Deep Learning for Generic Object Detection: A Survey

Li Liu  Wanli Ouyang  Xiaogang Wang  Paul Fieguth  Jie Chen  Xinwang Liu  Matti Pietikäinen

1 Sequence

Object detection (Fig. 1) is a computer vision problem that has been tackled for decades. In recent years, deep learning has achieved remarkable results. Objects in images are detected and classified with high precision. This has led to the creation of various object detection algorithms. In this survey, we cover the history of object detection, from early approaches to recent advances in deep learning. We discuss the performance of object detection algorithms on various datasets, including PASCAL VOC, MS COCO, and ILSVRC.

Fig. 1 2016年－2018年の ICCV, CVPR 論文の最頻出キーワード。各単語の大きさはそのキーワードの出現頻度に比例。物体検出 (object detection) が近年大きな注目を集めていることが分かる。

Fig. 2 物体検出には、特定の物体インスタンスの位置推定（上）と、一般的な物体カテゴリーの検出（下）が含まれる。本サーベイでは、後者の一般物体検出問題に関する近年の進歩に焦点を当てている。

Fig. 3 近年の検出性能の概要、2012年の基準法を基準に、微小な平均精度（mean average precision）の大幅な向上が見られる。

(a) VOC 2007－2012 コンペティション結果の検出性能。
(b) ILSVRC 2013－2017の検出性能のその性能。

Communicated by Bernt Schiele. © The Author(s) 2019

日本語訳 第 0.2 版（2020 年 2 月 2 日）

概要

物体検出は、コンピュータビジョンにおける最も基本的なかっ

チャレンジングな問題の一つであり、多数の既定カテゴリーから

自然画像内の物体インスタンス位置を見つけるタスクである、

深層学習技術は、データから特徴表現を直接学習するための強

力な戦略として登場し、一般物体検出の分野に顕著なブレーク

スルーをもたらした。本論文の目的は、深層学習技術により急

速な進化を遂げた本分野の近年の成績について、包括的なサー

ベイを提供することである。本サーベイでは、一般物体検出の

多くの側面（検出フレームワーク、物体特徴表現、物体領域候

補生成、コンテキストのモデル化、訓練戦略、評価指標）をカ

バーサーする300本以上の論文をまとめ、最後に将来の研究の有望

な方向性を示す。

キーワード：物体検出、深層学習、曇り込みニューラルネット

ワーク、物体認識
2012年，Krizhevskyら[140]はAlexNetと呼ばれる深層畳み込みニューラルネットワーク（Deep Convolutional Neural Network, DCNN）を提案し，ImageNet Large Scale Visual Recognition Challenge（ILSVRC）[234]で記録破りの画像分類精度を達成した。それ以来コンピュータビジョンの研究は，一般物体検出[85, 99, 84, 239, 230]の領域を含めほぼ全面的に，特に深層学習手法に重点を置いている。図3のように目覚ましい進歩が達成されたが，過去5年間にわたる本テーマの包括的サーベイを我々は把握していない，非常に急速な進歩を受け，本稿では，現在の一般物体検出の全景をより明確に把握のため，近年の進歩の追跡とその成果の要約を試みる。

1.1 先行レビュー論文との比較

表1に示すように，注目すべき物体検出サーベイはこれまでにも多数発表されてきた。歩行者検出[66, 79, 59]，顔検出[294, 301]，車両検出[258]，テキスト検出[295]など，特定物体検出問題に関する優れた調査が多数行われてきた。一方，一般物体検出問題に直接焦点を当てた近年の調査は比較的少ない，例として物体クラス検出のトピックにおけるサーベイを行ったZhangら[310]の論文が挙げられる。しかし，[310]や[91, 5]でレビューされている研究は，ほとんどが2012年以降の研究である。つまり，それらの研究は，深層学習とその関連手法が著しく成功し成長する以前よりも前の研究である。

深層学習は非常に複雑で微細で抽象的表現の学習を可能とし，視覚認識・物体検出・音声認識・自然言語処理・医用画像解析・創薬・ゲノミクスなどの幅広い問題における社会的な発展を推進している。様々な種類の深層ニューラルネットワークの中で，DCNN[148, 140, 149]が画像・動画・音声・音響の処理にプレクスールをもたらした，深層学習におけるサーベイも多数発表されている。例えば，Bengioら[13]，LeCunら[149]，Litjensら[170]，Guら[92]のサーベイ，より最近ではICCVやCVPRのチュートリアルが挙げられる。

対照的に，深層学習ベースの物体検出手法が多数提案されていているにも関わらず，それらの近年の包括的なサーベイ我々は把握していない。既存研究の徹底的レビューと要約は，物体検出の新しい進歩のため，特に本分野に参入したい研究者のために不可欠である。本論文の焦点は一般物体検出である。顔検出[154, 306, 116]，歩行者検出[307, 109]，車両検出[322]，交通標識検出[329]など，特定物体検出用DCNNの研究は考慮しない。

1.2 サーベイの対象範囲

深層学習ベースの一般物体検出論文は息をのむほど数が多く，最も先端技術（state-of-the-art）の包括的レビューやシナリオを編集しようとすると妥当な長さの論文に収まらない，そのため，単純な選択基準を定める必要があり，本サーベイではトピック・ジャーナル論文とトッグカンファレンス論文に焦点を絞ることとした。この制限により本サーベイに含まれていない論文の著者には心よりお詫び申し上げる。関連トピックの研究のサーベイについては，表1の論文を参照されたい，本サーベイの過去5年間に黒静止物体検出の主要な進歩を焦点にしており，重要テーマの動的物体検出は今後個別に検討すべきトピックとして残っている。

本論文の主目標は，深層学習ベースの一般物体検出技術の包括的サーベイを提供し，主に，一般的データセット，評価指標，コンテキストのモデル化，物理領域候補生成手法に基づき，ある程度の分類法と高レベルの視点・体系化を提示することで，本稿で提示する分類によって，多様な戦略の類似度，相違点を読者に理解しやすくすることが，その目的である，提案する分類法は，現在の研究を理解し将来の研究に対する未解決問題を特定するための枠組みを研究者に提供する。

図4 一般物体検出関連の問題の解釈．(a) 画像レベルの物体分類，(b) bounding box レベルの物体検出，(c) 画面単位セマンティックセグメンテーション，(d) インスタンスレベルセマンティックセグメンテーション。

次節以降の本論文の構成は以下の通りである。2節で関連研究の背景と過去20年間の進歩をまとめ，3節で深層学習について簡単に紹介する。4節でデータセットと評価基準をまとめ，5節で物体検出フレームワークのマイストーンについて説明する。6節から9節では物体検出器設計に関する基本的な部分問題と関連する論点について議論する。最後に10節で，物体検出に関する総合的な議論，最先端（state-of-the-art）の性能，そして将来の研究の方向性を述べ本論文を締める。

2 一般物体検出

2.1 問題

一般物体検出（generic object detection）は，一般物体カテゴリ検出・物体クラス検出・物体カテゴリ検出とも呼ばれる[310]，次のように定義される。画像が与えられたとき，所定カテゴリ（通常は多数のカテゴリで，例えばILSVRC物体検出チャレンジでは200カテゴリ）の物体インスタンスがあるかどうかを判定し，存在する場合は各インスタンスの空間的な位置と範囲を返す（顔・歩行者・車など）所定カテゴリの対象狭い特定物体カテゴリ検出とは対比的に，幅広い自然カテゴリの検出に重点が置かれている。私たちの生きる視覚世界には何千もの物体があるが，現在の研究コミュニティに於いて，（空・雲・草などの）構造化されていないシーンではなく，（車・自転車・飛行機などの）高度に構構化された物体や（人間・牛・馬などの）関節を持つ物体の位置指定（localization）に関心を向けている。

物体の空間的な位置と範囲は，図4のように，bounding box（以下bboxやBBと記す）[68, 234]で，主に正確な画角ごとの領域分割マスク[310]や閉じた境界[166, 235]で定義できる。ここでbboxは，びっくりと物体を囲む各辺が正方形である，我々の知る限り，一般物体検出アルゴリズムの評価において現在の文献で最も広く使用されているのはbboxであり[68, 234]，本サーベイでもそのようなアプローチを採る。ただし，研究コミュニティは，（画像レベルの物体分類から，単一物体の位置情報，一般物体検出，画素単位の物体セグメンテーションなど）より深いシーン理解に向けているため，将来の課題は画素レベル[166]にあると予想される。

※注：Axis-Aligned Bounding Box（AABB）と呼ばれる，2次元画像の一般物体検出ではAABBが用いられることが多いが，位置と範囲を粗く表す方法はトスカやデータセットによる。
A Review

On Road Vehicle Detection: A Review

Monocular Pedestrian Detection: Survey and Experiments [66] 2009 PAMI 3 种的行者検出器の評価
Survey of Pedestrian Detection for Advanced Driver Assistance Systems [79] 2010 PAMI 先進運転支援システム用歩行者検出のサーベイ
Pedestrian Detection: An Evaluation of the State of The Art [59] 2012 PAMI 単眼画像検出器の徹底的かつ詳細の評価
Detecting Faces in Images: A Survey [294] 2002 PAMI 単一顔からのは検出の最初のサーベイ
A Survey on Face Detection in the Wild: Past, Present and Future [301] 2015 CVIU 2000 年以降の“in the wild”（非制御環境、非真実環境）検出のサーベイ
On Road Vehicle Detection: A Review [258] 2006 PAMI ビジュアルベース道路車両検出システムのレビュー
Text Detection and Recognition in Imagery: A Survey [295] 2015 PAMI カラー画像中のテキストの検出、認識に関するサーベイ
Toward Category Level Object Recognition [215] 2007 書籍 物体の分類・検出・領域分割に関する代表的論文
The Evolution of Object Categorization and the Challenge of Image Abstraction [56] 2009 書籍 物体分類の 40 年間の進化の軌跡

Context based Object Categorization: A Critical Survey [78] 2010 CVIU 物体分類用コンテキスト情報のレビュー
50 Years of Object Recognition: Directions Forward [5] 2013 CVIU 物体認識システムの 50 年間の進化のレビュー
Visual Object Recognition [91] 2011 指導書 物体認識（インテリジェント認識・キャプチャ認識）技術
Object Class Detection: A Survey [310] 2013 ACM CS 2011 年以降の一般的物体検出手法のサーベイ
Feature Representation for Statistical Learning based Object Detection: A Review [160] 2015 PR 統計的学習ベースの物体検出における、ハンドクラフト特徴・深層学習ベース特徴を含む特徴表現手法

Salient Object Detection: A Survey [19] 2014 arXiv 顕著物体検出のサーベイ
Representation Learning: A Review and New Perspectives [13] 2013 PAMI 確率モデル、自己符号化器、多様体学習、深層ネットワークなどによる、教師無しの学習と深層学習

Deep Learning

A Survey on Deep Learning in Medical Image Analysis [170] 2017 MIA 医用画像解析における物体分類・物体検出・領域分割・画像レジストレーション用深層学習のサーベイ
Recent Advances in Convolutional Neural Networks [92] 2017 PR CNN の近時の進歩とコンピュータビジョン・音声・自然言語処理における応用に関する雑誌のサーベイ
Tutorial: Tools for Efficient Object Detection － 2015 ICCV15 近時のマイリストロンのみをカバーする物体検出の短期講習
Tutorial: Deep Learning for Objects and Scenes － 2017 CVPR17 物体とシーンの視覚認識のための深層学習に関する近年の概要
Tutorial: Instance Level Recognition － 2017 ICCV17 物体検出、インテリジェントセグメンテーション・物体ポジション予測など、インテリジェントレベルの認識に関する近時の進歩についての短期講習
Tutorial: Visual Recognition and Beyond － 2018 CVPR18 画像分類、物体検出、インテリジェントセグメンテーションの背景にある手法と原理に関するチュートリアル

Deep Learning for Generic Object Detection

本論文 2019 VISI (IJCV) 一般物体検出のための深層学習の包括的サーベイ
一般物体検出と密接に関連する問題は多くある2)。図 4 (a) の物体分類（object classification, object categorization）は、画像内にある所定クラス（所定の物体クラスのセットに含まれる物体クラス）の物体の存在を評価することを目的とする。つまり、与えられた画像に一つ以上の物体クラスを割り当てることで存在を判定することであり、位置は不要である。検出は、画像内のインスタンスの位置を判定する必要があるため、分類よりもチャレンジングなタスクである。物体認識（object recognition）問題は、画像内に存在する全ての物体の識別・位置推定を行うより一般的な問題を表し、物体検出・物体分類の問題を包含する [68, 234, 198, 5]。図 4 (c) のセマンティックセグメンテーション（意味的領域分割、意味分割、semantic segmentation, semantic image segmentation）は、画像の各画素への意味的クラススケラル分割を目的としており、一般物体検出と密接に関連する。図 4 (d) のインスタンスセグメンテーション（個体領域分割、個体分割、instance segmentation, object instance segmentation）は、同一物体クラスの異なるインスタンスの区別を目的としており、そのような区別を行わないセマンティックセグメンテーションとは対照的である。

2.2 主要課題

一般物体検出の理想的な物体は、高品質・高精度かつ効率性という相反する目標を達成する汎用アルゴリズムの開発である（図 5）。図 6 に示すように、高品質な検出のためには、実世界の多彩な物体カテゴリを区別できる高い識別性と、外観（アピアランス）のクラス内変動がある中でも同一カテゴリの物体インスタンスを同一カテゴリと認識し位置推定できる高い頑健性を両立しながら、画像または動画フレーム内の物体を正確に位置推定し認識する必要がある。また、必要なメモリストレージを許容可能な範囲内に抑えつつ、検出タスク全体をリアルタイムに実行可能な高い効率が求められる。

2.2.1 精度関連の課題

検出精度の課題は、多様なクラス内変動と膨大な物体カテゴリ数に起因する。クラス内変動は、内因性因子と撮像条件の 2 種に分類できる。内因性因子に関しては、各物体カテゴリが多くの異なる物体インスタンスを含むがおり、図 6 (i) の「椅子」カテゴリのように、色・テクスチャ・素材・形状・サイズの一つ以上が異なる可能性がある。人間や馬などのように明確に定義されたクラスタえでさえ、非剛体変形や衣服により物体インスタンスは異なる

姿勢で表示される。撮像条件の変動は、lighting（夜明け・日中・夕暮れ・屋内）・物理的位置・気象条件・カメラ・背景・照明・遮蔽・視距離など非制約環境が物体の視覚的な影響を与えることにより引き起こされる。図 6 (a)～(h) に示す照明・姿勢・スケール・遮蔽・混雑・陰・プラー動作などのように、これらの全条件が物体の視覚的な変化をもたらす。また、デジタル化によるアーティファクト、ノイズによる破損、解像度の低下、フィルタリングによる歪みによって、課題は更に更に難しいとなる。クラス内変動に加えて、10^4 ～ 10^10 のオーダーの多数の物体カテゴリの検出には、図 6 (j) に示すような微妙に異なるクラスタ間変動を区別できる高い識別性を必要とする。実際には現在の検出者が、主に構造化された物体のカテゴリに焦点を当てており、例えば、PASCAL VOC [68], ILSVRC [234], MS COCO [166] のカテゴリ数は順に 20, 200, 91 である。既存のベンチマークデータセットの考慮する物体カテゴリ数は、明らかに人間が認識可能な数よりもはるかに少ない。

2.2.2 効率とスケーリビリティに関連する課題

ソーシャルメディアネットワークとモバイルデバイス・ウェアラブルデバイスの普及により、視覚データ分析の需要が高まっており、モバイルデバイス・ウェアレルデバイスの計算能力とストレージ容量は限られているため、効率的な物体検出が必要である。

効率性が課題となるのは、位置推定と認識を両方行う必要があるためである。計算複雑性は物体カテゴリ数（多いかもしれない）に応じて増大し、また、図 6 (c, d) のように、単一画像内の位置とスケールの数（非常に多いかもしれない）に応じて増大する。

更にスケーリビリティの課題もある。検出器はこれまで見たことのない初見の物体や未知の状況を、高いデータレートで処理できる必要がある。画像数とカテゴリ数の増加が続くと、それに伴うメモリの増加や計算時間の増加になる。

2) 我々の知の限り、視覚認識の様々な部分タスクの定義に関して文献間の明確さが不足しているが、基本的分類に定義される分類は、一応、以下の 6 種に分類できる。detection, localization, recognition, classification, categorization, verification, identification, annotation, labeling, understanding などの定義は、一応、上記分野の定義がなされる [5]。
図7 物体の検出・認識のマイルストーン．特徴表現 [47, 52, 101, 140, 147, 178, 179, 212, 248, 252, 263, 276, 279]，検出フレームワーク [74, 85, 239, 271, 276]，データセット [68, 166, 234] のマイルストーンを示している．2012年まではハンドクラフト特徴に支配されていたが，2012年にKrizhevskyら [140] が開発した画像分類用DCNNを転用に，2012年以降の手法は深層ネットワークに支配方である．記載手法の多くは多数引用されており，ICCVやCVPRの主要な賞を受賞している．詳細は2.3節を参照されたい．

図8 (a) 典型的なCNNで繰り返し適用される3層の演算：多数の線形フィルタによる幅込み，非線形性（例：ReLU），局所ブーリング（例：最大ブーリング）．前層からのM枚の特徴マップは，N枚の異なるフィルタ（ここではサイズ3×3×M）でストライド1で幅込みされる．結果として得られたN枚の特徴マップは，ReLUなどの非線形関数を介して逆算される（例えば2×2の領域の最大値を取ることでブーリングされ，解像度の減じてN枚の特徴マップを提供する．(b) VGGNet [248] のアーキテクチャ（VGG-11）．重みを持つ層が11層ある典型的なCNN．3つのカラーチャンネルを持つ画像が入力として与えられる．ネットワークは，8つの幅込み層，3つの全結合層，5つの最大ブーリング層，1つのsoftmax分類器を持つ，最後3層の全結合層は，最後の幅込み層の出力をベクトル化したベクトルを入力として受け取る．最後の層，クラス数Cのsoftmax関数である．ネットワーク全体は，レベル付き訓練データを用い，確率的勾配降下（Stochastic Gradient Descent，SGD）により目的関数（例：平均二乗誤差・交差エントロピー損失）を最適化することで学習できる．（カラー図はオンラインで参照可．）

2.3 過去20年間の進歩
物体認識に関する初期の研究は，テンプレートマッチング技術と単純なパーソナーモデル [76] に基づいており，空間的な配置がほぼ固定されている頚などの特定物体に焦点を当てていた，1990年代以前は，物体認識の主なパラダイムは幾何学的表現に基づいていた [190, 215]．その後，幾何や事前モデルから，外観特徴 [191, 236] 基于統計的分類器（例：ニューラルネットワーク [233]，SVM [201]，Adaboost [276, 290]）へと焦点が移行した．この成功した物体検出器群により，本分野の多数の後続研究のための準備が整えられた．

近年の物体検出のマイルストーンを，2つの主要な時代（SIFT vs. DCNN）を強調して図7に示す．外観特徴は大域表現 [192, 260, 267] から，平行移動・スケール・回転・照明の視点・遮蔽の変化に不変であるように設計された局所表現へと移行した．ハンドクラフトの局所不変特徴は，Scale Invariant Feature Transform（SIFT）特徴 [178] に降非常に大き
な人気を博した．様々な視覚認識タスクの進歩は概ね，Haar-like特徴 [276]，SIFT [179]，Shape Contexts [12]，Histogram of Gradients（HOG）[52]，Local Binary Patterns（LBP）[196]，region covariances [268] などの局部記述子 [187] の使用に基づっていた．これらの局所特徴は通常，単純に結合されるか，Bag of Visual Words [232, 47]，BoWモデルのSpatial Pyramid Matching（SPM）[147]，Fisher Vectors [212] などの特徴ブーリングエンコーディングにより集約される．

ハンドクラフト局所記述子と識別的な分類器による手動調整の多段バイバンドは，物体検出を含むコンピュータビジョンの様々なメソッドを長年にわたり支配した．DCNN [140] が画像分類で記録破りの結果を達成した2012年の重要転換点まで，その支配方は変わなかった．

検出・位置推定でのCNNの使用 [233]は1990年代にさかのぼることことができ，使用される分野はあまり多くなかった [272, 233, 238] が，検出などの制限されたドメインで
成功を収めた。しかし最近になって、もっと深いCNNがより一般的な物体カテゴリの検出に記録破りの改善をもたらし、DCNNの画像分類での成功[140]が物体検出に転用され、マイルストーンであるGirshickら[85]のRegion-based CNN（RCNN）検出器と繋がった。

深層検出器の成功は、大量の訓練データと、数百、時には数十億ものパラメータを持つ大規模ネットワークを大きく依存している。非常に高計算能力を持つGPUと、ImageNet[54], 234]やMS COCO[166]などの大規模検出データセットが利用可能なことが、それぞれの成功に重要役割を果たしてきた。大規模データセットにより、大きなクラス内変動とクラス間類似性を有する画像を用いて、より実際の複雑な問題を対象とする研究が可能となった[166, 234]。ただし、正確なアノテーションを得るには多くの手間が必要となる。そのため、アノテーションの難しさを緩和したり、より小さい訓練データセットでの学習を可能にする手法を検討する必要がある。

多数の物体カテゴリを検出する能力が人間に匹敵する、汎用物体検出システムの構築というチャレンジングな目標に研究コミュニティは向け始めており、これは大きな課題である。認知科学者によると、人間は約3,000のエントリレベルカテゴリ（entry-level category）を、全体で30,000の視覚カテゴリを識別できる。ドメインの専門知識で区別できるカテゴリの数は10^5オーダーにも及び[15]。過去数年で目覚ましい進歩があったものの、10^4～10^5カテゴリで人間レベルの性能に匹敵する正確・効率的な検出・認識システムを設計することは、間違いないと未解決問題である。

3 深層学習の簡単な紹介

深層学習は、画像分類・動画処理から音声認識・自然言語理解に至るまで幅広い機械学習タスクに革新を与えた。このようにも急速な進化を踏まえて、深層学習に関する近年のサーベイ論文が多数ある[13, 89, 92, 149, 170, 216, 287, 297, 313, 320, 325]。これらのサーベイ論文での深層学習技術のレビューは、異なる視点から[13, 89, 92, 149, 170, 216, 287, 320]考えられている。医用画像解析[170]、自然言語処理[297]、音声認識システム[313]、リモートセンシング[325]への応用について行われている。

最も代表的な深層学習モデルである畳み込みニューラルネットワーク（Convolutional Neural Network, CNN）は、自然界の信号に抗する基本特徴である、平行移動不変性、局所連続性、組成上の階層性[149]を利用できる。図8に示すように、典型的なCNNは階層構造を有し、複数の抽象化レベルでデータの表現を学ぶためのいくつかの層で構成されている[149]。まず、畳み込み

\[ x^{(1)} = w^0 \]  

(1)

を考える。ここでは、前層である1－番目の層からの1枚の入力特徴マップ\( x^{(1)} \) がある。1枚の2次元畳み込みフィルタ（フィルタや重みとも呼ばれる）\( w^0 \) で畳み込まれている。この畳み込みは、非線形演算を \( \sigma \) として次のように一連の層で適用される。

\[ x^{(1)}_j = \sigma \left( \sum_{i=1}^{N^2} x^{(1)}_i \ast w^{(1)}_{i,j} + b^{(1)}_j \right) \]  

(2)

ここで、畳み込みは \( N^2 \) 枚の入力特徴マップ\( x^{(1)}_i \) と対応するカーネル\( w^{(1)}_{i,j} \)で行われ、\( b_j \)はバイアス項である。要素ごとの非線形関数\( \sigma() \)は、典型的には各要素に対するrectified linear unit（ReLU）であり。

\[ \sigma(x) = \max(x, 0) \]  

(3)

で計算される。最後に、ブーリングは特徴マップのダウンサンプリングまたはアップサンプリングに対応する。これら3種の演算（畳み込み、非線形性、ブーリング）を図8(a)に示す。多数の層を持つ「深い」CNNはDeep CNN（DCNN）と呼ばれる。典型的なDCNNアーキテクチャを図8(b)に示す。

CNNのほとんどの層は、各画素がニューロンのように機能する多数の特徴マップで構成される。畳み込み層の各ニューロンは、重みのセット\( w^{(1)}_{i,j} \)（基本的には2次元フィルタのセット）を介して前層の特徴マップに接続される。図8(b)のように、CNNの前の層（低層）は通常畳み込み層とブーリング層で構成され、後の層（高層）は通常全結合されている。前層の層から後の層まで入力画像は繰り返し畳み込まれ、各層で収容野（サポート領域）が増加する。一般に、CNNの最初の数層はエッジなどの低レベル特徴を抽出し、後方の層が複雑性の増しより一般的な特性を抽出する[303, 13, 149, 199]。

DCNNには多くの傑出した利点があり、複数レベルの抽象化でデータの表現を学習する階層構造を持ち、非常に複雑な数を数理化する能力があり、最小限のドメイン知識でデータから直接かつ自動的に特徴表現を学習できる。特に、大規模なラベル付きデータセットと、非常に高い計算能力を持つGPUを利用可能になったことが、DCNNを成功させた。

大成功の一方で既知の欠陥が残っている。特に、非常に多くのラベル付き訓練データと高価な計算リソースが必要である。適切な訓練バランマタとネットワークアーキテクチャの選択には、依然としてかなりのスキルと経験が必要である。また、訓練されたネットワークの、解釈性の不足、劣化に対する頑健性の欠如、攻撃に対する脆弱性の脆弱性[88]が、DCNNの実世界応用の制約となっている。

4 データセットと性徵評価

4.1 データセット

データセットは、競合するアルゴリズムの性能を測定し比較するための共通の場としてだけでなく、より一層複雑でチャレンジングな問題へと分野を押し進めめるものとして、物体認識研究の歴史の中で重要な役割を果たしてきた。特に近年では、深層学習技術が多様な視覚認識問題に大成功をもたらしたが、その成功に重要な役割を果たしたのが大量のアノテーション付きデータである。インターネット上での多数の画像へのアクセスが、豊富で多様な物体を捉える包括的なデータセットの構築を可能にし、空間の物体認識性能を実現している。

一般物体検出のために、PASCAL VOC[68, 69]、ImageNet[54], MS COCO[166], Open Images[143]という4つの有名データセットがある。これらのデータセットの目的は、表2にまとめ、選択したサンプル画像を図9に示す。大规模なアノテーション付きデータセットの作成には以下のような3つの工程がある。まず対象物体をデータベースに決定し、次に選択したカテゴリを表す複数の候補画像をインターネット上で収集し、最後に（通常はクラウドソーシング）で収集した画像にアノテーションを付ける。スケジュールの分から、これらのデータセットの構築と特性に関する詳細な説明については原著論文[68, 69, 166, 234, 143]を参照された。

4つのデータセットは、それぞれの検出チャレンジの根幹をなす。各チャレンジは、公開画像データセットと、正解（ground truth）のアノテーション、標準化された評価ソフトウェア、および年次コンペティションとそれに対応するワークショップで
表 2 物体認識用の一般的なデータセット、PASCAL VOC、ImageNet、MS COCO、Open Images の画像例を図 9 に示す。

| データセット名   | 総画像数 | カテゴリ数 | 画像あたりの物体数 | 画像サイズ | 開始年 | 注目点 |
|----------------|----------|------------|------------------|-----------|--------|--------|
| PASCAL VOC (2012) [69] | 11,540   | 20         | 2.4              | 470×380   | 2005   | 日常生活で一般的な 20 カテゴリのみカバー。多数の訓練画像、実世界のアプリケーションに近い。クラス内変動が非常に大きい。シーンコンテキスト内の物体、1枚の画像に複数の物体、多数の難しいサンプルを含む。 |
| ImageNet [234]    | 1400万+  | 21,841     | 1.5              | 500×400   | 2009   | 多数の物体カテゴリ、画像あたりの物体のインスタンス数とカテゴリ数が多い、PASCAL VOC よりチャレンジング、ILSVRC チャレンジの枠組み。画像は物体が中心。 |
| MS COCO [166]     | 32.8万+  | 91         | 7.3              | 640×480   | 2014   | 実世界のシーンに更に近い、各画像により多くの物体インスタンスとなり、豊富な物体アノテーション情報が含まれる。ImageNet データセットでは使用できない物体セグメンテーションのアノテーションデータを含む。 |
| Places [319]       | 1000万+  | 434        | -               | 256×256   | 2014   | シーン認識用の最大のラベル付きデータセット、Places365 Standard、Places365 Challenge、Places 205、Places88 の 4 つのサブセットをベンチマークとして使用。 |
| Open Images [143]  | 900万+   | 6,000+     | 8.3              | 多様    | 2017   | 画像レベルのラベル、物体のbbox、visual relationship（物体間関係、視覚的関係）のアノテーションが含まれる。Open Images V5 は大規模な、物体検出、物体インスタンスセグメンテーション、visual relationship detection（物体間関係検出、視覚的関係検出）をサポートしている。

図 9 PASCAL VOC、ILSVRC、MS COCO、Open Images の物体アノテーション付きの画像の例。データセットの概要については表 2 を参照された。

構成されている。検出チャレンジに関する訓練・検証・テスト（training, validation, testing）データセットの画像・物体インスタンスの数の統計を表 3 に示す。VOC, COCO, ILSVRC, Open Images の各検出データセットでの最頻出物体クラスを表 4 に示す。

PASCAL VOC [68,69] は、複数年にわたり作成・保守に努力が捧げられた、分類や物体検出のための一連のベンチマークデータセットであり、年次コンペティションの形で認識アルゴリズムの標準化された評価を行う前例を作った。2005 年にわずか 4 カテゴリから始まり、その後データセットは日常生活での一般的な 20 カテゴリに増加した。2009 年以降画像数は毎年増加しているが、テスト結果を年ごとに比較できるよう過去の全画像が維持されている。ImageNet、MS COCO、Open Images などの大規模データセットが利用可能になったことで、PASCAL VOC は徐々に時代遅れになっている。

ILSVRC（ImageNet Large Scale Visual Recognition Challenge）[234] は ImageNet [54] から派生し、検出アルゴリズムの標準化された訓練・評価という PASCAL VOC の目標を、物体クラス数・画像数の面で 1 柄以上拡大する。ImageNet1000 は、1000 個の異なる物体カテゴリと合計 120 万枚以上の画像を含む ImageNet 画像のサブセットであり、ILSVRC 画像分類チャレンジの標準ベンチマークを提供するために定められた。

MS COCO [166] は、ImageNet データセットへの批判に対応し、より豊かな画像理解を推進するために作成されたデータセットである。ImageNet データセットは物体が大きく中心に映っていることが多い、実世界のシナリオを代表していないのに対し、COCO データセットは複雑な日常シーンを含んでおり、一般的な物体が実生活に近い自然なコンテキストで映っている。更に、より正確な検出器評価のため、完全にセグメンテーションされたインスタンスで物体がラベル付けされている。

COCO 検出チャレンジ [166] は、bbox 出力のタスクと物体セグメンテーション出力のタスクがあり、3 つの新しい課題を導入した。

1. 幅広いスケールの物体が含まれ、小さな物体の割合が高い [249]。
2. 物体はあまりアイコンのように（大きく中心に）映っておらず、複雑な背景の中である。
3. より正確な物体位置推定を奨励する評価指標（表 5）を採用。

ImageNet 同様、MS COCO は今日の物体検出の標準となっていて、OICOD（Open Image Challenge Object Detection）は、Open
Images V4（2019年現在はV5）[143]から派生した現在最大の公開機体検出データセットである。OICOD は、ILSVRC や MS COCO のようなそれまでの大規模機体検出データセットと比べ、クラス数、画像数、bbox アノテーション数、インスタンスセグメンテーションのマスクアノテーション数が大幅に増加しているだけでなく、アノテーション工程に達している。ILSVRC と MS COCO ではデータセット内の全クラスのインスタンスに徹底的にアノテーションが付与されるが、Open Images V4 では各画像に分布異なる画像数に相当したクオリティの高いアノテーションが付与されているが、そのため OICOD では、正しいラベルだと人間が確認した機体インスタンスにのみアノテーションが付けられる。

4.2 評価基準
検出アルゴリズムの性能評価基準として、Frames Per Second (FPS) での検出速度、precision (適合率)、recall (再現率) の 3 つが挙げられる。最も一般的に用いられる評価基準は、precision と recall から計算される Average Precision (AP) である。AP は通常、各機体カテゴリに対し個別に計算され、カテゴリごとに評価される、全機体カテゴリにわたる性能の比較には、全機体カテゴリの AP の平均である mean AP (mAP) が性能の最終的な尺度として採用される。これらの評価基準の詳細については [68, 69, 234, 108] を参照されたい。

テスト画像 I に適用された検出器の標準的な出力は、予測検出 \{ (b_j, c_j, p_j) \} で表せる。ここで、i は予測検出の複数の機体カテゴリ A の位置、大小、およびその信念を表す。

\[ IOU(b, b') = \frac{\text{area}(b \cap b')}{\text{area}(b \cup b')} \]  

が、所定閾値 \( \varepsilon \) 以上である。ここで、\( \cap \) は intersection（積集合、共通部分）、\( \cup \) は union（和集合）を表す。\( \varepsilon \) の代表的な値としては 0.5 が用いられる。

## Table 3

一般的に使用される機体検出データセットの統計。VOC チャレンジの機体数の統計には、評価で使用された 'difficult' でない物体数を含む非常に、アノテーションの付いた全物体数を括弧内に掲載している。（注: PASCAL VOC のアノテーションには、認識困難と考えられる物体が否かを指す 'difficult' という項目がある。）2017年より前のCOCO チャレンジでは、test セットにはそれぞれ約2万枚の画像から54决の分割（Dev, Standard, Reserve, Challenge）があった。2017年以降、train セットとval セットの配分が異なる。2017年以降、test セットには Test Dev と Test Challenge のみがあり、残り 2 つの分割は削除された。2017年と2015年とで、Test Dev/Challenge はそれぞれ同じ画像で構成されるため、異なる年数も直接比較可能なことに注意された。

| チャレンジ   | 物体クラス | 画像数 | アノテーション付き物体数 | 合計 (Train+Val) |
|------------|---------|-------|---------------------|-----------------|
| PASCAL VOC Object Detection Challenge | | | | |
| VOC07 | 20 | 2,501 | 5,102 | 6,539 |
| VOC08 | 20 | 2,111 | 4,374 | 6,300 |
| VOC09 | 20 | 3,473 | 7,108 | 10,581 |
| VOC10 | 20 | 4,998 | 10,496 | 11,494 |
| VOC11 | 20 | 5,717 | 11,234 | 16,951 |
| VOC12 | 20 | 5,717 | 11,234 | 16,951 |
| ILSVRC Object Detection Challenge | | | | |
| ILSVRCL3 | 200 | 395,909 | 395,909 | 395,909 |
| ILSVRCl4 | 200 | 456,567 | 456,567 | 456,567 |
| ILSVRCl5 | 200 | 456,567 | 456,567 | 456,567 |
| ILSVRCl6 | 200 | 456,567 | 456,567 | 456,567 |
| MS COCO Object Detection Challenge | | | | |
| MS COCO15 | 80 | 82,783 | 82,783 | 82,783 |
| MS COCO16 | 80 | 82,783 | 82,783 | 82,783 |
| MS COCO17 | 80 | 118,287 | 118,287 | 118,287 |
| MS COCO18 | 80 | 118,287 | 118,287 | 118,287 |
| Open Images Challenge Object Detection (OICOD) | | | | |
| OICOD18 | 500 | 1,643,042 | 1,643,042 | 1,643,042 |

Texts are written in Japanese.
表4 各検出チャレンジの最頻出物体クラス、各単語の大きさは、訓練データセット内のそのクラスの頻度に比例。

(a) PASCAL VOC（20クラス）（b）MS COCO（80クラス）（c）ILSVRC（200クラス）
（d）Open Images Detection Challenge（500クラス）

それ以外の場合、False Positive（FP）と見なされる。信頼度pは、予測クラスラベルが受け入れられるかどうかを判断するために、通常何らかの閾値βと比較される。

APはPrecisionとRecallに基づいて物体クラスごとに計算される。ある物体クラスに関して、あるテスト画像Iに対する検出器の検出結果を信頼度pにより順に並べて{(pj,pj)}j=1で表すこととする。各検出(pj,pj)がTPとFPのどちらであるかは、図10のアルゴリズム48で決定される。TPとFPの検出に基づいてprecision P(β)とrecall R(β)65は信頼度閾値βの関数として計算される。そのため、信頼度閾値を変化させることで異なる(P,R)のペアが得られ、原則的にはprecisionをrecallの関数P(R)として見なすことができ、そこでAverage Precision（AP）68,234を計算できる。

MS COCOの導入以来、bbox位置の精度と注目が集まっている。MS COCOは固定IOU閾値を使用する代わりに、物体検出器の性能を特徴付けるいくつかの評価基準を導入している（表5をまとめてみる）。例えば、一つのIoU 0.5で計算された従来のmAPとは対照的に、AP cocoは、全体でカテゴリー毎の平均をとるだけでなく、複数のIOU値（0.05刻みで0.5から0.95まで）で平均をとる。MS COCOの物体の41%は非常に24%は大きいため、評価基準APsmall, APmedium, APlargeも導入されている。最後に、表5にPASCAL, ILSVRC, MS COCOの物体検出チャレンジで使用される主な評価基準を、[143]で提案されたOpen Images challenges用の修正とともにまとめる。

5 検出フレームワーク

ハンダクラフト特微[276, 52, 72, 98, 275]から学習されたDCNN特微[85, 203, 84, 229, 50]への劇的な変化から分かるように、認識のための物体特徴表現と分類器には着実な進歩があった。対照的に位置推定に関しては、全探索の回避[145, 271]も試みられているが、基本的な「スライディングウィンドウ」戦略[52, 74, 72]が主流のままである。しかし、ウィンドウの数は多く、画素数の二乗オーダーで増加する上、複数のスケールとアスペクト比で探索する必要がある程度探索空間は更に増加する。したがって、効率の良い効果的な検出フレームワークの設計は、この計算コストを削減する上で重要な役割を果たす。一般的に採用される戦略として、カスケード化、特徴計算の共有、ウィンドウごとの計算の削減が挙げられる。

本節では図11と表11に記載している検出フレームワークのレビューを行う。深層学習の本分野参入以降のマイルストーンアップローチは、以下の2つの主要カテゴリに分類される。

a. 物体領域候補生成のための前処理工程を含む、2段階の検出フレームワーク。
b. 物体領域候補生成の工程を分離しない単一の提案手法を持つ、1段階の検出フレームワーク（物体領域候補生成が不要なフレームワーク）。

5.1 領域ベース（2段階）フレームワーク

領域ベースのフレームワークでは、カテゴリ非依存の物体領域候補99を画像から生成し、それらの領域からCNN[140]特微を抽出した後、カテゴリ特化の分類器を使用して物体領域候補のカテゴリラベルを決定する。図11から分かるように、ほぼ同時期に独自にDetectorNet[281], OverFeat[239], MultiBox[67], RCNN[85]が一般物体検出のためのCNNの使用を提案した。

RCNN[85]: Girshickらは一般的物体検出用CNNの探求の初期期に、AlexNet[140]とselective searchによる物体領域候補生成[271]を組合せたRCNNを開発した[85,87]。これは、CNNによって得られた画期的な分類結果と、ハンダクラフト特微での物体領域候補生成におけるselective searchの成功[271]を触発されたものである。図12に詳細に示すように、RCNNフレームワークの訓練は多段階バイプライドで構成される。

1. 物体領域候補の計算：selective search[271]によりクラス非依存の物体領域候補（物体を含む可能性がある候補領域）を得る。
2. CNNモデルのfine-tuning：物体領域候補は、画像からクロップされ同じサイズにワルプされてから、ImageNetなどの大规模データセットで事前学習されたCNNモデルをfine-tuningするための入力として使用される。この段階では正解形状とのオーバーラップがIOU100.5以上である全物体領域候補が、正解形状のクラスに対してはpositiveと定義され、残りはnegativeと定義される。
3. クラス特徴SVM分類器の訓練：CNNで抽出された固定長特徴を使用して訓練されたクラス特徴の線形SVM分類

99 物体領域候補（物体提案、領域提案、検出提案とも呼ばれる）は、物体を含む候補である。画像内の領域またはbboxのセットである[27,110]。
100 IOUの定義については4.2節を参照された。
表5 物体検出器評価のために一般的に使用される評価基準の要約。

| 評価基準 | 意味 | 定義と説明 |
|---------|------|------------|
| TP      | True Positive | 図10によるtrue positive検出（真陽性の検出、正検出、正検知） |
| FP      | False Positive | 図10によるfalse positive検出（偽陽性の検出、誤検出、偽検知、過検出、過検知） |
| β      | 信頼度閾値 | $P(β)$, $R(β)$計算用の信頼度（confidence）閾値 |
| $ε$    | IOU閾値 | 典型的に0.5前後 |
| ILSVRC | min(0.5, $\frac{w}{h}$)ここで, $w \times h$は正解箱のサイズ |
| MS COCO | 10個のIOU閾値 $ε \in \{0.5:0.05:0.95\}$ |
| $P(β)$ | Precision | 少なくとも$β$の信頼度で検出器から出力された検出の総数のうち、正しい検出数の割合 |
| $R(β)$ | Recall | $N_s$個の全検出数のうち、検出器によって少なくてごく$β$の信頼度で検出された検出数の割合 |
| AP     | Average Precision | 信頼度閾値$β$を変化させて得られた様々なレートのrecallにわたって計算される |
| mAP    | mean Average Precision | VOC 単一のIOUでのAPを全クラスにわたって平均した値 |
| AR     | Average Recall | 画像あたり所定数の検出が許される場合のrecallの最大値を、全てのカテゴリとIOU閾値で平均した値 |
| AR     | Average Recall | MS COCO で $AP_{coco}=10$個のIOU：{0.5:0.05:0.95}でのmAPを平均した値 |

図11 一般物体検出のマイルストーン。
図 12 RCNN 検出フレームワーク [85, 87] の図解。

補に対する固定長の特徴を生成できるようになり、RCNN は検出品質を犠牲にすることなく著しく高速化した。SPPNet は RCNN の評価を数倍加速するが、検出器の訓練には同等の高速化をもたらす。また、SPPNet での fine-tuning [99] は SPP 層までの拡張を必要とせず、非常に深いネットワークの精度が制限される。

Fast RCNN [84]：Girshick は、RCNN と SPPNet のいくつかの短所に対処しつつ検出の速度・品質を改善する Fast RCNN [84] を提案した。図 13 に示すように、Fast RCNN は RCNN や SPPNet のように softmax 分類器・SVM・bbox 回帰器を個別に訓練するのではなく、softmax 分離器とクラス依存の bbox 回帰を同時に学習する合理化された訓練工法により、end-to-end の検出器訓練を可能にする。Fast RCNN は、物体領域候補全体で最高精度を計算するというアイディアを採用し、最後の CONV 層と最初の FC 層の間に Region of Interest（RoI）pooling 層を追加して各物体領域候補の固定長特徴を抽出する。本質的に、RoI pooling は特徴レベルでのワークイングを使用して画像レベルでのワークイングを近似するものである。

RoI pooling 層後の大特徴は一連の FC 層に送られ、その最後で 2 つの兄弟出力層に分岐し、物体タグリオ予測の softmax 確率と、物体領域候補部位のクラス依存の bbox 回帰信号とを出力する。RCNN/SPPNet と比較して Fast RCNN は効率を大幅に改善し、通常は訓練で 3 倍、テストで 10 倍高速である。また、検出品質はより高く、単一の訓練工法でネットワークの全層を更新でき、特徴のキャッシュング用のストレージが必要となる。

Faster RCNN [229, 230]：Fast RCNN は検出処理を大きく高速化したが、外部の物体領域候補生成への依存は続いており、その計算が Fast RCNN の新たな速度ボトルネックとなっていた。近年の研究により、CNN の CONV 層には物体の位置推定を行う顕著な能力があり [317, 318, 46, 200, 97]。FC 層ではその能力が弱まっていることが示されている。したがって、物体領域候補生成の selective search を CNN で代替することが可能。Ren ら [229, 230] により提案された Faster RCNN フレームワークは、物体領域候補（領域提案）生成用に効率的で
正確な Region Proposal Network（RPN）を提示した。図 13 に示すように，物体領域候補生成用の RPN と領域分類用の Fast RCNN のタスクを実行するために，同一の backbone ネットワークが利用され最後の共有畳み込み層からの特徴が使用される。

RPN はまず，CONV 特徴マップの各位置で，異なるスケールとアスペクト比の k 個の参照矩形（つまり，いわゆる anchor（アンカー））を初期化する。anchor の所定位置は画像コンテンツに非依存だが，anchor から抽出された特徴ベクトル自体は画像コンテンツに依存する。各 anchor は低次元ベクトルとしてマッピングされ，2 つの兄弟 FC 層（物体カテゴリ分類層と box 回帰層）に供与される。Fast RCNN での検出とは対照的に，RPN で回帰に使用される特徴は anchor box と同一形状であるため，anchor が k 個あれば回帰器も k 個となる。RPN は Fast RCNN と CONV 特徴を共有するため，物体領域候補生成の非常に効率的な計算が可能となる。RPN は事実上 Fully Convolutional Network（FCN，全層畳み込みネットワーク）[177，241] の一種であり，そのため Faster RCNN はハンドクラフト特徴を使用しない純粋な CNN ベースフレームワークである。

VGG16 モデル [248] の場合，Faster RCNN は（全ステージ込みで）GPU で 5 FPS でテストできるより，画像あたり 300 個の物体領域候補を使用して PASCAL VOC 2007 で最前述の物体検出精度を達成する。最初の Faster RCNN [229] にはいくつかの相互の訓練段階が含まれていたが，後に単純化された [230]。

Faster RCNN の開発と同時に，Lenc と Vedaldi [151] は，selective search などの物体領域候補生成手法の役割に疑念を抱き，CNN の検出器の物体領域候補生成の役割を評価し，CNN には正確な物体検出用の幾何情報が FC 層ではなく CONV 層に十分含まれていることを示した。彼らは，CNN のみに依存する単純で高速で統合された物体検出器を構築し，selective search などの物体領域候補生成手法を排除する可能性を示した。

RFCN（Region based Fully Convolutional Network）: Faster RCNN は Fast RCNN より術外に高速だが，領域ごとのサブネットワークを RoI（画像あたり数百個）ごとに適用する必要がある。この事実が，Dai ら [50] による RFCN の提案の主な理由となり，RFCN は全層畳込み（fully convolutional）の（全結合法の層も含まない）検出器であり，ほとんどの計算が画像全体で共有される，図 13 に示すように，RFCN は RoI サブネットワークのみが Faster RCNN と異なる。Faster RCNN では RoI pooling 層後の計算を共有できないため，Dai ら [50] は共有 RoI サブネットワークの構築のために全層 CONV 層を使用し，予測出力の直前のある最終 CONV 層の特徴自体 RoI クロップを取得することを提案した。しかし，この単純化された設計は検出精度がかなり劣ることが判明した [50]。これにより，深い CONV 層ほどカテゴリの意味に敏感であり平行移動に敏感でないが，物体検出には平行移動による変化を尊重する位置推定用表現が必要であるため推測される。Dai ら [50] はこの観察に基づいて，特殊な CONV 層のセクトルを FCN 出力とし，使用してposition-sensitive score map のネットワークを構築し，この上に position-sensitive RoI pooling 層を追加した。RFCN と ResNet101 [101] を用いると，Faster RCNN に匹敵する精度を，多くの場合より高速な実行で達成できることが示されてい

Mask RCNN: He ら [102] は，画像単位の物体インスタンスセグメンテーションに取り組むため，Faster RCNN を拡張した Mask RCNN を提案した。2 段階バイプライムを採用し，第 1 段階で RPN を使用するのは Faster RCNN 同様だが，Mask RCNN は第 1 段階で，クラスと box オフセットの予測と並行して各 RoI に対するバイナリマスクを出力するランチを追加する。新しいランチは CNN 特徴マップの上に追加される FCN [177，241] である。オリジナルの RoI pooling（RoIPool）層によって引き起こされる位置ずれを回避し，画素レベルの空間の対応を維持するため RoIAllign 層が提案された。Mask RCNN は ResNetXt101-FPN [291，167] を backbone ネットワークに用い，COCO の物体インスタンスセグメンテーションと bbox 物体検出で最高精度を達成した。Mask RCNN は訓練が簡単であり，よく汎化する，Faster RCNN にわずかなオーバーヘッドを追加するだけであり，5 FPS で実行できる [102]。

Chained Cascade Network と Cascade RCNN: カスクード [73，20，159] の本質は，多段階の分類器を使用し，後期段階がより難しい例の処理に集中できるよう早期段階で多数の簡単な負荷を排除することで，識別性の高い分類器を学習することである。2 段階の物体検出はカスクードと見なすことができ，最初の検出器は大量の背景を除去し 2 段階目は残りの領域を分類する。近年，2 つ以上のカスクード化分類器と一般物体検出用 DCNN の end-to-end 学習が，Chained Cascade Network [205] で提案され Cascade RCNN [23] で拡張された。さらに最近，物体検出とインスタンスセグメンテーションを同時に行う Hybrid Task Cascade [31] が，そのような学習を利用して COCO 2018 Detection Challenge で優勝している。

Light Head RCNN: RFCN [50] の検出を更に高速化するため，Li ら [165] は検出ネットワークの head をできるだけ軽くし RoI ごとの計算を削減する Light Head RCNN を提案した。具体的には，チャネル数の小さい薄い特徴マップ（例えば COCO では 490 チャンネル）を生成する深層込みと安定した RCNN サブネットワークを用いることで，優れた速度・精度トレードオフを実現した。

5.2 統合（1段階）フレームワーク

RCNN [85] 以来，5.1 節で述べた領域ベースプライマル溶液の戦略が支配的であったため，人間のベンチマークデータセット上でいい結果を収めた検出器を全て Faster RCNN [229] に基づいている。しかし，ストレージと計算能力が限られている現在のモバイルウェアラブルデバイスにとって，領域ベースのアプローチは計算コストが高い，そのため，複雑な領域ベースプライムの個々の構成要素の最適化を試みる代わりに，研究者は統合された検出戦略の開発を始めた。

統合プライムは，物体領域候補生成や後段の分類/特徴検索部分を端末のフレームワークで行う行われない単一のフィードフォワード手順 CNN で，クラスの確率と bbox のオフセットを直接予測するアーキテクチャを指し，全計算を单一のネットワークに入れ込む，パイプライン全体が单一のネットワークであるため，検出性能に直接影響を及ぼすend-to-endで最適化できる。

DetectoNet: Szegedy ら [261] の DetectoNet は，最初期に提案された物体検出用 CNN の一つであれ，物体検出をbbox マスクへの過渡時期として定式化した，AlexNet [140] が最後の softmax 分類層を回帰層で置き換えて使用された，画像ウィンドウが与えられると，1 つのネットワークを使用して粗いグリッドの背景画像を予測し，4 つの追加ネットワークを，物体の上半分・下半分・左半分・右半分を予測する，次に，グローバル化処理で予測マスクを，bbox に変換する，ネットワークは物体タイプとマスクタイプごとに訓練が必要があり，複雑クラスには拡張されない，DetectoNet は画像クロップを多数行い，全クラスの各部分に対して複数のネットワークを実行する必要があるため低評価である。

OverFeat: Sermanet ら [239] が提案した OverFeat は，図 14 に示すように，深層 FCN に基づく最初の単一ステージ物体検出器の 1 つと見なされる，これは最も影響力のある物体検出フレームワークの 1 つであり，ILSVRC2013 の位置推定と検出のコンペティションで優勝した。OverFeat はネットワーク内
図14 OverFeat [239]検出フレームワークの図解。

の完全な畳み込み層（図14 (a) の "Feature Extractor") を通る単一のフォワードパスを介して物体検出を実行する。テスト時
の物体検出の主要な工程を以下にまとめる。

1. マルチスケール画像上でスライディングウィンドウ方式の物体分類を行い物体候補を生成。OverFeat は、全
結層のため固定サイズの入力画像を必要とするAlexNet [140]などのCNNを使用する。スライディングウィ
ンドウのアプローチの計算を効率化するために、OverFeat
はそのネットワークを図14 (a) に示すように FCN に
キャストし、全結合層を 1×1 のカーネルを持つ畳み込
み層として見ることで任意のサイズの入力を受け取る。
OverFeat は全体的な性能向上のためマルチスケール特徴を
活用する。具体的には、図14 (b) に示すように元画像
を拡大した最大 8 スケールの画像をネットワークに通
し、評価されるコンテキストの視野の数を大幅に増やす。
マルチスケールの各入力に対し、分類器はスケールと
信頼度のグリッドを出力する。

2. オフセット最大ブーリングによる予測数の増加。分解度増
加のため、OverFeat は最後の CONV 層後にオフセット
最大ブーリング層を適用する。すなわち、全てのオフセットで
サブサンプリングを行い投票用の視野をより多く生成し、
効率を維持しながら頑健性を向上させることになる。

3. bbox 図示、物体が認識されると単一の bbox 回帰器が適
用される。分類器と回帰器は同一の特徴抽出（CONV）層
を共有し、FC 層のみを分類ネットワーク計算後に再計算
する必要がある。

4. 予測の結合。OverFeat は貪欲なマージング戦略を使用し、全て
の位置とスケールにわたる個々の bbox 予測を結合する。
OverFeat は速度の面で大きな利点を持つが、当時 FCN の訓練
が困難だったため RCNN [85] より精度が低い。速度の利点は、
FCN 内の重なり合わせのネットワーク構築により検出の計算を
有効にできる。しかし、マルチスケールは既存のアーキテクチャ
上で実装され、特にマップ構築が必要である。YOLO 227
や SSD [175] などの後続フレー
ムワークに類似している。

YOLO：Redmon ら 227 は統合検出器である YOLO (You
Only Look Once, 一度しか見ること) を提案した。YOLO は物体検出を、図 13
に示すように、空間的に区切られたbbox をそれぞれに細かくクラス
確率を示す画像から得られるに沿って、物体領域候補
生成の段階が完全に削減されているため、YOLO は少数の候
補領域セットを使用して検出を直接予測する 12、局所領域

特徴に基づいて検出を予測する領域ベースのアプローチ（例：
 Faster RCNN）とは異なり、YOLO は画像全体の特徴を大域
的に使用する。YOLO は画像を S × S のグリッドに分割し、
各グリッドから C 個のクラス確率、B 個の bbox 位置、信頼
度スコアを予測する。物体領域候補生成の工程を完全に捨て
ることで YOLO は設計上高速であり、YOLO は 45 FPS、Fast
YOLO [227] は 155 FPS でリアルタイムに実行できる。YOLO
は予測時に画像全体を見るため、物体クラスに関するコンテキ
スト情報を暗黙的にエンコードし、背景で false positive を予測
する可能性が低くなる。YOLO はbbox の位置とスケール・ア
スペクト比の分割が粗いため、Fast RCNN よりも多くの位置
検出エンジンを用いる。227 で議論されているように、YOLO
は一部の物体、特に小さな物体の位置推定に失敗することがあ
る。これはおそらく粗がグリッド分割のためであり、各
グリッドセルには 1 つの物体しか含まれていないためである。保
物検出は多数の物体を含む MS COCO のデータセットで、
YOLO がどの程度の良好性になるかは不明である。

YOLOv2 と YOLO9000: Redmon and Farhadi 226 は
YOLO の改良版である YOLOv2 を提案した。YOLO で使わ
れていた GoogLeNet [263] ベースのカスケードネットワークは
より単純な DarkNet19 に置き換えられ、バッチ正規化 (batch
normalization) [125] が追加され、全結合層が除去され、平均
法 (k-means clustering) とマルチスケール訓練で学習した適
切な anchor box の使用が行われる。YOLOv2 は標準の検出タス
クで最先端の性能を達成した。Redmon and Farhadi [226] は、
9000 以上 (over 9000) の物体カテゴリをリアルタイムで検
出できる YOLO9000 を導入した。そのために、WordTree を
使用した複数サイズからのデータ統合により、ImageNet 分類
データセットと COCO 検出データセットでの同時訓練を行う
共通最適化手法が提案された。YOLO9000 はこのような共同
訓練により、強健性を向上し、bbox アノセーションの無
い物体クラスの検出が可能である。

SSD: 検出精度をより犠牲にすることなくリアルタイムの
高度を維持するために、Liu ら [175] は、YOLO [227] より速
く Faster RCNN [229] などの領域ベースの検出器に匹敵する
精度を持つ SSD (Single Shot Detector) を提案した。SSD は
高品質の検出を維持しつつ高速な検出を達成するため、Faster
RCNN [229] の RPN, YOLO [227], マルチスケール CONV 特
徴 [97] のアイディアを効果的に組み合わせる。SSD は YOLO
同様、一定数の bbox とスコアを予測した後、NMS を行い最
終的な検出を出力する。SSD の CNN ネットワークは全層が
必要であり、早期層（比較的低層の部分）は VGG [248] な
どの標準的なアーキテクチャに基づいており、サイズが徐々
に小さくなるいくつかの補助コンボ層が続け、最終層の情報
は正確な位置推定を行うため空間的に粗を含む可能性がある
ため、適切なサイズの bbox に対するオフセットとカタログい
コアを複数の CONV 特徴マップ上で予測することで、SSD はマ
ルチスケールへの検出を行う。検出精度 300×300 の入力用
いた VOC2007 test set の評価において、Faster RCNN が 7 FPS
で mAP 73.2%，YOLO が 45 FPS で mAP 63.4% になる。SSD は
59 FPS で mAP 74.3% を達成する。

CornerNet: 最近 Law ら [146] は、SoTA 物体検出フレーム
ワーク [84, 102, 227, 175] における anchor box が果たしてきた
支配的な役割に疑問を抱いた。特に 1 段階検出器 [77, 168, 175,
227] で正例・負例数に大きな不均衡を引き起こし、訓練を遅
くし、余分なハイパラマータを導入するなど、anchor box
の使用には欠点がある [146, 168] と主張した。Law ら [146]

11 検出：YOLO (You Only Live Once, 一度生きの人生) のもじり。
12 YOLO の使用する bbox は、Selective Search の約 2000 個と比べるとさらに少
なく、画像あたり 98 個しか使用されていない。
14 物体候補となる様々なサイズ・アスペクト比のbbox。
15 例示：ドローンポールのペース・エリアは「8000 以上だ…」の英語ローカ
ライフ版「It’s Over 9000」を由来する数値表現。
6 物体表現

任意の検出器の主要構成要素の一つとして、優れた特徴表現は物体検出で最も重要である [56, 85, 82, 324]。過去には、局所記述子の設計（例：SIFT [178], HOG [52] や、識別的な部分を抽出させるために高レベルの表現に記述子をグループ化・抽象化するアプローチの探求（例：Bag of Words [252], Fisher Vector [212]）に、新たな努力が賛された。しかし、これからの特徴表現手法には注意深いエンジニアリングとかなりのドメイン専門知識が必要である。

一方従来学習手法（特に深層 CNN）は、複数の抽象化レベルの強力な特徴表現を原画像から直接学習できる [13, 149]。伝統的な特微エンジニアリングで必要とされた特定のドメイン知識と複数の手順への依存が軽減されたため [13, 149], 特微表現に割かれてきた負担は、より良いネットワークアーキテクチャと訓練手順の設計にかわるものとなっている。

5 節でレビューユーした主要フレームワーク (RCNN [85], Fast RCNN [84], Faster RCNN [229], YOLO [227], SSD [175]) は、検出の精度と速度を制約的に向上させてきた。CNN アーキテクチャで 6.1 節, 表 15 がその大きな役割を果たしていると一般的に受けられている。その結果、近年の検出精度向上の多くは新規ネットワーク開発に関する研究による。

そこで本節ではまず、一般物体検出で使用される人間の CNN アーキテクチャをレビューする。その後、物体のスケール・姿勢・変形の幾何変動に対応するための変数特徴の開発や、応用範囲での物体検出を改善するためのマルチスケール分析など、物体特徴表現の改善に費やされた取り組みをレビューする。

6.1 人間の CNN アーキテクチャ

CNN アーキテクチャ (3 節) は 5 節の検出フレームワークで使用される backbone ネットワークとして役立つ。AlexNet [141], ZFNet [303], VGGNet [248], GoogLeNet [263], Inception シリーズ [125, 264, 265], ResNet [101], DenseNet [118], SENet [115] を含め、代表的なフレームワークを表 6 にまとめ、時間経過にもとづく性能改善を図 15 に示す。近年の CNN の進歩に関するさらなるレビューは [92] を参照された。

アーキテクチャの進化では、特に AlexNet が 8 層だったのに対し、VGGNet は 16 層、より最近の ResNet と DenseNet ではともに 100 層を突破している。深さを増加させることで表現力を向上できることは VGGNet [248] と

13 物体検出のためにキーボートを使うというアイディアは、DeNet [269] で既出。
表6 一般物体検出に一般的に使用される DCNN アーキテクチャ、「バラメータ数」、「層数」の統計については、最後の FC 予測層を考慮されていない。「テストエラー」列は、ImageNet1000 の分類の Top 5 テストエラーを示している。このアーキテクチャを指しているか明示的な場合、「バラメータ数」「層数」「テストエラー」は OverFeat (accurate model), VGGNet16, ResNet101, DenseNet201 (Growth Rate 32, DenseNet-BC), ResNetXt50 (32*4d), SE ResNet50 について言及している。（誤訳: アンサンブルの詳細は原著論文を参照されたい。）

| No. | DCNN アーキテクチャ | バラメータ数（×10^8） | 層数 (CONV+FC) | テストエラー (Top 5) | 検出での初使用 | 注目点 |
|-----|---------------------|---------------------|----------------|---------------------|---------------|---------|
| 1   | AlexNet [141]       | 57                  | 5 + 2          | 15.3%               | [85]          |         |
| 2   | ZFNet (fast) [303]  | 58                  | 5 + 2          | 14.8%               | [99]          |         |
| 3   | OverFeat [239]      | 140                 | 6 + 2          | 13.6%               | [239]         |         |
| 4   | VGGNet [248]        | 134                 | 13 + 2         | 6.8%                | [84]          |         |
| 5   | GoogLeNet [263]     | 6                   | 22             | 6.7%                | [263]         |         |
| 6   | Inception v2 [125]  | 12                  | 31             | 4.8%                | [112]         |         |
| 7   | Inception v3 [264]  | 22                  | 47             | 3.6%                | separable convolution と空間解像度を抑え入れた。 |         |
| 8   | YOLONet [227]       | 64                  | 24 + 1         | –                   | GoogLeNetに触発されたネットワーク、YOLO検出器で使用される。 |         |
| 9   | ResNet50 [101]      | 23.4                | 49             | 3.6%                | identity mapping（特有で深いネットワークを学習できる。） | [101] |
| 10  | ResNet101 [101]     | 42                  | 100            | –                   | GoogLeNetで導入されたglobal average poolingをとポトネットを使用することことで、必要なパラメータ数がVGGより少ない。 | [101] |
| 11  | InceptionResNet v1 [265] | 21                | 87             | –                   | identity mapping（インピュータの組み合わせ、Inception v3 様的計算コストが訓練過程を高速。） | [101] |
| 12  | InceptionResNet v2 [265] | 30                | 95             | 3.1%                |認識精度が大幅に向上し、より計算コストの高い residual connection 付き Inception。 | [112] |
| 13  | Inception v4 [265]  | 41                  | 75             | –                   | Inceptionの悪態でresidual connection無し。InceptionResNet v2 ほぼ同等の認識性能だが滑らかに。 | [112] |
| 14  | ResNeXt [291]       | 23                  | 49             | 3.0%                | 同じアーキテクチャを持つ変換の集合を構築する building block（アーキテクチャの Entranceに使用される）ブロック、積み重ね、構成要素を繰り返し使用。 | [291] |
| 15  | DenseNet201 [118]   | 18                  | 200            | –                   | 輔助フィードフォワード方式で各層と他の全ての層を結合。 GPUsの失効問題を軽減し、特徴再利用をし、パラメータ数を削減。 | [321] |
| 16  | DarkNet [226]       | 20                  | 19             | –                   | VGGNetに似ているがバラメータ数が大幅に少ない。 | [226] |
| 17  | MobileNet [112]     | 3.2                 | 27 + 1         | –                   | depth-wise separable convolutionを使用した軽量のDCNN。 | [112] |
| 18  | SE ResNet [115]     | 26                  | 50             | 2.3%                | Adaptive Squeeze-and-Excitation blockという新規ブロックによるchannel-wise attention。既存の backbone CNNと相補的。 | [115] |

持つ（通常多重の視覚的カテゴリを含む大規模データセットでの CNN の事前学習が一般的に行われる。事前学習済み CNN は汎用的な特徴抽出器 [223, 8, 60, 296] として大規模データセットに直接適用でき、幅広い視覚認識タスクの下支えとなる。検出では一般的に、事前学習済みネットワークは所定の検出データセットで fine-tuning される [60, 85, 87]。いくつかの大規模画像分類データセット（例：1000 の物体カテゴリの 120 万枚以上の画像からなる ImageNet1000 [54, 234]、ImageNet1000 から数個がクラス数は Place [319]、Places-ImageNetの混合 [319]、JFT300M [106, 254]）は CNN の事前学習のために使用される。

fine-tuning 無しの事前学習済み CNN については、[60, 87, 1] で物体の分類、検出のためを探求され、一部の層から抽出した特徴で検出精度が異なることが示される。例えば、ImageNetで事前学習された AlexNet の場合、検出精度は FC6 / FC7 / Pool5 の順に低下する [60, 87]。事前学習済みネットワークの fine-tuning は検出性能を大幅に向上させることができる [85, 87]。AlexNetの場合、fine-tuning による性能向上は Pool5 / FC6 / FC7 ではかかる大きいことが示されており、これは Pool5 の方が汎用的な（ドメインに特化していない）特徴であることを示唆している。また、ソースデータセットとターゲットデータセットの関係が重要な役割を果たし、例えば ImageNet ベースの CNN 特徴は周辺移動よりも物体検出で良い性能を示す [317, 8]。

6.2 物体表現改善手法
RCNN [85], Fast RCNN [84], Faster RCNN [229], YOLO [227] などの deep CNN ベース検出器は、典型的には表6に記載した deep CNN アーキテクチャを backbone ネットワークとして使用し、CNNの最上層の特徴を物体表現として使用する。しかし、広範囲のスケールにわたりて物体を検出すことは重要な課題である。この問題に対処するための古典的な戦略は、多数のスケーリングされた入力画像上で検出器を実行することができる（例：画像ピラミッド） [74, 85, 99]。通常この戦略はより正確な検出を出すが、推論時間とメモリの面では明らかに不利である。

6.2.1 物体スケール変異への対処
CNN は層ごとにその特徴階層を算出するため、特徴階層の中のサブサンプリング層が既に入力するマトリックスの変化にてより、様々な空間解像度で特徴マップを生成するが、課題が生じやすい [97, 177, 247]。具体的には、高い分離性と強力な意味図式を持ち、物体姿勢、照明、パターン変形などの変動に最も頑健だが、解像度が高く幾何的な詳細が失われている。逆に低層は小さな受容野と豊富な幾何的詳細を

17 ImageNet のようなラベル付き大規模データセット用に最適化された重みでネットワークを初期化した後、ターゲットタスクの訓練課題を用いてネットワークの重みを更新することで fine-tuning は行われる。
| 検出器名 | 物体領域 | backbone | DCNN | パイプライン | mAP@IoU=0.5 | mAP@IoU=0.75 | 発表者 | 注目点 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ION [11] | SS+TB | MCG+RPN | VGG16 | Fast RN | 79.4 | 74.6 | 55.7 | CVPR16 |
| HyperNet [135] | RPN | VGG16 | Faster RN | 76.3 | 71.4 | CVPR16 |
| PVANet [132] | RPN | PVANet | Faster RN | 84.9 | 84.2 | NIPS16 |

| グループ (2)：複数層での検出 |
| --- |
| SDP+CRC [293] | EB | VGG16 | Fast RCN | 69.4 | 64.9 | 9.4 | CVPR16 |
| MSCNN [24] | RPN | VGG | Faster RCN | KITTI | Caltech のみテスト | ECCV16 |
| MPN [302] | SharpMask | VGG16 | Fast RCN | – | 51.9 | BMVC16 |
| DSOD [242] | Free | DenseNet | SSD | 77.7 | 72.2 | ICVC17 |
| REFNet [173] | Free | VGG16 | SSD | 82.2 | 81.2 | ICVC17 |

| グループ (3)：上記 (1), (2) の組み合わせ |
| --- |
| DSOD [247] | Free | ResNet101 | SSD | 81.5 | 80.0 | X17 | |
| TDM [247] | RPN | ResNet101 | SSD | 81.5 | 80.0 | X17 | |
| ZIF [156] | RPN | Inceptionv2 | Faster RN | 79.8 | 74.9 | X17 | |
| STDN [321] | Free | DenseNet169 | SSD | 80.9 | 76.8 | X17 | |
| RefineDet [308] | RPN | ResNet161 | SSD | 83.8 | 81.2 | X17 | |
| PANet [174] | RPN | ResNetx101 | Mask | RCNN | 81.3 | 77.9 | X17 | |
| DetNet [164] | RPN | ResNet59 | Faster RN | 79.8 | 74.9 | X17 | |
| FPR [137] | – | ResNet101 | SSD | 82.4 | 81.1 | X17 | |
| M2Det [315] | – | ResNet101 | SSD | 82.4 | 81.1 | X17 | |

| グループ (4)：幾何学的変換のモデル化 |
| --- |
| DeepIDNet [203] | SS+EB | AlexNet | RCNN | 79.0 | 78.0 | 25.6 | CVPR15 |
| DCN [51] | RPN | ResNet101 | IRN | 82.6 | 82.0 | X17 | |
| DPCF [188] | AttractionNet | ResNet | IRN | 83.3 | 82.2 | X17 | |
持つが、解像度が高く意味に対する感受度はなにかほどに対し、直観的に、物体サイズに応じて物体の意味的概念は様々な層に現れる可能性がある。対象物体が小さい場合、より早期の層で微細な詳細情報が必要な上、後層ではほぼ消え得るため、原理上小さな物体の検出は非常にチャレンジングである。そのため、他の特徴の解像度を向上させる dilated convolution ("atrous" convolution) [298, 50, 33] などの技法が提案されているが、それらは計算の複雑さを増大させる、一方、対象物体が大きい場合、意味的概念は後の層に現れる。CNN の複数層を利用して検出精度を向上させるために、多数的手法 [247, 314, 167, 136] が提案されており、マルチスケール物体検出は大まかに以下の 3 種に分類される。

1. 複数層の特徴の組み合わせによる検出
2. 複数層での検出
3. 上記 2 手法の組み合わせ

1. CNN の複数層の特徴の組み合わせによる検出: Hypercolumns [97], HyperNet [135], ION [11] など多くのアプローチが、予測を行う前に複数層からの特徴を組み合わせる。このような特徴の組み合わせは、一般的に concatenation (連結、結合) により実現される。異なる層からの特徴を組み合わせるためには統合的ゆールネットワークのアイデアであり、また、近年ではセマンティックセグメンテーションのアイデアで呼ばれる。近年では、セマンティックセグメントにおける解像度改善を念頭において、HyperNet [135] は同様のアイデアに従っており、深い特徴と中間特徴と浅い特徴を統合、end-to-end の共同訓練戦略で物体領域候補の生成と物体の予測を行う。連続特徴はより叙事的で位置推定と分類による有益であるが、計算の複雑さが増す。

2. コンボスティックの検出: 近年多くのアプローチが、様々な層で様々な解像度の物体を予測し、それぞれの予測を組み合わせることで検出を改善している。RLFBNet [173], MSDCNN [24], RFBNet [173], DSOD [242], SSD [175] は CNN 内の複数層に異なるスケールの default box を消し撒き、各層を特定スケールの物体の予測に注力させる。RFBNet [173] は SSD の方向方の積み込み層を Receptive Field Block (RFB) で置き換え、特徴の識別性や顕著性を強化する。RFB は Inception ブロック [263] と同様複数ブロックの積み込みブロックだが、複数ブロックをサイジングの異なる複数のカーネルと dilated convolution 層 [133] で組み合わせる。MSSCN [24] は CNN の複数層を用いて物体領域候補生成を学習し、また、解像度向上のため

に RoI pooling 前の特徴マップに deconvolution を適用する。TridentNet [163] は、RFBNet [173] 同様に受容野の異なる複数ブロックが並列にアーキテクチャを構築するが、各ブロックが同じ変換パラメータ (畳み込み層の重み) を共有する。様々なスケールの物体に受容野を適応させるために、dilation rate の異なる複数の dilated convolution が使用される。

3. 上記 2 手法の組み合わせ: Hypercolumns [97], HyperNet [135], ION [11] で示されるように、異なる層からの特徴は相補的であり検出精度を向上させることができる。しかし一方で、様々なスケールの物体を検出すのにほんの少しサイズの特徴を使用するのは自然である。これは、縮小した特徴マップから大きな物体を検出し、拡大された特徴マップから小さな物体を検出することで実現する。そこで、両者の長短を組み合わせるために、異なる層からの特徴を組み合わせて得られる特徴を使用し、かつ複数層で物体を検出することを、近年のいくつかの研究が提案している。このアプローチはセグメントーション [177, 241] と人物姿勢推定 [194] で有効性が見出され、物体インスタンス間のスケール変動の問題を軽減するため、1段階検出器と 2 段階検出器の両方に広く利用されている。

代表的な手法として、SharpMask [214], Deconvolutional Single Shot Detector (DSSD) [77], Feature Pyramid Network (FPN) [167], Top Down Modulation (TDM) [247], Reverse connection with Objectness prior Network (RON) [136], ZIP [156], Scale Transfer Detection Network (STDN) [321], RefineDet [308], StairNet [283], Path Aggregation Network (PANet) [174], Feature Pyramid Reconfiguration (FPR) [137], DetNet [164], Scale Aware Network (SAN) [133], Multiscale Location aware Kernel Representation (MLKLP) [278], M2Det [315] が挙げられる。手法の概要を表 7 に、対比図を図 17 に示す。

FPN [167], DSSD [77], TDM [247], ZIP [156], RON [136], RefineDet [308] などの早期の研究は、backbone に内在するマルチスケールのピラミッドアーキテクチャに従って特徴バイアスを構築し、有望な結果を達成した。これらの手法は図 17 (a1)～(a2) から見て取れるように、トップダウンネットワークと lateral connection の組み合わせによる特徴の強化を図っている。dilated convolution と lateral connection の組み合わせによる特徴の強化を図っている。
図 17 砂時計型アーキテクチャ。以下に示す近年のアプローチで一般的に使用される、多種の Feature Fusion Block (FFB) を比較している。FPN [167], TDM [247], DSSD [77], RON [136], RefineDet [308], ZIP [156], PA Net [174], FPR [137], DetNet [164], M2Det [315]。FFM：Feature Fusion Module。TUM：Thinned U-shaped Module。Conv1 ~ Conv5：VGG や ResNet などのブロックバックネットワークの主たる Conv ブロック。
に、各物体領域候補のために全ての特徴レベルから特徴を集約する adaptive feature pooling を行う。そのうえ、マスク予測サブネットワークに全結合層を含む補完ブロックを追加することで、各物体領域候補を異なる視点から捉え、マスク予測を更に改善する。これらの追加は計算オーバーヘッドを微増させるが効果的であり、PANet は COCO 2017 Challenge のインスタンスセグメンテーションタスクで 1 位、物体検出タスクで 2 位を獲得した。Kong らは、非線形性が高いが効率的な方法で特徴ピクセルマップ構築処理（例：FPN [167]）を特徴再構成閾数として明示的に再定式化する、FPR [137] を提案した。FPN のように最高上段層からの強力な意味的特徴をトップダウンパスで下に伝播する代わりに、FPR は図 17 (h1) に示すように、まず backbone ネットワークの複数層からの特徴を適応的に連絡して抽出し、次により複雑な設計の FFB module（図 17 (h2)) で強力なセマンティクスを全スケールに広げる。Li らは、深い層で高い空間解像度を維持するために backbone ネットワークの後方の層に dilated convolution を導入した DetNet [164]（図 17 (ii)) を提案した。Zhao ら [315] は、異なる特徴の物体の検出により効果的な特徴ピクセルマップを構築するために、MultiLevel Feature Pyramid Network（MLFPN）を提案した。図 17 (j1) から分かるように、まず backbone の異なる 2 つの層の特徴がベース特徴として融合された後、ベース特徴から lateral connection を用いてトップダウンパスに特徴ピクセルマップ構築のため作成される。図 17 (j2), (j5) に示すように、FPN などと比べてさらに複雑な FFB module を使用し、FFB 内の Thinned U-shaped Module（TUM）で 2 番目のピクセルマップを生成、その後、複数の TUM の同等サイズの特徴マップを物体検出のために組み合わせる。MLFPN を SSD に統合して M2Det が提案され、他の 1 段階検出器より優れた検出性能を達成した。

6.3 他のクラス内変動の対処

強力な物体表現は弁別性と頑健性を併せ持つ必要がある。6.2.1 節でレビューした通り、近年多くの研究が物体のスケール変化への対処に専念してきた。2.2 節で議論した図 5 まで進んでいるように、物体検出はスケールの変動だけでなくそれ以外の実世界の変動に対しても頑健な必要がある。これらの変動を以下の 3 カテゴリーに分類する。

- 幾何学的変換
- 遮蔽
- 画像の劣化

これらのクラス内変動対処のために最も単純なアプローチは、十分な量の変動を加えて訓練データセットを増強することである。例えば回転に対する頑健性は、数多くの向きに回転した物体を訓練データに追加するかのようにできるだけ、頑健性はしばしばこのような方法で学習できるが、これは通常として物体検出のコストがかかりエネルギーパラメータが複雑になる。そのため、研究者はこれからの問題を解決する代替案を提案してきた。

幾何学的変換の対処：DCNN の本質的な制約として、人工データの幾何学的変換に対して空間的に不变となる能力が不足している [152, 172, 28]。局所最大プールリングの層は DCNN に多少の平移変形不変性を与えが、実際には中間特徴マップは人工データの大きな幾何学的変換に対して不変ではない [152]。そのため、頑健性向上のために多くのアプローチが提案され、スケール [131, 21], 回転 [21, 42, 284, 323], またはその両方 [126] など、様々な種類の変換に対して不変である CNN 表現の学習が目指されてきた。代表的な研究の一つが、大域のパラメータ変換によりスクリーンリーニング・クロッピング・回転、非剛体変形に対処する、学習可能な新規モジュールを導入した Spatial Transformer Network（STN）[126] である。STN は現在、回転したテキストの検出 [126]、回転した顔の検出および一般物体検出 [280] で使用されている。

回転不変性はシーンテキスト検出 [103, 184]、顔検出 [243]、航空画像 [57, 288] などの特定のアプリケーションでは魅力的かもしれない、しかしこれ一般物体検出では、人間のペンマークの検出データセット（例：PASCAL VOC, ImageNet, COCO）で回転誤差が実際には提示されないため、回転不変性に焦点を当たった研究は限定されている。

深層学習より前に、Deformable Part based Model（DPM）[74] は、变形可能な構築で配置された構成パーツによって物体を表現して一般物体検出で成果を収めた、DPM は性能で近年の物体検出器に大きな注目を浴びたが、依然その精神は近年の多くの検出器で特に影響を与えている。DPM のモデル化は物体形状・視点・非剛体形状の変換の影響を受けにくい、この性質が、CNN ベースの検出を改善するために研究者 [51, 86, 188, 203, 277] が明示的に物体構成をモデル化する動機となった。最初の試み [86, 277] では、AlexNet で学習した深層特徴を DPM ベースの検出が使用することで DPM と CNN を組み合わせたが、物体領域候補生成は無かった。物体パーツの変形をモデル化する。その内蔵された能力の恩恵を CNN に提供される。DeepIDNet [203], DCN [51], DPFCCN [188]（表 7）を含め多くのアプローチが提案された。神経は似ているが変形は様々な方法で計算される、DeepIDNet [206] は、様々な物体クラスにおける共有視覚パターンとその変形特性を学習するため、通常の最大アブリック・アイデアを用いる変化構造付けアブリック（deformation constrained pooling）層を設計し、DCN [51] は、規則的な格子状のサミング位置オブジェクトを追加することで特徴マップのサミング位置を増大させるという考え方に基づいて、deformable convolution 層と deformable RoI pooling 層を設計した。

遮蔽の対処：実世界の画像では遮蔽は一般的に起こり、物体インスタンスからの情報が失われる、デフォーム・アブリック・アイデアは遮蔽処理に有用であるとされる、deformable RoI Pooling [51, 188, 202] や deformable convolution [51] が提案されており、通常は固定されている各物体構造の柔軟性を上げることで遮蔽の影響を軽減する。Wang ら [280] は、遮蔽と変形に含む例を生成する敵対的ネットワークの学習を提案している。また、コンテキストが遮蔽の対処に役立つ可能性がある [309]。これらの努力にもかかわらず遮蔽の問題は解決にはほど遠い、この問題への GAN の適用は研究の方向性として有望である。

画像の劣化の対処：画像ノイズは実世界の多くのアプリケーションで一般的な問題である、多くの場合、十分な照明、低品質のカメラ、画像圧縮、エッジデバイスとウェアラブルデバイスの意図で低コストなセンサーが原因である。低画質は視覚認識の性能を低下させるとは予想されるが、PASCAL VOC, ImageNet, MS COCO, Open Images の全てに比較的高品質の画像に焦点を当てている事実から明らかのように、現在のほとんどの手法は劣化の無視できる的な環境で評価される、我々の知る限り、この問題に対処する研究は今このところ非常に限られている。

7 コンテキストのモデル化

実世界では、視覚物は特定の環境に存在し、通常は他の関連物体と共に存在する。コンテキストが人間の物体認識において不可欠な役割を果たすことは強く心理的根拠がある [14, 10]。また、小さな物体サイズ、物体の遮蔽、画質の悪さが原因で物体の外観の特徴が不十分な場合は特に、コンテキ
| 検出器名 | 物体領域候補生成 | backbone | バイナリライン | mAP@IoU=0.5 VOC07 | mAP@IoU=0.5 VOC12 | 発表先 | 注目点 |
|----------|----------------|----------|---------------|---------------------|---------------------|--------|--------|
| 大域コンテキスト | | | | | | | |
| SegDeepM [326] | SS+CMPC | VGG16 | RCNN | – | – | – | CVPR15 |
| DeepIDNet [203] | SS+EB | AlexNet | RCNN | 69.0 | (07) | – | CVPR15 |
| ION [11] | SS+EB | VGG16 | Fast | 80.1 | 77.9 | 33.1 | CVPR16 |
| CPF [245] | RPN | VGG16 | Faster | 76.4 | 72.6 | – | ECCV16 |
| 局所コンテキスト | | | | | | | |
| MRCNN [82] | SS | VGG16 | SPPNet | 78.2 | (07+12) | 73.9 | (07+12) | – | ICCV16 |
| GBDNet [304, 305] | CRAFT [292] | Inception v2 | ResNet101 | PolyNet [311] | Fast | 77.2 | (07+12) | – | 27.0 | ECCV16 |
| ACCNN [157] | SS | VGG16 | Fast | 72.0 | (07+12) | 70.6 | (07+12) | – | TMM17 |
| CoupleNet [327] | RPN | ResNet101 | RFCN | 82.7 | 80.4 | (07+12) | 34.4 | ICCV17 |
| SMN [35] | RPN | VGG16 | Faster | 70.0 | (07) | – | – | ICCV17 |
| ORN [114] | RPN | ResNet101 | Faster | – | – | 39.0 | – | CVPR18 |
| SIN [176] | RPN | VGG16 | Faster | 76.0 | 73.1 | (07+12) | 23.2 | – | CVPR18 |

図18 各局所コンテキスト特徴について探究する代表的アプローチ：MRCNN [82], GBDNet [304, 305], ACCNN [157], CoupleNet [327].表8も参照されたい。
ストの適切なモデル化が物体の検出と認識に役立つと考えられている [266, 197, 33, 32, 58, 78]。多くの異なる種類のコンテキストが議論されてきた [58, 78] が、それらは大まかに以下の 3 カテゴリのいずれかに分類できる。

1. 意味的コンテキスト：ある物体が、一部のシーンでは見られ、また別のシーンでは見られない場合、

2. 空間的コンテキスト：物体が移動する場所での、物体が見られ、また別の場所での物体が見られない場合、

3. スケールコンテキスト：物体の取りうるサイズの範囲を、シーン内の他の物体との相対的なサイズによって定義されている。

深層学習の普及以前にかなり多くの研究 [34, 58, 78, 185, 193, 220, 207] が行われた。これらの研究の多くはまた DCNN ベースの物体検出器で提案されてきている [35, 114]。

物体検出における現在の最先端技術 [229, 199, 102] は、何のコンテキスト情報も明示的に利用せずに物体を検出する、複数のレベルで抽象化された階層的表現を学習するため、DCNN は暗黙的にコンテキスト情報を使用していると広く示されている [303, 316]。それでもかわらぬ、DCNN ベースの検出器で明示的にコンテキスト情報を誘導する提案がある [114, 35, 305]。そこで以下では、DCNN ベースの物体検出器でコンテキストの手がかりを活用する近年の研究をレビューする。早期の研究 [310, 78] から発展して、大域コンテキストと局所コンテキストのカテゴリーに分けて、代表的なアプローチを表 8 にまとめる。

7.1 大域コンテキスト

大域コンテキスト [310, 78] は、物体検出の手がかりとして役立てる画像レベルまたはシーンレベルのコンテキストを指す（例：寝室（ベッドルーム）にはベッドが存在することが予測される）。DeepIDNet [203] では、画像分類スコアをコンテキスト特徴として使用し、検出結果改善のため画像分類スコアと物体検出スコアを連結する。ION [11] で Bell らは、画像全体のコンテキスト情報を探索するため空間的リカレントニューラルネットワーク（Recurrent Neural Network, RNN）の使用を提案した。SegDeepM [326] で Zhu らは、各検出器を対象としてコンテキストをスコア付けるマルチクロザ率モデルを提案した。この提案モデルは、候補矩形が多数の物体セグメーション候補からセグメーション候補からセグメーション候補とセグメーション間の一致をスコア付けでできるようにしている。[245] では、セマンティックセグメーションがコンテキストプライミングの一つ階層として使用された。

7.2 局所コンテキスト

局所コンテキスト [310, 78, 220] は、局所的に近くにある物体間の関係性、および物体とその周囲の領域間の相互作用を考える。物体に属する領域のモデル化はチャレンジであり、クラスタ・位置・スケールなどの異なるbboxについて reasoning (推論、理由付け) を必要とする。物体の関係を明示的にモデル化する深層学習の研究は非常に限られており、代表的なものは Spatial Memory Network（SMN）[35]、Object Relation Network [114]、Structure Inference Network（SIN）[176] である。SMN では、空間の情報が全体の情報の一部である。クリスタルインスタンスを組み込みで疑似画像表現に展開するもの。疑似画像表現は物体の関係の推論のため別の CNN に簡単に入力できる。これにより画像とメソッドを並行処理し、メソッドを更新し、新たな入力検出を取得することで順次推論を行っていく新規アーキテクチャが生まれた。ORN [18] は、近頃の自然言語処理でのattentionモジュールの成功 [274] に触発され、外観特徴と幾何特徵の相乗作用を通じて物体のセットを同時に処理する。ORN は追加の教師情報が必要とせず、既存ネットワークを簡単に組み込み、近年の物体検出バイアスラインでの物体認識、重複除去工程の改善に有効で、最初の完全な end-to-end の物体検出器を生み出す。SIN [176] は、シーンのコンテキスト情報と単一画像内の物体の関係性という 2 種類のコンテキストを検討した。物体検出はグラフ推論の問題として定式化され、物体はグラフのノードとして扱われ、物体間の関係性はエッジとしてモデル化される。

検出ウィンドウのサイズを拡大して何らかの形で局所コンテキストを抽出するという、より単純なアイディアに基づいてコンテキストの課題に取り組む手法が、より広範に研究されてきた。代表的なアプローチとして、MRCNN [82]、Gated BiDirectional CNN（GBDNet）[304, 305]、Attention to Context CNN（ACNN）[157]、CoupleNet [327]、SemanNet [238] の研究がある。MRCNN [82] (図 18(a) で) Girardis と Komodakis らは、より豊かで頑健な物体表現を獲得するために、元々の物体領域候補から抽出された backbone の最終 CONV 層の特徴を加えて、物体領域候補内外の多数の異なる領域（半分の領域、境界領域、中央領域、コンテキスト領域、セマンティックセグメンテーションされた領域）から特徴抽出することを提案した。これらの特徴は全て連結によって結合される。

それ以来、MRCNN と密切に関連する手法が多い数提案されてきた。[302] の手法は foveal structure で編集された 4 つのコンテキスト領域（中心窓を移動して 1, 1.5, 2, 4 乗のサイズでクロップした領域）のみを使用し、複数のパスで連続的に SBDNet [304, 305] (図 18(b) を提案した。各領域に対する CNN 特徴を個別に学習してからそれらを連結するやや単純なアプローチは対照的に、GBDNet は異なるコンテキスト領域の特徴をメッセージを受け渡すメッセージの受け渡し操作を常に役立つとは限らず、個々のサンプルに依存しているため、Zeng ら [304] はメッセージの伝達を制御するゲート関数を使用すると注目され、Li ら [157] は大域コンテキストと局所コンテキストの両方の情報を利用する ACCNN (図 18(c) を提案した。大域コンテキストを捉えるため、入力画像に対するアテンショニングマップを縦横に生成して有望なコンテキストの位置を強調する Multiscale Local Contextualized（MLC）抽象ネットワークが使用され、局所コンテキストには MRCNN [82] と同様の手法が採用された。図 18(d) に示すように、CoupelNet [327] は ACCNN [157] から概念的に類似しているが、位置 sensitivRoI pooling データベースで物体の情報を捉える RFCNN [50] をベースとしており、RoI pooling で大域コンテキストをエンドコードするブランチが追加されている。

8 物体領域候補生成手法

物体は画像内の任意の位置に任意のスケールで存在する。ハンドクラフト特徴記述子（SIFT [179]、HOG [52]、LBP [196] の全盛期では、最も成功した物体検出手法（例：DPM [72]）はスライディングウィンドウ法を使用した [276, 52, 72, 98, 275]。しかし、ウィンドウの数は膨大で画像の画素数をともに増加する。また、複数のスケールとアスペクト比を探索する必要があります" が、探索空間の増加に伴って計算コストが高まるため、サポートを有するソフトウェアを開発する必要がある。このように、ウィンドウの数は膨大で、物体検出のための計算負荷が増加する。
2011年の頃、物体領域候補検出提案20を使用して、計算容易性と高い検出品質を両立させることができ提案された [273, 271]。物体領域候補は、[2]で提案されたobjectness [21]のアイディアに由来する。物体を含む可能性が高い画像内の検査領域のセットである。あまり多くない数（例：100）の物体領域候補で高い物体リコールを達成できる場合、ステイディングウィンドウアプローチと比べ大幅な高速化が得られ、より高度な分類器が使用可能になる。物体領域候補生成は前処理工程として通常使用され、検出器で評価する必要がある領域数を制限する。物体領域候補生成は以下の特性を持つ必要がある。

1. ごく少数の物体領域候補で高いリコールを達成できる。
2. 物体領域候補が物体のbboxに一致するよう、位置推定ができるだけ正確である。
3. 計算コストが低い。

物体領域候補生成に基づく物体検出の成功 [273, 271] は幅広い関心を集めている [25, 7, 3, 43, 330, 65, 138, 186]。物体領域候補生成は物体検出の域を超えたアプリケーションを持つ [6, 93, 328] ため、物体領域候補生成アルゴリズムの評価のレビューユは本論文の範囲を超えてる。興味のある読者は、多数の古典的な物体領域候補生成アルゴリズムとそれらが検出性能に与える影響について詳細に解説した近年のセミナー [110, 27] を参照されたい。このでの我々の関心は、DCNNベースであり、クラス非依存の物体領域候補を出し、かつ一般物体検出に応用する物体領域候補生成手法のレビューユである。

2014年、物体領域候補生成 [273, 271] とDCNN特徴 [140]が統合され、一般物体検出におけるマイルストーンであるRCNN [85]につながった。それがすぐに物体領域候補生成は標準的な前処理工程になった。この中には、PASCAL VOC [68]、ILSVRC [234]、MS COCO [166]物体検出チャレンジ2014年以降の優勝エントリーの全てが物体領域候補生成を改善した [85, 203, 84, 295, 305, 102] ことから裏付けられる。

伝統的な低レベルの手法が、例：色、テクスチャ、エッジ・勾配に基づく物体領域候補生成アプローチの中では、Selective Search [271]、MCG [7]、Edge Boxes [330]の人気が高。しかし、本分野が急速に進歩するにつれて、検出精度を向上させ、楕円から形状を独立した外部モデルとして採用された伝統的な物体領域候補生成アプローチ [271, 110, 330] が、検出バイアスを重視のポトルネックとなった [229]。新しく登場したDCNNを使用する物体領域候補生成アルゴリズム [67, 299, 142, 81, 213, 292]が幅広い注目を集めている。

近年のDCNNベースの物体領域候補生成手法は、一般的にbboxベースと物体セグメントベースの2カテゴリーに分類される。代表的な手法を表9にまとめること。

bbox候補生成手法は、図19に示すRenらの提案 [292]が最も良い例である。RPNは、最終共有CONV層の特徴マップ上で小さなネットワークをスライドさせることで物体領域候補を予測する。ストライデングウィンドウの各位置で、anchor boxを使用してk個の物体領域候補を予測する。ここで、各anchor box22は画像内部の特定の位置に基づき、特徴のスケールとアスペクト比に関連付けられる。Renら292は、多様な構成によりRPNとFast RCNNを一つのネットワークに統合することを提案した。これにより、最初のend-to-end検出パイプラインである Faster RCNNが与えられた。表7から分かるように、物体領域候補生成手法としてRPNは多くの最急端の物体検出器として採用されている。

MultiBox [67, 262]やRPN [229]のように先端的な特徴を棄去するanchorのセットを用いる代わりに、Luら [181]は、物体を含む可能性の高い部分領域に焦点を当てるように適切な計算リソースをガイドすることができる、再帰的探索戦略を使用したanchor位置生成を提案した。画像全体から始め、探索処理中で遭遇したanchor領域については、領域をさらに分割するかどうかを決定するためのzoom indicatorを用いる。そして、プランチを追加してRPNを拡張し既存プランチャと並行してzoom indicatorを計算するAdjacency and Zoom Network（AZNet）によって、objectnessスコアを持つbboxのセットが計算される。

さらに、複数層の畳み込み特徴の利用によって物体領域候補生成が試みられている。RPN [229]と同時期に、Ghodrati ら [81]は複数層の畳み込み特徴のマスクを使用して物体領域候補を生成するDeepProposalを提案した。DeepProposalは、最も有力な物体位置を推定しその結果をcoarse-to-fineな方法で精緻化する。最終畳み込む層から逆にしたもの追加カーブを構成する。RPNを改良したHCN [135]は、複数層の畳み込み特徴を強調のHyper Featuresを設計し、end-to-endの共同訓練戦略を介してこれらを物体領域候補生成と物体検出の両者で共有する。Yangらは同じカーブスケート戦略を使用するCRAFT [292]を提案した。CRAFTはまずRPNネットワークを訓練して物体領域候補を生成し、次にそれを使用して、物体と背景を更に区別するための2クラスマップを生成する一つのFast RCNNネットワークを訓練する。Liら [156]はRPNを改良したZIPを提案した。ZIPは、低レベルの詳細と高レベルのセマンティクスの両者の統合を目的とするため、深さの異なる複数の畳み込み特徴マップを使用して物体領域候補を予測する。ZIPでは、conv-deconv構造 [177]に触発された“zoom out and in”ネットワークがbackboneに使用される。

最後に、特筆に値する近年の研究を挙げると、DeepBox [142]は、Edge Boxesが生成した物体領域候補の再構成付けを容易にするために大幅にCNNを提案した。DeNet [269]は、効率的に物体領域候補を予測するbboxコーナーペド文に提案した。Fast RCNN形の検出器のRPNを兼ねる。

物体セグメント候補生成手法 [213, 214]は、物体に対する可能性が高いセグメント候補の生成を目標とする。セグメント候補はbbox候補よりも情報数が多く、物体セグメント化ステージ [96, 49, 162]へと一歩近付く。また、インスタンスセグメント化の教師情報を使用することでbbox検出の性能を向上できる。その先駆けであるPinheiroら [213]によって提案されたDeepMaskは、未加工の画像データから深層ネットワークで直接学習された物体領域候補をセグメント化する。DeepMaskは、RPNと同様多数の共有畳み込み層の後でネットワークを2つのプランチに分割し、クラス非依存


### 表9 DCNNを使用する物体領域候補生成手法の概要

| 物体領域候補生成手法 | backbone DCNN | テスト REC@IoU(VOC07) | 検出結果(mAP) | 発表元 | 注目点 |
|---------------------|---------------|----------------------|----------------|--------|--------|
| bbox 候補生成手法    |               |                      |                |        |        |
| MultiBox [67]        | AlexNet       | RCNN                 | 29.0           | CVPR14  |
| DeepBox [142]        | VGG16         | Fast RCNN (1000)     | 37.8           | ICCV15  |
| RPN [229, 230]       | VGG16         | Faster RCNN (300)    | 73.2           | NIPS15  |
| DeepProposal [81]    | VGG16         | Fast RCNN (100)      | 53.2           | ICCV15  |
| CRAFT [292]          | VGG16         | Faster RCNN (300)    | 75.7           | CVPR16  |
| AZNet [181]          | VGG16         | Faster RCNN (300)    | 70.4           | CVPR16  |
| ZIP [156]            | Inception v2 | Faster RCNN (300)    | 79.8           | IJCV18  |
| DeNet [269]          | ResNet101     | Faster RCNN (300)    | 77.1           | ICCV17  |

### セグメント候補生成手法

| ステップ1 RCNN | ステップ2 RCNN | ステップ3 RCNN | 発表元 |
|---------------|---------------|---------------|--------|
| DeepMask [213] | VGG16         | Fast RCNN     | NIPS15 |
| InstanceFCN [48] | VGG16         | –             | ECCV16 |
| SharpMask [214] | MPN [302]     | Fast RCNN     | ECCV16 |
| FastMask [113] | ResNet59      | –             | CVPR17 |

### DCNNのマスクとそれに紐付くobjectnessスコアを予測する

マスクとそれに紐付くobjectnessスコアを予測する。また，OverFeat [239]の効率的なスライディングウィンドウ戦略と同様，訓練済みのDeepMaskネットワークは推論時，画像（およそスクール変更版）をスライディングウィンドウ方式で適用される。より最近ではPinheiroら [214]が，DeepMask アーキテクチャに精緻化モジュール（reﬁnement module）を追加することでSharpMaskを提案した。SharpMaskのアーキテクチャは図17 (b) , (b) 同様であり，フィードフォワードネットワークにトップダウンの精緻化処理を追加している。

SharpMaskは，早期（低層）の特徴からの空間的に豊かな情報と，後方の層でエンコーディングされた強力な意味情報を効率的に統合でき，成績の高い物体マスクを生成する。

セマンティックセグメンテーション用のFully Convolutional Network (FCN) [177]とDeepMask [213]に動機づけられ，Daiらはインスタンスセグメント候補を生成するInstanceFCN [48]を提案した。

DeepMask同様 InstanceFCNのネットワークは2つの全層畳込みブランチに分割され，1つはinstance-sensitive score mapを生成し，もう1つはobjectnessスコアを予測する。

Huらは，マルチスケールの畳込み特徴を使用するため，SSD [175]と同様のone-shotの方法で効率的にインスタンスセグメント候補を生成するFastMask [113]を提案した。マルチスケールの畳込み特徴マップから密に抽出されたスライディングウィンドウが，セグメンテーションマスクとobjectnessスコアを予測するために，scale-tolerant attentional head moduleに入力された。

FastMaskは，解像度800 × 600の画像に対し13FPSで実行できると主張されている。

### 9 その他の問題

データ拡張として，DCNNを学習するためにデータ拡張（data augmentation）を行う [25, 84, 85] ことば，視覚認識によって重要であると一般に認識されている，平たんなデータ拡張は，クロッピング，反転，回転，スケーリング，平行移動，色摂動，ノイズ付加など，根本にあるカテゴリを変更しない変換によっ

23

23]]マスコンの変化を伴わずに，この段落は [63] に依頼する記述が多いため詳細はそちらを参照されたい。
表 10 訓練戦略とクラス不均衡対処の代表的手法、COCO の結果は Test Dev で報告される。COCO の検出結果は mAP@IoU[0.5, 0.95]に基づく。

| 検出器名     | 物体領域 | backbone   | DCNN | バイライン | VOC07結果 | VOC12結果 | COCO結果 | 発表先 | 注点 |
|-------------|----------|------------|------|------------|-----------|-----------|----------|-------|------|
| MegDet [209]| RPN      | ResNet50 +FPN | Faster RCNN | – | – | 52.5 | CVPR18 | Cross-GPU Batch Normalization を導入することで、以前よりも大きなミニバッチサイズでの訓練を可能にする。128 GPU で COCO の訓練を 4 時間で終了でき、精度向上を遂げる。COCO2017 検出チャレンジで優勝。 |
| SNIP [249]  | RPN      | DPN [37] +DCN [51] | RFCN | – | – | 48.3 | CVPR18 | 新しいマルチスケールの訓練スキーム、小さな物体の検出のためのアクサセンプリングの効果を実証的に調査。訓練時、特徴のスケールに合った物体のみを正しく選択。 |
| SNIPER [251]| RPN      | ResNet101 +DCN | Faster RCNN | – | – | 47.6 | NeurIPS18 | 効率的なマルチスケールの訓練戦略、正解インターセクタの周囲のコンテキスト領域を適切なスケールで処理。 |
| OHEM [2-46]| SS       | VGG16 | Fast RCNN | 78.9 (07+12) | 76.3 (07+12) | 22.4 | CVPR16 | Online Hard Example Mining アルゴリズム |
| FactorNet [204]| SS       | GoogLeNet | RCNN | – | – | – | CVPR16 | 様々な物体カテゴリのサンプル数の不均衡を特定、分割統治による表現学習スキームを提案。 |
| Chained Cascade [205]| SS | CRAFT | VGG Inception2 | Fast RCNN, Faster RCNN | 80.4 (07+12) | (SS+VGG) | – | ICCV17 | DCNN とカスクード分割器の複数のステージを共同で学習、Fast RCNN と Faster RCNN の両方で、PASCAL VOC 2017 と ImageNet での検出精度を向上。データセットによって異なる物体領域不均衡手法を使用。 |
| Cascade RCNN [23]| RPN      | VGG | ResNet101 +FPN | Faster RCNN | – | – | 42.8 | CVPR18 | DCNN とカスクード分割器の複数のステージを共同で学習、これらのステージの学習では、正例選択の位置設定精度として異なる囲を用意、複数のスヌーク bbox 回帰を積み重ねる |
| RetinaNet [168]| – | ResNet101 +FPN | RetinaNet | – | – | – | ICCV17 | hard example での検出に焦点を当てる視覚の Focal Loss を提案。1段階検出器を訓練する際の正例・負例の不均衡問題にうまく対処。 |

立つ。テスト時、またまたはその両方で使用できる。しかし、訓練に必要な時間が大幅に増加するという明らかに欠点がある。データ拡張は完全に新しい訓練画像を作成してもよい [210, 280] が、合成画像が実画像にうるそくを欠いていることを保証するのは困難である。一部の研究者 [64, 94] は、セグメント化されたリアルな物体を自然画像に貼り付けることでデータセットを拡張することを提案した。また、Dvornikら [63] は、物体を適切に配置するには、物体を観察する視覚コンテキストの適切なモデリングが必要であると示し、データ拡張法の適切な新物体配置位置を自動的に見つけるコンテキストモデルを提案した。

従来の訓練戦略。広範囲のスケール変動下での物体の検出、特に非常に小さい物体の検出は、主要課題として突出している。画像解像度が検出精度に大きな影響を持つことが示されており [120, 175]、高解像度な方が小さい物体の検出可能性が高いため [120]、スケーリングはデータ拡張の中でも特に一般的に使用される。近年 Singh らは、スケール変動の不均一を説明し、先進的で効率的なデータ拡張手法である SNIP [249] および SNIPER [251] （表 10）を提案した。SNIP の特徴は、小さな物体の検出と大きな物体の検出を並行して訓練することで、物体の検出精度を向上させる。また、SNIPER は、画像ピラミッド全体を処理することによって、正解物体が複数のコンテキスト領域をもつ適切なスケールで処理することで、効率的なマルチスケール訓練を可能にする。Peng ら [209] は訓練の重要な要素であるリバサマツエイズについて研究し、Large MiniBatch 物体検出器である MegDet を提案し、以前よりはるかに大きなミニバッチサイズ（RetinaNet や Mask RCNN の 16 に対する MegDet では最大 256）での訓練を可能にした。収束の遅さの回避と計算効率の大幅な高速化のために、Peng ら [209] は学習率ポリシーと Cross-GPU Batch Normalization を提案し、128 GPU を効率的に活用した。これにより、MegDet は 128 GPU を使って 4 時間で COCO の訓練を完了できるようになり、COCO 2017 検出チャレンジで優勝した。

位置推定エラー低減、物体検出では、検出されたbbox とその正解矩形の Intersection Over Union 25（IOU）が最も一般的な評価基準であり、正例と負例を定義するには IOU 閾値（例：典型価の 0.5）が必要である。図 13 から分かるように、最先端の検出器 [84, 175, 102, 229, 227] のほとんどで物体検出はマルチタスク学習問題として定式化される。つまり、物体領域候補にクラスタペルを割り当てる softmax 分類器と、検出結果と正解の IOU またはその他の指標を最大化することで物体を位置推定するbbox 回帰器を共に最適化する。bbox は結合さされた物体の不均一性に近似を極力さないために、背景画像がしばしばbbox に含まれる分類と位置設定の精度に影響。[108]の研究では、類似物体間の混同に加えて物体の位置推定エラーが最も影響の大きいエラーの形式の一つであることが示されている。位置推定エラーは、オーバーラップ不足（図 20 の绿色の矩形のように必要とされる IOU 閾値よりも小さい状態や重複検
出（つまり、一つの物体インスタンスに対して重なり合った複数の検出がなされる状態）が原因となり得る。通常、重複検出を除去するために非最大抑制（Non-Maximum Suppression, NMS）[18, 111]のような何らかの処理工程が用いられる。しかし、(図20)に示す紫色の矩形のように位置推定精度の良いbboxがNMS中の照準ミスで抑制され、位置推定の品質が低下する可能性がある。そのため、位置推定エラー低減による検出性能向上を目的とした手法が多くある。

MRCNN[82]はRCNNを数回適用する反復bbox回帰を導入する。RAFT[292]とAttractioNet[83]はマルチステージ検出サブネットワークを使用し、正確な物体領域候補を生成してFastRCNNに転送する。CaiとVasconcelosはRCNNを多段階に拡張し、close false positive（正確な形に近いが正解とはならない信値のbbox）に対して順次選択性を高めるために、一次の検出器をIOU関数を増加させながら順次訓練するCascadeRCNN[23]とした。これは、特定のIOUで訓練された検出器の出力が、その次より高いIOU関数を検出器を訓練するのに適した分野にいるという観点に基づいており、このアプローチは任意のRCNNベースの検出器で構築でき、わずかな計算の増加で、ベースライン検出器の強さに関係なく一貫した精度向上（約2～4ポイント）を達成することが保証されている。また、最近では直接IOUを最適化的目的の関数として定式化する研究[128, 232, 121]や、SoftNMS[18]やlearningNMS[111]のように改良NMSの結果を提示する研究[18, 104, 111, 270]もある。

クラス不均衡の対処。画像分類にはない物体検出特有の別の問題がある。ラベル付きの物体インスタンスの数と背景の例（どの対象物体クラスにも属さない物体インスタンスの数）の2つの値が不均衡である。これは、背景の例のほとんどは簡単な負例だが、この不均衡により訓練が極めて非効率的になり多数の簡単な負例が訓練を難しくしている。過去には、この問題は通常boosting[259]などの技術で対処してきた。最近では、この問題にもいくつか注目が集まっている[153, 168, 246]。物体領域候補生成の段階ではほとんどの背景領域が速くな Busy で除外され少数の物物体候補が生成されるため、このクラス不均衡問題は2段階検出器[85, 84, 229, 102]ではある程度軽減される。前処理の選択肢やバランスを維持するために、Online Hard Example Mining（OHEM）[246]などのsampling技術が提案されている。このアプローチを使用してもよい。1段階物体検出器[227, 175]の場合には、この不均衡は非常に深刻である（例えば、全体対象に対して背景の例が100,000）。Liuら[168]はFocalLossを提案し、正しく分類された例について割り当てられた損失への重み付けを小さくするように、交差エントロピー損失を修正することでこの不均衡に対処した。Liuら[153]は勾配ノルム分布の観点からこの問題を研究し、対処のためにGradient Harmonizing Mechanism（GHM）を提案した。

10 論議（考察）と結論

一般物体検出はコンピュータビジョンにおける重要かつチャレンジングな問題であるがその注目を集めている。深層学習技術の発展により、物体検出の分野は劇的に進化した。一般物体検出のための深層学習に関する包括的なサバイバートイで、本論文では近年の成果に焦点を当て、検出における役割に基づく構造的・手法分類を提供し、既存の一般的なデータセットと評価基準をまとめ、最も実用的な手法の性能について説明した。以下、10.1節で最前端技術について議論し、10.2節で主要な問題の総合的な議論を行い、最後に10.3節で将来の研究の方向性を提案して本レビューを締めくくる。

図21 COCOの物体検出性能の進歩（Test-Dev の結果）。結果は[84, 102, 230]から引用。backboneネットワーク、検出フレームワークの設計、そして適切な大规模データセットを利用してできるが、検出精度において最も重要な3要素である。

10.1 最先端（state-of-the-art）の性能

多様な検出器が過去数年間に登場し、PASCAL VOC[68, 69]、ImageNet[234]、COCO[166]などの標準ベンチマークの導入により検出器の比較が容易になった。前述した5節から9節での議論からわかるように、元々報告された性能（例：精度・速度）での検出器の比較は誤解を招くことがある。何故なら、以下の選択肢を含む根本的または文脈上の点でそれらは異なるためである。

- RCNN[85]、FastRCNN[84]、RFPN[229]、RFCN[102]、MaskRCNN[102]、YOLO[227]、SSD[175]などのメタ検出フレームワーク
- VGG[248]、Inception[263, 125, 264]、ResNet[101]、ResNetXt[291]、Xception[45]などの表6に記載したbackboneネットワーク
- 異なる層の特徴の組み合わせ[167, 247, 77]、従来のconvolutional network[51]、従来のRoI pooling[203, 51]、重いhead[231, 209]、軽いhead[165]などの技術革新
- ImageNet[234]、COCO[166]、Places[319]、JFT[106]、OpenImages[139]などのデータセットでの事前学習
- 物体領域候補生成手法の差異と物体領域候補の差異
- 訓練時/テスト時のデータ拡張、新しいマスクケアール訓練戦略[249, 251]など、およびモデルのアンサンブル

最近提案された全ての検出器を比較することは現実的かもしれない。それでも、公開されている代表的な検出器を共通プラットフォームに統合し適用された方法で比較することは有益である。backboneネットワーク、画像解像度、矩形候補数を変化させて3種の主要な検出器（FastRCNN[229]、RFCN[50]、SSD[175]）を比較したHuangらの研究[120]でさらに、この点では研究が非常に限られている。

表7、8、9、10、11から分かることは、我々は多数の手法について広く使用される3つの標準ベンチマークで報告された最高性能をまとめている。これらの手法の結果、上記の1つ以上の差異に異なるにもかかわらず同じテストベンチマークで報告されている。

図5、21は、PASCAL VOC、ILSVRC、MSCOCOチャレンジの最高の検出結果をまとめ、最前端技術の概要を非常に
10.2 まとめと議論

本論文を通じて多数の参考文献と手法を議論してきたため、ここでは、深層学習に基づく一般物体検出で出現した主要な要因に焦点を当てる。

(1) 検出フレームワーク：2段階 vs. 1段階

5 節では、領域ベース（2段階）検出器と統合された（1段階）検出器という2つの主要な検出フレームワークのカテゴリーを特定した。

- 大きな計算コストが許容される場合、2段階検出器の方が構造が柔軟かつ領域ベースの分類に適しているため、一般に1段階検出器よりも高い検出精度が出る。このことでは、有名な検出チャレンジで使用されるアプローチのほとんどが、主に2段階フレームワークに基づいているという事実から明らかである。最も広く使用されているフレームワークは、Faster RCNN [229]、RFCN [50]、Mask R-CNN [102]である。

- [120]で示されてのように、1段階検出器である SSD [175]の検出精度は、代表的な2段階のフレームワークと比べ backbone ネットワークの品質の影響を受けない。

- YOLO [227]や SSD [175]のような1段階検出器は、処理のアルゴリズムを回避し、軽量の backbone ネットワークを使用し、より少ない候補領域で予測を実行し、分類サブネットワークを全領域に組み込みしているため、一般に2段階検出器よりも高速である。ただし、2段階検出器に同様の技術を導入することでリアルタイムで実行できる。1段階と2段階のいずれかにせよ、最も時間かかる工程は特徴抽出（backbone ネットワーク）である [146, 229]。

- [120, 227, 175]で示されように、YOLOや SSDのような1段階フレームワークは Faster RCNN や RFCNのような2段階のアーキテクチャと比べ、大きな物体の検出では競争力があるが、小さな物体の検出時の性能は通常は劣っている。

検出フレームワークの各段階を攻めることで、より良い（高速・正確・頼頼）検出器を構築することが求められている。1段階か2段階かマルチステージかに関係なく、検出フレームワークの設計はいくつかの重要な設計上の選択に取束していている。

- 全層畳込みのパイプライン
- 他の関連タスクからの補足情報の探求（例：Mask R-CNN [102]）
- スライディングウィンドウ [229]
- backbone の異なる層からの情報の融合

COCO や他のチャレンジでの物体検出 [23, 40, 41] とインスタンススケーミングに影響のある近年のバケッテの成功が確認されており、マルチステージ物体検出は速度・精度・トレードオフ改善のための将来のフレームワークとなり得る。2018 WIDER Challenge [180] でティーザー調査が行われている。

(2) backbone ネットワーク

6.1 節で説明したように識別的な物体特徴表現が重要な役割を果たすため、backbone ネットワークは特性を連続的に変化させ、個々のマップを特徴量を基にした特徴表現を構築するためのリングストリームのフレームワークとなり得る。2018 WIDER Challenge [180] でティーザー調査が行われている。

(3) 物体表現の特徴量の向上

実世界の画像の変動は物体認識の主要課題である。変動には、照明・姿勢・変形・背景の乱達・遮蔽・ブラー・解像度・ノイズ・カメラの歪みが含まれる。

(3.1) 物体スケールと小さな物体サイズ

物体スケールの大きな変動は、特に小さな物体で大きな課題となる。ここでは、6.2 節で特定した主な戦略に関する要約と議論を述べる。

- 画像ビラミッドの使用：単純かつ効果的で、小さな物体の拡大や小さな物体の縮小に役立つ。計算コストが高いにもかかわらず、精度向上のために推奨的に一般的に使用される。

- 異なる解像度の畳込み層からの特徴の使用：SSD [175]のような初期の研究では予測は独立で実行され、他の層からの情報を組み合わせたり混合させたりしない。今日では、例えば FFN [167]のように様々な層の特徴の組み合わせがより実パフォーマンスに優れる。

- dilated convolution [164, 163]の使用：より広いコンテンツを収容するための手法で、一般的に使用される。

- 様々なスケールとアンペド比の anchor box の使用：多数のパラメータを持つ欠点がある上、anchor box のスケールとアンペド比は推奨的プロミックスに決定される。

- 高解像度ネットワーク [255, 256]の発展させることができる。超解像技術が検出精度を改善するかどうかは不明のままである。

近年の進歩にもかかわらず、依然小さな物体の検出精度は大き、物体の検出精度よりは低く低い。したがって、小さな物体の検出は物体検出における主要課題の1つであり、自律運転などの特定のアプリケーションは、大きな領域内の小さな物体の存在の識別のみを必要とし、正確な位置推定は不要であるため、おそらく位置推定の要件をスケールの関数として一般化す
これらの課題に基づき今後の研究の方向性を以下に述べる。

(1) オープンワールドの学習：究極の目標は、人間の視覚システムに匹敵するレベルで、オープンワールドシーンにおける数千万以上の物体カテゴリのインスタンスを正確かつ効率的に認識し位置推定をできる物体検出を開発することである。物体検出アルゴリズムは、従来の物体カテゴリを認識[144, 95]する能力を持つべきだ。一般的には訓練データセット外の物体カテゴリを認識できない。現在の検出データセット[68, 234, 166]は、人間が認識できるカテゴリより少なくない数から数百万のカテゴリが含まれていない。かななり多くのカテゴリを持つ新たな大規模データセット[107, 250, 226]を構築する必要がある。

(2) より良くより効率的な検出フレームワーク：領域ベースの検出器（RCNN[85]、Fast RCNN[84]、Faster RCNN[229]、Mask RCNN[102]と1段階検出器（YOLO[227]、SSD[175]）の両方を、優れた検出フレームワークが開発されてきたことが一般物体検出の成功の一つである。領域ベースの検出器は精度がある。1段階検出器は一般に高速かつ単純である。物体検出器は下層にあるbackboneネットワークに大きく依存している。backboneは画像分類用に最適化されているため、学習にバイアズが生じる可能性がある。物体検出器のクラスチェッキング学習が新しい検出フレームワークの開発に役立つだろう。

(3) コンペティションで効率的なCNN特徴：CNNは数値（AlexNet[141]）から数百層（ResNet[101]、DenseNet[118]）へと著しく深さが増加している。これらのネットワークには数百から数百万のパラメータがあり、訓練に大量のデータとGPUを要する。ネットワークの冗長性を低減しまたは除去する。コンペティションで軽量ネットワークの設計[29, 4, 119, 112, 169, 300]やネットワークの高速化[144, 122, 253, 155, 188, 282]を行う研究への関心が高い。

(4) ニューラルネットワークのアーキテクチャの自動探索：深層学習は、強大なドメイン知識を持つ人間の専門家が必要な人手による特徴エンジニアリングを回避する。しかし、DCNNも同様にかなりの専門知識を必要とする。画像分類や物体検出に適用されている[22, 39, 80, 171, 331, 332]年の自動化学習：AutoML[219]のように、検出backboneのアーキテクチャの自動設計を検討するのは自然である。

(5) 物体インスタンスセグメンテーション：画像コンテンツの豊かでより詳細な理解のため、画素レベルの物体インスタンスセグメンテーション[166, 102, 117]に取り組む必要がある。物体インスタンスセグメンテーションは、個々の物体が正確な境界を必要とする潜在的なアーキテクチャで重要な役割を果たす。

(6) 弱教師あり検出：現在の最先端の検出器は、bboxまたはセグメンテーションマスクのラベル付きデータ[69, 166, 234]から学習した完全な教師ありモデルを採用している。しかし、特にbboxアノテーションの収集が多くの人手を要する場合や画像数が多い場合に、完全な教師あり学習は深刻な制約となる。完全なラベル付き訓練データがなければ完全な教師あり学習はスケールしないため、弱教師ありデータや部分的アノテーションされたデータのみが与えられる[17, 55, 244]場合に、CNNの力かいかにして引き出すかの理解が不可欠である。

(7) 少数サンプル物体検出とゼロショット検出：検出器の成功率は膨大な量のアノテーション付き訓練データを大々依存している。ラベル付きデータの不足している場合、検出器の性能はしばしば低下し性能に依存する。これは対照的に、人間は（とくに子供でも）非常に少ない例から即座に視覚的な概念を学ぶことができ、たいていうまく処理（一般化）できる[16, 144, 71]。そのため、わずか数例から
学習する少数ショット（few shot）検出の能力は非常に魅力的である [30, 61, 75, 129, 144, 228, 237]。さらに制約されたセリオドット（zero shot）物体検出は、これまでに見ることのない物体クラスを位置推定し認識する [26, 9, 53, 222, 221]。これは、新しい物体カテゴリーを知るかかつ段階的に発見する必要がある生涯学習（life-long learning）を行う機械に不可欠である。

(8) 他のモザリティでの物体検出：ほとんどの検出器は2次元の静止画に基づくが、他のモザリティでの物体検出は、自動運転車・無人航空機・ロボット工学などの分野との関連が深い。デブス（暗 nir） [36, 211, 289, 266], 動画 [70, 130], 点群 [217, 218]の効果的な使用に関して、これらのモザリティは新たな課題を提起する。

(9) 普通検出：近年、自然画像・動画・航空画像・医用CT 画像など、複数の画像ドメインで有効 [224, 225] な普通表現（universal representation）を学習する取り組みが増えており、そのような研究のほとんどの画像分析の焦点を占めており、物体検出を対象とすることは限界が少なく、通常開発された検出器はドメイン特化である。画像ドメインでの非既存の物体検出とユースコンドミの物体検出は、重要な将来の方向性を示している。

一般物検出の研究分野はまだ完全にはほど遠い。しかし、過去5年間のブレースルールを考え、我々は将来の発展と契機について楽観的である。

謝辞 オープンアクセス資金はオウル大学病院を含むオウル大学によって提供されている。一般物検出と他の関連分野の先駆者である研究者に感謝する。また、コメントと提案をいただいた編集委員の Jiri Matas 教授と匿名の査読者に心から感謝する。本研究は、オウル大学（フィンランド）Center for Machine Vision and Signal Analysis, および中国国家自然科学基金（Grant 61872379）の助成を受けている。

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made。

参考文献

[1] Agrawal, P., Girshick, R., Malik, J. (2014) Analyzing the performance of multi-layer neural networks for object recognition. In: ECCV, pp. 329–344
[2] Alexei B., Deselaers T., Ferrari V. (2010) What is an object? In: CVPR, pp. 73–80
[3] Alexei B., Deselaers T., Ferrari V. (2012) Measuring the objectness of image windows. IEEE TPAMI 34(4):2189–2202
[4] Alvarez J., Salzmann M. (2016) Learning the number of neurons in deep networks. In: NIPS, pp. 2270–2278
[5] Andropoulos A., Tsotsos J. (2013) 50 years of object recognition: Directions forward. Computer Vision and Image Understanding 117(8):827–891
[6] Arbeláez P., Hanharian H., Gu C., Gupta S., Bourdev L., Malik J. (2012) Semantic segmentation using regions and parts. In: ECCV, pp. 3378–3385
[7] Arbeláez P., Pont-Tuset J., Barron J., Marques F., Malik J. (2014) Multi-scale combinatorial grouping. In: CVPR, pp. 328–335
[8] Azzopardi H., Razavan A., Sullivan J., Maki A., Carlsson S. (2016) Factors of transferability for a generic convnet representation. IEEE TPAMI 38(9):1790–1802
[9] Bansal A., Sikka K., Sharma G., Chellappa R., Divakaran A. (2018) Zero shot object detection. In: ECCV 28
[10] Bar M. (2004) Visual objects in context. Nature Reviews Neuroscience 5(8):617–629
[11] Bell S., Lawrence Z., Bala K., Girshick R. (2016) Inside Outside Net: Detecting objects in context with skip pooling and recurrent neural networks. In: CVPR, pp. 2874–2883
[12] Belongie S., Malik J., Puzicha J. (2002) Shape matching and object recognition using shape contexts. IEEE TPAMI 24(4):509–522
[13] Bengio Y., Courville A., Vincent P. (2013) Representation learning: A review and new perspectives. IEEE TPAMI 35(8):1798–1828
[14] Biederman I. (1972) Perceiving real world scenes. IJCV 177(7):77–80
[15] Biederman I. (1987) Recognition by components: a theory of human image understanding. Psychological Review 94(2):115
[16] Biederman I. (1987) Recognition by components: a theory of human image understanding. Psychological Review 94(2):115
[17] Bilen H., Vedaldi A. (2016) Weakly supervised deep detection networks. In: CVPR, pp. 2846–2854
[18] Bodla N., Singh B., Chellappa R., Davis L. S. (2017) SoftNMS improving object detection with one line of code. In: ICCV, pp. 5562–5570
[19] Borji A., Chen M., Jiang H., Li J. (2014) Salient object detection: A survey. arXiv:14115878v1:1–26
[20] Bourdev L., Brandt J. (2005) Robust object detection via soft cascade. In: CVPR, vol 2, pp. 236–243
[21] Bruna J., Mallat S. (2013) Invariant scattering convolution networks. IEEE TPAMI 35(8):1872–1886
[22] Cai H., Yang J., Zhang W., Han S., Yu Y. (2018) Path level network transformation for efficient architecture search. In: CVPR
[23] Cai Z., Vasconcelos N. (2018) Cascade RCNN: Delving into high quality object detection. In: CVPR, pp. 2846–2854
[24] Cai Z., Fan Q., Feris R., Vasconcelos N. (2016) A unified multiscale deep convolutional neural network for fast object detection. In: ECCV, pp. 354–370
[25] Carreira J., Simo-suscescu C. (2012) CMPC: Automatic object segmentation using constrained parametric mincuts. IEEE TPAMI 34(7):1132–1138
[26] Chatfield K., Simonyan K., Vedali A., Zisserman A. (2014) Return of the devil in nets: Delving deep into convolutional nets. In: BMVC 23
[27] Chavali N., Agrawal H., Mahendru A., Batra D. (2016) Object proposal detection using shape contexts. IEEE TPAMI 24(4):509–522
[28] Chen Y., Li J., Xiao H., Jin X., Yan S., Feng J. (2017) Dual path networks. In: CVPR, pp. 2846–2854
[29] Chen X., Gupta A. (2017) Spatial memory for context reasoning in object detection. In: CVPR
[30] Chen X., Kundu K., Zhu Y., Han M., Fidler S., Urtasun R. (2015) 3D object proposals for accurate object class detection. In: NIPS, pp. 424–432
[31] Chen Y., Li J., Xiao H., Jin X., Han S., Feng J. (2017) Dual path networks. In: NIPS, pp. 4467–4475
[32] Chen Y., Rohrbach M., Yan Z., Yan S., Feng J., Kalantidis Y. (2019) Graph based global reasoning networks. In: CVPR
[33] Chen Y., Tong T., Zhang X., Meng G., Pan C., Sun J. (2019) DeNets: Neural architecture search on object detection. arXiv:190310979
[34] Cheng B., Wei Y., Shi H., Feris R., Xiong J., Huang T. (2018) Decoupled classification refinement: Hard false positive suppression for object detection. arXiv:181004002
[35] Cheng B., Wei Y., Shi H., Feris R., Xiong J., Huang T. (2018) Revisiting RCNN: on awakening the classification power of faster RCNN. In: ECCV
[36] Cheng G., Zhou P., Han J. (2016) RIFDCNN: Rotation invariant and fisher discriminative convolutional neural networks for object detection. In: CVPR, pp. 2884–2892
[37] Cheng M., Zhang Z., Lin W., Torr P. (2014) BING: Binarized normed gradients for objectness estimation at 300fps. In: CVPR, pp. 3286–3295
[38] Cheng Y., Wang D., Zhao P., Zhang T. (2018) Model compression and acceleration for deep neural networks: The principles, progress, and challenges. IEEE Signal Processing Magazine 35(1):126–136
[39] Chollet F. (2017) Xception: Deep learning with depthwise separable convolutions. In: CVPR, Perceiving real world scenes. IJCV 177(7):77–80
[40] Cimino R., Verbeck J., Schmid C. (2017) Weakly supervised object localization with multi-fold multiple instance learning. IEEE TPAMI 39(1):189–203
[41] Csurka G., Dance C., Fan L., Willamowski J., Bray C. (2004) Visual categorization with bag of keypoints. In: ECCV Workshop on statistical learning in computer vision 5
[42] Dai J., He K., Li Y., Ren S., Sun J. (2016) Instance sensitive fully convolutional networks. In: ECCV, pp. 534–549
[43] Dai J., He K., Sun J. (2016) Instance aware semantic segmentation via multitask network cascades. In: CVPR, pp. 3150–3158

26ただし wikipediaのページや属性 untouchedなどの補助情報(side information) が与えられる場合がある。
are convolutional neural networks. In: CVPR, pp. 437–446

13

Deng J., Xue N., Long Y., Xia G., Lu Q. (2018) Learning Rol for detecting oriented objects in aerial images. In: CVPR

19

Deng J., Li Z., Li Y., Zhang X., Bai C., Song G. (2014) An empirical study of context in object detection. In: CVPR. pp. 1271–1278 21, 27

10

Dollar P., Wojek C., Schiele B., Perona P. (2012) Pedestrian detection: An evaluation of the state of the art. IEEE TPAMI 34(4):743–761 2, 3

13

Donahue J., Jia Y., Vinyals O., Hoffman J., Zhang N., Tzeng E., Darrell T. (2014) DeCAF: A deep convolutional activation feature for generic visual recognition. In: ICML, vol 32, pp. 647–655 15

13

Donahue J., Deng J., Jia Y., Sohn S., Braida L., Shlens J., Darrell T. (2015) DeCAF: a deep convolutional feature model size. In: arXiv preprint arXiv:1602.07360

10

Donahue J., Proppe S., Torralba A., LeCun Y., Baltrušaitis T., Girshick R. (2017) Deep convolutional networks for mobile vision applications. In: CVPR

13

Divvala S., Bhattacharya A. M., He K., Van Gool L. (2017) Weakly supervised cascaded convolutional networks. In: CVPR, vol 3, p 27

13

Dogrusoz U., Dahan D., Gurevich Y., Goldman S., Girshick R., Dollár P. (2019) The Pascal visual object classes (VOC) challenge. IJCV 88(2):303–338

13

Dollar P., Wojek C., Schiele B., Perona P. (2012) Pedestrian detection: An evaluation of the state of the art. IEEE TPAMI 34(4):743–761 2, 3

13

Dollar P., Wojek C., Schiele B., Perona P. (2012) Pedestrian detection: An evaluation of the state of the art. IEEE TPAMI 34(4):743–761 2, 3

13

Dollar P., Wojek C., Schiele B., Perona P. (2012) Pedestrian detection: An evaluation of the state of the art. IEEE TPAMI 34(4):743–761 2, 3

13

Dollar P., Wojek C., Schiele B., Perona P. (2012) Pedestrian detection: An evaluation of the state of the art. IEEE TPAMI 34(4):743–761 2, 3

13

Dollar P., Wojek C., Schiele B., Perona P. (2012) Pedestrian detection: An evaluation of the state of the art. IEEE TPAMI 34(4):743–761 2, 3

13

Dollár P., Malladi R., Girshick R., Malik J., Lin T. Y. (2014) Mask RCNN. In: ICCV. pp. 297–312 1

13

Dollár P., Malladi R., Girshick R., Malik J., Lin T. Y. (2014) Mask RCNN. In: ICCV. pp. 297–312 1

13

Dollár P., Malladi R., Girshick R., Malik J., Lin T. Y. (2014) Mask RCNN. In: ICCV. pp. 297–312 1

13

Dollár P., Malladi R., Girshick R., Malik J., Lin T. Y. (2014) Mask RCNN. In: ICCV. pp. 297–312 1

13

Dollár P., Malladi R., Girshick R., Malik J., Lin T. Y. (2014) Mask RCNN. In: ICCV. pp. 297–312 1

13

Dollár P., Malladi R., Girshick R., Malik J., Lin T. Y. (2014) Mask RCNN. In: ICCV. pp. 297–312 1

13

Dollár P., Malladi R., Girshick R., Malik J., Lin T. Y. (2014) Mask RCNN. In: ICCV. pp. 297–312 1

13

Dollár P., Malladi R., Girshick R., Malik J., Lin T. Y. (2014) Mask RCNN. In: ICCV. pp. 297–312 1

13

Dollár P., Malladi R., Girshick R., Malik J., Lin T. Y. (2014) Mask RCNN. In: ICCV. pp. 297–312 1

13

Dollár P., Malladi R., Girshick R., Malik J., Lin T. Y. (2014) Mask RCNN. In: ICCV. pp. 297–312 1

13

Dollár P., Malladi R., Girshick R., Malik J., Lin T. Y. (2014) Mask RCNN. In: ICCV. pp. 297–312 1

13

Dollár P., Malladi R., Girshick R., Malik J., Lin T. Y. (2014) Mask RCNN. In: ICCV. pp. 297–312 1

13

Dollár P., Malladi R., Girshick R., Malik J., Lin T. Y. (2014) Mask RCNN. In: ICCV. pp. 297–312 1

13

Dollár P., Malladi R., Girshick R., Malik J., Lin T. Y. (2014) Mask RCNN. In: ICCV. pp. 297–312 1

13

Dollár P., Malladi R., Girshick R., Malik J., Lin T. Y. (2014) Mask RCNN. In: ICCV. pp. 297–312 1

13

Dollár P., Malladi R., Girshick R., Malik J., Lin T. Y. (2014) Mask RCNN. In: ICCV. pp. 297–312 1

13

Dollár P., Malladi R., Girshick R., Malik J., Lin T. Y. (2014) Mask RCNN. In: ICCV. pp. 297–312 1

13

Dollár P., Malladi R., Girshick R., Malik J., Lin T. Y. (2014) Mask RCNN. In: ICCV. pp. 297–312 1

13

Dollár P., Malladi R., Girshick R., Malik J., Lin T. Y. (2014) Mask RCNN. In: ICCV. pp. 297–312 1

13

Dollár P., Malladi R., Girshic...
pose estimation. In: ECCV, pp. 483–499 14, 17, 18

[195] Newell A., Huang Z., Deng J. (2017) Associative embedding: end to end learning for joint detection and grouping. In: NIPS, pp. 2277–2287 13

[196] Ojala T., Pietikäinen M., Maenpää T. (2002) Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE TPAMI 24(7):971–987 5, 21

[197] Oliva A., Torralba A. (2007) The role of context in object recognition. Trends in cognitive sciences 11(2):520–527 21, 27

[198] Opelt A., Pinz A., Fussenegger M., Auer P. (2006) Generic object recog-

[199] Oquab M., Bottou L., Laptev I., Sivic J. (2014) Learning and transferring

[200] Oquab M., Bottou L., Laptev I., Sivic J. (2015) Is object localization for free?

[201] Osuna E., Freund R., Girosi F. (1997) Training support vector machines: an

[202] Ouyang W., Wang X. (2013) Joint deep learning for pedestrian detection.

[203] Ouyang W., Wang X., Zeng C., Yang X. (2016) Factors in finetuning

[204] Ouyang W., Wang K., Zhu X., Wang X. (2017) Chained cascade network

[205] Ouyang W., Zeng X., Wang X., Qiu S., Luo P., Tian Y., Li H., Yang S., Wang Z., Loy C.-C., et al. (2015) DeepIDNet: Deformable deep convolu-

[206] Paszke A., Gross S., Chintala S. (2017) PyTorch: An imperative style,

[207] Perreault D., Zitnik M., Pecevski D., Tenenbaum J. B. (2012) Exploring thin images: The roles of appearance and contextual information for machine and human object recognition. IEEE TPAMI 34(10):1978–1991 21

[208] PASCAL VOC detection leaderboard (2018) http://host.robots.

[209] Peng C., Xiao T., Liu J., Yang Z., Xing J., Jia X., Yu G., Sun J. (2018)

[210] Peng X., Sun B., Ali K., Saenko K. (2015) Learning deep object detectors from ground models. In: CVPR, pp. 1278–1286 24

[211] Pepik B., Bener S., Ritschel T., Schiele B. (2015) What is holding back

[212] Perronnin F., Sánchez J., Messink M. (2010) Improving the fisher kernel for large scale image classification. In: ECCV, pp. 143–156 5, 14

[213] Pinheiro P., Collobert R., Dollar P. (2015) Learning to segment object can-

[214] Pinheiro P., Lin T., Collobert R., Dollar P. (2015) Learning to refine object

[215] Ponce J., Hebert M., Schmid C., Zisserman A. (2007) Toward Category

[216] Pouyanfar S., Sadig S., Yan Y., Tian H., Tao Y., Reyes M.P., Shyu M., Chen S., Iyengar S. (2018) A survey on deep learning: Algorithms, tech-

[217] Qi C.-R., Su H., Mo K., Guibas L. J. (2017) PointNet: Deep learning on

[218] Qiu C.-R., Liu W., Wu C., Su H., Guibas L. J. (2018) Frustum pointnets for

[219] Quanming Y., Mengshuo W., Hugo J. E., Isabelle G., Yuqi H., Yufeng L., Weiwei T., Qiang Y., Yang Y. (2018) Taking human out of learning applications: A survey on automated machine learning. arXiv:181013306 27

[220] Rabinovich A., Vedaldi A., Galleguillo C., Wiewiora E., Belongie S. (2007) Objects in context. In: ICCV, 21, 27

[221] Rahman S., Khan S., Barnes N. (2018) Polarity loss for zero shot object
detection. arXiv preprint arXiv:1811.02356 25

[222] Rahman S., Khan S., Polino F. (2019) Zero shot object detection: Learning to simultaneously recognize and localize novel concepts. In: ACCV 28

[223] Razavian A., Azizpour H., Sullivan J., Carlsson S. (2014) CNN features off

[224] Rebuffi S., Bilen H., Vedaldi A. (2017) Learning multiple visual domains

[225] Ren M., Triantafillou E., Ravi S., Snell J., Swersky K., Tenenbaum J. B.,

[226] Ren S., He K., Girshick R., Sun J. (2015) Faster R-CNN: Towards real

time object detection with region proposal networks. In: NIPS, pp. 91–99 9, 11, 12, 13, 14, 15, 21, 22, 23, 24, 25, 26, 27, 34

[227] Ren S., He K., Girshick R., Zhang X., Sun J. (2017) Object detection net-

[228] Rezatofigh H., Tsoi N., Gwak J., Sadehian A., Reed L., Savenas R. (2019)

[229] Redmon J., Farhadi A. (2017) YOLO9000: Better, faster, stronger. In: CVPR, 13, 15, 27, 34

[230] Redmon J., Divvala S., Girshick R., Farhadi A. (2016) You only look once: Unified, real-time object detection. In: CVPR, pp. 779–788 14, 15, 24, 25, 26, 27, 34

[231] Ren M., Triantafillou E., Ravi S., Snell J., Swersky K., Tenenbaum J. B.,

[232] Ren S., He K., Girshick R., Sun J. (2015) Faster R-CNN: Towards real
transformations for deep neural networks. In: CVPR
wild: Past, present and future. Computer Vision and Image Understanding
neuron importance score propagation. CVPR
lusions
ences Magazine 13(3):55–75
138:1–24 2, 3
307 Zagoory S., Lerner A., Lin T., Pinheiro P., Gross S., Chantala S., Dollar P. (2015) A multiscale object detector. In: BMVC 16, 23, 21
306 Zeiler M., Fergus R. (2014) Visualizing and understanding convolutional networks. In: ECCV, pp. 818–833 6, 14, 15, 21
305 Zeng X., Ouyang W., Yang B., Yan J., Wang X. (2016) Gated bidirectional recurrent networks for object detection. In: CVPR
304 Zeng X., Ouyang W., Yan J., Li, Xiao T., Wang K., Liu Y., Zhou Y., Yang B., Wang Z., Zhou H., Wang X. (2017) Crafting gdbnet for object detection. IEEE TPAMI 20, 21, 22, 27
303 Zhang K., Zhang Z., Li, Z., Zeng X., Qiao Y. (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE SPI 23(10):1499–1503 2
302 Zhang L., Lin L., Liang X., He K. (2016) Is faster RCNN doing well for pedestrian detection? In: ECCV, pp. 443–457 2
301 Zhang S., Wen L., Bux J., Lei Y., Li S. (2018) Single shot refinement neural network for object detection. In: CVPR 16, 17, 18
300 Zhang S., Yang J., Schiele B. (2018) Occluded pedestrian detection through guided attention in CNNs. In: CVPR, pp. 2056–2063 19
299 Zhang X., Yang Y., Han Z., Wang H., Gao C. (2013) Object class detection: A survey. ACM Computing Surveys 46(1):10.10–15, 1, 23, 27
298 Zhang X., Li, Z., Change Loy C., Lin D. (2017) PolyNet: a pursuit of structural diversity in very deep networks. In: CVPR, pp. 718–726 20
297 Zhang X., Zhuo X., Lin M., Sun J. (2018) ShuffleNet: an extremely effi cient convolutional neural network for mobile devices. In: CVPR 26
296 Zhang Z., Geiger I., Pohjalainen J., Mousa A. E., Schuller B. (2018) Deep learning for environmentally robust speech recognition: An overview of recent developments. ACM Trans Intell Syst Technol 9:40–26 21
295 Zhang Z., Qiao S., Xie C., Shen W., Wang B., Yuille A. (2018) Single shot object detection with enriched semantics. In: CVPR 17
294 Zhao Q., Sheng T., Wang Y., Tang Z., Chen Y., Cai L., Ling H. (2019) M2Det: A single shot object detector based on multilevel feature pyramid networks. In: ICCV, pp. 6369–6378
293 Zhou J., Cui G., Zhang Z., Yang C., Liu Z., Sun M. (2018) Graph convolutions and learnable filters for graph neural networks. In: CVPR, pp. 851–859
292 Zhou B., Kolesnikov A., Lapedriza A., Oliva A., Torralba A. (2015) Object detectors emerge in deep scene CNNs. In: ICLR 11, 15
291 Zhou B., Kolesnikov A., Lapiedra A., Oliva A., Torralba A. (2016) Learning deep features for discriminative localization. In: CVPR, pp. 2921–2929 11
290 Zhu B., Lapiedra A., Kolesnikov A., Oliva A., Torralba A. (2017) Places: A million image database for scene recognition. IEEE Trans Pattern Analysis and Machine Intelligence 7, 15, 25
289 Zhu J., Cui G., Zhang Z., Yang C., Liu Z., Sun M. (2018) Graph neural networks: A review of methods and applications. arXiv preprint arXiv:181208346
288 Zhu, P., Ni, B., Geng C., Hu, J., Xu Y. (2018) Scale transferable object detection. In: CVPR 15, 16, 17
287 Zhu, Y., Liu L., Shao L., Mellor M. (2016) DAVE: A unified framework for fast vehicle detection and annotation. In: ECCV, pp. 278–293 2
286 Zhu, Y., Ye Q., Qiu J., Jiao L. (2017) Oriented response networks. In: CVPR, pp. 4961–4970 19
285 Zhu X., Vondrick C., Fowlkes C., Ramanan D. (2016) Do we need more training data? IJCV 119(1):76–92
284 Zhu X., Tuia D., Mou L., Xia G., Zhu Y., Urtasun R., Salakhutdinov R., Fidler S. (2015) SegDeepM: Exploiting segmentation to improve deep neural networks for object detection. In: CVPR, pp. 4703–4711 20, 21
283 Zhu Y., Zhao C., Wang J., Zhao X., Wu Y., Lu H. (2017) CoupleNet: Coupling global structure with local parts for object detection. In: ICCV 20, 21
282 Zhu Y., Zhou L., Shao L., Mellor M. (2016) DAVE: A unified framework for fast vehicle detection and annotation. In: ECCV, pp. 278–293 2
281 Zhu, Y., Ye Q., Qiu J., Jiao L. (2017) Oriented response networks. In: CVPR, pp. 4961–4970 19
280 Zhu X., Vondrick C., Fowlkes C., Ramanan D. (2016) Do we need more training data? IJCV 119(1):76–92
279 Zhu X., Tuia D., Mou L., Xia G., Zhu Y., Urtasun R., Salakhutdinov R., Fidler S. (2015) SegDeepM: Exploiting segmentation to improve deep neural networks for object detection. In: CVPR, pp. 4703–4711 20, 21
278 Zhu Y., Zhao C., Wang J., Zhao X., Wu Y., Lu H. (2017) CoupleNet: Coupling global structure with local parts for object detection. In: ICCV 20, 21
277 Zhu Y., Zhou L., Shao L., Mellor M. (2016) DAVE: A unified framework for fast vehicle detection and annotation. In: ECCV, pp. 278–293 2
276 Zhu, Y., Ye Q., Qiu J., Jiao L. (2017) Oriented response networks. In: CVPR, pp. 4961–4970 19
275 Zhu X., Vondrick C., Fowlkes C., Ramanan D. (2016) Do we need more training data? IJCV 119(1):76–92
274 Zhu X., Tuia D., Mou L., Xia G., Zhu Y., Urtasun R., Salakhutdinov R., Fidler S. (2015) SegDeepM: Exploiting segmentation to improve deep neural networks for object detection. In: CVPR, pp. 4703–4711 20, 21
273 Zhu Y., Zhao C., Wang J., Zhao X., Wu Y., Lu H. (2017) CoupleNet: Coupling global structure with local parts for object detection. In: ICCV 20, 21
272 Zhu Y., Zhou L., Shao L., Mellor M. (2016) DAVE: A unified framework for fast vehicle detection and annotation. In: ECCV, pp. 278–293 2
Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.
表11 一般物体検出のマイルストーンである検出フレームワークの特性と性能の概要。詳細な説明は5節を参照されたい。一部のアーキテクチャは図13で図解している。backbone DCNNの特性は表6に示している。訓練データの略記はそれぞれ、“07”：VOC2007のtrainval、“07+”：VOC2007のtrainvalとtest、“12”：VOC2012のtrainval、“CO”：COCOのtrainvalを意味する。"速度"列は単一のNVIDIA Titan X GPUでの検出速度を概算した値。SS：Selective Search、RPN：Region Proposal Network。RCNN>Rは“RCNN minus R”を意味し、詳細な物体領域検出手法の使用を含んでいる。

| 検出器名 | 領域候補 | backbone DCNN | 入力画像サイズ | VOC07結果 | VOC12結果 | 速度(FPS) | 発表年 | ソースコード | 注目点と欠点 |
|----------|-----------|---------------|---------------|-----------|-----------|----------|--------|------------|-------------|
| RCNN [85] | SS        | AlexNet       | 固定          | 58.5 (07) | 53.3 (12) | < 0.1    | CVPR14 | Caffe       | MatLab      |
| 注目点：最初にCNNと物体領域検出の手法を統合、従来の先端技術から新たな性能向上。 |
| 欠点：順次訓練される多層のパイプライン（外部の物体領域検出、CNN fine-tuning、ワークイングされた物体領域検出のCNN通報、SYMとbbox回帰器の訓練）、空間計算時間計算機の両面で訓練コストが高い、テストが遅い。 |
| SPPNet [59] | SS        | ZFNet         | 任意          | 60.9 (07) | -         | < 1      | ECCV14 | Caffe       | MatLab      |
| 注目点：最初にSPPをCNNとアーキテクチャに導入。囊み込み特徴の共有を可能にする、性能を犠牲にすることなくRCNNの評価を数倍加速、OverFeatより高速。 |
| 欠点：RCNNの欠点を継承、訓練がより高速化しない、Fine-tuningでSPP層でのCONV層を更新できない。 |
| Fast RCNN [84] | SS VGG16  | AlexNet       | 任意          | 70.0 (VGG) (07+12) | 68.4 (VGG) (07+12) | < 1      | ICCV15 | Caffe       | Python      |
| 注目点：最初にend-to-end（物体領域検出生成を無視）の検出器訓練を可能にした、RoI pooling層を設計、SPPNetよりるるかに高速で正確、特徴のキャッシュ用のディスクストレージが不要。 |
| 欠点：外部の物体領域検出計算が新たなポルテックとなる、依然リアルタイムアプリケーションには選ばない。 |
| Faster RCNN [229] | RPN       | ZFNet         | 任意          | 73.2 (VGG) (07+12) | 70.4 (VGG) (07+12) | < 5      | NIPS15 | Caffe       | Python      |
| 注目点：selective searchに対決する、ほとんど追加計算コスト無しに高品質な物体領域検出を生成するRPNを提案、RPNの参照用形態として平行移動変不変マルチスケールのanchor boxを導入、CONV層を共有することでRPNとFast RCNNを単一のネットワークに統合、性能低下無しにFast RCNNより1倍高速、VGG16では5FPSでテストを実行可能。 |
| 欠点：訓練が複雑、訓練時間が不合理化していない、依然リアルタイムに選ばない。 |
| RCNN+C [151] | 新規      | ZFNet + SPP   | 任意          | 59.7 (07) | -         | < 5      | BMVC15 | -           |             |
| 注目点：selective searchを有効な物体領域検出生成方法を提案するCNNのみに依存される統合された単純で高速な検出器を構築する可能性を認める。 |
| 欠点：リアルタイム化する、面倒な物体領域検出生成により精度が低い。 |
| RFCN [50] | RPN       | ResNet101     | 任意          | 80.5 (07+12) | 77.6 (07+12) | < 10     | NIPS16 | Caffe       | Python      |
| 注目点：全層囊み込みの検出ネットワーク、特殊なCONV層のネットを使用してposition sensitive score mapのセットを設計、あまり精度を犠牲にすることなくFaster RCNNより高速。 |
| 欠点：訓練が不合理化されていない、依然リアルタイムに選ばない。 |
| Mask RCNN [102] | RPN       | ResNet101 ResNetX101 | 任意          | 50.3 (ResNetX101) (600) | 50.3 (ResNetX101) (COCOでの結果) | < 5      | ICCV17 | Caffe       | Python      |
| 注目点：単純で柔軟で効率的な物体インスタンスセグメンテーション用フレームワーク。bbox予測用の既存プロトタイプを追加することでFaster RCNNを拡張、Feature Pyramid Network (FPN)が活用される、傑出した性能。 |
| 欠点：リアルタイムアプリケーションには選ばない。 |

| 統合フレームワーク（5.2節） |
|---------------------------|
| OverFeat [239] | AlexNet like |  | 固定 | - | - | < 0.1 | ICLR14 | C++ |
| 注目点：囊み込み特徴の共有、マルチスケールの画像ピラミッドでのCNN特徴抽出、ISLVRC2013位置推定コンペティションで優勝、RCNNよりさらに高速。 |
| 欠点：順次訓練される多層パイプライン、単一bbox回帰器、同一クラスの複数の物体インスタンスに対処できず、リアルタイムアプリケーションには選ばない。 |
| YOLO [227] | GoogLeNet like | 固定 | 66.4 (07+12) | 57.9 (07+12) | < 25 | CVPR16 | DarkNet |
| 注目点：最新の効率的な統合検出器、物体領域検出生成の工程を完全に削除、エレガントで効率的な検出フレームワーク、従来の検出器よりもさらに高速。YOLOは45FPSで、Fast YOLOは155FPSで実行される。 |
| 欠点：精度は最高性能の検出器に達ばず、小さな物体の位置推定に苦労。 |
| YOLOv2 [226] | DarkNet | 固定 | 78.6 (07+12) | 73.5 (07+12) | < 50 | CVPR17 | DarkNet |
| 注目点：より高速なDarkNet19を提案、多数の既存戦略を使用して精度と速度の両方を向上、高精度と高速度を両立。YOLO90000は9000以上の物体カタログをリアルタイムで検出可能。 |
| 欠点：小さな物体の検出が苦手。 |
| SSD [175] | VGG16 | 固定 | 76.8 (07+12) | 74.9 (07+12) | 8.00 (07+12+CO) | < 60 | ECCV16 | Caffe |
| 注目点：最良の効率的な統合検出器、RPNとYOLOのアイデアを効率的に組み合わせ、マルチスケールのCONV層を検出を実行、YOLOより高速かつはるかに正確、59FPSで実行可能。 |
| 欠点：小さな物体の検出が苦手。 |