

完全教師あり学習手法を用いた 弱教師あり領域分割におけるシード領域生成方法の改良

下田 和^{1,†1,a)} 柳井 啓司^{1,b)}

概要：弱教師あり領域分割は精度が向上している。特に、弱教師あり領域分割結果を教師情報としてさらに学習する手法が提案され精度が大きく向上した。しかし、依然として、ノイズを含むアノテーションを用いた教師情報で学習を行うと精度が低下するという問題がある。本手法では、領域分割結果の一貫性から、領域分割の精度を推定し、領域分割の容易な画像を学習データから収集した。また、この推定結果から学習データのセレクションを行い、セレクションした学習データで再学習を行った。また、セレクションした画像を活用して data augmentation を行うとより精度が向上することを示した。

1. 導入

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り、教師情報の付与は大きなコスト、時間を要する。一般に、高度な教師情報が必要とする手法を完全教師あり学習領域分割、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領域分割と呼ぶ。弱教師あり学習による領域分割が可能となれば、大幅な学習データを収集するためのコストの削減が可能である。

弱教師あり領域分割を行う方法として画像の識別結果の可視化がある。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るため、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。CNN の認識結果の可視化による弱教師あり領域分割精度は、CNN 以前の完全教師あり領域分割制度を超えているが、CNN を活用した完全教師あり領域分割の精度とは大きな精度差がある。ただし、画像ラベルの識別に必要な領域と、実際の物体の領域には強い相関

があるものの、その間にはギャップが存在する。例えば、犬か猫かの認識においては、顔のみの情報で十分であり、手足を考慮する必要はない場合がほとんどである。また、ボートの認識などにおいては、海などの背景も共起の強い物体が認識に貢献する場合があります、可視化結果は共起の強い物体の領域にも反応する傾向がある。このように可視化結果と物体の領域の間にはギャップがあり、弱教師あり学習における大きな課題となっている。

近年、弱教師あり学習による領域分割結果を教師情報として領域分割用のネットワークを学習することでより高い精度で領域分割が可能であることが明らかになっている [1]。再学習を行う場合には全体のロスを最小化するためにパラメータを更新するので、学習データが十分大きければ突発的な誤差は学習における優先度合が下がる。これにより突発的な誤差が全体の領域の再学習により吸収され、全体の領域分割結果の精度が向上していると予測される。これはノイズを含む教師情報について CNN を活用した EM アルゴリズムの一種であると捉えることができる。この領域分割モデルの繰り返し学習手法は、初期の教師情報に含まれる誤りの傾向に一貫性があれば、その誤りの傾向も学習してしまうという欠点がある。再学習は小さな誤差を吸収する手法として有効であるが、分類結果の可視化結果と物体の領域のギャップを根本的に埋めることはできていない。

これを解決するために、近年はクラスラベル以外の物体領域と相関のあるもの、物体の大きさ、人のアテンション、物体らしさなどを教師情報として学習に取り入れており、精度向上が可能であることが示されている。これらの物体領域と相関のあるものを教師情報として扱う際には、新た

¹ 電気通信大学 総合情報学科

^{†1} 現在、同大学院 情報理工学研究科 情報学専攻所属

^{a)} shimoda-k@mm.inf.uec.ac.jp

^{b)} yanai@cs.uec.ac.jp

な教師情報の付与にコストがさらに必要となる。領域の教師情報と関連のあるものを活用し、精度を向上させる試みは有意義な試みであり、新たな教師情報の付与に必要なコストの大小と精度向上の度合いについての議論もされている。しかし、一方で SEC[2] などの教師情報と関連のあるものを活用せずに精度が向上可能であるという報告もある。教師なしのアプローチによる弱教師あり領域分割の検討はまだ不十分であり、議論の余地がある。そこで、本手法は教師なしの手法を弱教師あり領域分割に取り入れることに着目し、弱教師あり領域分割結果の領域分割精度を教師なしの手法により推定し、よい初期値のセレクションを行うことで再領域分割精度の向上を目指した。特に可視化手法による弱教師あり領域分割を行い、それぞれの一貫性を活用し、領域分割結果の容易性を推定した。以下に本研究の contribution を示す。

- 領域分割の容易性を領域分割結果から推定する手法を提案した。
- 提案手法が特に data augmentation と組み合わせることで有効に活用できることを示した。
- Pascal VOC 2012 benchmark における弱教師あり領域分割精度で他の既存手法と比較しても高精度を達成した。

2. 関連研究

2.1 クラス分類器の可視化手法

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関連があり、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 [3] 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 に対して、逆伝搬・backward 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 [4] 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 [4] は可視化結果を Grabcut の種とすることで高精度な弱教師あり領域分割を達成した。CNN の可視化結果から弱教師あり領域分割が可能であることを示した派生手法に guided backpropagation [5] がある。

2.2 CNN の activation を活用した弱教師あり領域分割

Oquab ら [6], Pathak ら [7] 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 は大まかな物体の位置を推定することが可能であり、逆伝搬による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 [8] や Zhou ら [9] により、GP の派生手法についても研究がなされている。また、Chen ら [10], Pathak ら [11] は GP 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。一方で、著者らは FCN を GP で学習した認識結果を backward を用いて可視化することで高精度な可視化を行った [12]。また、Simonyan ら [4] 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が、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ている。類似研究として、Jianming ら [13] による研究がある。

2.3 領域分割結果の再学習手法

その後、Wei ら [1] が Simple to Complex (STC) フレームワークを発表した。Wei ら [1] 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、弱教師あり領域分割の研究に大きな変化が生まれた。特に、その後の弱教師あり領域分割において、学習画像の領域分割手法（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堅牢な学習手法の重要性が増した。Chen ら [10], Pathak ら [11] は画像ラベルを活用して教師情報を動的に生成していたが、その方向性が強くなったという見方もできる。Kolesnikov ら [2] は、GP [9] 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 [14] は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 [15] 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。本手法においては、物体の領域と関連のある情報の活用ではなく、一貫性によ

る教師なしのアプローチに取り組んだ。

3. 提案手法

近年、領域の再学習により弱教師あり領域分割精度が向上している [1]。領域分割の再学習においては、以下の手順により領域分割のモデルを学習する。

- 既存の弱教師あり領域分割手法を用いて、学習画像について領域分割を行う。
- 学習画像の領域分割結果を領域の教師情報として、領域分割の識別モデルを学習する。

領域分割結果の再学習においては、学習画像を領域分割する点がこれまでの弱教師あり領域分割手法と大きく異なり、領域分割の際に学習画像のクラスラベルが既知のものとなる。本手法においては、領域分割結果を再学習するうえで、より高い精度で再学習を行う方法として、学習画像を領域分割した際の精度を推定し、領域分割の容易な画像を活用するというアプローチを行った。

3.1 領域分割結果の精度の推定

本研究においては、可視化結果が物体の領域を反映しているかどうかを、可視化結果の一貫性から推定した。特に、クラス応答の差分による可視化結果の一貫性と、入力画像サイズの変化についての可視化結果の一貫性の2つを用いて、領域分割結果の精度を推定した。また、この精度の推定結果から、学習画像について容易性の順序をつけ、順序の高いものから学習を行う。

3.1.1 差分による変化の一貫性

一般に、可視化結果は物体の顕著性マップのようになり、直接領域の推定を行うことは難しかったが、逆伝搬値について差分をとることによりクラス応答を鮮明にすることが可能である [12]。本研究では、この逆伝搬値について差分をとった場合の変化に着目した。差分をとった際に大きく結果が変化する場合には以下のような原因が考えられる。

- クラス分類に失敗し、クラス応答の勾配が消失した
- 対象のクラスの識別とは別に、顕著性のある物体が存在する

つまり、差分をとった場合に変化が小さい場合には、クラス分類が容易であり、顕著性マップと領域分割結果が一致していることが期待できる。そこで、本研究では、差分をとった際の変化から、領域分割結果の精度を推定した。画像 I における、差分なしによる領域分割結果を $V_o(x)$ 差分ありによる可視化領域分割結果を $V_w(x)$ としたとき、その領域分割結果の信頼度 $R_{sub}(x)$ を以下の式で定義する。

$$R_{sub}(x) = \sum_{c \in C} IoU(V_o^c(x), V_w^c(x)) \quad (1)$$

このとき、 $IoU(.,.)$ は各領域の Intersection over Union(IoU) を返す関数であるとする。

3.1.2 入力サイズの変化における一貫性

完全教師あり領域分割において、複数のスケールの入力から得られる結果を統合することで、精度を向上することが可能であるということが報告されている。また、認識結果の可視化による弱教師あり領域分割においても同様の報告がある。これは、dilation や spatial pyramid pooling と同様に、領域分割結果の各ピクセルで、局所的な情報と、大局的な情報を同時に表現し、分類結果に反映させることによるものであると考えられる。識別結果の可視化においては、画像の入力サイズを大きくした場合には局所的な認識の可視化結果、小さくした場合には大局的な認識の可視化結果となる傾向がある。本研究においては、この変化に着目し、この変化が大きい場合には、物体の大きさ、見え方の変化が大きいことが予測される。もし、入力画像のサイズを変化させて得られる可視化結果に一貫性があれば、領域分割が容易であり、高い精度で領域分割を行っていると期待できる。本研究ではこれを二つ目の領域分割の精度推定の評価指標とした。 $s_n = 320, 416, 512 (n = 0, 1, 2)$ 、を入力サイズ、それぞれのサイズにおいてクラス応答の差分により得られる領域分割結果を $V_{s_n}(x)$ とする。このとき、入力サイズの変化における一貫性 $R(I)_{size}$ を以下の式で定義する。

$$R_{size}(x) = \frac{1}{C} \sum_{c \in C} IoU(V_{s_0}^c(x), V_{s_1}^c(x), V_{s_2}^c(x)) \quad (2)$$

このとき、 $IoU(.,.,.)$ は3つの領域の IoU を返す関数であるとする。最終的な領域分割の信頼度は以下の式で計算した。

$$score = \lambda_1 \cdot R_{sub}(x) + \lambda_2 \cdot R_{size}(x) \quad (3)$$

R_{sub}, R_{size} は01の範囲の数値であるが、 R_{sub}, R_{size} のどちらかの指標の値の大きさの傾向からどちらかの指標に偏らないように、掛けあわせたものを評価指標とした。本研究においては、単純に $\lambda_1 = 0.5, \lambda_2 = 0.5$ とした。図1に本手法における領域分割が容易であると推測された例、容易でないと推測された例を示した。左のブロックが差分なしの可視化結果、右のブロックが差分ありの可視化結果である。また、それぞれについて、入力画像サイズの変化における可視化結果を示している。本手法においては、これらの条件の変化において一貫性がある場合に領域分割結果が信頼できるものとした。

3.2 領域の推定結果の一貫性を活用した領域評価

3.1節においては画像の領域分割の容易性を評価したが、本節においては、画像の各領域の信頼性評価する。Kolesnikov ら [2] は領域の再学習を行う際に GAP のアテンションにおける強く反応している極一部分のみの領域を教師情報として学習を行うことで高精度を達成した。本手

法においては、これを応用し 3.1 節における領域分割共通部分、一貫している領域を用いた。 $V_o(x)$, $V_w(x)$, $V_{s_n}(x)$ における共通領域のみを学習中に評価し、その他の領域については評価をスキップした。また、本研究においては、背景領域の影響が強く出る傾向があったために、背景領域は、2 クラス識別器における可視化結果との共通領域のみを評価するというパターンについて検証を行った。図 2 に生成したマスクの例を示す。

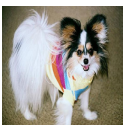
	input size	320	416	512	320	416	512
	visualization						
	CRF result						
		w/o sub			w/ sub		
	visualization						
	CRF result						
		w/o sub			w/ sub		
	visualization						
	CRF result						
		w/o sub			w/ sub		

図 1 異なる条件下における可視化手法による領域分割結果の例である。上段における結果は領域分割結果に一貫性があり、領域分割精度が高いと推定された例、中段、下段は領域分割結果に一貫性がないとして、領域分割精度が低いと推定された例である。

3.3 領域分割マスクの生成

学習における領域分割マスクを手法 [12] により生成する。[12] に従い、各種パラメータを設定し可視化を行い、全結合 CRF を行った。また、本研究では問題を簡単化するために、対象の物体が存在するか存在しないかの 2 クラス識別器を用い、20 クラスのカテゴリについてそれぞれ Softmax で学習した。また、本研究では異なる条件により得られた領域分割結果の共通部分を論文 [2] で用いられた localization cue として学習に用いた。図 2 は生成された領域分割マスクの例である。

本手法では、セクション 3.2 で求めた領域分割結果の容易性を用いて、生成した領域分割結果から評価の高い教師情報を mining することにより精度向上を狙う。また、本研究では、本手法により収集した精度の高い学習データについてのみ data augmentation を適用を行い、その精度の変

化を検証する。入力画像を小さな変化を与えることによる教師情報の水増しが深層学習において有効であることは現在広く知られている。しかし、物体検出において、より大きな変化を与えることで精度が大幅に向上していることなどが報告されている [16](論文におけるポスター発表)。data augmentation については、[16] を参照し、random crop と random padding を行った。ただし、crop された画像については、可視化において用いたマルチクラス識別器で識別し、これが画像と対応する 3.2 節で生成した領域の教師情報と一致していた場合に学習に活用した。

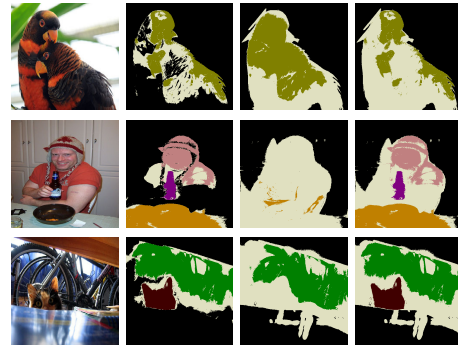


図 2 (1) 入力画像, (2) single-class model により生成されたマスク, (3) multi-class model により生成されたマスク, (4) (2),(3) の統合により得られたマスク

4. 実験

4.1 データセット

領域分割のベンチマークのデータセットとして PASCAL VOC 2012 [17] を用いた。PASCAL VOC 2012 における領域分割ベンチマークでは、20 の異なるクラスと back ground クラスを含む 21 のクラスの領域分割の精度を比較する。また、1464 枚の train 画像、1449 枚の validation 画像、1456 枚の test 画像があるが、近年は [18] により提供された 10582 枚の train_aug 画像を用いるのが一般的となっている。本実験においてもこれを学習画像として用いた。評価指標は mean IoU を用い、ピクセルの一致度を評価する。ただし、評価は一般に公開されている Pascal のサーバーを用いて行った。また、可視化を行う際には [12] と同様に VGG16 モデル [19] を用いた。再学習の際に用いる領域分割の識別モデルとしては VGG16 モデル [19] の派生である Deeplab モデル [20] を用いた。

4.2 Segmentation model の学習

マルチクラス識別器とシングルクラス識別器の学習における設定については、[12] に従った。fully-supervised segmentation model としては DeepLab-CRF model [20] を採用した。学習には SGD を使い、バッチサイズは 16, momentum は 0.9, weight decay is 0.0005, learning rate

の初期値は最終層を 0.01, それ以外を 0.001 とした。learning rate は 2000 iteration ごとに 0.1 さげた。各モデルは NVIDIA GeForce Titan GPU(12GB memory) を用いて 7~8 時間で学習を行った。また、これらの実験はオープンソースとして公開されている Chen ら [20] によって提供されている Caffe framework [21] による実装を活用した。

4.3 領域分割の容易性の推定結果の評価

図 3 は本手法により、収集された各クラスにおけるトップ 5 の学習データである。本手法によるよい領域分割結果のマイニング結果は Un-supervised なアプローチであるにもかかわらず高い精度の領域分割結果を推定することができているとわかる。特に, aero や bus, car, cow, dog などは ground truth に近い結果を収集できている。一方で, sofa, chair, table などは上位の結果であっても誤りを含む結果が多い。一般にこれらのクラスは弱教師あり領域分割で低い精度なるため, これが直接関係していると予測される。

表 1 は教師あり領域分割のモデルを学習する際に, 本手法により収集した学習データを活用してセレクションを行い, data augmentation を適用した場合の精度変化についての表である。なお, 各画像について, crop, padding を行い, 各 10 枚ずつ計 20 枚データの水増しを行い, 選択する学習データの枚数は式 3 の Threshold により決定した。結果としては, 学習するデータをセレクションにより 8760 枚, data augmentation に用いる画像枚数をさらにセレクションし 2105 枚にした場合の精度がもっとも高くなり, 51.3 %となった。一方で, セレクションを行わずにすべてのデータで data augmentation を行った場合には 48.8 であったことを考えると本手法が有効であったことがわかる。また, data augmentation を行わない画像についてもセレクションが有効であったことが分かる。ただし, data augmentation に用いる画像枚数をさらに減らし 730 枚とすると精度が下がってしまった。深層学習における学習サンプルの数とクオリティはトレードオフであるが, セレクションと data augmentation を活用することで学習サンプルが少ない場合にも最大限に生かすことができている。また, これらの結果は validation set における評価結果である。

表 3 に他の弱教師あり領域分割手法との比較を示した。本手法は, 画像ラベルのみを用いる手法において, もっとも高い精度を達成している。特に, 領域の再学習を行っている F/B prior[14], STC [1], SEC [2] と比較しても高い精度を達成した。また, 本手法は CNN を用いない完全教師あり領域分割手法である O2P [22], プロポーザルと CNN を組み合わせた完全教師あり領域分割手法である SDS [23] の精度を上回った。実験は Pascal VOC test set における

ものであり, 評価に用いた画像は他の表における結果のものと異なっている。図 4 は本手法の領域分割結果の例である。

表 1 学習に用いる画像枚数を変化させた場合の領域分割精度

setting	base image N	aug image N	mIoU
(a)	8760 (th ≥ 0.3)	730 (th ≥ 0.8)	50.1
(b)	10582 (all)	730 (th ≥ 0.8)	48.9
(c)	8760 (th ≥ 0.3)	2105 (th ≥ 0.7)	51.3
(d)	10582 (all)	2105 (th ≥ 0.7)	49.9
(e)	8760 (th ≥ 0.3)	8760 (th ≥ 0.3)	49.7
(f)	10582 (all)	10582 (all)	48.8

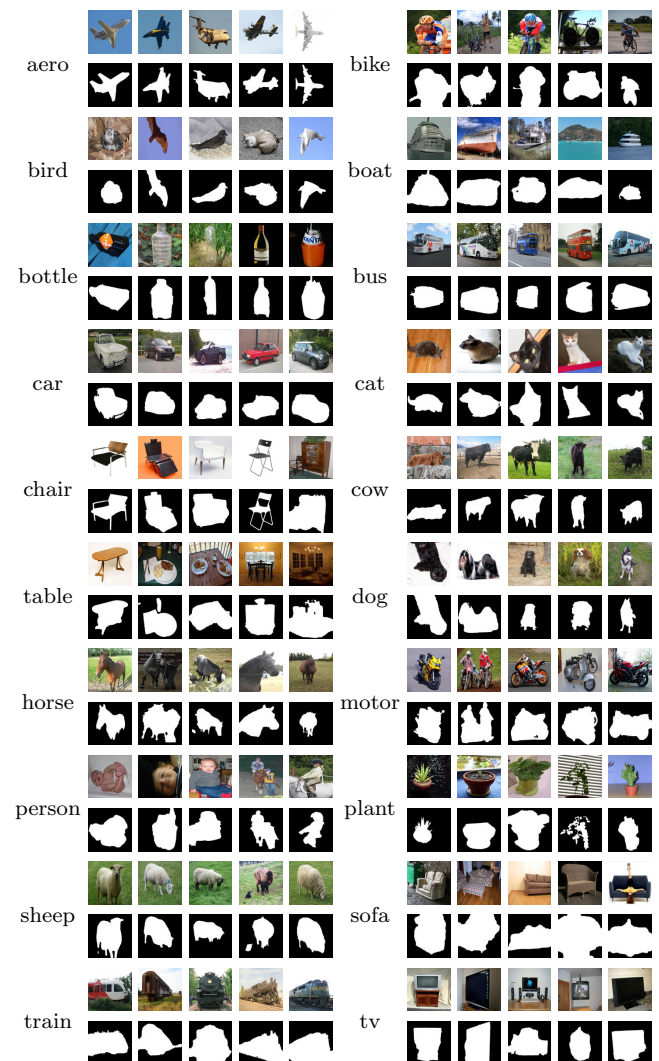


図 3 Pascal VOC 2012 train_aug dataset における容易性の推定結果から収集されたよい領域分割結果の top5。

5. 結論

本研究においては, 認識結果の可視化結果から, 領域分割の精度の善し悪しを推定し, これを活用したデータの水増しにより精度を向上させたが, Pascal VOC における weakly supervised segmentation において最高精度を達成した。今

後は容易な画像の精度の推定結果をさらに応用させていきたい。特に、半教師あり領域分割や few shot learning など、weakly-supervised の手法以外への応用を考えていきたい。

参考文献

- [1] Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M., Zhao, Y. and Yan, S.: STC: A Simple to Complex Framework for Weakly-supervised Semantic Segmentation, *ECCV* (2016).
- [2] Kolesnikov, A. and H.Lampert, C.: Seed, Expand and Constrain: Three Principles for Weakly-Supervised Image Segmentation, *ECCV* (2016).
- [3] Zeiler, M. and Fergus, R.: Adaptive deconvolutional networks for mid and high level feature learning, *ICCV* (2011).
- [4] Simonyan, K., Vedaldi, A. and Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps, *ICLR WS*, (online), available from <http://arxiv.org/abs/1312.6034> (2014).
- [5] Springenberg, J. T., Dosovitskiy, A., Brox, T. and Riedmiller, M.: Striving for Simplicity: The All Convolutional Net, *ICLR WS*, (online), available from <http://arxiv.org/abs/1412.6806> (2015).
- [6] Oquab, M., Bottou, L., Laptev, I. and Sivic, J.: Learning and Transferring Mid-level Image Representations Using Convolutional Neural Networks, *CVPR* (2014).
- [7] Pathak, D., Shelhamer, E., Long, J. and Darrell, T.: Fully convolutional multi-class multiple instance learning, *ICLR* (2015).
- [8] Pedro, P. and Ronan, C.: From Image-level to Pixel-level Labeling with Convolutional Networks, *CVPR* (2015).
- [9] Zhou, B., Khosla, A., Lapedriza, A. and Oliva, A. Torralba, A.: Learning Deep Features for Discriminative Localization, *CVPR* (2016).
- [10] Papandreou, G., Chen, L.-C., Murphy, K. and Yuille, A. L.: Weakly-and semi-supervised learning of a dcnn for semantic image segmentation, *ICCV* (2015).
- [11] Pathak, D., Krahenbuhl, P. and Darrell, T.: Constrained convolutional neural networks for weakly supervised segmentation, *ICCV* (2015).
- [12] Shimoda, W. and Yanai, K.: Distinct Class Saliency Maps for Weakly Supervised Semantic Segmentation, *ECCV* (2016).
- [13] Jianming, Z., Zhe, L., Jonathan, B., Xiaohui, S. and Sclaroff, S.: Top-down Neural Attention by Excitation Backprop, *ECCV* (2016).
- [14] Saleh, F., Akbarian, M., Salzmann, M., Petersson, L., Gould, S. and M.Alvares, J.: Built-in Foreground/Background Prior for Weakly-Supervised Semantic Segmentation, *ECCV* (2016).
- [15] Tokmakov, P., Alahari, K. and Schmid, C.: Weakly-Supervised Semantic Segmentation using Motion Cues, *ECCV* (2016).
- [16] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y. and Berg, A. C.: SSD: Single Shot MultiBox Detector, *ECCV* (2016).
- [17] Everingham, M., Eslami, S. M. A., Van Gool, L., Williams, C. K. I., Winn, J. and Zisserman, A.: The Pascal Visual Object Classes Challenge: A Retrospective, *International Journal of Computer Vision*, Vol. 111, No. 1, pp. 98–136 (2015).
- [18] Hariharan, B., Arbelaez, P., Bourdev, L., Maji, S. and J., M.: Semantic contours from inverse detectors, *ICCV* (2011).
- [19] Simonyan, K., Vedaldi, A. and Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, *ICLR* (2015).
- [20] Chen, L., Papandreou, G., Kokkinos, I., Murphy, K. and L., Y. A.: Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs, *ICLR* (2015).
- [21] Jia, Y.: Caffe: An Open Source Convolutional Architecture for Fast Feature Embedding, <http://caffe.berkeleyvision.org/> (2013).
- [22] Carreira, J., Caseiro, R., Batista, J. and Sminchisescu, C.: Semantic Segmentation with Second-Order Pooling, *ECCV* (2012).
- [23] Hariharan, B., Arbeláez, P., Girshick, R. and Malik, J.: Simultaneous Detection and Segmentation, *ECCV* (2014).
- [24] Long, J., Shelhamer, E. and Darrell, T.: Fully Convolutional Networks for Semantic Segmentation, *CVPR* (2015).
- [25] Bearman, A., Russakovsky, O., Ferrari, V. and Fei-Fei, L.: What 's the Point: Semantic Segmentation with Point Supervision, *ECCV* (2016).

表 2 Results on PASCAL VOC 2012 *val set*.

Methods	additional images	bg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU
MIL-FCN [7]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	25.7
EM-Adapt [10]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	38.2
CCNN [11]	-	65.9	23.8	17.6	22.8	19.4	36.2	47.3	46.9	47.0	16.3	36.1	22.2	43.2	33.7	44.9	39.8	29.9	33.4	22.2	38.8	36.3	34.5
MIL-spxl [8]	✓	77.2	37.3	18.4	25.4	28.2	31.9	41.6	48.1	50.7	12.7	45.7	14.6	50.9	44.1	39.2	37.9	28.3	44.0	19.6	37.6	35.0	36.6
MIL-bb [8]	✓	78.6	46.9	18.6	27.9	30.7	38.4	44.0	49.6	49.8	11.6	44.7	14.6	50.4	44.7	40.8	38.5	26.0	45.0	20.5	36.9	34.8	37.8
MIL-seg [8]	✓	79.6	50.2	21.6	40.6	34.9	40.5	45.9	51.5	60.6	12.6	51.2	11.6	56.8	52.9	44.8	42.7	31.2	55.4	21.5	38.8	36.9	42.0
DCSM w/ CRF [12]	-	76.7	45.1	24.6	40.8	23.0	34.8	61.0	51.9	52.4	15.5	45.9	32.7	54.9	48.6	57.4	51.8	38.2	55.4	32.2	42.6	39.6	44.1
F/B prior[14]	-	79.2	60.1	20.4	50.7	41.2	46.3	62.6	49.2	62.3	13.3	49.7	38.1	58.4	49.0	57.0	48.2	27.8	55.1	29.6	54.6	26.6	46.6
STC[1]	✓	84.5	68.0	19.5	60.5	42.5	44.8	68.4	64.0	64.8	14.5	52.0	22.8	58.0	55.3	57.8	60.5	40.6	56.7	23.0	57.1	31.2	49.8
SEC [2]	-	82.4	62.9	26.4	61.6	27.6	38.1	66.6	62.7	75.2	22.1	53.5	28.3	65.8	57.8	62.3	52.5	32.5	62.6	32.1	45.4	45.3	50.7
Ours	-	81.6	64.9	25.8	71.4	29.2	57.8	75.2	68.0	72.7	15.2	46.6	33.8	56.7	57.1	60.9	60.7	24.1	65.4	31.5	43.9	35.3	51.3

表 3 Results on PASCAL VOC 2012 *test set*.

Methods	bg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mIoU
Fully Supervised:																						
O2P [22]	85.4	69.7	22.3	45.2	44.4	49.6	66.7	57.8	56.2	13.5	46.1	32.3	41.2	59.1	55.3	51.0	36.2	50.4	27.8	46.9	44.6	47.6
SDS [23]	86.3	63.3	25.7	63.0	39.8	59.2	70.9	61.4	54.9	16.8	45.0	48.2	50.5	51.0	57.7	63.3	31.8	58.7	31.2	55.7	48.5	51.6
FCN-8s [24]	-	76.8	34.2	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1	62.2
DeepLab Large FOV [20]	92.6	83.5	36.6	82.5	62.3	66.5	85.4	78.5	83.7	30.4	72.9	60.4	78.5	75.5	82.1	79.7	58.2	82.0	48.8	73.7	63.3	70.3
Using Additional Supervision:																						
CCNN w/size [11]	-	42.3	24.5	56.0	30.6	39.0	58.8	52.7	54.8	14.6	48.4	34.2	52.7	46.9	61.1	44.8	37.4	48.8	30.6	47.7	41.7	45.1
One point[25]	80.6	50.2	23.9	38.4	33.1	38.5	52.0	50.9	55.4	18.3	38.2	37.7	51.0	46.1	54.7	43.2	35.4	45.1	33.0	49.6	40.0	43.6
F/B prior + CheckMask[14]	87.4	65.7	26.0	64.2	43.7	53.2	72.6	63.6	59.5	17.1	48.0	43.7	61.2	52.0	69.3	54.8	43.0	50.3	34.6	59.2	42.0	52.9
Weakly Supervised:																						
MIL-FCN [7]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	24.9
EM-Adapt [10]	76.3	37.1	21.9	41.6	26.1	38.5	50.8	44.9	48.9	16.7	40.8	29.4	47.1	45.8	54.8	28.2	30.0	44.0	29.2	34.3	46.0	39.6
CCNN [11]	-	21.3	17.7	22.8	17.9	38.3	51.3	43.9	51.4	15.6	38.4	17.4	46.5	38.6	53.3	40.6	34.3	36.8	20.1	32.9	38.0	35.5
MIL-ILP-seg [8]	78.7	48.0	21.2	31.1	28.4	35.1	51.4	55.5	52.8	7.8	56.2	19.9	53.8	50.3	40.0	38.6	27.8	51.8	24.7	33.3	46.3	40.6
DCSM w/ CRF [12]	78.1	43.8	26.3	49.8	19.5	40.3	61.6	53.9	52.7	13.7	47.3	34.8	50.3	48.9	69.0	49.7	38.4	57.1	34.0	38.0	40.0	45.1
F/B prior[14]	80.3	57.5	24.1	66.9	31.7	43.0	67.5	48.6	56.7	12.6	50.9	42.6	59.4	52.9	65.0	44.8	41.3	51.1	33.7	44.4	33.2	48.0
STC [1]	85.2	62.7	21.1	58.0	31.4	55.0	68.8	63.9	63.7	14.2	57.6	28.3	63.0	59.8	67.6	61.7	42.9	61.0	23.2	52.4	33.1	51.2
SEC [2]	83.5	56.4	28.5	64.1	23.6	46.5	70.6	58.5	71.3	23.2	54.0	28.0	68.1	62.1	70.0	55.0	38.4	58.0	39.9	38.4	48.3	51.7
Weakly Supervised:																						
Ours	83.0	67.5	29.7	69.7	28.8	59.7	71.2	66.4	69.8	18.6	49.8	44.7	49.4	60.5	73.5	61.8	32.7	62.7	39.0	34.3	36.5	52.8

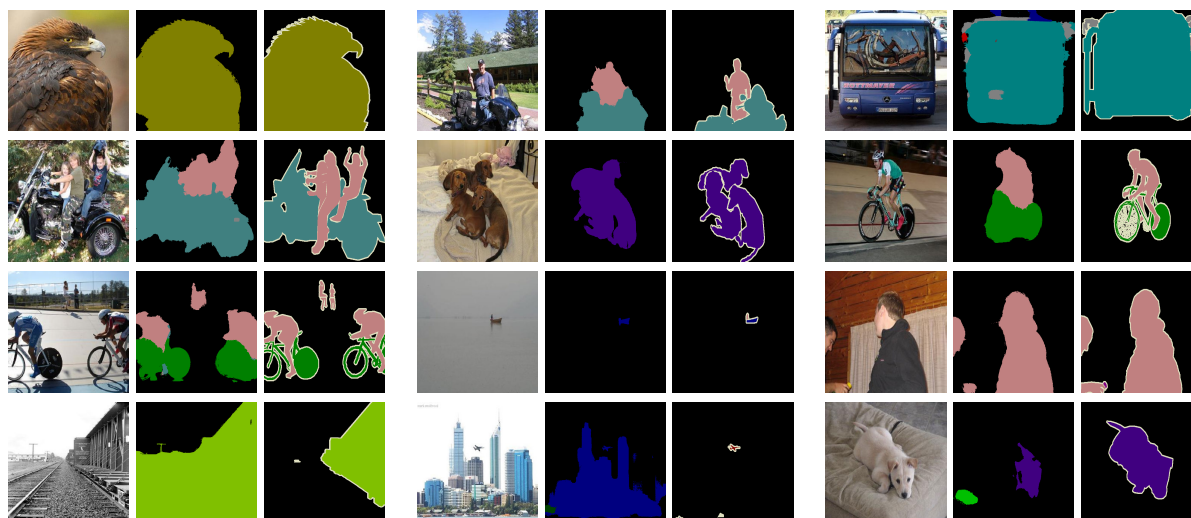


図 4 左から入力画像, 推定結果, 真値である。