カメラ間の整合性を考慮した全周囲画像のセグメンテーション

○後藤 圭汰†，平川 翼†，山下 隆義†，藤吉 弘亘†
†: 中部大学
{keita510@mprg.cs., hirakawa@mprg.cs., takayoshi@isc., fujiyoshi@isc.}chubu.ac.jp

概要 : 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把握といった自動運転支援において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．そこで本研究では，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と空間方向及びカメラ間に情報を伝搬させることで，従来のセグメンテーション手法より高精度なセグメンテーションが可能であることを確認した．

セマンティックセグメンテーション，情報伝搬ネットワーク，カメラ間伝搬

1. はじめに

Deep Convolutional Neural Network (DCNN)[1]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．中でも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り，数多くの手法が提案されている．DCNN を用いたセグメンテーション手法として，Fully Convolutional Network (FCN)[2]や Pyramid Scene Parsing Network (PSPNet)[3], Encoder-Decoder 構成の SegNet[4]や U-Net[5]などがある．また, DCNN と RNN[6]を組み合わせたセグメンテーション手法として, DAG-RNN[7]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[8][9][10][11]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システムの実現に向けて，大規模なデータセット[8][9][12]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，車載環境において一部の状況しか理解することができない．

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．全周囲の画像は，位置により物体の見え方が異なる性質を持っている．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．このような物体の見え方の変化を考慮した情報を伝搬させる情報伝搬ネットワーク[14]が提案されている．情報伝搬ネットワークは時間方向と空間方向に情報を伝搬させる．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]を導入している．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．

しかし，360 度全周囲の画像は 4 つのカメラで撮影された画像を連結して構成されているので，カメラ端でのセグメンテーションが困難となる問題がある．特に画像を分割して学習する際，セグメンテーション精度が低下する傾向がある．そこで，各カメラで撮影された画像においてセグメンテーションの整合性を保つために，カメラ端領域が近づくように学習を行う．これにより，従来の情報伝搬ネットワークと比べて，カメラ端のセグメンテーション精度の向上が期待できる．

本研究の有効性を確認するため，360 度データセットで評価する．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ものを連結させた画像から構成されている．

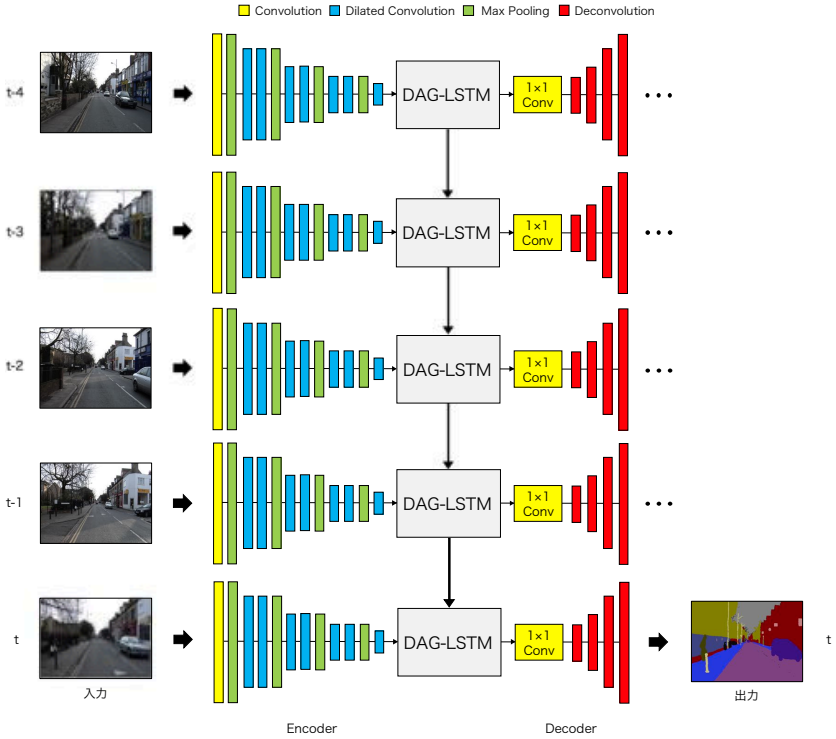

図 1:情報伝搬ネットワークのネットワーク構造

2. 関連研究

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手法として，Fully Convolutional Network (FCN)[2]や SegNet[4]がある．FCN は畳み込み層とプーリング層のみで構成されたセグメンテーション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，異なるサイズの入力データに対応することができる. また, プーリング層では, プーリング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．FCN ではこの問題を解決するために，スキップ構造を導入している．スキップ構造は，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．これにより，画像から大局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[15]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，畳み込み層とプーリング層で構成されており，入力画像を圧縮し特徴マップを抽出する．SegNet は，プーリング層で Max Pooilng 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 のプーリング時に記憶していた位置で値を復元し，Upsampling を行う．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.

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[7]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．DAG-RNN は畳み込み処理とプーリング処理によって特徴マップを抽出する．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，大局的な特徴を捉えることができる手法として 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になる．一般的な Encoder-Decoder 構造のネットワーク[4]では，プーリング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になる問題がある．また，DCNN をベースとしたネットワーク[2]では，局所的な情報からクラス識別を行うので，大局的

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 Convolution 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.

3. 情報伝搬ネットワーク

360 度全周囲の画像は位置により物体の見え方が異なるため，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．情報伝搬ネットワークでは，物体の見え方の変化に対応するため，時間方向と空間方向に情報を伝搬させる機構を導入している．情報伝搬ネットワークの構造を図 1 に示す．情報伝搬ネットワークは，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，フィルタ数はプーリング処理を行うごとに 2 倍に増加する．プーリング処理として Max Pooling を採用している．また，8 層目の畳み込み層と 9 層目の Deconvolution 層の間に DAG-LSTM を導入する. DAG-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，空間情報の伝搬が可能となる．時間方向の伝搬には 5 フレーム間の特徴マップを利用する．1 層目以降の畳み込み層では, 通常の畳み込み処理ではなく, Dilated Convolution を行う．これにより広範囲の情報を考慮した畳み込み処理が可能になる．以下に時間方向の情報伝搬と空間方向の情報伝搬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)及び Dilated Convolution を導入する. 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の手法である．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

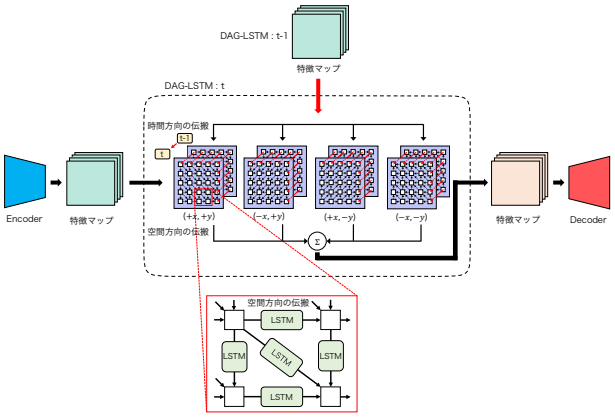

図 2：DAG-LSTM の構造

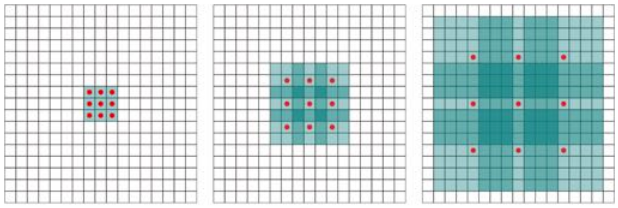

図 3 ：従来の畳み込み処理と Dilated Convolution の違い

可能となる．従来の DAG-RNN は，Elman 型の RNN[6]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[16]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.

時間方向と空間方向の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サイズ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，隣接した密な位置との畳み込み処理となる．一方，Dilated Convolution は図 3(b)のように 3×3 のフィルタを画像中の同サイズの位置に対して畳み込み処理を行う．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．離れた位置に対して畳み込むによって，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る.

4. カメラ間の情報伝搬

360 度全周囲の画像は，前方・後方・左右のカメ

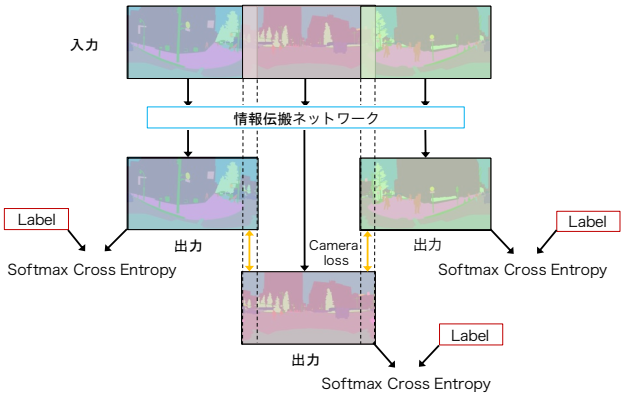

図 4：カメラ間の情報伝搬

ラで撮影された画像を連結し構成される．このとき，各カメラの画像における端部分でセグメンテーションの整合性が保たれず，精度が低下する問題がある．例として，前方のカメラと左のカメラで撮影された画像に同一の車が画像のつなぎ目に写っていた場合，セグメンテーションが困難になる．そこで図 4 のように，カメラ間で整合性を保つようにする．はじめに，全周囲画像を分割する．分割を行う際，分割後の画像に隣接するカメラで撮影された画像の端が含まれるようにする．各カメラ間の端領域において，隣接するカメラの端領域に近づくように学習を行う．前方と左右のカメラで考えた場合，損失関数は式(1)のように表すことができる．ここで，L_{SCE}は各ラベルとの Softmax Cross Entropy を表し，$S_{CL}\cdot S_{L1}\cdot S_{R1}$は各カメラの端領域である．それらを Jensen-Shannon Divergense (JSD)で近づけるように学習する．

$$L = L_{SCE} + JSD(S_{CL} \parallel S_{L1}) + JSD(S_{CR} \parallel S_{R1}) \quad (1)$$

これにより各カメラ間の整合性が保たれ，カメラ端領域におけるセグメンテーション精度の向上が期待できる．

5. 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．

5.1. データセット

評価データには 360 度データセットを用いる．360 度データセットの画像例を図 5 に示す．360 度データセットは，車両の周囲を 4 つのカメラで撮

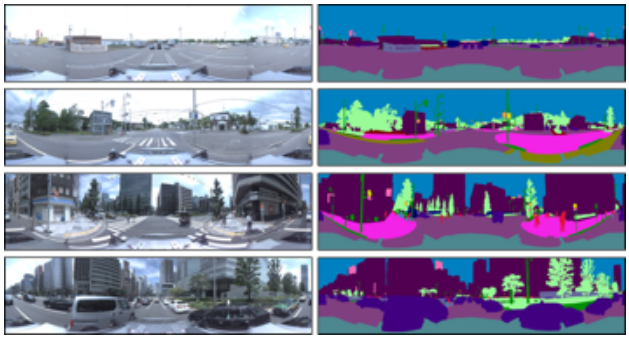

図 5：360 度データセットの例

影し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方が異なる．本データセットは日本国内の複数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 1466 枚，評価用に 104 枚用いる．学習データに対して Data Augmentation を行い，学習サンプルのバリエーションを増加させている．最適化関数には Adam[17]を用いている．

5.2. 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．以下に各評価指標について述べる．

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(2)で表すことができる．

$$A_g = \frac{P_c}{P_a} \quad (2)$$

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(3)で表すことができる．

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

平均 IoU は，クラス精度に対象クラスの誤推定領域を考慮した精度である．平均 IoU は，セマンテ

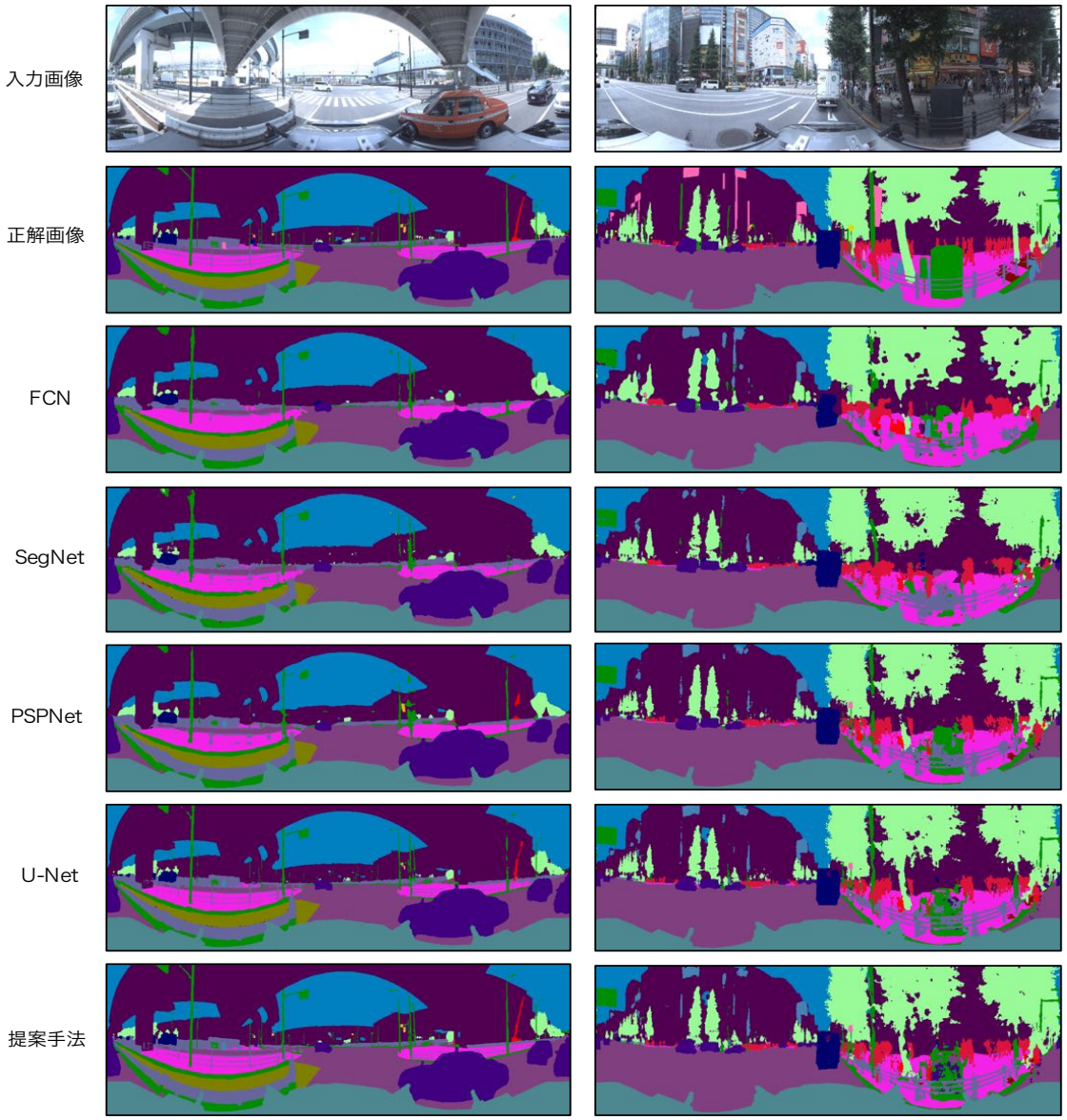

図 6:360 度データセットにおけるセグメンテーション結果例

表 1 ：情報伝搬の有無による比較結果[%]

情報伝搬	畳み込み	全体精度	クラス精度	平均 IoU
なし	従来	92.4	60.3	52.8
あり	従来	91.9	58.6	50.7
なし	Dilated	92.2	61.5	54.1
あり	Dilated	**94.3**	**62.4**	**56.2**

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(4)で表すことができる．

$$M_i = \frac{1}{K}\sum_{i=1}^{K}\left(\frac{H_i}{C_i + E_i}\right) \tag{4}$$

5.3. 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有効性を検証するため，各伝搬処理の有無による精度と従来手法との識別精度を比較する．情報伝搬の有無による識別精度の比較を表 1 に示す．表 1 より，Dilated Convolution と情報伝搬処理を導入することで，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精度が低下している．一方, Dilated Convolution を行うことで広範囲の情報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．

従来手法との比較結果を表 2 に示す. 表 2 より，各評価指標において従来手法より精度が向上していることがわかる．中でも SegNet と比較して平均

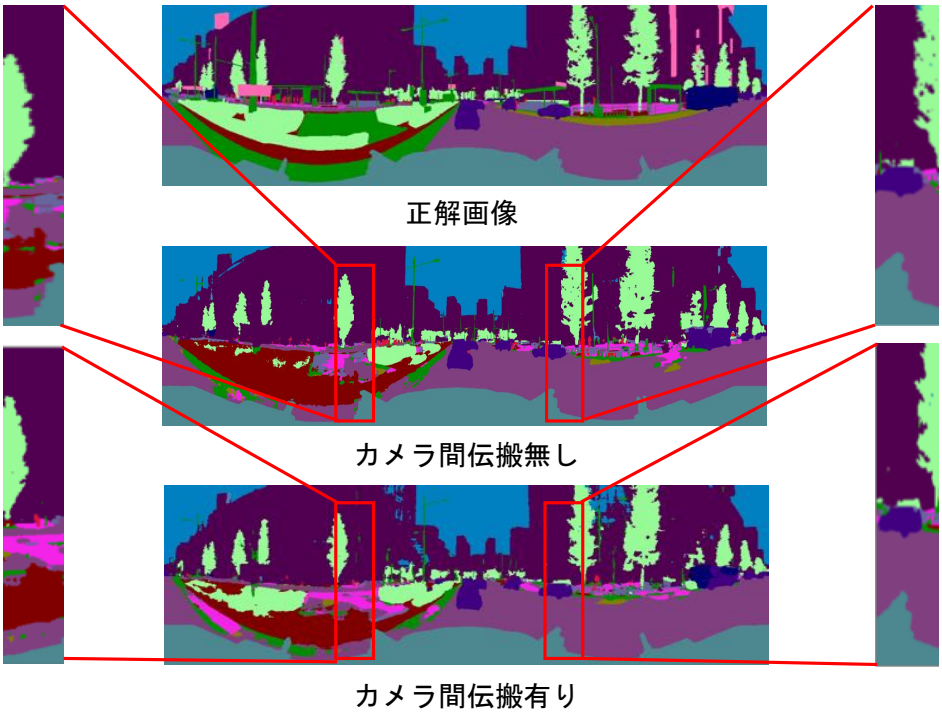

図 7:カメラ間伝搬の有無によるセグメンテーション結果の比較

表 2：従来手法との比較結果[%]

	全体精度	クラス精度	平均 IoU
FCN	92.5	58.5	52.1
SegNet	92.4	55.2	48.9
U-Net	93.8	60.2	54.4
PSPNet	93.5	61.6	55.1
提案手法	**94.3**	**62.4**	**56.2**

IoU が約 7%向上していることが分かる．したがって，情報伝搬処理は有効であるといえる．

図 6 に各手法のセグメンテーション結果を示す．図 6 より，道路や空といった大きなクラスのセグメンテーションができていることがわかる．このことから提案手法では，大局的な特徴を捉えることができているといえる．また，SegNet や FCN と比較して提案手法は，クレーン車や柵といった小さな領域のクラスに対しても高精度なセグメンテーションができていることが分かる．このことから局所的な特徴も捉えることができているといえる．

5.4. カメラ間伝搬の有効性の確認

カメラ間の伝搬の有無によるセグメンテーション結果の比較を図 7 に示す．図 7 より，カメラ間伝搬無しの出力結果ではカメラ端領域において，

表 3：カメラ間伝搬の有無による精度の比較[%]

	全体精度	クラス精度	平均 IoU
伝搬無し	93.8	61.2	55.1
伝搬有り	**94.3**	**62.4**	**56.2**

車や歩道クラスのセグメンテーションの整合性が保たれてないことがわかる．一方，カメラ間伝搬有りの出力結果ではカメラ端領域において，セグメンテーションの整合性が保たれていることが確認できる．また，表 3 にカメラ間の伝搬処理の有無による識別精度の比較を示す．表 3 より，伝搬無しと比べて伝搬有りでは，約 1%精度が向上している．したがって，360 度データに対するカメラ間の伝搬処理は有効的であるといえる．

6. おわりに

本研究では，時間方向と空間方向及びカメラ間に情報を伝搬させる情報伝搬ネットワークを提案した．時間方向の伝搬処理では，5 フレーム間の情報を伝搬させることで，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．空間方向の伝搬処理では，DAG-LSTM 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，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 をあわせて導入す

ることで，周辺の情報を広範囲に捉えることができた．カメラ間の伝搬処理では，各カメラ端の領域が一致するように学習することで，カメラ間の整合性を保つことが可能になった．情報を伝搬させることで，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．今後は，伝搬処理の改良及び標識や信号機といった識別が困難なクラスの精度向上を課題に研究を行う．

参考文献

[1] 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.

[2] J. Long, E. Shelhamer, T. Darrell, “Fully convolutional networks for semantic segmentation,’’ IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[3] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia,‘‘Pyramid Scene Parsing Network,’’ IEEE Conference on Computer Vision and Pattern recognition (CVPR), 2017.

[4] V. Badrinarayanan, A. Kendall, R. Cipolla, “SegNet : A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation,’’ arXiv:1612.01105, 2016.

[5] O.Ronneberger, P. Fischer, T. Brox, ‘‘U-Net : Convolutional Networks for Biomedical Image Segmentation,’’ Medical Image Computing and Computer-Assisted Intervention (MICCAI), 2015.

[6] J. Elman, ‘‘Finding Structure in Time,’’ Cognitive Science. Vol. 14, no. 2, pp.1735-1780, 1997.

[7] B. Shuai, Z. Zuo, G. Wang, B. Wang, ‘‘DAG-Recurrent Neural Networks For Scence Labeling,’’ Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[8] G. Brostow, J. Shotton, J. Fauqueur, R. Cipolla, ‘‘Segmentaion and Recognition Using Structure from Motion Point Clouds,’’ European Conference on Computer Vision (ECCV), 2008.

[9] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benensen, B. Schiele, “The cityscapes dataset for semantic urban scene understanding,’’ IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[10] 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 Retro spective,’’ International Journal of Computer Vision (IJCV), 2014.

[11] N.Silberman, D. Hoiem, P. Kohil, R. Fergus, “Indoor Segmentation and Support Inference from RGBD Images,’’European Conference on Computer Vision (ECCV), 2012.

[12] G. Neuhold, T. Ollmann, S. R. Bulo, P. Kontschieder, “The Mapillary Vistas Dataset for Semantic Understanding of Street Scenes,’’ International Conference on Computer Vision (ICCV), 2017.

[13] F. Yu, V. Koltun, “Multi-scale context aggregation by dilated convolutions,’’ International Conference on Learning Representation (ICLR), 2016.

[14] 後藤圭汰, 山下隆義, 藤吉弘亘, 成岡健一, 堀田隆介, 玉津幸政, “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”, 画像センシングシンポジウム, 2018.

[15] K. Simonyan, A. Zisserman, “Very deep convolutional networks for large-scale image recognition,’’ International Conference on Learning Representation (ICLR), 2015.

[16] S.Hochreiter, “Long Short-Term Memory,’’ Neural Computation. Vol. 9, no. 8, pp.1735-1780,1997.

[17] D. Kingma, J. Ba, “Adam : A method for stochastic optimization,’’ arXiv preprint arXiv:1412:6980, 2014.

後藤圭汰 : 2018 年中部大学工学部情報工学科卒業，現在同大学大学院工学研究科修士課程在学中．全周囲画像を用いたセマンティックセグメンテーションの研究に従事．

平川翼 : 2013 年広島大学大学院博士課程前期修了．2014 年広島大学大学院博士課程後期入学，2017 年中部大学研究員(~2019 年), 2017 年広島大学大学院博士後期課程修了．2019 年中部大学特任助教．コンピュータビジョン，パターン認識，医用画像処理の研究に従事．

山下隆義 : 2002 年奈良先端科学技術大学大学院博士前期課程修了．2002 年オムロン株式会社入社，2009 年中部大学大学院博士後期課程修了(社会人ドクター)，2014 年中部大学講師，2017 年より同大学准教授．人の理解に向けた動画像処理，パターン認識・機械学習の研究に従事．

藤吉弘亘 : 1997 年中部大学大学院博士後期課程修了．1997~2000 年米カーネギメロン大学ロボット工学研究所 Postdoctoral Fellow.2000 年中部大学講師．2004 年より同大学教授．2005~2006 年米カーネギメロン大学ロボット工学研究所客員研究員．計算機視覚，動画像処理，パターン認識・理解の研究に従事．