Attention と意味情報補強処理による高速な物体検出手法の提案

山重 雄哉† 青野 雅樹‡

†豊橋技術科学大学 情報・知能工学専攻 〒441-8580 愛知県豊橋市天白町雲雀ヶ丘 1-1
‡豊橋技術科学大学 情報・知能工学系 〒441-8580 愛知県豊橋市天白町雲雀ヶ丘 1-1
E-mail: †yamashige@kde.cs.tut.ac.jp ‡aono@tut.jp

あらまし 深層学習に基づく画像上の物体検出技術が注目され，盛んに研究が行われている．代表的な手法である SSD: Single Shot Multibox Detector は高精度な検出を可能とし，多くの改良モデルが提案されている．一方，これらのモデルは VGG-16 や ResNet-101 を特徴抽出のネットワークとして用いるため計算コストが高く，高速な検出には高性能な GPU を必要とする．これに対して，先行研究では，畳み込み層および特徴マップの次元数を削減することで計算量を大幅に削減した SSD7 が提案された．本稿では SSD7 に対して高精度化改良を行う．具体的には，Attention による物体領域の強調，FPN および Skip Concatenation による意味情報補強処理を組み合わせることで精度向上を図る．従来手法(SSD7，FPSSD7，SSD300)との比較実験では Udacity Annotated Driving Dataset，Pascal VOC Dataset を用いて精度，速度の両面で評価を行った．その結果，提案手法の優位性が示された．また，アブレーション実験および特徴マップの可視化によって提案モジュールの有効性を確認した．

キーワード 物体検出，深層学習， SSD: Single Shot MultiBox Detector，Attention，FPN: Feature Pyramid Networks

1. はじめに

深層学習を用いた画像上の物体検出技術が注目され，様々な研究が行われている．物体検出は様々なビジネスシーンに応用可能であり，監視カメラによる自動監視,自動車やロボット,ドローンなどの自動運転，工場での外見検査,がん検出などが挙げられる. Single Shot MultiBox Detector (SSD)[1]は比較的高速で高精度な検出手法であり,様々な改良手法が提案されている．しかし，これらの改良手法は精度の向上に重きを置いたものが多く低速化している．また，特徴抽出ネットワークとして VGG-16[4]や ResNet-101[20]等を用いるため計算コストが高く，高速な検出には高性能な GPU を必要とする．一般的に深層学習による物体検出は精度と速度のトレードオフが課題とされ，特に自動運転シーンでは精度だけでなくリアルタイム性も求められる．また，ドローンなどの小型デバイスを用いる場合はマシンコスト面も考慮しなければならない.そこで，本稿では新たな物体検出手法として Attention Feature Pyramid SSD7 (AFPSSD7)を提案する．本手法は，特徴抽出部の畳み込み層を 7 層に減らし，特徴マップ次元数を削減した SSD7[16]を用いることで計算量を大幅に削減している．また，Attention および意味情報補強処理を導入することで，計算量削減による精度低下を改善している．比較実験ではデータセットとして Udacity Annotated Driving Dataset[17] および The PASCAL Visual Object Classes Challenge 2007[5] の Pascal VOC 2007 test を用いて精度，速度の両面から提案手法の性能を評価した．また，アブレーション実験によって提案モジュールの有効性の検証した.さらに，検出結果および特徴マップの可視化を行うことで Attention の動作を確認した．

2 節において関連研究として代表的な物体検出手法，および SSD の改良手法を説明する．3 節では提案手法について説明する．4 節では比較実験およびアブレーション実験の結果を示す．また，5 節では検出結果および特徴マップの可視化によって Attention の動作を確認する. 6 節において本実験の考察を行い, 7 節では結論および今後の課題について述べる．

2. 関連研究

本節では関連研究を示し，2.1 節では深層学習を用いた代表的な物体検出手法の概要を説明する．また，2.2 節では SSD について説明する．

2.1. 深層学習を用いた物体検出手法

深層学習を用いた物体検出手法は two-stage 型と one-stage 型に大別される. two-stage 型は物体の候補領域を提案するステップと，その候補領域のクラスを特定するステップに分かれている点が特徴である．代表的な手法としては Regions with CNN features (R-CNN)[6], Fast R-CNN[7], Faster R-CNN[8]などがある．Faster R-CNN は PASCAL VOC2007 test で 73.2mAP を記録したが検出速度が非常に低速であり，リアルタイム性の向上が課題であった．

one-stage 型は領域提案とクラス分類のステップを同時に行う点が特徴であり，モデルが単一のネットワークであるためエンドツーエンドに訓練することが可能である．さらに one-stage 型は YOLO 系と SSD 系に分類され，YOLO 系は You Only Look Once (YOLO v1) [9]，その改良モデルの YOLO v2[10]，YOLO v3[11]な

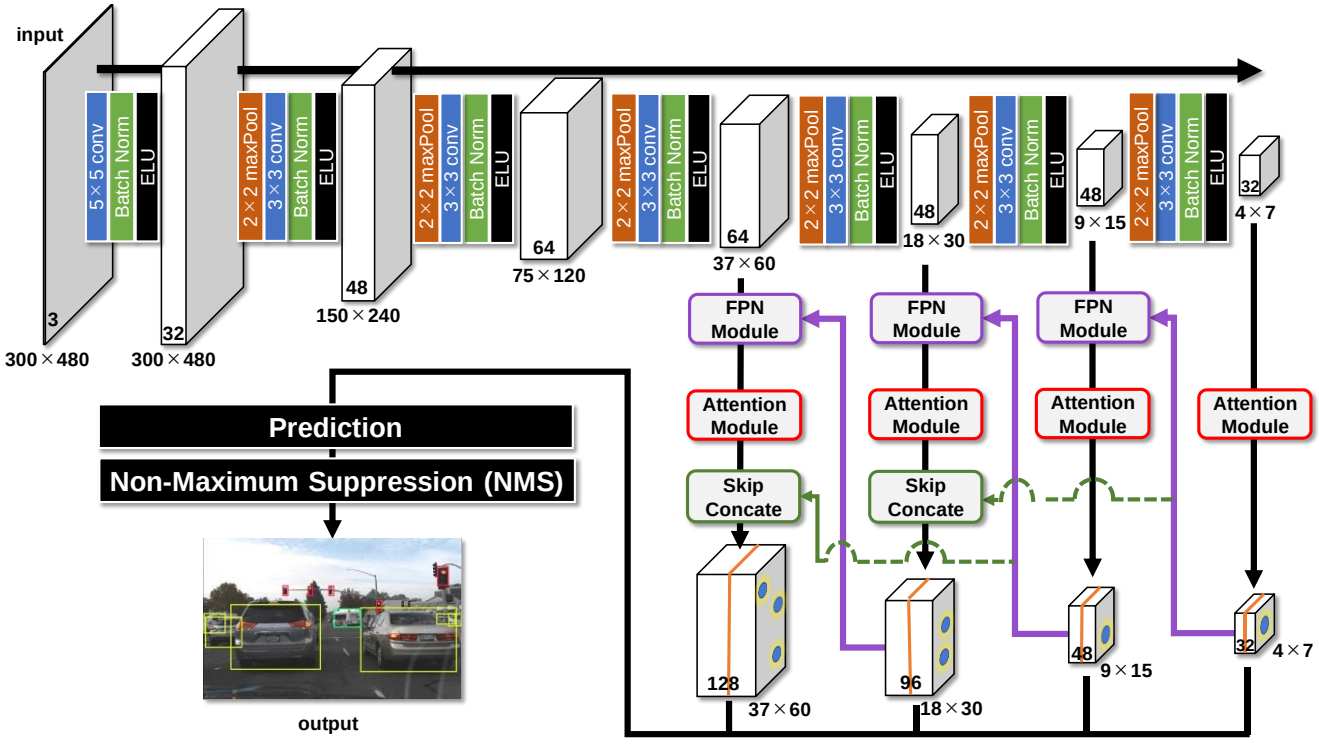


図 1. Attention Feature Pyramid SSD7 (AFPSSD7)のモデルアーキテクチャ

どがある．YOLO v1 は特徴抽出を行うバックボーンネットワークとして GoogleLeNet[12]を使用し，特徴マップの生成後，この特徴マップを S×S のグリッドに分割する．その後，各セルについてバウンディングボックスの座標計算とクラス確率計算を同時に行う．YOLO の利点としては高速性が挙げられ，YOLO v1 は Pascal VOC 2007 test において 45FPS の速度で 63.4mAP の検出を可能とした．一方，グリッドサイズが固定であるため，セル内に小さい物体が複数存在した場合の検出が困難になるという欠点もある．

2.2. SSD : Single Shot Multibox Detector

高精度な検出が可能な SSD (SSD300)は Pascal VOC 2007 test において 46FPS の速度，77.2mAP の精度を記録した．SSD は、物体が存在する画像を VGG-16 に入力し，各層で解像度の異なる特徴マップを得る．その後，計 6 層の特徴マップを物体の予測層に入力する．そして，特徴マップをグリッドに分割し，各セルにおいてバウンディングボックスのオフセット値およびクラス確率を計算する．ここで，低解像特徴マップでは大きい物体を，高解像特徴マップでは小さい物体の予測を行うため，物体サイズにロバストな検出を可能としている．最終的に，Non-Maximum Suppression (NMS)を用いて重なったバウンディングボックスを統合する．

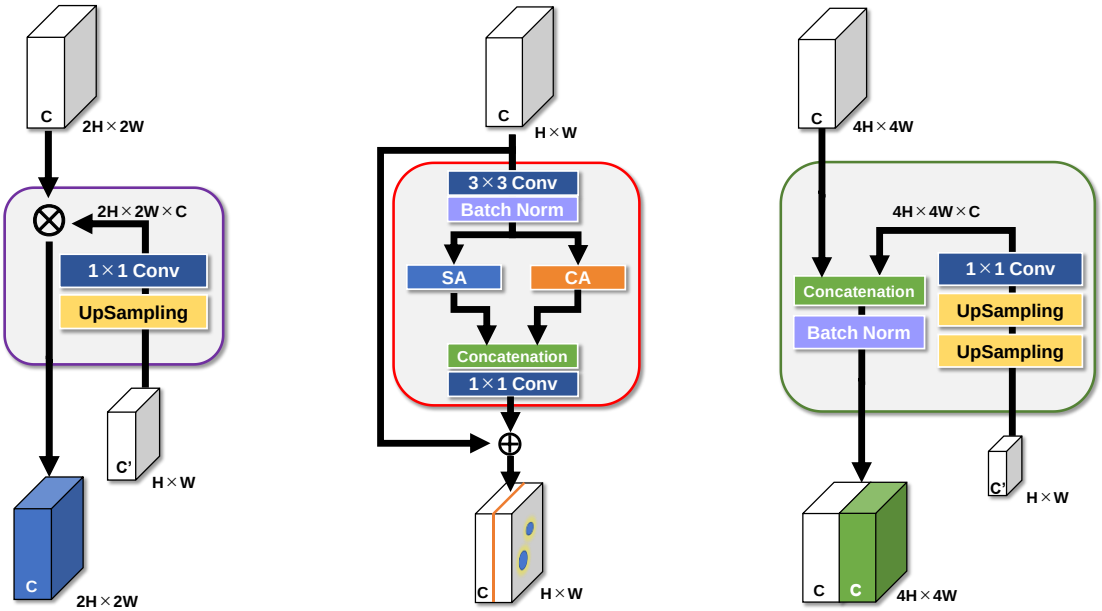
SSD は様々な改良手法が提案されており，DSSD[13], M2Det[14], RefineDet[15]などが代表例として挙げられる．例えば，RefineDet は Pascal VOC 2007 test において 40.3FPS，80.0mAP を達成し，リアルタイム性を維持しつつ高精度な検出を可能とした．なお，上記手法の検出速度は NVIDIA Titan X を用いた場合の検出速度であり，リアルタイムな検出には高性能な GPU を必要とする．これに対して，先行研究では SSD の高速改良モデルが提案された．例えば，SSD7 は畳み込み層を 13 層から 7 層に減らし，特徴マップの次元数を削減することで計算量を大幅に削減したモデルである．SSD7 を改良した Feature Pyramid SSD7 (FPSSD7) [2]は SSD7 に Feature Pyramid Networks (FPN)[3]を用いた意味情報補強処理を導入し，計算量削減による精度低下を改善した．しかしながら SSD300 や，その他の改良手法と比較すると精度は低く，更なる改良が必要とされた．

3. 提案手法

我々は小規模なモデルで高速な検出が可能な Attention Feature Pyramid SSD7 (AFPSSD7)を提案する．3.1 節では提案手法の概要を示し，3.2 節では意味情報補強処理，3.3 節では Attention の手法を説明する．

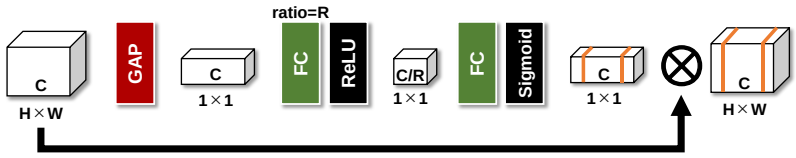
3.1. AFPSSD7 : Attention Feature Pyramid SSD7

SSD は特徴抽出ネットワークに VGG-16 を使用しており，ネットワーク全体のモデルパラメータは約 2600 万に及ぶ(20 クラス分類時)．そのため，推論時および訓練時の計算コストが膨大になってしまう．そこで，本節では高速検出が可能な FPSSD7 に更なる改良を加えた Attention Feature Pyramid SSD7 (AFPSSD7)を提案する．図 1 は提案手法のモデルアーキテクチャであり，

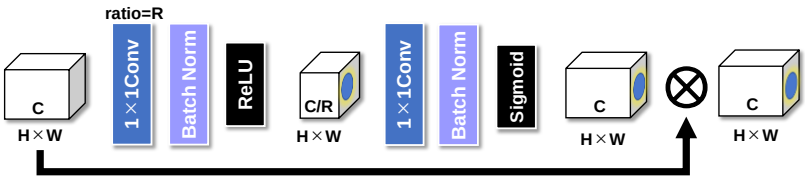


(a)FPN Module (FPNM)　(b) Attention Module (AM)　(c) Skip Concatenation Module (SCM)

図 2. 提案手法内モジュールのアーキテクチャ



(a) Channel Attention (Squeeze and Excitation)



(b) Spatial Attention

図 3. Channel Attention および Spatial Attention のアーキテクチャ

従来の FPN Module (FPNM)に加え，Attention Module (AM)，Skip Concatenation Module (SCM)が新たに追加されている．まず 7 層の畳み込み層を通して特徴マップを生成した後，FPNM において，低解像特徴マップのセマンティックな情報(意味情報)を高解像特徴マップに伝播させる．その後 AM において，特徴マップのチャンネル方向の強調，および物体が存在する領域の強調が行われる．さらに，SCM で高解像特徴マップの意味情報を補強し，表現度を向上させる．なお，上記の処理は物体予測に用いる 4 層目から 7 層目の特徴マップに対して行われる．

3.2. 意味情報補強処理

提案モジュールは FPN Module (FPNM)，Attention Module (AM)，Skip Concatenation Module (SCM)から構成され，アーキテクチャを図 2 に示す．本節では，FPNM および SCM について説明する．これらのモジュールの目的は，深い層の特徴マップが持つ意味情報を浅い層の特徴マップに伝播させることであり，我々は意味情報補強処理と呼んでいる．FPNM のアーキテクチャは図 2(a)に示す通りである．まず，H×W の特徴マップに対して解像度を 2 倍にする線形アップサンプリングとカーネルサイズが 1×1 の畳み込み(1×1Conv)を行う．その後，2H×2W の特徴マップと要素積を計算することで意味情報を補強する．ここで，1×1Conv 層は要素積計算時のチャンネル数一致を目的として使用される．また，アップサンプリング時，解像度が異なる場合はゼロパディングにより調節する．

図 2(c)に SCM のアーキテクチャを示す．入力は H×W と 4H×4W の特徴マップであり，意味情報が補強された 4H×4W の特徴マップが出力される．SCM のアプローチは基本的には FPNM と同様であるが，4 倍のアップサンプリングを使用する点と，特徴マップ同士の結

表 1. Udacity Annotated Driving Dataset 使用時の性能比較

method	inference time [ms]	fps	params	mAP	car	truck	pedestrian	bicyclist	light
SSD7 [16]	15.330	65.233	213,904	25.4	52.7	36.7	10.2	9.1	18.1
SSD13	15.991	62.534	833,437	30.7	57.9	43.9	16.1	16.2	21.0
SSD300 [1]	23.446	42.650	24,280,556	**34.6**	**58.4**	**45.4**	**23.1**	**19.8**	**26.3**
FPSSD7 [2]	**13.043**	**76.669**	299,472	27.8	54.4	35.6	16.3	12.2	20.7
AFPSSD7 (ours)	15.444	64.751	384,176	31.8	56.3	42.1	18.1	17.7	24.7

表 2. Pascal VOC 2007 test 使用時の性能比較

method	backbone	fps	GPU	params	input size	mAP
Faster R-CNN [8]	VGG-16	0.5	Titan X	-	～1000×600	73.2
YOLO [9]	GoogleLeNet	45.0	Titan X	-	448×448	63.4
YOLO v2 [10]	Darknet-19	40.0	Titan X	-	544×544	78.6
SSD300 [1]	VGG-16	46.0	Titan X	-	300×300	77.2
SSD512 [1]	VGG-16	19.0	Titan X	-	512×512	79.5
SSD300 (implementation)	VGG-16	34.7	P6000(3GB)	26,285,486	300×300	76.5
SSD7 [16]	7th conv	42.0	P6000(3GB)	317,824	300×300	36.5
FPSSD7 [2]	7th conv	42.4	P6000(3GB)	403,392	300×300	43.5
AFPSSD7 (ours)	7th conv	32.9	P6000(3GB)	548,576	448×448	49.6

合(Concatenation)により意味情報を補強する点が異なる.

3.3. Attention

Attention Module(AM)の構造は，Hu らの CSAR block[18]から着想を得たものであり，アーキテクチャは図 2(b)に示す通りである．AM は特徴マップのチャンネル方向の強調を行う Channel Attention (CA) 層と，領域の強調を行う Spatial Attention (SA) 層から構成されている．まず，入力された特徴マップに対してカーネルサイズが 3×3 の畳み込みを行う．その後，CA 層および SA 層において Attention を行なう．そして，Attention された特徴マップを結合し，1×1 Conv 層を通してこれらを融合する．なお，このモジュールは ResNet[20]に基づくショートカット構造を用いている．これは，層数増加に基づく劣化問題の改善，および訓練の容易性向上を目的としている.

各 Attention のアーキテクチャは図 3 に示す通りである．CA 層については Squeeze and Excitation (SE)機構 [19] を使用している. なお, ハイパーパラメータである ratio 値は 8 に設定している. また, SA 層も SE 機構を基本構造としているが，Global Average Pooling (GAP)層および全結合層(Dense)を 1×1Conv 層に変更することで領域方向の Attention を実現している.

4. 比較実験

車載カメラ画像データセットの Udacity Annotated Driving Dataset および一般物体画像データセットの Pascal VOC Dataset を用いて各モデルの検出性能を比較した．さらに，アブレーション実験も行い，提案モジュールの有効性を検証した.

4.1. データセット

Udacity Annotated Driving Dataset は車載カメラ画像のデータセットである. 物体は car, truck, pedestrian, bicyclist, light が存在し, 比較的少クラスなデータセットである. 解像度は 300×480, 訓練データは 18000 枚, テストデータは 4241 枚から構成されている.

また，より多クラスなデータセットとして Pascal VOC Dataset を用いた. 物体は car, airplane, sofa, bus, bottle などの 20 クラスが存在する. 訓練データは Pascal VOC 2007 trainval，2012 trainval を組み合わせた 16551 枚，テストデータは Pascal VOC 2007 test の 4952 枚である.

4.2. 実験環境

比較に用いたモデルは SSD7，SSD13，SSD300，FPSSD7 および提案手法の AFPSSD7 である．ここで，SSD13 は本実験で新たに構築したモデルであり, SSD7 の畳み込み層を 7 層から 13 層に拡張したものである. 速度の評価指標は，領域統合処理を含む画像 1 枚当たりの平均推論時間および FPS である．また，精度の評価指標として各クラスの Average Precision (AP)および mean Average Precision (mAP)を用い，予測領域と正解領域の Intersection over Union (IoU)が 0.5 以上かつクラスが一致した場合に正解とみなす．また，GPU は NVIDIA Quadro P6000 を用い，推論時の使用可能メモリは 3GB に制限した.

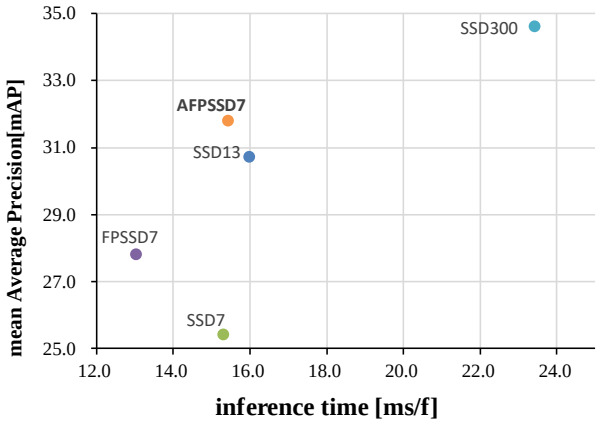


図 4. 各モデルの推論時間に対する精度

4.3. Udacity Annotated Driving Dataset 実験

Udacity Annotated Driving Dataset を用いた検出性能比較を行った．各モデルのバッチサイズは 32，エポック数は 100 として訓練を行った．最適化手法は Adam [21] を用い，学習係数の初期値は 0.001 とした．また，推論時のバッチサイズは 1 とし，確信度閾値は 0.02 とした．ここで，確信度閾値とは NMS におけるハイパーパラメータであり，閾値より低い確信度のバウンディングボックスは統合前に破棄される．

実験結果を表 1 および図 4 に示す．ここで，図 4 は各モデルの推論時間に対する精度をプロットしたものである．横軸は画像 1 枚当たりの平均推論時間，縦軸は mAP である．提案手法は SSD13 や SSD300 と比較してモデルパラメータが小規模であるにも関わらず，31.8mAP を記録し，FPSSD7 に対して精度が 4.0 ポイント向上した．さらに，提案手法は SSD13 よりも高い mAP を記録しており，モデルパラメータ数を増加することで精度を向上させるアプローチよりも，提案手法のアプローチが優れていることを示した．また，推論時間は SSD300 と比較して非常に短く，SSD7 と同程度の推論時間で高い精度を記録した．

4.4. Pascal VOC 2007 test 実験

本実験では Pascal VOC 2007 test を用いて多クラス物体検出時の性能を検証した．なお，実験方法やハイパーパラメータについては 4.3 節と同様である．また，提案手法の入力画像サイズは 448×448 とした．

実験結果を表 2 に示す．提案手法の mAP は従来の FPSSD7 に対して 6.1 ポイント向上し，49.6mAP を記録した．しかしながら他手法と比較すると精度は低く，また，速度面についても再現実装した SSD300 と同程度な速度という結果となった．

4.5.アブレーション実験

提案モジュールの有効性を検証するために，アブレーション実験を行った．対象モジュールは FPNM, AM, SCM で，使用データセットは Udacity Annotated Driving Dataset，実験環境は 4.3 節と同様である．表 3 に実験結果を示す．ここで，「✓」はそのモジュールがモデル内に含まれていることを表す．表 3 より，どのモジュールを取り除いた場合においても精度の低下が見られ，全モジュールが精度向上に寄与していることが分かる．特に FPNM と AM を組み合わせたモデルの精度が高いため，両モジュールの重要性も確認できる．また，Attention を取り除くと light (信号機)の AP が大きく低下しており，Attention は信号機のような比較的サイズの小さい物体に対して有効であると推測できる．

表 3. 提案手法のアブレーション実験結果

FPNM	AM	SCM	mAP	car	truck	pedestrian	bicyclist	light
	✓	✓	28.4	54.5	35.7	16.0	14.3	21.5
✓		✓	28.8	54.1	40.2	16.0	14.9	18.9
✓	✓		31.2	56.5	40.3	18.5	16.8	24.1
✓	✓	✓	31.8	56.3	42.1	18.1	17.7	24.7

5. 可視化

本節では，検出結果および特徴マップの可視化を行うことで，Attention の有効性を検証する．

5.1. 検出結果の可視化

図 5 に検出結果の可視化例を示す．左から SSD7 (a), FPSSD7 (b)，提案手法の AM を除いたモデル(AM 除去モデル) (c), 提案手法 (d)の順である．SSD7 や FPSSD7 は，車の前方に存在する 2 つの信号機(light)や右端の歩行者(pedestrian)を検出できていない．また，AM 除去モデルについては，左側の信号機は検出できているが，もう一方の信号機や歩行者の検出には失敗している．これに対して提案手法は 2 つの信号機および 3 人の歩行者を正確に検出できている．

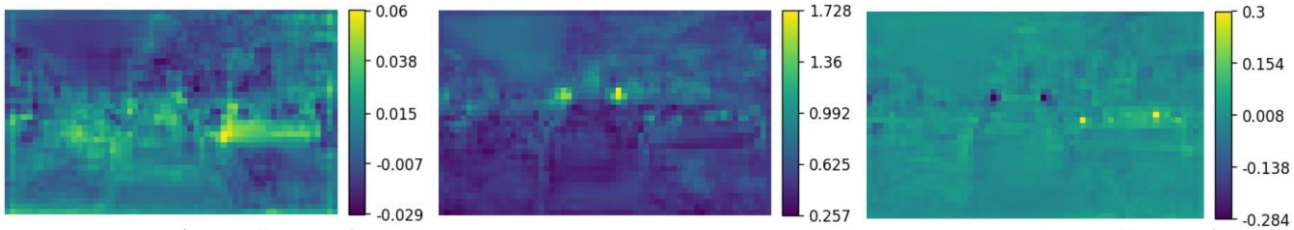
図 8 は，上記の 4 つの手法の検出結果の違いが分かる例である．なお，この図のバウンディングボックスの色は黄が car，緑が truck，青が pedestrian，紫が bicyclist，赤色が light クラスを表している．まず 1 行目の図から，提案手法は従来手法や AM 除去モデルと比較して，信号機に対する検出漏れが最も少なく，2 台の自転車も正確に検出できていることが分かる．2 行目の図では，小さく映った歩行者を検出できているのは提案手法のみであり，従来手法の検出漏れを改善している．3 行目の図においては，従来手法は建物の誤検出や検出領域のオーバーラップの多さが目立っていたのに対して，提案手法は的確な検出に成功している．また，4 行目の図は提案手法の小さい物体に対する検出漏れ，誤検出，オーバーラップの少なさを確認できる例である．

5.2. 特徴マップの可視化

Attention の有効性を確認するために図 5 の特徴マップの可視化を行った．ここで，可視化対象は 4 層目の

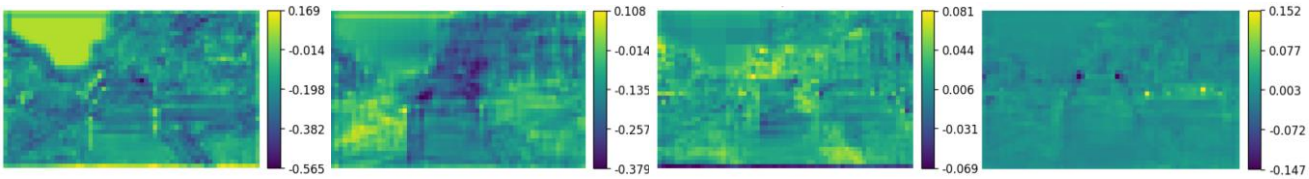
(a) SSD7　(b) FPSSD7　(c) Attention 除去モデル　(d) AFPSSD7 (提案)

図 5. 検出結果の可視化例



(a)Attention 前の特徴マップ　(b) SA 層の特徴マップ　(c) Attention 後の特徴マップ

図 6. Attention の可視化 (4 層目)



(a) SSD7　(b) FPSSD7　(c) Attention 除去モデル　(d) AFPSSD7 (提案)

図 7. 予測に用いた特徴マップの可視化 (4 層目)

解像度の特徴マップとした．なお，特徴マップは複数のチャンネルを持つため，同画素位置でチャンネル方向の平均値を計算し，単一化することで可視化した．

図 6 (a)に Attention が行われる前の特徴マップを示し，目立った物体の強調は行われていないことが分かる．ここで，各画素の描画色は右側に示すカラーバーに対応付けられている．なお，縦軸の値は特徴マップの画素の最大値が黄色，最小値が黒色となるようにスケーリングされている．図 6 (b)に SA 層における特徴マップを示し，信号機や歩行者の領域に反応を示している様子が確認できる．図 6 (c)に CA を含めた Attention 後の特徴マップを示し，信号機と歩行者の領域がより強調されている．また，図 7 は各モデルの物体予測に使用した特徴マップを示している．これらの図から，提案手法以外の特徴マップでは物体領域が強調されている様子は見られず，提案手法で生成された特徴マップの表現度の高さが視覚的に確認できる．

6. 考察

比較実験より，提案手法は少クラスな物体検出シーンにおいては有効性を発揮し，精度と速度のバランスが取れたアプローチであると言える．これは，意味情報補強処理と Attention の相互作用が，検出の補助機構として機能したためであると考えられる．また，アブレーション実験および特徴マップの可視化から Attention は比較的小さい物体の検出精度向上に有効であると考えられる．これは，特徴マップの領域強調だけでなく，チャンネルの強調が行われたためであると思われる．例えば図 6(b)では物体が存在する領域に反応を示しているが，チャンネルの強調が加わった図 6(c)では，信号機と歩行者のクラスの違いを考慮した強調が行われていることが分かる．これより，チャンネルの強調が物体の意味理解に寄与し，クラスの分類を容易にしたと推測される．

一方で，提案手法は多クラスの一般物体検出では速度も低く，モデルの軽量性を生かすことができなかった．この要因として，モデルパラメータ数が少ないためにモデルの表現度が不足している点が挙げられる．モデルの表現度が低いことで物体のクラスを分類することができず，確信度の高い領域も得られないため，領域統合処理の計算コストが膨大になったと推測される．

7. おわりに

本稿では，意味情報補強処理と Attention を用いた高速な物体検出手法 AFPSSD7 を提案した．Udacity Annotated Driving Dataset を用いた比較実験では，小規模なモデルで高速性を保ちつつ，従来手法の精度を上

(a) SSD7　(b) FPSSD7　(c) Attention 除去モデル　(d) AFPSSD7

図 8. 検出結果の比較例

回る性能を示した．これは，図 5 および図 8 に示す検出結果例からも確認できる．また，アブレーション実験を行うことで提案モジュールの有効性を示し，特徴マップの可視化によって Attention が信号機などの比較的小さい物体の検出精度向上に寄与していることを確認した．一方，Pascal VOC Dataset のような多クラス検出時において精度の向上は見られたものの，SSD300 に対しては精度，速度の両面で優位性を示すことができなかった．今後の課題としてはモデルの表現度向上が挙げられ，意味情報補強処理や Attention の改良，および特徴抽出ネットワークの変更など，様々なアプローチを模索し改良を重ねていきたい．

謝　辞

本研究の一部は，科研費基盤 (B)17H01746 の支援を受けて遂行した．

参　考　文　献

[1] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg. ”SSD: Single Shot MultiBox Detector”, In ECCV, 2016

[2] Y. Yamashige, M. Aono. “FPSSD7: Real-time Object Detection using 7 Layers of Convolution based on SSD”, In ICAICTA, 2019

[3] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. “Feature Pyramid Networks for Object Detection,” In CVPR, 2017

[4] K. Simonyan, A. Zisserman. “Very deep convolutional networks for large-scale image recognition”, In NIPS, 2015

[5] M. Everingham, L. J. V. Gool, C. K. I. Williams , J. M. Winn , and A. Zisserman. ”The PASCAL Visual Object Classes (VOC) Challenge", In IJCV, 88(2), pp.303-338, 2010

[6] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik. “Rich feature hierarchies for accurate object detection and semantic segmentation”, In CVPR, 2014

[7] R. B. Girshick. “Fast R-CNN”, In ICCV, 2015

[8] S. Ren, K. He, R. B. Girshick, and J. Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, In NIPS, 2015

[9] J. Redmon, S. Divvalay, R. B. Girshick, and A. Farhadiy. “You Only Look Once : Unified, Real-Time Object Detection”, In CVPR, 2016

[10] J. Redmon and A. Farhadi. “YOLO9000:Better, Faster, Stronger”, In CVPR, 2017

[11] J. Redmon and A. Farhadi. “yolov3: An incremental improvement”, arXiv preprint arXiv:1804.02767, 2018

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. “Going deeper with convolutions”, CoRR, abs/1409.4842, 2014

[13] C. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg. ”DSSD : Deconvolutional single shot detector”,

CoRR, abs/1701.06659, 2017

[14] Q. Zhao, T. Sheng, Y. Wang1, Z. Tang, Y. Chen, L. Cai, and H. Ling. “M2Det: A Single-Shot Object Detector based on Multi-Level Feature Pyramid Network”, In AAAI, 2019

[15] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li. “Single-Shot Refinement Neural Network for Object Detection”, In CVPR, 2018

[16] Pierluigiferrarr. “keras_ssd7.py,” [Online] https://github.com/pierluigiferrari/ssd_keras/tree/master/models/keras_ssd7.py, [Accessed: 5-August 2018]

[17] “ Udacity, Annotated Driving Dataset,” [Online] Available:https://drive.google.com/open?id=1tfBFavijh4UTG4cGqIKwhcklLXUDuY0D,[Accessed:1-August 2018].

[18] Y. Hu, J. Li, Y. Huang, and X. Gao. “Channel-wise and Spatial Feature Modulation Network for Single Image Super-Resolution”, arXiv: 1809.11130, 2018

[19] J.Hu, L. Shen, G. Sun. “Squeeze-and-Excitation Networks”, In CVPR, 2018

[20] K. He, X. Zhang, S. Ren, J. Sun. “Deep Residual Learning for Image Recognition”, In CVPR, 2016

[21] D. P. Kingma, J. Ba. “Adam: A Method for Stochastic Optimization”, In ICLR, 2015