運転計測情報を用いたデータ選択と半教師あり学習によるセマンティックセグメンテーション

山田裕大† 正木翔大† 平川翼† 山下隆義† 藤吉弘亘†
†中部大学
E-mail: nell@mprg.cs.chubu.ac.jp

1 はじめに

セマンティックセグメンテーションは，自動運転車の走行可能領域の把握や周辺状況の把握のために必要な技術である．通常のセマンティックセグメンテーションの学習には，画像の各ピクセルに対して正解ラベルを付与する必要があり，人的コストが高いという問題点がある．

この問題を解決するために，教師ありデータと教師なしデータの両方を用いて学習を行う半教師あり学習が注目されている [1]．しかし，半教師あり学習において教師ありデータの割合が極端に少ない場合，精度が低下するという問題がある．

そこで本研究では，教師なしデータの一部に擬似ラベルを付与した半教師あり学習法を提案する．擬似ラベルは，教師ありデータのみを使用して学習したモデルを用いて生成する．生成した擬似ラベルを教師なしデータに付与し，教師ありデータとして扱う．この時，少量の教師ありデータで学習したモデルは，学習データと大きく異なるシーンのデータに対して精度が低く，擬似ラベルとして用いることが適切ではないことが考えられる．そこで，提案手法では，データ撮影時に獲得した計測情報をもとに，擬似ラベルを付与する教師なしデータ選択することによって，学習データと似たシーンの擬似ラベルを使用する．これにより，教師ありデータの割合を増やし，精度の向上を実現する事を示す．

2 半教師あり学習

半教師あり学習とは，ラベルの無い大量のデータを併用することで，少量の教師ありデータでも高い精度を獲得することを目的とした手法である．主なアプローチとして，エントロピー最小化 [2] や，擬似ラベリング [3] などがある．

セマンティックセグメンテーションを対象とした半教師あり学習法として，Cross-Consistency Training (CCT) [4] が提案されている．CCT のネットワーク構造を図 1 に示す．CCT は通常の教師あり学習と並行する形で，ラベルなしデータを用いた一貫性学習を行う．一貫性学習は，ラベルなしデータに対するエンコーダ出力に摂動を加えたものを入力する補助デコーダを使用し，摂動を適用していないメインデコーダの出力と一貫性を持たせるように学習する．これにより，単純な構造のネットワークモデルを用いた場合でも高い精度を得ることが可能となる．エンコーダは，ImageNet で事前学習された ResNet-50 [5] と，PSP モジュール [6] により構成される．従来のセマンティックセグメンテーション手法 [6, 7, 8] に倣って，ResNet の最後の 2 つの stride convolution を dilated convolution に変更している．デコーダは，効率とパラメータ数を考慮して，1×1 畳み込みのみを使用する．クラス数に合わせて深さを合わせるために畳み込み処理を行った後，出力を入力サイズにアップサンプリングするために，Rectified Linear Unit (ReLU) を持つ 3 つのサブピクセル畳み込みを適用する．

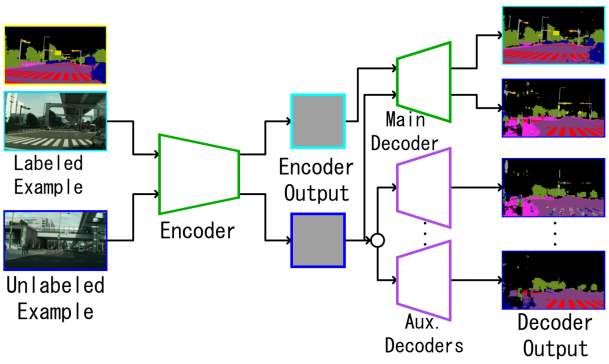

図 1 ネットワーク構造

3 提案手法

半教師あり学習には，教師なしデータに対し教師ありデータが極端に少ない場合，精度が低下する問題がある．そこで本研究では，擬似ラベルを用いた半教師あり学習法を提案する．擬似ラベルは，事前に少量の学習データで学習したモデルを用いて生成し，教師なしデータに対して付与する．しかし学習データと大き

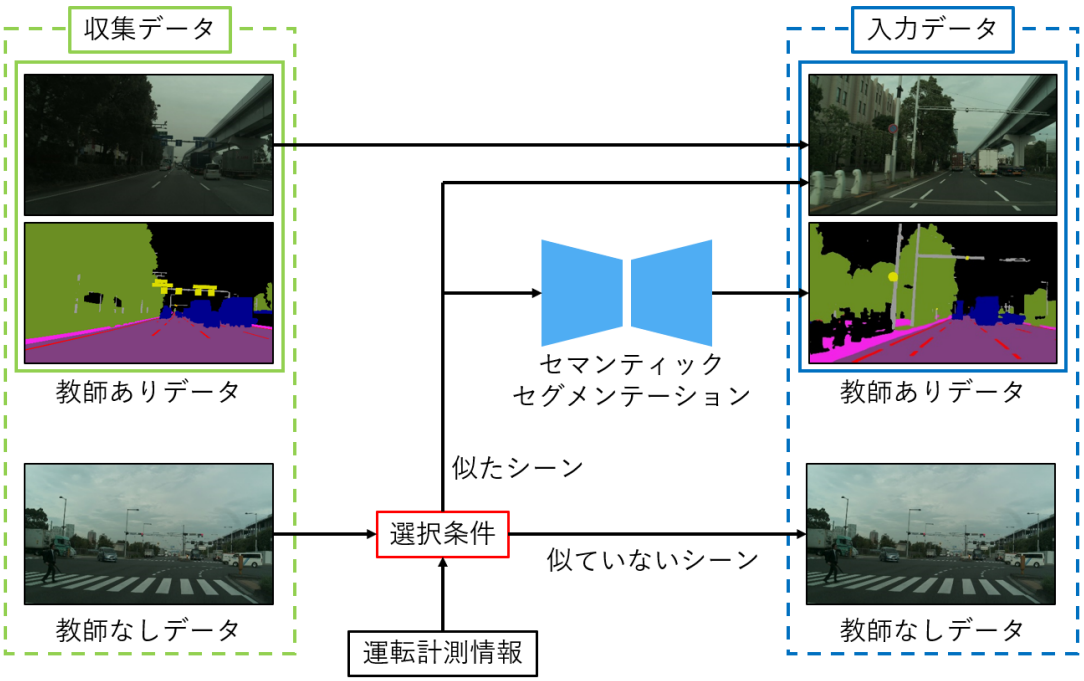

図 2　データの選択と擬似ラベルの付与

く異なるシーンのデータに対する擬似ラベルは精度が低く，学習に最適ではない．そのため，データ撮影時に同時に獲得した速度などの計測情報を用いて，学習データに似たシーンを選択する．選択された教師なしデータに擬似ラベルを付与し，教師あり学習に組み込むことで，精度の高い擬似ラベルを使用し，精度向上を目的とする．

3.1　計測情報によるデータ選択と擬似ラベルの付与

そこで提案手法では，図 2 に示すように，選択条件により教師ありデータと類似するシーンの教師なしデータを選択する．選択条件は，データ撮影時に獲得した計測情報をもとに，位置と速度から以下のように定義する．

位置情報: 教師なしデータと，全ての教師ありデータの撮影位置の距離の最小値を位置情報データとして扱う．教師ありデータと位置情報の近いデータを選択することによって，教師ありデータに似たシーンのデータを用いた学習が期待できる．

速度情報: 一定の範囲の速度で走行しているデータを選択する．教師ありデータには直線の道路を走行中のシーンが多いため，速度を基準にすることによって似たデータを選択することが期待できる．

上記の条件により選択されないデータは教師なし学習に用いる．

3.2　損失関数

半教師あり学習手法として，CCT を用いる．CCT は，教師あり学習と教師なし学習を同時に行う．教師あり学習は，過学習を抑制するために確率によって損失を 0 にする annealed bootstrapped-Cross Entropy (ab-CE) 損失 L_s を用いて学習を行う．ab-CE 損失は式 (1) により求める．

$$L_s = \frac{1}{|D_s|} \sum_{x_i^s, y_i \in D_s} \{f(x_i^s) < \eta\}_1 H(y_i, f(x_i^s)) \quad (1)$$

このとき D_s は教師ありデータ集合，x^s は教師ありデータ，y はラベルデータ，η はしきい値，f はネットワーク出力，H は Cross-Entropy 損失である．

教師なし学習は，入力データ x_i^u から獲得した特徴マップ z_i に対するセグメンテーション結果 $g(z_i)$ と摂動を加えたものに対する結果 $g_a^k(z_i)$ の平均二乗誤差を基に学習する．

教師なし学習における損失 L_u は式 (2) となる．

$$L_u = \frac{1}{|D_u|} \frac{1}{K} \sum_{x_i^u \in D_u} \sum_{k=1}^{K} d(g(z_i), g_a^k(z_i)) \quad (2)$$

このとき D_u は教師なしデータ集合，K は摂動を適用した補助デコーダの数，d は平均二乗誤差である．

L_s と L_u の総和を複合損失とする．複合損失は式 (3) により求める．このとき，L_u には重み ω_u を与える．

$$L = L_s + \omega_u L_u \quad (3)$$

3.3　摂動関数

エンコーダ出力 z に適用する摂動は 7 種類あり，特徴に基づく摂動，予測に基づく摂動，ランダムな摂動の 3 つに分類できる．

3.3.1 特徴に基づく摂動

エンコーダ出力 z に対しノイズを付与するか，その一部を削除する処理を行う．

F-Noise: z と同じサイズのノイズテンソル $N \sim \mathcal{U}(-0.3, 0.3)$ を一様にサンプリングする．その振幅を z と掛け合わせて調整した後，z に対しノイズを付与する．関数は式 (4) となる．

$$\tilde{z} = (z \odot N) + z \tag{4}$$

F-Drop: z と同じサイズのしきい値 $\gamma \sim \mathcal{U}(0.6, 0.9)$ を一様にサンプリングする．特徴マップのチャンネル方向に対する総和を行い，z を正規化して z' を得た後，マスク $M_{drop}\{z' < \gamma\}_1$ を生成し，式 (5) のように適用する．これにより，特徴マップのアクティブな領域の 10% から 40% をマスクする．

$$\tilde{z} = z \odot M_{drop} \tag{5}$$

3.3.2 予測に基づく摂動

メインデコーダの予測値，または補助デコーダの予測値に基づいて摂動を加える．

Guided Masking: 複雑なシーンの理解にはコンテキストが重要であることを仮定すると，ネットワークがコンテキストに依存しすぎている可能性がある．これを制限するために，検出されたオブジェクトとコンテキストをマスキングする 2 つの摂動関数を使用する．メインデコーダの予測値から，検出された前景オブジェクトをマスクするオブジェクトマスク M_{obj} と，コンテキストマスク $M_{con} = 1 - M_{obj}$ を生成し，それぞれ z に式 (6) と式 (7) のように適用する．

$$\tilde{z} = z \odot M_{obj} \tag{6}$$

$$\tilde{z} = z \odot M_{con} \tag{7}$$

Guided Cutout: オブジェクトの特定部分への依存度を減らすために摂動を加える．メインデコーダ出力から各オブジェクトのバウンディングボックスを見つける．対応する特徴マップから，各オブジェクトのバウンディングボックス内のランダムな位置をクロップし，ゼロ埋めをする．

Intermediate VAT (I-VAT): 出力分布を滑らかにするために．Virtual Adversarial Training (VAT) [9] を使用する．与えられた補助デコーダに対し，その予測を最も変化させるであろう敵対的摂動 r_{adv} を発見する．それを式 (8) のようにノイズとしてエンコーダ出力に付与することにより，摂動を適用する．

$$\tilde{z} = r_{adv} + z \tag{8}$$

3.3.3 ランダムな摂動

Spatial dropout [10] をランダムな摂動として適用する．これは画素単位で判定を行う通常のドロップアウトとは違い，特徴マップ単位での判定を行い，特定のクラスの特徴マップをドロップアウトさせる処理を行う．

4 評価実験

評価実験では，各データ選択の手法による精度を検証する．

4.1 実験概要

学習及び評価には，東京臨海部を走行して収集したデータを用いる．教師ありデータは 380 枚，教師なしデータは 5747 枚である．教師ありデータは学習用データ 280 枚と検証用データ 100 枚に分割して使用する．運転計測情報は，画像データと一対一対応するデータではないため，画像データのタイムスタンプに近い時刻の運転計測情報を自己位置情報として扱う．

ベースラインはデータの選択をしない場合とし，各選択手法との比較を行う．CCT の補助デコーダの数は，従来手法 [4] に従う．Object Mask と Context Mask は確率性がないため 2 個，I-VAT は計算コストが高いため 2 個，それ以外は 6 個とし，合計 30 個の補助デコーダを使用する．

評価指標には mIoU を用いる．

4.2 実験結果

各条件における最高精度と，ベースラインを比較する．定量評価を表 2，クラス精度を表 1，定性的評価を図 3 に示す．表 2 より，速度情報でデータ選択した提案手法の精度が 51.2%と最も高く，ベースラインと比較して 3.3pt 向上した．教師ありデータは，走行中のデータが多いため，速度情報による定速走行時の選択によって似たデータを選択でき，高精度になったと考えられる．クラス精度より，いずれの手法でも全体的に精度が向上していることがわかる．また，位置情報では人や電柱など，道路上以外の精度が大きく向上し，速度では歩道や白線，車など道路付近の精度が向上していることがわかる．定性的評価より，ベースラインでは認識できていなかった電柱などが認識できていることがわかる．

4.2.1 位置情報における比較

位置情報を用いた実験では，最も近くの教師ありデータとの距離を基準に選択し，擬似ラベルを付与する．表 3 に結果を示す．教師ありデータとの距離が 0m の時の精度が 49.18%と最も高くなった．これは，教師ありデータと全く同じ位置で撮影されたことにより，全く

表 1　クラス IoU

選択条件		Road	Sidewalk	Whiteline	Vegetation	Car	Person	Pole	Traffic sign	Traffic light
選択なし		**84.4**	**47.0**	31.4	81.5	58.5	11.1	37.0	42.6	27.4
提案手法	位置情報	79.1	45.9	38.3	79.5	**65.7**	**27.2**	**40.9**	51.6	0.0
	速度情報	78.9	44.7	**39.6**	**83.2**	63.9	20.8	40.8	**52.0**	**37.0**

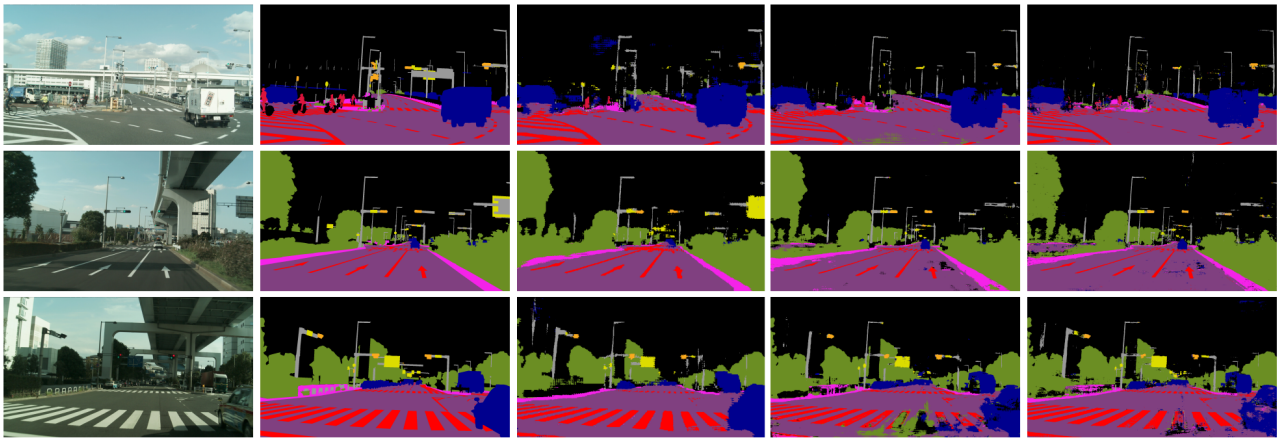

図 3　定性的評価

同じシーンの擬似ラベルを学習に使用できたためだと思われる．

4.2.2　速度情報における比較

速度による実験では，教師なしデータを 10km/h 毎に分割し，擬似ラベルを付与する．表 4 に結果を示す．選択されたデータの撮影時の速度が時速 50km 以上 60km 未満の時の精度が 51.21%と最も高くなった．これは，定速走行時は直線の道路を走っているシーンが多く，似たシーンをうまく選択できたためだと思われる．そのため，時速 60km 以上の場合でも同様に，他の条件より精度が向上していることが確認できる．時速 0km の場合は，位置情報と同様の理由で精度が向上していると考えられる．

5　おわりに

本研究では，擬似ラベルを用いた半教師あり学習法を提案した．評価実験により，提案手法を用いることで，擬似ラベルを用いない場合と比較し精度が向上した．また，データ選択のために用いる計測情報を比較し，速度情報を用いることで，擬似ラベルを用いない場合と比較して 3.3pt の精度向上を確認した．今後の予定としては，より高精度な結果を得ることができるデータ選択方法の検討などが挙げられる．

6　謝辞

本研究は，総合科学技術・イノベーション会議の戦略的イノベーション創造プログラム (SIP) 第 2 期/自動運転 (システムとサービスの拡張)「自動運転技術 (レベル 3，4) に必要な認識技術等に関する研究」(管理法人：NEDO) によって実施されました．

参考文献

[1] S. Laine and T. Aila: “Temporal ensembling for semi-supervised learning”, International Conference on Learning Representations (ICLR) (2017).

[2] Y. Grandvalet and Y. Bengio: “Semi-supervised learning by entropy minimization”, Advances in Neural Information Processing Systems, Vol. 17, pp. 529–536 (2005).

[3] D.-H. Lee: “Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks”, Workshop on Challenges in Representation Learning, Vol. 3 (2013).

[4] Y. Ouali, C. Hudelot and M. Tami: “Semi-supervised semantic segmentation with cross-consistency training”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020).

[5] K. HE, X. Zhang, S. Ren and J. Sun: “Deep residual learning for image recognition”, IEEE Confer-

表 2　実験結果

選択条件		教師ありデータ数		教師なしデータ数	mIoU[%]
		実ラベル	擬似ラベル		
-		280	0	5,747	47.9
全て		280	5747	0	47.8
提案	位置情報	280	616	5,131	47.5
手法	速度	280	291	5,456	**51.2**

表 3　位置情報を用いた実験

選択条件	教師ありデータ数		教師なしデータ数	mIoU[%]
[m]	実ラベル	擬似ラベル		
0	280	99	5,648	**49.18**
0〜9	280	1,508	4,239	41.82
9〜18	280	923	4,824	42.38
18〜27	280	616	5,131	47.58
27〜36	280	360	5,387	44.34
36〜45	280	616	5,131	42.64
45〜54	280	206	5,541	47.92

ence on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016).

[6] H. Zhao, J. Shi, X. Qi, X. Wang and J. Jia: “Pyramid scene parsing network”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6230–6239 (2017).

[7] H. Judy, D. Wang, F. Yu and T. Darrell: “Fcns in the wild: Pixel-level adversarial and constraint-based adaptation” (2016).

[8] J. Ahn and S. Kwak: “Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation”, IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4981–4990 (2018).

[9] T. Miyato, S. Maeda, M. Koyama and S. Ishii: “Virtual adversarial training: A regularization method for supervised and semi-supervised learning”, IEEE Transactions on Pattern Analysis and Machine Intelligence, **41**, 8, pp. 1979–1993 (2019).

[10] J. Tompson, R. Goroshin, A. Jain, Y. LeCun and C. Bregler: “Efficient object localization using convolutional networks”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 648–656 (2015).

表 4　速度情報を用いた実験

選択条件	教師ありデータ数		教師なしデータ数	mIoU[%]
[km/h]	実ラベル	擬似ラベル		
0	280	519	5,228	49.82
0〜10	280	592	5,155	45.14
10〜20	280	796	4,951	47.59
20〜30	280	870	4,877	44.58
30〜40	280	1,052	4,695	44.59
40〜50	280	1,377	4,370	42.97
50〜60	280	291	5,456	**51.21**
60〜	280	250	5,497	49.94