Vol.2018-CG-172 No.4
Vol.2018-DCC-20 No.4
Vol.2018-CVIM-214 No.4
2018/11/7

セマンティックセグメンテーションにおけるハイパーパラメータの自動選択と室内画像からの床領域抽出への適用

大石 涼火[1]　数藤 恭子[1,a)]　佐藤 文明[1,b)]

概要：室内の廊下や部屋において，障害物がない床の領域をカメラから認識することが，車椅子への応用等で必要とされている．近年，道路や室内の画像を対象に高精度なセマンティック・セグメンテーションを可能とする各種の深層学習のモデルが示されているが，ユーザの所望の環境に適合させるには，学習データのアノテーションにコストを要するため，学習データが十分に得られないことが多い．本研究では，室内画像の領域を床・ドア・人・障害物などのカテゴリ別に認識することを目的として，実際の運用時に想定可能な程度の少ない学習データのもとで，ハイパーパラメータが性能にどのように寄与しているのかについて検討を行った．セマンティックセグメンテーションのモデルの一つである U-Net において，エンコード側からのデコード側へのパスに Dropout を挿入し，その割合を含むいくつかのハイパーパラメータを遺伝的アルゴリズムと Dice 係数での評価により決定した．室内画像への適用実験を行った結果，100 枚程度の少ない学習データでも，オリジナルの U-Net と提案のいずれについても床領域の抽出結果は良好であった．提案手法により，精度の改善はわずかであるが，学習の収束については安定性がみられた．

Hyperparameter Selection in Semantic Segmentation and its Application to the Floor Extraction from Indoor Image

Ryoka Oishi[1]　Kyoko Sudo[1,a)]　Fumiaki Sato[1,b)]

Abstract: Recognizing the area of the floor with no obstacle at indoor scenes like corridors or rooms is required for the application of self-driving of wheel chairs. This work aims to recognize the area of an indoor image into a floor, doors, persons, and obstacles. In recent years, several models of semantic segmentation for recognition of roads or indoor images that achieve high performance with deep learning approach are presented, however, they need large training dataset to adapt the arbitrary environment that requires large cost for annotation. In this paper, we indicate the optimization method of the hyper-parameters of U-Net, one of the semantic segmentation models, by genetic algorithms (GA). Our experiment shows that the stability of convergence in training phase is improved with the optimized hyper-parameters, though the accuracy of segmentation is not largely improved.

Keywords: deep learning, U-Net, genetic algorithms, semantic segmentation

1. はじめに

画像からオブジェクトの属性と領域を認識するセマンティックセグメンテーションの技術は，様々な認識アプリケーションへの幅広い応用を目指して研究されている．主な応用例として，自動車の自動走行のための屋外シーンからの道路や車両や人，障害物等の認識が知られているが，これと同様に，車椅子の自動走行を目的として，室内シーンから床領域や障害物等を認識することが求められている．本研究は，このような目的のもとに，室内シーンのカメラ入力画像から，床・ドア・人・障害物などのカテゴリ別に認識し，特に障害物がない床の領域を精度良く認識す

[1] 東邦大学
Faculty of Science, Toho University Miyama 2–2–1, Funabashi-shi, Chiba, 274–8510 Japan
[a)] kyoko.sudo@sci.toho-u.ac.jp
[b)] fsato@is.toho-u.ac.jp

ることを目的としている．

近年，道路や室内の画像を対象に高精度なセマンティック・セグメンテーションを可能とする各種の深層学習のモデルが示されている．こうした手法をユーザの所望の環境に適合させるには，一般に大量の学習データが必要である．しかし，画像ごとにピクセル単位でラベルを与えるアノテーションのコストは大きく，少ない学習データでも性能を上げる技術が望まれる．そこで本研究では，セマンティックセグメンテーションモデルの改良とハイパーパラメータの効率的な決定により性能向上が可能かどうかを検討する．

深層学習におけるセマンティックセグメンテーションのモデルでは FCN(Fully Convolutional Networks)[1],SegNet[2], PSPNet[3], U-Net[4] などの有効性が知られているが，こうした入出力をともに画像とするエンコーダ・デコーダ型のモデルは，計算コストやメモリ使用量が高い．本研究では，リアルタイム認識への応用を想定しているため，よりテスト時の計算コストが低いモデルを選択する必要がある．そこで，比較的新しいモデルである PSPNet と U-Net のテスト時の計算時間を計測する予備実験を行い，より計算時間が抑えられたことと，室内画像の床の抽出というタスクにおいて目視レベルでやや優る傾向がみられた U-Net を用いることとした．ネットワークの学習をユーザの所望の環境に適合させるには，大量の学習データが必要であるが，セマンティクセグメンテーションの学習データの教師情報であるピクセル単位のラベルづけ（アノテーション）を行うには，コストを要するため，学習データが十分に得られないことも多い．そこで，そのような状況のもとで，汎化性能に関わるといわれているネットワークの構成上の変更や，それに関するハイパーパラメータの決定がどの程度性能に影響するかの検討を行った．まず，セグメンテーションの精度向上を期待し，U-Net に次のような軽微な変更を加えた．(1) ダウンサンプリング側の中間層からアップサンプリング側の中間層への結合の際に Dropout[5] を導入する．(2)U-Net のダウンサンプリング時の Pooling において，Maxpooling または平均プーリングを用いる．これらの変更に関わるハイパーパラメータは，従来からの知見がないため，最適な値を効率よく決定する必要がある．

ニューラルネットワークのハイパーパラメータを自動決定する方法として，近年，遺伝的アルゴリズム (以下，GA と呼ぶ) を用いた試みが多く行われており，有効性が知られている．[6] 自動決定の対象とされているハイパーパラメータは，例えばフィルタ数やカーネルサイズなどである．本研究では，U-Net の改良に関わる 1.Dropout の比率及び 2. プーリングの種類の混合比率をハイパーパラメータとして GA により決定する．

2. 提案手法

2.1 U-Net の概要と改良点

U-Net は一般に入力，出力ともに画像であるエンコーダ・デコーダ型のニューラルネットワークの一種である．まず入力画像に対して CNN(Convolutional Neural Network) と画像サイズを縮小する Pooling を複数回施し，入力画像に対して画像サイズの小さい特徴マップを得る．その後特徴マップに画像サイズを拡大する upsampling と CNN を複数回施すことによって入力画像と同じ画像サイズにして出力する．このとき，Pooling によって位置不変性を得るが入力画像と同じ解像度にする際，領域判別の境界線を表現することが難しくなってしまう．そこで，Pooling 時のサイズの小さい画像を upsampling 時にチャンネル単位で結合することでそれを回避している．提案手法では，更に，ダウンサンプリング時の Pooling 層の 1 つ前の情報を保存しておき，その情報をアップサンプリング時の中間層に結合する際に Dropout を行い，チャンネル単位で接続する．このネットワークの構造を図 1 に示す．

Dropout とは，ニューラルネットワークの中間層の一部に対し適用され，学習時に一定確率 $p(0 \leq p < 1)$ でニューロンを選択し順伝搬及び逆伝搬時のニューロンの値を 0 にする手法である．また，選ばなかったニューロンに対しては $\frac{1}{1-p}$ 倍を行う．これはテスト時に期待値をとるためであり，期待値の値を $E[out] = (1-p) \times (\frac{1}{1-p} \times x) + p \times 0 = x$ とするためである．CNN に適用すると中間層の特徴マップに欠落処理を行うことになり，汎化に寄与すると考えられている．そこでこの Dropout を U-Net のダウンサンプリング側からアップサンプリング側に送る画像に加え，Dropout の比率をハイパーパラメータとして GA による推定の対象とする．また，中間層にあるプーリング層の手法の選択を GA により最適化を図る．MaxPooling と AveragePooling をある割合で合わせ用いるものとし，その混合比率をハイパーパラメータとして GA による推定の対象とする．これを混合比率パラメータと呼ぶことにする．さらに，本手法では学習の最適化時間削減のために U-Net の中間層の CNN と活性化関数 ReLU の間に Batch Normalization[7] を追加する．Batch Normalization とは主に中間層の出力に対して行われる処理であり，(1) 中間層の出力をバッチごとに平均を 0 に，標準偏差を 1 に正規化を行う．(2) 正規化されたデータに対し改めて平均 β，標準偏差 γ になるように処理する．β, γ は Batch Normalization のパラメータであり，それぞれ初期値 0, 1 として誤差逆伝播法により学習を行う．これにより，例えば中間層の標準偏差が大きな値である場合 (1) により正規化がされているため，次の層での入力において絶対値の大きな値になることが抑えられ，誤差逆伝播時に絶対値の大きな値によってコンピュータが

Vol.2018-CG-172 No.4
Vol.2018-DCC-20 No.4
Vol.2018-CVIM-214 No.4
2018/11/7

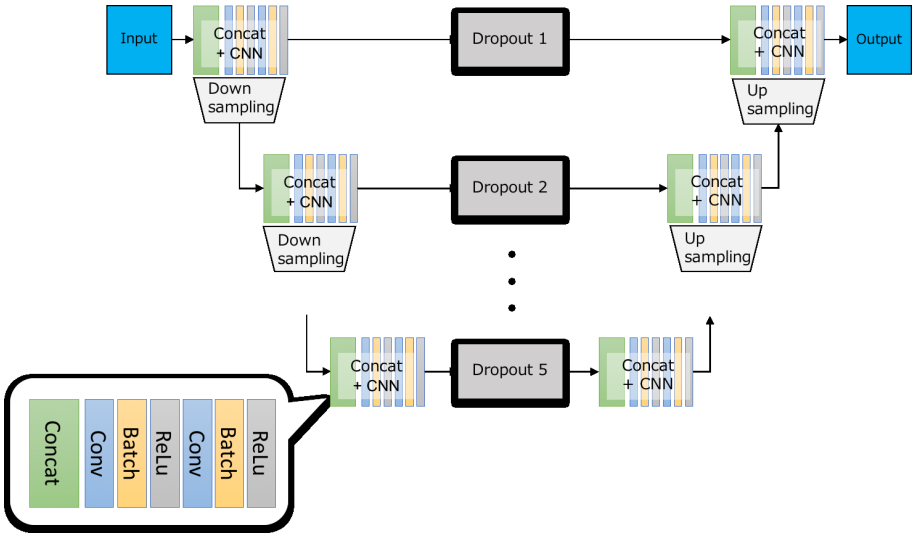


図 1 U-Net の構造．本研究ではダウンサンプリング時の Pooling 層の 1 つ前の情報を保存しておき，その情報をアップサンプリング時の中間層に結合 (concatinate) する際に Dropout を行い，チャンネル単位で接続する．

勾配を計算できなくなる勾配爆発問題を抑えられると考えられる．また，次の層の入力を (2) により平均 β，標準偏差 γ にすることで，中間層の表現力が上がることが分かっている．本手法では従来の U-Net に Batch Normalization を追加したネットワークを U-Net with BN と，従来の U-Net に Batch Normalization と Dropout と混合比率パラメータを追加したネットワークを U-Net with BN drop と呼ぶことにする．

2.2 データセット

2 種類のデータセットを用意する．1 つは公開データセットである ADE20K Data の training set の corridor カテゴリを使用する．(以下 ADE20K Corridor Dataset と呼ぶことにする．)ADE20K Corridor Dataset は室内の RGB 画像及び，室内画像に対して各ピクセルごとに室内画像のカテゴリのインデックスが格納されている白黒画像（以下，ラベル画像と呼ぶ．）のペアが 109 組存在する．このデータセットのカテゴリ数は壁，床，ドアなどの 13 種類である．また，ADE20K Corridor Dataset 全体に対し 99 組を学習データとし，10 組をテストデータとする．もう 1 つは，我々が大学内の複数の棟の複数の階で撮影した 98 枚の廊下の画像である．(以下，室内データセットと呼ぶ．)この撮影画像については，手動のアノテーションを行い，ADE20K Dataset と同様のラベル画像を生成し，元の画像とラベル画像とのペアとした．このデータセットのカテゴリ数は 7 種類であり，屋内データセット全体に対し 88 組を学習データ，10 組をテストデータとする．ADE20K Corridor Dataset と室内データセットの学習に用いた画像とラベル画像の例を図 2（オリジナルデータセット）及び図 3 （ADE20K データセット）に示す．

2.3 ハイパーパラメータの推定

中間層の Dropout の確率をまとめた集合 $P = (p_1, p_2, p_3, p_4, p_5)$ を GA を用いてデータセットに最も適した P を推定する．また，中間層にある Pooling 層について MaxPooling を取るか AveragePooling を取るかを同時に決定する．$\alpha \in [0, 1]$ とし Pooling 層の入力を x とすると，Pooling 層は

$Pooling(x, \alpha) = (1 - \alpha) \times MaxPooling(x) + \alpha \times AveragePooling(x)$

で表される．例えば，$\alpha = 0.25$ の場合 MaxPooling を 75%，AveragePooling を 25%となる．今回では全ての Pooling 層において同じ α を用いる．Pooling 層を図で示すと図 4 のようになる．

これらを合わせた $G = (p_1, p_2, p_3, p_4, p_5, \alpha)$ を GA の遺伝子として GA を訓練する．このアルゴリズムを図 5 に示す．

GA の各個体に用いた評価は Dice 係数を採用する．Dice 係数の式は，以下のようになる．

$$Dice(X, Y) = \frac{2 \times |X \cap Y|}{|X| + |Y|}$$

本手法では入出力がともに画像であることから，以下の式を Dice 係数とした．

$$Dice(target, predict) = \frac{2 \times sum(target \odot predict)}{sum(target) + sum(predict) + 10^{-8}}$$

target を正解画像，predict を U-Net での出力画像とする．正解画像は one-hot ラベルであり，画像にクラスのインデックスが格納されている．クラスの数だけチャンネル数

Vol.2018-CG-172 No.4
Vol.2018-DCC-20 No.4
Vol.2018-CVIM-214 No.4
2018/11/7

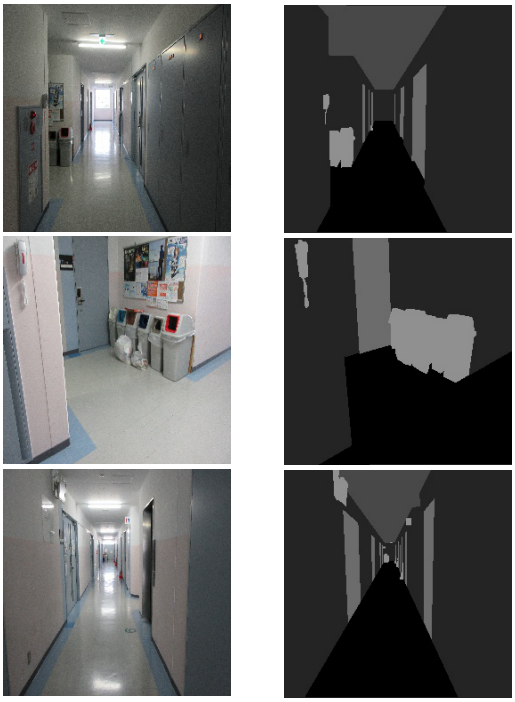


図 2 オリジナルデータセットから，学習に用いた画像とラベル画像の例．左側が入力画像，右側がラベル画像（ラベル画像は可視化のために定数倍した）

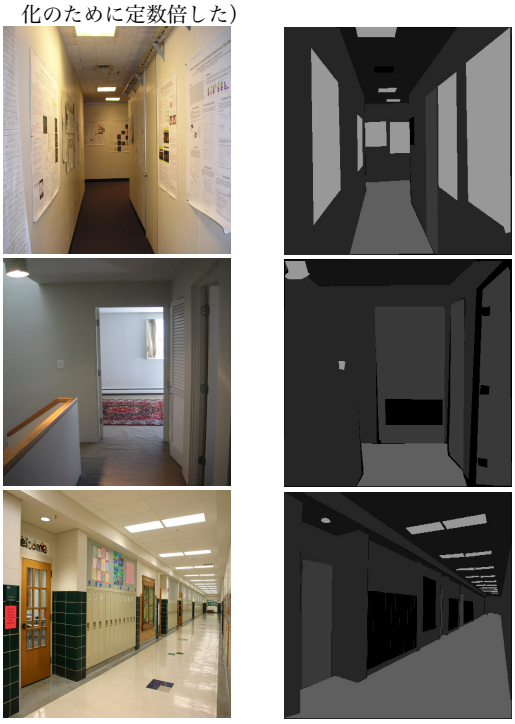


図 3 ADE20K Corridor dataset から，学習に用いた画像とラベル画像の例：左側が入力画像，右側がラベル画像（ラベル画像は可視化のために定数倍した）

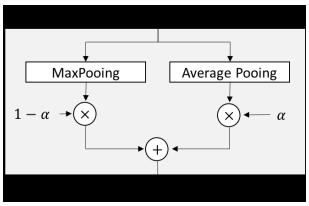


図 4 提案手法におけるプーリング層の構成．乗算ノードでは各テンソルに定数倍，加算ノードではテンソル同士の和をとるものとし，混合比率 α を GA による推定対象とする．

をもつ，値 0 で埋められた画像配列を作成し，各ピクセル値に対応する番号を 0 から始まるインデックスとみなし，インデックスに対応するチャンネルにおける対応するピクセル値を 1 にする．以下，one-hot 画像と呼ぶことにする．また，ラベル画像におけるインデックスの最大値から 1 引いた値をクラス数と呼ぶことにする．$\odot$ はアダマール積で各要素の対応する成分同士の積である．sum はすべての成分の総和をとり，結果はスカラーである．また 10^{-8} はゼロ除算防止用の数値である．target 画像が one-hot 画像の場合，分母は必ず 0 より大きくなる．ここでは分母に 10^{-8} を加えているが，それによる大小関係への影響は少ない．Dice 係数の性質より，target 画像と predict 画像が完全一致するとき Dice 係数はほぼ 1 をとり，target 画像と predict 画像がほとんど一致していなければ 0 をとることがわかる．また，Dice 係数の値は 0 以上 1 以下となる．

U-Net の最適化に用いる関数は以下のような式とする．

$$Loss(target, predict) = 1 - Dice(target, predict)$$

GA の最適化のアルゴリズムを図 6 に示す．gene を GA のある個体とする．各 Epoch ごとにすべてのテストデータに対し Dice 係数を計算し，各テストデータにおける Dice 係数の平均値を計算する．十分な Epoch 数学習したのち Dice 係数の平均値の最大値を各個体の評価値とする．その評価値が大きくなるような個体を決定するために GA を最適化する．

2.4 セマンティックセグメンテーション

推定ステップによって得られた $\tilde{G}$ を基にしたネットワークで室内データセットで学習する．学習は 2500Epoch 行い，前節と同様 Dice 係数を最適化する．また，出力画像は one-hot 画像であるため，可視化の際は図 7 のようにラベル画像に変換して出力することとする．

ハイパーパラメータ推定の学習を室内データセット，セグメンテーションの学習を室内データセット及び ADE20K Dataset[8][9] の Corridor カテゴリ（以下，ADE20K Corridor Dataset と呼ぶ) で実験を行った．両データとも室内の画像とそれに対応する壁，床，ドアなどでクラス分けが行われている画像 (以下，ラベル画像と呼ぶ) がある．室内の画像と画像に対するラベルのインデックスがピクセル

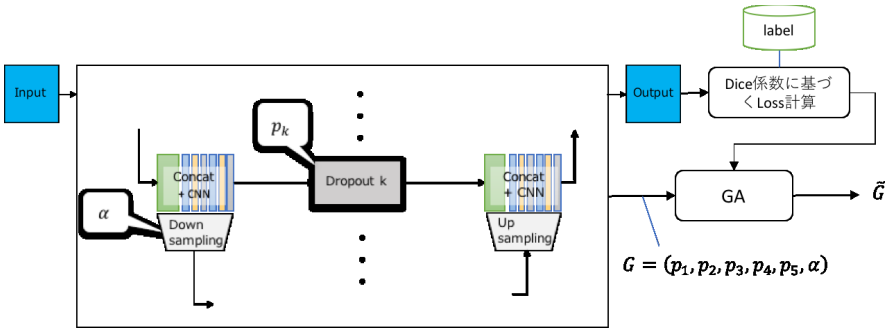


図 5 GA で学習するパラメータと学習の流れ.

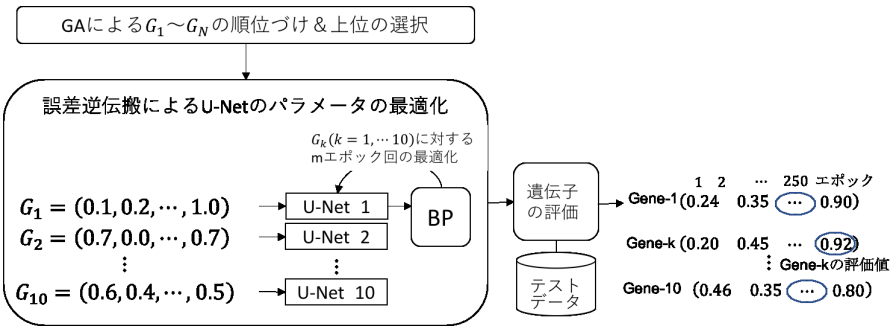


図 6 GA の最適化のアルゴリズム.

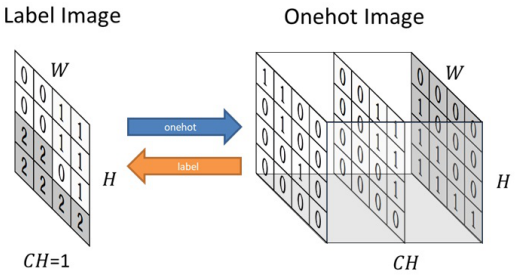


図 7 3 クラスの領域分割の例，ラベル画像には各画素に対応したチャンネルのインデックス，one-hot 画像には各画素をチャンネル方向に考えたとき，一つのチャンネルの画素値のみ 1 であり，それ以外の画素値は 0 とする.

ごとに保存されているラベル画像のペアで与えられている．クラス数はオリジナルデータセットでは 7，ADE20K Corridor Dataset のクラス数は 13 である.

3. 評価実験

3.1 ハイパーパラメータ推定の評価実験

ハイパーパラメータ推定を室内データセットを用いて学習を行う．データセットの各組に対して，画像サイズをバイニリア補完で 286 × 286 へリサイズ，Random Crop により 256 × 256 にする．また，50%の確率で左右反転，室内画像にのみ輝度変化のデータ拡張 (Data Augmentation) を行う．従って，室内データセットの入力サイズとチャンネルは 256 × 256 × 3 であり，出力時のカテゴリ数が 7 種類のため，256 × 256 × 7 となる．U-Net の最適化は誤差逆伝播法によって得られる勾配情報を Adam(Adaptive moment estimation)[10] を用いた.

学習には各世代 10 個体，各個体で 250Epoch ずつ学習を行い各 Epoch ごとに Dice 係数を測定する．Epoch ごとに測定した Dice 係数の最大値を個体の評価値とする．次の世代には評価値の高い上位 4 個体を選択，残り 6 個体を一様交叉法で選択を行い，10 世代まで行った．全世代における最優秀個体の Dice 係数の変化は以下の図のようになった．最優秀個体 $\tilde{G} = (\tilde{p_1}, \tilde{p_2}, \tilde{p_3}, \tilde{p_4}, \tilde{p_5}, \tilde{\alpha})$ を元にしたハイパーパラメータは以下のような結果を得られた.

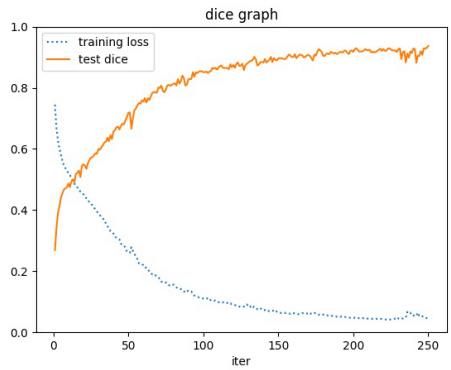


図 8 推定ステップに室内データセットを用いた際の学習データ Loss 及び，テストデータに対する Dice 係数の変化

$\tilde{G} = (0.45, 0.00, 0.25, 0.5, 0.00, 1.0)$

3.2 セグメンテーションの評価実験

セグメンテーションを室内データセット及び ADE20K

Vol.2018-CG-172 No.4
Vol.2018-DCC-20 No.4
Vol.2018-CVIM-214 No.4
2018/11/7

Dataset を用いて評価実験を行う．Dropout 層を用いない U-Net との比較を同データセットで行った際の Dice 係数の関係を表 1 で示す．

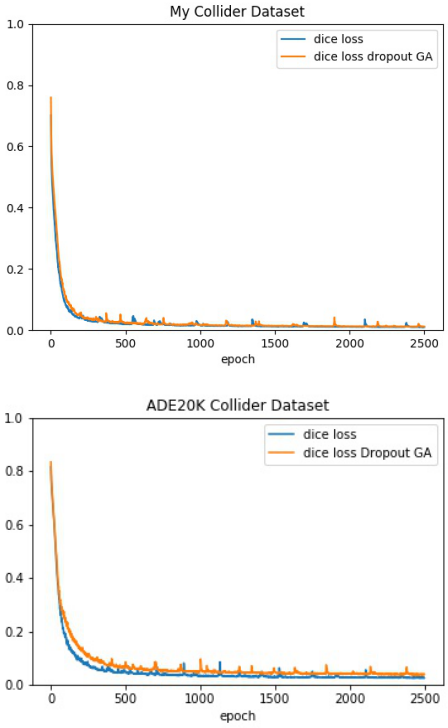


図 9 室内データセット (上)，ADE20K Corridor Dataset(下) での学習データにおける loss の変化

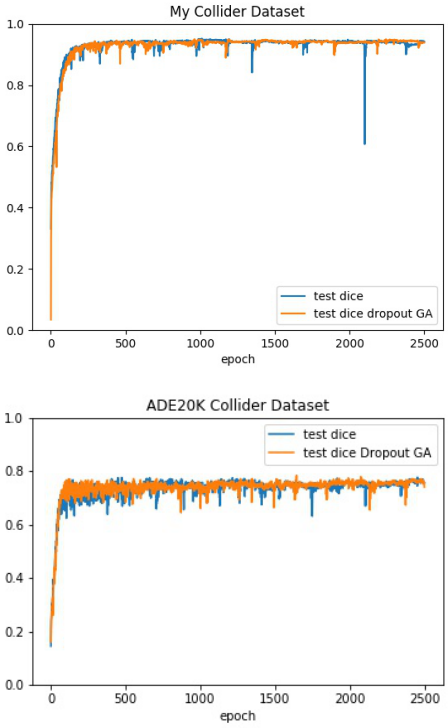


図 10 室内データセット (上)，ADE20K Corridor Dataset(下) におけるテストデータでの Dice 係数の変化

表 1 オリジナルの U-Net と Dropout を入れた場合とでの Dice 係数の比較

データセット	ネットワーク構成	Dice 係数
室内データセット	U-Net with BN	0.951
	U-Net with BN drop	0.953
ADE20K Corridor Dataset	U-Net with BN	0.776
	U-Net with BN drop	0.783

4. 考察

GA により推定した $\tilde{G} = (0.45, 0.00, 0.25, 0.5, 0.00, 1.0)$ を用いて学習した場合，Dropout 率は 0.00 から 0.5 の間を選択する傾向がみられた．Dropout 率を高めると特徴マップの欠落の割合が上がるため，0.5 を超えるような数値が GA によって選択された個体は，テストデータにおける Dice 係数の向上はあまり見られない傾向が確認された．

Dropout 率に関して，推定したパラメータを用いた場合とそうでない場合の学習の収束については，Dropout を導入したことで，従来の U-Net と比較して学習中の loss の減少は提案手法のほうが収束が遅くなった．

テストデータに対しては学習中の Dice 係数の変化が少ないことから，従来の U-Net と比較して提案手法のほうが安定して学習できていることが考えられる．

混合比率パラメータついては，世代が進むごとに混合比率パラメータが 1.0 に近い値をとる個体がほとんどになり，最終的に GA で選択された個体では 1.0 を出力していた．これは U-Net におけるセマンティックセグメンテーションタスクにおいて最大プーリングと比較して平均プーリングが良い精度を得られると考えられる．これは，セマンティックセグメンテーションでは入力画像の形状を維持すると同時に特徴量を抽出すべきと考えられているためである．従って，一部の特徴のみが次層の入力に伝搬する最大プーリングと比較して，全ての特徴が次層の入力に伝搬する平均プーリングのほうが優れていると考えられる．

また，既存の U-Net よりテストデータに対して安定した学習ができることが示された．

5. まとめ

セマンティックセグメンテーションの汎化性能の向上を期待し，従来の知見を参考にした軽微な改良を U-Net に施し，それに関わるハイパーパラメータを学習前に GA によって推定した．オリジナルのデータと公開データの 2 種類の室内画像データセットに対して適用し，改良の有無，ハイパーパラメータの推定の有無について，学習の収束性とセグメンテーションの精度を評価した結果，テストデータにおける Dice 係数の収束については図 9 のグラフより安定性がみられた．未知の入力データに対する精度は，大きな差はみられなかったが，いずれも提案手法がわずかに上回った．

Vol.2018-CG-172 No.4
Vol.2018-DCC-20 No.4
Vol.2018-CVIM-214 No.4
2018/11/7

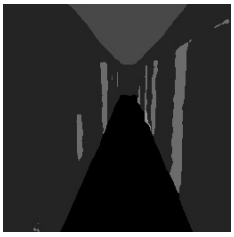
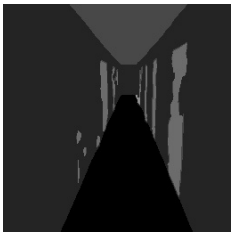
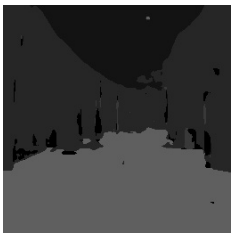
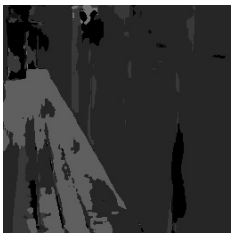
image	original	ground truth	base line	GA
オリジナルデータ (a)				
オリジナルデータ (b)				
オリジナルデータ (c)				
ADE20K(a)				
ADE20K(b)				
ADE20K(c)				

図 11 オリジナルデータセット及び ADE20K データセットのテスト画像に対する処理結果例. 左から，テスト画像，正解のアノテーション，U-Net with BN によるセグメンテーション結果，提案手法（U-Net with BN drop) によるセグメンテーション結果.

参考文献

[1] Jonathan Long, Evan Shelhamer and Trevor Darrell, "Fully Convolutional Networks for Semantic Segmentation", arX iv, 2014
[2] Vijay Badrinarayanan, et al., "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation", arXiv, pp.1-14, 2015
[3] Hengshuang Zhao, et al., "Pyramid Scene Parsing Network", pp.1-11, 2016
[4] Olaf Ronneberger, Philipp Fischer and Thomas Brox, "U-Net: Convolutinal Networks for Biomedical Image Segmentation,", arXiv, pp.1-8, 2015
[5] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. The Journal of Machine Learning Research, Volume 15 Issue 1, January 2014 Pages 1929-1958
[6] 藤野　紗耶, 森　直樹, 松本　啓之亮, "3 分岐畳み込みニューラルネットワークによる 4 コマ漫画の順序識別 Recoginizing the Order of Four-sence Comics bt Three-Path Convolutional Neural Networks", 2018 年度人工知能学会全国大会（第 32 回）, 2018
[7] Sergey Ioffe and Christian Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,", arXiv pp.1-11, 2015
[8] Scene Parsing through ADE20K Dataset. Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso and Antonio Torralba. Computer Vision and Pattern Recognition (CVPR), 2017.
[9] Semantic Understanding of Scenes through ADE20K Dataset. Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso and Antonio Torralba. arXiv:1608.05442.
[10] D.P Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv, pp.1-15, 2014