Multi Head 構造を導入した マルチドメイン・セマンティックセグメンテーション

正木翔大† 平川翼† 山下隆義† 藤吉弘亘†
†中部大学
E-mail: masaki@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，ピクセルレベルでの識別を行うため，オブジェクトの種類だけでなく物体の位置，形状も認識することができる．一方で，セマンティックセグメンテーションは，学習と異なるシーンやカメラの位置などドメインの変化によって認識精度が著しく低下する．そのため，様々な地域で運用される自動運転システム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が複数必要となる．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，異なるドメインのデータセットを同時に学習するために，Domain Attention Module と Multi Head 構造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．本手法では，エンコーダ・デコーダ構造のモデルをベースとし，エンコーダおよび，一部のデコーダは全てのドメインで共有する．その際，Domain Attention Module を ResNet に導入することで，単一のモデルでは得られることができないドメイン固有の特徴抽出が可能となる．また，各 Head は，図 1(b) のようにデータセット固有のクラスを出力をする．これにより，図 1(a) の Single Head 構造では学習不可能だった，Cityscapes [1] と Mapillary [2] のように対象とするオブジェクトクラスが異なるデータセットを同時に学習可能となる．学習時には，1 つのデータセットに偏った学習を回避するために，各データセットの損失を同時に逆伝播する Mix Loss を導入する．この単一モデルによって，わずかなパラメータ増加で複数のドメインに対応可能なモデルを学習可能である．複数のデータセットを用いた実験により，提案手法の有効性を実証する．

本論文の貢献は次の通りである．

- セマンティックセグメンテーションにおける Multi Head 構造を採用したマルチドメイン学習手法を提案する．各ドメイン固有の出力ヘッドを用意することで，異なるオブジェクトクラスを持つデータセットも同時に学習可能とする．
- ドメイン情報を共有する Domain Attention Module とデータセットごとの損失を同時に逆伝播する Mix Loss を導入する．これにより，複数のドメインを均等に学習し認識精度を向上させる．
- 複数のデータセットを用いた実験により，同一のオブジェクトクラスを持つデータセットと異なるオブジェクトクラスを持つデータセットの認識性能を測り，マルチドメイン学習が可能なことを示す．

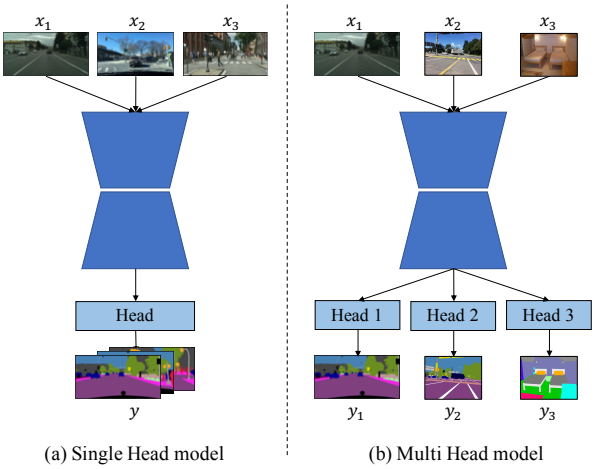

図 1 (a) Single Head モデルおよび (b) Multi Head モデルのネットワーク概要

2 関連研究

Fully Convolutional Network (FCN) [3] の登場によって，CNN を用いたセマンティックセグメンテーションは，高い認識精度を達成し研究が活発に行われるようになっている [4, 5, 6]．その中でもエンコーダ・デコーダ構造を採用した SegNet [7]，U-Net [8] は，省メモリ化に貢献している．また，Dilation convolution [9, 10] は，フィルタのストライドを広くすることで，広い特徴を捉えることが可能である．そのため，多くのセグメンテーションの手法に取り入れられている．一方，PSPNet [11]

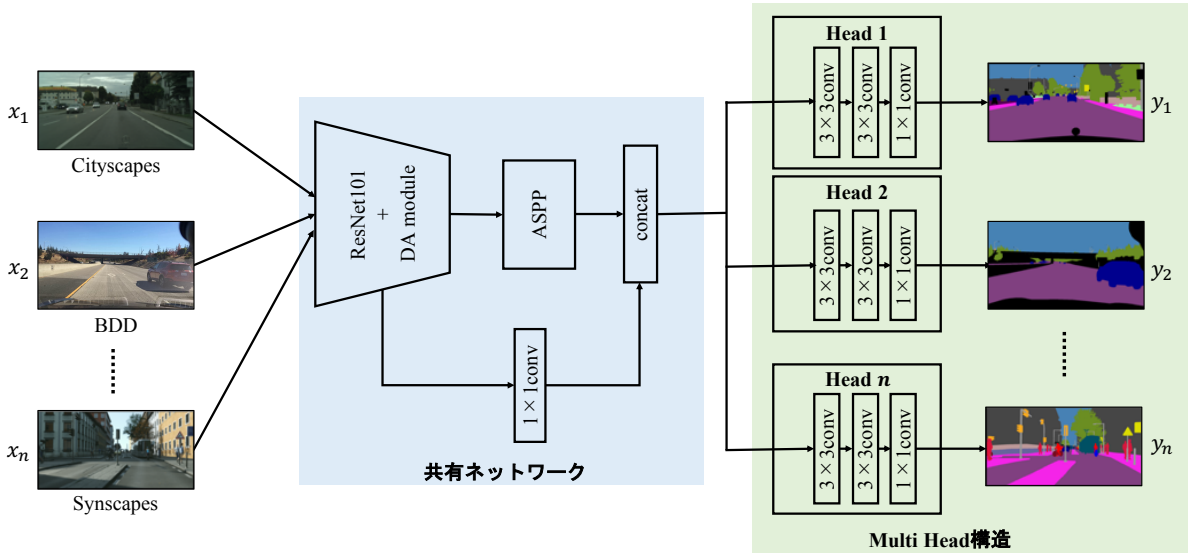

図 2　提案手法のネットワーク構造

や DeepLab [12, 13, 14] では，エンコーダとデコーダの間に Spatial Pyramid Pooling [15] を採用している．これは，特徴マップに対して異なるサイズの Pooling を行うことで，マルチスケールなコンテキストを獲得できる．また，物体認識タスクにおいて使用されている Channel-wise Attention がセマンティックセグメンテーションにおいても活用されている [16, 17, 18, 19]．これにより，特徴マップに対して重要度与えて強調することで，各オブジェクトクラスの認識精度を向上させることができる．

一方，単一モデルで複数のドメインに適応するためにマルチドメイン学習の研究も行われている [20, 21, 22]．多くの手法は，物体認識を対象としている [23, 24]．これらは，ドメイン固有の畳み込み層や BN 層を導入して各ドメインを学習している．物体検出を対象とした手法では，マルチドメイン学習における最適なネットワーク構造を提案している [25]．特に，共有ネットワーク内に各ドメイン固有のパラメータを排除して，ドメイン情報を共有する module を追加することで，複数のドメイン情報を獲得する単一モデルを学習可能としている．また，セマンティックセグメンテーションでは，MSeg [26] のように複数のデータセットをひとまとめにした複合データセットを提案し，異なるドメインを持つデータセットを同時に学習している．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，一部のクラスを再アノテーションや削除，統合を行うため，多くの手間がかかる問題がある．

3　提案手法

本研究では，複数のドメインを同時に学習するために Domain Attention Module と Multi Head 構造を導入したセグメンテーション手法を提案する．提案手法のネットワーク構造を図 2 に示す．本手法のベースネットワークには，ResNet101 [27] を backbone にした DeepLab v3+ [14] を用いる．DeepLab v3 +は，Atrous Spatial Pyramid Pooling (ASPP) を採用したネットワークである．ASPP は，異なる Dilation の畳み込み処理を並列で行い統合することで，マルチスケールな特徴を獲得できる．1 × 1 の畳み込み，Dilation を 6，12，18 に設定した 3 × 3 畳み込み，Global Average Pooling (GAP) を並列に行い，獲得した 5 つの特徴マップを連結して，1 × 1 の畳み込みを行う．また，各オブジェクトの境界周りの認識精度を向上のため，低次元層の特徴マップを利用する．この特徴マップは，Backebone の ResNet の 1 ステージ目の特徴マップを用いて，1 × 1 の畳み込みを行い，ASPP で獲得した特徴マップと連結する．これにより，深い層で不明瞭になるオブジェクトの境界部分の特徴を獲得できる．

3.1　Multi Head Model

一般的なセグメンテーションネットワークは，エンコーダおよびデコーダで獲得した特徴マップを出力 Head に入力することで，クラス数分の確率マップを出力する Single Head 構造である．Single Head 構造は，あらかじめ定義したクラスに対する出力しかできないため，クラス数が異なるデータセットを同時に学習できない．そこで本研究では，クラス数が異なるデータセットを同時に学習するため，Multi Head 構造を採用する．これにより，データセットごとに出力 Head を用意するため，データセット固有のクラスにも対応することが

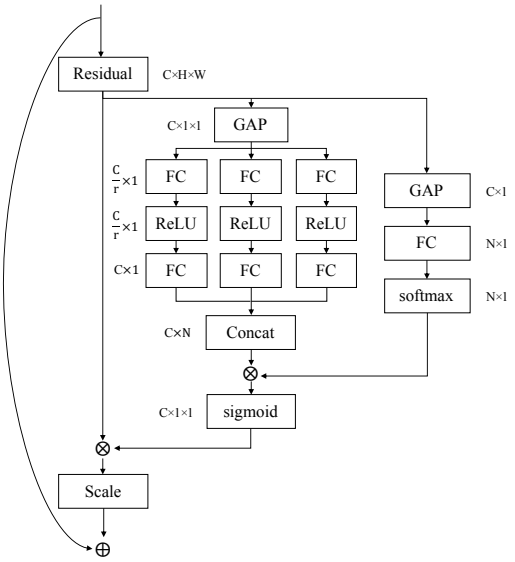

図 3　Domain Attention (DA) Module の構造

できる．共有ネットワークで，ASPP で獲得した特徴マップと ResNet の 1 ステージ目の特徴マップと連結して，データセット固有の出力 Head に入力する．各出力 Head 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，獲得した確率マップを入力サイズに戻すため，バイリニアアップサンプリングを行い出力結果を獲得する．出力結果は，入力画像を共有のネットワークに入力して，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に入力することで獲得する．入力データ x としたとき式 (1) のようになる．

$$y_n = \mathrm{F}_{\mathrm{head}_n}\left(\mathrm{F}_{\mathrm{FE}}\left(x\right)\right) \tag{1}$$

F_{FE} は共有ネットワーク，$\mathrm{F}_{\mathrm{head}_n}$ はデータセット固有の出力 Head である．この構造で用意するデータセット固有のパラメータは，出力 Head のみである．これにより，特徴抽出器は共有となるため，データセット増加によるパラメータ数の増加を抑えることが可能である．

3.2　Domain Attention Module

Domain Attention (DA) Module [25] は，ResNet の Residual Block に適用され，各ドメインの情報を共有しながら学習を行うことで，複数の特徴表現を獲得できる．DA Module の構造を図 3 に示す．DA Module は，SE Adapter と Domain Assignment によって構成される．SE Adapter は，複数の SE Module [28] で構成されており，各 SE Module は各ドメインに特化している．それぞれの出力を連結することで全ドメインの表現空間を形成できる．各 SE Module で入力 x から獲得する重みベクトルは式 (2) で求めることができる．

$$x_{\mathrm{SE}} = \mathrm{F}_{\mathrm{SE}}\left(\mathrm{F}_{\mathrm{avg}}\left(x\right)\right), \tag{2}$$

ここで，$\mathrm{F}_{\mathrm{avg}}$ は Global Average Pooling (GAP)，F_{SE} は，FC+ReLU+FC 層である．Domain Assignment は，GAP と全結合層とソフトマックス層で構成されておりドメインに適応する重みを獲得する．Domain Assignment の重みは，式 (3) により求めることができる．

$$w_{da} = \mathrm{softmax}\left(\mathrm{W}_{\mathrm{DA}}\mathrm{F}_{\mathrm{avg}}\left(x\right)\right), \tag{3}$$

ここで，x は特徴マップ，$\mathrm{F}_{\mathrm{avg}}$ は GAP，W_{DA} はソフトマックス層の重み行列である．GAP の後の全結合層からの出力は，SE Adapter の SE Module の数と同じになる．獲得した重みは SE Adapter の出力と乗算し，Sigmoid 関数で算出する．これにより，SE Adapter からドメインに適応した重みベクトルが獲得できる．

3.3　損失関数

一般的なマルチドメイン学習では，異なるドメインデータを順番に入力して，各 Head でクロスエントロピー誤差を算出し，毎回逆伝播を行う．このとき，データセット毎にパラメータが更新されるため，逆伝播させる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，全てのドメインデータを入力して，各 Head で出力した損失を合計してから逆伝播する Mix Loss を使用する．N 個のデータセットを学習するとき，逆伝播する誤差 L は式 (4) のように求めることができる．

$$L = \sum_{n=1}^{N} L_n \tag{4}$$

全てのドメインの誤差を求めてから逆伝播することで，各ドメインを同時にパラメータ更新して特定のデータセットのみの認識精度向上を防ぐことができる．

3.4　学習方法

各データセットに含まれるデータ数が異なるため，1 epoch あたりの学習枚数が不均等になる．全てのデータセットをバランスよく学習を行うため，学習時にデータ枚数の不均等を回避する必要がある．そのため，1 epoch あたりに用いるデータ数を学習枚数が最も多いデータセットに合わせることでデータ数を調整する．学習時のミニバッチは，同じデータセットのデータのみで構成する．これは，各 Head がドメイン固有のパラメータを持つため，一度の入力に複数のドメインを混在させないためである．学習時は，全データセットのミニバッチを逐次に入力し，誤差を累積して同時に逆伝播する．しかし，学習するデータセット数が増えるにつれて，学習時間と使用メモリが膨大になる．そのため，学習時間とメモリ使用量を削減のために自動混合精度 [29] を用いて学習を行う．

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 つのデータセットを用いた実験，異なるクラス数で構成される

表 1　使用データセットの情報

データセット	Cityscapes	BDD	Synscapes	A2D2	Mapillary	ADE20K
Domain	Driving (Europe)	Driving (USA)	Driving (simulator)	Driving (Europe)	Driving (Worldwide)	Everyday objects
クラス数	19	19	19	18	63	150
学習データ	2,975	7,000	23,000	26,955	18,000	20,210
評価データ	500	1,000	2,000	4,493	2,000	2,000

表 2　Single Domain と提案手法の比較 [%]

Train/Test	Cityscapes	BDD	Synscapes
Cityscapes	77.57	39.81	63.06
BDD	59.05	61.55	55.78
Synscapes	39.04	12.66	**91.55**
提案手法	**78.49**	**62.63**	90.18

3 つのデータセットを用いた実験．5 つのデータセットを用いた実験を行う．各データセットは，水平方向のランダム反転，[0.5, 2.0] の範囲でのランダムスケール，512×512 ピクセルでランダムにクロップして入力する．最適化には，モーメンタムを 0.9，重み減衰を 0.0001 に設定した SGD を用いる．そして，初期学習率を 0.01 に設定し，$(1-\frac{itertotal}{iter})^{0.9}$ を乗算して学習率をスケジューリングする．学習回数は 100 epoch とする．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 1 に示す．使用するデータセットは，車載画像データセットと日常シーンに分けることができる．Cityscapes と A2D2 [30] は同じヨーロッパで撮影されたデータセットだが，Cityscapes は都市で撮影されたデータのみに対して，A2D2 は高速道路や，田舎道といったデータも含まれている．データセット内の画像サイズが異なる Mapillary [2]，ADE20K [31] は，短辺を 720 ピクセルにリサイズする．Cityscapes [1]，BDD [32]，Synscapes [33] は同一の 19 クラスから構成されている．Mapillary は，自車両などが含まれている void カテゴリを除いた 63 クラスのデータを使用し，A2D2 は 18 クラスに再定義したデータを使用する．

4.2　同一クラスを持つデータセットでの結果比較

本実験では，Cityscapes，BDD，Synscapes を対象データセットとして用いる．

Single Domain との比較　表 2 にデータセット単体で学習する Single Domain との比較を示す．学習と評価が異なる Domain 情報の時，全ての場合において精度が大幅に低下していることが確認できる．このことから，セマンティックセグメンテーションが未学習の Domain に対応できないことがわかる．しかし，提案手法では 3 つのデータセットを同時に学習したことにより，Cityscapes と BDD では Single Domain 以上の精度を達成し，Synscapes においても同等の精度を達成した．この結果から，提案手法は Multi Domain 学習に有効であるといえる．

Single Head 構造との比較　Multi Head 構造の有効性を確認するために，Single Head 構造との精度比較を行う．表 3 に Single Head モデル，Mix Loss のみ，Multi Head モデルのみ，Multi Head モデルに Mix Loss と DA Module を導入する提案手法での認識精度の比較を示す．Multi Domain 学習を行うとき，Multi Head 構造でない場合は，BDD は高い精度を達成しているが，Cityscapes と Synscapes は精度が低下していることがわかる．このことから，1 つのデータセットに偏ったモデルとなっているといえる．Mix Loss のみを導入した場合においては，Single Head 構造と同様に BDD データセットのみが高い認識精度を達成している．また，Multi Head 構造のみを適用した場合においては，Cityscapes と Synscapes は高い認識精度を達成しているが，BDD の認識精度が低下している．一方，Mix Loss と Multi Head 構造の両方を取り入れた提案手法では，全てのデータセットがバランスよく学習されていることがわかる．この結果から，同一ラベルを持つデータセットを学習する場合，Mix Loss と Multi Head 構造の導入が有効であるといえる．

DA Module の適用による比較　表 3 より，Multi Head 構造のネットワークに，DA module を適用することで，全てのデータセットで DA module なしの Multi Head 構造ネットワークよりも認識精度が向上した．また，Cityscapes と BDD では，Single Domain で学習したモデルと比較して認識精度が向上した．これは，異なる Domain 情報を DA module によって共有，活用できるためだと考えられる．この結果から，Multi Domain 学習において DA module が有効であるといえる．

パラメータ数の比較　表 4 にパラメータ数の比較を示す．データセット単体で学習した Single Domain のモデルは，パラメータ数がデータセットの数だけ増加する．提案手法では共有のネットワークを使用することで，Single Domain のモデルを複数用意する場合よりもパラメータ数を 57.10% 削減した．Single Head 構造

表 3　同一クラスのデータセットでの精度比較 [%]

	DA module	Multi Head	Mix Loss	Cityscapes	BDD	Synscapes	Mean
Single Domain	-	-	-	77.57	61.55	91.55	76.89
Multi Domain	-	-	-	75.55	63.14	86.91	75.34
	-	-	✓	75.86	**63.33**	88.10	75.76
	-	✓	-	77.30	59.47	**90.20**	75.66
	-	✓	✓	77.92	62.51	90.14	76.86
	✓	✓	✓	**78.49**	62.63	90.18	**77.10**

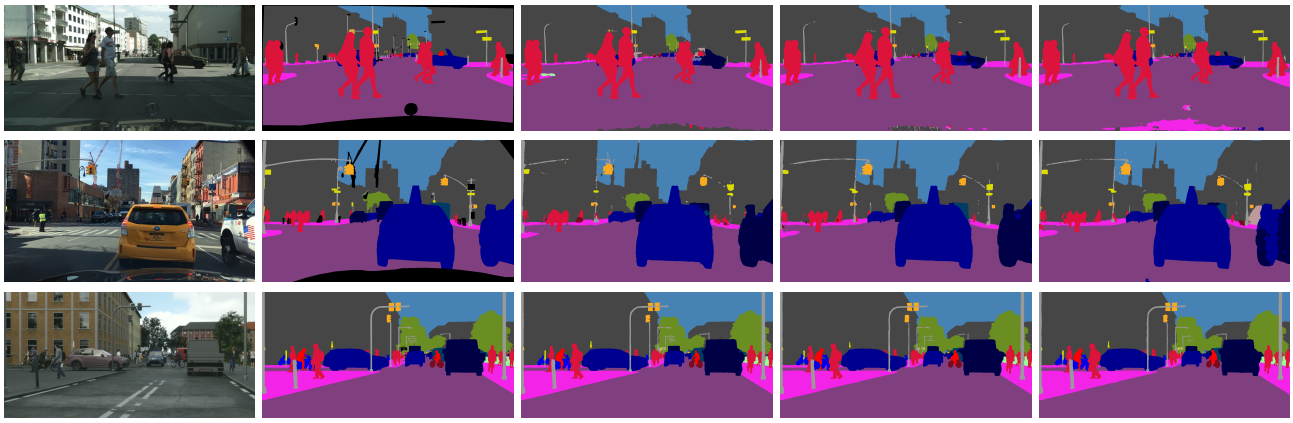

図 4　出力結果の比較

表 4　パラメータ数の比較

	Params.	削減率 [%]
Single Domain	178.02M	-
Single Head	59.34M	66.67
Multi Head	61.94M	65.21
提案手法	76.37M	57.10

からのパラメータ数の増加率をみると，Multi Head 構造の場合 1.04 倍，DA module を適用した場合 1.28 倍のパラメータ数となった．この結果から，Multi Head 構造と DA module は，各データセットで学習したモデルを複数用意する場合よりも，僅かなパラメータ数の増加で複数ドメインを学習できることがわかる．

出力結果の比較　図 4 に出力結果を示す．図 4 は，1 段目が Cityscapes，2 段目が BDD，3 段目が Synscapes の出力結果である．Single Domain の結果では，Cityscapes と BDD で車をトラックと誤識別していることがわかる．一方，複数のデータセットで学習を行った提案手法では，認識結果を修正できていることがわかる．これは，複数データセットを共有して学習したことで，各クラスより多くの特徴を学習できたためだと考えられる．

これらの結果より，同一ラベルで学習を行う場合では，Mix Loss，Multi Head 構造，DA module を採用することでベースの精度以上，または同等の精度を達成することが可能であることを確認した．

4.3　異なるクラスを持つデータセットでの実験

次に，異なるクラス数で構成されているデータセットを同時に学習できるか実験する．実験には，Cityscapes, Mapillary，ADE20K を用いる．また，モデルには同一ラベルでの実験で高精度であった，DA module を適用した Multi Head モデルを使用する．

Single Domain との精度比較を表 5 に示す．クラス数やドメイン情報が異なるデータセットで学習した場合でも，Single Domain と同等の精度であることを確認した．この結果より，提案手法は異なるクラス数で構成されているデータセットでも同時に学習可能であることを確認した．

4.4　5 つのデータセットを使用した比較実験

学習に使用するデータセットの数を 5 つにして実験を行う．実験には，Cityscapes, BDD, Synscapes, A2D2, mapillary を用いる．また，モデルには DA module を適用した Multi Head モデルを使用する．

Single Domain との精度比較を表 6 に示す．学習するデータセットが 5 つの場合，Single Domain よりも Cityscape は 1.94 ポイント，BDD は 4.04 ポイント，Mapillary は 1.25 ポイントの認識精度向上を確認した．また，精度が低下した Synscapes と A2D2 でも Single Domain と同等の認識精度であることを確認した．この

表 5　異なるクラスを持つデータセットでの精度比較 [%]

	Cityscapes	Mapillary	ADE20K	Mean
Single Domain	77.57	43.71	36.42	52.57
提案手法	76.01	43.31	37.16	52.16

表 6　5 つのデータセットでの実験 [%]

	Cityscapes	BDD	Synscapes	A2D2	Mapillary	Mean
Single Domain	77.57	61.55	**91.55**	**78.08**	43.71	70.49
提案手法	**79.51**	**65.59**	89.47	76.56	**44.96**	**71.22**

結果より，提案手法は 5 つのデータセットを学習する場合でも有効であることを確認した．

5 おわりに

本研究では，異なるドメインを同時に学習する Multi Head 構造のセマンティックセグメンテーション手法を提案した．ドメイン情報を共有する DA Module とデータセットごとの損失を同時に逆伝播する Mix Loss を適用し，データセットごとに出力 Head を用意する Multi Head 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，同一のクラスを持つデータで学習した場合において，Single Head 構造よりも高い認識精度を達成した．また，Single Domain で学習した場合と比較しても同等以上の精度を達成した．学習が困難であるクラス数が異なるデータセットを同時に学習する場合でも，Single Domain と同等の精度を達成した．学習するデータセットを 5 つにした実験においても，Single Domain よりも高い認識精度を達成した．今後は，クラス数が異なるデータセットを学習した場合でも，Single Domain の認識精度を超えることや，他のベースネットワークへの適用により汎用性を確認する．

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] 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”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3213–3223 (2016).

[2] G. Neuhold, T. Ollmann, S. Rota Bulò and P. Kontschieder: “The mapillary vistas dataset for semantic understanding of street scenes”, International Conference on Computer Vision (ICCV), pp. 4990–4999 (2017).

[3] J. Long, E. Shelhamer and T. Darrell: “Fully convolutional networks for semantic segmentation”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3431–3440 (2015).

[4] T. Takikawa, D. Acuna, V. Jampani and S. Fidler: “Gated-scnn: Gated shape cnns for semantic segmentation”, Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 5229–5238 (2019).

[5] F. Zhang, Y. Chen, Z. Li, Z. Hong, J. Liu, F. Ma, J. Han and E. Ding: “Acfnet: Attentional class feature network for semantic segmentation”, Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 6798–6807 (2019).

[6] A. Kirillov, Y. Wu, K. He and R. Girshick: “Pointrend: Image segmentation as rendering”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9799–9808 (2020).

[7] V. Badrinarayanan, A. Kendall and R. Cipolla: “Segnet: A deep convolutional encoder-decoder architecture for image segmentation”, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), pp. 2481–2495 (2017).

[8] O. Ronneberger, P. Fischer and T. Brox: “U-net: Convolutional networks for biomedical image segmentation”, Medical Image Computing and Computer-Assisted Intervention (MICCAI),

pp. 234–241 (2015).

[9] F. Yu and V. Koltun: "Multi-scale context aggregation by dilated convolutions", International Conference on Learning Representations (ICLR) (2016).

[10] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille: "Semantic image segmentation with deep convolutional nets and fully connected crfs", International Conference on Learning Representations (ICLR) (2015).

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

[12] L. Chen, G. Papandreou, I. Kokkinos, K. Murphy and A. L. Yuille: "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs", IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), pp. 834–848 (2018).

[13] L.-C. Chen, G. Papandreou, F. Schroff and H. Adam: "Rethinking Atrous Convolution for Semantic Image Segmentation", arXiv: 1706.05587 (2017).

[14] L. Chen, Y. Zhu, G. Papandreou, F. Schroff and H. Adam: "Encoder-decoder with atrous separable convolution for semantic image segmentation", Proceedings of the European Conference on Computer Vision (ECCV), pp. 833–851 (2018).

[15] K. He, X. Zhang, S. Ren and J. Sun: "Spatial pyramid pooling in deep convolutional networks for visual recognition", IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), pp. 1904–1916 (2015).

[16] H. Zhang, H. Zhang, C. Wang and J. Xie: "Co-occurrent features in semantic segmentation", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 548–557 (2019).

[17] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang and H. Lu: "Dual attention network for scene segmentation", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3146–3154 (2019).

[18] Z. Zhong, Z. Q. Lin, R. Bidart, X. Hu, I. B. Daya, Z. Li, W.-S. Zheng, J. Li and A. Wong: "Squeeze-and-attention networks for semantic segmentation", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13065–13074 (2020).

[19] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi and A. Agrawal: "Context encoding for semantic segmentation", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7151–7160 (2018).

[20] 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).

[21] 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).

[22] 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).

[23] S.-A. Rebuffi, H. Bilen and A. Vedaldi: "Learning multiple visual domains with residual adapters", Advances in Neural Information Processing Systems 30 (NIPS).

[24] S.-A. Rebuffi, H. Bilen and A. Vedaldi: "Efficient parametrization of multi-domain deep neural networks", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8119–8127 (2018).

[25] X. Wang, Z. Cai, D. Gao and N. Vasconcelos: "Towards universal object detection by domain attention", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7289–7298 (2019).

[26] 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).

[27] 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).

[28] J. Hu, L. Shen and G. Sun: "Squeeze-and-excitation networks", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141 (2018).

[29] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh and H. Wu: "Mixed precision training", International Conference on Learning Representations (ICLR) (2018).

[30] J. Geyer, Y. Kassahun, M. Mahmudi, X. Ricou, R. Durgesh, A. S. Chung, L. Hauswald, V. H. Pham, M. Mühlegg, S. Dorn, T. Fernandez, M. Jänicke, S. Mirashi, C. Savani, M. Sturm, O. Vorobiov, M. Oelker, S. Garreis and P. Schuberth: "A2D2: Audi Autonomous Driving Dataset", arXiv, **2004.06320**, (2020).

[31] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso and A. Torralba: "Scene parsing through ade20k dataset", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 633–641 (2017).

[32] 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", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2636–2645 (2020).

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