〈一般研究課題〉　少ない教師付き画像による
セマンティックセグメンテーション
助 成 研 究 者　名城大学　　堀田　一弘

少ない教師付き画像による セマンティックセグメンテーション

堀田　一弘
（名城大学）

Semantic Segmentation with A Small Number of Training Images

Kazuhiro Hotta
(Meijo University)

Abstract：

In general, semantic segmentation by deep learning requires training images with pixel-level ground truth labels. Manual pixel-level annotation needs a lot of time and cost. Thus, in this paper, we propose semantic segmentation method from a small number of training images. To train good features for semantic segmentation from a small number of training images, we utilize the relation of pixels in addition to standard pixel-wise training. When we select certain pixel in an image, we increase the similarity of pixels with the same class and decrease the similarity of pixels with different classes. Because the proposed training method is used at only training phase, the computational cost of test phase does not decrease. This is the merit of the proposed method. We confirmed the effectiveness of our method by experiments on two kinds of datasets.

1. はじめに

セマンティックセグメンテーションは自動運転や医療用画像処理などの高精度な画素単位の識別を必要とする業界で応用が期待され，近年活発に研究が行われている[1-4]．セマンティックセグメンテーションではほとんどの場合，学習データに正解ラベルをつけて学習する「教師あり学習」を用いる．教師あり学習を用いるためには画素単位でのラベル付け作業が必要となり，ここに大きな手間と時間がかかってしまう．

本論文では，ラベル付けに伴うコスト削減を目的とし，少数の教師付き画像からセグメンテーションの学習を行う方法を提案する．具体的には，少ない教師付き画像から多様な特徴量を抽出するために，画素間の関係を用いる．

画素間の関係とは，ある画素と同じクラスの画素の特徴量は互いに類似し，異なるクラスの画素の特徴量は互いに類似していないということである．Cos類似度により画像中の画素間の類似度を計算し，同じクラスの画素の特徴量の類似度を1に近づけ，異なるクラスの画素間の類似度を-1に近づけるように学習する．画素間の類似度を学習させることにより，同じクラスの画素から得られる特徴量が類似したものとなる．学習画像の枚数が少ない場合でも画素間の組み合わせは膨大となるため，少ない教師付き画像からでもセマンティックセグメンテーションの精度を向上させることができるのではないかと考えられる．

画素間での類似度の計算する際，1枚の画像内の画素間の類似度のみを用いて学習する場合と，2枚の画像内の画素間の関係を用いて類似度の学習を行う場合が考えられる．本稿ではその両方を試した．また，セマンティックセグメンテーションでは面積の大きいクラスと小さいクラスが存在し，画素間の関係性を用いる場合でもランダムに画素を選択すれば面積の影響を受けてしまう．そこで，全てのクラスの画素を均等に選択する工夫も行った．

評価実験では，COVID-19データセット[5]とショウジョウバエの細胞画像[6]の2種類のデータセットを用いて学習用画像を4枚だけ用いた場合において実験を行った．COVID-19データセットを用いた実験では，従来法であるU-NetのmIoUは31.33%であったが，提案手法では，1枚の画像内の画素間の関係を用いた場合が35.05%，2枚の画像の画素間の関係を用いた場合が34.55%となり，どちらの場合も従来法と比較して3%以上精度を向上することができた．また，ショウジョウバエの細胞画像を用いた場合も従来法は56.78%，２つの提案手法の精度は63.41%と64.20%となり，どちらの場合も従来法と比較して6.5%以上精度を向上することができた．

本稿の構成は以下の通りである．2節では関連研究について述べ，3節では提案手法の詳細を述べる．4節で実験結果を示し，5節で結論と今後の課題を述べる．

２．関連研究

2.1節ではセマンティックセグメンテーションについて述べる．2.2節ではContrastive Learningの説明を行う．

2.1 セマンティックセグメンテーション

セグメンテーションでは，エンコーダ・デコーダ構造を持つSegNet[2]やU-Net[1]が有名である．また，深い層まで勾配を伝えることができるResNet[7]をバックボーンとして利用し，Spatial Pyramid Poolingの構造を用いたPSPNet[3]やDeeplab[8]など多くのセマンティックセグメンテーション法が提案されている．しかし，一般に，複雑なネットワークを用いる場合には大量の教師付き画像が必要となるという問題がある．特に，セグメンテーションでは1画素ごとに人間が手作業でラベル付けする必要があるため，学習に十分な教師画像を用意するには多大なコストがかかり，Cityscapesデータセットのアノテーションには画像1枚あたり90分かかるとの報告もある[9]．そこで本論文では，ラベル付けのコストを削減するために，少ない教師付き画像からセグメンテーションの精度向上を目指す．

2.2 Contrastive Learning

深層学習を用いたセグメンテーションの研究のほとんどが教師あり学習を用いている．しかし，2.1節で述べたようにアノテーションのコストが非常に大きい．これに対し，SimCLR[10]やBYOL[11]などの教師なし表現学習で用いられるContrastive Learning[12]では画像同士の類似度を学習することにより，識別に適した空間を作成する．一般に，Contrastive Learningは画像間の類似度の学習に利用されるが，本稿では少ない教師付き画像からセグメンテーションの学習を行うために，Contrastive Learningの考え方を画素間の類似度の学習に応用する．学習枚数が少なくても画素間の組み合わせは膨大にあるため，セグメンテーションのための良い学習ができると考えられる．

３．提案手法

本節では少ない教師付き画像からセマンティックセグメンテーションの学習を行う方法を提案する．具体的には，少ない教師付き画像から多様な特徴量を抽出するために画素間の関係を用いる．3.1節で画素間の関係について述べ，3.2節で画素間の関係を用いた学習法について述べる．

3.1 画素間の関係

上述のように，Contrastive Learningが画像間で特徴量の学習していたのを画素間にも利用できないかと考え，提案するのが本稿の手法になる．同じクラスの画素間の特徴量の類似度を上げ，異なるクラスの画素間の特徴量の類似度を下げることにより，少数の教師付き画像から学習を行う．画像間で比較して学習を行う場合，画像変換を行ったとしても比較する画像がいくつも必要になるが，画素間で比較することにより1枚の画像内でも膨大な組み合わせを学習できる．

画素間の関係と言っても，1枚の画像内の画素間の関係を用いた場合と異なる2枚の画像の画素間の関係を用いた場合がある．本稿では両方とも試してみる．提案手法で画素間の関係を利用するのは学習時のみのため，テスト時の計算コストが変わらないのも利点の１つである．

3.2 1枚の画像内の画素間の関係を用いた学習法

1枚の画像内の画素間の関係を用いた学習法を説明する．学習の概要を図1に示す．今回は医用画像と細胞画像を用いるため，U-Netをベースラインとして用いた．まず初めに，入力画像をU-Netに入力し，最終層の１つ前の層でクラス数分の特徴マップを得る．次に，クラス数分の特徴マップから図1のように各クラスから1画素ずつ選ぶ．選択した画素と特徴マップ内の全ての画素に対してCos類似度を計算する．例えば，4クラスのセグメンテーションの場合には，4枚の類似度マップ(sim0, sim1, sim2, sim3)が得られる．Cos類似度は特徴量が似ていると1，似ていないと-1に近い値が出力される．そこで，画素の正解ラベルを用いて選択した画素と同じクラスの画素には1，異なる画素には-1を入れた4枚の正解マップ(map0, map1, map2, map3)を作成し，クラス毎に類似度マップと正解マップの二乗誤差を学習させる．例えば，クラス0の学習には下記のロスを用いる．

$$loss0 = \ (sim0 - map0)^2$$

このロスをクラス数分用意し，通常のセグメンテーションの学習に用いられる各画素のクロスエントロピーロスと各クラスの二乗誤差のロスの和を用いて学習する．

ただし，セグメンテーションではクラス毎に面積が大きく異なり，ランダムに画素を選択すると

その影響を受けやすい．そこで，特徴マップをクラス数分用意し，1回の学習で全てのクラスの画素を選択することにより，面積の違いによる影響を低減しながら学習する．最後に，得られた特徴マップをチャンネル方向に連結し，畳み込みをすることによりセグメンテーションを行う．

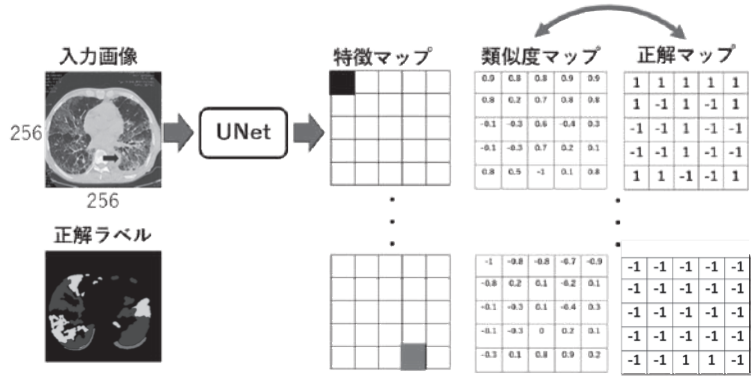

図1　1枚の画像内の画素間の関係を用いた学習法

3.3 異なる2枚の画像内の画素間の関係を用いた学習法

異なる2枚の画像内の画素間の関係を用いた学習法について説明する．図2に学習の概要を示す．この方法は前節の学習法と類似度マップの作成方法が異なる．前節の方法では，1枚の画像から得られた特徴マップから選択した画素と同じ画像内の全ての画素間のCos類似度を計算したが，本項で提案する学習法では異なる2枚の画素間の関係を利用し，片方の画像の特徴マップから選んだ画素ともう一方の画像の全画素とのCos類似度を基に類似度マップを作成する．以降は前節と同様の手順で正解マップを作成し，クロスエントロピーロスに加えてクラス毎に類似度マップと正解マップの二乗誤差を学習する．

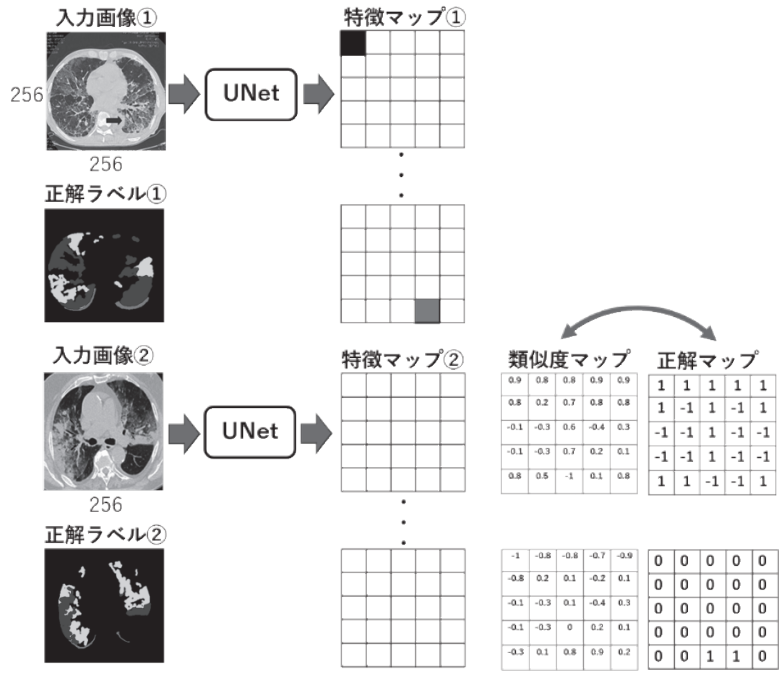

図2　異なる2枚の画像内の画素間の関係を用いた学習法

４．評価実験

本節では評価実験の結果を示す．まず使用したデータセットについて4.1節で説明する．4.2節で実験条件，4.3節で実験結果について述べる．

4.1 データセット

データセットにはCOVID-19データセット，ショウジョウバエの細胞画像の2種類のデータセットを用いた．COVID-19データセットには，256×256画素の画像データに対して4クラスがアノテーションされており，学習用画像70枚，検証用画像10枚，評価用画像20枚を用いた．ショウジョウバエの細胞画像では 3クラスがアノテーションを用いた．元画像は1024×1024画素であるが，そこから256×256画素の領域を切り出し，学習用画像192枚，検証用画像48枚，評価用画像80枚を用いた．

4.2 実験条件

医用画像のセグメンテーションに利用されるU-Netをベースラインとした．学習時のBatch Sizeは4，最適化手法にAdam($lr=1\times10^{-4}$)を用いた．またパラメータの初期値の乱数を変更して5回実験を行い，5回の平均値を評価に用いた．評価指標はMean IoU (mIoU)とした．

4.3 実験結果

表1および図3にCOVID-19データセットを用いた検証結果を示す．全体的なセグメンテーション結果としてあまり差は見られないが，画像中で最もラベルが少ない「Consolidations」のクラスを示す赤色の結果に着目した場合，U-Netではほとんど識別することができていないが，提案手法では識別できている箇所もある．表1からは，2種類の提案手法のどちらも従来法のmIoUから3%以上精度が向上することを確認した．特に「Consolidations」のクラスにおいて大幅に精度が向上した．通常，面積の小さいクラスは予測が難しく，精度が低くなってしまう．しかし，提案手法では画素間の関係を用いることにより，面積の小さいクラスである「Consolidations」において7%以上の大幅な精度向上が確認できた．

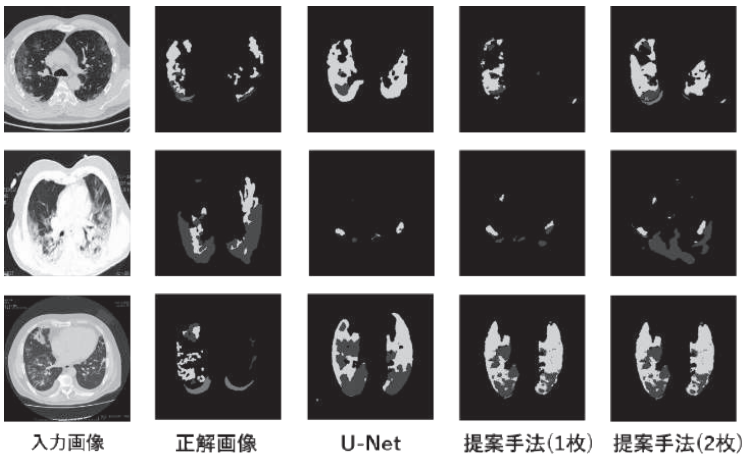

図3　COVID-19の出力結果の比較

表1　COVID-19のIoU比較

	IoU(%)		
Label	従来手法	提案手法 (1 枚)	提案手法 (2 枚)
Back ground	**92.88**	91.89	91.92
Lung other	6.05	6.61	**6.85**
Ground glass	27.72	**34.17**	23.51
Consolidation	0.03	7.52	**8.33**
mean	31.33	**35.05**	34.55

次に，表2および図4にショウジョウバエの細胞画像を用いた検証結果を示す．図4を見ると，正解画像と比較してU-Netの出力画像はクラスの境目となる箇所が粗いセグメンテーションとなっていることが分かる．一方，提案手法の出力結果では，一部誤って予測をしている箇所はあるがU-Netの出力結果に比べてクラス境界のセグメンテーション精度が高い．表2の結果から，2種類の提案手法のどちらも従来手法のmIoUから6.5%以上精度を向上していることを確認した．また，特に「Mitochondria」のクラスでは10%以上も精度が向上した．提案手法ではmIoUの精度向上はもちろん，面積の小さいクラスにおいて10%以上の精度を向上させることができた．これは画素間の関係の学習およびクラスを均等に選択することの有効性を示している．また，表1，2のデータセット間で比較した場合，２つの提案手法の優越に違いがでた．この理由は学習画像の個体差だと考えられる．表2のショウジョウバエの画像は細胞レベルであるため個体差の影響は少なく，異なる画像の画素間の特徴量を上手く学習することができ，精度向上に繋がったと考えられる．一方，表1のCOVID-19データセットは人の肺のCT画像のため，学習画像内の個体差が原因で2枚を用いる方法では学習が難しくなったと考えられる．

2つのデータセットの検証結果から，クラス毎の画素間の関係を用いた学習は，少数の教師付き学習において精度を向上させ，従来では識別が難しかった面積の小さいクラスのセグメンテーション精度を向上させることができた．

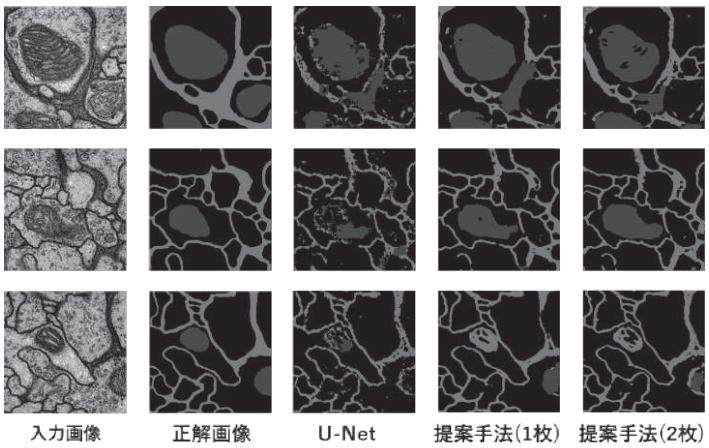

図4　ショウジョウバエの細胞画像の出力結果の比較

表2　ショウジョウバエの細胞画像のIoU比較

	IoU(%)		
Label	従来手法	提案手法(1 枚)	提案手法(2 枚)
Back ground	87.02	88.68	**88.87**
Membrane	54.98	62.17	**63.45**
Mitochondria	28.34	40.01	**40.27**
mean	56.78	63.41	**64.20**

5．おわりに

本論文では，セグメンテーションの学習に必要な教師付き画像の枚数を削減することを目的とし，画素間の関係を用いた少数の教師付き画像からのセグメンテーション法を提案した．クラス毎に画素間の関係を用いて類似度を学習することにより，学習に用いる教師付き画像のデータ数を抑えつつ精度を向上させることができた．今後はクラス数が多いデータセットでも評価し，クラス数が増えて予測が難しくなっても提案手法の有効性が示せることを確認したい．また，教師なしデータに擬似ラベルを付与することにより，さらなる精度の向上を目指していきたい．

参考文献

[1] J. Long, E. Shelhamer, and T. Darrell, Fully convolutional networks for semantic segmentation, *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431- 3440, 2015.

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp2481-2495, 2017.

[3] H. Zhao, X. Qi, X. Wang, and J. Jia, Pyramid Scene Parsing Network, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.2881-2890, 2017.

[4] H. Tsuda and K. Hotta, Cell Image Segmentation by Integrating Pix2pix2 for Each Class, *CVPR Workshop on Computer Vision for Microscopy Image Analysis*, pp. 1065-1073, 2019.

[5] COVID-19 CT segmentation dataset, https://medicalsegmentation. com/COVID19/ , 2020.

[6] S. Gerhard, J. Funke, J. Martel, A. Cardona, R. Fetter, Segmented anisotropic ssTEM dataset of neural tissue. figshare. Retrieved 16:09, 2013.

[7] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.770-778, 2016.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.40, No.4, pp.834-848, 2017.

[9] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, The cityscapes dataset for semantic urban scene understanding, *IEEE Conference on Computer Vision and Pattern Recognition*, pp.3213-3223, 2016.

[10] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, A simple framework for contrastive learning of visual representations, International Conference on Machine Learning, pp.1597-1607, 2020.

[11] P. H. Richemond, et al., Bootstrap your own latent-a new approach to Self-supervised Learning, Advances in Neural Information processing Systems, Vol. 33, pp. 21271- 21284, 2020.

[12] A. Jaiswal, AR. Babu, MZ. Zadeh, D. Banerjee, F. Madedon, A Survey on Contrastive Self-supervised Learning, Technologies, Vo.9, No.1, pp.2, MDPI, 2020.