セマンティックセグメンテーションにおけるシミュレーションデータの活用

EP18102 堀川薫斗　　指導教授：山下隆義

1. はじめに

高性能なセマンティックセグメンテーションを実現するには，大量の画像データによる事前学習が必要である．これまで事前学習には，ImageNet などの異なるタスクのデータセットが用いられてきた．本研究では，セマンティックセグメンテーションタスクのためのデータセットを，シミュレーションにより生成し，事前学習に有効か調査する．また，生成したデータと実画像データを組み合わせた学習について調査する．

2. データセット

2.1. 物体認識データセット

代表的な物体認識用データセットである ImageNet は，データ数が約 1,200,000 枚，クラス数が 1,000 のデータセットである．ImageNet を用いて事前学習することで，汎用性のあるモデルを獲得できる．そのため物体認識以外のセマンティックセグメンテーションタスクでも事前学習モデルとして利用されている．

2.2. セグメンテーションデータセット

Cityscapes は，セグメンテーションタスクの学習に用いられる代表的なデータセットである．セマンティックセグメンテーションのデータは画素単位で正解ラベルをアノテーションしなければならないため，データセットの構築コストが高い．そのため，物体認識データセットよりもデータ枚数が少ない．

3. CARLA によるデータセット作成

車載環境での大規模なセマンティックセグメンテーションデータセットの構築を目指して，自動運転シミュレータの CARLA を用いてセグメンテーション用データセットの作成を行う．CARLA はセグメンテーションの正解データを容易に取得できるため，アノテーションコストの削減が可能である．CARLA 上の様々なマップ内を自動運転で周回し，10 秒毎にデータを収集する．画像サイズは，2,048×1,024 ピクセルであり，マップ内の時間と天候は常に変動させて収集する．周回する軌跡と天候の例を図 1 に示す．軌跡上に 50 台の車をランダムに配置し，その中の 1 台の車からの画像を収集する．これにより，様々な風景，天候，時間の要素をもつデータセットを自動で作成する．

図 1：周回する軌跡と天候の例

4. 学習に使用するネットワーク

作成したデータセットを，CNN ベースと Transformer ベースのネットワークで学習する．これにより，本データセットの有効性を 2 つのネットワークで調査する．

4.1. HRNet

CNN ベースの手法として HRNet[1] を使用する．HRNet は，セマンティックセグメンテーションなどのタスクにおいて，高解像度の特徴マップの維持を目的としたネットワークである．低解像度と高解像度の畳み込み処理を並列に行い，高解像度の特徴マップを維持する．また，異なる解像度のネットワーク間で特徴マップの情報を共有する．出力部分では，すべての解像度の特徴マップを結合し畳み込み処理をすることで，マルチスケールな情報を考慮した出力が可能である．

4.2. SegFormer

Transformer ベースの手法として SegFormer[2] を使用する．SegFormer は Transformer エンコーダと MLP デコーダを組み合わせたネットワークである．SegFormer は，階層型 Transformer エンコーダで構成されており，マルチレベルな特徴マップを獲得可能である．MLP デコーダは，MLP 層のみで構成されており，軽量かつシンプルな構造なため，計算量を削減できる．

5. 評価実験

CARLA で作成したデータセット 100,000 枚と ImageNet1,200,000 枚のそれぞれで事前学習を行ったモデルを用いて Cityscapes のファインチューニングを行う．また，作成したデータセットと Cityscapes を合わせたデータセットをファインチューニングに用いる．ネットワークに HRNet と SegFormer を用いて，CNN ベースと Transformer ベースに対する本手法の有効性の調査を行う．

5.1. 実験結果

定量的評価を表 1 に示す．CARLA を事前学習に用いたモデルをファインチューニングした HRNet と SegFormer の精度は，それぞれ 75.11%と 77.36%である．ImageNet を使用した場合と比較して，約 4 ポイント低い精度となった．そのため，事前学習には ImageNet を用いることが効果的であると言える．一方，CARLA と Cityscapes を同時に学習した場合，SegFormer は 79.66%となり，CARLA を学習データに加えることで精度が向上した．

表 1：定量的評価

手法	事前学習済みモデル	データセット	MIoU[%]
HRNet	CARLA	Cityscapes	75.11
	ImageNet	Cityscapes(13 クラス)	**79.52**
		Cityscapes + CARLA	75.38
SegFormer	CARLA	Cityscapes	77.36
	ImageNet	Cityscapes(13 クラス)	79.54
		Cityscapes + CARLA	**79.66**

定性的評価を図 2 に示す．HRNet では，電車を誤認識しているが，SegFormer では正確に認識している．これは，SegFormer がより広い受容野で注目することで，大きい物体に対して正確に認識を行えるためだと考えられる．

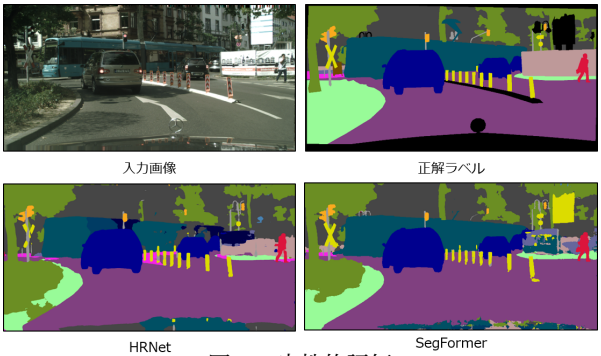

図 2：定性的評価

6. おわりに

本研究では，シミュレーションデータがセマンティックセグメンテーションに有効であるかの調査を行った．事前学習には，少量のデータ数の場合，大幅な精度の向上を見込めないことを確認した．また，ファインチューニングの際にデータセットを組み合わせて学習することで，SegFormer では精度が向上することを確認した．今後は，事前学習に自己教師あり学習を用いることを検討する予定である．

参考文献

[1] J. Wang, *et al.*, “Deep High-Resolution Representation Learning for Visual Recognition”, CVPR, 2019.

[2] E. Xie, *et al.*, “SegFormer：Simple and Efficient Design for Semantic Segmentation with Transformers”, arXiv, 2021.