Multiple Dilated Convolutional Blocksによるセマンティクセグメンテーション

山下隆義 † 山内悠司 †　藤吉弘亘 †
† 中部大学
E-mail: yamashita@cs.chubu.ac.jp

Abstract

セマンティックセグメンテーションは，ピクセル単位でクラスを推定する問題であり，ディープラーニングをベースとした高精度な手法が提案されている．車載カメラ映像のセマンティックセグメンテーションの場合，歩行者や車両などの物体はカメラまでの距離により大きさが変動する．本稿では，このような変動に対応する方法として，1)Multiple Dilated Convolution Blocksにより様々な物体の大きさを考慮した畳み込み処理，2)スキップ結合により階層を跨いだ特徴伝搬，をエンコーダ・デコーダ構成のネットワークに導入する．これにより，Cityscapesデータセットにおいて，一般的なセグメンテーション手法よりも高い精度を得ることができた．また，最新の手法と比較して，同程度の平均カテゴリIoU精度を達成することができた．

1 はじめに

畳み込みニューラルネットワーク [1] は物体認識問題において，非常に高い認識精度を達成している [2][8][16]．また，畳み込みニューラルネットワークは物体認識だけでなく，物体検出 [5] やセマンティックセグメンテーション [9][10][11][12][13] にも応用されている．セマンティックセグメンテーションは，ピクセル単位でクラスを推定する問題である．畳み込みニューラルネットワークを利用した方法として，Fully Convolutional Neural Network (FCN)[9] やエンコーダ・デコーダ構成のSegNet[11][12] など，様々な手法が提案されている [18][20][21]．畳み込みニューラルネットワーク以前にも，色やエッジ情報などのあらかじめ定義した特徴量を利用してクラスタリングを行い，領域を連結させるボトムアップな手法が提案されている [3]．畳み込みニューラルネットワークを利用した方法は，画像全体を入力すると各クラスの確率マップを出力するための特徴量を学習により獲得することができる．

セマンティックセグメンテーションは，自動運転支援に向けた走行可能領域の抽出や，歩行者や車両などの検出に応用可能である．また，ロボットの自律走行に向けて屋内のテーブルや椅子などの配置を理解することにも応用できる．これらの応用シーンには，様々な大きさの物体が存在している．また，同一クラスの物体でも位置により大きさや見えが異なる．このように，セメンティックセグメンテーションは，大きさや見えが異なる様々な物体の領域を抽出することが求められている．我々は，様々な物体の大きさに対応するために，Multiple Dilated Convolution Blocks (MDC Blocks)を提案する．Dilated Convolutionは，畳み込み処理において，一定間隔離れた要素を畳み込む [15]．そして，間隔が異なるDilated Convolutionを並列に複数行うことで，様々な大きさの物体情報を同時に抽出することができる．また，見えの変化に対応するために，一般物体認識で高い精度を達成しているResidual Network[16]のSkip Connectionの構造を導入する．これにより，詳細なセグメンテーションが可能となる．

2 関連研究

ディープラーニングによる物体認識は，ImageNet Large Scale Visual Recognition Challenge (ILSVRC)を通じて，高精度なネットワーク構造が提案されている．5層の畳み込み層，3層の全結合層の8層構造をしているAlexNetは，従来の物体認識手法よりも高精度な手法であり，ディープラーニングが注目される先駆けとなったネットワークである [2]．AlexNetは，活性化関数にRectified Linear Unit (ReLU)，汎化性能を向上させる工夫としてDropoutを用いている．また，GPUでの学習を行うことで，深いネットワーク構造を現実的な時間で学習することを可能としている．VGG16は13層の畳み込み層，3層の全結合層の16層から構成されるより深いネットワーク構造である [8]．このネットワーク構造では，各畳み込み層のフィルタサイズを3×3とし，2層または3層の畳み込み層後にプーリング層を配置している．フィルタサイズを小さくすることでAlexNetよりも深い構造にも関わらずパラメータ数を減らすことができている．GoogleNetは，フィルタサイズが異なる畳み込み処理を並列で行うInception module

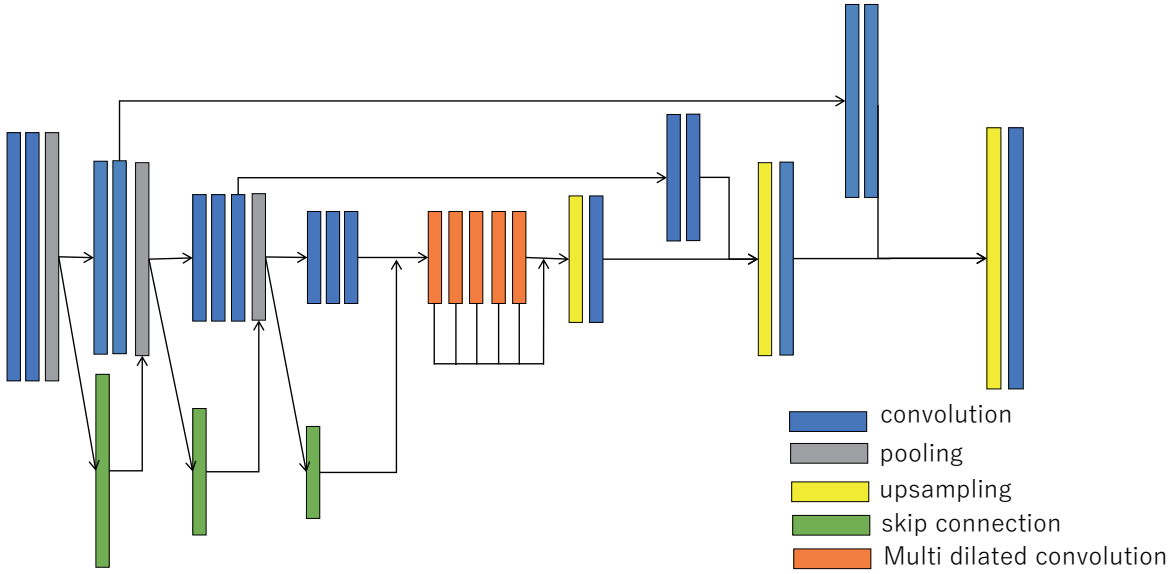

図 1　提案手法のネットワーク構造

を 9 層積層した，22 層構造となっている [6]．Inception module では，1×1，3×3 または 5×5 のフィルタを畳み込んで得られた特徴マップを連結することで，着目領域の異なる特徴を同時に捉えることができる．Residual Network は，152 層と非常に深いネットワーク構造となっている [16]．ネットワーク構造を深くした場合，誤差の消失及び爆発問題があり，精度が向上しない問題があった．Residual Network では，複数の層をまたいでバイパスを通すようなスキップ結合を導入することで，誤差を入力層近くまで逆伝搬できるようにしている．また，推論時は入力層近くの情報をスキップ結合により上位へ順伝搬できる．

セマンティックセグメンテーションは，これらのネットワーク構造をベースにしてピクセルごとのクラスラベルを推定するネットワークを end-to-end で学習する．Fully Convolutional Network (FCN) は，ImageNet dataset を用いて物体認識用に学習されたネットワークを初期値として利用し，物体認識を対象としたのネットワークをセマンティックセグメンテーション用に転移している [9]．その際，異なる入力データサイズに対応するために，全結合層を 1×1 のフィルタで構成される畳み込み層に置き換えている．また，画像全体のグローバルな情報とクラスごとのローカルな情報を捉えるために，スキップ構造を採用している．ネットワーク構造の入力層に近い層は画像の細かな情報を捉えている．一方，出力層に近づくとプーリング処理を繰り返すことで特徴マップは小さくなり画像全体の特徴を捉えている．クラスラベルを推定する際に，中間層の情報を合わせて利用するスキップ構造にすることで，クラスごとの細かな領域までセグメンテーションすることを可能としている．Segnet はエンコーダ・デコーダのネットワーク構造をしている [11]．エンコーダは，VGG16 の畳み込み層部分と同等の構造をしており，入力画像から特徴を抽出する．デコーダはエンコーダと対となる構造をしており，アップサンプリングと畳み込み処理から構成されるデコンボリューション層が畳み込み層の代わりに用いられる．また，対となるプーリング層で選択された位置を記憶しておき，アップサンプリング時は記憶している位置に値を値を代入し，それ以外の位置は 0 を代入する．これにより，デコードされた特徴から詳細なクラスラベルを推定することができる．FCN や SegNet で得られた各クラスの確率マップを入力してより詳細なセグメンテーションを行う後処理に CRF-RNN がある [10]．隣接画素間の各クラスの確率分布を考慮して局所的なセグメンテーションの誤り訂正を繰り返し行うことで，高精度化を図っている．CRF-RNN は，各クラスの確率マップを出力するネットワークと合わせて end-to-end で学習することができる．CRF-RNN は隣接画素に着目しているが，大局的な情報を利用する方法として Dilated Convolution がある [15]．Dilated Convolution は畳み込み処理を行う位置を一定間隔離すことで，広範囲の領域を考慮した畳み込みが可能となる．

3　提案手法

我々は，物体の詳細までセグメンテーションすることが可能なネットワーク構造を提案する．提案するネットワーク構造を図 1 に示す．ベースとなるネットワーク構造は，エンコーダ･デコーダ構成となっている．エンコーダ時に局所的な情報が欠落しないように，スキップ構造を導入する．さらに，物体の詳細情報がデコーダ側に伝

表 1　各層のフィルタ構成

層	フィルタサイズ	フィルタ数	活性化関数	プーリング
1 層目	3 × 3	32	ReLU	–
2 層目	3 × 3	32	ReLU	max pooling
3 層目	3 × 3	64	ReLU	–
4 層目	3 × 3	64	ReLU	max pooling
5 層目	3 × 3	128	ReLU	–
6 層目	3 × 3	128	ReLU	–
7 層目	3 × 3	128	ReLU	max pooling
8 層目	3 × 3	256	ReLU	–
9 層目	3 × 3	256	ReLU	–
10 層目	3 × 3, s=1	256	ReLU	–
11 層目	3 × 3, s=2	512	ReLU	–
12 層目	3 × 3, s=4	512	ReLU	–
13 層目	3 × 3, s=8	512	ReLU	–
14 層目	3 × 3, s=16	512	ReLU	–
15 層目	3 × 3, s=32	512	ReLU	–
16 層目	3 × 3	256	ReLU	upsampling
17 層目	3 × 3	128	ReLU	upsampling
18 層目	3 × 3	クラス数	ReLU	upsampling

搬されるように，エンコーダ側の特徴マップを対となるデコーダ側の層に連結する．この時，1 × 1 の畳み込み処理を行った特徴マップを連結する．また，大局的な情報を考慮するために，エンコーダ側とデコーダ側の間に Multiple Dilated Convolution Block (MDC Block) を追加する．MDC Block は，複数の Dilated Convolution 層から構成されている．また，各畳み込み層は，Batch Normalization[14] を行なった後で畳み込み処理を行う．Batch Normalization は，ミニバッチ学習においてバッチ間のデータのばらつきを抑え，学習の収束性を早めるとともに，明るさなどの変動に頑健となる．以下に，各構成の詳細を述べる．

3.1　エンコーダ・デコーダ構成

図 1 に示すようにエンコーダ側は 10 層の畳み込み層，デコーダ側は 3 層のデコンボリューション層から構成されている．各層のフィルタサイズおよびフィルタ数を表 1 に示す．各層のフィルタ数は 3 × 3 であり，フィルタ数はプーリング処理を行うごとに 2 倍している．プーリングは 2 × 2 の max pooling である．これらのエンコード側の処理は VGG16 の conv4_3 までと同様の構造である．デコード側はプーリングを行なった回数分，デコンボリューション行う．各デコンボリューションでは，特徴マップを 2 倍にアップサンプリングして畳み込み処理を行う．畳み込むフィルタサイズは 3 × 3 である．Segnet では，各デコンボリューション時にエンコード側と同等数の畳み込み層があるが，本ネットワーク構造では 1 層としている．エンコーダ側とデコーダ側の各層の活性化関数には ReLU を用いている．

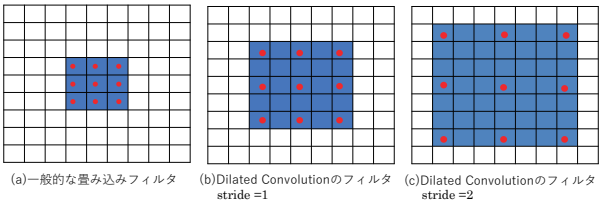

図 2　従来の畳み込み処理と Dilated Convolution 処理の違い

3.2　Multiple Dilated Convolution Block

エンコーダとデコーダの間に大局的な特徴を捉える Multiple Dilated Convolution Blocks (MDC blocks) を配置する．Dilated Convolution は，図 2 のように，畳み込む要素を一定間隔で離す畳み込み処理である．3×3 のフィルタを畳み込む際，一般的な畳み込み処理は図 2(a) のように 3 × 3 の領域に対して，入力の値とフィルタの値を要素ごとに乗算して合計値を求める．このように，局所的に密な結合となっている．一方，Dilated Convolution は，図 2(b) のように，stride を 1 とした場合，5 × 5 の領域に対して，フィルタの値を要素ごとに乗算する．Dilated Convolution は，一般的な畳み込み処理よりも広い範囲に対して疎な結合となっている．stride を 2 にした場合，図 2(c) のように 7 × 7 の領域に対して畳み込み処理を行う．Dilated Convolution はフィルタサイズは従来の畳み込み処理と変わらないものの，より広範囲に対して疎に畳み込むため，大局的な情報を捉えることが可能となる．MDC Blocks は，間隔を変えた 5 つの Dilated Convolution 層を積層している．このように積層することで，従来の畳み込み層よりも広範囲の情報を捉えることが可能となる．

3.3　スキップ結合

Residual Network では，スキップ結合を導入することで，深いネットワーク構造でも誤差を入力層近くまで伝搬できるようになり，高精度な物体認識を実現している．また，FCN では，中間層の特徴マップを利用して高解像度なセグメンテーション結果を得ている．このアイデアも一種のスキップ結合とみなすことができる．本ネットワークでは，Residual Network のスキップ結合をエンコーダ側に，FCN のスキップ結合を対となるコンコーダとデコーダの間に連結させる構造とする．エンコーダ側のスキップ結合では，特徴マップの各要素の値を加算する．その際，特徴マップのチャネル数が異なる場合，上位層のチャネル数と同じだけ 1 × 1 のフィルタサイズの畳み込みを行い，チャネル数を合わせる処理を行う．すなわち，2 層目で得られた 32 チャンネルの特徴マップを 4 層目の畳み込み層で得られた 64 チャンネルの特徴マップに加算する場合，2 層目で得られた特徴マップに対して 64 枚の 1 × 1 の畳み込みを行って

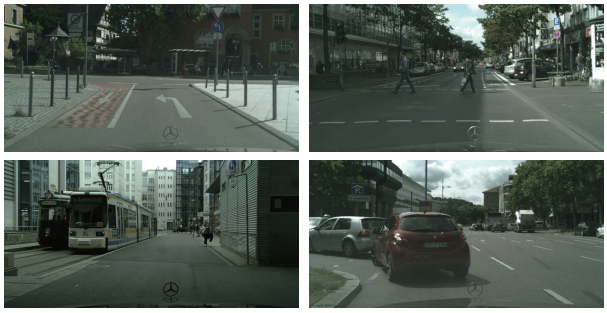

図 3　Cityscapes データセットの例

64 チャンネルの特徴マップとする．FCN のスキップ結合は，エンコーダ側の特徴マップをデコーダ側の特徴マップに要素ごとに加算する．Residual Netowrk などのスキップ結合は，特徴マップを連結する方法を用いているが，文献 [19] で連結と加算は等価であることがわかっている．スキップ結合を加算で行うことで，特徴マップ数が増えずにメモリ使用量を抑えることができる．これらのスキップ結合により，物体の詳細な情報が 2 つの経路で伝搬できるようになる．また，MDC blocks の各層の出力をエンコーダ側にスキップ結合させる構造も検討する．これにより，様々な範囲の情報を捉えた特徴マップをデコーダ側に入力することができる．

3.4　ネットワークの学習

本ネットワークは，end-to-end に学習を行う．その際，他手法のように事前学習されたネットワークを用いない．これにより，ネットワーク構造を柔軟に変えることが可能となる．学習時のバッチサイズは 16，学習の最適化方法には Adam[7] を用いる．学習データは，画像全体を入力せずに，一定サイズで切り出した領域を入力する．これにより，シーンの様々な構図を作り出すことができ，学習データのバリエーションを増やすことができる．切り出すサイズは 720×720 とし，切り出し位置はランダムに指定する．また，切り出す領域は，切り出しサイズの 0.75 倍から 1.25 倍の範囲とする．これらの実装は，chainer を用いて行い，学習は NVIDIA DGX-1 で行う．DGX-1 に搭載されている Tesla P100 のメモリサイズは 16GB となっており，1 枚の GPU で処理可能なミニバッチサイズは 2 である．そこで，8 枚の GPU を利用してデータ並列でミニバッチ学習を行う．

4　評価実験

本ネットワークの有効性を確認するために，車載カメラで撮影されたシーンを対象としてセマンティックセグメンテーションの評価実験を行う．評価データには，Cityscapes データセット [17] を用いる．図 3 に Cityscapes の画像例を示す．本データセットは，ヨーロッパの 50 都市で日中の天気の良い日に撮影されており，クラス数は 30 クラスである．そのうち，一部のクラスは頻出頻度が低い．そのため，評価には 19 クラスを用いている．アノテーションは Fine annotations と Coarse annotations がある．Fine annotations は 5000 枚のデータに対して，詳細にアノテーションされており，Coarse annotations は，20000 枚のデータに対して，領域を大まかに囲むようなラフなアノテーションとなっている．本実験では，Fine annotations を用いる．Fine annotations のデータは，学習用に 2975 枚，検証用に 500 枚，評価用に 1525 枚含まれている．評価用の画像に対する annotation データは公開されておらず，WEB サイトに結果を登録することで評価結果を得ることができる．WEB サイトへの登録は 48 時間に 1 回のみ，30 日間で 6 回までと規定が定められている．そのため，本実験では，検証用のデータを用いて比較実験を行う．

4.1　評価方法

セマンティックセグメンテーションの評価は，画素ごとにアノテーションされたクラスと一致しているかどうかを判定して，平均 IoU (Intersection over Union) を算出する [4]．各クラスの IoU は式 (1) のように求める．ここで，TP は正解した画素数，FP は別クラスの画素に対して誤判定した画素数，FN は別クラスとして誤判定した画素数となる．

$$IoU = \frac{TP}{TP + FP + FN} \tag{1}$$

Cityscapes では，評価対象の 19 クラスを 7 つのカテゴリに分類しており，平均カテゴリ IoU についても評価している．一方，IoU は，領域の小さなクラスに対する精度が低下しやすいという問題がある．対象とする車載カメラの映像には，歩行者や車，標識などはカメラからの距離により大きさが変化するクラスが多い．そこで，Cityscapes では，インスタンスレベルでの IoU を算出し，その平均を評価する基準がある．インスタンスレベルでの IoU は，iIoU として，式 (2) のように求める．

$$iIoU = \frac{iTP}{iTP + iFP + iFN} \tag{2}$$

ここで，iTP はインスタンス内における正解画素数，iFP は，インスタンス内における別クラスの画素に対して誤判定した画素数，iFN はインスタンス内において別クラスと誤判定した画素数である．iIoU は，IoU と同様にクラスとカテゴリそれぞれの平均を求めて評価を行う．

4.2　ネットワーク構造の比較

提案手法では，エンコーダ・デコーダ構成のネットワークに MDC Blocks とスキップ結合を導入している．本実験では，これらの構造の有無による精度比較を行う．評価結果を表 2 に示す．これより，スキップ結合

表 2　ネットワーク構成による精度比較

DMC Blocks	スキップ結合	クラス [%] IoU	クラス [%] iIOU	カテゴリ [%] IoU	カテゴリ [%] iIoU
なし	なし	54.9	37.8	83.6	73.2
なし	あり	56.1	40.2	84.3	76.1
あり 直列	なし	67.3	45.8	87.8	74.1
あり 直列	あり	72.5	52.5	89.2	78.2
あり スキップ結合	あり	73.0	55.6	89.2	81.9

表 3　学習サイズによる精度比較

学習サイズ	クラス [%] IoU	クラス [%] iIoU	カテゴリ [%] IoU	カテゴリ [%] iIoU
540 × 540	73.1	55.8	89.1	81.7
640 × 640	72.2	53.8	89.1	79.5
720 × 720	73.0	55.6	89.2	81.9

のみを導入することで，各精度が 2%から 3%程度向上している．5 つの Dilated Convolution 層を直列に積層した MDC Blocks を導入することで，平均クラス IoU は 54.9%から 67.3%，平均クラス iIoU は，37.8%から 45.8%に大きく精度向上している．また，カテゴリ IoU およびカテゴリ iIoU も 5%程度，精度向上している．スキップ結合と MDC Blocks の両方を導入した場合，さらに精度は向上し，平均クラス IoU が 72.5%，平均クラス iIoU が 52.5%，平均カテゴリ IoU は 89.2%，平均カテゴリ iIoU は 78.2%である．これより，これらの 2 つの処理は精度向上に大きく寄与していることがわかる．MDC Blocks に対してもスキップ結合を導入するした場合，クラスおよびカテゴリの iIoU 精度が 3%程度向上している．MDC Blocks は，直列につなげることで上位層では視野を広げて広範囲の特徴を捉える．MDC Blocks にスキップ結合を導入することで，視野の範囲が異なる特徴を同時に捉えることが可能となり，物体の大きさに対して頑健なセグメンテーションができるようになったと言える．

4.3　学習サイズによる精度比較

本ネットワークは学習時，画像を 720×720 に切り出して入力している．Cityscapes の画像サイズは 2048×1024 であり，画像全体を一度に入力して学習できない．そこで学習時の入力画像サイズが精度に影響するかを確認するための精度比較を行う．表 3 に入力画像サイズを変えた場合の精度比較結果を示す．これより，平均クラス IoU および平均クラス iIoU は 540 × 540 の場合が最も精度が良い．一方で，平均カテゴリ IoU と平均カテゴリ iIoU は 720 × 720 が最も精度が良い．画像サイズにより精度の差があるものの，これらの差は学習時の誤差逆伝搬における誤差範囲であると考えられる．よって，入力画像サイズによる大きな性能差はないと言える．

表 4　Cityscapes のテストデータに対する精度

手法	クラス [%] IoU	クラス [%] iIoU	カテゴリ [%] IoU	カテゴリ [%] iIoU
SegModel	78.5	56.1	89.8	75.9
ResNet-38[20]	78.4	59.1	90.9	81.1
Dilation10 [15]	67.1	42.0	86.5	71.1
FCN-8s[9]	65.3	41.7	85.7	70.1
Segnet basic[11]	57.0	32.0	79.1	61.9
提案手法	71.6	49.4	89.3	78.3

4.4　テストデータに対する精度比較

Cityscapes のテストデータセットに対する精度比較を行う．精度比較結果を表 4 に示す．これより，Segnet や FCN のような一般的なセグメンテーション手法と比較して，各評価指標において大幅に精度を向上させることができている．また，Dilated Convolution を用いた手法と比較しても提案手法の方が良い結果を得ることができている．一方，Cityscapes のベンチマーク結果に登録されている上位の手法 (SegModel，ResNet-38) と比較すると，クラス IoU およびクラス iIoU は上位手法の方が優れている．しかしながら，カテゴリ IoU およびカテゴリ iIoU は，これらの手法と比較して同等か上回ることがある．提案手法は，カテゴリレベルでの分類ができていることから，MDC blocks やスキップ結合は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効果があると言える．これらの比較手法の各クラスに対する IoU を図 4 に示す．これより，road や building, sky などのように出現頻度の高いクラスに対する精度は非常に高く，上位の手法と同等の精度となっている．また，person や car のように大きさの変動が大きなクラスに対する精度も上位の手法と同等である．これより，自動運転に向けて重要となる road や person，car などのクラスに対しては高精度なセマンティックセグメンテーションを実現することができている．一方，truck や bus，train, motorcycle は上位の手法と比較すると大きく精度が低下している．この原因は 2 つ考えられる．1 つ目は，学習済みモデルに起因する特徴抽出性能の違いである．上位の手法は，Imagenet のデータセットで学習済みのモデルを利用している．bus や motorcycle などのクラスは認識対象クラスであり，これらのクラスに対応した特徴抽出ができるようになっていることが考えられる．提案手法は，学習済みモデルを利用して

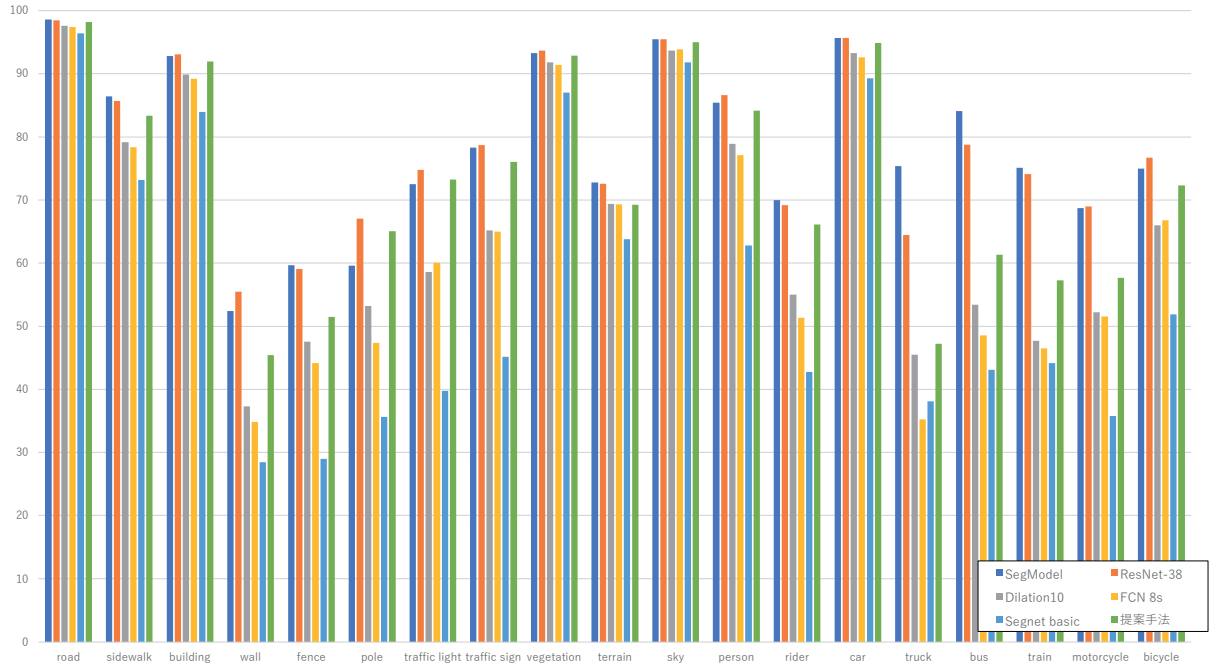

図 4　Cityscapes テストデータセットに対する各クラスの IoU 精度

おらず，Cityscapes データセットのみ利用して学習している．これらのクラスは学習データセットにおいて出現頻度が非常に低い．そのため，出現頻度の低いクラスの学習が不十分である可能性がある．２つ目は，入力画像サイズに起因する大きな物体への精度低下である．本ネットワークは GPU のメモリサイズの関係上，学習時の入力サイズを 640×640 としてる．そのため，画像中に train や bus が大きく写った場合，その一部分のみを学習している．全体を包含した学習データを入力されておらず，精度が低下している可能性がある．

図 5 にテストデータに対するセマンティックセグメンテーション結果を示す．これより，精度が高い road や building，person，car などのクラスに対して詳細までセグメンテーションできていることがわかる．また，撮影シーンや小さな物体のクラスに対しても対応できている．一方，bus や truck は，大まかにはセグメンテーションできているものの，物体中の領域を別のクラスとしてセグメンテーションされている．そのため，これらのクラス IoU 精度が低下している．

5　まとめ

本稿では，複数の Dilated Convolution を組み合わせた Multiple Dilated Convolution block と FCN および Residual Network のスキップ結合を導入したエンコーダ・デコーダ型のセマンティックセグメンテーションのネットワークを提案した．これらの組み合わせにより，Cityscapes データセットにおいて，一般的なセグメンテーション手法である FCN や Segnet よりも高精度なセグメンテーションを実現した．また，最新の手法と比較して，平均カテゴリ IoU 精度および平均カテゴリ iIoU 精度は同等の精度を達成することができた．一方で，出現頻度が低いクラスの物体が大きく写る場合，誤ったセグメンテーションを行うことがある．今後は，大きな物体に対して精度向上させる方法を検討する必要がある．

参考文献

[1] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, ”Gradient-Based Learning Applied to Document Recognition” , Proceedings of the IEEE, 1998.

[2] A. Krizhevsky, I. Sutskever, G. E. Hinton, ”Imagenet classification with deep convolutional neural networks”, Advances in neural information processing systems (NIPS2012), 2012.

[3] C. Farabet, C. Couprie, L. Najman, Y. LeCun, ”Learning hierarchical features for scene labeling”, IEEE transactions on pattern analysis and machine intelligence (PAMI), 2013.

[4] M. Everingham, A. S. M. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, ”The Pascal Visual Object Classes challenge: A retrospective”, International Journal of Computer Vision (IJCV), 2014.

[5] R. Girshick, J. Donahue, T. Darrell, J. Malik, ”Rich feature hierarchies for accurate object detection and semantic segmentation”, IEEE conference on computer vision and pattern recognition (CVPR2014), 2014.

図 5　Cityscapes テストデータセットに対する結果例

[6] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, A. Rabinovich, "Going deeper with convolutions", IEEE Conference on Computer Vision and Pattern Recognition (CVPR2014), 2014.

[7] D. Kingma, J. Ba, "Adam: A method for stochastic optimization", arXiv preprint arXiv:1412.6980, 2014.

[8] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition", International Conference on Learning Representation (ICLR2015), 2015.

[9] J. Long, E. Shelhamer, T. Darrell, "Fully convolutional networks for semantic segmentation", IEEE Conference on Computer Vision and Pattern Recognition (CVPR2015), 2015.

[10] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du,P. H. Torr, "Conditional random fields as recurrent neural networks", IEEE International Conference on Computer Vision (CVPR2015), 2015.

[11] V. Badrinarayanan, A. Kendall, R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation", arXiv preprint arXiv:1511.00561, 2015.

[12] A. Kendall, V. Badrinarayanan, R. Cipolla, "Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding", arXiv preprint arXiv:1511.02680, 2015.

[13] H. Noh, S. Hong, B.Han, "Learning deconvolution network for semantic segmentation", IEEE International Conference on Computer Vision (ICCV2015), 2015.

[14] S. Loffe, C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift", arXiv preprint arXiv:1502.03167, 2015.

[15] F. Yu, V. Koltun, "Multi-scale context aggregation by dilated convolutions", International Conference on Learning Representation (ICLR2016), 2016.

[16] K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition", IEEE Conference on Computer Vision and Pattern Recognition (CVPR2016), 2016.

[17] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, B. Schiele, "The cityscapes dataset for semantic urban scene understanding", IEEE Conference on Computer Vision and Pattern Recognition (CVPR2016), 2016.

[18] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs", arXiv preprint arXiv:1606.00915, 2016.

[19] P. O. Pinheiro, T. Y. Lin, R. Collobert, P. Dollar, "Learning to refine object segments", European Conference on Computer Vision (ECCV), 2016.

[20] Z. Wu, C. Shen, A. van den Hengel, "Wider or Deeper: Revisiting the ResNet Model for Visual Recognition", arXiv preprint arXiv:1611.10080, 2016.

[21] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, "Pyramid Scene Parsing Network", arXiv preprint arXiv:1612.01105, 2016.