マルチドメインに対応した
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

中村泰智† 平川翼† 山下隆義† 藤吉弘亘†
†中部大学
E-mail: taichi@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，入力画像の画素単位でクラス分類を行うタスクである．自動運転のセマンティックセグメンテーションを用いる場合，都市部や山間部などの様々な環境において高い識別精度が要求される．そのため，様々な環境で取得した膨大な車載画像を学習する必要がある．車載画像データは，車載カメラの普及により容易に収集することが可能となった．しかしながら，画素単位でラベル付けするアノテーション作業が必要であるため，ラベルありデータの作成に多くの時間を必要となる．

また，学習データセットと評価するデータセットが異なるシーンやカメラの位置の違いによるドメインの変化によって認識精度が著しく低下する問題がある．そのため，様々な地域で運用される自動運転にセマンティックセグメンテーションを用いる場合，地域ごとのデータで学習したモデルを複数用意する必要がある．これにより，メモリコストの増加や使用するモデルを選択する機構が必要となるなどの問題がある．そこで本研究では，マルチドメインに対応した出力が可能であるマルチヘッド構造を導入することで単一のモデルで学習可能になり，少量のラベルありデータと大量のラベルなしデータを用いた学習方法である半教師あり学習を導入することで，アノテーションコストの削減を目指す．

2 関連手法

2.1 マルチドメイン学習

マルチドメイン学習は，複数のドメインのデータセットを同時に学習する手法である．一般的なセグメンテーションネットワークでは, クラス数が異なるデータセットを単一のモデルで学習することは難しい．そこで，主なマルチドメイン学習手法として [1][2][3] では，ドメイン固有の畳み込み層を導入して各ドメインを学習する方法がある．中でも MSeg[4] では，異なるドメインを持つ複数のデータセットをまとめた複合データセットを提案した．複合データセットは，異なるデータセットを共通のラベルに変換して作成されている．そのため，ラベルの再定義が必要となり，コストがかかる問題がある．そこで，Multi-Head Model[5] は全てのドメインで共通する共有ネットワークと各ドメインごとの出力を行うマルチヘッド構造を提案した．各ドメインごとに出力ヘッドを用意することで，異なるラベルを持つデータセットを再定義を必要なく同時に学習できる．しかし，文献 [5] は教師あり学習のため，すべてのドメインに対する大量のラベルあり学習データが必要である．

2.2 セマンティックセグメンテーションにおける半教師学習

半教師あり学習とは，大量のラベルなしデータを用いることで，少量のラベルありデータでも高い精度を得る手法である．セマンティックセグメンテーションを対象とした主な半教師あり学習法として擬似ラベリング [6][7][8] がある．擬似ラベリングは，ラベルありデータを用いた教師あり学習を行ったモデルを用いて教師ラベルの持たない画像に対して，擬似ラベルを付与することで精度の向上を目指す手法である．中でも U2PL[9] は，正確性が低い擬似ラベルも付与することを提案した．これは正確性が低い擬似ラベルも十分な数の擬似ラベルを与えることで精度が向上する．CAC[10] は，画像のピクセル単位で対照学習を行うことを提案した．対照学習では，異なる 2 つの画像ペアから同じクラスに属するペアは近く，異なるクラスに属するペアは離れた特徴空間になるように学習する手法である．CAC では学習時に 1 枚の正解なし学習データから重複領域が存在するランダムなパッチを 2 枚を抽出する．2 枚のパッチからピクセル単位で対照学習を行うことでラベルなし学習データの整合性を保つことができ，環境の変化に対するロバスト性が向上する．このように半教師あり学習はラベルありデータが少量で良いためアノテーションコストの削減が可能である．

3 提案手法

本研究では，各ドメインに対応した出力を行うためのマルチヘッド構造に，アノテーションコストを削減のために少量のラベルありデータ画像と大量のラベル

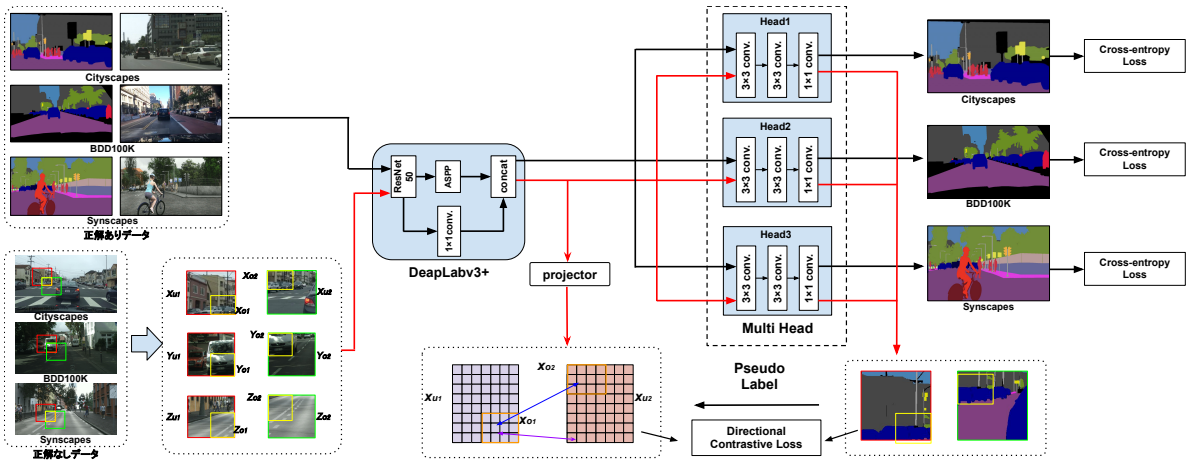

図 1　提案手法のネットワーク構造

なし画像による半教師あり学習を導入したセマンティックセグメンテーション手法を提案する．

3.1　ネットワーク構造

提案手法のネットワーク構造を図 1 に示す. 本手法の共有ネットワークは，単一のモデルで複数ドメイン情報を共有する部分である．この共有ネットワークには，ResNet50[11] を Backbone にした DeepLabv3+[12] を用いる．DeepLabv3+は，Atrous Spatial Pyramid Pooling (ASPP) を採用した Encoder Decoder 構造のネットワークである．ASPP は，異なる間隔の畳み込み処理を並列で行い，特徴マップを統合することで，広範囲のコンテキストを獲得できる．

3.2　半教師あり学習

半教師あり学習は，周囲のピクセル情報に着目した教師あり学習と教師なし学習を組み合わせて学習を行う方法である．教師あり学習の損失関数は式 (1) の Cross-entropy Loss をデータセットごとに用いる．このとき，ネットワークの出力は q，教師ラベルは p，クラス数は N である．

$$L_{ce} = -\sum_{i=1}^{N} p(i) log(q(i)) \tag{1}$$

教師なし学習の損失関数は式 (2) の Directional Contrastive Loss（DCLoss）をデータセットごとに用いる．まず，ラベルなしデータから重複領域を含む 2 つのパッチをランダムに切り出して，モデルに入力し各画素に対するクラス分類を行う．そして，重複領域に着目しクラス分類の信頼度が低い特徴を信頼度が高い特徴に近づくように学習を行う．特徴の信頼度は入力画像に対して回転や拡大縮小などの変換を加えた場合，同じ物体でもピクセルの位置や周辺領域の情報が異なる場合があるため，特徴の信頼度が低下する可能性がある．その時，正確な擬似ラベルの生成が困難になるため，信頼度の高い特徴が必要である．しかし，信頼度が高い

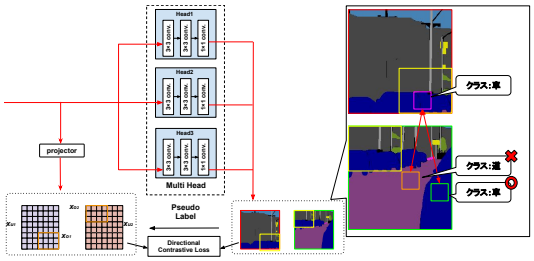

図 2　Directional Contrastive Loss（DCLoss）の構造

特徴に近づくように学習する過程で特徴の信頼度が他の周りの特徴の信頼度に比べ低いことがある．これが精度低下の原因になるため，$M_{d,\,p}$ は式 (3) より閾値 γ よりも低い信頼度を持つ特徴は計算に含めない．また，図 2 に示すように重複領域外のピクセルでは，異なる擬似ラベルを持つ特徴の時は離れるように学習を行うのに対して，道や空などの画像の広範囲を占めるものでは重複領域外のピクセルで同じ擬似ラベルを持つ特徴の場合がある．そこで $M_{n,1}$ は式 (4) より重複領域外で同じ擬似ラベルを持つ特徴では近づくように学習を行う．このとき γ は閾値，$maxC(x_{o1})$，$maxC(x_{o2})$ はラスの確率の最大値，$argmaxC(x_{o1})$，$argmaxC(x_{o2})$ はクラスの確率が最大となるクラスのインデックスを計算，N は重複空間のピクセル数，r は類似度の指数関数，h，w は 2 次元空間の位置，x_{o1}，x_{o2} それぞれ重複空間の 1 枚目，2 枚目である．

$$L_{dc}(x_{o1}, x_{o2}) = -\frac{1}{N}\sum_{h,w} M_{d,p} log \frac{r(x_{o1}, x_{o2})}{r(x_{o1}, x_{o2}) \sum M_{n,1} \cdot r(x_{o1}, x_{u2})} \tag{2}$$

$$M_{d,p} = (maxC(x_{o2}) > (maxC(x_{o1}) \cdot (maxC(x_{o2}) > \gamma) \tag{3}$$

$$M_{n,1} = (argmaxC(x_{o1}) \neq (argmaxC(x_{u2})) \tag{4}$$

表 1　教師ありマルチドメイン学習と提案手法 (半教師ありマルチドメイン学習) の比較

	ラベルあり画像の割合	ラベルなし画像の割合	Cityscaps[%]	Synscapes[%]	BDD[%]	平均 [%]
教師あり学習	$\frac{1}{8}$	0	63.24	73.57	50.18	62.33
提案手法		$\frac{7}{8}$	65.52	77.60	53.07	65.39
教師あり学習	$\frac{1}{4}$	0	66.11	77.15	53.99	65.75
提案手法		$\frac{3}{4}$	66.86	78.11	54.53	66.83
教師あり学習	1	0	75.44	91.85	62.15	76.48

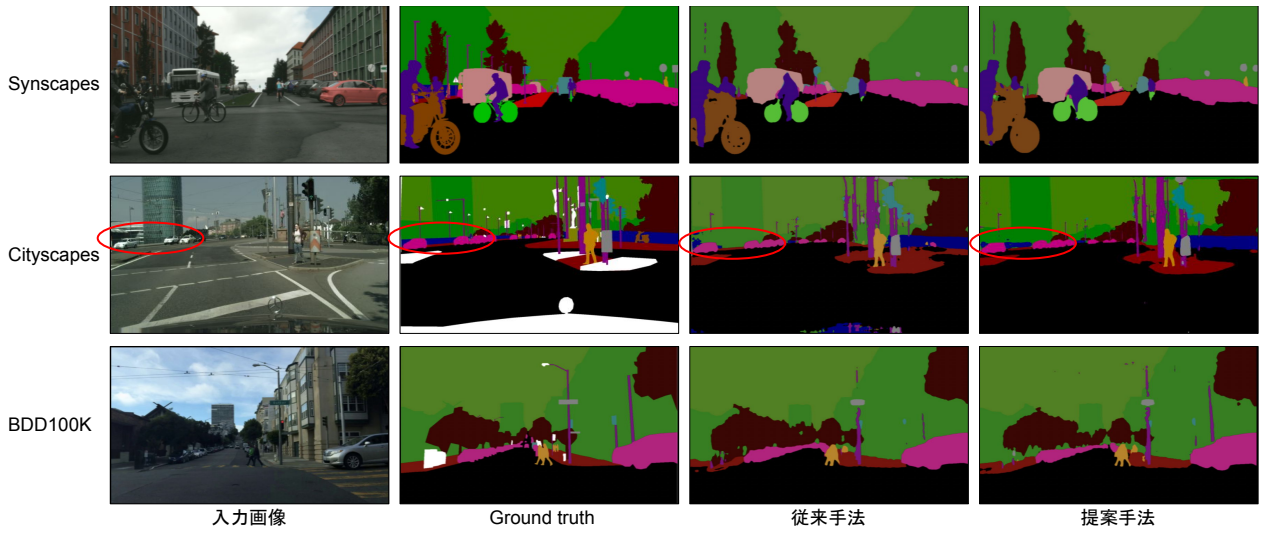

図 3　教師ありマルチドメイン学習と提案手法の定性的評価

表 2　使用データセットの情報

データセット	Cityscapes	BDD	Synscapes
Domain	Driving (Europe)	Driving (USA)	Driving (simulator)
クラス数	19	19	19
学習データ	2,975	7,000	23,000
評価データ	500	1,000	2,000

3.3　マルチヘッド構造

マルチヘッド構造は，共有のネットワークから獲得した特徴マップをデータセットに対応した出力ヘッドごとに入力することでデータセットに対応した出力結果を獲得できる．各出力ヘッドは 2 層の 3 × 3 の畳み込み層と 1 × 1 の畳み込み層で構成されており，各ドメイン情報を使用することで高精度な出力結果を得ることができる．各データセットの Loss ごとに逆伝播を行う場合，データセットの順番によって特定のデータセットにバイアスがかかる可能性がある．そこで，全ての Loss を合計し逆伝播を行う Mix Loss を導入する全てのドメインデータを入力し，各ヘッドで計算した Loss を合計してから逆伝播を行う．Mix Loss を式 (5) に示す．このとき，L_{ce} は CELoss，l_{dc} は DCLoss，x_n は各データセットの入力サンプル，N はデータセットの数，λ は DCLoss に対しての重みである．

$$L = \sum_{n=1}^{N} L_{ce}(x_n) + \lambda L_{dc}(x_n) \tag{5}$$

4　評価実験

本章では，複数のデータセットを用いて提案手法の有効性を示す．実験には，同一クラスを持つ 3 つのデータセットで実験を行う．ネットワークは，ResNet-50 を backbone にした DeepLabv3+を使用する．最適化手法は SGD を使用し，重み減衰 0.0001，モーメンタムを 0.9，初期学習率を 0.01 とする．ラベルありデータの画像サイズをを 0.5 から 2 倍の倍率の間でランダムに変換し，720 × 720 にリサイズしてから使用する．教師なし損失重み λ は 0.1 に設定し，全てのモデルを各データセットにおいて 150 エポック学習する．評価指標には mIoU を用いる．

4.1　データセット

使用するデータセットのクラス数，データ枚数，ドメイン情報を表 2 に示す．Cityscapes[13] はドイツの 50 都市で撮影されたデータセットである．BDD[14] データセットはアメリカの様々な時間や天候で撮影された車載画像データセットである．Synscapes[15] はフォトリアリスティックレンダリング技術を使って作成され

表 3　半教師ありシングルドメイン学習と提案手法の比較

	ラベルあり画像の割合	Cityscaps[%]	Synscapes[%]	BDD[%]	平均 [%]
シングルドメイン学習	$\frac{1}{8}$	63.85	81.45	43.85	63.05
提案手法		65.52	77.60	53.07	65.39
シングルドメイン学習	$\frac{1}{4}$	61.65	**82.40**	43.51	62.51
提案手法		**66.86**	78.11	**54.53**	**66.83**

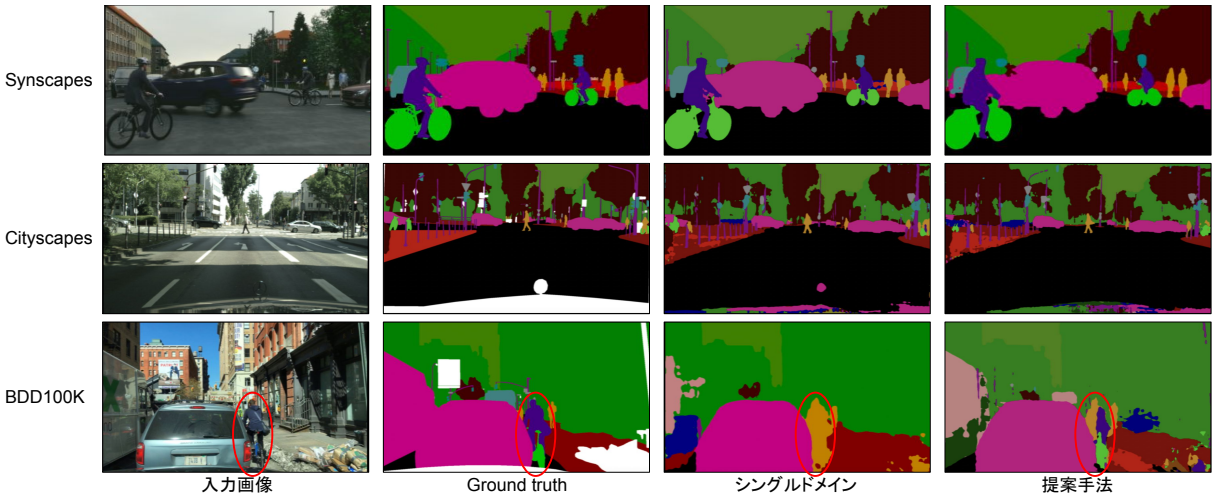

図 4　半教師ありシングルドメイン学習と提案手法の定性的評価

たストリートシーン解析用の合成データセットである．Cityscapes，BDD，Synscapes データセットは同一の 19 クラスから構成されている．

5　実験結果

定量的評価を表 1，表 3，定性的評価を図 3，図 4 に示す．表 1 は従来手法（教師ありマルチドメイン学習）と提案手法（半教師ありマルチドメイン学習）の比較であり，図 3 はこの定性的評価である．表 3 はシングルドメインとマルチドメインの比較であり，図 4 はこの定性的評価である．

5.1　教師ありマルチドメイン学習との比較

表 1 より，教師ありマルチドメイン学習と提案手法（半教師ありマルチドメイン学習）の比較ではラベルありデータの割合が $\frac{1}{8}$ の場合，半教師あり学習に比べ半教師あり学習よりも mIoU が Cityscapes は 2.28 ポイント，Synscapes は 4.03 ポイント，BDD は 2.89 ポイント向上し，平均は 3.06 ポイント向上した．また，ラベルありデータの割合が $\frac{1}{4}$ の場合，表 1 も同様の傾向となった．しかし，ラベルありデータの割合が 1 の場合では半教師あり学習比べ全てのデータセットで高精度であることが確認できる．そして，定性的評価である図 3 では，Cityscapes データセットの赤丸で示すようにフェンスが正解クラスですが，教師ありマルチドメイン学習では柵を認識できていないことに対して提案手法では柵を認識できていることが確認できる．これらのことから，ラベルなし画像を使用することにより精度が向上し，ラベルあり画像が少ないほど精度の向上を確認することができる．

5.2　半教師ありシングルドメイン学習との比較

表 3 より，半教師ありシングルドメイン学習と提案手法（半教師ありマルチドメイン学習）の比較ではラベルあり画像の割合が $\frac{1}{8}$ の場合において，マルチドメイン学習はシングルドメイン学習に比べ Synscapes の mIoU が 3.85 ポイント低下しているが，Cityscapes の mIoU は 1.67 ポイント，BDD の mIoU は 9.22 ポイント向上し，全体の平均の mIoU は，2.30 ポイント向上した．ラベルあり画像の割合が $\frac{1}{4}$ の場合も同様に Synscapes データセットでは精度が低下し，Cityscapes,BDD データセットでは精度が向上した．なかでも，ラベルあり画像の割合 $\frac{1}{8}$ の BDD100K データセットを選択した際に識別精度がシングルドメインと比較して mIoU が 9.22 ポイント向上した．また，定性的評価である図 4 では，BDD データセットの赤丸で示すようにライダーとバイクが正解クラスですが，半教師ありシングルドメイン学習では 人と誤認識しています，一方，提案手法ではライダーとバイクのように正しく認識していることが確認できる．これらのことから，Cityscapes と BDD は実環境データセットに対し，Synscapes は CG データセットであるため，CG 環境と実環境とのドメイン差が原因だと考えられる．そのためマルチドメインで学習

するデータセットの種類を検討する必要がある．

6 おわりに

本研究では，マルチドメインに対応した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手法を提案した．アノテーションコストの削減のため，ラベルなし画像に擬似ラベリングを使用した半教師あり学習にデータセットごとに出力 Head を用意するマルチヘッド構造を導入することで，異なるクラスを持つデータセットに対しても単一のモデルで学習することを可能とした．実験ではラベルあり画像が少ない場合では半教師あり学習が教師あり学習に比べ全てのデータセットで高い認識精度を達成し，マルチヘッド構造はシングルドメイン構造と比べ，Synscapes データセットでは精度が低下し，Cityscapes,BDD データセットでは精度が向上した．今後は，実環境とのドメイン差を考えてのデータセットの選択や新たな学習方法を考案し，より高い精度を得られる条件を検討する．

参考文献

[1] M. Joshi, M. Dredze, W. W. Cohen and C. Rosé: “Multi-domain learning: When do domains matter?”, Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 1302–1312 (2012).

[2] H. Nam and B. Han: “Learning multi-domain convolutional neural networks for visual tracking”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4293–4302 (2016).

[3] T. Kim, M. Cha, H. Kim, J. K. Lee and J. Kim: “Learning to discover cross-domain relations with generative adversarial networks”, Proceedings of the 34th International Conference on Machine Learning, Vol. 70 of Proceedings of Machine Learning Research, PMLR, pp. 1857–1865 (2017).

[4] J. Lambert, Z. Liu, O. Sener, J. Hays and V. Koltun: “Mseg: A composite dataset for multi-domain semantic segmentation”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2879–2888 (2020).

[5] S. Masaki, T. Hirakawa, T. Yamashita1, H. Fujiyoshi：“Multi-domain semantic-segmentation using multi-head model”, ITSC (2021).

[6] Z. Feng, Q. Zhou, Q. Gu, X. Tan, G. Cheng, X. Lu, J. Shi and L. Ma: “Dmt: Dynamic mutual training for semi-supervised learning”, Pattern Recognition, **130**, p. 108777 (2022).

[7] D. Guan, J. Huang, A. Xiao and S. Lu: “Unbiased subclass regularization for semi-supervised semantic segmentation”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9968–9978 (2022).

[8] S. Liu, S. Zhi, E. Johns and A. J. Davison: “Bootstrapping semantic segmentation with regional contrast” (2022).

[9] Y. Wang, H. Wang, Y. Shen, J. Fei, W. Li, G. Jin, L. Wu, R. Zhao and X. Le: “Semi-supervised semantic segmentation using unreliable pseudo labels” (2022).

[10] X. Lai, Z. Tian, L. Jiang, S. Liu, H. Zhao, L. Wang, J. Jia：“Semi-supervised semantic segmentation with directional context-aware consistency”, CVPR (2021).

[11] K. He, X. Zhang, S. Ren and J. Sun: “Deep residual learning for image recognition”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016).

[12] L. Chen, Y. Zhu, G. Papandreou, F. Schroff and H. Adam: “Encoder-decoder with atrous separable convolution for semantic image segmentation”, Proceedings of the ECCV, pp. 833–851 (2018).

[13] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth and B. Schiele: “The cityscapes dataset for semantic urban scene understanding”, pp. 3213–3223 (2016).

[14] F. Yu, H. Chen, X. Wang, W. Xian, Y. Chen, F. Liu, V. Madhavan and T. Darrell: “Bdd100k: A diverse driving dataset for heterogeneous multitask learning”, pp. 2636–2645 (2020).

[15] M. Wrenninge and J. Unger: “Synscapes: A photorealistic synthetic dataset for street scene parsing”, arXiv, **1810.08705**, (2018).