

役割分担ネットワークによる高効率学習

菅原俊^{†1} 田口賢佑^{†1} 渡邊正人^{†1} 船津陽平^{†1}

概要 : Fully Convolutional Network を用いた Semantic Segmentation はピクセル単位のクラス推定問題であり、十分な量のデータを用いて学習する事で高精度な認識精度が得られる。車載認識に向けた Semantic Segmentation では、高解像度画像に対して、遠方に写る小さな対象物にアノテーションを行う必要がある。その為、十分な量の学習データを作成する為には多くのコストが生じる。本論文では学習に必要なアノテーションコストを削減する為に、高精度な Semantic Segmentation を、汎化性能を獲得するタスクと緻密な Segmentation を獲得するタスクに分割する手法と、タスク毎に Network を分割し、それぞれの Network で異なる品質の教師データを学習する Sub Module Network(SMN)を提案する。提案手法は Cityscapes Dataset を用いた検証実験によって、従来手法と同等の認識精度を得る為に必要なアノテーションコストを下げられる事を示した。

キーワード : Semantic Segmentation, 車載認識カメラ, アノテーションコスト

1. はじめに

Semantic Segmentation はピクセル単位のクラス分類問題であり、ロボティクスや自動運転に応用可能なコンピュータビジョンのタスクである。Semantic Segmentation は Random Forest を用いた手法[1]や Texton 特徴量と CRF を組み合わせた手法[2]等の手法が提案されていたが、2014 年に Fully Convolutional Network[3]が提案され、従来の精度を大幅に塗り替えてから、殆どの手法が CNN ベースの手法になった。最新の Semantic Segmentation の研究では、大規模な公開データセット[4][5]によるベンチマークが行われ、日々新しい手法が提案されている。中でも PSPNet[6]や ResNet[7], RefineNet[8]は高い認識精度を記録しており、応用的な活用が期待されている。

Semantic Segmentation の応用展開先に車載認識カメラがある。車載認識カメラとは Advanced Driving Assistant System(ADAS)や自動運転に向けた走行環境の認識を目的とした認識機能付き車載カメラであり、衝突の危険がある歩行者や車両等の障害物、白線や走行可能領域をリアルタイムに認識する事を目的としている。Semantic Segmentation を車載認識に用いると、ピクセル単位で走行空間を意味別に領域分割する事が可能となり、車両制御や走行計画処理を容易に行う事ができる。

Semantic Segmentation を車載認識カメラに実用化するにあたっていくつかの課題が残されている。中でも計算コスト、ロバスト性、学習コストは特に重要な課題であると我々は考えている。

車載認識カメラは高速道路等の自車両や他車が高速に移動するシーンでの使用や、突然の飛び出しを検出する必要がある為、高フレームレートでの動作が要求される。しかしながら、CNN を用いた Semantic Segmentation は時間、

空間計算量が共に高く、リアルタイム処理を行うことが難しい[10]。また、CNN における時間計算量は一般に入力画像サイズに比例して大きくなる。車載認識カメラでは前方に搭載するカメラにおいて遠方の対象物を認識するケースや広範囲に渡って周辺を認識するケース等、入力画像サイズは大きくなる傾向にあり、ますます計算量は増加する。

車載認識カメラにおけるロバスト性とは、逆光や暗所等の照度変動、気候変動、雨滴や泥、傷等のレンズ変動、雪や路面反射等の走行空間変動等の変動要因に対する耐性を指す。車載認識カメラは搭乗者の安全に直結するシステムであり高いロバスト性が要求されるが、あらゆるシーンでの利用が想定される為、すべての変動に対策を施す事は困難である。

学習コストとは、認識モデルの学習に使用するデータセットの撮影コストやそれらに対するアノテーションコストを指す。車載認識カメラではロバスト性向上の為、多くのシーンでの撮影データが必要になり撮影コストが肥大化する。また Semantic Segmentation のアノテーション作業は、画像分類や Object Detection のアノテーション作業と比較して、作業工程が複雑で時間がかかる。

我々は上記の課題に対し、Semantic Segmentation のタスクを複数の簡易なタスクに分解することで学習効率を向上させる役割分担学習と、専用のネットワーク構造の Sub Module Network (SMN)を提案する。役割分担学習はデータあたりの学習効率が向上する為、学習コストに対する認識モデルのロバスト性が向上する。また、役割分担学習は Semantic Segmentation をアノテーションコストの低いタスクとアノテーションコストの高いタスクに分解するが、アノテーションにかかるリソースを各タスクに適切に配分する事で、総合的なアノテーションコストを低減する事ができる。SMN はタスク毎に Module(小さな Network)が分けれ

^{†1} 京セラ株式会社 研究開発本部 システム研究開発統括部
ソフトウェア研究所 画像処理研究部

れたネットワーク構造であり、Module 毎に異なるアノテーションが付与されたデータセットで学習が行われる。Module は他の Module の出力結果を参照し、学習、推論を行う構造である為、誤認識を行った際、各 Module の出力を参照すれば原因を解析できる可能性がある。また、SMN ではアノテーションコストの低いタスクにおいて、入力画像として高い解像度が必要でないため、Early Down Sampling[10]を行う事ができ高速化を図れる。これにより、計算コスト、ロバスト性、学習コスト課題を解決する事ができる。

2. 関連研究

Fully Convolutional Network[3]が提案されて以来、多くの Semantic Segmentation は CNN ベースの手法になった。FCN は VGG16[12]の全結合層を 1×1 Convolution に変更し、出力層をピクセル毎の Softmax Layer に置き換えたネットワークである。VGG16 は画像分類用に提案されたネットワークであり、効率的に大局的な特徴を捉える為に繰り返し Pooling 処理を行う特徴がある。Pooling 処理を施すと出力の Feature map のサイズは小さくなっていくが、Semantic Segmentation の出力は入力画像と同等の解像度を求められる。その為、縮小した Feature Map をもとの解像度に拡大する Up Sampling が必要である。一般に Pooling 処理を行うことによって、Feature Map から局所的な特徴が抜け落ちる。そこで、FCN では Pooling を行う前の中間層を Skip 接続する事により、Up Sampling する際に局所的な特徴を参照している。SegNet[13] や DeconvNet[14]は Encoder-Decoder 構造のネットワーク構造である。これらのネットワークも FCN と同様に、Up Sampling を行う際に、対応する中間層から拡大に必要な Pooling Index を参照している。また、UNet[19]は Encoder-Decoder 構造に Skip 接続を組み合わせた手法である。PSPNet は ResNet[7]ベースのネットワークに、局所的な特徴と大局的な特徴を同時に捉える Pyramid Convolution Layer が接続されている。これにより、Global な Context を効果的に捉えられる事ができる。

ENet[10]は Real-time な Semantic Segmentation を目指した研究である。ENet は計算効率を高める為にネットワークの浅い層で速やかに Down Sampling を行う手法と、ResNet の Bottleneck Module が Semantic Segmentation に有効である事を示した。ICNet[11]は Real-time 処理と認識性能の両立を目指したネットワークである。マルチスケールに変換した入力画像を、それぞれ異なるブランチに与える事で、効率的に大局的な特徴と局所的な特徴を捉える事ができる。

ERFNet[15]は ENet ベースのネットワークであり、ResNet の Non-bottleneck を用いてネットワークを広さ方向に拡張し、高速、高精度な認識を実現した。ShuffleSeg[16]は Deconvolution を行う際はより高解像度の中間層に Skip 接続を行う事と Depthwise Convolution が有効である事を示した。

また ContextNet[17]は ICNet の様にオリジナルの解像度の画像と低解像度の画像を受け付ける 2 つのブランチで構成されたネットワークである。低解像度の入力を受け付けるブランチは Global な特徴をもう一方のブランチでは局所的な特徴量を獲得する。

3. 提案手法

Semantic Segmentation の学習にかかる総合的なコストを削減する役割分担学習と、それに適したネットワーク構造である Sub Module Network (SMN)を提案する。総合的な学習コストとは、学習データのアノテーションコストやモデルの学習時間を指す。これらのコストを削減することでコストに対するデータ量を増やすことが可能となり、従来と同等の学習コストに対し、より多量のデータを学習する事ができる。それによって、学習データの網羅性が高まり認識モデルのロバスト性が向上する。

また、SMN は高解像度 Module と汎化性 Module の 2 つのブランチで構成されたネットワーク構造である。役割分担学習では Semantic Segmentation をより簡易なタスクの組み合わせに分解する事で学習効率の改善を図っているが、SMN の各 Module がそれぞれ分解されたタスクに対応している。

3.1 役割分担学習

Semantic Segmentation はピクセル単位のクラス分類である。画像単位でクラス分類を行う物体認識に対し、Localization の要素も含まれるのでより難易度が高いタスクである。また、Localization に類似したタスクとして Object Detection がある。Semantic Segmentation は認識対象物の画像上の位置だけでなく、より緻密な領域分割を行う必要がある為、Object Detection に対してより難しいタスクである。この様に Semantic Segmentation は Localization タスクと緻密な領域分割が複合した難易度の高いタスクであると考えられる。一般に Semantic Segmentation の学習は、単一のデータセットを用いて学習を行う。Semantic Segmentation は難易度の高いタスクである為、十分な性能を満たすために、高い品質のアノテーションが必要である。しかしながら、Semantic Segmentation のアノテーションコストは非常に高いため、十分なデータ量を用意するのは難しい。そこで我々は Semantic Segmentation のタスクを Localization タスクと緻密な領域分割タスクに分解する事で、上記の課題を解決する。Localization タスクとは画像中から大まかに認識対象物の位置を求めるタスクである。また、緻密な領域分割タスクとは、入力画像と認識対象物の位置に関する大まかなヒントが与えられている状態から、認識対象の領域を詳細に分割するタスクである。Localization タスクは Semantic Segmentation と比べ緻密な領域分割を行う必要がない為、比較的難易度の低いタスクである。また、緻密な領域分割タスクも Semantic Segmentation と類似したタスクであるが、

認識対象に関する大まかなヒントが与えられている為、比較的難易度の低いタスクになっている。

Semantic Segmentation を Localization タスクと緻密な領域分割タスクに分割する事で、それぞれのタスクに見合った品質のデータを学習させる事が可能になる。Localization のアノテーションは Semantic Segmentation 用のアノテーションより作成効率が高い為、アノテーションコストを同等にした場合、より多くの教師データを作成することができる。また、緻密な領域分割タスクに用いる教師データは、Semantic Segmentation のものと同一である。その為、作成コストが高いが、役割分担学習を行うことで、緻密な領域分割タスクの難易度は相対的に低いものとなり、より少量のデータでも学習ができると考えられる。

3.2 Sub Module Networks

Sub Module Networks (SMN) とは役割分担学習向けのネットワークであり、複数の小さなネットワーク (Module) で構成されている。各 Module はそれぞれ分解したタスクに一对一で紐付いており、Semantic Segmentation タスクにおいては Localization タスクに対応する汎化性 Module と緻密な領域分割に対応する高解像 Module で構成される。ネットワークのパイプラインを図 1 に示す。

SMN に与えられた入力画像は、まず汎化性 Module に与えられる。汎化性 Module では入力画像について Rough に領域分割を行い、Localization 情報を出力する。高解像 Module は入力画像と同時に汎化性 Module の出力も受け取り、汎化性 Module をヒントにしながら緻密な領域分割を行う。

タスク毎に汎化性 Module と高解像 Module が別れている為、それぞれの Module で異なるデータセットを用いて学習を行うことができる。その為、マルチタスク学習の様にデータに対してすべてのアノテーションが付いている必要はなく、汎化性 Module 用のデータセットと高解像 Module 用のデータセットに違いがあっても問題なく学習を行う事ができる。

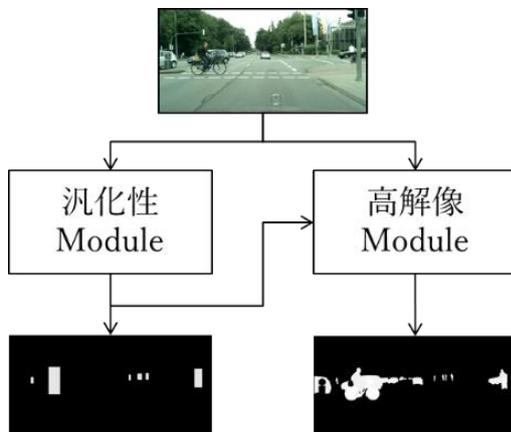


図 1 提案手法のパイプライン

3.3 汎化性 Module

汎化性 Module は入力画像に対して大まかに Semantic

Segmentation を行う。教師データとして作成コストが低く緻密性の低いアノテーションを使用する。汎化性 Module の役割は高解像 Module が緻密な推論をする際のヒントを与える事である。その為、アノテーションコストの低いデータを大量に学習し未検知が生じないようにする。ネットワーク構造は一般的な Encoder-Decoder 構造であり、出力として入力と同解像度のピクセル単位のラベル情報を返す。汎化性 Module で学習する教師データは緻密性が低く、低解像の Feature Map でも学習を行う事ができる。その為、提案手法は, ICNet や ContextNet の様に汎化性 Module の入出力画像を低解像化することができ、計算量を削減できる。また、ネットワークの出力としてピクセル単位のラベル情報を出力するが、これは汎化性 Module の出力を人間の目でも理解できる様にしておくことで、ネットワークの解析可能性が向上する。

3.4 高解像 Module

高解像 Module の役割は、入力画像とそれに対応する Localization に関するヒントを用いて、緻密な Segmentation を推論することである。高解像 Module は大局的な情報として汎化性 Module の出力が得られるので、Down Sampling を繰り返し施す必要がなく、浅いネットワークになる。また、役割分担学習によってタスクが簡易化している。その為、一般的な Semantic Segmentation 用のネットワークより低い表現力でも学習を行う事ができ、ネットワークの計算量を削減することができる。

4. 検証実験

本手法は Cityscapes データセット及び CityPersons データセット[18]を用いて有効性を検証した。

4.1 評価方法

対象物の未検出、誤検出を詳細に解析するために Precision と Recall を用いた。Precision と Recall は歩行者と Void の 2 クラスに関する Confusion Matrix 算出し、式(1), (2)を用いて求めた。

$$Precision = \frac{True\ positive}{True\ positive + False\ positive} \quad (1)$$

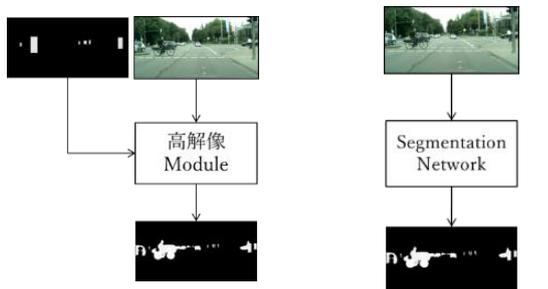
$$Recall = \frac{True\ positive}{True\ positive + False\ Negative} \quad (2)$$

4.2 高解像 Module の実現可能性

提案手法では、Semantic Segmentation のタスクを分解し、簡易化したタスクをそれぞれ汎化性 Module と高解像 Module で学習する。本実験では後段の高解像 Module を用いて、分解前後でタスクの難易度が変化しているか比較を行う。実験は高解像 Module に対して Localization のヒントが与えられたヒントありモデルと、ヒントを与えないヒントなしモデルを用いる。それぞれのモデルに対し、Cityscapes の歩行者のみの Segmentation を学習する。

図2にヒント付きモデルとヒントなしモデルの概略図を示す。

また、ヒントは入力画像に対応した CityPersons データセットの Grand Truth を用いた。



ヒントありモデル ヒントなしモデル

図2 ヒントあり、ヒントなしモデルの概略

各モデルのネットワークの表現能力は、学習難易度を比較するために可能な限り同等にして実験を行った。表1に評価結果を示す。

表1 ヒントの有無に対する認識精度の比較

	Recall	Precision
ヒントありモデル	0.817	0.612
ヒントなしモデル	0.756	0.122

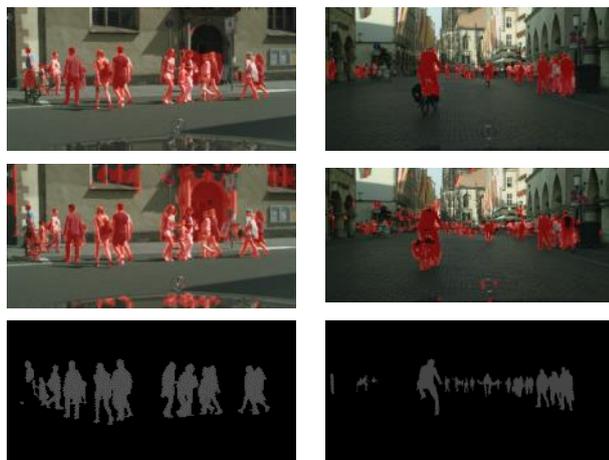


図3 ヒントありモデルとヒントなしモデルの認識結果
1行目:ヒントありモデルの認識結果, 2行目:ヒントなしモデルの認識結果, 3行目:Grand Truth

これにより、ヒントを与える事で Precision は 0.49, Recall は 0.061 の改善が見られた。特に Precision が大きく改善しているが、これはヒントを与える事で誤検出が抑制された為だと考えられる。ヒントなしモデルにおいて、誤検出は特に複雑な背景やポールや看板等の立体物に対して多く発生していた。これは今回使用したネットワークの表現能力が不足していた為、歩行者とそれに類似した物体をうまく分離する事ができなかった事が原因だと考えられる。それに対し、ヒントを与える事により大幅に精度が改善された事から、ヒント付き Semantic Segmentation はヒントなしの Semantic Segmentation より、簡易なタスクになったと考え

られる。

4.3 ヒント付き Semantic Segmentation の信頼性 (1)

前実験では高解像 Module に与えるヒントを、CityPersons データセットの Grand Truth とした。しかしながら、実際に運用する際は汎化性 Module の出力がヒントとして与えられる為、ヒントの形状にばらつきが生じたり、未検出が生じたりする可能性がある。本実験では高解像 Module がどのようなばらつきに対し弱いのか確認する為に、学習時に与えられたヒントの形状や検出率に対し、ヒント情報を機械的に変動させて評価実験を行った。変動として CityPersons の Grand Truth に対し、表2に示した7種類の処理を施した。

表2 変動内容

変動	処理内容
Base Line	処理を行わない
Fill	すべての領域をヒントありにする
Void	すべての領域をヒントなしにする
ロスト小	小さな対象物のみヒントを削除
ロストランダム	ランダムにヒントを削除
ロスト大	大きな対象物のみヒントを削除
膨張	ヒント領域を縦横 1.25 倍にする
縮小	ヒント領域を縦横 0.75 倍にする

なお、ロスト小、ロスト大における小さな対象物及び大きな対象物は、歩行者の高さによって決定した。CityPersons に登録された全歩行者の高さの中央値に対し、それより高さの高い歩行者を大きな対象物、低い歩行者を小さな対象物とした。表3に各変動に対する評価結果を示す。

表3 ヒントの変動に対する認識精度の比較

変動	Recall	Precision
Base Line	0.817 (+0.000)	0.612 (+0.000)
Fill	0.941 (-0.124)	0.029 (-0.582)
Void	0.486 (-0.331)	0.258 (-0.353)
ロスト小	0.777 (-0.040)	0.614 (+0.002)
ロストランダム	0.539 (-0.278)	0.607 (-0.005)
ロスト大	0.172 (-0.644)	0.484 (-0.127)
膨張	0.841 (+0.024)	0.461 (-0.151)
縮小	0.294 (-0.523)	0.621 (+0.009)

実験により、ヒントの変動が起こることで Recall か Precision のどちらか一方、あるいは両方が Base Line より悪化することが分かる。Recall のみが大幅に悪化する変動として Void, ロストランダム, 縮小が挙げられる。これらの変動に共通する性質は、Base Line のヒントに対し情報が欠損している事である。Precision のみが大幅に悪化する変動として、Fill, Void, 膨張が挙げられる。これらの変動に共通する性質は、Base Line のヒントに対して、過

剩に情報が付与されている事である。これらの結果は、高解像 Module がヒントを参照して学習を行っている為、正しく認識できなくなっているのだと考えられる。

図 4 は消失した対象物のサイズに対して、認識精度を比較した結果である。

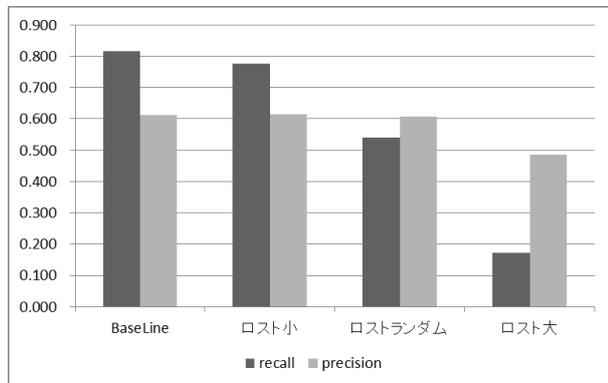


図 4 対象物の消失に対する認識精度

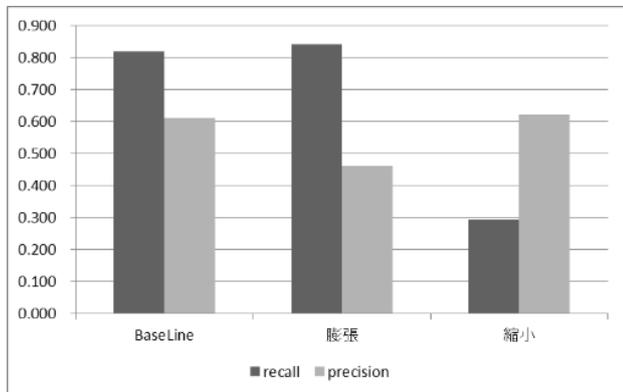


図 5 膨張と縮小に対する認識精度

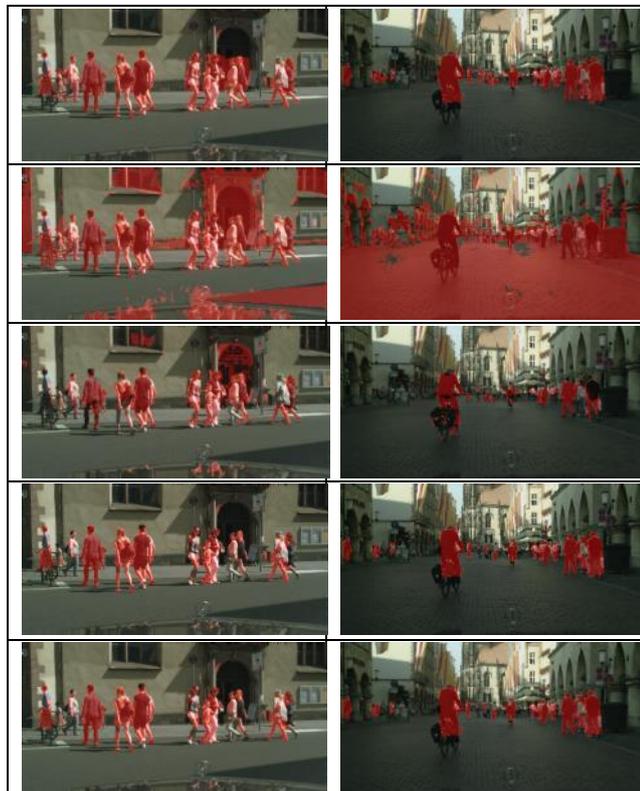


図 6 各変動に対する認識結果

1 行目: Base Line, 2 行目: Fill, 3 行目: Void, 4 行目: ロスト小, 5 行目: ロストランダム, 6 行目: ロスト大, 7 行目: 膨張, 8 行目: 縮小

消失した対象物のサイズが大きくなるに連れて、Recall は大きく低減するが、Precision は相対的に低減していないことがわかる。

図 5 は膨張と縮小に対して、認識精度を比較した結果である。ヒントが学習時より膨張している場合、Recall が増加し Precision が悪化するが、縮小している場合、Recall が極端に悪化する。この事からヒントの出力は縮小しているより膨張している方が比較的に良い事がわかる。

4.4 ヒント付き Semantic Segmentation の信頼性 (2)

前実験によりヒントの品質にばらつきが生じた時、認識精度が低下する事が分かった。実験 4.3 では学習時のヒント出力の品質と推論時の品質に差がある場合を想定した検証であったが、本実験では学習時におけるヒント品質のばらつきに対する耐性の検証を行った。実験方法として Train データの Rough アノテーションに対し、表 4 の変動を与えて高解像 Module を学習した。

表 4 変動内容

Baseline	処理を行わない
膨張 (学習)	ヒント領域を縦横 1.25 倍にする
膨張 (確率的な学習)	各対象物に対し 50%の確率で ヒント領域を縦横 1.25 倍にする
縮小 (学習)	ヒント領域を縦横 0.75 倍にする
縮小 (確率的な学習)	各対象物に対し 50%の確率で ヒント領域を縦横 0.75 倍にする
ロスト小 (学習)	小さな対象物のみヒントを削除
ロスト小 (確率的な学習)	各対象物に対し 50%の確率で 小さな対象物のみヒントを削除
ロスト小+膨張 (確率的な学習)	各対象物に対し 50%の確率で ロスト小と膨張処理を施す
ロスト小+縮小 (確率的な学習)	各対象物に対し 50%の確率で ロスト小と縮小処理を施す

これはヒントに関するデータオーギュメンテーションに当たる。評価を行う際に、ロストに関して小のみで実験を行っている理由は、大きな対象物の認識が汎化性 Module の役割である為であり、高解像 Module の役割でない為である。表 5 に実験結果を示す。

表 5 オーギュメンテーションに対する認識精度比較

	Recall	Precision
Baseline	0.817 (+0.000)	0.612(+0.000)
膨張 (学習)	0.742 (-0.074)	0.573 (-0.243)
膨張 (確率的な学習)	0.810 (-0.007)	0.544 (-0.272)
縮小 (学習)	0.676 (-0.141)	0.477 (-0.340)
縮小 (確率的な学習)	0.675 (-0.142)	0.565 (-0.251)
ロスト小 (学習)	0.806 (-0.01)	0.588 (-0.229)
ロスト小 (確率的な学習)	0.826 (0.009)	0.597 (-0.219)
ロスト小+膨張 (確率的な学習)	0.792 (-0.024)	0.581 (-0.235)
ロスト小+縮小 (確率的な学習)	0.743 (-0.074)	0.572 (-0.244)

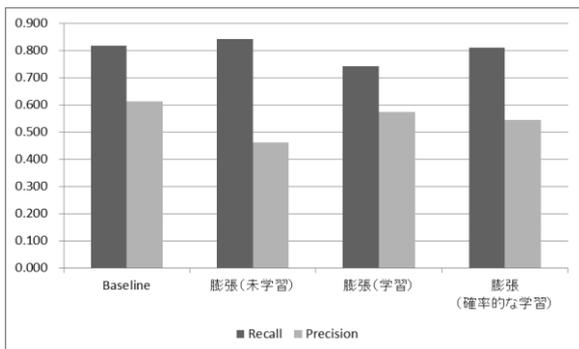


図 7 膨張に関する精度の比較

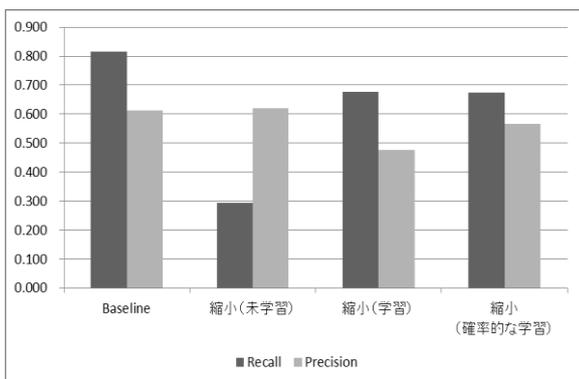


図 8 縮小に関する精度の比較

図 7 は膨張に関する精度比較である。未学習時は、ヒントが膨張していると学習時には想定していない部分にヒントが付く為、Precision が低減していたが、学習を行うことによって改善する事が分かった。また、すべての対象物に

対して膨張処理を施し学習を行うと、やや Recall が低減する事が分かった。これは Fine アノテーションに対するヒントのオーバーラップ領域が増える為、ネットワークが過剰に検出した状態を抑制する様に学習する事が原因だと考えられる。この副作用は学習を行う際に確率的に変動を与えると、抑制できる事が確認できる。

図 8 は縮小に関する精度比較である。未学習時は、ヒントが縮小していると認識に必要な領域のヒントが欠損する為、Recall が大きく低減していたが、学習を行うことで Recall が大きく改善する事が確認できたが、Precision は逆に悪化している。これは Fine アノテーションに対してヒント領域が小さい為、高解像 Module を学習する際に、ヒントが与えられていない領域から対象物を認識する必要が生じたことが原因だと考えられる。膨張と同様にこの副作用は確率的に学習する事で抑制できる。

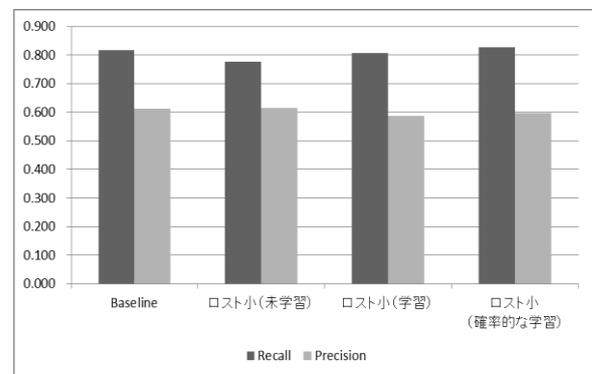


図 9 ロストに関する精度の比較

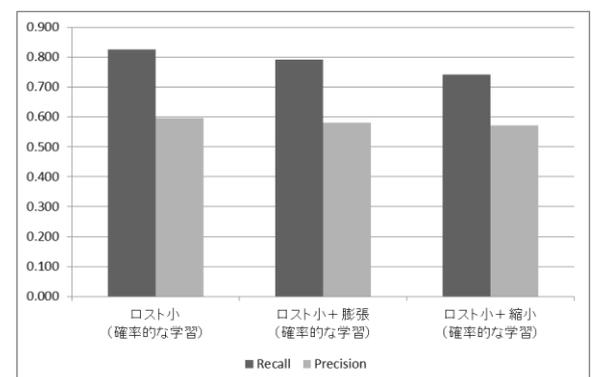


図 10 ロスト小と膨張、縮小の複合に関する精度比較

図 9 はロストに関する精度の比較である。ロスト小は未学習時でも Precision, Recall 共に Base Line と大きく変わらないが、確率的に学習する事でほぼ Base Line と同等の結果になった。これは高解像 Module がロストした歩行者を認識している為だと考えられ、表現力でも小さな対象物なら学習可能である事を示唆している。

以上の結果からオーギュメンテーションを行うことで未学習時に対して Precision もしくは Recall が改善する事と確率的に学習を行う事で学習時の副作用を低減できる事が分かった。また、確率的学習について縮小よりも膨張の

ほうが Precision と Recall の Base Line との差が小さいと分かった。したがって、汎化性 Module を学習する際の Rough なアノテーションは対象物よりも大きく囲んであげれば良い事と分かる。また、小さな対象物は高解像 Module で学習可能であるため、汎化性 Module のアノテーションにおいては比較的に見落としでも問題ないと考えられる。

また、図 10 はロスト小と膨張、縮小の複合に関する精度比較である。前述の通りロスト小と膨張の組み合わせの方が Recall, Precision 共に高い精度であると確認できる。

4.5 Module の結合

実験 4.4 は高解像 Module のみについて検証を行っていたが、本実験では汎化性 Module と高解像 Module を結合し学習を行う。また、学習に用いるモデルは表 6 に示した条件で学習し比較した。

表 6 SMN の学習方法

従来法	End-to-End で学習を行う
役割分担 (freezing)	汎化性 Module を学習したあと重みを Freezing して、高解像 Module を学習する
役割分担 (trainable)	汎化性 Module を学習したあと重みを Freezing せずに、高解像 Module を学習する

従来法とは SMN に対し直接 Fine アノテーションを End-to-End で学習するモードであり、本検証における Base Line となる。役割分担はまず汎化性 Module を Rough アノテーションで学習した後、高解像 Module と接続し Fine アノテーションで高解像 Module のみを学習する。この時、学習量に関して従来法とフェアにする為に、汎化性 Module と高解像 Module で学習するデータは同一のものとし、従来法と役割分担について互いに十分に収束するまで学習を行った。また、freezing と trainable は高解像 Module を学習する際に、汎化性 Module を Freezing するか否かである。

表 7 結合学習の精度比較

	Recall	Precision
従来法	0.693 (+0.00)	0.848 (+0.00)
役割分担 (freezing)	0.661(-0.032)	0.785(-0.063)
役割分担 (trainable)	0.718(+0.025)	0.880(+0.032)

表 7 に実験結果を示す。実験によりアノテーションコストが同等の時、役割分担学習(freezing)より従来法の方が Recall で 0.032, Precision が 0.063 高い事が分かった。これは役割分担(freezing)の汎化性 Module を学習する際に高解像の Semantic Segmentation を想定していない事が原因だと考えられる。

また、役割分担(trainable)の様に高解像 Module の学習にあわせて汎化性 Module を追加的に学習する事によって、従来手法より Recall で 0.025, Precision が 0.032 高くなり、従来方より高効率に学習できることが分かった。

これは高解像 Module の学習にあわせて、汎化性 Module がチューニングされている事が要因だと考えられる。以上の結果から、アノテーションコストが同等の時、役割分担学習を行う事で高効率に学習を行えることがわかった。

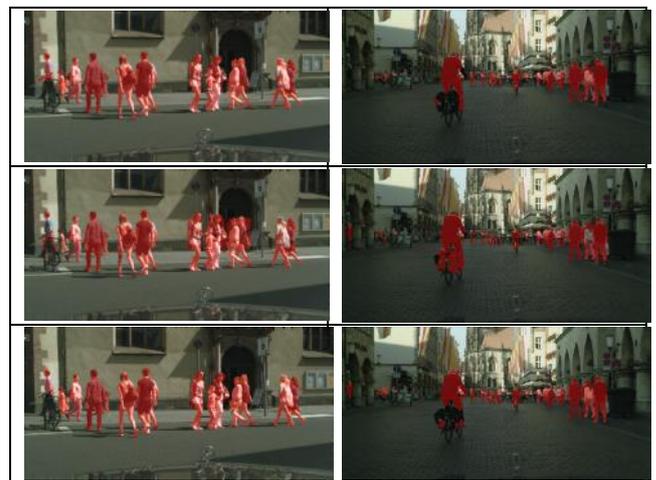


図 11 結合学習の認識結果

1 行目: 従来法, 2 行目: 役割分担(freezing),
 3 行目: 役割分担(trainable)

4.6 学習効率の比較実験

本実験では、Sub Module Network と役割分担学習による学習効率の改善効果を確認する。本実験では Fine アノテーションと Rough アノテーションのアノテーションコストを一定にした際の学習効率を比較する。

事前検証の結果 Fine と Rough のアノテーションコストはそれぞれ Fine が 7.5 分/枚, Rough が 2.1 分/枚となった。Cityscapes と CityPersons データセットは共に同じ画像に対してアノテーションが付けられている。そこで総アノテーションコストを学習用データの総数である 2975 枚に対し、Rough のアノテーション時間を積算した 6248 分とし、Rough と Fine の比率に応じてアノテーションコストを分配した。表 8 に学習データの分配比率を示す。

表 8 Fine と Rough のデータ数内訳

	Fine と Rough の割合		枚数	
	Fine	Rough	Fine	Rough
役割分担	80%	20%	666	595
従来法	100%	0%	833	0

役割分担は汎化性 Module を Rough データ 595 枚で学習し、高解像 Module を 666 枚で学習する。従来法は汎化性 Module と高解像 Module に対し Fine データ 833 枚で学習した。また、役割分担は高解像 Module を学習する際に汎化

性 Module を学習しない freezing と学習を行う trainable の 2 つの設定で実験した。実験結果を表 9 に示す。

表 9 役割分担学習の学習効率

	Recall	Precision
従来法	0.397 (+0.000)	0.820 (+0.000)
役割分担 (freezing)	0.541 (+0.144)	0.659 (-0.161)
役割分担 (trainable)	0.608 (+0.211)	0.822 (+0.002)

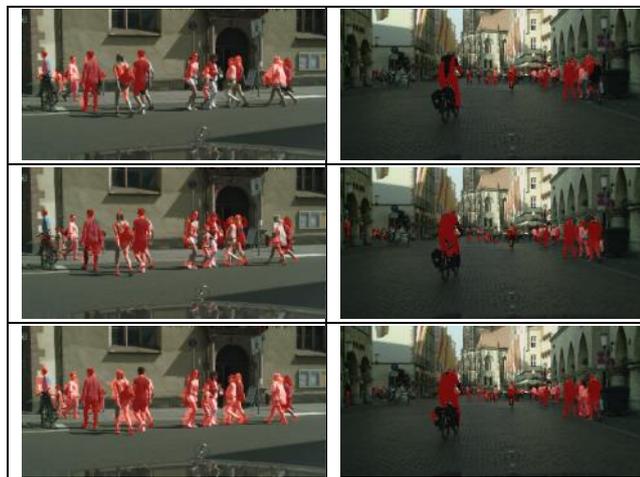


図 12 各モデルの認識結果

1 行目: 役割分担(freezing), 2 行目: 役割分担(training), 3 行目: 従来法

実験により役割分担学習を行う事で、従来法より Recall が改善していることが分かる。これは汎化性 Module を低コストのアノテーションで大量に学習する事により、未検出率が低減しているためだと考えられる。また、役割分担 (freezing) は従来法に対して Precision が悪化している。これは実験 4.6 と同様に汎化性 Module の学習時に緻密な Segmentation を考慮できていないことが要因だと考えられるが、高解像 Module を学習する際に汎化性 Module も学習を行う事で、従来法と同等以上の Precision になった。以上の結果から、学習コストを一定にした際に、役割分担学習を行う事で、従来法よりも高効率で学習可能である事が分かった。

5. 結論

本論文では、Semantic Segmentation のタスクを複数の簡易なタスクに分解することで学習効率を向上させる役割分担学習と、専用のネットワーク構造の Sub Module Network を提案した。タスク分解をすることで、Semantic Segmentation を 2 つの簡易なタスクの組み合わせにすることができた。また、それぞれのタスクに対し個別に学習を行う事で、従来手法よりアノテーションコスト辺りの学習効率が向上することを確認した。今後の課題として、役割分担学習における Rough, Fine の配合比率の詳細な調査と、Rough アノテーションに求められる性質の調査を行う必要

がある。また、その一環として Cityscapes の Coarse を用いた実験を行い、本手法の汎用性を調査する必要がある。

参考文献

- [1] Shotton, Jamie, Matthew Johnson, and Roberto Cipolla. "Semantic texton forests for image categorization and segmentation." Computer vision and pattern recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008.
- [2] Shotton, Jamie, et al. "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context." International journal of Computer Vision 81.1(2009): 2-23.
- [3] Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition, 2015.
- [4] Cordts, Marius, et al. "The cityscapes dataset for semantic urban scene understanding." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
- [5] Lin, Tsung-Yi, et al. "Microsoft coco: Common objects in context.. "European conference on computer vision. Springer, Cham, 2014
- [6] Zhao, Hengshuang, et al. "Pyramid scene parsing network". IEEE Conf. on Computer Vision and Pattern Recognition, 2017
- [7] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.
- [8] Lin, Guosheng, et al. "RefineNet: Multi-part Refinement Networks for High-Resolution Semantic Segmentation" CVPR. Vol. 1. No. 2. 2017
- [9] Kendall, Alex, Vijay Badrinarayanan, and Roberto Cipolla. "Bayesian segnet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding " arXiv preprint arXiv: 1511.02680:
- [10] Paszke, Adam, et al. "Enet: A deep neural network architecture for real-time semantic segmentation." arXiv preprint
- [11] Zhao, Hengshuang, et al. "Icnnet for real-time semantic segmentation on high-resolution images." arXiv preprint arXiv:1704.08525(2017)
- [12] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition", International Conference on Learning Representation (ICLR2015), 2015.
- [13] V. Badrinarayanan, A. Kendall, R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation", arXiv preprint arXiv: 1511.00561, 2015.
- [14] Hoh, Hyeonwoo, Seunghoon Hong, and Bohyung Han. "Learning deconvolution network for semantic segmentation." Proceedings of the IEEE international conference on computer vision. 2015.
- [15] Romera, Eduardo, et al. "Erfnet: Efficient residual factorized convent for real-time semantic segmentation." IEEE Transactions on Intelligent Transportation Systems 19.1(2018): 263-272
- [16] Gamal, Mostafa, Mennatullah Siam, and Moemen AbdelRazek. "ShuffleSeg: Real-time Semantic Segmentation Network." arXiv preprint arXiv:1803.03816, 2018
- [17] Poudel, Rudra PK, et al. "ContextNet: Exploring Context and Detail for Semantic Segmentation in Real-time." arXiv preprint arXiv: 1805.04554 2018.
- [18] Zhang, Shanshan, Rodrigo Benenson, and Bernt Schiele. "CityPersons: A diverse dataset for pedestrian detection." The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Vol. 1. No. 2. 2017
- [19] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.