日本画像学会誌　第 56 巻　第 2 号：168–176（2017）

解　　説

画像を生成する深層学習ネットワーク —領域分割と画像生成・変換—

柳井　啓司*，下田　　和*

（2017.2.24　受理）

Deep Convolutional Neural Networks Which Output Images —Semantic Segmentation and Image Generation/Transformation—

Keiji YANAI* and Wataru SHIMODA*

Initially, the effectiveness of CNNs（Convolutional Neural Network）was proved for image categorization tasks for which a CNN accepts an image as an input and outputs a class probability vector as an output in general. Recently the way to use of CNNs becomes diverse, and CNNs which output images have been commonly used for semantic image segmentation, image transformation and image generation. Then, in this article, we explain CNNs for semantic segmentation in case of weakly supervision as well as full supervision, and CNNs for image generation and transformation which are typically decoder-style CNNs and encoder-decoder-style CNNs.

Keywords: Deep neural network, Convolutional neural network, Semantic segmentation, Image generation, Image transformation

画像認識において畳み込みネットワーク（Convolutional Neural Network, CNN）の有効性が最初に実証された画像カテゴリ分類問題においては，CNN はクラス確率ベクトルを出力とすることが一般的であったが，近年は入力画像から別の画像を出力したり，低次元ベクトルから画像を出力するような，画像が出力となるような CNN の利用法が盛んに研究されている．そこで本稿では，画像カテゴリ分類以外の CNN の利用例として，領域分割と，画像生成・変換について最近のトレンド及び我々の取り組みについて紹介する．領域分割については，通常のピクセルレベルでのラベルの付いた学習データを用いた完全教師あり領域分割と，画像レベルでのラベルのみから学習する弱教師あり領域分割を説明し，CNN の利用によって急激な性能向上が実現されていることを解説する．さらに，最近，急速に研究が進展しているデコーダ型ネットワークによる画像生成，エンコーダデコーダ型ネットワークによる画像変換についても解説を行う．

キーワード：深層学習，畳み込みネットワーク，領域分割，画像生成，画像変換

1. はじめに

2012 年に登場した大規模深層畳み込みネットワーク（Deep Convolutional Neural Network, CNN）はその圧倒的な認識性能の高さから，わずか数年で画像認識における中心な手法となり，2015 年には遂に 1000 種類一般物体のカテゴリ認識問題において人間を超える性能を実現するに至った．こうした状況の中，深層学習を用いた画像認識の技術的進歩は留まるところを知らず，かつてないほどのスピードで日々発展を続けている．クラス分類問題が解決された今，その派生問題である物体検出や領域分割も人間超えを目指して CNN を応用した研究が続けられている．その一方で CNN は「認識」だけではなく，「画像生成」や，「認識」と「生成」を組み合わせた「画像変換」についても大きな可能性があることが近年示されており，例えば，人の表情を自由に変化させるといったような画像の意味的な操作が CNN による画像変換で可能となっており，CNN 時代の新たな研究分野として注目を集めている．

本稿では，認識以外の CNN の応用として，領域分割と，画像生成・変換について最近のトレンド及び我々の取り組みについて紹介する．なお，領域分割については第 2 著者の下田の修士論文の一部を抜粋し，加筆修正したものである．

2. 領域分割

本節では，画素単位でラベル情報が付いた学習データから学

* 電気通信大学大学院情報理工学研究科情報学専攻
* Department of Informatics, The University of Electro-Communications, Tokyo, Japan

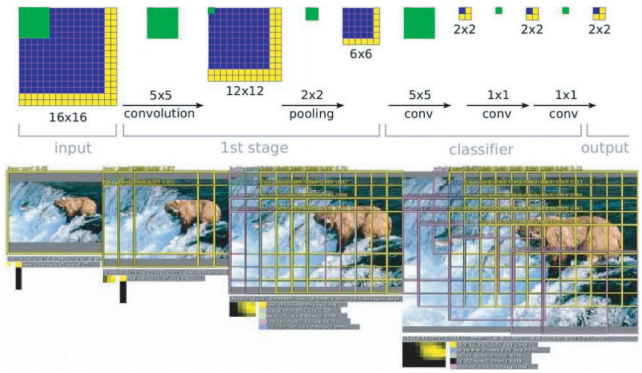

Fig. 1　Overfeat（4）より引用）

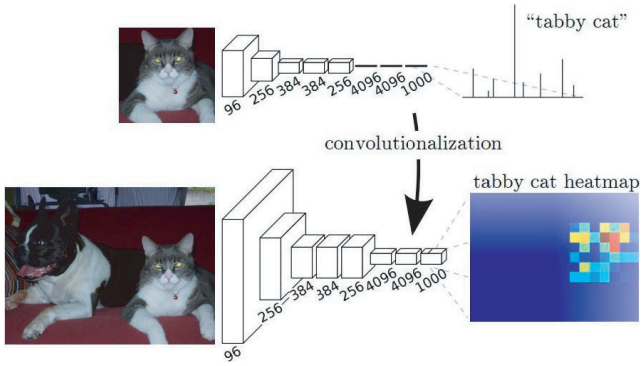

Fig. 2　FCN（5）より引用）

習する完全教師あり学習による方法と，画像に複数のクラスラベルが付いただけの学習データから学習する弱教師あり学習による方法の 2 通りの領域分割について解説する．

2.1　完全教師あり領域分割

Semantic Segmentation とは画像から物体の領域と物体のカテゴリの両方を推定するタスクである．Convolutional Neural Network（CNN）がカテゴリ分類問題において従来手法を大幅に上回る精度を達成してから，CNN は様々な分野に応用されるようになったが，特に領域分割においては CNN を活用する研究が早くからされており，CNN の幅広い応用性を示す先駆けとなった．CNN を活用した領域分割の初期の研究としては，Region Convolutional Neural Network（R-CNN）[1], Simultaneous Detection and Segmentation（SDS）[2]がある．これらの研究はともに画像中から大量の領域候補を抽出し，その領域候補についての CNN の適用結果から最良の領域候補を選出することで領域分割を行う．領域候補を用いた領域分割については古くから研究がされていたが，領域候補と CNN の画像分類精度を組み合わせた手法は既存手法の精度を大きく上回った．

RCNN と SDS はシンプルに CNN の分類精度を領域分割に生かすという形で高精度を達成した．しかし，一般に領域候補を用いた領域分割においては，領域候補が画像内の物体の領域を網羅している必要があり，領域候補の数は約 2000 と膨大になる．約 2000 の領域候補について CNN を適用する計算コストはアプリケーションへの応用を考える際に大きな障害となるため，領域候補に頼らない物体の領域分割手法が注目を集めるようになった．He ら[3]，Sermanet ら[4]は CNN を単なるクラス識別器として扱わずに，畳み込み演算の特性を活用することで効率の良い物体の位置の推定が可能であることを示した．He ら[3]は CNN における位置情報の圧縮された中間層の特徴マップを活用している．元の入力画像についての領域候補と対応する領域の中間層の特徴マップを用いて領域候補の識別が可能であることを示し，これを複数のスケールで Pooling する Spatial Pyramid Pooling（SPP）を提案した．SPP により各領域候補における計算を共有させることが可能になり，CNN を用いた物体の位置の推定は大幅に高速化された．Sermanet らが文献 4）で提案した Overfeat のネットワークは，CNN における最終層の全結合層を，その要素数とカーネル数が等しい 1×1 の畳み込み層と置き換えることによって入力画像のサイズを任意とすることができ，最終層の出力がマップ状になることを示した．Overfeat の出力は粗い格子状となっており，スライディングウィンドウを用いた領域候補の認識結果との類似点があった．一般にスライディングウィンドウを用いた認識には膨大な計算コストが必要であるが，この Overfeat のネットワーク構造は一度の CNN の認識からスライディングウィンドウとほぼ等価な結果を出力しており，CNN を用いた物体の位置の推定がより高速化可能であることを示唆していた（**Fig. 1**）．これが CNN の順伝搬による物体位置検出および領域分割の基本的なアイデアになっている．

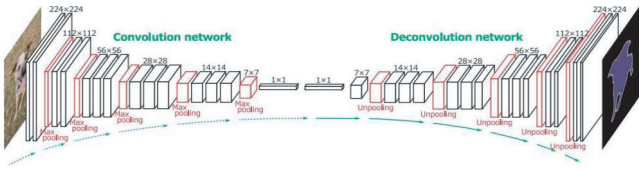

Fig. 3　DeconvolutionalNet.（6）より引用）

Long ら[5]は Overfeat[4]に着目し，全結合を取り去り出力をマップ状としたネットワークのことを Fully Convolutional Network（FCN）と呼称して紹介し，この FCN の出力が物体領域の教師情報を用いて直接最適化可能であることを示した（**Fig. 2**）．当時 CNN はクラス分類ラベルの教師情報を学習させるためのネットワークと考えられていた側面があり，領域の教師情報を用いて領域分割を行うネットワークを直接学習できるという事実は，その後の CNN を活用する研究に大きな影響を与えた．FCN のように一つのネットワークで一つのタスクを簡潔させるネットワークは，これまでの RCNN や SDS と比較して，end-to-end network と呼ばれ，領域分割にとどまらず他の CNN の応用分野でも end-to-end network が広く利用される先駆けとなった．また，Long らは FCN の出力を領域の教師情報のサイズに近づけるため，逆畳み込み演算を行うことで粗い FCN の出力マップを拡大する Deconvolution を同論文内[5]で提案したが，これは後に Deconvolutional Networks[6]（**Fig. 3**）や SegNet[7]，U-Net[8]などで広く応用されている．さらに，Deconvolution は，次章で述べる画像生成，画像変換の基本技術ともなっている．

Long ら[5]は FCN により既存の手法と比較して高精度な領域

分割を達成したが，FCN の出力は物体の中心が強く反応し，領域の境界が曖昧になる傾向があった．Chen ら[9]は Deep-lab ネットワークを提案し，境界の導出において有効であると知られている既存手法の Conditional Random Field（CRF）を FCN と組み合わせた．Chen らは FCN の出力を確率分布とみなし，これを単項ポテンシャル，低次特徴量を平滑化項ポテンシャルとすることで CRF を適用し，より高精度な領域分割を達成した．また，Chen ら[9]は FCN の出力と同じ大きさに領域の教師情報を縮小することで，逆畳み込み演算を必要とせずに FCN の学習が可能であることを示した．一方で，Zheng ら[10]は Recurrent Neural Network を活用し，CRF を FCN の学習に組み込むことで，CRF のパラメーターを同時に学習させた．

Yu ら[11]は Deep-lab[9]において用いられた Atrous convolution に着目した．Yu ら[11]は Atrous convolution を Dilation と呼称し，領域分割における Dilation の重要性について言及し，その有効性を検証した．通常の畳み込み演算では近傍のピクセルを畳み込むが，Dilation においては畳み込みのフィルターをスパースにすることで離れたピクセルを畳み込む．FCN の出力の一つ一つのピクセルは局所特徴量のように近傍の情報の影響を強く受けており，意味的な表現力が不十分で領域分割精度が低下する傾向があったが，Dilation によりこの問題が緩和された．Dilation は広域の情報を一つのピクセルで表現するうえで重要な役割を果たしていることが推測される．

最新の研究としては Chen らによる Deep-lab V2[12]，Zhao らによる PSP-Net[13]がある．Chen ら[12]は複数のスケールの入力画像から得られる出力の統合，複数のスケールの Dilation から得られる出力の統合により，さらに精度が改善可能であることを示した．Zhao ら[13]は Spatial Pyramid Pooling[3]を Dilation と組み合わせることが可能であることを示し，FCN の出力のみで CRF などの低次特徴量を用いた後処理による領域の境界の探索などが不要なほど高精度な領域分割を達成した．これらの最新の研究動向から，いかに局所的な情報と俯瞰的な情報の両方を一つのピクセルで表現するかという問題が，現在の CNN を用いた領域分割の主題となっていることが見受けられる．局所的な特徴と意味的な認識結果を統合するという方針は，低次特徴量による領域分割において既に取り入れられており，特にボトムアップとトップダウンの統合という形で広く研究がなされている[14,15]．低次特徴量の領域分割において有用であった概念が，CNN における領域分割においても有用であったといえる．領域分割は，CNN 登場前は精度向上に限界が見られていたが，CNN を活用することで，標準データセットである Pascal VOC 2012 segmentation dataset において，この 2 年間で 40% 程度から 80% 以上にまで精度が急速に向上している．まだ性能向上は止まっておらず，さらなる精度の改善の余地は十分にあるであろう．なお，CNN による手法では，ほぼすべての手法で ImageNet 1000 種類の 100 万枚の画像で事前学習した VGG-16 や ResNet などの大規模ネットワークをベースとなるネットワークとして利用しており，認識用の大規模学習画像セットの転移学習による効果が領域分割に生かさせていると考えることもできる．

2.2　弱教師あり領域分割

深層学習においては，高精度な認識を実現するために，膨大な教師付き画像が必要であり，認識対象の拡張などの面で大きな障害となっている．また，物体の位置推定を初めとした応用分野では，より高度な教師情報が必要となるのが一般的であり，この問題が顕著になる．特に，領域分割においては，学習画像における物体のカテゴリごとにピクセル単位の領域の教師情報が必要である．膨大な画像に対して，このような高度な教師情報の付与は大きなコスト，時間を要する．一般に，高度な教師情報を必要とする手法を完全教師あり学習，画像における物体のカテゴリ情報のみから学習する手法を弱教師あり学習と呼ぶ．弱教師あり学習の場合，必要なのはカテゴリ情報のみであるので，Web 画像を学習画像として扱うことが可能であり，収集が容易である．弱教師あり学習による物体の位置推定が可能となれば，大幅なコスト削減が可能である．

弱教師あり領域分割を行う方法として CNN の認識結果の可視化手法がある．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．Zeiler ら[16]は畳み込み演算と同じ学習パラメータによる逆畳み込み演算と逆 Pooling により，出力を入力空間に戻した際に，物体の位置に対応するピクセルが強く応答することを示した．この応答は，オクルージョンを用いて入力画像の一部を隠した場合の認識結果の変化領域と対応しており，逆畳み込み演算と逆 Pooling が CNN の認識結果の可視化において有効であることが分かった．この逆畳み込み演算と逆 Pooling は，認識における順伝搬（forward propagation）に対して，逆伝搬（backward propagation）と呼ばれるようになり，後の CNN の認識結果の可視化手法の基礎となった．Simonyan ら[17]は，Zeiler 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．この可視化結果は物体の顕著性マップと類似した結果となっており，Simonyan ら[17]は可視化結果を GrabCut[18]の seed とすることで高精度な弱教師あり領域分割を達成した．派生手

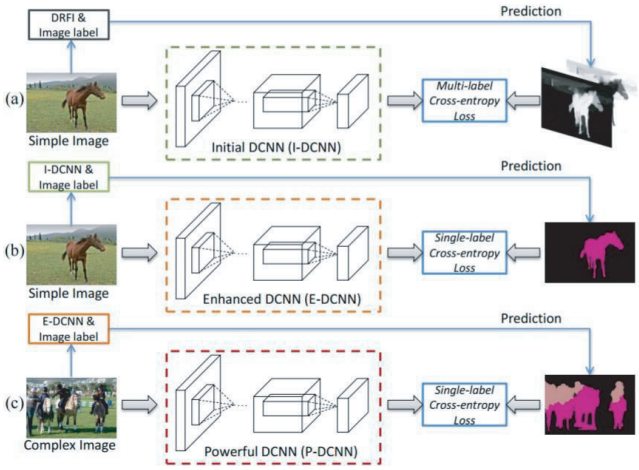

Fig. 4　Simple-To-Complex framework.（29）より引用）

法に Guided Back-propagation[19]，Layer-Wise Relevance Propagation[20]などがある．

Simonyan ら[17]による Backward の可視化手法においては，シングルクラスの場合に有効であり，マルチクラスの場合に極端に精度が下がるという問題があった．それに対して，Shimoda ら[21]は各クラスの逆伝搬値について差分をとることでマルチクラスにも応用可能であることを示した．類似研究としては，Jianming ら[22]による研究がある．

一方で，Oquab ら[23]，Pathak ら[24]は FCN の最終層に Global Pooling（GP）を用いることで，出力マップをクラス分類の CNN と同じ次元に変換し，画像ラベルのみを用いて FCN を学習させた．GP により学習させた FCN は精度が落ちるものの，大まかな物体の位置を推定することが可能であり，認識結果の可視化とは異なる形で弱教師あり領域分割を実現した．その後，Pinheiro ら[25]や Zhou ら[26]により，Global Pooling の派生手法についても研究がなされている．また，Chen ら[27]，Pathak ら[28]は，Global Pooling を用いずに，弱教師ありで FCN の出力を直接学習させた．FCN の出力の各ピクセルが画像ラベルに含まれていないクラスを出力している場合には修正を加えるような形で，FCN の出力と画像ラベルから動的に領域の教師情報を生成し，これを用いて FCN の学習を行った．

その後，Wei ら[29]が Simple to Complex（STC）フレームワークを発表した（**Fig. 4**）．STC では，第 1 ステップとして学習データのうちクラスラベルが 1 つしかないような簡単な学習画像に対して，低次特徴量による物体顕著性マップを用いて領域分割を行い，その結果を第 2 ステップの完全教師ありの領域分割手法による学習の学習データとする．さらに，複数ラベルが含まれる比較的難しい学習画像に対して第 2 ステップでの学習モデルを適用し，さらに多くの領域分割結果を得る．これによってほぼすべての学習画像に，完全教師あり手法の学習に必要な領域マスクデータが推定される．最終の第 3 ステップでは，領域マスクデータを利用して完全教師あり手法を再度学習して，最終的な領域分割モデルを得る．この手法は，弱教師あり専用の手法を利用することなく，既存手法の顕著性マップと完全教師あり手法の組合せのみで構成されていて，単純にも関わらず既存の弱教師あり領域分割の精度を大きく上回り，弱教師あり領域分割の研究に大きな変化をもたらした．特に，その後の弱教師あり領域分割において，最初の種となる領域マスクデータを作成するための教師なしの領域分割手法（カテゴリは既知である画像の領域分割手法），ノイズを含む領域の教師情報について頑健な学習手法の重要性が増した．Kolesnikov ら[30]は，Global Pooling[26]による学習画像の大まかな位置推定結果について再学習を行った．また，Kolesnikov らは多くの領域の評価をスキップすることによるノイズに堅牢な学習，KL-Divergence による CRF による平滑化結果との近似などを用いて高精度な弱教師あり領域分割を達成した．Saleh ら[31]は

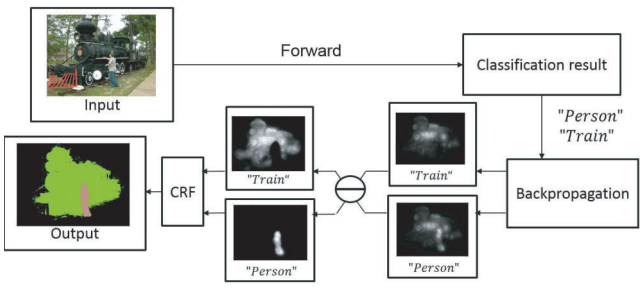

Fig. 5　Distinct Class Saliency Maps（DCSM）[21].

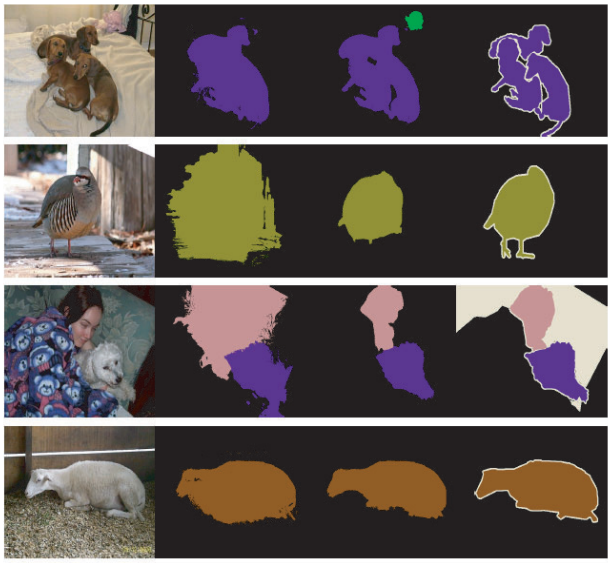

Fig. 6　結果例．DCSM＋CRF は[21]の結果．

表 1　PASCAL VOC 2012 *test set* の領域分割タスクにおける結果比較．（†RCNN は VOC2011 の結果，‡DeepLabV2 は MSCOCO の学習データも利用，のため条件が異なる．）

Methods	mean IoU
Fully Supervised :	
O2P［34］	47.6
（R-CNN）［1］†	47.9
SDS［2］	51.6
FCN-8s［5］	62.2
DeepLab［9］	71.6
CRF as RNN［10］	72.0
（DeepLabV2）［12］‡	79.7
PSPNet［13］	82.6
Weakly Supervised :	
MIL-FCN［24］	24.9
EM-Adapt［27］	39.6
CCNN［28］	35.5
MIL-ILP-seg［25］	40.6
DCSM w/o CRF［35］	41.0
DCSM w/ CRF［35］	45.1
F/B prior［31］	48.0
STC［29］	51.2
SEC［30］	51.7
Our results（unpublished）	51.2
Ours＋Data Augmentation	52.8

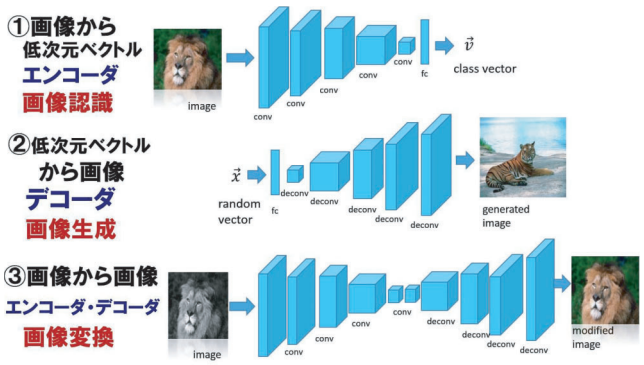

Fig. 7　3 種類の CNN.

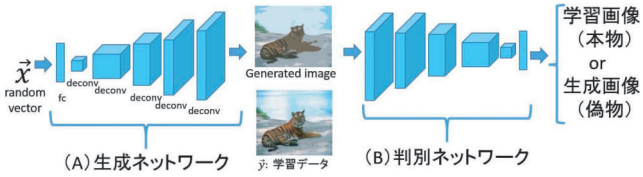

Fig. 8　Deep Convolutional Generative Adversarial Network（DCGAN）.

特徴マップについて CRF を適用した結果から再学習を行い高精度を達成した．Tokmakov ら[32]は動画から得られるモーションの Segmentation と Gaussian Mixture Model における動画のフレームの Segmentation 結果を用いて領域分割結果の再学習を行った．

2.3　弱教師あり領域分割の研究例

ここでは我々の研究グループで提案した Distinct Class Saliency Maps（DCSM）[21]について紹介する．

CNN の特徴マップを用いた弱教師あり領域分割手法は，複数候補領域の認識による領域分割と比較して，高速で領域分割を行うことができた．しかし，弱教師ありの学習による特徴マップからは粗い位置推定結果しか得ることができず，back-propagation を用いた位置推定結果の補正が必要であった．

そこで提案手法の DSCM では，17）の手法を改良することで，back-propagation に基づいた手法のみで領域分割を完結させる．17）の手法は，各カテゴリについてのサリエンシーマップを得ることができるが，各カテゴリのサリエンシーマップには曖昧な違いしかない．そこで，我々は以下の 3 点の改良により，各カテゴリごとに鮮明なサリエンシーマップを生成し，back-propagation に基づいた手法のみで領域分割を完結させ，より高精度な弱教師あり領域分割を実現した．

・伝搬値における up-sampling
・各カテゴリの伝搬値の差分
・複数の異なるサイズの結果の統合

特に，各カテゴリの伝搬値の差分をとることで，各カテゴリの領域の推定結果を明確化し，特徴マップの利用を必要とせずに，詳細な物体の位置推定を行うことができた．最後に後処理として全結合型 CRF（Dense CRF）[33]を適用している．**Fig. 5** に処理の流れを示す．

さらに，STC[29]の考え方を採り入れて，拡張を行った．STC では初期学習用の領域マスク推定に顕著性マップ手法を用いているが，DSCM を代わりに利用し，出力結果の領域分割結果の確率値を利用して学習画像の水増しを行い，STC を上回る精度を実現した．**Fig. 6** に領域分割結果の例を示す．

最後に，領域分割の標準ベンチマークである PASCAL VOC 2012 データセットに対する，完全教師あり及び弱教師ありの主な手法による認識結果を **Table 1** にまとめて示す．2014 年の完全教師あり手法 SDS[2]の精度 51.6% を最新の弱教師あり手法の精度 52.8% が上回っている点は特筆すべき点である．

3.　画像生成・変換

深層畳み込みネット（Convolutional Neural Network, CNN）は画像認識に導入された当初は画像を認識をするのが目的であった．つまり，画像を入力とし，出力はクラス確率などの低次元ベクトルであるエンコーダ型のネットワーク（**Fig. 7**（1））が一般的であった．それがここ最近 2 年でエンコーダネットワークに加えて，その逆の低次元ベクトルからの画像生成を行うデコードネットワーク（Fig. 7（2）），さらにはその両者の合わせた形になっている，入力画像を一度エンコードしてからデコードして異なる画像を生成するエンコーダ・デコーダネットワーク（Fig. 7（3））が様々なタスクに応用されており，CNN の応用範囲は認識以外にも大きく広がっている．

3.1　画像生成ネットワーク

低次元ベクトルから画像を生成するのがデコーダ型ネットワークである．エンコーダが識別モデルだとすれば，デコーダは生成モデルと言うことになる．画像生成ネットワークでは，学習画像に対してある一定の大きさの次元（例えば 128 次元）のランダムベクトルを割り当て，それを入力として学習するのが一般的で，学習時に与えていないベクトル値の入力であっても何らかの意味のある画像を出力できるようなるように学習することが目的となる．画像認識で利用されるエンコーダ型ネットワークでは入力データが持ち合わせている情報を低次元化するのが目的であるのでネットワークの学習は比較的容易であるが，逆のデコーダ型ネットワークでは入力ベクトルに対して出力の情報量が極めて大きいため，学習が困難であるという問題点があった．例えば，エンコーダネットワークの回帰問題を扱う場合に一般的な 2 乗誤差による損失関数では，重要な物体の輪郭や細部の構造が学習されず，全体的にぼやけた画像しか生成できないネットワークが学習されることが知られていた．

この問題は，Generative Adversarial Network（GAN）（生成型敵対ネットワーク）[36]によって解決された．GAN では，生成ネットワーク以外に，学習画像と生成された画像を見分けるエンコーダ型の認識 CNN を判定ネットワークとして用意する．判別ネットワークは，学習画像を本物，生成された画像を偽物として 2 クラス分類を行う CNN として学習する．一方，生成ネットワークは判別ネットワークにおいて生成された画像が本物と判定されるように学習を行う．以上のように，互いに反対の目的を持った敵対する 2 つのネットワークを競わせて繰り返し学習することによって，物体の細部までがうまく学習され，生成される画像が本物に徐々に近付いていくことになる．

当初は小さい画像しか生成できなかったが，Deep Convolutional GAN（DC-GAN）[37]の登場によって，64×64 程度の大きさの高品質な画像が生成可能となった．**Fig. 8** に一般的な DCGAN のネットワークを示す．DCGAN では，デコーダ型ネットワークを Deconvolution と Batch Normalization[38]で構成することによって高精度画像の生成を実現している．さらに，入力のランダムベクトルの演算ができることでも注目された．例えば，眼鏡の男性の写真に対応するベクトルから男性の写真のベクトルを引いて，女性の写真のベクトルを足すと，眼鏡の女性の画像が生成される，という具合いである．

DCGAN では，学習画像に学習時に生成したランダムベクトルを割り当てていたが，Conditional GAN（cGAN）[39]では属性をバイナリ値として条件ベクトルに割り当て，それをランダムベクトルと結合して GAN で生成ネットワークを学習することによって，例えば，髭の生えた顔を生成することが簡単にできるようになった．Reed ら[40]は文章を Recurrent Neural Network でベクトル化し cGAN の条件ベクトルとすることによって，入力文章に応じた画像生成を実現した．また，InfoGAN[41]では，conditional GAN で必要だった明示的な属性情報を必要とせず，条件ベクトルの代りに潜在変数ベクトルを導入し，生成画像の分布と潜在変数の相互情報量最大化によって学習しながら属性の自動推定が行われる．

Invertible Conditional GAN（IcGAN）[42]では，cGAN と入力画像からランダムベクトルと属性ベクトルを求める逆変換ネットワークを組み合わせることで，入力画像の属性を変更した画像の生成ができることを示した．入力画像からランダムベクトルと属性ベクトルを求め，属性ベクトルだけを修正して，同じランダムベクトルで画像を生成すると，例えば，髭の生えた人の顔を髭のない顔に修正することができる．Interactive GAN[43]はドローエディターに GAN による画像生成機能を追加した．描画中の画像からランダムベクトルを求め，それを用いて画像を生成すると，画像の輪郭を描いただけで，画像が表示されるような新しい画像生成のインタフェースが実現できる．

なお，画像の表情を変えることは IcGAN でできるが，同様の変換が VGG-16 の特徴マップに特定のベクトルを加算して，その特徴マップを逆変換して画像に戻すことによってもできることが，最近示されている[44]．普通の表情の人を笑った顔にするような変換が可能であることが特徴マップの操作で可能である．IcGAN ではランダムベクトルをエンコーダ CNN で求めていたが，VGG などの一般のエンコーダ型 CNN で求める特徴マップも本質的には GAN のランダムベクトルと同じであり，特徴マップ空間での演算による画像操作が可能で，しかも画像の意味的な操作が可能であるということが，この結果から示唆されている．また，これは，スタイル画像のスタイルに入力画像を変化させるスタイル変換とも関係があって，入力画像の VGG-16 の特徴マップの分布を，スタイル画像の特徴マップの分布（チャネル平均と分散）に合わせるように変換して，画像に逆変換で戻すと，スタイル変換された画像になること[45,46]と類似性がある．

DCGAN で生成画像の質が向上したが，解像度は 64×64 であった．最近登場した Stack GAN[47]では，超解像処理を行うネットワークと組み合わせることによって 256×256 の画像生成を可能とした．なお，GAN については，提案者である Goodfellow によるチュートリアル資料[48]に詳しく書かれている．

他にも，短時間の動画生成[49,50]や 3D voxel データの生成[51]など 2 次元画像に留まらず様々な応用が提案されており，今後，さらなる発展が期待される．

3.2 画像変換ネットワーク

画像変換ネットワークは，エンコーダ・デコーダ型ネットワーク，もしくは Conv-Deconv ネットと一般に呼ばれ，画像を一度エンコードして低次元の中間表現に変換して，それをデコードすることによって，入力画像に何らかの処理を施した画像を生成することができる．元々は前の章で触れた領域分割の手法として提案された SegNet[7]が最初の画像変換のためのエンコーダデコーダ型ネットワークであるとされている．SegNet ではエンコーダは学習せずに pre-train 済の VGG-16 ネットワークが利用された．DeconvNet[6]でも，ほぼ同様の構造が提案されている．一方，U-Net[8]では，エンコードによる低次元で細部の情報が失われて，生成される出力画像の解像度が劣化するのを防ぐために，エンコーダの各中間特徴マップと対応するサイズのデコーダの特徴マップを直接結合する skip connection をエンコーダデコーダネットに導入した．実際これによって細部の構造が保持されたまま，最終出力にそれを反映させることが可能となった．

元々は領域分割用に提案された画像変換用のエンコーダデコーダ型ネットワークであるが，その後，入力画像と出力画像のペアが大量にあれば，あらゆる画像入力画像出力タスクに適用できることが分かってきた．白黒画像のカラー化[52]，超解像度処理[53]，単一画像からの奥行き推定[54]，2 枚の画像からのオプティカルフローの推定 FlowNet[55]，高速スタイル変換ネットワーク（詳細は後述）[56]，HOG，BOF，AlexNet の中間信号などの特徴表現ベクトルからの画像復元[57]，など様々なタスクにほとんど標準的なエンコーダデコーダネットワークをそのまま利用するだけで応用可能であり，「早いもの勝ち」の状況が起っていた．

さらに，こうしたエンコーダデコーダネットの汎用性をさらに一歩進めたのが UberNet[58]である．UberNet は Universal Network の略で，エンコーダを共通なものとして，デコーダのみをタスク毎に用意することによって，物体抽出，領域分割，物体境界抽出，人間のパーツ推定，面の法線方向推定，saliency map，エッジ抽出の 7 つのタスクを同時に行うネットワークを提案した．

このエンコーダデコーダネットワークは，入力と出力のペアさえあれば様々な変換を学習可能であることより，特に従来の computer vision が対象としてきた一般には逆問題であるとされる 2 次元画像からの 3 次元物理世界の推定に利用することが可能である．大量の入出力のペアは順問題のモデルで生成可能であるので，順問題で生成した大量データをネットワークを学習すれば逆問題のモデルをエンコーダデコーダネットで構築できることになる．そのため，認識分野ではなく，3 次元復元や光学推定分野にも今後，ますます CNN が普及していくことが

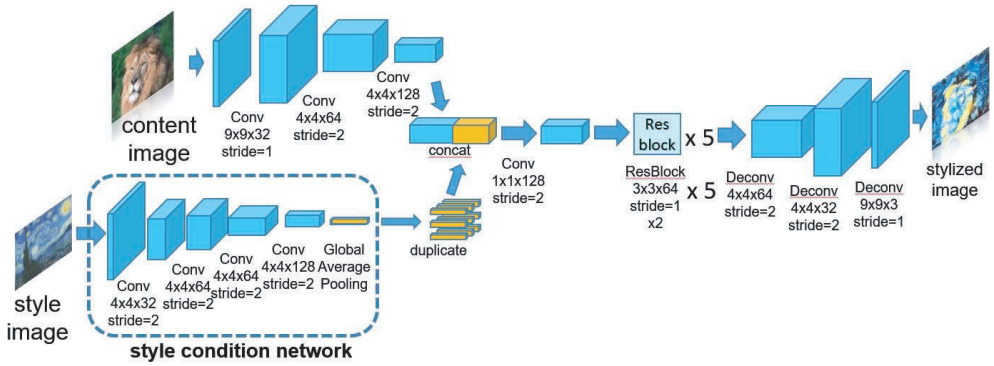

Fig. 9　高速任意スタイル変換ネットワーク．

Fig. 10　任意スタイル変換の結果例．

予想される．

万能と思えるエンコーダデコーダネットであるが，従来は，領域分割結果から本画像を生成するなど，著しく情報量が縮退している場合は学習が困難であった．ところが 2016 年末に，それも画像生成の GAN の枠組みで提案された Adversarial Loss を利用することで解決できることが示された[59]．GAN はエンコーダネットワークにおいて低次元ベクトルから画像生成するための手法であるので，この手法で学習すると画像生成的な要素が強くネットワークに導入され，情報量が著しく減っていようともまったく関係なく，地図から航空写真，エッジ画像から写真画像，領域分割結果のマスク画像から元画像など，従来は難しかった逆変換も自由自在に学習可能となることが示された．またこの方法は画像の一部を消去して描画するインペインティングにも利用可能である．

3.3　研究例

ここで我々のグループで，画像変換ネットを使った研究例を示す．これは Neural Style Transfer[60] を高速に実行するネットワークで，**Fig. 9** に示すように変換対象画像入力とスタイル画像入力の 2 つの入力をもち，スタイル画像のスタイルを変換対象画像に転送する end-to-end な変換ネットワークである[61]．入力が 2 つあるので，エンコーダ部分が 2 つあることが特徴となっている．このネットワークは，Johnson らによって提案されたエンコーダデコーダ型の高速スタイル変換ネットワーク[56] にスタイルエンコーダを追加して実現したものであり，学習も画像変換も end-to-end で行うことができる．

CNN はこのように複雑な構造となっていても，一般的な誤差逆伝搬法とミニバッチを用いた確率的勾配降下法というシンプルな手法で学習が可能であり，これはある意味，驚くべきことである．結果例を **Fig. 10** に示す．上部に横に並んでいるのがスタイル画像で，左の 1 列が入力画像，スタイル変換された画像が 2 列 2 行以降に示されている．元々の Neural Style Transfer[60] ではバックプロパゲーションで画像を徐々に変化させるため 1 分程度の処理時間が掛ったが，学習したネットワークによって，2 枚の画像を入力するだけで瞬時に画像スタイル変換が可能となった．また，実行時間の短縮だけでなく，複数のスタイルを混ぜ合わせることも容易にできるという特徴もある．

4.　おわりに

本稿では，画像を出力する CNN の応用例として，通常のピクセルレベルでのラベルの付いた学習データを用いた完全教師あり領域分割と，画像レベルでのラベルのみから学習する弱教師あり領域分割を説明した．さらに，最近，急速に研究が進展しているデコーダ型ネットワークによる画像生成，エンコーダデコーダ型ネットワークによる画像変換に関しても解説した．

今後の展望としては，領域分割に関しては，弱教師あり手法が完全教師あり手法に近いレベルまで発展することが予想される．完全に一致するところまでは教師データの情報量が大きく違うため困難が予想されるが，一部の画像のみに全ピクセルラベルを付けるなどの部分的な教師情報の追加をおこなう準教師あり学習（semi-supervised learning）を導入することによって，ほぼ同程度の精度が実現できる可能性が高い．

本解説の最後に述べたように，今後の画像生成は何と言っても動画生成である．短時間の動画生成[49,50] は研究されているが，結果の自然さや解像度，生成されるビデオの長さなどいずれもまだまだ不十分であり，今後のさらなる発展が期待される．また言語処理と融合した文章からの動画生成は，将来の動画コンテンツの自動作成にもつながるもので，今後ますます研究が盛んになると思われる．

画像変換では，現在は特定の決まった変換を行う研究が多いが，最後に示した研究例のように入力に応じて変換を変化させることのできる conditional な画像変換が今後，多く研究されるようになるであろう．言語など画像以外の入力も組み合わせたマルチモーダルなエンコーダデコーダネットワークなど，ネットワークは複雑化し，複雑なタスクへの応用が進むであろう．

参 考 文 献

1) R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, pp. 580–587, 2014.
2) B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Simultaneous detection and segmentation. In *ECCV*, 2014.
3) K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, pp. 346–361, 2014.
4) P. Sermanet, D. Eigen, X. Zhang, M. l Mathieu, R. Fergus, and Y. LeCun. Overfeat : Integrated recognition, localization and detection using convolutional networks. In *ICLR*, 2014.
5) J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.
6) H. Noh, S. Hong, and B. Han. Learning deconvolution network for semantic seg-mentation. In *ICCV*, 2015.
7) A. Kendall, V. Badrinarayanan, and R. Cipolla. Segnet : A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. In *CVPR*, 2015.
8) O. Ronneberger, P. Fischer, and T. Brox. U-net : Convolutional networks for biomed-ical image segmentation. In *Medical Image Computing and Computer-Assisted In-tervention* (*MICCAI*), 2015.
9) L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and Yuille A.L. Semantic image segmentation with deep convolutional nets and fully connected CRFs. In *ICLR*, 2015.
10) S. Zheng, S. Jayasumana, B.R. Paredes, V. Vineet, and Z. Su. Conditional random fields as recurrent neural networks. In *CVPR*, 2015.
11) F. Yu and V. Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2016.
12) L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille. DeepLab : Seman-tic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *arXiv : 1606.00915*, 2016.
13) H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. *arXiv : 1612.01105*, 2016.
14) E. Borenstein, E. Sharon, and S. Ullman. Combining top-down and bottom-up segmentation. In *CVPR*, 2004.
15) E. Borenstein, E. Sharon, and S. Ullman. Combining top-down and bottom-up segmentation. In *CVPR*, 2004.
16) M. Zeiler and R. Fergus. Adaptive deconvolutional networks for mid and high level feature learning. In *ICCV*, 2011.
17) K. Simonyan, A. Vedaldi, and A. Zisserman. Deep inside convolutional networks : Visualising image classification models and saliency maps. In *ICLR WS*, 2014.
18) C. Rother, V. Kolmogorov, and A. Blake. Grabcut : Interactive foreground ex-traction using iterated graph cuts. *ACM Trans. on Graphics*, Vol. 23, No. 3, pp. 309–314, 2004.
19) J.T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for sim-plicity : The all convolutional net. In *ICLR WS*, 2015.
20) S. Bach, A. Binder, G. Montavon, F. Klauschen, K-R. Muller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. In *Plos One 10*, 2015.
21) W. Shimoda and K. Yanai. Distinct class-specific saliency maps for weakly super-vised semantic segmentation. In *ECCV*, 2016.
22) Z. Jianming, L. Zhe, B. Jonathan, S. Xiaouhui, and S. Sclaroff. Top-down neural attention by excitation backprop. In *ECCV*, 2016.
23) M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014.
24) D. Pathak, E. Shelhamer, J. Long, and T. Darrell. Fully convolutional multi-class multiple instance learning. In *ICLR*, 2015.
25) P. Pedro and C. Ronan. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, 2015.
26) B. Zhou, A. Khosla, A. Lapedriza, and A. Oliva, A. Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.
27) G. Papandreou, L.-C. Chen, K. Murphy, and A.L. Yuille. Weakly-and semi-supervised learning of a dcnn for semantic image segmentation. In *ICCV*, 2015.
28) D. Pathak, P. Krahenbuhl, and T. Darrell. Constrained convolutional neural net-works for weakly supervised segmentation. In *ICCV*, 2015.
29) 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. In *ECCV*, 2016.
30) A. Kolesnikov and C.H. Lampert. Seed, expand and constrain : Three principles for weakly-supervised image segmentation. In *ECCV*, 2016.
31) F. Saleh, M. Akbarian, M. Salzmann, L. Petersson, S. Gould, and J.M. Alvares. Built-in foreground/background prior for weakly-supervised semantic segmentation. In *ECCV*, 2016.
32) P. Tokmakov, K. Alahari, and C. Schmid. Weakly-supervised semantic segmentation using motion cues. In *ECCV*, 2016.
33) P. Krahenbuhl and V. Koltun. Efficient inference in fully connected CRFs with gaussian edge potentials. In *NIPS*, 2011.
34) J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, 2012.
35) S. Wataru and Y. Keiji. Distinct class saliency maps for weakly supervised semantic segmentation. In *ECCV*, 2016.
36) I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. In *NIPS*, 2014.
37) A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.
38) S. Ioffe and C. Szegedy. Batch normalization : Accelerating deep network training by reducing internal covariate shift. In *ICML*, pp. 448–456, 2015.
39) M. Mirza and S. Osindero. Conditional generative adversarial nets. In *arXiv : 1411.1784*, 2014.
40) S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text-to-image synthesis. In *ICML*, 2016.
41) X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel. In-fogan : Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*, 2016.
42) Jun-Yan Zhu, Philipp Kräahenbuühl, Eli Shechtman, and Alexei A. Efros. Generative visual manipulation on the natural image manifold. In *ECCV*, 2016.
43) G. Perarnau, J. van de Weijer, B. Raducanu, and J.M. Alvarez. Invertible condi-tional gans for image editing. In *NIPS*, 2016.
44) P. Upchurch, K. Gardner, K. Bala, R. Pless, N. Snavely, and K. Weinberger. Deep feature interpolation for image content

changes. In *arXiv : 1611.05507*, 2016.
45) T.Q. Chen and M. Schmidr. Fast patch-based style transfer of arbitary style. In *arXiv : 1612.04337*, 2016.
46) Y. Li, N. Wang, J. Liu, and X. Hou. Demystifying neural style transfer. In *arXiv : 1701.01036*, 2017.
47) H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. StackGAN : Text to photo-realistic image synthesis with stacked generative adversarial networks. In *arXiv : 1613.03242*, 2016.
48) I.J. Goodfellow. NIPS 2016 tutorial : Generative adversarial networks. In *arXiv : 1701.00160*, 2017.
49) Y. Zhou and T.L. Berg. Learning temporal transformations from time-lapse videos. In *ECCV*, 2016.
50) H. Vondrick, C. amd Pirsiavash and A. Torralba. Generating videos with scene dynamics. In *NIPS*, 2016.
51) J. Wu, C. Zhang, T. Xue, W.T. Freeman, and J.B. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *NIPS*, 2016.
52) S. Iizuka, E. Simo-Serra, and H. Ishikawa. Let there be color! : Joint end-to-end learning of global and local image priors for automatic image colorization with si-multaneous classificati-on. *ACM Trans. on Graphics* (*SIGGRAPH*), Vol. 35, No. 4, 2016.
53) W. Shi, J. Caballero, F. Huszar, et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016.
54) R. Garg, V.B.G. Kumar, and I.D. Reid. Unsupervised CNN for single view depth estimation : Geometry to the rescue. In *ECCV*, 2016.
55) A. Dosovitskiy, P. Fischer, E. Ilg, et al. Flownet : Learning optical flow with convo-lutional networks. In *ICCV*, 2015.
56) J. Johnson, A. Alahi, and L.F. Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.
57) A. Dosovitskiy and T.T. Brox. Inverting visual representations with convolutional networks. In *CVPR*, 2016.
58) I. Kokkinos. UberNet : Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *arXiv : 1609.02132*, 2016.
59) P. Isola, J.Y. Zhu, T. Zhou, and A.A. Efros. Image-to-image translation with conditional adversarial nets. In *arXiv : 1611.07004*, 2016.
60) L.A. Gatys, A.S. Ecker, and M. Bethge. A neural algorithm of artistic style. In *arXiv : 1508.06576*, 2015.
61) K. Yanai : Unseen Style Transfer Based on a Conditional Fast Style Transfer Network, In *ICLR WS*, 2017.

柳井啓司

1995 年東京大学工学部計数工学科卒業．1997 年東京大学大学院情報工学専攻修士課程修了．1997 年電気通信大学情報工学科助手．2003 年～2004 年文部科学省在外研究員として米国アリゾナ大学に滞在．2006 年電気通信大学情報工学科准教授．2015 年電気通信大学大学院情報学専攻教授．博士（工学）．一般物体認識，深層学習，マルチメディアデータマイニングなどに興味がある．情報処理学会，電子情報通信学会，人工知能学会の会員．

下田　和

2015 年電気通信大学総合情報学科卒業．2017 年電気通信大学大学院総合情報学専攻博士前期課程修了．現在，電気通信大学大学院情報理工学研究科情報学専攻博士後期在学中．日本学術振興会特別研究員 DC1．CNN を用いた画像の領域分割，およびその食事画像分析への応用，などに興味を持つ．