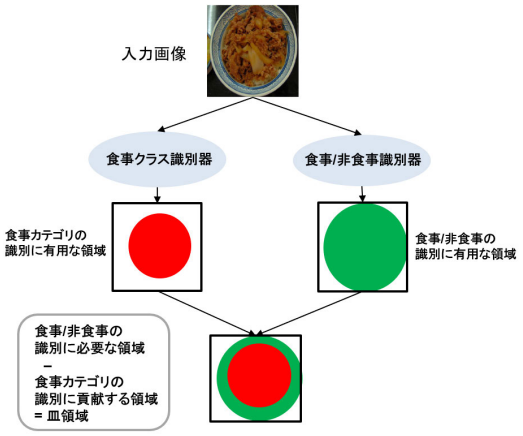

図 1　提案手法のアイディア図.

皿領域を領域レベルのアノテーションを用いずに推論するた

めに、本研究では食事クラスのクラス識別器のみでなく、食事か非食事かの識別器の学習を同時に行う。食事か非食事かの認識においては、食事と共起が強い皿領域が反応するはずである。一方で、食事カテゴリの識別結果においては、皿は常に画像中に含まれるので食事カテゴリの認識への寄与は大きくない。つまり、食事識別器と、食事カテゴリ識別器の可視化において、皿の領域の反応が異なっていることが期待できる。本研究では、この二種類の識別器における皿領域の可視化結果の違いを用いて、皿領域を領域レベルのアノテーションなしで推論する。

2. 関 連 研 究

本研究では、関連の深い研究として、食事画像認識についての研究と、弱教師あり領域分割の研究を紹介する。

2.1 食 事 認 識

食事画像の認識はカロリー推定や食生活の改善に役立つものとして、画像認識におけるアプリケーションの一つとして研究され、これまで多くの [2]〜[8] 研究が発表されてきた。しかしながら、多くの研究は画像に一枚しか食事が含まれていない場合を想定している。複数の食事を認識し、位置の推定を行うことで、より詳細なカロリー推定への応用を行うことが可能であり、これまでいくつかの研究が発表されている [5], [9]〜[11]。

松田ら [5] は複数の食事を検出するために、Deformable Part Model (DPM) [12]、円検出、JSEG による領域分割 [13] を用いた。He ら [11] は Local Variation [14] を食事の領域分割に活用しカロリー推定を行った。モバイル食事認識アプリとしては河野ら [9], [10] の研究があり、ユーザーのインタラクティブな操作から GrabCut により食事の領域分割を行った。近年の研究としては、Myers ら [15] による、“Im2Calorie” と呼ばれるモバイルフレームワークがある。Myers らはピクセルの深度情報を Deep learning により推定し、体積に基づいてカロリーを推定した。

2.2 弱教師あり領域分割

クラス分類結果を可視化する手法を用いることで弱教師ありの条件で領域分割が可能である。認識結果の可視化においては，画像におけるクラス分類に寄与した領域を推定する。クラス分類に寄与した領域と領域分割における対象領域との間には相関があり，認識結果の可視化は弱教師あり領域分割の手法として活用できる。Simonyan ら [16] は，Zeilar らと類似した手法で特定のクラスについての信号を逆伝搬させることで，CNN の認識結果に対するクラス応答を可視化させた。派生手法に Guided Backpropagation [17] がある。これの拡張手法としては下田ら [18], Jianming ら [19] の手法がある。近年は Backpropagation を用いなくても Forward の Activation から認識の可視化が可能であることが知られるようになった。Class Activation Mapping(CAM) [20] は可視化の基本的な手法として近年幅広く用いられている。

領域分割モデルを学習する際に，領域分割モデルの出力自身を教師情報にすることで精度改善が可能であることが知られている。Chen ら [21]，Pathak ら [22] は領域分割モデルの出力を領域の確信度とみなして，画像ラベル情報と前景領域と背景領域の割合などから領域分割結果で得られる結果を制限することで， 領域分割モデルを弱教師ありの条件下で学習させた。その後， Wei ら [23] は低次特徴量による物体顕著性マップを用いて学習画像の領域分割を行い，その領域分割結果を領域の教師情報として再学習を行った。Wei らの手法は単純ながら既存の弱教師あり領域分割の精度を大きく上回り，事前に領域分割マスクで学習する手法が広く行われるようになった。

一般に弱教師ありの条件で学習された DCNN から得られる物体の領域の確率分布は物体の輪郭が曖昧である。これを色特徴やエッジを用いた低次の特徴量を用いることでより正確な領域がえられることが知られている。特に CRF を使った領域の補正手法が有効であることが知られている。Chen ら [21]，Pathak ら [22] は後処理として CRF を採用し， 弱教師あり領域分割の精度が改善することを示した。Ahn ら [24] は， CRF で得た領域を教師情報として，ピクセルレベルの特徴量の類似度を学習する手法を提案した。このように CRF は弱教師あり領域分割の精度向上に大きく貢献している。しかしながら CRF は必ずしも結果の改善を保証するものではなく，むしろ結果を悪化させることがあり， これが弱教師あり領域分割の精度向上を妨げている側面がある。下田ら [1] は上記の問題点に着目し、２つの領域候補について変化を推論するモデルを学習し， ２つの候補領域からより多くの正解領域を抽出する手法を提案した。本手法においては、下田ら [1] の手法をベースとしてこれを食事画像に適用した。

3. 手　　法

本研究においては、食事画像における弱教師あり領域分割手法を提案する。既存の弱教師あり領域分割において高い精度を達成している手法を [1] 食事画像について適用する。また、これをよい精度で達成するために、皿の領域についてもピクセルのアノテーションなしで推論を行い、この皿の領域の推論結果を活用し、変化領域の推論による手法に適用し、弱教師あり領域分割を行う。セクション 3.1 において皿領域の推論方法の提案、セクション 3.2 において、今回ベースとして扱う手法 [1] について紹介、セクション 3.3 において皿領域の推論結果を活用した食事の弱教師あり領域分割の改善手法を提案する。

3.1 皿領域分割モデルの学習

本研究においては食事領域分割結果の精度を向上させるために、皿領域を推論する領域分割モデルを学習するための領域分割マスクを生成する。皿領域の推論を行うために、本研究では食事クラス識別器と非食事画像識別器の可視化結果を用いる。入力画像 x について Class Activation map(CAM) を用いた食事クラスの識別器による可視化結果を $v_L = CAM(x;\theta_L) \in \mathbb{R}^{C\times H\times W}$、非食事画像識別器による可視化結果を $v_F = CAM(x;\theta_F) \in \mathbb{R}^{2\times H\times W}$ とする。ただし、C は食事のクラス数、θ_L, θ_F は識別器についてのパラメータである。v_F は食事か食事でないかの認識を行うため、可視化結果は食事についての領域に対応しているはずである。一方で、食事のカテゴリ識別器の可視化結果 v_L は、クラス識別のために重要な領域が反応する。この二つの可視化結果は両方食事の領域が反応するはずであるが、これらの可視化の間には違いが存在する。特に、食事識別器の可視化においては認識対象に非食事画像が含まれているので、食事と共起の強い領域も同様に反応するのに対して、食事カテゴリの識別においては食事の画像に

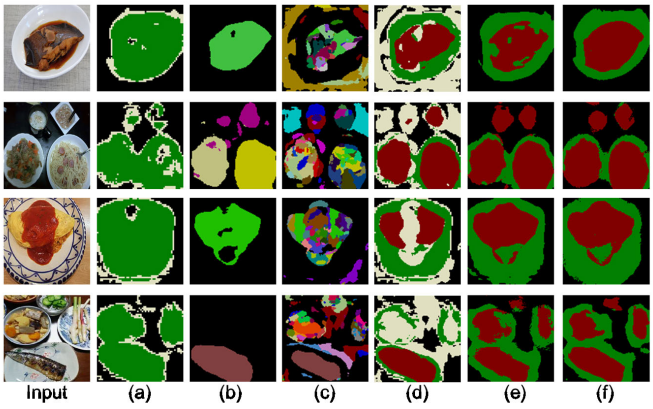

図 2　(a):食事/非食事識別器の可視化結果。(b):ラベルセット y における食事クラス識別器の可視化結果。(c):ラベルセット y における上位 k クラスについての食事クラス識別器の可視化結果。(d):生成された皿領域マスク。(e):皿領域の推論結果。(f):皿領域の推論結果+CRF。

ついて共起の強い物体は常に画像中に存在するので共起の強い領域はクラス分類に貢献しない。つまり、これらの可視化結果においては、食事について共起の強い、例えば食器や皿といった物体が反応するか否かについて大きな違いが生まれるはずである。本研究においては、非食事識別器の可視化結果と、食事カテゴリの可視化結果の違いから食事について共起の強い領域を皿領域であると仮定してこの抽出結果を皿領域とする。

本研究においては、二つの食事についての識別器の可視化結果の差分から皿領域の領域分割マスクを生成する。まず、食事か食事でないかの二値領域分割結果 $m_{F,cam}$ を v_F から得る (図 2-(a))。次に、画像に付与されたカテゴリラベル y についての領域 $m^y_{L,cam}$ をこれに相当する可視化結果 $v^y_L \in \mathbb{R}^{c^y \times H \times W}$ から得る (図 2-(b))。仮に、この $m_{F,cam}$ と $m^y_{L,cam}$ が正しく抽出することができたとすれば、この領域の差が食事に共起の強い領域となっているはずである。しかしながら、食事分類は詳細クラス分類であり、可視化結果は信頼できる精度ではない。そこで、本研究ではクラスラベルに対応する可視化結果に加えて、認識結果の上位 k クラスまでの可視化結果 $v^{r^k}_L$ から得られる領域 $m^{r^k}_{L,cam}$ を不確かな領域として設定した (図 2-(c))。実際には、$m^{r^k}_{L,cam}$ によって得られる $m^y_{L,cam}$ と重複しない領域を学習には使用しない領域として扱った。以上の処理により生成した領域分割マスクを $m_{P,cam}$(図 2-(d)) とした。この生成した背景、皿領域、食事領域の三値の領域分割モデルを $m_{P,cam}$ を用いて学習する。この学習により得られる皿領域分割結果に CRF を適用した結果を $m_{P,out}$(図 2-(f)) とし、これを食事の領域分割モデルの精度向上のために活用する。

3.2　変化領域の推論による弱教師あり領域分割 (SSDD module)

本研究においては、[1] の手法をベースとして食事の弱教師あり領域分割を行う。この手法においては、二つの候補領域マスクについて変化領域の推論を行い、二つの候補領域を統合する。図 3 に手法の概要図を示した。具体的には、セクション 3.1 における Class activation map により得られたマスク $m_{F,cam}$ と、食事領域分割結果と皿領域分割結果の統合結果 $m_{F,plt}$(3.3.1 において後述) を本手法により統合する。これらの変化領域を推論しピクセルレベルでどちらのラベルがよいかを評価し、領域

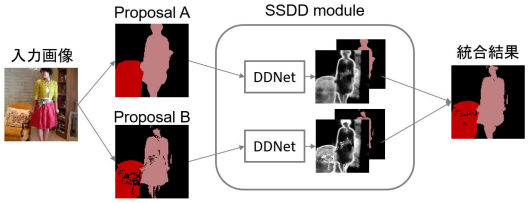

図 3　提案手法でベースとして扱った弱教師あり領域分割手法 [1] における SSDD モジュールの概要図。

分割結果の統合を行う。この二つの領域分割結果についての変化領域 $M(m_{F,cam}, m_{F,plt})$ を以下の式によって定義する。

$$M(m_{F,cam}, m_{F,plt}) = \begin{cases} 1 & \text{if } (m_{F,cam,u} = m_{F,plt,u}) \\ 0 & \text{if } (m_{F,cam,u} \neq m_{F,plt,u}) \end{cases} \quad (1)$$

ただし、$u \in \{1, 2, .., n\}$ はピクセルの位置、n はピクセル数を表す。変化領域の推定においては、バックボーンネットワークの最終ブロックと最初のブロックの出力から抽出される特徴量 (e_h、e_l) を活用する。変化領域の推論モデル (DD-net) とその推論結果 d を次式で表す。

$$d = \text{DDnet}(e_h, e_l, \hat{m}; \theta_d), d \in \mathbb{R}^{H \times W}, \quad (2)$$

ただし、$\hat{m}$ は、セグメンテーションマスク m の各ピクセルにおけるラベル情報をワンホットベクトルとしてエンコードしたものである。また、θ_d は、領域分割モデルの学習に依存しない変化領域の推論のためのパラメーターである。この変化領域の推論モデルを以下の式により学習する。

$$\mathcal{L}_{diff} = \frac{1}{|S|}\sum_{u \in S}(J(M, d_{F,cam}, u) + J(M, d^{F,plt}, u)), \quad (3)$$

$$J(M, d_{F,cam}, u) = M \log \sigma(d_{F,cam,u}) + (1 - M) \log(1 - \sigma(d_{F,cam,u})),$$
$$J(M, d_{F,plt}, u) = M \log \sigma(d_{F,plt,u}) + (1 - M) \log(1 - \sigma(d_{F,plt,u})).$$

推論時においては、この変化領域の推論モデルを、2 つの候補セグメンテーションマスク ($m_{F,cam}$、$m_{F,plt}$) に適用し、($d_{F,cam}$、$d_{F,plt}$) を得る。この変化領域の推論結果に基づいて、各ピクセルのラベルの信頼度 w を定義する。

$$w(d_{F,cam}, d_{F,plt}, u) = d_{F,cam,u} - d_{F,plt,u} + b_u, \quad (4)$$

$$b_u = \begin{cases} b_g \pm b_{cl} & either\ \ m_u\ \ belongs\ \ to\ \ C, \left(\forall c \in y, \frac{\sum |S^c_{v,O}|}{\sum |S^c_{g,O}|} < 0.5\right) \\ b_g & otherwise \end{cases} \quad (5)$$

ただし、b はハイパーパラメータである。($b_g = 0.4, b_{cl} = 1.0$). これらの値はグリッドサーチにより求めた。信頼度 w に基づいて、統合された領域分割結果 $m_{F,tch}$ を以下のようにして得る。

$$m_{F,tch,u} = \begin{cases} m_{F,cam,u} & \text{if } (w(d_{F,cam}, d_{F,plt}, u) \geqq 0) \\ m_{F,plt,u} & \text{if } (w(d_{F,cam}, d_{F,plt}, u) < 0) \end{cases} \quad (6)$$

この領域分割結果の統合結果 $m_{F,tch}$ を教師情報として、食事画像領域分割モデルの学習を行う。

3.3　皿領域の推論結果を活用した食事の弱教師あり領域分割の精度向上

皿領域は内側が食事領域で外側が非食事領域であるという性

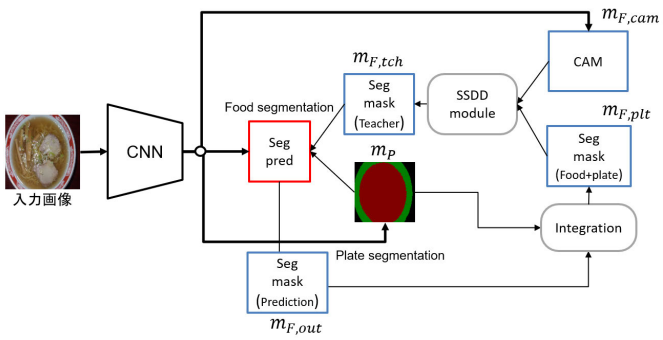

図 4　提案手法の概要図.

質がある。本研究ではこの皿領域と食事領域の関係性を活用して、食事領域の領域分割の精度改善を目指す。本研究では弱教師あり領域分割を行うために 3.2 における変化領域の推論を活用した手法を用いる。本研究ではこの手法を SSDD モジュールとして、二つの領域分割結果を入力とし、より高い精度の領域分割結果を返すモジュールとして扱う。図 4 に手提案手法の概要を示した。本手法においては [1] を活用するが、図に示すように、この手法においては二つの領域マスクを統合し、新しい領域マスクを得る。この統合された領域分割マスクを教師情報として領域分割モデルを学習する。本研究では入力として、CAM による領域分割結果 $m_{F,cam}$ と領域分割モデルの出力と皿領域の領域分割結果 $m_{F,plt}$ を使う。統合された領域分割のマスクの精度は入力に用いる二つの領域マスクの精度に依存するので、これらの入力の領域分割の改善は最終的な精度向上につながる。セクション 3.3.1 においては $m_{F,plt}$ の精度を改善する手法、セクション 3.3.2 においては $m_{F,cam}$ の精度を向上させる手法、セクション 3.3.3 では食事画像の領域分割においては背景カテゴリの出力が強くなる傾向を抑えるための手法の提案を行う。

3.3.1　皿領域の推論結果による食事領域の推論結果の制限

通常の食事領域分割では食事領域と皿領域が混合してしまうことがある。本研究においてはこれを防ぐために、皿領域の領域分割結果を食事領域分割結果に反映させる。食事の領域分割の出力について、CRF を適用した結果を $m_{F,out}$ とする。これを皿領域の領域分割マスク $m_{P,out}$ を以下の式で統合し、$m_{F,plt}$ を得る。

$$m_{F,plt} = \begin{cases} m_{F,out} & \text{if } (m_{P,out} = food\ class) \\ BG\ class & \text{if } (m_{P,out} = BG\ or\ plate\ class) \end{cases} \tag{7}$$

この食事領域の出力の制限により、SSDD モジュールの入力の精度向上が期待できる。

3.3.2　領域分割の補正結果のフィードバックによる CAM の精度改善

3.3.1 では、SSDD モジュールの出力における $m_{F,plt}$ の精度を改善したが、CAM の結果も SSDD モジュールの出力結果の精度に依存しており、高い精度の領域分割を目指すにはこれを改善する必要がある。CAM の結果を改善するために、本研究では SSDD モジュールによる領域分割の統合結果と CAM の結果が一致するように Classifier を訓練する手法を提案する。クラス分類においては特徴量全体を平均して一つのベクトルにするが、CAM においては特徴量をそのまま分類しており出力結果は各ピクセルについてのクラス分類結果であると考えることができる。そこで、本手法では領域分割結果と CAM の結果で異なる領域について、クラスラベルを生成し、このラベルを用いて識別器を学習することにより、CAM と SSDD モジュールの出力を近づける。これにより、CAM の精度が徐々に向上し、SSDD モジュールの統合結果 $m_{F,tch}$ の精度も向上することが期待できる。CAM についての領域分割マスク $m_{F,cam}$、SSDD モジュールで統合された領域分割マスク $m_{F,tch}$ についての差を $m_{F,dff}$ とする。本提案手法ではこのマスク $m_{F,dff}$ におけるクラス分類結果が $m_{F,tch}$ と一致するように学習させる。クラス分類における Global pooling 前の特徴量を $e_h(x;\theta_e)$ とする。これらについてマスク $m_{F,dff}$ の各クラス k についての領域ピクセルセットを $S^k_{F,dff}$ とする。e^k_h における $m_{F,dff}$ に対応する特徴 e^k_d は、以下の式によって計算することができる。

$$e^k_d(x;\theta_e) = -\frac{1}{|S^k_{F,dff}|}\sum_{u\in S^k_{F,dff}} e_h(x;\theta_e), \tag{8}$$

CAM の精度は入力の特徴量とクラス識別器のパラメータ θ_{cl} に依存している。CAM の結果を領域分割結果 $m_{F,tch}$ に近づけるために、本手法では $e^k_d(x;\theta_e)$ における識別器の出力が $m_{F,tch}$ のクラスに一致するように、次式により食事のクラス識別器を学習させる。

$$\mathcal{L}_{feedback} = -\frac{1}{|\hat{y}|}\sum_{k\in\hat{y}} \log(p^k_d(x;\theta_{cl})), \tag{9}$$

ここで、$\hat{y}$ は背景クラス（$y \in \hat{y}$）の注釈付きカテゴリラベルのセットであり、p^k_d は θ_{cl} により条件付けられた確率分布である。

3.3.3　皿領域の推論結果を活用した背景領域の出力の制限

食事の領域分割は詳細クラス分類であるために、一般の領域分割と比較して難易度が高く分類が難しい。領域分割が失敗する場合、食事の分類結果は背景領域として分類されることが多い傾向になる。そこで、本セクションでは皿領域の領域分割結果を用いて背景領域の推論について制限をかける。皿領域の領域分割結果の食事領域は、食事の領域分割結果の背景以外のクラスとなるはずである。そのため、皿領域の領域分割結果の食事領域に対応している部分は背景領域が出力されないはずである。そこで、本研究では食事の領域分割結果において、皿領域の食事領域に対応する領域において背景領域が出力されないように皿領域の領域分割結果を用いて食事の領域分割結果を補正した。食事領域の推論結果を $h(;\theta_s)$、皿の領域分割結果における食事領域を S_{p_f} とする。このとき、以下の式により、背景領域の出力について制限をかけた。

$$\mathcal{L}_{penalty} = -\frac{1}{|S^{food}_{P,out}|}\sum_{u\in S^{food}_{P,out}} \log(-h^{bg}_u(x;\theta_{seg})), \tag{10}$$

ただし、h^{bg}_u は $background$ クラスにおける食事領域分割結果の出力、$S^{food}_{P,out}$ は皿領域分割結果における食事クラス $food$ の領域の集合である。

3.4　最終的な領域分割モデルの学習におけるロス関数

本セクションでは最終的な弱教師あり領域分割モデルの学習

について述べる。食事領域分割モデルのパラメーター θ_{seg} は、以下の式により SSDD モジュール $m_{F,tch}$ の統合結果を用いて学習される。

$$\mathcal{L}_{main} = -\frac{1}{\sum_{k\in\hat{y}} |S^k_{F,tch}|} \sum_{k\in\hat{y}} \sum_{u\in S^k_{F,tch}} \log(h^k_u(x;\theta_{seg})). \tag{11}$$

また、これに加えてセクション 3.3.2 と 3.3.3 で提案したロス関数を以下の式により同時に学習する。

$$\mathcal{L}_{final} = \mathcal{L}_{main} + 0.1\mathcal{L}_{feedback} + 0.1\mathcal{L}_{penalty} \tag{12}$$

$\mathcal{L}_{feedback}$ と $\mathcal{L}_{bgpen}$ の係数についてはグリッドサーチにより決定した。

4. 実　　験

本実験においては UEC-FOOD100 データセット [5] を使用する。UEC-FOOD100 データセット [5] は 100 クラスの食品カテゴリで、各カテゴリには 100 枚の画像が含まれている。各食事画像はバウンディングボックスのアノテーションを保持しているが、領域分割マスクのアノテーションは付与されていない。そこで、本研究では評価のために領域分割のアノテーションを UEC-FOOD100 データセットの 10%に付与を行い、これを用いて実験の評価を行った。また、このアノテーションは評価のみに用いた。

4.1 提案手法の実装における詳細

領域分割モデルとして、[1] で使用されているアーキテクチャと同じ ResNet-38 モデルを用いた。入力画像サイズは、トレーニング画像とテスト画像どちらにおいても 448x448 とした。領域分割モデルは Pascal VOC dataset の画像および、ImageNet の画像を用いて pre-training されたモデルでパラメーターの初期化を行った。Learning rate を 1e-3 に設定し、Cosign worm up により学習中に学習率を減少させた [25]。

4.2 皿領域の推論

本研究においては食事の弱教師あり領域分割を行ううえで、皿了領域を推論することも重要であると考え、食事カテゴリ識別器と非食事識別器の可視化結果から皿領域モデルのための領域ラベルを生成した。本研究ではこの領域分割マスクを用いて皿領域推論モデルを学習した。図 5 に本手法により学習することで推論された皿領域の例を示した。単純に色特徴量に頼っていたり、円形の物体のみを推定しているのではなく、提案手法により様々なタイプの皿領域が推論されていることがわかる。これは、教師情報を用いていない結果であることを考慮すると、優れた結果であるといえる。

4.3 提案手法の効果の検証

本研究においてはセクション 3.3.3, セクション 3.3.1, セクション 3.3.2 において、それぞれ皿領域の推論結果を活用した弱教師あり領域分割結果の改善手法を提案した。表 1 にそれぞれの提案手法をとり入れた際の弱教師あり領域分割の精度を示した。全ての手法を取り入れた結果は mean IoU と Pixel acc の両方で最高精度を達成している。個々の手法は、単一では有効に働かないものもあったが、組み合わせると有効に働き精度向上に貢献していることがわかる。また、図 6 にそれぞれの結果の例を示した。それぞれの手法には特色があり、(I), (II) の手法は背景領域の推論結果に大きく寄与していることがわかる。手法 (III) においてはクラス識別器への feedback により、背景領域の推論の精度だけでなく、領域分割モデルの識別結果が改善されていることがわかる。

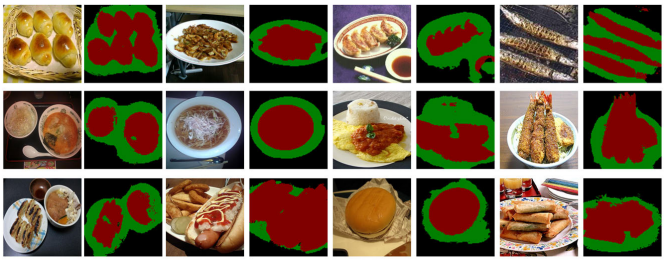

図 5　皿領域の推論結果の例。

表 1　提案手法の適用による精度の変化

Method	Sec3.3.3	Sec3.3.3	Sec3.3.3	mIoU	Pix acc
(I)	-	-	-	50.2	77.5
(II)	-	✓	✓	49.8	78.9
(III)	✓	-	✓	46.0	67.3
(IV)	✓	✓	-	51.2	78.2
(V)	✓	✓	✓	52.3	80.4

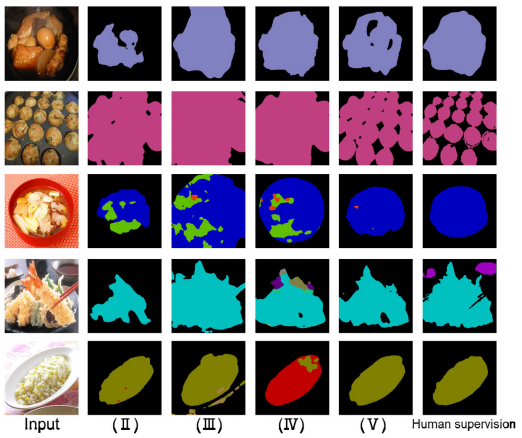

図 6　弱教師あり領域分割結果の例。 (II),(III),(IV)(V) の結果は表 1 の手法に対応している.

本研究では、セクション 3.3.3, セクション 3.3.1、セクション 3.3.2 における提案手法により、弱教師あり食事領域分割の精度向上を目指した。表 1 は、各提案手法を適用した際の弱教師あり食事領域分割の精度を示している。結果としては、全ての提案手法を組み合わせたアプローチが Mean IoU と Pixel acc の両方の評価尺度において最高精度を達成した。各提案手法は、単独で使用した場合は効果的に機能しない場合があったが、組み合わせることにより効果を発揮していることがわかる。各提案手法の効果を示す例を図 6 に示した。この例から提案手法はそれぞれ領域分割結果の外観に特定の効果を与えていることが見て取れる。特に、(I) と (II) の手法は、背景領域の推論結果に大きく寄与していることがわかる。(III) の手法におけるクラス識別器へのフィードバックにおいては、背景領域のみでなく、食事カテゴリの識別においても精度が向上していることが見て取れる。

表 2　他の弱教師あり領域分割手法との比較

Method	mIoU	Pix acc
Base method [1]	50.2	77.5
BB annotation + GrabCut [26]	51.1	81.9
Proposed	52.3	80.4

4.4　他の弱教師あり領域分割手法との比較

BB annotation + GrabCut は [26] において用いられているバウンディングボックスのアノテーションを活用した手法である。具体的には、画像に付与されているバウンディングボックスについて、グラブカットを適用し、グラブカットで得られた前景領域にクラスラベルを与えて領域分割モデルの教師情報とするアプローチである。UEC FOOD100 はバウンディングボックスを保持しているデータセットであるので、このアプローチが適用可能である。バウンディングボックスのアノテーションは領域分割と比較すると低コストであるが、クラスラベルと比較すると遥かに高コストである。バウンディングボックスを用いる手法はクラスラベルのみを用いる手法と比較して大きく有利であり、強力なベースラインであるといえる。表 2 に提案手法と既存手法の比較を示した。Base method は本提案手法がベースのフレームワークとして用いた手法 3.2 であり、表 1 における (I) の結果に相当する。驚くことに、Base method はクラスラベルのみを用いた手法であるものの、バウンディングボックスを用いた手法に近い精度を達成している。一方で、提案手法による皿領域の推論結果を活用したアプローチは、mIoU においては、Base method のみでなくバウンディングボックスを用いた手法より高い精度を達成しており、優れた結果になっていることがわかる。

5.　結　　論

本論文では、認識結果の可視化の差分から食事画像における皿領域を求める手法を提案した。具体的には、食事カテゴリ識別器と食事/非食事識別器、2 種類の食事画像についての識別器の視覚化結果の違いから、食事画像と共起の強い領域（皿領域）を抽出した。さらに、本論文ではこの皿領域から食事画像の皿領域推論モデルを学習し、これを用いて弱教師あり食事画像領域分割の精度が向上可能であることを示した。

謝辞: 本研究は, JSPS 科研費 17J10261, 15H05915, 17H01745, 19H04929, 17H06100 の助成を受けたものです。

文　　献

[1] W. Shimoda and K. Yanai, “Self-supervised difference detection for weakly supervised segmentation,” ICCV, 2019.
[2] L. Bossard, M. Guillaumin, and L.V. Gool, “Food-101 - mining discriminative components with random forests,” ECCV, 2014.
[3] H. Kagaya, K. Aizawa, and M. Ogawa, “Food detection and recognition using convolutional neural network,” ACM MM, pp.1085–1088, 2014.
[4] Y. Kawano and K. Yanai, “Foodcam: A real-time food recognition system on a smartphone,” Multimedia Tools and Applications, pp.1–25, 2014.
[5] Y. Matsuda, H. Hoashi, and K. Yanai, “Recognition of multiple-food images by detecting candidate regions,” ICME, pp.1554–1564, 2012.
[6] M. Chen, Y. Yang, C. Ho, S. Wang, S. Liu, E. Chang, C. Yeh, and M. Ouhyoung, “Automatic chinese food identification and quantity estimation,” SIGGRAPH Asia, 2012.
[7] M. Bosch, F. Zhu, N. Khanna, C.J. Boushey, and E.J. Delp, “Combining global and local features for food identification in dietary assessment,” ICIP, 2011.
[8] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, “Food recognition using statistics of pairwise local features,” CVPR, 2010.
[9] C. Morikawa, H. Sugiyama, and K. Aizawa, “Food region segmentation in meal images using touch points,” Proc. of ACM MM WS on Multimedia for Cooking and Eating Activities (CEA), pp.7–12, 2012.
[10] Y. Kawano and K. Yanai, “Real-time mobile food recognition system,” Proc. of IEEE CVPR International Workshop on Mobile Vision (IWMV), 2013.
[11] Y. He, C. Xu, N. Khanna, C.J. Boushey, and E.J. Delp, “Food image analysis: Segmentation, identification and weight estimation,” ICME, pp.1–6, 2013.
[12] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part based models,” IEEE Trans. on PAMI, vol.32, no.9, pp.1627–1645, 2010.
[13] Y. Deng and B.S. Manjunath, “Unsupervised segmentation of color-texture regions in images and video,” IEEE Trans. on PAMI, vol.23, no.8, pp.800–810, 2001.
[14] P.F. Felzenszwalb and D.P. Huttenlocher, “Image segmentation using local variation,” CVPR, pp.98–104, 1998.
[15] A. Meyers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K.P. Murphy, “Im2Calories: Towards an automated mobile vision food diary,” ICCV, 2015.
[16] K. Simonyan, A. Vedaldi, and A. Zisserman, “Deep inside convolutional networks: Visualising image classification models and saliency maps,” ICLR WS, 2014. http://arxiv.org/abs/1312.6034
[17] J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for simplicity: The all convolutional net,” ICLR WS, 2015. http://arxiv.org/abs/1412.6806
[18] W. Shimoda and K. Yanai, “Distinct class saliency maps for weakly supervised semantic segmentation,” ECCV, 2016.
[19] Z. Jianming, L. Zhe, B. Jonathan, S. Xiaouhui, and S. Sclaroff, “Top-down neural attention by excitation backprop,” ECCV, 2016.
[20] B. Zhou, A. Khosla, A. Lapedriza, and A. Oliva, A. Torralba, “Learning deep features for discriminative localization,” CVPR, 2016.
[21] G. Papandreou, L.-C. Chen, K. Murphy, and A.L. Yuille, “Weakly-and semi-supervised learning of a dcnn for semantic image segmentation,” ICCV, 2015.
[22] D. Pathak, P. Krahenbuhl, and T. Darrell, “Constrained convolutional neural networks for weakly supervised segmentation,” ICCV, 2015.
[23] Y. Wei, X. Liang, Y. Chen, X. Shen, M. Cheng, Y. Zhao, and S. Yan, “STC: A simple to complex framework for weakly-supervised semantic segmentation,” IEEE Trans. on PAMI, 2017.
[24] J. Ahn and S. Kwak, “Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation,” CVPR, 2018.
[25] L. Loshchilov and F. Hutter, “SGDR: Stochastic gradient descent with warm restarts,” ICLR, 2017.
[26] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, “Simple does it: Weakly supervised instance and semantic segmentation,” CVPR, 2017.