

情報伝搬ネットワークによる全周囲のセマンティックセグメンテーション

後藤圭汰† 山下隆義† 藤吉弘亘† 成岡健一‡ 堀田隆介‡ 玉津幸政‡

† 中部大学 ‡ 株式会社デンソー

E-mail: keita510@mprg.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識別する問題であり、走行領域や物体領域の抽出といった自動運転支援で重要な技術の一つである。車載映像からのセマンティックセグメンテーションを対象とした大規模なデータセットが公開されている。しかし、これまで公開されているデータセットは前方のみを対象としているため、車載環境における一部の状況しか理解することができない。自動運転支援においては、前方だけでなく、左右、後方を含めた 360 度全周囲の状況を把握する必要がある。そこで本研究では、左右、後方を含めた 360 度全周囲を対象とするセマンティックセグメンテーション手法を提案する。提案手法では、時間方向及び空間方向に情報を伝搬させることで、Deep Convolutional Neural Network (DCNN) をベースとした一般的なセグメンテーション手法より高精度なセマンティックセグメンテーションが可能となる。

1 はじめに

Deep Convolutional Neural Network (DCNN) [6] は、画像認識の分野において高い認識精度を達成している。DCNN は物体認識だけでなく、物体検出やセマンティックセグメンテーションの研究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、画像に含まれている物体をピクセル単位で識別する問題である。DCNN を利用したセグメンテーション手法として、Fully Convolutional Network (FCN) [7] や Pyramid Scene Parsing Network (PSPNet) [8], Encoder-Decoder 構成の SegNet [9] や U-Net [10] などがある。また、DCNN と RNN [11] を組み合わせたセグメンテーション手法として、DAG-RNN [15] がある。これらの手法は、様々なセマンティックセグメンテーションのベンチマークデータセット [1][2][3][4] で高い認識精度を実現している。中でも車載映像は、自動車の自動運転支援の実現に向けて、大規模なデータセット [1][2][5] が公開されており、近年活発的に研究が行われている。しかし、これまでに公開されているデータセットは前方を撮影した映像か

ら構成されているため、車載環境における一部しか理解することができない。自動運転支援においては、前方だけでなく、左右、後方を含めた 360 度全周囲の状況を理解する必要がある。

そこで本研究では、左右、後方を含めた 360 度全周囲を対象とするセマンティックセグメンテーション手法を検討する。全周囲の画像は、歪みの影響で位置により物体の見え方が異なる。例として、信号機は画像中央の進行方向の位置では垂直に写るが、横にいくほど湾曲して見える。これにより、同一物体でもバリエーションが増加し、セグメンテーションが困難となる問題がある。一方で、前方で撮影された物体は左右に流れていき、その後後方のカメラで撮影される。このような物体の見え方の変化を考慮した時間方向及び空間方向に情報を伝搬させる情報伝搬ネットワークを提案する。提案手法では、Long Short-Term Memory (LSTM) [12] を導入することで、短期・長期的な情報を記憶し、時間方向の情報伝搬を可能にする。また、時間方向における空間変化を考慮し情報伝搬の効果を高めるため、Dilated Convolution [13] を導入する。Dilated Convolution は畳み込み処理を行う位置を一定間隔に間引くことで、広範囲の領域を考慮した畳み込みが可能となる。これにより、360 度全周囲の画像から時間的及び空間的な変化を捉えることができ、道路と間違えやすい歩道や領域の小さな遠方のバイクを正しく識別できたりすることが期待できる。

本研究の有効性を確認するために、360 度データセットを構築する。本データセットは、車両の周囲を 4 つのカメラで撮影した画像を連結させた画像から構成されている。

2 関連研究

DCNN によるセマンティックセグメンテーションの学習では、推論結果と教師データの誤差を算出してパラメータの更新を行う。教師データには、ピクセル単位でオブジェクトごとにラベリングされた画像を使用する。DCNN を用いた代表的なセマンティックセグメンテーションの手法として、Fully Convolutional Network (FCN) [7] や SegNet [9] がある。

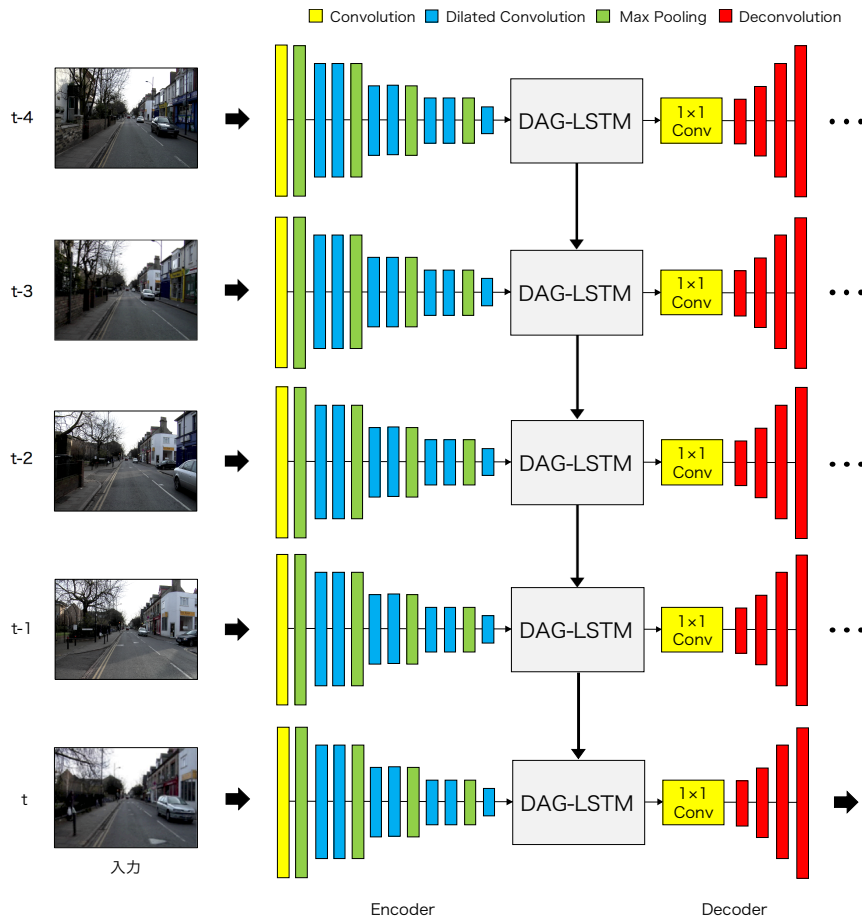


図1 提案手法のネットワーク構造

FCNは、畳み込み層とプーリング層のみで構成されたネットワークである。FCNでは、DCNNにおける全結合層をフィルタサイズ 1×1 の畳み込み層に置き換えている。畳み込み層に置き換えることで、異なる入力データサイズに対応できる。また、DCNNベースの手法は、Pooling処理を繰り返すにつれ局所的な特徴が欠落するという問題がある。この問題を解決するため、クラスラベルを推定する際に中間層の特徴マップと最終層の特徴マップを合わせて利用するスキップ構造を採用している。スキップ構造を導入することで、画像から大域的な特徴と局所的な特徴の双方を捉えることができ、高精度なセマンティックセグメンテーションが可能となる。

SegNetは、VGG16 [14]をベースとしたネットワークであり、EncoderとDecoderを組み合わせたネットワーク構造をしている。Encoderは、入力画像を圧縮し特徴マップを抽出する。また、Max Poolingを行う際、最大値だけでなく選択した最大値の位置も同時に記憶しておく。Decoderは、Encoderで抽出した特徴マップを入力画像サイズまで拡大する。この時、EncoderのPooling時に記憶していた位置でUpsamplingを行い、値を復元する。Upsampling時に値が復元されない位置には、0を補間する。これにより、正確な位置に情報を復元できる。また、Encoder-Decoder構造にすること

で、FCNと比較して高速化及び省メモリ化を実現している。

DCNNとRNNを組み合わせたセマンティックセグメンテーションの手法としてDAG-RNN [15]がある。DCNNは、注目ピクセルの近傍である局所的な情報からクラス識別を行う。RNNを組み合わせることによって広範囲の空間領域を文脈として捉えたクラス識別が可能となる。はじめに、畳み込み処理とPooling処理を繰り返し特徴マップを抽出する。次に、得られた特徴マップをDAG-RNNに入力し、注目ピクセルに対して前層の特徴マップの入力と近傍3方向からの入力を与え、周辺の画素の情報を考慮した特徴マップを生成する。生成された特徴マップをDeconvolution層に入力し、セグメンテーション結果を得る。また、大域的な特徴を捉え利用する手法としてDilated Convolution [13]がある。Dilated Convolutionは、畳み込み処理を一定間隔離れた位置に行うことで、広範囲の領域を考慮した畳み込みが可能となる。

一般的なEncoder-Decoderネットワーク [9]では、Pooling時に局所特徴量が欠落するため、小領域クラスの識別が困難である。また、DCNNをベースとしたネットワーク [7]では、局所的な情報からクラス識別を行うので、大域的な情報が考慮されにくい。提案手法では、DCNNとRNNを組み合わせ、Dilated Convo-

表 1 情報伝搬ネットワークの各層の詳細

層	処理内容	フィルタサイズ	フィルタ数	Pooling
1層目	Conv.	3×3	16	—
2層目	Dilated Conv.	3×3	32	Max Pooling
3層目	Dilated Conv.	3×3	32	—
4層目	Dilated Conv.	3×3	64	Max Pooling
5層目	Dilated Conv.	3×3	64	—
6層目	Dilated Conv.	3×3	128	Max Pooling
7層目	Dilated Conv.	3×3	128	—
8層目	Dilated Conv.	3×3	256	Max Pooling
9層目	LSTM	—	—	—
10層目	Deconv.	3×3	128	Upsampling
11層目	Deconv.	3×3	64	Upsampling
12層目	Deconv.	3×3	32	Upsampling
13層目	Deconv.	3×3	12	Upsampling

lutionを導入することによって、局所的な特徴と大局的な特徴を捉えることが可能となる。これにより、従来手法と比較して高精度なセマンティックセグメンテーションが期待できる。

3 提案手法

物体の見え方の変化に対応するため、時間方向及び空間方向に情報を伝搬させる情報伝搬ネットワークを提案する。提案手法のネットワーク構造を図1に、各層の詳細を表1に示す。提案手法のネットワークは、8層の畳み込み層と4層のDeconvolution層で構成されている。各層のフィルタサイズは 3×3 であり、フィルタ数はPooling処理を行うごとに2倍に増加する。Pooling処理として、局所領域における最大値を選択するMax Poolingを使用している。また、8層目の畳み込み層と9層目のDeconvolution層の間にLSTMを導入する。LSTMを導入することで、短期・長期的な情報を記憶し、時間方向の情報伝搬が可能となる。1層目以降の畳み込み層では、通常の畳み込み処理でなく、Dilated Convolutionを行う。Dilated Convolutionは、畳み込み処理を離れた位置に行うことで、空間情報を考慮した畳み込みが可能となる。以下に時間方向の情報伝搬、空間方向の情報伝搬およびネットワークの学習方法の詳細について述べる。

3.1 時間方向の情報伝搬

時間方向に情報を伝搬させるため、RNNを導入する。伝搬させる情報として、8層目の畳み込み処理で抽出された特徴マップを使用する。現時刻を t とした時、 $t-4$ から t までの5フレーム間の情報をRNNにより伝搬させる。時間方向に情報を伝搬させることで、画像中の物体の動きや自転車や他車の進行方向といった時間による変化を意味的に学習することができる。

3.2 空間方向の情報伝搬

空間方向に情報を伝搬させるため、DAG-RNN with LSTM (DAG-LSTM) およびDilated Convolutionを導入

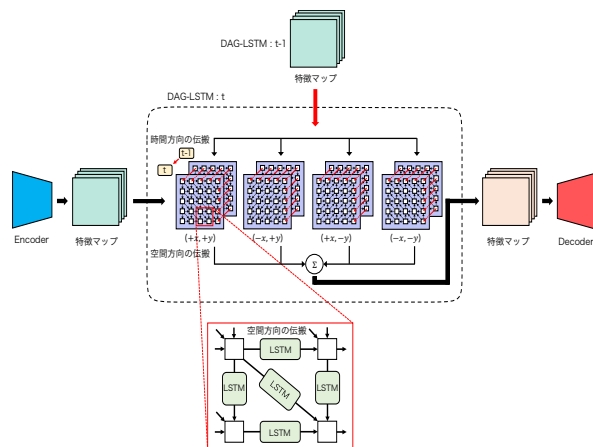


図 2 DAG-LSTM の内部構造

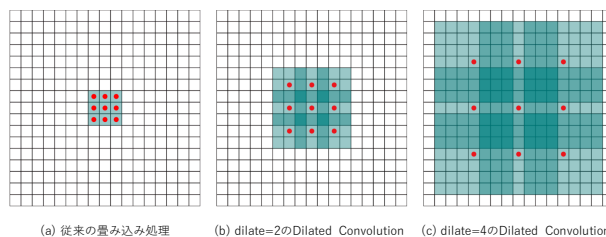


図 3 従来の畳み込み処理とDilated Convolutionの違い

入する。DAG-LSTMの詳細を図2に示す。DAG-RNNは、DCNNとRNNを組み合わせたセマンティックセグメンテーション手法である。2次元に拡張したRNNをDCNNに導入することで、広範囲の空間領域を文脈として捉えたクラス識別が可能となる。従来のDAG-RNNは、Elman型のRNNを利用している。しかし、RNNは数時刻分程度の入力しか出力に反映されず、長期的な記憶ができないという問題がある。本研究では、RNNの代わりにLSTMを導入したDAG-LSTMを提案する。LSTMを導入することで、短期・長期的な情報を記憶することができる。これにより、RNNと比べて情報伝搬の効果を高めることができる。

時間方向における空間変化を考慮し情報伝搬の効果を高めるため、Dilated Convolutionを導入する。Dilated Convolutionの概要を図3に示す。従来の畳み込み処理では、図3(a)のように、 3×3 のフィルタを画像中の同じ位置の位置に対して畳み込み処理を行う。フィルタの要素を赤点、畳み込む位置を緑とした時、隣接した密な位置への畳み込みとなる。一方、Dilated Convolutionは図3(b)のように、 3×3 のフィルタを画像中の2ピクセル離れた位置に畳み込む。または、図3(c)のように、4ピクセル離れた位置に畳み込む。離れた位置に畳み込むことによって、大局的な特徴を捉えることができる。したがって、自転車や他車が動いた場合でも周辺の情報を広範囲に捉えることができ、セグメンテーション精度の向上が期待できる。

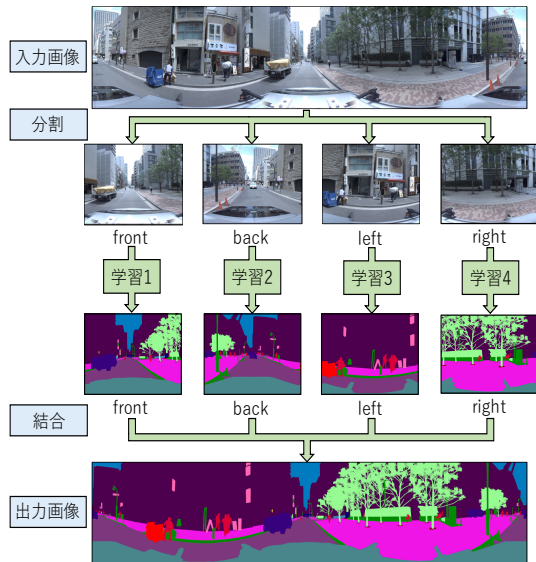


図4 360度データセットを用いた学習の流れ

3.3 ネットワークの学習方法

本ネットワークはエンコード部、デコード部およびLSTM部をend-to-endで学習する。学習時のバッチサイズは3、最適化関数はAdam[16]を用いる。360度データセットを学習に用いる場合、画像全体を入力せずに、情報伝搬ネットワークに適した入力を行う。360度データセットを用いた際の学習の流れを図4に示す。はじめに、全周囲画像を前後左右の4つの領域に分割し、画像として保存する。次に、分割された各領域に対して、個別のネットワークを用いて学習する。最後に、各領域の出力結果を結合し、全周囲画像を獲得する。

4 評価実験

情報伝搬ネットワークの有効性を検証するため、車載カメラで撮影されたシーンを対象としたセマンティックセグメンテーション用のデータセットを用いて、評価実験を行う。以下にデータセット、評価方法、実験結果の詳細について述べる。

4.1 データセット

評価データには、CamVid データセット [1] および 360 度データセットを用いる。CamVid データセットの画像例を図5に、360度データセットの画像例を図6に示す。CamVid データセットは、昼間および夕暮れのケンブリッジ付近を対象とした道路シーンの画像で構成されており、学習用の画像 367 枚と評価用の画像 233 枚が含まれている。CamVid データセットのクラス数は、32 クラスのものと 11 クラスに削減したものと 2 種類ある。本研究では、複数のクラスをグループ化し、11 クラスに削減した CamVid データセットを用いる。

360 度データセットは、車両の周囲を 4 つのカメラで撮影した画像を連結させた画像から構成されている。そのため、物体が撮影されたカメラの位置により見え

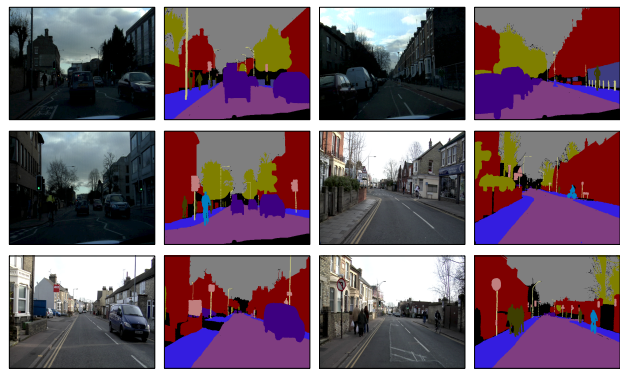


図5 CamVid データセットの例

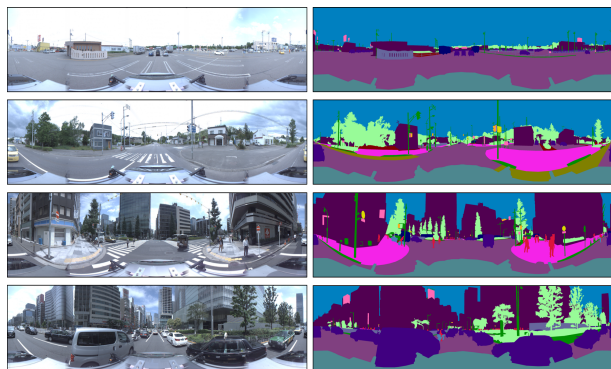


図6 360度データセットの例

方が異なる。本データセットは日本国内の3地域で撮影しており、クラス数は19クラスである。また、360度データの画像サイズは、 5171×1160 である。本研究では、360度データセットを学習用に970枚、評価用に475枚用いる。学習データはData Augmentationにより4850枚にした後、学習過程でランダムに一部分を切り出してパッチ学習を行なっている。

4.2 評価方法

セマンティックセグメンテーションにおける定量的評価指標として、全体精度、クラス精度、平均IoUがある。以下に各評価方法の詳細について述べる。

4.2.1 全体精度

全体精度は、1枚の画像全体に対する精度である。画像全体のピクセル数を P_a 、正解したピクセルの数を P_c 、全体精度を A_g とすると、式(1)で表すことができる。

$$A_g = \frac{P_c}{P_a} \quad (1)$$

4.2.2 クラス精度

クラス精度は、クラスごとに識別率を算出し、各クラスの識別率を平均した精度である。全体精度では、対象物体のスケールによって精度が偏る傾向があるのに対し、クラス精度は各クラスの識別率から最終的な精度を求めることで、偏りを軽減させることができる。ク

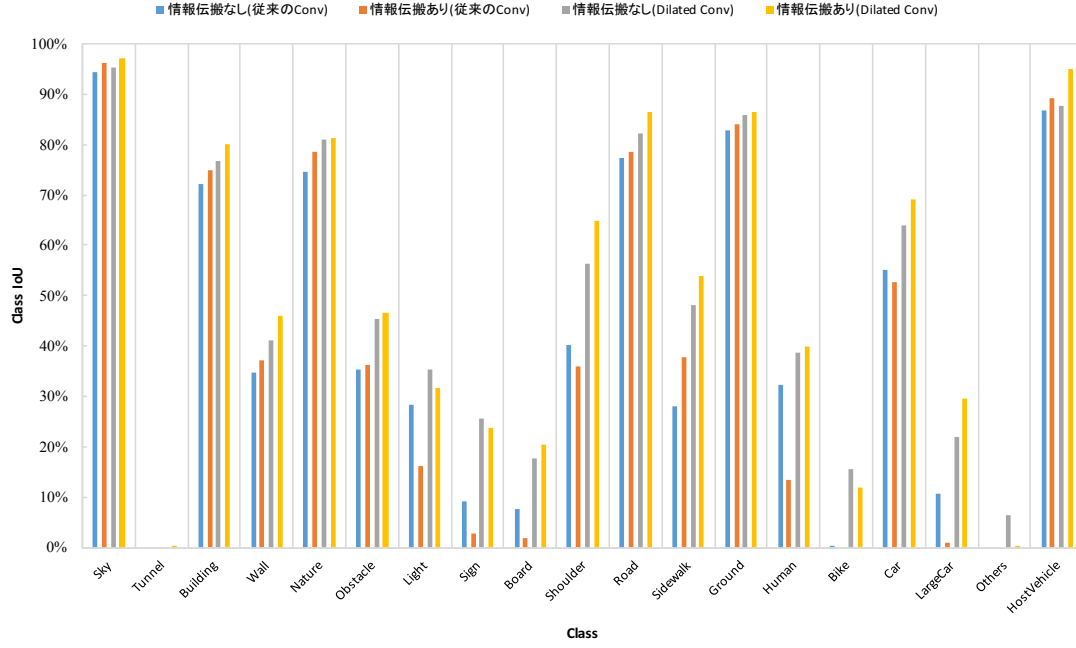


図 7 360 度データセットにおけるクラスごとの IoU 精度

表 2 CamVid データセットでの比較結果 [%]

手法	全体精度	クラス精度	平均 IoU
SegNet (論文)	88.50	65.20	55.60
SegNet (再学習)	79.31	66.61	46.27
DAG-RNN	85.09	57.64	47.74
DAG-RNN with LSTM	89.83	70.18	59.12
情報伝搬ネットワーク	89.25	71.73	61.50

ラス数を K ，対象クラスの正推定領域を H ，対象クラスの正解領域を C ，クラス精度を A_c とすると，式 (2) で表すことができる．

$$A_c = \frac{1}{K} \sum_{i=1}^K \left(\frac{H_i}{C_i} \right) \quad (2)$$

4.2.3 平均 IoU

平均 IoU は，クラス精度に対象クラスの誤推定領域を考慮させた精度である．平均 IoU は，セマンティックセグメンテーションの評価指標として広く利用されている．クラス数を K ，対象クラスの正推定領域を H ，対象クラスの正解領域を C ，対象クラスの誤推定領域を E ，平均 IoU を M_i とすると，式 (3) で表すことができる．

$$M_i = \frac{1}{K} \sum_{i=1}^K \left(\frac{H_i}{C_i + E_i} \right) \quad (3)$$

4.3 CamVid データセットにおける評価結果

提案手法が一般的な車載映像データセットにおいても有効であるかを確認するため，CamVid データセッ

表 3 360 度データセットでの比較結果 [%]

情報伝搬	畳み込み	全体精度	クラス精度	平均 IoU
なし	従来	87.41	47.70	40.50
あり	従来	88.50	45.19	38.74
なし	Dilated	89.87	57.16	48.67
あり	Dilated	91.89	58.24	50.72

トを用いて識別精度の比較を行う．空間方向および時間方向の伝搬を加えた情報伝搬ネットワークと空間方向のみの手法との精度比較結果を表 2 に示す．表 2 より，提案手法である情報伝搬ネットワークの精度が最も良い．DAG-RNN with LSTM と比較して全体精度は 89.83% から 89.25% と多少低下しているものの，クラス精度は 70.18% から 71.73%，平均 IoU は 59.12% から 61.50% と識別精度が向上している．このことから，一般的な車載映像データセットにおいても情報伝搬ネットワークが有効であるといえる．

4.4 360 度データセットにおける評価結果

360 度データセットにおける情報伝搬ネットワークの効果を確認するため，情報伝搬の有無と用いる畳み込み処理の種類による識別精度の比較を行う．360 度データセットでの識別精度の比較結果を表 3 に示す．表 3 より，Dilated Convolution と情報伝搬処理を導入することで，各評価指標で識別精度が向上している．従来の畳み込み処理の場合，情報伝搬処理を行うことでクラス精度，平均 IoU が約 2% 低下している．一方，Dilated Convolution を導入することで，時間方向における空間変化を考慮することができ，情報伝搬処理の効果を高めている．

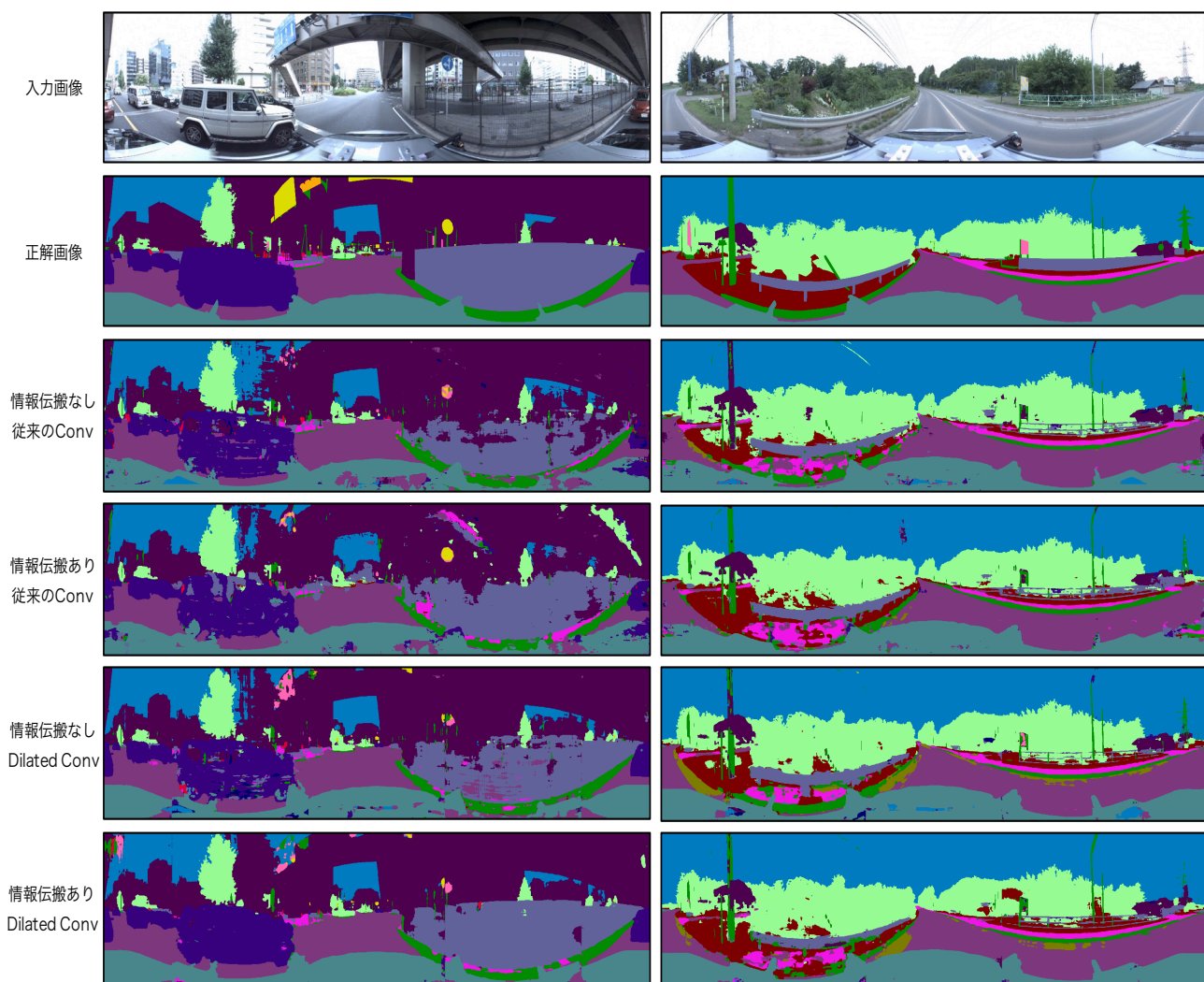


図 8 360 度データセットにおけるセグメンテーション結果例

360 度データセットにおけるクラスごとの IoU を図 7 に示す。情報伝搬処理に Dilated Convolution を導入することで、Shoulder や LargeCar といった大きな物体の精度が向上していることがわかる。このことから大局的な特徴を捉えることができているといえる。また、従来の畳み込み処理で情報伝搬を行うと精度が低下していた Human の精度が低下することなく、同等以上の精度となっている。これより、領域の小さなクラスと大きなクラスのどちらにおいても精度が向上していることがわかる。

図 8 に各手法のセグメンテーション結果を示す。図 8 より、各シーンにおいて、大きな領域の一部において誤識別が減少していることがわかる。これより、提案手法では 360 度データに対して高精度なセマンティックセグメンテーションが実現できたといえる。

5 おわりに

本研究では、時間方向および空間方向に情報を伝搬させる情報伝搬ネットワークを提案した。時間方向の情

報伝搬処理では、5 フレーム間の情報を RNN により伝搬させることで、画像中の物体の動きや自車や他車の進行方向といった時間による変化を捉えることができた。空間方向の情報伝搬処理では、DAG-RNN を導入した。従来の DAG-RNN では長期的な記憶ができないという問題があるが、本研究では RNN を LSTM に置き換えることで長期的な記憶を実現した。また、Dilated Convolution を導入することで、自車や他物体が動いた場合でも周辺の情報を広範囲に捉えることができた。情報を伝搬させることで物体の見え方の変化に対応することができ、360 度データにおけるセマンティックセグメンテーションの高精度化を実現した。なかでも Shoulder や LargeCar といった大きな物体の精度が向上した。一方、Sign や Light のように歪みの影響を受けやすいクラスの IoU 精度が、情報伝搬なしの時と比べて低下している。今後は、カメラ間で情報伝搬を行い各クラスの精度向上を図るとともに、歪みの影響を受けやすいクラスの精度を向上させる手法を検討する。

参考文献

- [1] G. Brostow, J. Shotton, J. Fauqueur, R. Cipolla, “Segmentation and Recognition Using Structure from Motion Point Clouds,” European Conference on Computer Vision (ECCV), 2008.
- [2] 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 (CVPR), 2016.
- [3] 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.
- [4] N. Silberman, D. Hoiem, P. Kohil, R. Fergus, “Indoor Segmentation and Support Inference from RGBD Images,” European Conference on Computer Vision (ECCV), 2012.
- [5] G. Neuhold, T. Ollmann, S. R. Bulò, P. Kotschieder, “The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes,” International Conference on Computer Vision (ICCV), 2017.
- [6] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, “Gradient-Based Learning Applied to Document Recognition,” in Proceedings of the IEEE, vol. 86, no. 11, pp.2278-2324, 1998.
- [7] J. Long, E. Shelhamer, T. Darrell, “Fully convolutional networks for semantic segmentation,” IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
- [8] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, “Pyramid Scene Parsing Network,” arXiv preprint arXiv:1612.01105, 2016.
- [9] V. Badrinarayanan, A. Kendall, R. Cipolla, “SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation,” arXiv:1511.00561, 2016.
- [10] O. Ronneberger, P. Fischer, T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015.
- [11] J. Elman, “Finding Structure in Time,” Cognitive Science. vol. 14, no. 2, pp.179-211, 1990.
- [12] S. Hochreiter, “Long Short-Term Memory,” Neural Computation. vol. 9, no. 8, pp. 1735-1780, 1997.
- [13] F. Yu, V. Koltun, “Multi-scale context aggregation by dilated convolutions,” International Conference on Learning Representation (ICLR), 2016.
- [14] K. Simonyan, A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” International Conference on Learning Representation (ICLR), 2015.
- [15] B. Shuai, Z. Zuo, G. Wang, B. Wang, “DAG-Recurrent Neural Networks For Scene Labeling,” Conference on Computer Vision and Pattern Recognition (CVPR), 2015.
- [16] D. Kingma, J. Ba, “Adam: A method for stochastic optimization,” arXiv preprint arXiv:1412.6980, 2014.