

Deep Learning for Generic Object Detection: A Survey

Li Liu Wanli Ouyang Xiaogang Wang Paul Fieguth Jie Chen Xinwang Liu Matti Pietik inen

日本語訳 第 0.1 版 (2019 年 12 月 23 日)

訳者：進矢 陽介

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かつ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果について、包括的なサービスを提供することである。本論文では、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、物体提案生成、コンテキストモデリング、訓練戦略、評価指標）をカバーする300本以上の論文のサービス結果をまとめた後、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, 置き込みニューラルネットワーク, 物体認識

1 序論

物体検出（図 1）は、コンピュータビジョンにおける根本的かつチャレンジングな問題として数十年にわたって活発に研究されてきた [76]。物体検出の目的は、（人間・車・自転車・犬・猫などの）所定カテゴリの物体インスタンスが画像内に存在するか判定し、存在する場合は各物体インスタンスの空間的な位置と範囲を（例えば bounding box を介して [68, 234]）返すことである。物体検出は画像理解とコンピュータビジョンの土台として、領域分割・シーン理解・物体追跡・画像キャプション生成・イベント検出・行動認識などの複雑または高レベルな視覚タスクを解決するための基盤を成し、また、ロボットビジョン・家電・セキュリティ・自動運転・ヒューマンコンピュータインターフェース・コンテンツベース画像検索・インテリジェントビデオ監視・拡張現実などの幅広いアプリケーションを支えている。

近年、データから特徴表現を自動的に学習するための強力な方法として登場した深層学習技術 [105, 149] が、物体検出に大幅な改善をもたらしている（図 3）。

図 2 に示すように、物体検出は特定インスタンスの検出と広範なカテゴリの検出の 2 種類に大別される [91, 310]。前者は特定の物体インスタンス（ドナルド・トランプの顔、エiffel 塔、隣家の犬など）の検出を目的とし、本質的にはマッチング問題である。後者の目的は、（人間・車・自転車・犬などの）いくつかの既定の物体カテゴリの（通常は初見の）インスタンスを検出することである。歴史的には単一または少数の特定カテゴリの物体検出に多くの労力が注がれてきた（顔検出・歩行者検出が典型）。一方ここ数年の研究コミュニティ

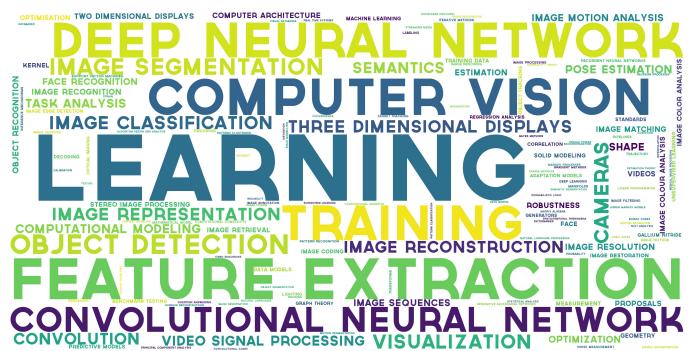


図1 2016年～2018年のICCV, CVPR論文の最頻出キーワード。各単語の大きさはそのキーワードの出現頻度に比例。object detection(物体検出)が近年大きな注目を集めていることが見て取れる。

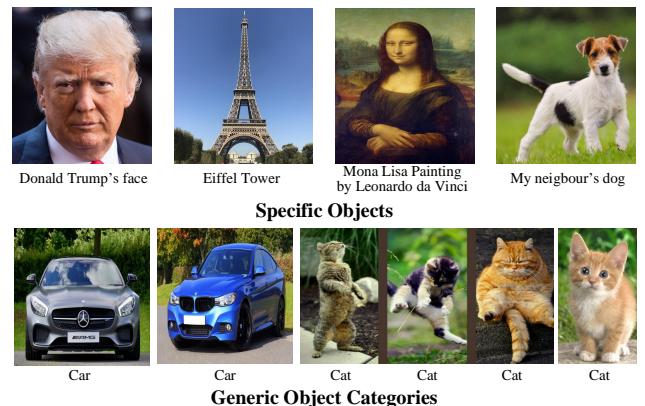


図2 物体検出には、特定の物体インスタンスの位置推定（上）と、一般の物体カテゴリの検出（下）が含まれる。本サービスは、後者の一般物体検出問題に関する近年の進歩に焦点を当てている。

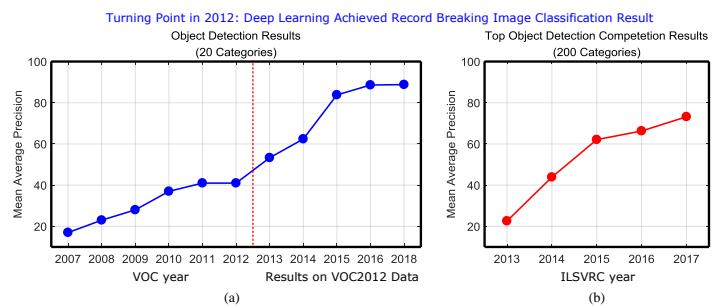


図3 近年の物体検出性能の概要：2012年のディープラーニング登場以降、性能（mean average precisionで評価）の大幅な向上が見られる。**(a)** VOC2007-2012コンペティション優勝エントリの検出性能。**(b)** ILSVRC2013-2017物体検出コンペティションの最高性能。（いずれも提供された訓練データのみ使用する条件での結果。）

Communicated by Bernt Schiele.

© The Author(s) 2019

本稿は上記著者らの “Deep Learning for Generic Object Detection: A Survey” (IJCV 2019) を日本語訳し補足修正したものである。

元論文：<https://doi.org/10.1007/s11263-019-01247-4>

元論文 URL : <https://doi.org/>
元論文ライセンス : CC BY 4.0

は、人間の幅広い物体検出能力に匹敵する一般物体検出システムを構築するという、よりチャレンジングな目標に向かい始めている。

2012 年、Krizhevsky ら [140] は AlexNet と呼ばれる深層畳

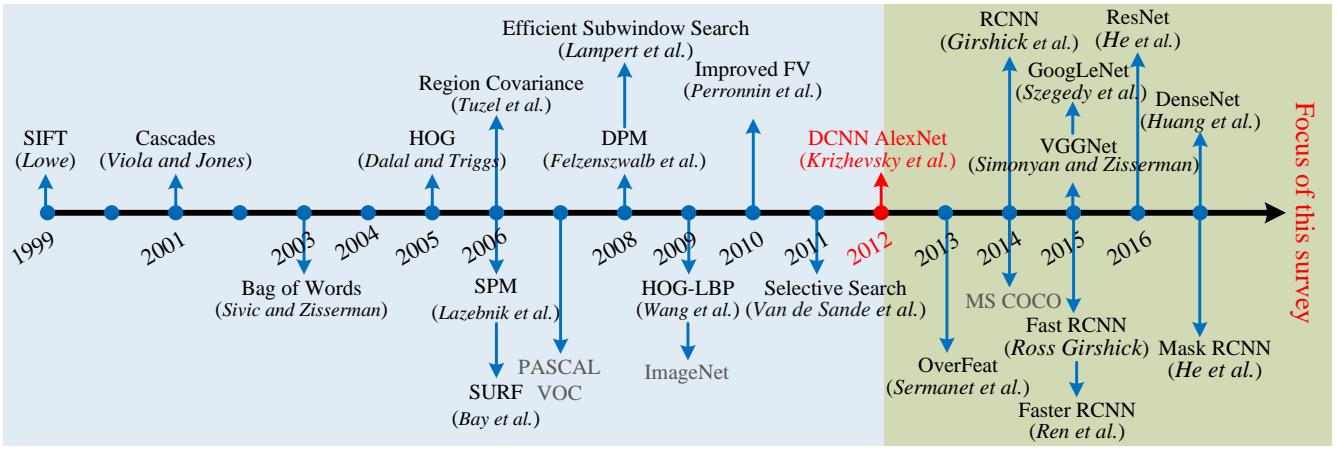


図4 物体の検出・認識のマイルストーン. 本図には、特徴表現 [47, 52, 101, 140, 147, 178, 179, 212, 248, 252, 263, 276, 279], 検出フレームワーク [74, 85, 239, 271, 276], データセット [68, 166, 234] を含めている. 2012年まではハンドクラフト特徴に支配されていたが、2012年に Krizhevsky ら [140] が開発した画像分類用 DCNN を転機に、2012年以降の手法は深層ネットワークに支配されている. 記載手法の多くは多数引用されており、ICCV や CVPR の主要な賞を受賞している. 詳細は 2.3 節を参照されたい.

み込みニューラルネットワーク (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、深層学習手法に特に重点を置いている. 図3のように自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. この非常に速い進歩速度を考慮して、本稿では現在の一般物体検出の全景をより明確に把握するため、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してきた. 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. 例外として物体クラス検出のトピックに関するサーベイを実施した Zhang ら [310] の研究が挙げられる. しかし、[310] や [91, 5] でレビューされているのは、ほとんどが 2012 年以前の、つまり近年の深層学習とその関連手法による顕著な成功と支配より前の研究である.

深層学習は非常に複雑で微細で抽象的な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著しい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] は画像・動画・音声・音響の処理にブレークスルーをもたらした. 深層学習についても多数のサーベイが発表されている（例えば、Bengio ら [13], LeCun ら [149], Litjens ら [170], Gu ら [92] や、より最近では ICCV や CVPR のチュートリアル）.

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るにも関わらず、それらについての近年の包括的なサーベイを我々は把握していない. 既存研究の徹底的なレビューと要約は、物体検出の更なる進歩のため、そして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ため、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.

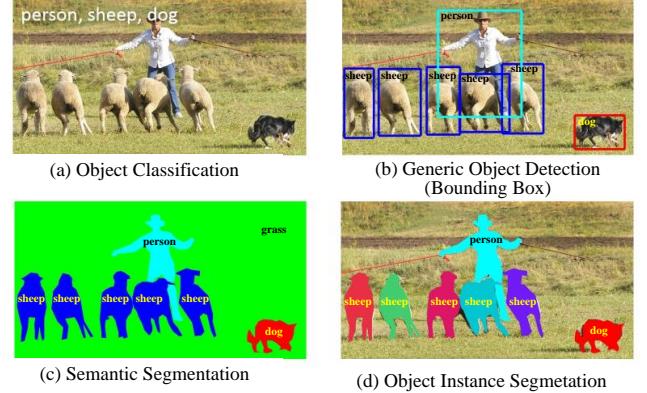


図5 一般物体検出関連の認識問題：(a) 画像レベル物体分類、(b) Bounding Box レベル一般物体検出、(c) 画素単位セマンティックセグメンテーション、(d) インスタンスレベルセマンティックセグメンテーション.

1.2 本サーベイの対象範囲

深層学習ベースの一般物体検出の論文数は息をのむほどであり、最先端技術 (state-of-the-art) の包括的レビューを編集しようとすると妥当な長さの論文の範囲を超えてしまう. そのため、論文の選択基準を定める必要があり、本サーベイではトップジャーナル論文とトップ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めら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. 本サーベイは過去5年間の静止画物体検出の主要な進歩に焦点を当てており、重要テーマの動画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、幾つかの分類法と高レベルの視点・組織化を提示することである. 主に、一般的なデータセット、評価指標、コンテキストモデリング、検出提案の手法に基づいてこれらを行う. 本稿の分類は多様な戦略の類似点・相違点の理解の助けになることを意図しており、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課題を特定するための枠組みを研究者に提供する.

以降、本論文は以下のように構成される. 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節で人気のデータセットと評価基準をま

表 1 2000 年以降の物体検出関連サービスの概要。

No. サービスタイトル	参考文献	発表年	発表先	内容
1 Monocular Pedestrian Detection: Survey and Experiments	[66]	2009	PAMI	3 種の歩行者検出器の評価
2 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ビス
3 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な評価
4 Detecting Faces in Images: A Survey	[294]	2002	PAMI	単一画像からの顔検出の最初のサービス
5 A Survey on Face Detection in the Wild: Past, Present and Future	[301]	2015	CVIU	2000 年以降の “in the wild” 顔検出のサービス
6 On Road Vehicle Detection: A Review	[258]	2006	PAMI	ビジョンベース路上車両検出システムのレビュー
7 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・認識に関するサービス
8 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
9 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
10 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
11 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
12 Visual Object Recognition	[91]	2011	指導書	物体認識（インスタンス認識・カテゴリ認識）技術
13 Object Class Detection: A Survey	[310]	2013	ACM CS	2011 年以前の一般物体検出手法のサービス
14 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
15 Salient Object Detection: A Survey	[19]	2014	arXiv	顕著物体検出のサービス
16 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、教師無しの表現学習と深層学習
17 Deep Learning	[149]	2015	Nature	深層学習とその応用の紹介
18 A Survey on Deep Learning in Medical Image Analysis	[170]	2017	MIA	医用画像解析における画像分類・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ビス
19 Recent Advances in Convolutional Neural Networks	[92]	2017	PR	CNN の近年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広範なサービス
20 Tutorial: Tools for Efficient Object Detection	—	2015	ICCV15	近年のマイルストーンのみをカバーする物体検出の短期講習
21 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研究の概略
22 Tutorial: Instance Level Recognition	—	2017	ICCV17	物体検出・インスタンスセグメンテーション・人物姿勢予測など、インスタンスレベルの認識に関する近年の進歩についての短期講習
23 Tutorial: Visual Recognition and Beyond	—	2018	CVPR18	画像分類、物体検出、インスタンスセグメンテーション、セマンティックセグメンテーションの背後にある手法と原理に関するチュートリアル
24 Deep Learning for Generic Object Detection	Ours	2019	VISI	一般物体検出のための深層学習の包括的サービス

とめる。5 節で物体検出フレームワークのマイルストーンについて述べる。6 節から 9 節では物体検出器設計に関する基本的なサブ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 [310]、次のように定義される。画像が与えられたら、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が狭い特定物体カテゴリ検出とは対照的に、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は主に、（空・草・雲などの）構造化されていないシーンではなく、（車・顔・自転車・飛行機などの）高度に構造化された物体や（人間・牛・馬などの）関節を持つ物体の位置推定（localization）に関心を向けている。

物体の空間的な位置と範囲は、図 5 のように、bounding box（ぴったりと物体を囲む各辺が座標軸と平行な矩形）で粗く、[68, 234]、または正確な画素ごとの領域分割マスク [310] や閉じた境界 [166, 235] で定義することができる。我々の知る限り、一般物体検出アルゴリズムの評価では bounding box が現在の文献で最も広く使用されており [68, 234]、本サービスでもそのアプローチを探る。ただし、研究コミュニティはより深いシーン理解に向かっている（画像レベルの物体分類 → 単一物体の位置推定 → 一般物体検出 → 画素単位の物体セグ

メンテーション）ため、将来の課題は画素レベル [166] にあると想される。

一般物体検出と密接に関連する問題は多くある^{*1}。図 5 (a) の物体分類 (object classification, object categorization) は、所定の物体クラスのセットから画像内の物体の存在を評価することを目的とする。つまり、位置は不要で、与えられた画像に一つ以上の物体クラスラベルを割り当てることで存在を判定する。検出は、画像内のインスタンスの位置も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 5 (c) のセマンティックセグメンテーション (semantic image segmentation) は、画像の各画素へのセマンティッククラスラベル割り当てを目的としており、一般物体検出と密接に関連する。図 5 (d) のインスタンスセグメンテーション (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は、高品質/高精度かつ高効率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 6）。図 7 に示すように、高品質な検出のためには、実世界の多種多様な物体カテゴリを区別できる高い識別性と、外観のクラス内変動があって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、

^{*1} 我々の知る限り、様々な視覚サブタスクの定義に関する文献間の普遍的合意は無い。detection, localization, recognition, classification, categorization, verification, identification, annotation, labeling, understandingなどの用語は、しばしば異なる定義がなされる [5]。

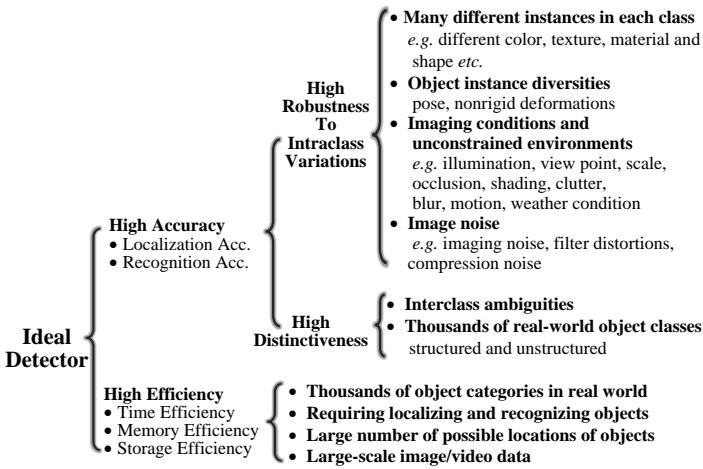


図 6 一般物体検出における課題の分類



図 7 撮像条件変動による同一クラスの外観変化 (a)-(h) . 一つの物体クラスとされるべきものには驚くべきバリエーションがある (i) . 一方で, (j) の 4 枚の画像は非常に似ているように見えるが, 実際には 4 種の異なる物体クラスの画像である. 画像の多くは ImageNet [234] または MS COCO [166] より.

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, 許容可能なメモリとストレージで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ことがあり, 図 7 (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. 人間や馬などのより狭く定義されたクラスできさえ, 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うる.

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外観に与える劇的な影響によって引き起こされる. 図 7 (a-h) に例示する照明・姿勢・スケール・遮蔽・混雑・陰影・ブラ - 動作などのように, これらの全条

件が物体の外観に大きな変化をもたらす. 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, 課題は更に加わりうる.

クラス内変動に加えて, $10^4 - 10^5$ のオーダーの多数の物体カテゴリの検出には, 図 7 (j) に示すような微妙に異なるクラス間変動を区別できる高い識別力を必要とする. 現在の検出器は主に構造化された物体のカテゴリに焦点を当てており, 実際の検出カテゴリ数はそれより少ない. 例えば, PASCAL VOC [68], ILSVRC [234], MS COCO [166] のカテゴリ数は順に 20, 200, 91 であり, 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 効率とスケーラビリティに関する課題

ソーシャルメディアネットワークとモバイル/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いる. しかし, モバイル/ウェアラブ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重要である.

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 (多い可能性がある) に応じて増大し, また, 図 7 (c, d) のように, 単一画像内には非常に多くの位置とスケールが含まれる可能性があるため更に増大する.

更にスケーラ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, それに手動でアノテーション (注釈付け) するのは不可能となり, 弱教師あり学習の戦略に頼らざるを得なくなる.

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ツベースモデル [76] に基づいており, 顔などの空間レイアウトがほぼ固定された特定物体に焦点を当てていた. 1990 年以前は, 物体認識の主要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] に基づく統計的分類器 (例: ニューラルネットワーク [233], SVM [201], Adaboost [276, 290]) へとフォーカスが移行した. 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 4 に示す. 外観特徴はグローバル表現 [192, 260, 267] から, 平行移動・スケール・回転・照明・視点・遮蔽の変化に不变であるように設計されたローカル表現へと移行した. ハンドクラフトの局所不变特徴は, Scale Invariant Feature Transform (SIFT) 特徴 [178] 以降非常に大きな人気を博し, 様々な視覚認識タスクの進歩は実質的に Haar-like 特徴 [276], SIFT [179], Shape Contexts [12], Histogram of Gradients (HOG) [52], Local Binary Patterns (LBP) [196], region covariances [268] などの局所記述子 [187] の使用に基づいていた. これらのローカル特徴は通常, 単純に結合されるか, Bag of Visual Words [252, 47], BoW モデルの Spatial Pyramid Matching (SPM) [147], Fisher Vectors [212] などの特徴プーリングエンコーダにより集約される.

DCNN [140] が画像分類で記録破りの結果を達成した 2012 年の重要な転換点まで, ハンドクラフト局所記述子と識別的分類器による手動調整の多段パイプラインは, 物体検出を含むコンピュータビジョンの様々なドメインを長年にわたり支配した.

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ができ, 使用される隠れ層はあまり多くなかつた [272, 233, 238] が, 顔検出などの制限されたドメインで

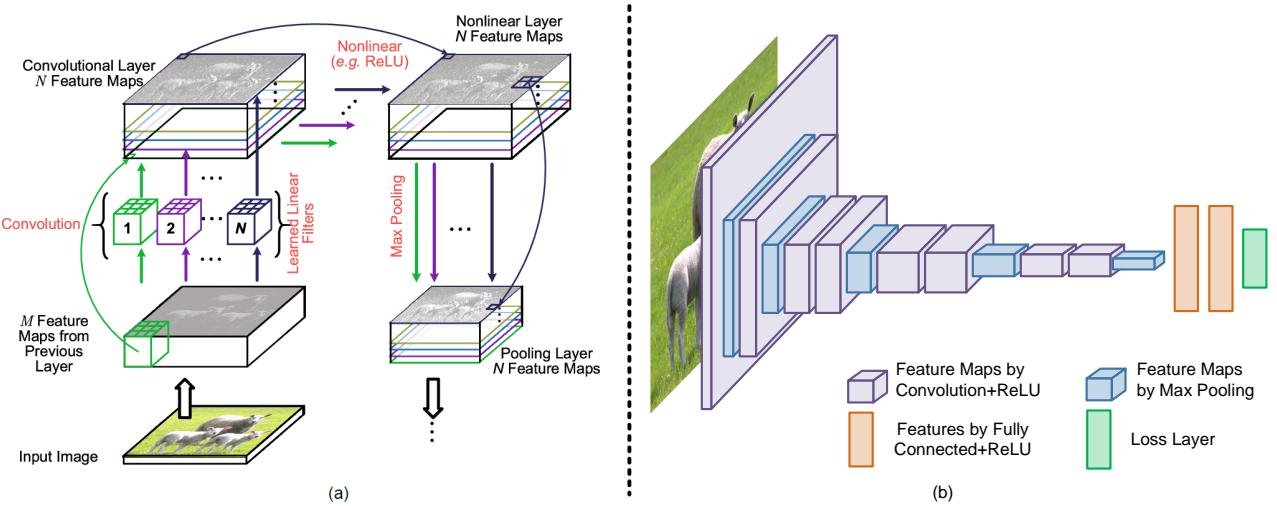


図 8 (a) 典型的な CNN で繰り返し適用される 3 種の演算：多数の線形フィルタによる畠み込み；非線形性（例：ReLU）；ローカルプーリング（例：最大プーリング）。前層からの M 枚の特徴マップは、 N 枚の異なるフィルタ（ここではサイズ $3 \times 3 \times M$ ）でストライド 1 で畠み込まれる。結果として得られた N 枚の特徴マップは、ReLU などの非線形関数を介して渡され、（例えば 2×2 の領域の最大値を取ることで）プーリングされ、解像度の減った N 枚の特徴マップを提供する。(b) VGGNet [248] のアーキテクチャ (VGG-11)。重みを持つ層が 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 つの畠み込み層、3 つの全結合層、5 つの最大プーリング層、1 つの softmax 分類層を持つ。最後 3 つの全結合層は、最後の畠み込み層の出力をベクトル化した特徴を入力として受け取る。最後の層は、クラス数 C の softmax 関数。ネットワーク全体は、ラベル付き訓練データを用い、確率的勾配降下により目的関数（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。）

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た。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) 検出器へと繋がった。

深層検出器の成功は、大量の訓練データと、数百万、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に大きく依存している。非常に高い計算能力を持つ 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になったことが、それらの成功に重要な役割を果たしている。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、より現実的で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人手が必要となる。そのため、アノテーションの難しさを緩和したり、より小さな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かい始めている。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、全体で 30,000 の視覚カテゴリを識別でき、ドメインの専門知識で区別できるカテゴリの数は 10^5 オーダーにも及ぶ [15]。過去数年で目覚ましい進歩を遂げたものの、 $10^4 - 10^5$ カテゴリで人間レベルの性能に近づく正確・頑健・効率的な検出・認識システムを設計することは、間違いなく未解決問題である。

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命をもたらした。この途方もなく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が多数ある [13, 89, 92, 149, 170, 216, 287, 297, 313, 320, 325]。これらのサーベイは異なる視点から [13, 89, 92, 149, 216, 287, 320]、または、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、深層学習技術を調査している。

最も代表的な深層学習モデルである畠み込みニューラルネットワーク (Convolutional Neural Network; CNN) は、自然界の信号に潜在する基本特性である、平行移動不变性、局所連結性、組成上の階層 [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、畠み込み

$$\mathbf{x}^{l-1} * \mathbf{w}^l \quad (1)$$

を考える。ここでは、前層である $l-1$ 番目の層からのある 1 枚の入力特徴マップ \mathbf{x}^{l-1} が、ある 1 枚の 2 次元畠み込みカーネル（フィルタや重みとも呼ばれる） \mathbf{w}^l で畠み込まれている。この畠み込みは、非線形演算を σ として次式のように一連の層で適用される。

$$\mathbf{x}_j^l = \sigma \left(\sum_{i=1}^{N^{l-1}} \mathbf{x}_i^{l-1} * \mathbf{w}_{i,j}^l + b_j^l \right). \quad (2)$$

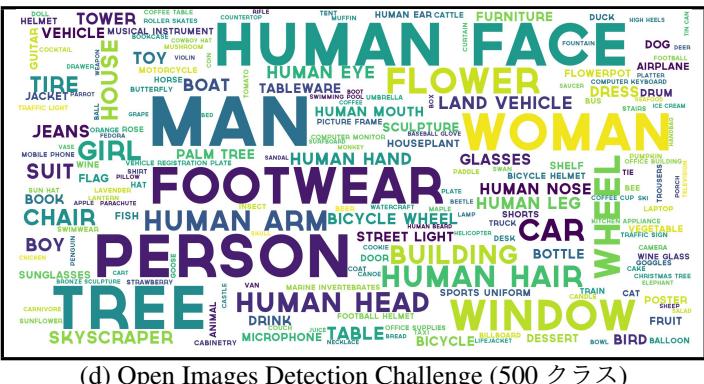
ここで、畠み込みは N^{l-1} 枚の入力特徴マップ \mathbf{x}_i^{l-1} と対応するカーネル $\mathbf{w}_{i,j}^l$ で行われ、 b_j^l はバイアス項である。要素ごとの非線形関数 $\sigma(\cdot)$ は、典型的には各要素に対する rectified linear unit (ReLU) であり、

$$\sigma(x) = \max\{x, 0\} \quad (3)$$

で計算される。最後に、プーリングは特徴マップのダウンサンプリング/アップサンプリングに対応する。これら 3 種の演算（畠み込み、非線形性、プーリング）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、各ピクセル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。畠み込み層の各ニューロンは、重みのセット $\mathbf{w}_{i,j}$ （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

表2 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。



4 データセットと性能評価

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、深層学習技術が多くの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注釈付きデータである。インターネット上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、空前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属性を表3にまとめ、選択したサンプル画像を図9に示す。大規模な注釈付きデータセットの作成には以下の3つの手順がある。まず対象物体カテゴリを決定し、次に選択したカテゴリを表す多様な候補画像セット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）収集した画像に注釈を付ける。スペースの制限のため、これらのデータセットの構築・特性に関する詳細な説明については元論文 [68, 69, 166, 234, 143] を参照されたい。

4つのデータセットは、それぞれの検出チャレンジの主要要素をなす。各チャレンジは、公開画像データセットと、アノテーション真値 (ground truth)，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4に示す^{*2}。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2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桁以上拡大する。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修正された。

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

(b) のように、CNN の前方の層（低層）は通常畳み込み層とブーリング層で構成され、後方の層（高層）は通常全結合されている。前方の層から後方の層まで入力画像は繰り返し畳み込まれ、各層で受容野（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たより一般的な特徴を抽出する [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、非常に複雑な関数を学習する能力があり、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な上、適切な訓練パラメー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、攻撃に対する深刻な脆弱性 [88] が、DCNN の実世界応用の制約となっている。

^{*2} PASCAL VOC2007以外はテストセットへのアノテーションは非公開。

表3 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	カテゴリあたりの 画像数	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	303 ~ 4087	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 チャレンジの基幹. 画像は物体が中心.
MS COCO [166]	328,000+	91	—	7.3	640 × 480	2014	実世界のシナリオに更に近い. 各画像にはより多くの物体インスタンスとより豊富な物体注釈情報が含まれる. 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注釈データを含む.
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+	6000+	—	8.3	varied	2017	画像レベルのラベル, 物体の bounding box, visual relationship (物体関係, 視覚的関係) のアノテーションがされている. 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 の物体注釈付きの画像の例. データセットの概要については表3を参照されたい.

ている. COCO 物体検出チャレンジ [166] は, bounding box 出力のタスクと物体インスタンス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, 混雑や重度の遮蔽の中にある.
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る.

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物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物体検出データセットと比べ, クラス・画像・bounding box 注釈・インスタンスセグメンテーションマスク注釈の数が大幅に増加しているだけでなく, 注釈プロセスに違いがある. 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注釈が付けられるが, Open Images V4 では各画像に分類器を適用して十分スコアの高かったラベルのみが人間の検証に回されている. そのため OICOD では, 正しいラベルだと人間が確認した物体インスタンスにのみ注釈が付けられる.

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. 最も一般的に使用される評価基準は, precision と recall から計算される Average Precision (AP) である. AP は通常, 各物体カテゴリに対して個別に計算され, カテゴリごとに評価される. 全物体カテゴリにわたる性能の比較には, 全物体カテゴリの 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^{*3}. 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 \mathbf{I} に適用された検出器の標準的な出力は, 予測検出 $\{(b_j, c_j, p_j)\}_j$ で表せる. ここで, j は予測検出の物体の index, b_j は Bounding Box (BB), c_j は予測カテゴリ, p_j は confidence である. 予測検出 (b, c, p) は, 以下の場合に True Positive (TP) と見なされる.

- 予測カテゴリ c が真値ラベル c_g と等しい.
- 予測 BB b と真値 BB b^g とのオーバーラップ比率を表す指標である IOU (Intersection Over Union) [68, 234]

$$\text{IOU}(b, b^g) = \frac{\text{area}(b \cap b^g)}{\text{area}(b \cup b^g)}, \quad (4)$$

が, 所定閾値 ε 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. ε の代表的な値としては 0.5 が用いられる.

^{*3} PASCAL VOC や ILSVRC などの物体検出チャレンジでは, 各物体カテゴリの AP スコアが最も高い参加者が各物体カテゴリの勝者となり, 最も多くの物体カテゴリで勝ったチームがチャレンジの勝者となる. mAP によるチームのランキングは勝利した物体カテゴリ数によるランキングと常に同じであったため, チーム成績の尺度として mAP を使用することは正当化される [234].

表 4 一般的に使用される物体検出データセットの統計。VOC チャレンジの物体統計には、評価で使用された ‘non-difficult’ の全注釈付き物体を掲載している。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 20K 枚の画像からなる 4 つの分割 (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り、また、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、異なる年の結果も直接比較可能なことに注意されたい。

チャレンジ	物体 クラス数	画像数			注釈付き物体数		合計 (Train+Val)		
		Train	Val	Test	Train	Val	Images	Boxes	Boxes/Image
PASCAL VOC Object Detection Challenge									
VOC07	20	2,501	2,510	4,952	6,301(7,844)	6,307(7,818)	5,011	12,608	2.5
VOC08	20	2,111	2,221	4,133	5,082(6,337)	5,281(6,347)	4,332	10,364	2.4
VOC09	20	3,473	3,581	6,650	8,505(9,760)	8,713(9,779)	7,054	17,218	2.3
VOC10	20	4,998	5,105	9,637	11,577(13,339)	11,797(13,352)	10,103	23,374	2.4
VOC11	20	5,717	5,823	10,994	13,609(15,774)	13,841(15,787)	11,540	27,450	2.4
VOC12	20	5,717	5,823	10,991	13,609(15,774)	13,841(15,787)	11,540	27,450	2.4
ILSVRC Object Detection Challenge									
ILSVRC13	200	395,909	20,121	40,152	345,854	55,502	416,030	401,356	1.0
ILSVRC14	200	456,567	20,121	40,152	478,807	55,502	476,668	534,309	1.1
ILSVRC15	200	456,567	20,121	51,294	478,807	55,502	476,668	534,309	1.1
ILSVRC16	200	456,567	20,121	60,000	478,807	55,502	476,668	534,309	1.1
ILSVRC17	200	456,567	20,121	65,500	478,807	55,502	476,668	534,309	1.1
MS COCO Object Detection Challenge									
MS COCO15	80	82,783	40,504	81,434	604,907	291,875	123,287	896,782	7.3
MS COCO16	80	82,783	40,504	81,434	604,907	291,875	123,287	896,782	7.3
MS COCO17	80	118,287	5,000	40,670	860,001	36,781	123,287	896,782	7.3
MS COCO18	80	118,287	5,000	40,670	860,001	36,781	123,287	896,782	7.3
Open Images Challenge Object Detection (OICOD) (Open Images V4 [143] に基づく統計)									
OICOD18	500	1,643,042	100,000	99,999	11,498,734	696,410	1,743,042	12,195,144	7.0

```

Input:  $\{(b_j, p_j)\}_{j=1}^M$ :  $M$  predictions for image  $\mathbf{I}$  for object class  $c$ ,  

       ranked by the confidence  $p_j$  in decreasing order;  

 $\mathcal{B} = \{b_k^g\}_{k=1}^K$ : ground truth BBs on image  $\mathbf{I}$  for object class  $c$ ;  

Output:  $\mathbf{a} \in \mathbb{R}^M$ : a binary vector indicating each  $(b_j, p_j)$  to be a TP or FP.  

Initialize  $\mathbf{a} = 0$ ;  

for  $j = 1, \dots, M$  do  

  Set  $\mathcal{A} = \emptyset$  and  $t = 0$ ;  

  foreach unmatched object  $b_k^g$  in  $\mathcal{B}$  do  

    if  $IOU(b_j, b_k^g) \geq \varepsilon$  and  $IOU(b_j, b_k^g) > t$  then  

       $\mathcal{A} = \{b_k^g\}$ ;  

       $t = IOU(b_j, b_k^g)$ ;  

    end  

  end  

  if  $\mathcal{A} \neq \emptyset$  then  

    Set  $\mathbf{a}(i) = 1$  since object prediction  $(b_j, p_j)$  is a TP;  

    Remove the matched GT box in  $\mathcal{A}$  from  $\mathcal{B}$ ,  $\mathcal{B} = \mathcal{B} - \mathcal{A}$ .  

  end  

end

```

図 10 物体検出結果と box 真値を貪欲にマッチングすることで TP・FP を決定するアルゴリズム。

それ以外の場合、False Positive (FP) と見なされる。confidence p は、予測クラスラベル c 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 c に関して、あるテスト画像 \mathbf{I}_i に対する検出器の検出結果を confidence p_{ij} で降順に並べ $\{(b_{ij}, p_{ij})\}_{j=1}^M$ で表すこととする。各検出 (b_{ij}, p_{ij}) が TP と FP のどちらであるかは、図 10 のアルゴリズム^{*4}で決定され

^{*4} 所定閾値 β に対して、画像内の同一物体に対する複数の検出の全てが正しい検出と見なされるわけではなく、最も confidence の高い検出のみが TP と見なされ、残りは FP と見なされることに注意されたい。

る。TP と FP の検出に基づいて precision $P(\beta)$ と recall $R(\beta)$ [68] は confidence 閾値 β の関数として計算される。そのため、confidence 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、そこから Average Precision (AP) [68, 234] を計算できる。

MS COCO の導入以来、bounding box 位置の精度に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 にまとめている）。例えば、単一の IoU 0.5 で計算された従来の mAP とは対照的に、 AP_{coco} は、全物体カテゴリで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% は小さく 24% は大きいため、評価基準 AP_{coco}^{small} , AP_{coco}^{medium} , AP_{coco}^{large} 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

ハンドクラフト特徴 [276, 52, 72, 98, 275] から学習された DCNN 特徴 [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ため探索空間は更に増加する。したがって、効率的かつ効果的な検

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（偽陽性的検出、誤検出、誤検知、過検出、過検知）	
β	Confidence 閾値			$P(\beta), R(\beta)$ 計算用 confidence 閾値。	
ε	IOU 閾値	VOC		典型的には 0.5 前後	
		ILSVRC		$\min(0.5, \frac{wh}{(w+10)(h+10)})$; $w \times h$ は真値 box のサイズ。	
		MS COCO		10 個の IOU 閾値 $\varepsilon \in \{0.5 : 0.05 : 0.95\}$	
$P(\beta)$	Precision			少なくとも β の confidence 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。	
$R(\beta)$	Recall			N_c 個の全物体数のうち、検出器によって少なくとも β の confidence で検出された物体数の割合。	
AP	Average Precision			confidence β を変化させて達成される様々なレベルの recall にわたって計算される。	
mAP	VOC			VOC 単一の IOU での AP を全クラスにわたって平均した値。	
	ILSVRC			修正 IOU (上記 ε 参照) での AP を全クラスにわたって平均した値。	
mAP	mean Average Precision	MS COCO		<ul style="list-style-type: none"> • AP_{coco}: 10 個の IOU: $\{0.5 : 0.05 : 0.95\}$ での mAP を平均した値 • $AP_{coco}^{IOU=0.5}$: $IOU=0.5$ での mAP (PASCAL VOC metric); • $AP_{coco}^{IOU=0.75}$: $IOU=0.75$ での mAP (strict metric); • AP_{coco}^{small}: 面積が 32^2 より小さい物体用の mAP • AP_{coco}^{medium}: 面積が 32^2 と 96^2 の間の物体用の mAP • AP_{coco}^{large}: 面積が 96^2 より大きい物体用の mAP 	
AR	Average Recall	MS COCO		画像あたり所定数の検出が許容される場合の 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		<ul style="list-style-type: none"> • $AP_{coco}^{max=1}$: 画像あたり 1 個の検出が許容される場合の AR • $AP_{coco}^{max=10}$: 画像あたり 10 個の検出が許容される場合の AR • $AP_{coco}^{max=100}$: 画像あたり 100 個の検出が許容される場合の AR • AP_{coco}^{small}: 面積が 32^2 より小さい物体用の AR • AP_{coco}^{medium}: 面積が 32^2 と 96^2 の間の物体用の AR • AP_{coco}^{large}: 面積が 96^2 より大きい物体用の AR 	

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、カスクード化、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プローチは、以下の 2 つの主要カテゴリに分類される。

- 物体提案生成のための前処理ステップを含む、2 段階の検出フレームワーク。
- 検出提案処理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（領域提案不要のフレームワーク）。

6 節から 9 節では、DCNN 特徴、検出提案、コンテキストモーデリングを含む、検出フレームワークに関わる基本的なサブ問題についてより詳細に論じる。

5.1 領域ベース（2 段階）フレームワーク

領域ベースフレームワークでは、カテゴリ非依存の領域提案⁵を画像から生成し、それらの領域から CNN [140] 特徴を抽出した後、カテゴリ特化の分類器を使用して提案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 [261], 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 らは一般物体検出用 CNN の探求の最初期に、AlexNet [140] と selective search による領域提案 [271] を統合した RCNN を開発した [85, 87]。これは、CNN によって得られた画期的な画像分類結果と、ハンドクラフト特徴での領域提案における selective search の成功 [271] に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パイプラインで構成される。

- 領域提案の計算：selective search [271] によりクラス非依存の領域提案（物体を含む可能性がある候補領域）を

⁵ 物体提案（領域提案や検出提案とも呼ばれる）は、物体を含みうる候補である、画像内の領域または bounding box のセットである。[27, 110]

得る。

- CNN モデルの finetuning :** 領域提案は、画像からクロップされ同一サイズにワープされてから、ImageNet などの大規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。この段階では真値 box とのオーバーラップが IOU ⁶ 0.5 以上である全領域提案が、真値 box 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。
- クラス特化 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化の線形 SVM 分類器のセットで、fine-tuning によって学習された softmax 分類器を置き換える。SVM 分類器の訓練では、各クラスの真値 box で正例が定義される。あるクラスの全真値インスタンスとのオーバーラップが IOU 0.3 未満の領域提案が、そのクラスに対する負例とされる。SVM 分類器の訓練用に定義された正例・負例は、CNN の fine-tuning 用のものとは異なることに注意されたい。
- クラス特化 bounding box 回帰器の訓練 :** 各物体クラス用に CNN 特徴で bounding box 回帰を学習する。

高品質の物体検出を達成したものの、RCNN は以下の顕著な欠点を持つ [84]。

- 訓練が多段階パイプラインであり遅く、各段階を個別に訓練する必要があるため最適化困難である。
- SVM 分類器と bounding box 回帰器の訓練において、各画像の各物体提案から CNN 特徴を抽出する必要があるため、ディスク容量と時間の両方でコストがかかる。これは大規模な検出で大きな課題となり、VGG16 [248] 等の非常に深いネットワークを用いる場合は特に問題である。
- 計算を共有せず、各テスト画像の物体提案ごとに CNN 特徴が抽出されるため、テストが遅い。

これらの全欠点は後続の技術革新の動機となり、後述する SPPNet, Fast RCNN, Faster RCNN などの多くの改良された検出フレームワークにつながっている。

SPPNet [99]: ワーピングした領域提案からの CNN 特徴抽出を画像ごとに数千回必要とすることが、テスト時の RCNN 検出パイプラインの主なボトルネックである。畳み込み層は任意のサイズの入力を受け取れるため、CNN において固定サイズの画像が必要となるのは全結合 (FC) 層が原因である。そこで He ら [99] は、CNN アーキテクチャに伝統的な spatial pyramid pooling (SPP) [90, 147] を取り入れ、FC 層用の固定長の特徴を得るために最後の畳み込み (CONV) 層の上に SPP 層を追加した。この SPPNet により、テスト画像全体に一度だけ畳み込み層の計算を実行すれば、任意のサイズの領域提案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 の評価を数桁加速するが、検出器の訓練には同等の高速化をもたらさない。また、SPPNet での fine-tuning [99] は SPP 層までの畳み込み層を更新できず、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・bounding box 回帰器を個別に訓練するのではなく、softmax 分類器とクラス依存の bounding box 回帰を同時に学習する合理化された訓練プロセスにより、end-to-end の検出器訓練を可能にする。Fast RCNN は、領域提案全体で畳み込みの計算を共有するという

⁶ IOU の定義については 4.2 節を参照されたい。

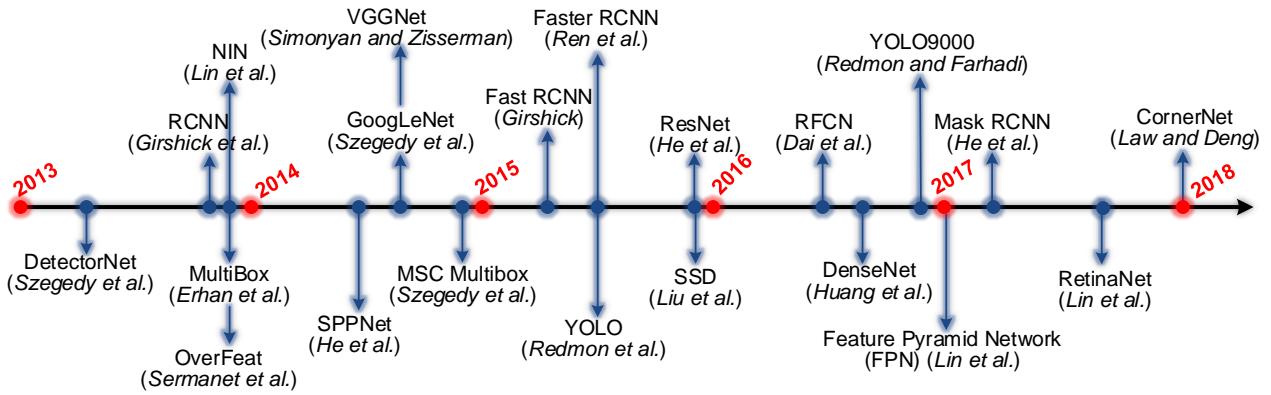


図 11 一般物体検出のマイルストーン.

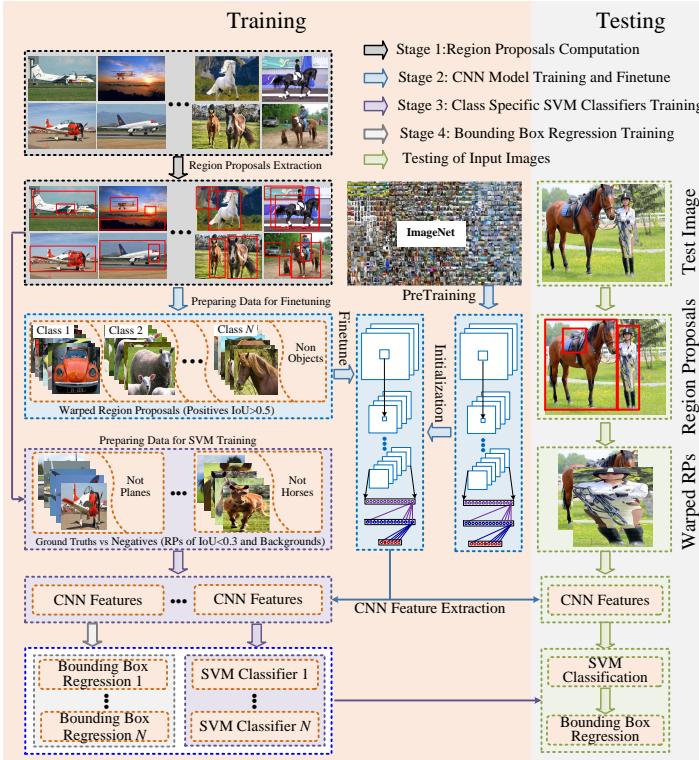


図 12 RCNN 検出フレームワーク [85, 87] の図解.

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 層を追加して各領域提案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ピングを使用して画像レベルでのワーピングを近似するものである。RoI pooling 層後の特徴は一連の FC 層に送られ、その後で 2 つの兄弟出力層に分岐し、物体カテゴリ予測用の softmax 確率と、提案精緻化用のクラス依存の bounding box 回帰オフセットを出力する。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。したがって、検出品質はより高く、単一の訓練プロセスでネットワークの全層を更新でき、特徴のキャッシング用のストレージが不要となる。

Faster RCNN [229, 230]: Fast RCNN は検出処理を大きく高速化したが、外部の領域提案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ることが示されている。したがって、領域提案生成時の selective search を CNN で置き換えることができる。Ren ら [229, 230] により提案された Faster RCNN フレーム

ワークは、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、領域提案用 RPN と領域分類用 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畠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 box (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 は低次元ベクトルにマッピングされ、2 つの兄弟 FC 層 (物体カテゴリ分類層と box 回帰層) に供給される。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 と同じ形状であるため、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、領域提案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) [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 でテストできる上、画像あたり 300 個の提案を使用して PASCAL VOC 2007 で最先端の物体検出精度を達成する。最初の Faster RCNN [229] にはいくつかの交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 and Vedaldi [151] は、selective search などの領域提案生成手法の役割に疑惑を抱き、CNN ベースの検出器での領域提案生成の役割を研究し、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発見した。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 などの領域提案生成手法を排除する可能性を示した。

RFN (Region based Fully Convolutional Network): Faster RCNN は Fast RCNN よりも桁違いに高速だが、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。この事実は、Dai ら [50] の提案した、ほぼ全ての計算が画像全体で共有される *fully convolutional* (隠れ FC 層無し) の RFN 検出器につながった。図 13 に示すように、RFN 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で CONV 層を使用し、予測出力の直前にある最終 CONV 層の特徴から RoI クロップを取得することを提案した。しかし、この単純な設計は検出精度がかなり劣ることが判明した [50]。これは、深い CONV 層ほどカテゴリの意味に敏

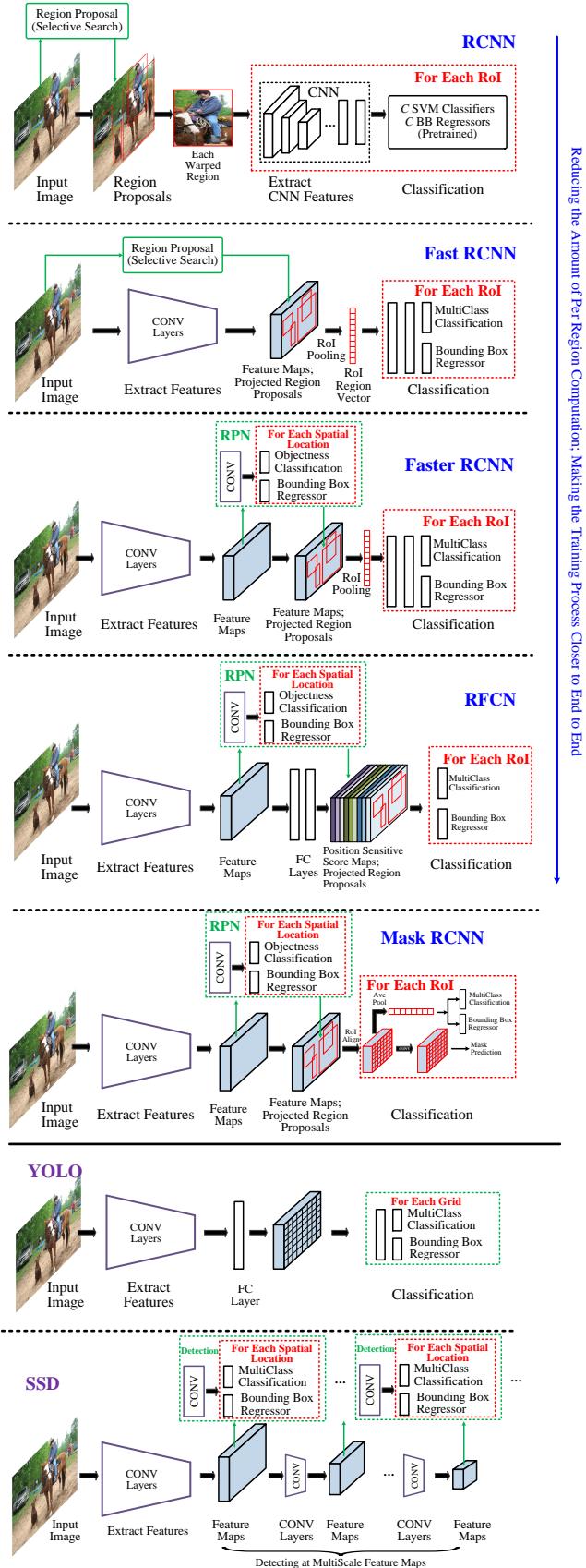


図 13 一般物体検出の主要なフレームワークの高レベルの図。これらの手法の特性は表 11 にまとめている。(訳注: RFCN の “FC Layers” は “CONV Layers” の誤植だろう。)

ing 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時間で達成できることが示されている。

Mask RCNN: He ら [102] は、画素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 段階パイプラインを採用し第 1 段階で RPN を使用するのは Faster RCNN 同様だが、Mask RCNN は第 2 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ブランチを追加する。新しいブランチは CNN 特徴マップの上に追加される Fully Convolutional Network (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、画素レベルの空間的対応を維持するために RoIAxis 層が提案された。Mask RCNN は ResNeXt101-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 bounding box 物体検出で最高精度を達成した。Mask RCNN は訓練が単純であり、よく汎化する上、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: カスケード [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い段階で多数の簡単な負例を廃棄することで、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ケ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 つ以上のカスケ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。より最近では、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、そのような学習を用いて 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 計算を削減する Light Head RCNN を提案した。具体的には、チャンネル数の小さな薄い特徴マップ（例えば COCO では 490 チャンネル）を生成する畳み込みと安価な 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パイプラインの戦略が支配的だったため、人気のベンチマークデータセット上でリードする成績を収めた検出器は全て 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、複雑な領域ベースパイプライン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。

統合パイプラインは、領域提案生成や後段の分類/特徴リサンプリングを行わない单一のフィードフォワード CNN で、クラスの確率と bounding box 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、検出性能に直接基づいて end-to-end で最適化できる。

DetectorNet: Szegedy ら [261] の DetectorNet は、最初期に探求された物体検出用 CNN の一つであり、物体検出を物体 bounding box マスクへの回帰問題として定式化した。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上の前景ピクセルを予測し、4 つの追加ネットワークを使用して物体の上半分・下半分・左半

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と推測される。Dai ら [50] はこの観察に基づいて、特殊な CONV 層のセットを FC 出力として使用して position-sensitive score map のセットを構築し、その上に position-sensitive RoI pool-

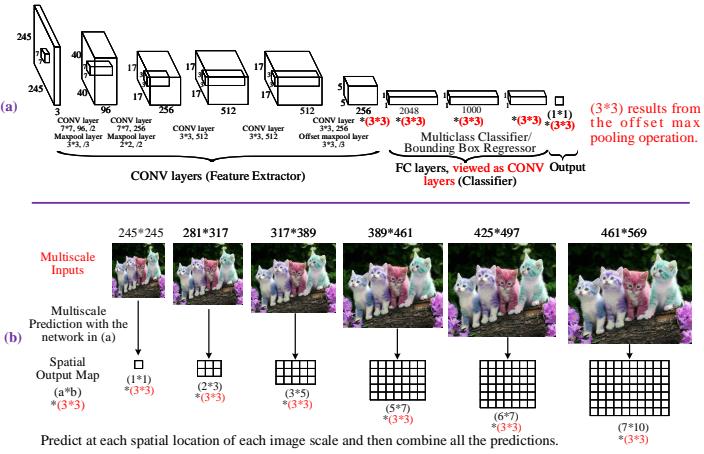


図 14 OverFeat [239] 検出フレームワークの図解。

分・右半分を予測する。次に、グループ化プロセスで予測マスクを検出 bounding box 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する必要があり、複数クラスには拡張されない。DetectorNet は画像クロップを多数行い、全クロップ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速である。

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 つと見なせる。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勝利した。OverFeat はネットワーク内の完全な畠み込み層（図 14(a) の “Feature Extractor”）を通る单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手順を以下にまとめる。

- 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、全結合層のために固定サイズの入力画像を必要とし得る AlexNet [140] 等の 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、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 のカーネルを持つ畠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 6 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、分類器は予測（クラスと confidence）のグリッドを出力する。
- オフセット最大プーリングによる予測数の増加。解像度増加のため、OverFeat は最後の CONV 層後にオフセット最大プーリングを適用する。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。
- bounding box* 回帰。物体が認識されると单一の *bounding box* 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。
- 予測の結合。OverFeat は貪欲なマージ戦略を使用し、全ての位置とスケールにわたる個々の *bounding box* 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 内の重なり合うウィンドウ間で畠み込みの計算を共有することに由来する。OverFeat は、分類器と回帰器が順次訓練される点以外は、YOLO [227] や SSD [175] 等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) 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 *bounding box* とそれに紐付くクラス確率を画素から回帰する問題に落とし込む。領域提案生成の段階が完全に削除されているため、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^{*7}。局所領域の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グローバルに使用する。YOLO は画像を $S \times S$ 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 *bounding box* 位置、confidence スコアを予測する。領域提案生成ステップ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 *bounding box* の位置・スケール・アスペクト比の分割が粗いため、Fast RCNN よりも多くの位置推定エラーを起こす。[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。これはおそらく粗いグリッド分割のためであり、また、各グリッドセルには 1 つの物体しか含められないためである。画像ごとに多数の物体を含む MS COCO などのデータセットで、YOLO がどの程度良い性能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] ベースのカスタム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) [100] が追加され、全結合層が除去され、 k 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^{*8} が使用された。YOLOv2 は標準の検出タスクで最先端を達成した。Redmon and Farhadi [226] は、9000 以上 (over 9000^{*9}) の物体カテゴリをリアルタイムで検出できる YOLO9000 も導入した。そのために、WordTree を使用した複数ソースからのデータ結合により、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同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、弱教師あり検出、つまり、*bounding box* 注釈の無い物体クラスの検出ができる。

SSD: 検出精度をあまり犠牲にすることなくリアルタイムの速度を維持するために、Liu ら [175] は、YOLO [227] より早く 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 *bounding box* 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 fully convolutional であり、早期の層（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 CONV 層が続

^{*7} YOLO の使用する *bounding box* は、Selective Search の約 2000 個と比べるに少なく、画像あたり 98 個しか使用されない。

^{*8} 物体候補となる様々なサイズ・アスペクト比の box。

^{*9} 訳注：ドラゴンボールのベジータの台詞「8000 以上だ…！」の英語ローカライズ版台詞 “It's Over 9000!” に由来する数量表現。

く。最終層の情報は正確な位置推定を行うには空間的に粗すぎる可能性があるため、適切なサイズの bounding box に対するオフセットとカテゴリスコアを複数の CONV 特徴マップで予測することで、SSD はマルチスケールでの検出を行う。解像度 300×300 の入力を用いた VOC2007 test での評価において、Faster RCNN が 7 FPS で mAP 73.2%，YOLO が 45 FPS で mAP 63.4% なのに対し、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・負例間に大きな不均衡を引き起こし、訓練を遅くし、余分なハイパーパラメ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] は、多人数姿勢推定の Associative Embedding [195] からアイディアを借り、左上・右下の一対のキーポイントの検出^{*10}として bounding box 物体検出を定式化することで CornerNet を提案した。CornerNet では、backbone ネットワークは 2 つの stacked Hourglass network [194] からなり、コーナーのより良い位置推定のために単純な corner pooling アプローチを用いる。CornerNet は MS COCO で 42.1% AP を達成し、以前のすべての 1 段階検出器より高性能である。しかし、平均推論時間は Titan X GPU で約 4 FPS であり、SSD [175], YOLO [227] より著しく遅い。CornerNet は、どのキーポイントのペアを同じ物体にグループ化するかの決定が困難なため不正確な bounding box を生成する。CornerNet を更に改善するために、Duan ら [62] は提案の中に 1 つの追加キーポイントを導入し、キーポイントの三つ組として各物体を検出する CenterNet を提案した。これにより MS COCO AP は 47.0% まで上がったが、推論速度は CornerNet よりも遅い。

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、ローカル記述子の設計（例：SIFT [178], HOG [52]）や、識別的な部分を出現させるためにより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、多大な努力が費やされた。しかし、これ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。

一方深層学習手法（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]。伝統的な特徴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雑な手順への依存が軽減されたため [13, 149]、特徴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割かれるようになっている。

5 節でレビュ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持続的に向上させてきた。CNN アーキテクチャ (6.1 節、表 15) がそのための重要な役割を果たしていると一般的に受け入れられている。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気の 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・視点・パーツ変形の幾何変動に対応するための不变特徴の開発や、広範囲のスケールで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

^{*10} 物体検出のためにキーポイントを使用するというアイディアは、DeNet [269] で既出。

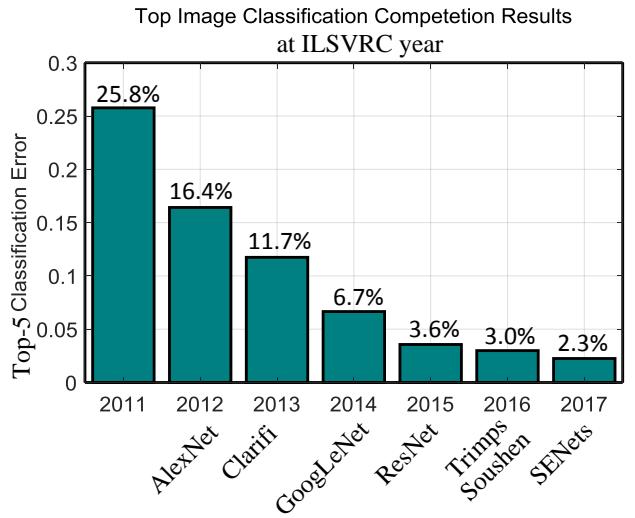


図 15 ILSVRC コンペティション画像分類タスクにおける、2011 年から 2017 年までの優勝エントリの性能。

た取り組みをレビューする。

6.1 人気の 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、時間経過にともなう性能改善を図 15 に示す。近年の CNN の進歩に関する更なるレビューは [92] を参照されたい。

アーキテクチャの進化では深化がトレンドである。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 GoogLeNet [263] で示されている。表 6 から分かるように、AlexNet, OverFeat, ZFNet, VGGNet などのネットワークは、層数は少ないにもかかわらず、FC 層のパラメータが大部分を占めるため膨大な数のパラメータを持つ。Inception, ResNet, DenseNet などの新しいネットワークは非常に深いものの、FC 層の使用を回避しておりパラメータ数ははるかに少ない。

注意深くトポロジーを設計された Inception モジュール [263] の使用により、GoogLeNet のパラメータ数は AlexNet, ZFNet, VGGNet と比べ劇的に低減される。ResNet は、数百層の非常に深いネットワークを学習するためのスキップ接続の有効性を実証し、ILSVRC 2015 分類タスクで優勝した。ResNet [101] に触発された InceptionResNet [265] は、ショートカット接続がネットワークの訓練を著しく高速化できることに基づき、Inception ネットワークにショートカット接続を組み合わせた。Huang ら [118] は ResNet を拡張し DenseNet を提案した。DenseNet は、各層が他の全ての層にフィードフォワード方式で接続する dense block から構築されており、パラメータ効率、implicit deep supervision^{*11}、特徴再利用などの説得力のある利点をもたらす。最近 Hu ら [115] は、畳み込み特徴のチャンネル間の相互依存関係を明示的にモデル化することで、チャンネルごとの特徴の応答を適応的に再較正する Squeeze and Excitation (SE) block を提案した。SE block は、既存の深いアーキテクチャに組み合わせることで最小限の追加計算コストで性能を向上でき、ILSVRC 2017 分類タ

^{*11} DenseNet は暗黙的な方法で deep supervision を実行する。すなわち、個々の層は他の層から短い接続を介して追加の supervision を受け取る。deep supervision の利点は、Deeply Supervised Net (DSN) [150] で先に実証されている。

表 6 一般物体検出に一般的に使用された DCNN アーキテクチャ。「パラメータ数」・「層数」の統計については、最後の FC 予測層は考慮されていない。「テストエラー」列は、ImageNet1000 分類の Top 5 テストエラーを示している。どのアーキテクチャを指しているか曖昧な場合、「パラメータ数」・「層数」・「テストエラー」は OverFeat (accurate model), VGGNet16, ResNet101, DenseNet201 (Growth Rate 32, DenseNet-BC), ResNeXt50 (32*4d), SE ResNet50 について言及している。

No.	DCNN アーキテクチャ	パラメータ数 ($\times 10^6$)	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]	ImageNet 分類に有効とわかった最初の DCNN。ハンドクラフト特徴から CNN への歴史的転換点。ILSVRC2012 画像分類コンペティションで優勝。
2	ZFNet (fast) [303]	58	5 + 2	14.8%	[99]	AlexNet に似ているが、畳み込みのストライド、フィルタサイズ、一部の層のフィルタ数が異なる。
3	OverFeat [239]	140	6 + 2	13.6%	[239]	AlexNet に似ているが、畳み込みのストライド、フィルタサイズ、一部の層のフィルタ数が異なる。
4	VGGNet [248]	134	13 + 2	6.8%	[84]	3 × 3 の畳み込みフィルタを積むことでネットワークの深さ（層数）を大幅に増加。また、ネットワークの深さを段階的に増加。
5	GoogLeNet [263]	6	22	6.7%	[263]	Inception モジュールを使用。このモジュールは、畳み込み層のフィルタサイズが異なる複数のブランチを使用し、これらのブランチによって生成された特徴マップを連結する。ボトルネック構造と global average pooling を最初に取り入れた。
6	Inception v2 [125]	12	31	4.8%	[112]	バッチ正規化 (Batch Normalization) の導入による高速な訓練。
7	Inception v3 [264]	22	47	3.6%		separable convolution と空間解像度削減を取り入れた。
8	YOLONet [227]	64	24 + 1	—	[227]	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% (ResNets)	[101]	identity mapping によりかなり深いネットワークを学習できる。
10	ResNet101 [101]	42	100		[101]	GoogLeNet で導入された global average pooling とボトルネックを使用することで、必要なパラメータ数が VGG より少ない。
11	InceptionResNet v1 [265]	21	87	3.1% (Ensemble)		identity mapping と Inception モジュールの組み合わせ。Inception v3 同様の計算コストだが訓練プロセスが高速。
12	InceptionResNet v2 [265]	30	95		[120]	認識性能が大幅に向上した、より計算コストの高い residual connection 付き Inception。
13	Inception v4 [265]	41	75			Inception の亜種で residual connection 無し。InceptionResNet v2 とほぼ同等の認識性能だが著しく遅い。
14	ResNeXt [291]	23	49	3.0%	[291]	同一トポロジを持つ変換のセットを集約する building block ((アーキテクチャの構築に使用される) ブロック、積み木、構成要素) を繰り返し使用。
15	DenseNet201 [118]	18	200	—	[321]	フィードフォワード方式で各層と他の全ての層を連結。勾配消失問題を軽減し、特徴再利用を促し、パラメータの数を削減。
16	DarkNet [226]	20	19	—	[226]	VGGNet に似ているがパラメータ数が大幅に少ない。
17	MobileNet [112]	3.2	27 + 1	—	[112]	depth-wise separable convolution を使用した軽量な DCNN。
18	SE ResNet [115]	26	50	2.3% (SENets)	[115]	Squeeze-and-Excitation block という新規ブロックによる channel-wise attention。既存のバックボーン CNN と相補的。

スクでの優勝に導いた。CNN アーキテクチャに関する研究は依然活発であり、Hourglass [146], Dilated Residual Network [299], Xception [45], DetNet [164], Dual Path Network (DPN) [37], FishNet [257], GLoRe [38]などのネットワークが登場している。

CNN の訓練にはクラス内の多様性を含むラベル付きの大規模データセットが必要である。画像分類と異なり、検出では画像内の（場合によっては多数の）物体の位置推定をする必要がある。画像レベルの注釈のみではなく物体レベルの注釈がある (ImageNet などの) 大規模データセットで深層モデルを事前学習すると、検出性能が向上することが示されている [206]。しかし、特に数十万カテゴリの場合 bounding box ラベルの収集コストは高い。そのため、画像レベルのラベルを持つ（通常は多数の視覚的カテゴリからなる）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] として小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^{*12}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 万枚の画像からなる ImageNet1000 [54, 234], ImageNet1000 より大きいがクラス数は少ない Places [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・検出のために探求され、どの層から抽出した特徴かで検出精度が異なることが示された。例えば、ImageNet で事前学習された AlexNet の場合、検出精度は FC6 / FC7 / Pool5 の順に低くなる [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 より FC6 / FC7 ではるかに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 特徴は人間行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

^{*12} 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セット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。

表 7 一般物体検出用に DCNN 特徴表現を改善する代表的手法の特性の概要. グループ (1), (2), (3) の詳細は 6.2 節で提示されている. 略語: Selective Search (SS), Edge Boxes (EB), InceptionResNet (IRN). Conv-Deconv は, 標準的なバックボーンネットワークを補完するために, アップサンプリング, 署み込み層, lateral connection を使用することを指す. VOC07, VOC12, COCO での mAP@IoU=0.5 による検出結果に加え, “mAP” 列に AP_{coco} (0.5 から 0.95 までの IoU 閾値に対する mAP の平均) による COCO での検出結果を示す. 訓練データの略記はそれぞれ, “07”: VOC2007 trainval, “07T”: VOC2007 trainval and test, “12”: VOC2012 trainval, “CO”: COCO trainval を意味する. COCO での検出結果は, COCO2015 Test-Standard で報告した MPN [302] を除き, COCO2015 Test-Dev で報告された.

グループ	検出器名	領域提案	Backbone	使用 DCNN バイオペライン	mAP@IoU=0.5			mAP	発表先	注目点
					VOC07	VOC12	COCO			
(1) 複数層の特徴を使用し 单一層で検出	ION [11]	SS+EB MCG+RPN	VGG16	Fast RCNN	79.4 (07+12)	76.4 (07+12)	55.7	33.1	CVPR16	複数層からの特徴を使用. コンテキスト情報モデル化のため空間的リカレントニューラルネットワークを使用. 2015 年の COCO 検出チャレンジの最優秀学生エントリーで総合 3 位.
	HyperNet [135]	RPN	VGG16	Faster RCNN	76.3 (07+12)	71.4 (07T+12)	—	—	CVPR16	領域提案と領域分類の両方で複数層からの特徴を使用.
	PVANet [132]	RPN	PVANet	Faster RCNN	84.9 (07+12+CO)	84.2 (07T+12+CO)	—	—	NIPS16	深いが軽量. concatenated ReLU [240], Inception [263], HyperNet [135] のアイディアを併用.
(2) 複数層での検出	SDP+CRC [293]	EB	VGG16	Fast RCNN	69.4 (07)	—	—	—	CVPR16	複数層の特徴を使用し CRC (cascaded rejection classifier) により easy negative を棄却した後, 残った提案を SDP (scale-dependent pooling) を使用して分類.
	MSCNN [24]	RPN	VGG	Faster RCNN	KITTI, Caltech でテスト			ECCV16	領域提案と分類を複数層で実行. 特徴のアップサンプリングを含む end-to-end 学習.	
	MPN [302]	SharpMask [214]	VGG16	Fast RCNN	—	—	51.9	33.2	BMVC16	様々な畳み込み層からの特徴と様々なコンテキスト領域の特徴を連結. 複数のオーバーラップ閾値用の損失関数. COCO15 の検出とセグメンテーションの両チャレンジで 2 位.
	DSOD [242]	Free	DenseNet	SSD	77.7 (07+12)	72.2 (07T+12)	47.3	29.3	ICCV17	DenseNet のように特徴を順次連結. 事前学習せずターゲットデータセットでスクラッチ学習.
	RFBNet [173]	Free	VGG16	SSD	82.2 (07+12)	81.2 (07T+12)	55.7	34.4	ECCV18	Inception [263] に似ているが dilated convolution を使用するマルチランチの畳み込みブロックを提案.
(3) 上記(1), (2)の組み合わせ	DSSD [77]	Free	ResNet101	SSD	81.5 (07+12)	80.0 (07T+12)	53.3	33.2	arXiv17	図 17 (c1, c2) に示すように Conv-Deconv を使用.
	FPN [167]	RPN	ResNet101	Faster RCNN	—	—	59.1	36.2	CVPR17	図 17 (a1, a2) に示すように Conv-Deconv を使用. 検出器で幅広く使用される.
	TDM [247]	RPN	ResNet101 VGG16	Faster RCNN	—	—	57.7	36.8	arXiv16	図 17 (b2) に示すように Conv-Deconv を使用.
	RON [136]	RPN	VGG16	Faster RCNN	81.3 (07+12+CO)	80.7 (07T+12+CO)	49.5	27.4	CVPR17	図 17 (d2) に示すように Conv-Deconv を使用. 物体の探索空間を大幅に削減するため objectness prior を追加.
	ZIP [156]	RPN	Inceptionv2	Faster RCNN	79.8 (07+12)	—	—	—	IJCIV18	図 17 (f1) に示すように Conv-Deconv を使用. 様々な層からの特徴用に map attention decision (MAD) unit を提案.
	STDN [321]	Free	DenseNet169	SSD	80.9 (07+12)	—	51.0	31.8	CVPR18	様々なスケールの特徴を同一スケールに並行してリサイズする新しい scale-transfer module.
	RefineDet [308]	RPN	VGG16 ResNet101	Faster RCNN	83.8 (07+12)	83.5 (07T+12)	62.9	41.8	CVPR18	より良くより少ない anchor を得るためカスクードを使用. 特徴の改善のため図 17 (e2) に示すように Conv-Deconv を使用.
	PANet [174]	RPN	ResNeXt101 +FPN	Mask RCNN	—	—	67.2	47.4	CVPR18	図 17 (g) に示す. FPN をベースに別のボトムアップパスを追加し, 下層と最上層間で情報を受け渡す. adaptive feature pooling. COCO 2017 のタスクで 1 位と 2 位.
	DetNet [164]	RPN	DetNet59+FPN	Faster RCNN	—	—	61.7	40.2	ECCV18	深い層で高解像度を維持するため, ResNet バックボーンに dilated convolution を導入. 図 17 (i) に示す.
	FPR [137]	—	VGG16 ResNet101	SSD	82.4 (07+12)	81.1 (07T+12)	54.3	34.6	ECCV18	様々な空間的位置・スケールにわたるタスク指向の特徴を, グローバルかつローカルに融合. 図 17 (h) に示す.
(4) 幾何学的変換のモデル化	M2Det [315]	—	SSD	VGG16 ResNet101	—	—	64.6	44.2	AAAI19	図 17 (j) に示す. 一連のマルチレベル (複数層) の特徴を学習するために新たに設計されたトップダウンパスが, 物体検出用の特徴ピラミッドを構築するために再結合される.
	DeepIDNet [203]	SS+ EB	AlexNet ZFNet OverFeat GoogLeNet	RCNN	69.0 (07)	—	—	25.6	CVPR15	既存の DCNN の畳み込み層と共同で学習される, 变形制約付きブーリング (deformation constrained pooling) 層を導入. end-to-end で訓練されない以下のモジュールを使用. カスクード, コンテキストモデルリング, モデル平均化 (model averaging), マルチステージ検出バイオペラインでの bounding box の位置調整.
	DCN [51]	RPN	ResNet101 IRN	RFCN	82.6 (07+12)	—	58.0	37.5	CVPR17	既存の DCNN にある通常の畳み込みを置き換える可能な deformable convolution と deformable ROI pooling モジュールを設計.
	DPFCN [188]	AttractioNet [83]	ResNet	RFCN	83.3 (07+12)	81.2 (07T+12)	59.1	39.1	IJCIV18	物体提案周辺の識別的な領域を明示的に選択するため, deformable part based ROI pooling 層を設計.

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って物体を検出する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である (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力するが, 推論時間とメモリの面では明らかに不利である.

6.2.1 物体スケール変動への対処

CNN は層ごとにその特徴階層を算出するため, 特徴階層中のサブサンプリング層が既に内在するマルチスケールピラミッドになっており, 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, 高層は

大きな受容野と強力な意味情報を持ち, 物体姿勢・照明・パース変形などの変動に最も頑健だが, 解像度が低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持つが, 解像度が高く意味に対する感度ははるかに低い. 直観的には, 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, 後の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, 特徴の解像度を向上させる dilated ("atrous") convolution [298, 50, 33] などの技巧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

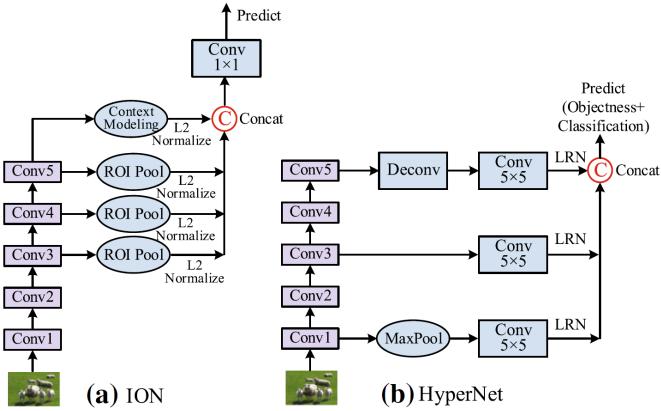


図 16 HyperNet と ION の比較. LRN は Local Response Normalization（局所応答正規化）であり、入力を局所領域（隣接チャンネル）にわたって正規化することで「側方抑制 (lateral inhibition)」の一種を行う [127, 140]. (訳注：arXiv(v4) 版では本図に誤りがあるため注意されたい.)

検出精度向上させるために、多数の手法 [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] など多数のアプローチが、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のは古典的なニューラルネットワークのアイディアであり、また、近年ではセマンティックセグメンテーションのアーキテクチャで普及している [177, 241, 97]。図 16 (a) に示すように、ION [11] は ROI pooling を使用して複数層から ROI 特徴を抽出し、selective search と Edge Boxes によって生成された物体提案を連結特徴を使用して分類する。図 16 (b) に示す HyperNet [135] も同様のアイディアに従っており、深い特徴・中間の特徴・浅い特徴を統合し、end-to-end の共同訓練戦略で物体提案の生成と物体の予測を行う。連結特徴はより叙事的位置推定と分類により有益だが、計算の複雑さが増す。

(2) CNN の複数層での検出 : 近年多数のアプローチが、様々な層で様々な解像度の物体を予測し、それらの予測を組み合わせることで検出を改善している(例: SSD [175], MSCNN [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 を撒き散らし、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 の後方の畠み込み層を Receptive Field Block (RFB) で置き換える、特徴の識別性と頑健性を強化する。RFB は Inception ブロック [263] 同様複数ブランチの畠み込みブロックだが、複数ブランチをサイズの異なる複数のカーネルと dilated convolution 層 [33] で組み合わせる。MSCNN [24] は CNN の複数層を用いて領域提案を学習し、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ランチが並ぶアーキテクチャを構築するが、各ブランチが同じ変換パラメータ (畠み込み層の重み) を共有する。様々なスケールの物体に受容野を適応するために、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補助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るのにほぼ同じ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、拡大した特徴マップから小さな物体を検出することで実現できる。そこで、両者の長所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テーション [177, 241] や人物姿勢推定 [194] で有効性が見いだされ、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 段階検出器の両方で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 (MLKP) [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ピラミッドを構築し、有望な結果を達成した。これらの手法は図 17 (a1)~(f1) から見て取れるように、トップダウンのネットワークと lateral connection (側方接続、側方結合) を組み入れて標準的なボトムアップのフィードフォワードネットワークを補完する、非常に類似した検出アーキテクチャを持つ。具体的には、ボトムアップパスを経た最上段の高レベルセマンティック特徴がトップダウンネットワークによって送り返され、lateral connection で処理されたボトムアップの中間層特徴と結合された後、結合された特徴が検出に使用される。図 17 (a2)~(e2) から分かるように、主な違いは、異なる層からの特徴を選択し複数層の特徴を組み合わせる単純な Feature Fusion Block (FFB) の設計にある。

FPN [167] は汎用的な特徴抽出器として、物体検出 [167, 168] やインスタンスセグメンテーション [102] を含むいくつかのアプリケーションで重要な改善を示している。基本的(ベーシック)な Faster RCNN システムでの使用で、FPN は COCO 検出データセットで最先端の結果を達成した。STDN [321] は DenseNet [118] を使用して異なる層の特徴を組み合わせ、異なる解像度の特徴マップを得るために scale transfer module を設計した。scale transfer module はわずかな追加コストで DenseNet に直接組み込むことができる。

PANet [174], FPR [137], DetNet [164], M2Det [315] などのより最近の研究は、図 17 (g)~(j) に示すように、FPN のようなピラミッドアーキテクチャを更に改善する様々な方法を提案している。Liu らは FPN をベースに PANet [174] (図 17 (g1)) を設計した。PANet はまず、情報パスを短縮し特徴ピラミッドを強化するために、低レベルからトップレベルへのボトムアップパスを clean lateral connection 付きでもう一つ追加する。次に、各提案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。更に、マスク予測の更なる改善のため、各提案について異なる視点から捉える補完ブランチを提案サブネットワークに追加する。これらの追加手順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

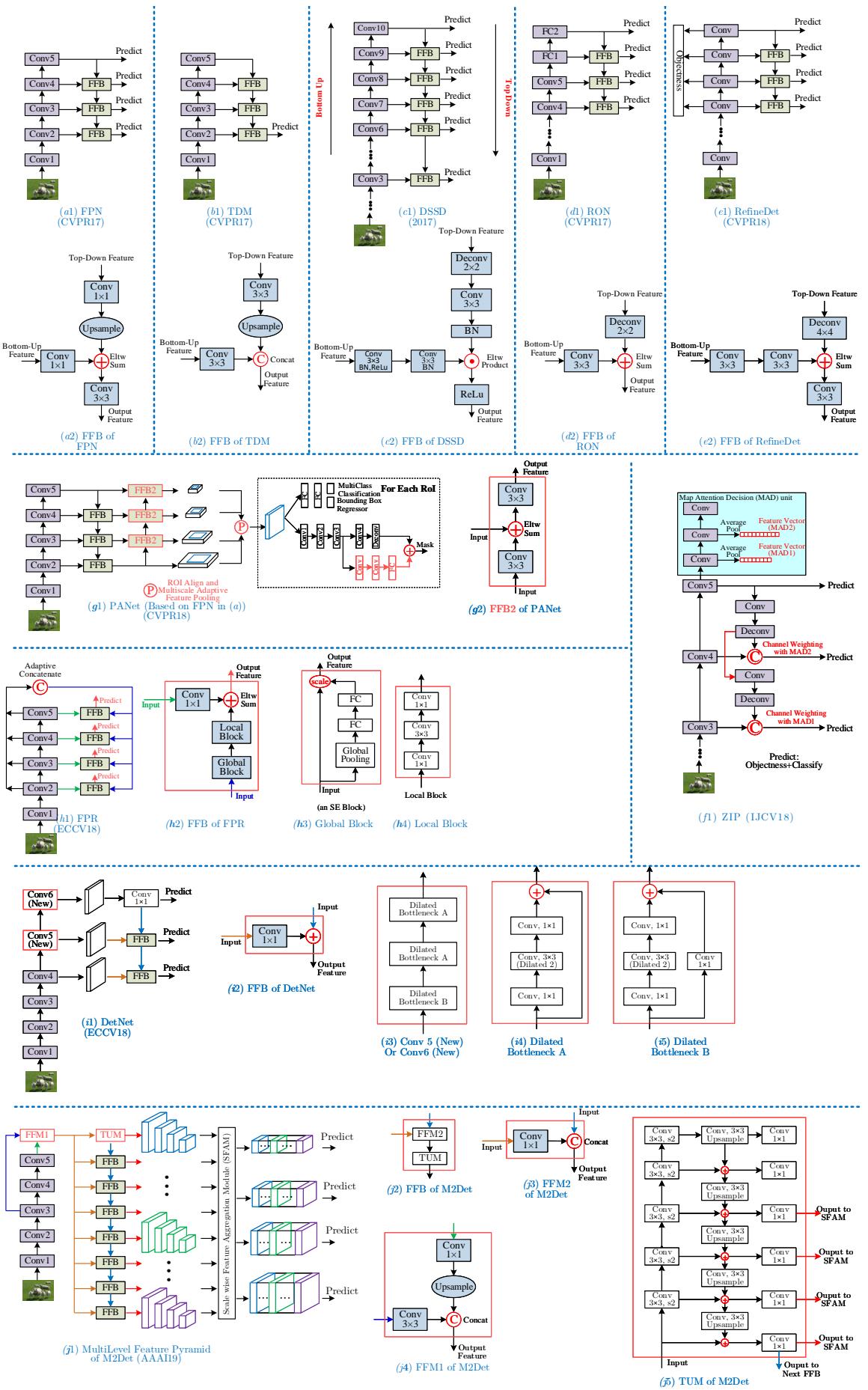


図 17 Hourglass アーキテクチャ。以下に示す近年のアプローチで一般的に使用される、多種の Feature Fusion Block (FFB) を比較している。FPN [167], TDM [247], DSSD [77], RON [136], RefineDet [308], ZIP [156], PANet [174], FPR [137], DetNet [164], M2Det [315]。FFM: Feature Fusion Module, TUM: Thinned U-shaped Module。Conv1~Conv5: VGG や ResNet などの backbone ネットワークの主たる Conv ブロック。

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ラミッド構築プロセス（例：FPN [167]）を特徴再構成関数として明示的に再定式化する、FPR [137] を提案した。FPN のように最上位層からの強力なセマンティック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結して抽出し、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 (i1)）を提案した。Zhao ら [315] は、異なるスケールの物体の検出により効果的な特徴ピラミッドを構築するために、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、ベース特徴からの lateral connection を持つトップダウンパスが特徴ピラミッド構築のために作成される。次に図 17 (j2), (j5) に示すように、FPN 等と比べてはるかに複雑な FFB module を使用し、FFB 内の Thinned U-shaped Module (TUM) で 2 番目のピラミッド構造を生成する。その後、複数の TUM からの同等サイズの特徴マップが物体検出のために組み合わされる。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 節で議論し図 6 にまとめているように、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。それらの変動を以下の 3 カテゴリに分類する。

- 幾何学的変換
- 遮蔽
- 画像の劣化

これらのクラス内変動に対するための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、数多の向きに回転した物体を訓練データに追加すれば実現できるだろう。頑健性はしばしばこの方法で学習できるが、通常は代償として訓練コストが上がりモデルパラメータが複雑になる。そのため、研究者はこれ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、入力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]。局所最大プーリング層の導入は DCNN に多少の平行移動不变性を与えたが、実際には中間特徴マップは入力データの大きな幾何学的変換に対して不变ではない [152]。そのため、頑健性向上のために多くのアプローチが提示され、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变な CNN 表現の学習が目指されてきた。代表的な研究の一つが、大域的パラメータ変換によりスケーリング・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。

回転不变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。しかし一般物体検出では、人気のベンチマークの検出データセット（例：PASCAL VOC, ImageNet, COCO）

で回転画像が実際は提示されないため、回転不变性に焦点を当て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、変形可能な構成で配置された構成パートによって物体を表現して一般物体検出で成功を収めた。DPM は性能では近年の物体検出器に大きく凌駕されたが、依然その精神は近年の多くの検出器に深い影響を与えている。DPM のモデリングは物体姿勢・視点・非剛体変形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] は、AlexNet で学習した深層特徴を DPM ベースの検出で使用することで DPM と CNN を組み合わせたが、領域提案は無かった。物体パートの変形をモデル化する、その内蔵された能力の恩恵を CNN が受けられるよう、DeepIDNet [203], DCN [51], DPFCN [188]（表 7）を含め多くのアプローチが提案された。精神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わたって共有視覚パターンとその変形特性を学習するため、通常の最大プーリングに代わる変形制約付きプーリング (deformation constrained pooling (def-pooling)) 層を設計した。DCN [51] は、規則的な格子状のサンプリング位置にオフセットを追加することで特徴マップのサンプリング位置を増大させるという考えに基づいて、deformable convolution 層と deformable ROI pooling 層を設計した。DPFCN [188] は、全パートの潜在変位を同時に最適化することで物体提案周囲の識別的な物体パートを選択する deformable part-based ROI pooling 層を提案した。

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。deformable parts のアイディアは遮蔽対処に有用となり得るため、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幾何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、遮蔽と変形を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、不十分な照明、低品質のカメラ、画像圧縮や、エッジデバイスとウェアラブルデバイスの意図して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予想されるが、PASCAL VOC, ImageNet, MS COCO, Open Images の全てが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なように、現在のほとんどの手法は劣化の無いクリーン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のところ非常に限られている。

7 コンテキストモデリング

物理的な世界では、視覚物体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には強力な心理学的証拠がある [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、コンテキストの適切なモデリングが物体の検出と認識に役立つと認識されている [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。

表8 コンテキスト情報を活用する検出器の概要。略語の詳細は表7同様。

グループ	検出器名	領域提案	Backbone DCNN	使用 パイプライン	mAP@IoU=0.5	mAP	発表先	注目点
					VOC07	VOC12		
グローバルコンテキスト	SegDeepM [326]	SS+CMPC	VGG16	RCNN	—	—	CVPR15	拡大した物体提案から抽出した特徴をコンテキスト情報として追加。
	DeepIDNet [203]	SS+EB	AlexNet ZFNet	RCNN	69.0 (07)	—	CVPR15	各物体提案の検出スコアを調整するため、画像分類スコアをグローバルコンテキスト情報として使用。
	ION [11]	SS+EB	VGG16	Fast RCNN	80.1	77.9	33.1	CVPR16
	CPF [245]	RPN	VGG16	Faster RCNN	76.4 (07T+12)	72.6 (07T+12)	—	ECCV16
ローカルコンテキスト	MRCNN [82]	SS	VGG16	SPPNet	78.2 (07T+12)	73.9 (07T+12)	—	ICCV15
	GBDNet [304, 305]	CRAFT [292]	Inception v2 ResNet269 PolyNet [311]	Fast RCNN	77.2 (07T+12)	—	27.0	ECCV16 TPAMI18
	ACCNN [157]	SS	VGG16	Fast RCNN	72.0 (07T+12)	70.6 (07T+12)	—	TMM17
	CoupleNet [327]	RPN	ResNet101	RFCN	82.7 (07T+12)	80.4 (07T+12)	34.4	ICCV17
	SMN [35]	RPN	VGG16	Faster RCNN	70.0 (07)	—	—	ICCV17
	ORN [114]	RPN	ResNet101 +DCN	Faster RCNN	—	—	39.0	CVPR18
	SIN [176]	RPN	VGG16	Faster RCNN	76.0 (07T+12)	73.1 (07T+12)	23.2	CVPR18

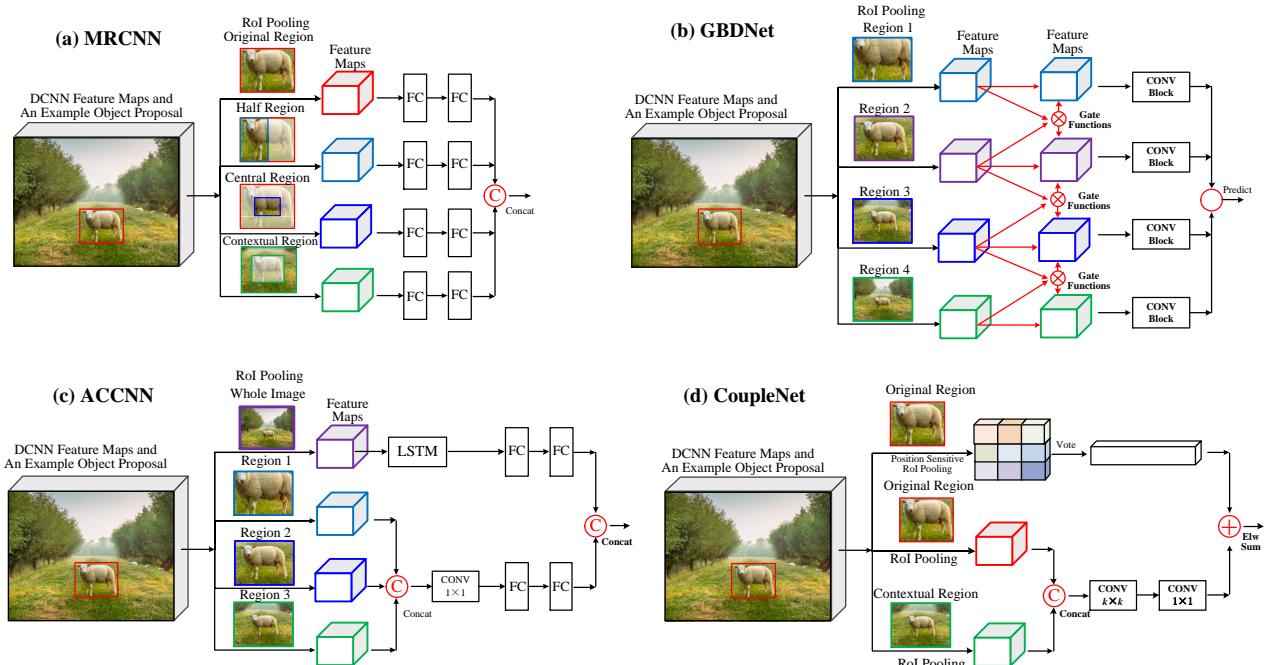


図18 局所周囲コンテキスト特徴について探求する代表的アプローチ：MRCNN [82], GBDNet [304, 305], ACCNN [157] and CoupleNet [327]。表8も参照されたい。

1. セマンティックコンテキスト：ある物体が、一部のシーンでは見られ他のシーンでは見られない尤度。
2. 空間的コンテキスト：どこかの位置で、ある物体は見られシーン内の他の物体は見られない尤度。
3. スケールコンテキスト：物体の取りうるサイズの範囲は、シーン内の他の物体との相対的なサイズにより限定されている。

深層学習の普及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だ DCNN

ベースの物体検出器で探求されていない [35, 114]。

物体検出における現在の最先端技術 [229, 175, 102] は、何のコンテキスト情報も明示的に利用せず物体を検出する。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合意されている [303, 316]。それにもかかわらず、DCNN ベースの検出器で明示的にコンテキスト情報を探求する価値がある [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] に動機づけられ、グローバルコンテキストとローカルコンテキストのカテゴリに分けていている。代表的なアプローチを表 8 にまとめる。

7.1 グローバルコンテキスト

グローバルコンテキスト [310, 78] は、物体検出の手がかりとして役立てられ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と予測される）。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 Recurrent Neural Network (RNN) の使用を提案した。SegDeepM [326] で Zhu らは、各検出に対して外観とコンテキストをスコア付けするマルコフ確率場モデルを提案した。この提案モデルは、各候補矩形が多数の物体セグメンテーション提案からセグメントを選択し、候補矩形とセグメント間の一一致をスコア付けできるようにしている。[245] では、セマンティックセグメンテーションがコンテキストプライミングの一形態として使用された。

7.2 ローカルコンテキスト

ローカルコンテキスト [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慮する。一般に物体の関係のモデル化はチャレンジングであり、クラス・位置・スケールなどの異なる bounding box について reasoning（推論、理由付け）する必要がある。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、空間メモリが本質的に行うのは、物体インスタンスを集めて疑似画像表現に戻すことである。疑似画像表現は物体の関係の推論のために別の CNN に簡単に入力できる。これにより画像とメモリを並行処理し、メモリを更に更新する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もたらされる。ORN^{*13}は、近年の自然言語処理での attention モジュールの成功 [274] に触発され、外観特徴と幾何特徴間の相互作用を通じて物体のセットを同時に処理する。ORN は追加の教師情報を必要とせず、既存ネットワークに簡単に組み込め、近年の物体検出パイプラインでの物体認識・重複除去ステップの改善に有効であり、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单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、物体はグラフ内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式でローカル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、より広範に研究してきた。代表的なアプローチとして、MRCNN [82], Gated BiDirectional CNN (GBDNet) [304, 305], Attention to Context CNN (ACCNN) [157], CoupleNet [327], Sermanet ら [238] の研究がある。MRCNN [82]（図 18 (a)）で Gidaris and Komodakis は、より豊かで頑健な物体表現を獲得するために、元々の物体提案から抽出されたバックボーンの最終 CONV 層の特徴に加えて、多数の異なる物体提案の領域（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 と密接に関連する手法が非常に多く提案してきた。[302] の手法は foveal structure で編成された 4 つのコンテキスト領域（中心窓を模して 1, 1.5, 2, 4 倍のサイズでクロップした領域）のみを使用し、複数パスの途中の分類器を end-to-end で共同で訓練する。Zeng らは検出性能を改善するために、物体提案を囲むマルチスケールのコンテキスト領域から特徴を抽出する GBDNet [304, 305]（図 18 (b)）を提案した。各領域に対して 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とは対照的に、GBDNet は異なるコンテキスト領域の特徴間でメッセージを受け渡す。メッセージの受け渡しは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したことに注意されたい。Li ら [157] はグローバルコンテキストとローカル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。グローバル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ンマップを繰り返し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) サブネットワークが使用され、ローカルコンテキストには MRCNN [82] と同様の手法が採用された。図 18 (d) に示すように、CoupleNet [327] は ACCNN [157] と概念的に類似しているが、position sensitive RoI pooling で物体の情報を捉える RFCN [50] をベースとしており、RoI pooling でグローバルコンテキストをエンコードするブランチが追加されている。

8 検出提案の手法

物体は画像内の任意の位置に任意のスケールで存在し得る。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用手法（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とともに増加する。また、複数のスケールとアスペクト比で探索する必要があるため探索空間は更に増加する^{*14}。そのため、高度な分類器を適用するには計算コストが高すぎる。

2011 年頃、検出提案^{*15}を使用して計算の扱いやすさと高い検出品質の間の緊張を緩和することが提案された [273, 271]。物体提案は [2] で提案された *objectness* 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候補領域のセットである。あまり多くない数（例：100）の物体提案で高い物体リコールを達成できる場合、スラ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。検出提案は前処理ステップ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。検出提案は以下の特性を持つ必要がある。

1. ごく少数の提案で高いリコールを達成できる。
2. 提案が物体の bounding box 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。

検出提案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]。物体提案は物体検出の域を越えたアプリケーションを持つ [6, 93, 328] ため、物

^{*14}スライディングウィンドウベースの検出では、画像あたり約 $10^4\text{--}10^5$ 個のウィンドウを分類する必要がある。複数のスケールとアスペクト比を考慮すると、ウィンドウ数は画像あたり $10^6\text{--}10^7$ まで大幅に増加する。

^{*15}本稿では、検出提案 (*detection proposals*)、物体提案 (*object proposals*)、領域提案 (*region proposals*) の各用語を同じ意味で用いる。

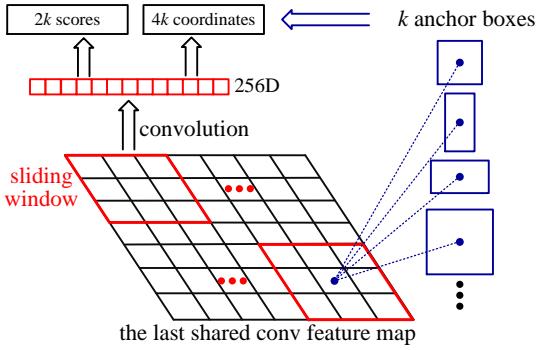


図 19 [229] で導入された Region Proposal Network (RPN) の図解。

体提案アルゴリズムの包括的レビューは本論文の範囲を超えており、興味のある読者は、多数の古典的な物体提案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析した近年のサーベイ [110, 27] を参照されたい。ここでの我々の関心は、DCNN ベースであり、クラス非依存の提案を出し、かつ一般物体検出に関する物体提案手法のレビューである。

2014 年、物体提案 [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。それからすぐに検出提案は標準的な前処理ステップになった。このことは、PASCAL VOC [68], ILSVRC [234], MS COCO [166] 物体検出チャレンジで、2014 年以降の優勝エントリの全てが検出提案を使用した [85, 203, 84, 229, 305, 102] ことから裏付けられる。

伝統的な低レベルの手がかり（例：色・テクスチャ・エッジ・勾配）に基づく物体提案のアプローチの中では、Selective Search [271], MCG [7], Edge Boxes [330] の人気が高い。しかし、本分野が急速に進歩するにつれて、検出器と独立した外部モジュールとして採用された伝統的な物体提案のアプローチ [271, 110, 330] が、検出パイプラインの速度のボトルネックとなった [229]。新しく登場した DCNN を使用する物体提案アルゴリズム [67, 229, 142, 81, 213, 292] が幅広い注目を集めている。

近年の DCNN ベースの物体提案手法は一般に、*bounding box* ベースと物体セグメントベースの 2 カテゴリに分類される。代表的な手法を表 9 にまとめる。

bounding box 提案の手法 は、図 19 に示す Ren らの RPN [229] が最も良い例となる。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提案を予測する。スライディングウインドウの各位置で、 k 個の anchor box を使用して k 個の提案が予測される。ここで、各 anchor box^{*16} は画像内の特定の位置を中心とし、特定のスケールとアスペクト比に関連付けられる。Ren ら [229] は、畳み込み層の共有により 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 がもたらされた。表 7, 8 から分かるように、物体提案の手法として RPN は多くの最先端の物体検出器で採用されている。

MultiBox [67, 262] や RPN [229] のように先駆的な固定された 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サブ領域に焦点を当てるよう適応的に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。画像全体から始めて、探索処理中に訪れた全領域が anchor として機能する。探索手順

の中で遭遇した 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 zoom indicator が使用される。そして、ブランチを追加して RPN を拡張し既存ブランチと並行して zoom indicator を計算する Adjacency and Zoom Network (AZNet) によって、objectness スコアを持つ bounding box のセットが計算される。

更に、複数層の畳み込み特徴の利用による物体提案生成が試みられている。RPN [229] と同時期に、Ghodrati ら [81] は複数の畳み込み特徴のカスケードを使用して物体提案を生成する DeepProposal を提案した。DeepProposal は、最も有望な物体位置を選択しそれらの矩形を coarse-to-fine な方法で改良するため、（最終畳み込み層から逆にたどる）逆カスケードを構築する。RPN を改良した亞種である HyperNet [135] は、複数層の畳み込み特徴を集約する Hyper Features を設計し、end-to-end の共同訓練戦略を介してそれらを提案生成と物体検出の両者で共有する。Yang らは同じくカスケード戦略を使用する 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提案を生成し、次にそれらを使用して、物体と背景を更に区別するための 2 クラス分類を行うもう一つの Fast RCNN ネットワークを訓練する。Li ら [156] は RPN を改善する 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を統合するために、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提案を予測する。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。DeepBox [142] は、Edge Boxes が生成した提案の再ランク付けを学ぶために軽量の CNN を提案した。DeNet [269] は、効率的に物体提案を予測する bounding box コーナー推定を導入し、Faster RCNN 形式の検出器の RPN を置き換える。

物体セグメント提案の手法 [213, 214] は、物体に対応する可能性が高いセグメント提案の生成を目標とする。セグメント提案は bounding box 提案よりも情報量が多く、物体インスタンスセグメンテーション [96, 49, 162] へと一步近づく。また、インスタンスセグメンテーションの教師情報を使用することで bounding box 物体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提案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 つのブランチに分割し、クラス非依存の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インドウ戦略と同様、訓練済みの DeepMask ネットワークは推論時、画像（およびそのスケール変更版）にスライディングウイ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 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 (b1), (b2) 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ドされた強力な意味情報を効率的に統合でき、忠実度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提案を生成する Instance-FCN [48] を提案した。DeepMask 同様 Instance-FCN のネットワークは 2 つの fully convolutional 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み込み特徴を使用するため、SSD [175] と同様の one-shot の方法

*16 “anchor”的概念は [229] で初めて登場した。

表 9 DCNN を使用する物体提案手法の概要. 青字の値は物体提案数を表す. 特記のない限り COCO の検出結果は mAP@IoU[0.5, 0.95] に基づく.

Bounding Box 物体提案の手法	提案器名	Backbone Network	テストした 検出器	Recall@IoU (VOC07)			検出結果 (mAP)			発表先	注目点
				0.5	0.7	0.9	VOC07	VOC12	COCO		
	MultiBox [67]	AlexNet	RCNN	—	—	—	29.0 (10) (12)	—	—	CVPR14	所定の 800 個の anchor box からなる小さなセットでクラス非依存の回帰器を学習. 検出と特徴を共有しない. (誤注: 800 個使用するのは MSC-MultiBox [262] であり MultiBox [67] は 100 個または 200 個を使用する.)
	DeepBox [142]	VGG16	Fast RCNN	0.96 (1000)	0.84 (1000)	0.15 (1000)	—	—	37.8 (500) (IoU@0.5)	ICCV15	Edge Boxes が生成した提案の再ランク付けを学ぶために軽量の CNN を使用. 画像あたり 0.26 秒で実行可能. 検出と特徴を共有しない.
	RPN [229, 230]	VGG16	Faster RCNN	0.97 0.98 (1000)	0.79 0.84 (1000)	0.04 0.04 (1000)	73.2 (300) (07+12)	70.4 (300) (07+12)	21.9 (300)	NIPS15	画像全体の畳み込み特徴を検出と共有することで物体提案を生成した最初の手法. 最も広く使用される物体提案手法. 検出速度を大幅に改善.
	DeepProposal [81]	VGG16	Fast RCNN	0.74 0.92 (1000)	0.58 0.80 (1000)	0.12 0.16 (1000)	53.2 (100) (07)	—	—	ICCV15	マルチスケール方式で DCNN 内で提案を生成. 検出ネットワークと特徴を共有.
	CRAFT [292]	VGG16	Faster RCNN	0.98 (300)	0.90 (300)	0.13 (300)	75.7 (07+12)	71.3 (12)	—	CVPR16	RPN の後に分類ネットワーク (つまり, 2 クラス分類の Fast RCNN) のカスクードを導入. 検出用に抽出された特徴を共有しない.
	AZNet [181]	VGG16	Fast RCNN	0.91 (300)	0.71 (300)	0.11 (300)	70.4 (07)	—	22.3	CVPR16	coarse-to-fine 探索を使用. 大きな領域から開始し, 物体を含む可能性のあるサブ領域を再帰的に探索. 適応的に計算リソースをガイドし可能性のあるサブ領域に注力する.
	ZIP [156]	Inception v2	Faster RCNN	0.85 COCO	0.74 COCO	0.35 COCO	79.8 (07+12)	—	—	IJCV18	多層の conv-deconv ネットワークを使用して提案を生成. 様々な層からの特徴に重みを割り当てるため map attention decision (MAD) ユニットを提案.
	DeNet [269]	ResNet101	Fast RCNN	0.82 (300)	0.74 (300)	0.48 (300)	77.1 (07+12)	73.9 (07+12)	33.8	ICCV17	Faster RCNN よりはるかに高速. RPN に代わる効率的に物体提案を予測するための bounding box コーナー推定を導入. 事前定義された anchor が不要.
セグメント提案の手法	提案器名	Backbone Network	テストした 検出器	Box 提案 (AR, COCO)		セグメント提案 (AR, COCO)		発表先	注目点		
				0.33 (100), 0.48 (1000)	—	0.26 (100), 0.37 (1000)	—				
	DeepMask [213]	VGG16	Fast RCNN	—	—	0.32 (100), 0.39 (1000)	—	NIPS15	DCNN で物体マスク提案を生成した最初の手法. 推論時間が遅い. 訓練にセグメンテーションのアノテーションが必要. 検出ネットワークと特徴を共有しない. Fast RCNN で 69.9% (500) の mAP を達成.		
	InstanceFCN [48]	VGG16	—	—	—	0.32 (100), 0.39 (1000)	—	ECCV16	FCN [177] と DeepMask [213] のアイデアを組み合わせる. instance-sensitive score map を導入. ネットワークの訓練にセグメンテーションのアノテーションが必要.		
	SharpMask [214]	MPN [302]	Fast RCNN	0.39 (100), 0.53 (1000)	—	0.30 (100), 0.39 (1000)	—	ECCV16	トップダウンの refinement module を導入することで複数の畳み込み層の特徴を活用. 検出ネットワークと特徴を共有しない. 訓練にセグメンテーションのアノテーションが必要.		
	FastMask [113]	ResNet39	—	0.43 (100), 0.57 (1000)	—	0.32 (100), 0.41 (1000)	—	CVPR17	SSD [175] 同様の one-shot 方式で効率的にインスタンスセグメント提案を生成. マルチスケールの畳み込み特徴を使用. 訓練にセグメンテーションのアノテーションを使用.		

で効率的にインスタンスセグメント提案を生成する FastMask [113] を提案した. マルチスケールの畳み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 の画像に対して 13 FPS 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¹⁷. DCNN を学習するためにデータ拡張 (data augmentation) を行う [26, 84, 85] ことは, 視覚認識にとって重要であると一般に認識されている. 平凡なデータ拡張は, クロッピング・反転・回転・スケーリング・平行移動・色揃動・ノイズ付加など, 根底にあるカテゴリを変更しない変換によって画像に摂動を加えることを指す. データ拡張はサンプル数を人為的に増大させ, 過適合を軽減し汎化を改善するのに役立つ. 訓練時, テスト時, 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な欠点がある. データ拡張は完全に新しい訓練画像を合成してもよい [210, 280] が, 合成画像が実画像にうまく汎化す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¹⁸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, 物体を適切な環境に配置するためには物体を囲む視覚コンテキストの適切なモデル化が重要であると示し, デー

タ拡張用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.

新規の訓練戦略. 広範囲のスケール変動下での物体の検出, 特に非常に小さな物体の検出は, 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な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, スケール不変性の問題を説明し, 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 は, 小さな物体と大きな物体はそれぞれより小さなスケールとより大きなスケールでの検出が困難という直感的な理解に基づき, 訓練サンプルを削減することなく訓練中のスケール変動を低減できる新しい訓練スキームを導入する. SNIPER は, 画像ピラミッド全体を処理する代わりに, 適切なスケールで真値物体の周囲のコンテキスト領域のみを処理することで, 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ミニバッチサ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 での 16 に対し MegDet では最大 256) での訓練を可能にした. 収束の失敗の回避と訓練プロセスの大幅な高速化のため, Peng ら [209] は学習率ポリシーと Cross-GPU Batch Normalization を提案し 128 GPU を効果的に活用した. これにより, MegDet は 128 GPU を使って 4 時間で 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 bounding

*17 誤注: この段落は [63] に依拠する記述が多いため詳細はそちらを参照されたい.

*18 誤注: [64] では実写画像から切り抜かれた物体が使用されるが, [94] では合成されたテキストが使用される.

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	領域提案	Backbone DCNN	使用パイプライン	VOC07 結果	VOC12 結果	COCO 結果	発表先	注目点
MegDet [209]	RPN	ResNet50 +FPN	Faster RCNN	—	—	52.5	CVPR18	Cross-GPU Batch Normalization を導入することで、以前よりはるに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、精度向上を達成。COCO2017 検出チャレンジで優勝。
SNIP [249]	RPN	DPN [37] +DCN [51]	RCNN	—	—	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ップサンプリングの効果を実験的に調査。訓練時、特徴のスケールに合った物体のみを正例として選択。
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。真値インスタンスの周囲のコンテキスト領域を適切なスケールで処理。
OHEM [246]	SS	VGG16	Fast RCNN	78.9 (07+12)	76.3 (07++12)	22.4	CVPR16	領域ベースの検出器の訓練を改善するための単純で効果的な Online Hard Example Mining アルゴリズム。
FactorNet [204]	SS	GooglNet	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 Inceptionv2	Fast RCNN, Faster RCNN	80.4 (07+12) (SS+VGG)	—	—	ICCV17	DCNN とカスケード分類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 2007 と ImageNet での検出精度を向上。データセットによって異なる領域提案の手法を使用。
Cascade RCNN [23]	RPN	VGG ResNet101 +FPN	Faster RCNN	—	—	42.8	CVPR18	DCNN とカスケード分類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、正例選択用の位置推定精度として異なる閾値を使用。複数のステージの bounding box 回帰を積み重ねる。
RetinaNet [168]	—	ResNet101 +FPN	RetinaNet	—	—	39.1	ICCV17	hard example での訓練に注力する新規の 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。

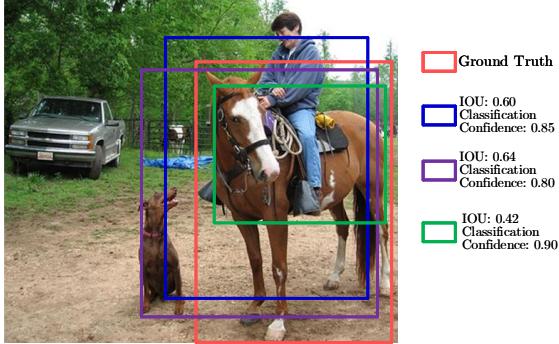


図 20 位置推定エラーは、オーバーラップ不足や重複検出で起こることがあり、しばしば過検出の原因となる。(カラー図はオンラインで参照可。)

box とその真値 box 間の Intersection Over Union^{*19} (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：典型値の 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、物体提案にクラスラベルを割り当てる softmax 分類器と、検出結果と真値間の 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 bounding box 回帰器を共同で最適化する。bounding box は連結された物体の大雑把な近似に過ぎないため、背景画素がほぼ常に bounding box に含まれ分類と位置推定の精度に影響する。[108] らの研究は、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なエラーの形式の一つであることを示している。位置推定エラーは、オーバーラップ不足（図 20 の緑色の box 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] のような何らかの後処理ステップが用いられる。し

かし、(図 20 に示す紫色の矩形のような) 位置推定精度の良い bounding box 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かなり多くある。

MRCNN [82] は RCNN を数回適用する反復 bounding box 回帰を導入する。CRAFT [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、正確な提案を生成して Fast RCNN に転送する。Cai and Vasconcelos は RCNN を多段階に拡張し、close false positive (真値 box に近いが正解とはならない惜しい bounding box) に対して順次選択性を高めるために、一連の検出器を IOU 閾値を増加させながら順次訓練する Cascade RCNN [23] を提案した。これは、特定の IOU で訓練された検出器の出力が、その次のより高い IOU 閾値の検出器を訓練するのに適した分布になっているという観察に基づいている。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）を達成することが実証されている。また、最近では直接 IOU を最適化の目的関数として定式化する研究 [128, 232, 121] や、Soft NMS [18] や learning NMS [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画像領域）の数との間の重大な不均衡である。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覆い尽くしがちである。過去には、この問題は通常 bootstrapping [259] などの技術で対処してきた。最近では、この問題にもいくらか注目が集まっている [153, 168, 246]。領域提案段階でほとんどの背景領域が速やかに除外され少数の物体候補が提案されるため、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。前景と背景の適切なバランスを維持するために、Online Hard Example Mining (OHEM) [246] などの example mining 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、全物体に対して背景の例

*19 IOU の定義の詳細は 4.2 節を参照されたい。

が 100,000). Lin ら [168] は Focal Loss を提案し, 正しく分類された例に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. Li 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 議論（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なチャレンジングな問題でありかなりの注目を集めている. 深層学習技術の著しい発展により, 物体検出の分野は劇的に進歩した. 一般物体検出のための深層学習に関する包括的なサービスとして, 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な手法分類法を提供し, 既存の一般的なデータセットと評価基準をまとめ, 最も代表的な手法の性能について説明した. 以下, 10.1 節で最先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る.

10.1 最先端の性能

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な点/文脈上の点でそれらは異なるためである.

- RCNN [85], Fast RCNN [84], Faster RCNN [229], RFCN [50], Mask RCNN [102], YOLO [227], SSD [175] などのメタな検出フレームワーク
- VGG [248], Inception [263, 125, 264], ResNet [101], ResNeXt [291], Xception [45] などの表 6 に記載した backbone ネットワーク
- 複数層の特徴の組み合わせ [167, 247, 77], deformable convolutional network [51], deformable 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], Open Images [139] などのデータセットでの事前学習
- 検出提案の手法の差異と物体提案数の差異
- 訓練時/テスト時のデータ拡張, 新しいマルチスケール訓練戦略 [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のは非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て統一された方法で比較することは有益である. バックボーンネットワーク, 画像解像度, bounding box の提案数を変化させて 3 種の主要な検出器 (Faster RCNN [229], RFCN [50], SSD [175]) を比較した Huang らの研究 [120] を除き, この点では研究が非常に限られている.

表 7, 8, 9, 10, 11 から分かるように, 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は, 上記の 1 つ以上の面で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た.

図 3, 21 は, PASCAL VOC, ILSVRC, MSCOCO チャレンジの最高の検出結果をまとめ, 最先端の概要を非常に簡潔に示している(より多くの結果は検出チャレンジの web サイトで見られる [124, 189, 208]). open image challenge の物体検出タスクコンペティションの勝者は, Fast RCNN [84], Faster RCNN

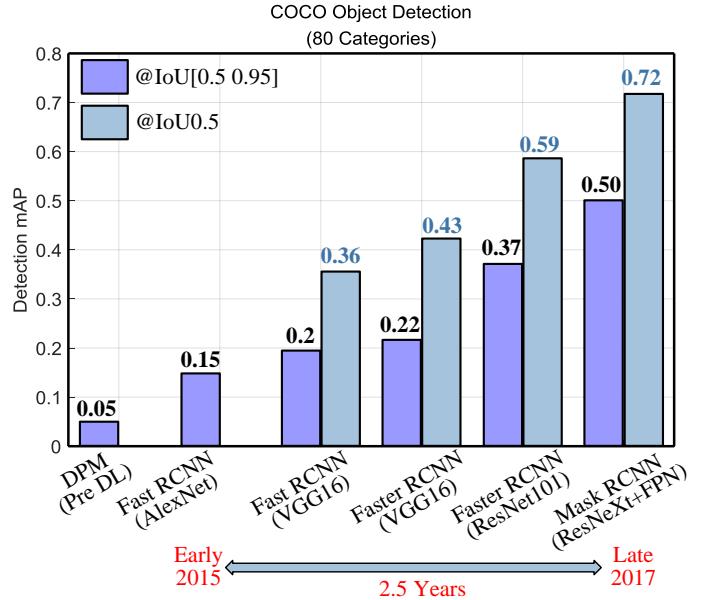


図 21 COCO の物体検出性能の進歩 (Test-Dev の結果). 結果は [84, 102, 230] から引用. バックボーンネットワーク, 検出フレームワークの設計, そして適切な大規模データセットを利用できることが, 検出精度において最も重要な 3 要素である.

[229], FPN [167], Deformable RCNN [51], Cascade RCNN [23] を含むいくつかの 2 段階検出器の検出結果を組み合わせ, public リーダーボードで 61.71% mAP, private リーダーボードで 58.66% mAP を達成した. 要約すると, バックボーンネットワーク, 検出フレームワーク, そして大規模データセットを利用できることが, 検出精度において最も重要な 3 要素である. 複数のモデルのアンサンブル, コンテキスト特徴の取り込み, データ拡張は, いずれもより良い精度の達成に役立つ.

図 15 に示すように, AlexNet [140] の提案から 5 年で, ImageNet [234] 1000 クラス分類の Top5 error は 16% から 2% に減少した. しかし, わずか 80 クラスを検出するよう訓練された COCO [166] での最高性能検出器 [209] の mAP は, IoU 閾値 0.5 での評価であってもわずか 73% である. これは物体検出が画像分類よりもはるかに難しいことを示している. 最先端の検出器によって達成される精度と頑健性は, 実世界のアプリケーションの要求を満たすにはほど遠いため, 今後の改善の余地は大きい.

10.2 まとめと議論

本論文を通し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を出す. このことは, 有名な検出チャレンジで使用される勝利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 RCNN [102] である.
- [120] で示されたように, 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バックボーンネットワークの品質の影響を受けにくい。

- YOLO [227] や SSD [175] のような1段階検出器は、前処理のアルゴリズムを回避し、軽量のバックボーン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 fully convolutional 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段階検出器も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段階のいずれにせよ、最も時間のかかるステップは特徴抽出器（バックボーンネットワーク）である [146, 229]。
- [120, 227, 175] で示されたように、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るかに劣っている。

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・頑健な）検出器を構築する試みが多くなされている。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収束している。

- fully convolutional 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 RCNN [102]）
- スライディングウィンドウ [229]
- バックボーン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] とインスタンスセグメンテーション [31] における近年のカスケードの成功が証拠として示すように、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、バックボーンネットワークは検出性能の急速な改善の背後に主な推進力の1つである。一般に、ResNet [101], ResNeXt [291], InceptionResNet [265] などの深い backbone の方が高性能である。しかし、それらは（推論の）計算コストが高く、訓練にはるかに多くのデータと大量の計算を要す。逆に、速度を重視する backbone [112, 123, 312] もいくつか提案されている^{*20}。例えば MobileNet [112] は、わずか $\frac{1}{30}$ の計算コストとモデルサイズで、VGGNet16 とほぼ同等の ImageNet 精度を達成すると示されている。より多くの訓練データとより良い訓練戦略が利用可能になった [285, 183, 182] ため、backbone のクラッチ学習（ランダム初期値からの訓練）が可能になるかもしれない。

(3) 物体表現の頑健性の向上

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・背景の乱雑さ・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。

- 画像ピラミッドの使用：単純かつ効果的で、小さな物体の

拡大や大き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論時に一般的に使用される。

- 異なる解像度の畳み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して実行され、他の層からの情報を組み合わせたり混ぜ合わせたりはしない。今日では、例えば FPN [167] のように様々な層の特徴の組み合わせがごく標準的に行われる。
- dilated convolution [164, 163] の使用：より広いコンテキストを取り込みつつ高解像度の特徴マップを維持するための単純で効果的な手法である。
- 様々なスケールとアスペクト比の anchor box の使用：多数のパラメータを持つ欠点がある上、anchor box のスケールとアスペクト比は通常ヒューリスティックに決定される。
- アップスケーリング：特に、小さな物体の検出のために high-resolution network [255, 256] を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な物体の検出精度よりはるかに低い。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つのままである。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なため、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る必要がある。

(3.2) 変形・遮蔽・その他の要因

2.2 節で説明したように、幾何学的変換・遮蔽・変形に対処するアプローチは主に2つのパラダイムに基づく。1つ目は spatial transformer network であり、回帰を使用して変形場を得た後その変形場によって特徴をワーピングする [51]。2つ目は deformable part-based model [74] に基づくもので、空間的制約を考慮に入れたペーパーフィルタに対する応答が最大となる位置を見つける [203, 86, 277]。

回転不变性は特定のアプリケーションでは魅力的かもしれないが、一般的なベンチマークの検出データセット (PASCAL VOC, ImageNet, COCO) には大きな回転の変動がないため、回転不变性に焦点を当てた一般物体検出の研究は限られている。遮蔽の対処は顔検出と歩行者検出で集中的に研究されているが、一般物体検出のための遮蔽対処に専念する研究はほとんどない。近年の進歩にもかかわらず、一般に深層ネットワークは様々な変動に対する頑健性が欠如しており依然性能が悪い。そのため、実世界のアプリケーションが大幅に制約されている。

(4) コンテキストの推論 (reasoning)

7 節で紹介したように、“in the wild” の物体は通常他の物体や環境と共に存在する。特に小さな物体、遮蔽された物体、低画質の場合に、コンテキスト情報（物体の関係性、グローバルシーン統計）が物体の検出と認識に役立つ [197] と認識されている。深層学習の普及前に広範な研究が行われ [185, 193, 220, 58, 78]、深層学習の時代にもかなりの数の研究が行われている [82, 304, 305, 35, 114]。コンテキスト情報を効率的かつ効果的に取り込む方法はまだ探求する必要があり、人間の視覚がいかにしてコンテキストを使用しているかが指針となるだろう。そのため、シンググラフ [161] に基づくことや、panoptic segmentation [134] による物体とシーンの完全なセグメンテーションを介すことが考えられる。

(5) 検出提案

検出提案は探索空間を大幅に削減する。将来の検出提案では

^{*20} 訳注：SqueezeNet [123] が重視しているのはパラメータ数やモデルサイズである。

[110] で推奨されているように、繰り返し精度、recall、位置推定精度、速度を確実に改善する必要がある。共通のフレームワークに提案生成と検出を統合した RPN [229] の成功以来、CNN ベースの検出提案生成手法が領域提案で支配的になっている。新しい検出提案は、検出提案を単体で評価するのではなく物体検出のための評価を行うことが推奨される。

(6) その他の要因

9 節で説明したように、物体検出の品質に影響するその他の要因が以下のように多数ある。データ拡張、新しい訓練戦略、バックボーンモデルの組み合わせ、複数の検出フレームワーク、他の関連タスクからの情報の取り込み、位置推定エラーの低減手法、正例・負例間の巨大な不均衡の対処、ハードネガティブマイニング、損失関数の改善。

10.3 研究の方向性

物体検出技術は、本分野における近年の驚異的な進歩にもかかわらず、人間の視覚に比べるかに原始的なままであり、2.2 節のような実世界の課題にまだ十分対処できない。以下のような長年の課題がいくつかある。

- オープンワールドで動作すること：あらゆる環境の変化に頑健であり、進化または適応できること。
- 制約条件下での物体検出：弱教師ラベル付きデータまたは少数の bounding box アノテーションからの学習、ウェアラブルデバイス、未知の物体カテゴリ、等。
- 他のモダリティでの物体検出：動画、RGBD 画像、3 次元点群、LiDAR、リモートセンシング画像、等。

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、理想的には新規の物体カテゴリを認識 [144, 95] する能力を持つべきだが、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もかなり少ない数十から数百のカテゴリしか含まない。か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]) の両方で、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理由の一つである。領域ベースの検出器は高精度であり、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、学習にバイアスが生じる可能性がある。物体検出器のスクラッチ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) コンパクトで効率的な CNN 特徴：CNN は数層 (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百万から数億のパラメータがあり、訓練には大量のデータと GPU を要す。ネットワークの冗長性を低減または除去するために、コンパクトで軽量な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 [44, 122, 253, 155, 158, 282] を行う研究への関心が高まっている。

(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、強力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] 近年の Automated Machine Learning (AutoML) [219] のように、検出バックボーンのアーキテクチャの自動設計を検討するのは自然なことである。

(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、個々の物体の正確な境界を必要とする潜在的なアプリケーションで重要な役割を果たせる。

(6) 弱教師あり検出：現在の最先端の検出器は、物体の bounding box 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 bounding box 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に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 の力をいかにして引き出すかの理解が不可欠である。

(7) 少数 / ゼロショット (Few / Zero Shot) 物体検出：深層検出器の成功は膨大な量の注釈付き訓練データに大きく依存している。ラベル付きデータが不足している場合、深層検出器の性能はしばしば低下し汎化に失敗する。これとは対照的に、人間は（たとえ子供でも）非常に少数の既定の例から即座に視覚的な概念を学ぶことができ、たいていうまく汎化（一般化）できる [16, 144, 71]。そのため、わずか数例から学習する few shot (少数ショット) 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 zero shot (ゼロショット) 物体検出は、これまでに見たことのない物体クラスを位置推定し認識する^{*21} [9, 53, 222, 221]。これは、新しい物体カテゴリを知的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) 他のモダ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。他のモダ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。デプス（奥行き、深度）[36, 211, 289, 286]、動画 [70, 130]、点群 [217, 218] の効果的な使用に関して、これらのモダリティは新たな課題を提起する。

(9) 普遍物体検出：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] な普遍表現 (*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は画像分類に焦点を当てており、物体検出を対象とすることは滅多になく [281]、通常開発された検出器はドメイン特化である。画像ドメインに非依存の物体検出とクロスドメインの物体検出は、重要な将来の方向性を示している。

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

Acknowledgements 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。一般物体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 Jiří 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。

^{*21} ただし wikipedia のページや属性ベクトルなどの補助情報 (side information) が与えられる場合がある。

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 multilayer neural networks for object recognition. In: ECCV, pp. 329–344 [14](#)
- [2] Alexe B., Deselaers T., Ferrari V. (2010) What is an object? In: CVPR, pp. 73–80 [20](#)
- [3] Alexe B., Deselaers T., Ferrari V. (2012) Measuring the objectness of image windows. IEEE TPAMI 34(11):2189–2202 [20](#)
- [4] Alvarez J., Salzmann M. (2016) Learning the number of neurons in deep networks. In: NIPS, pp. 2270–2278 [26](#)
- [5] Andreopoulos A., Tsotsos J. (2013) 50 years of object recognition: Directions forward. Computer Vision and Image Understanding 117(8):827–891 [2, 3](#)
- [6] Arbeláez P., Hariharan B., Gu C., Gupta S., Bourdev L., Malik J. (2012) Semantic segmentation using regions and parts. In: CVPR, pp. 3378–3385 [20](#)
- [7] Arbeláez P., Pont-Tuset J., Barron J., Marques F., Malik J. (2014) Multi-scale combinatorial grouping. In: CVPR, pp. 328–335 [20, 21](#)
- [8] Azizpour H., Razavian A., Sullivan J., Maki A., Carlsson S. (2016) Factors of transferability for a generic convnet representation. IEEE TPAMI 38(9):1790–1802 [14](#)
- [9] Bansal A., Sikka K., Sharma G., Chellappa R., Divakaran A. (2018) Zero shot object detection. In: ECCV [26](#)
- [10] Bar M. (2004) Visual objects in context. Nature Reviews Neuroscience 5(8):617–629 [18](#)
- [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 [15, 16, 19, 20](#)
- [12] Belongie S., Malik J., Puzicha J. (2002) Shape matching and object recognition using shape contexts. IEEE TPAMI 24(4):509–522 [4](#)
- [13] Bengio Y., Courville A., Vincent P. (2013) Representation learning: A review and new perspectives. IEEE TPAMI 35(8):1798–1828 [2, 3, 5, 6, 13](#)
- [14] Biederman I. (1972) Perceiving real world scenes. IJCV 177(7):77–80 [18](#)
- [15] Biederman I. (1987) Recognition by components: a theory of human image understanding. Psychological review 94(2):115 [5](#)
- [16] Biederman I. (1987) Recognition by components: a theory of human image understanding. Psychological review 94(2):115 [26](#)
- [17] Bilen H., Vedaldi A. (2016) Weakly supervised deep detection networks. In: CVPR, pp. 2846–2854 [26](#)
- [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 [23](#)
- [19] Borji A., Cheng M., Jiang H., Li J. (2014) Salient object detection: A survey. arXiv: 14115878v1 1:1–26 [3](#)
- [20] Bourdev L., Brandt J. (2005) Robust object detection via soft cascade. In: CVPR, vol 2, pp. 236–243 [11](#)
- [21] Bruna J., Mallat S. (2013) Invariant scattering convolution networks. IEEE TPAMI 35(8):1872–1886 [18](#)
- [22] Cai H., Yang J., Zhang W., Han S., Yu Y. (2018) Path level network transformation for efficient architecture search [26](#)
- [23] Cai Z., Vasconcelos N. (2018) Cascade RCNN: Delving into high quality object detection. In: CVPR [11, 23, 24, 25](#)
- [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 [15, 16](#)
- [25] Carreira J., Sminchisescu C. (2012) CMPC: Automatic object segmentation using constrained parametric mincuts. IEEE TPAMI 34(7):1312–1328 [20](#)
- [26] Chatfield K., Simonyan K., Vedaldi A., Zisserman A. (2014) Return of the devil in the details: Delving deep into convolutional nets. In: BMVC [22](#)
- [27] Chavali N., Agrawal H., Mahendru A., Batra D. (2016) Object proposal evaluation protocol is gameable. In: CVPR, pp. 835–844 [9, 21](#)
- [28] Chellappa R. (2016) The changing fortunes of pattern recognition and computer vision. Image and Vision Computing 55:3–5 [18](#)
- [29] Chen G., Choi W., Yu X., Han T., Chandraker M. (2017) Learning efficient object detection models with knowledge distillation. In: NIPS [26](#)
- [30] Chen H., Wang Y., Wang G., Qiao Y. (2018) LSTD: A low shot transfer detector for object detection. In: AAAI [26](#)
- [31] Chen K., Pang J., Wang J., Xiong Y., Li X., Sun S., Feng W., Liu Z., Shi J., Ouyang W., et al. (2019) Hybrid task cascade for instance segmentation. In: CVPR [11, 25](#)
- [32] Chen L., Papandreou G., Kokkinos I., Murphy K., Yuille A. (2015) Semantic image segmentation with deep convolutional nets and fully connected CRFs. In: ICLR [18](#)
- [33] Chen L., Papandreou G., Kokkinos I., Murphy K., Yuille A. (2018) DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE TPAMI 40(4):834–848 [15, 16, 18](#)
- [34] Chen Q., Song Z., Dong J., Huang Z., Hua Y., Yan S. (2015) Contextualizing object detection and classification. IEEE TPAMI 37(1):13–27 [19](#)
- [35] Chen X., Gupta A. (2017) Spatial memory for context reasoning in object detection. In: ICCV [19, 20, 25](#)
- [36] Chen X., Kundu K., Zhu Y., Berneshawi A. G., Ma H., Fidler S., Urtasun R. (2015) 3d object proposals for accurate object class detection. In: NIPS, pp. 424–432 [26](#)
- [37] Chen Y., Li J., Xiao H., Jin X., Yan S., Feng J. (2017) Dual path networks. In: NIPS, pp. 4467–4475 [14, 23](#)
- [38] Chen Y., Rohrbach M., Yan Z., Yan S., Feng J., Kalantidis Y. (2019) Graph based global reasoning networks. In: CVPR [14](#)
- [39] Chen Y., Yang T., Zhang X., Meng G., Pan C., Sun J. (2019) DetNAS: Neural architecture search on object detection. arXiv:190310979 [26](#)
- [40] 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 [25](#)
- [41] 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 [25](#)
- [42] Cheng G., Zhou P., Han J. (2016) RIFDCNN: Rotation invariant and fisher discriminative convolutional neural networks for object detection. In: CVPR, pp. 2884–2893 [18](#)
- [43] Cheng M., Zhang Z., Lin W., Torr P. (2014) BING: Binarized normed gradients for objectness estimation at 300fps. In: CVPR, pp. 3286–3293 [20](#)
- [44] Cheng Y., Wang D., Zhou 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 [26](#)
- [45] Chollet F. (2017) Xception: Deep learning with depthwise separable convolutions. In: CVPR, pp. 1800–1807 [14, 24](#)
- [46] Cinbis R., Verbeek J., Schmid C. (2017) Weakly supervised object localization with multi-fold multiple instance learning. IEEE TPAMI 39(1):189–203 [10](#)
- [47] Csurka G., Dance C., Fan L., Willamowski J., Bray C. (2004) Visual categorization with bags of keypoints. In: ECCV Workshop on statistical learning in computer vision [2, 4](#)
- [48] Dai J., He K., Li Y., Ren S., Sun J. (2016) Instance sensitive fully convolutional networks. In: ECCV, pp. 534–549 [21, 22](#)
- [49] Dai J., He K., Sun J. (2016) Instance aware semantic segmentation via multitask network cascades. In: CVPR, pp. 3150–3158 [21](#)
- [50] Dai J., Li Y., He K., Sun J. (2016) RFCN: object detection via region based fully convolutional networks. In: NIPS, pp. 379–387 [8, 10, 11, 15, 20, 24, 32](#)
- [51] Dai J., Qi H., Xiong Y., Li Y., Zhang G., Hu H., Wei Y. (2017) Deformable convolutional networks. In: ICCV [15, 18, 23, 24, 25](#)
- [52] Dalal N., Triggs B. (2005) Histograms of oriented gradients for human detection. In: CVPR, vol 1, pp. 886–893 [2, 4, 8, 13, 20](#)
- [53] Demirel B., Cinbis R. G., Izkiler-Cinbis N. (2018) Zero shot object detection by hybrid region embedding. In: BMVC [26](#)
- [54] Deng J., Dong W., Socher R., Li L., Li K., Li F. (2009) ImageNet: A large scale hierarchical image database. In: CVPR, pp. 248–255 [5, 6, 14](#)
- [55] Diba A., Sharma V., Pazandeh A. M., Pirsiavash H., Van Gool L. (2017) Weakly supervised cascaded convolutional networks. In: CVPR, vol 3, p. 9 [26](#)
- [56] Dickinson S., Leonardis A., Schiele B., Tarr M. (2009) The Evolution of Object Categorization and the Challenge of Image Abstraction in *Object Categorization: Computer and Human Vision Perspectives*. Cambridge University Press [3, 13](#)
- [57] Ding J., Xue N., Long Y., Xia G., Lu Q. (2018) Learning RoI transformer for detecting oriented objects in aerial images. In: CVPR [18](#)
- [58] Divvala S., Hoiem D., Hays J., Efros A., Hebert M. (2009) An empirical study of context in object detection. In: CVPR, pp. 1271–1278 [18, 19, 25](#)
- [59] 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](#)
- [60] 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 [14](#)
- [61] Dong X., Zheng L., Ma F., Yang Y., Meng D. (2018) Few example object detection with model communication. IEEE TPAMI [26](#)
- [62] Duan K., Bai S., Xie L., Qi H., Huang Q., Tian Q. (2019) CenterNet: Keypoint triplets for object detection. arXiv preprint arXiv:190408189 [13](#)
- [63] Dvornik N., Mairal J., Schmid C. (2018) Modeling visual context is key to augmenting object detection datasets. In: ECCV, pp. 364–380 [22](#)
- [64] Dwibedi D., Misra I., Hebert M. (2017) Cut, paste and learn: Surprisingly easy synthesis for instance detection. In: ICCV, pp. 1301–1310 [22](#)
- [65] Endres I., Hoiem D. (2010) Category independent object proposals [20](#)
- [66] Enzweiler M., Gavrila D. M. (2009) Monocular pedestrian detection: Survey and experiments. IEEE TPAMI 31(12):2179–2195 [2, 3](#)
- [67] Erhan D., Szegedy C., Toshev A., Anguelov D. (2014) Scalable object detection using deep neural networks. In: CVPR, pp. 2147–2154 [9, 21, 22](#)
- [68] Everingham M., Gool L. V., Williams C., Winn J., Zisserman A. (2010)

- The pascal visual object classes (voc) challenge. IJCV 88(2):303–338 1, 2, 3, 4, 6, 7, 8, 21, 24, 26
- [69] Everingham M., Eslami S., Gool L. V., Williams C., Winn J., Zisserman A. (2015) The pascal visual object classes challenge: A retrospective. IJCV 111(1):98–136 6, 7, 24, 26
- [70] Feichtenhofer C., Pinz A., Zisserman A. (2017) Detect to track and track to detect. In: ICCV, pp. 918–927 26
- [71] FeiFei L., Fergus R., Perona P. (2006) One shot learning of object categories. IEEE TPAMI 28(4):594–611 26
- [72] Felzenszwalb P., McAllester D., Ramanan D. (2008) A discriminatively trained, multiscale, deformable part model. In: CVPR, pp. 1–8 8, 20
- [73] Felzenszwalb P., Girshick R., McAllester D. (2010) Cascade object detection with deformable part models. In: CVPR, pp. 2241–2248 11
- [74] Felzenszwalb P., Girshick R., McAllester D., Ramanan D. (2010) Object detection with discriminatively trained part based models. IEEE TPAMI 32(9):1627–1645 2, 8, 15, 18, 25
- [75] Finn C., Abbeel P., Levine S. (2017) Model agnostic meta learning for fast adaptation of deep networks. In: ICML, pp. 1126–1135 26
- [76] Fischler M., Elschlager R. (1973) The representation and matching of pictorial structures. IEEE Transactions on computers 100(1):67–92 1, 4
- [77] Fu C.-Y., Liu W., Ranga A., Tyagi A., Berg A. C. (2017) DSSD: Deconvolutional single shot detector. In: arXiv preprint arXiv:1701.06659 13, 15, 16, 17, 24
- [78] Galleguillos C., Belongie S. (2010) Context based object categorization: A critical survey. Computer Vision and Image Understanding 114:712–722 3, 18, 19, 20, 25
- [79] Geronimo D., Lopez A. M., Sappa A. D., Graf T. (2010) Survey of pedestrian detection for advanced driver assistance systems. IEEE TPAMI 32(7):1239–1258 2, 3
- [80] Ghiasi G., Lin T., Pang R., Le Q. (2019) NAS-FPN: learning scalable feature pyramid architecture for object detection. CVPR 26
- [81] Ghodrati A., Diba A., Pedersoli M., Tuytelaars T., Van Gool L. (2015) DeepProposal: Hunting objects by cascading deep convolutional layers. In: ICCV, pp. 2578–2586 21, 22
- [82] Gidaris S., Komodakis N. (2015) Object detection via a multiregion and semantic segmentation aware CNN model. In: ICCV, pp. 1134–1142 13, 19, 20, 23, 25
- [83] Gidaris S., Komodakis N. (2016) Attend refine repeat: Active box proposal generation via in out localization. In: BMVC 15, 23
- [84] Girshick R. (2015) Fast R-CNN. In: ICCV, pp. 1440–1448 2, 8, 9, 13, 14, 21, 22, 23, 24, 26, 32
- [85] Girshick R., Donahue J., Darrell T., Malik J. (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR, pp. 580–587 2, 5, 8, 9, 10, 11, 12, 13, 14, 15, 21, 22, 23, 24, 26, 32
- [86] Girshick R., Iandola F., Darrell T., Malik J. (2015) Deformable part models are convolutional neural networks. In: CVPR, pp. 437–446 18, 25
- [87] Girshick R., Donahue J., Darrell T., Malik J. (2016) Region-based convolutional networks for accurate object detection and segmentation. IEEE TPAMI 38(1):142–158 9, 10, 14
- [88] Goodfellow I., Shlens J., Szegedy C. (2015) Explaining and harnessing adversarial examples. In: ICLR 6
- [89] Goodfellow I., Bengio Y., Courville A. (2016) Deep Learning. MIT press 5
- [90] Grauman K., Darrell T. (2005) The pyramid match kernel: Discriminative classification with sets of image features. In: ICCV, vol 2, pp. 1458–1465 9
- [91] Grauman K., Leibe B. (2011) Visual object recognition. Synthesis lectures on artificial intelligence and machine learning 5(2):1–181 1, 2, 3
- [92] Gu J., Wang Z., Kuen J., Ma L., Shahroudny A., Shuai B., Liu T., Wang X., Wang G., Cai J., Chen T. (2017) Recent advances in convolutional neural networks. Pattern Recognition pp. 1–24 2, 3, 5, 13
- [93] Guillaumin M., Küttel D., Ferrari V. (2014) Imagenet autoannotation with segmentation propagation. International Journal of Computer Vision 110(3):328–348 20
- [94] Gupta A., Vedaldi A., Zisserman A. (2016) Synthetic data for text localisation in natural images. In: CVPR, pp. 2315–2324 22
- [95] Hariharan B., Girshick R. B. (2017) Low shot visual recognition by shrinking and hallucinating features. In: ICCV, pp. 3037–3046 26
- [96] Hariharan B., Arbeláez P., Girshick R., Malik J. (2014) Simultaneous detection and segmentation. In: ECCV, pp. 297–312 21
- [97] Hariharan B., Arbeláez P., Girshick R., Malik J. (2016) Object instance segmentation and fine grained localization using hypercolumns. IEEE TPAMI 10, 12, 15, 16
- [98] Harzallah H., Jurie F., Schmid C. (2009) Combining efficient object localization and image classification. In: ICCV, pp. 237–244 8, 20
- [99] He K., Zhang X., Ren S., Sun J. (2014) Spatial pyramid pooling in deep convolutional networks for visual recognition. In: ECCV, pp. 346–361 2, 9, 14, 15, 32
- [100] He K., Zhang X., Ren S., Sun J. (2015) Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: ICCV, pp. 1026–1034 12
- [101] He K., Zhang X., Ren S., Sun J. (2016) Deep residual learning for image recognition. In: CVPR, pp. 770–778 2, 11, 13, 14, 24, 25, 26
- [102] He K., Gkioxari G., Dollár P., Girshick R. (2017) Mask RCNN. In: ICCV 11, 13, 16, 19, 21, 23, 24, 25, 26, 32
- [103] He T., Tian Z., Huang W., Shen C., Qiao Y., Sun C. (2018) An end to end textspotter with explicit alignment and attention. In: CVPR, pp. 5020–5029 18
- [104] He Y., Zhu C., Wang J., Savvides M., Zhang X. (2019) Bounding box regression with uncertainty for accurate object detection. In: CVPR 23
- [105] Hinton G., Salakhutdinov R. (2006) Reducing the dimensionality of data with neural networks. science 313(5786):504–507 1
- [106] Hinton G., Vinyals O., Dean J. (2015) Distilling the knowledge in a neural network. arXiv:150302531 14, 24
- [107] Hoffman J., Guadarrama S., Tzeng E. S., Hu R., Donahue J., Girshick R., Darrell T., Saenko K. (2014) LSDA: large scale detection through adaptation. In: NIPS, pp. 3536–3544 26
- [108] Hoiem D., Chodpathumwan Y., Dai Q. (2012) Diagnosing error in object detectors. In: ECCV, pp. 340–353 7, 23
- [109] Hosang J., Omran M., Benenson R., Schiele B. (2015) Taking a deeper look at pedestrians. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4073–4082 2
- [110] Hosang J., Benenson R., Dollár P., Schiele B. (2016) What makes for effective detection proposals? IEEE TPAMI 38(4):814–829 9, 21, 26
- [111] Hosang J., Benenson R., Schiele B. (2017) Learning non-maximum suppression. In: CVPR 23
- [112] Howard A., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Andreetto M., Adam H. (2017) Movenets: Efficient convolutional neural networks for mobile vision applications. In: CVPR 14, 25, 26
- [113] Hu H., Lan S., Jiang Y., Cao Z., Sha F. (2017) FastMask: Segment multi-scale object candidates in one shot. In: CVPR, pp. 991–999 22
- [114] Hu H., Gu J., Zhang Z., Dai J., Wei Y. (2018) Relation networks for object detection. In: CVPR 19, 20, 25
- [115] Hu J., Shen L., Sun G. (2018) Squeeze and excitation networks. In: CVPR 13, 14
- [116] Hu P., Ramanan D. (2017) Finding tiny faces. In: CVPR, pp. 1522–1530 2
- [117] Hu R., Dollár P., He K., Darrell T., Girshick R. (2018) Learning to segment every thing. In: CVPR 26
- [118] Huang G., Liu Z., Weinberger K. Q., van der Maaten L. (2017) Densely connected convolutional networks. In: CVPR 13, 14, 16, 26
- [119] Huang G., Liu S., van der Maaten L., Weinberger K. (2018) CondenseNet: An efficient densenet using learned group convolutions. In: CVPR 26
- [120] Huang J., Rathod V., Sun C., Zhu M., Korattikara A., Fathi A., Fischer I., Wojna Z., Song Y., Guadarrama S., Murphy K. (2017) Speed/accuracy trade offs for modern convolutional object detectors. In: CVPR 14, 22, 24, 25
- [121] Huang Z., Huang L., Gong Y., Huang C., Wang X. (2019) Mask scoring rcnn. In: CVPR 23
- [122] Hubara I., Courbariaux M., Soudry D., ElYaniv R., Bengio Y. (2016) Binarized neural networks. In: NIPS, pp. 4107–4115 26
- [123] Iandola F., Han S., Moskewicz M., Ashraf K., Dally W., Keutzer K. (2016) SqueezeNet: Alexnet level accuracy with 50x fewer parameters and 0.5 mb model size. In: arXiv preprint arXiv:1602.07360 25
- [124] ILSVRC detection challenge results (2018) <http://www.image-net.org/challenges/LSVRC/> 24
- [125] Ioffe S., Szegedy C. (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning, pp. 448–456 13, 14, 24
- [126] Jaderberg M., Simonyan K., Zisserman A., et al. (2015) Spatial transformer networks. In: NIPS, pp. 2017–2025 18
- [127] Jia Y., Shelhamer E., Donahue J., Karayev S., Long J., Girshick R., Guadarrama S., Darrell T. (2014) Caffe: Convolutional architecture for fast feature embedding. In: ACM MM, pp. 675–678 16
- [128] Jiang B., Luo R., Mao J., Xiao T., Jiang Y. (2018) Acquisition of localization confidence for accurate object detection. In: ECCV, pp. 784–799 23
- [129] Kang B., Liu Z., Wang X., Yu F., Feng J., Darrell T. (2018) Few shot object detection via feature reweighting. arXiv preprint arXiv:181201866 26
- [130] Kang K., Ouyang W., Li H., Wang X. (2016) Object detection from video tubelets with convolutional neural networks. In: CVPR, pp. 817–825 26
- [131] Kim A., Sharma A., Jacobs D. (2014) Locally scale invariant convolutional neural networks. In: NIPS 18
- [132] Kim K., Hong S., Roh B., Cheon Y., Park M. (2016) PVANet: Deep but lightweight neural networks for real time object detection. In: NIPS/W 15
- [133] Kim Y., Kang B.-N., Kim D. (2018) SAN: learning relationship between convolutional features for multiscale object detection. In: ECCV, pp. 316–331 16
- [134] Kirillov A., He K., Girshick R., Rother C., Dollár P. (2018) Panoptic segmentation. arXiv:180100868 25
- [135] Kong T., Yao A., Chen Y., Sun F. (2016) HyperNet: towards accurate region proposal generation and joint object detection. In: CVPR, pp. 845–853 15, 16, 21
- [136] Kong T., Sun F., Yao A., Liu H., Lu M., Chen Y. (2017) RON: Reverse connection with objectness prior networks for object detection. In: CVPR 15, 16, 17
- [137] Kong T., Sun F., Tan C., Liu H., Huang W. (2018) Deep feature pyramid reconfiguration for object detection. In: ECCV, pp. 169–185 15, 16, 17, 18
- [138] Krähenbühl P., Koltun V. (2014) Geodesic object proposals. In: ECCV 20
- [139] Krasin I., Duerig T., Alldrin N., Ferrari V., AbuElhaija S., Kuznetsova A.,

- Rom H., Uijlings J., Popov S., Kamali S., Malloci M., PontTuset J., Veit A., Belongie S., Gomes V., Gupta A., Sun C., Chechik G., Cai D., Feng Z., Narayanan D., Murphy K. (2017) OpenImages: A public dataset for large scale multilabel and multiclass image classification. Dataset available from <https://storage.googleapis.com/openimages/web/indexhtml> 24
- [140] Krizhevsky A., Sutskever I., Hinton G. (2012) ImageNet classification with deep convolutional neural networks. In: NIPS, pp. 1097–1105 1, 2, 4, 5, 9, 11, 12, 16, 21, 24
- [141] Krizhevsky A., Sutskever I., Hinton G. (2012) ImageNet classification with deep convolutional neural networks. In: NIPS, pp. 1097–1105 13, 14, 26
- [142] Kuo W., Hariharan B., Malik J. (2015) DeepBox: Learning objectness with convolutional networks. In: ICCV, pp. 2479–2487 21, 22
- [143] Kuznetsova A., Rom H., Alldrin N., Uijlings J., Krasin I., PontTuset J., Kamali S., Popov S., Malloci M., Duerig T., et al. (2018) The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. arXiv preprint arXiv:181100982 6, 7, 8
- [144] Lake B., Salakhutdinov R., Tenenbaum J. (2015) Human level concept learning through probabilistic program induction. Science 350(6266):1332–1338 26
- [145] Lampert C. H., Blaschko M. B., Hofmann T. (2008) Beyond sliding windows: Object localization by efficient subwindow search. In: CVPR, pp. 1–8 8
- [146] Law H., Deng J. (2018) CornerNet: Detecting objects as paired keypoints. In: ECCV 13, 14, 25
- [147] Lazebnik S., Schmid C., Ponce J. (2006) Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR, vol. 2, pp. 2169–2178 2, 4, 9
- [148] LeCun Y., Bottou L., Bengio Y., Haffner P. (1998) Gradient based learning applied to document recognition. Proceedings of the IEEE 86(11):2278–2324 2
- [149] LeCun Y., Bengio Y., Hinton G. (2015) Deep learning. Nature 521:436–444 1, 2, 3, 5, 6, 13
- [150] Lee C., Xie S., Gallagher P., Zhang Z., Tu Z. (2015) Deeply supervised nets. In: Artificial Intelligence and Statistics, pp. 562–570 13
- [151] Lenc K., Vedaldi A. (2015) R-CNN minus R. In: BMVC15 10, 32
- [152] Lenc K., Vedaldi A. (2018) Understanding image representations by measuring their equivariance and equivalence. IJCV 18
- [153] Li B., Liu Y., Wang X. (2019) Gradient harmonized single stage detector. In: AAAI 23, 24
- [154] Li H., Lin Z., Shen X., Brandt J., Hua G. (2015) A convolutional neural network cascade for face detection. In: CVPR, pp. 5325–5334 2
- [155] Li H., Kadav A., Durdanovic I., Samet H., Graf H. P. (2017) Pruning filters for efficient convnets. In: ICLR 26
- [156] Li H., Liu Y., Ouyang W., Xiaogang Wang (2018) Zoom out and in network with map attention decision for region proposal and object detection. IJCV 15, 16, 17, 21, 22
- [157] Li J., Wei Y., Liang X., Dong J., Xu T., Feng J., Yan S. (2017) Attentive contexts for object detection. IEEE Transactions on Multimedia 19(5):944–954 19, 20
- [158] Li Q., Jin S., Yan J. (2017) Mimicking very efficient network for object detection. In: CVPR, pp. 7341–7349 26
- [159] Li S. Z., Zhang Z. (2004) Floatboost learning and statistical face detection. IEEE TPAMI 26(9):1112–1123 11
- [160] Li Y., Wang S., Tian Q., Ding X. (2015) Feature representation for statistical learning based object detection: A review. Pattern Recognition 48(11):3542–3559 3
- [161] Li Y., Ouyang W., Zhou B., Wang K., Wang X. (2017) Scene graph generation from objects, phrases and region captions. In: ICCV, pp. 1261–1270 25
- [162] Li Y., Qi H., Dai J., Ji X., Wei Y. (2017) Fully convolutional instance aware semantic segmentation. In: CVPR, pp. 4438–4446 21
- [163] Li Y., Chen Y., Wang N., Zhang Z. (2019) Scale aware trident networks for object detection. arXiv preprint arXiv:190101892 16, 25
- [164] Li Z., Peng C., Yu G., Zhang X., Deng Y., Sun J. (2018) DetNet: A backbone network for object detection. In: ECCV 14, 15, 16, 17, 18, 25
- [165] Li Z., Peng C., Yu G., Zhang X., Deng Y., Sun J. (2018) Light head RCNN: In defense of two stage object detector. In: CVPR 11, 24
- [166] Lin T., Maire M., Belongie S., Hays J., Perona P., Ramanan D., Dollár P., Zitnick L. (2014) Microsoft COCO: Common objects in context. In: ECCV, pp. 740–755 2, 3, 4, 5, 6, 7, 21, 24, 26
- [167] Lin T., Dollár P., Girshick R., He K., Hariharan B., Belongie S. (2017) Feature pyramid networks for object detection. In: CVPR 11, 15, 16, 17, 18, 24, 25
- [168] Lin T., Goyal P., Girshick R., He K., Dollár P. (2017) Focal loss for dense object detection. In: ICCV 13, 16, 23, 24
- [169] Lin X., Zhao C., Pan W. (2017) Towards accurate binary convolutional neural network. In: NIPS, pp. 344–352 26
- [170] Litjens G., Kooi T., Bejnordi B., Setio A., Ciompi F., Ghafoorian M., J. van der Laak B. v., Sánchez C. (2017) A survey on deep learning in medical image analysis. Medical Image Analysis 42:60–88 2, 3, 5
- [171] Liu C., Zoph B., Neumann M., Shlens J., Huo W., Li L., FeiFei L., Yuille A., Huang J., Murphy K. (2018) Progressive neural architecture search. In: ECCV, pp. 19–34 26
- [172] Liu L., Fieguth P., Guo Y., Wang X., Pietikäinen M. (2017) Local binary features for texture classification: Taxonomy and experimental study. Pattern Recognition 62:135–160 18
- [173] Liu S., Huang D., Wang Y. (2018) Receptive field block net for accurate and fast object detection. In: ECCV 15, 16
- [174] Liu S., Qi L., Qin H., Shi J., Jia J. (2018) Path aggregation network for instance segmentation. In: CVPR, pp. 8759–8768 15, 16, 17
- [175] Liu W., Anguelov D., Erhan D., Szegedy C., Reed S., Fu C., Berg A. (2016) SSD: single shot multibox detector. In: ECCV, pp. 21–37 12, 13, 16, 19, 21, 22, 23, 24, 25, 26, 32
- [176] Liu Y., Wang R., Shan S., Chen X. (2018) Structure Inference Net: Object detection using scene level context and instance level relationships. In: CVPR, pp. 6985–6994 19, 20
- [177] Long J., Shelhamer E., Darrell T. (2015) Fully convolutional networks for semantic segmentation. In: CVPR, pp. 3431–3440 10, 11, 15, 16, 21, 22
- [178] Lowe D. (1999) Object recognition from local scale invariant features. In: ICCV, vol 2, pp. 1150–1157 2, 4, 13
- [179] Lowe D. (2004) Distinctive image features from scale-invariant keypoints. IJCV 60(2):91–110 2, 4, 20
- [180] Loy C., Lin D., Ouyang W., Xiong Y., Yang S., Huang Q., Zhou D., Xia W., Li Q., Luo P., et al. (2019) WIDER face and pedestrian challenge 2018: Methods and results. arXiv:190206854 25
- [181] Lu Y., Javidi T., Lazebnik S. (2016) Adaptive object detection using adjacency and zoom prediction. In: CVPR, pp. 2351–2359 21, 22
- [182] Luo P., Wang X., Shao W., Peng Z. (2018) Towards understanding regularization in batch normalization. In: ICLR 25
- [183] Luo P., Ren J., Peng Z., Zhang R., Li J. (2019) Switchable normalization for learning to normalize deep representation. IEEE TPAMI 25
- [184] Ma J., Shao W., Ye H., Wang L., Wang H., Zheng Y., Xue X. (2018) Arbitrary oriented scene text detection via rotation proposals. IEEE TMM 20(11):3111–3122 18
- [185] Malisiewicz T., Efros A. (2009) Beyond categories: The visual memex model for reasoning about object relationships. In: NIPS 19, 25
- [186] Manen S., Guillaumin M., Van Gool L. (2013) Prime object proposals with randomized prim's algorithm. In: CVPR, pp. 2536–2543 20
- [187] Mikolajczyk K., Schmid C. (2005) A performance evaluation of local descriptors. IEEE TPAMI 27(10):1615–1630 4
- [188] Mordan T., Thome N., Henaff G., Cord M. (2018) End to end learning of latent deformable part based representations for object detection. IJCV pp. 1–21 15, 18
- [189] MS COCO detection leaderboard (2018) <http://cocodataset.org/#detection-leaderboard> 24
- [190] Mundy J. (2006) Object recognition in the geometric era: A retrospective. in book Toward Category Level Object Recognition edited by J Ponce, M Hebert, C Schmid and A Zisserman pp. 3–28 4
- [191] Murase H., Nayar S. (1995) Visual learning and recognition of 3D objects from appearance. IJCV 14(1):5–24 4
- [192] Murase H., Nayar S. (1995) Visual learning and recognition of 3d objects from appearance. IJCV 14(1):5–24 4
- [193] Murphy K., Torralba A., Freeman W. (2003) Using the forest to see the trees: a graphical model relating features, objects and scenes. In: NIPS 19, 25
- [194] Newell A., Yang K., Deng J. (2016) Stacked hourglass networks for human pose estimation. In: ECCV, pp. 483–499 13, 16
- [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 4, 20
- [197] Oliva A., Torralba A. (2007) The role of context in object recognition. Trends in cognitive sciences 11(12):520–527 18, 25
- [198] Opelt A., Pinz A., Fussenegger M., Auer P. (2006) Generic object recognition with boosting. IEEE TPAMI 28(3):416–431 3
- [199] Oquab M., Bottou L., Laptev I., Sivic J. (2014) Learning and transferring midlevel image representations using convolutional neural networks. In: CVPR, pp. 1717–1724 6
- [200] Oquab M., Bottou L., Laptev I., Sivic J. (2015) Is object localization for free? weakly supervised learning with convolutional neural networks. In: CVPR, pp. 685–694 10
- [201] Osuna E., Freund R., Girosi F. (1997) Training support vector machines: an application to face detection. In: CVPR, pp. 130–136 4
- [202] Ouyang W., Wang X. (2013) Joint deep learning for pedestrian detection. In: ICCV, pp. 2056–2063 18
- [203] Ouyang W., Wang X., Zeng X., Qiu S., Luo P., Tian Y., Li H., Yang S., Wang Z., Loy C.-C., et al. (2015) DeepIDNet: Deformable deep convolutional neural networks for object detection. In: CVPR, pp. 2403–2412 8, 15, 18, 19, 20, 21, 24, 25
- [204] Ouyang W., Wang X., Zhang C., Yang X. (2016) Factors in finetuning deep model for object detection with long tail distribution. In: CVPR, pp. 864–873 23
- [205] Ouyang W., Wang K., Zhu X., Wang X. (2017) Chained cascade network for object detection. ICCV 11, 23
- [206] Ouyang W., Zeng X., Wang X., Qiu S., Luo P., Tian Y., Li H., Yang S., Wang Z., Li H., Wang K., Yan J., Loy C. C., Tang X. (2017) DeepIDNet: Object detection with deformable part based convolutional neural networks. IEEE TPAMI 39(7):1320–1334 14, 18
- [207] Parikh D., Zitnick C., Chen T. (2012) Exploring tiny images: The roles of appearance and contextual information for machine and human object recognition. IEEE TPAMI 34(10):1978–1991 19
- [208] PASCAL VOC detection leaderboard (2018) <http://host>.

- robots.ox.ac.uk:8080/leaderboard/_main_bootstrap.php 24
- [209] Peng C., Xiao T., Li Z., Jiang Y., Zhang X., Jia K., Yu G., Sun J. (2018) MegDet: A large minibatch object detector. In: CVPR 22, 23, 24
- [210] Peng X., Sun B., Ali K., Saenko K. (2015) Learning deep object detectors from 3d models. In: ICCV, pp. 1278–1286 22
- [211] Pepik B., Benenson R., Ritschel T., Schiele B. (2015) What is holding back convnets for detection? In: German Conference on Pattern Recognition, pp. 517–528 26
- [212] Perronnin F., Sánchez J., Mensink T. (2010) Improving the fisher kernel for large scale image classification. In: ECCV, pp. 143–156 2, 4, 13
- [213] Pinheiro P., Collobert R., Dollar P. (2015) Learning to segment object candidates. In: NIPS, pp. 1990–1998 21, 22
- [214] Pinheiro P., Lin T., Collobert R., Dollár P. (2016) Learning to refine object segments. In: ECCV, pp. 75–91 15, 16, 21, 22
- [215] Ponce J., Hebert M., Schmid C., Zisserman A. (2007) Toward Category Level Object Recognition. Springer 3, 4
- [216] Pouyanfar S., Sadiq S., Yan Y., Tian H., Tao Y., Reyes M. P., Shyu M., Chen S., Iyengar S. (2018) A survey on deep learning: Algorithms, techniques, and applications. ACM Computing Surveys 51(5):92:1–92:36 5
- [217] Qi C. R., Su H., Mo K., Guibas L. J. (2017) PointNet: Deep learning on point sets for 3D classification and segmentation. In: CVPR, pp. 652–660 26
- [218] Qi C. R., Liu W., Wu C., Su H., Guibas L. J. (2018) Frustum pointnets for 3D object detection from RGBD data. In: CVPR, pp. 918–927 26
- [219] Quanming Y., Mengshuo W., Hugo J. E., Isabelle G., Yiqi H., Yufeng L., Weiwei T., Qiang Y., Yang Y. (2018) Taking human out of learning applications: A survey on automated machine learning. arXiv:181013306 26
- [220] Rabinovich A., Vedaldi A., Galleguillos C., Wiewiora E., Belongie S. (2007) Objects in context. In: ICCV 19, 20, 25
- [221] Rahman S., Khan S., Barnes N. (2018) Polarity loss for zero shot object detection. arXiv preprint arXiv:181108982 26
- [222] Rahman S., Khan S., Porikli F. (2018) Zero shot object detection: Learning to simultaneously recognize and localize novel concepts. In: ACCV 26
- [223] Razavian R., Azizpour H., Sullivan J., Carlsson S. (2014) CNN features off the shelf: an astounding baseline for recognition. In: CVPR Workshops, pp. 806–813 14
- [224] Rebuffi S., Bilen H., Vedaldi A. (2017) Learning multiple visual domains with residual adapters. In: Advances in Neural Information Processing Systems, pp. 506–516 26
- [225] Rebuffi S., Bilen H., Vedaldi A. (2018) Efficient parametrization of multidomain deep neural networks. In: CVPR, pp. 8119–8127 26
- [226] Redmon J., Farhadi A. (2017) YOLO9000: Better, faster, stronger. In: CVPR 12, 14, 26, 32
- [227] Redmon J., Divvala S., Girshick R., Farhadi A. (2016) You only look once: Unified, real time object detection. In: CVPR, pp. 779–788 12, 13, 14, 23, 24, 25, 26, 32
- [228] Ren M., Triantafillou E., Ravi S., Snell J., Swersky K., Tenenbaum J. B., Larochelle H., Zemel R. S. (2018) Meta learning for semisupervised few shot classification. In: ICLR 26
- [229] 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 8, 10, 11, 12, 13, 14, 19, 21, 22, 23, 24, 25, 26, 32
- [230] Ren S., He K., Girshick R., Sun J. (2017) Faster RCNN: Towards real time object detection with region proposal networks. IEEE TPAMI 39(6):1137–1149 2, 10, 22, 24
- [231] Ren S., He K., Girshick R., Zhang X., Sun J. (2017) Object detection networks on convolutional feature maps. IEEE TPAMI 24
- [232] Rezatofighi H., Tsoi N., Gwak J., Sadeghian A., Reid I., Savarese S. (2019) Generalized intersection over union: A metric and a loss for bounding box regression. In: CVPR 23
- [233] Rowley H., Baluja S., Kanade T. (1998) Neural network based face detection. IEEE TPAMI 20(1):23–38 4
- [234] Russakovsky O., Deng J., Su H., Krause J., Satheesh S., Ma S., Huang Z., Karpathy A., Khosla A., Bernstein M., Berg A., Li F. (2015) ImageNet large scale visual recognition challenge. IJCV 115(3):211–252 1, 2, 3, 4, 5, 6, 7, 8, 14, 21, 24, 26
- [235] Russell B., Torralba A., Murphy K., Freeman W. (2008) LabelMe: A database and web based tool for image annotation. IJCV 77(1-3):157–173 3
- [236] Schmid C., Mohr R. (1997) Local grayvalue invariants for image retrieval. IEEE TPAMI 19(5):530–535 4
- [237] Schwartz E., Karlinsky L., Shtok J., Harary S., Marder M., Pankanti S., Feris R., Kumar A., Gries R., Bronstein A. (2019) RepMet: Representative based metric learning for classification and one shot object detection. In: CVPR 26
- [238] Sermanet P., Kavukcuoglu K., Chintala S., LeCun Y. (2013) Pedestrian detection with unsupervised multistage feature learning. In: CVPR, pp. 3626–3633 4, 20
- [239] Sermanet P., Eigen D., Zhang X., Mathieu M., Fergus R., LeCun Y. (2014) OverFeat: Integrated recognition, localization and detection using convolutional networks. In: ICLR 2, 9, 12, 14, 21, 32
- [240] Shang W., Sohn K., Almeida D., Lee H. (2016) Understanding and improving convolutional neural networks via concatenated rectified linear units. In: ICML, pp. 2217–2225 15
- [241] Shelhamer E., Long J., Darrell T. (2017) Fully convolutional networks for semantic segmentation. IEEE TPAMI 10, 11, 16
- [242] Shen Z., Liu Z., Li J., Jiang Y., Chen Y., Xue X. (2017) DSOD: Learning deeply supervised object detectors from scratch. In: ICCV 15, 16
- [243] Shi X., Shan S., Kan M., Wu S., Chen X. (2018) Real time rotation invariant face detection with progressive calibration networks. In: CVPR 18
- [244] Shi Z., Yang Y., Hospedales T., Xiang T. (2017) Weakly supervised image annotation and segmentation with objects and attributes. IEEE TPAMI 39(12):2525–2538 26
- [245] Shrivastava A., Gupta A. (2016) Contextual priming and feedback for Faster RCNN. In: ECCV, pp. 330–348 19, 20
- [246] Shrivastava A., Gupta A., Girshick R. (2016) Training region based object detectors with online hard example mining. In: CVPR, pp. 761–769 23
- [247] Shrivastava A., Sukthankar R., Malik J., Gupta A. (2016) Beyond skip connections: Top down modulation for object detection. arXiv:161206851 15, 16, 17, 24
- [248] Simonyan K., Zisserman A. (2015) Very deep convolutional networks for large scale image recognition. In: ICLR 2, 5, 9, 10, 12, 13, 14, 24
- [249] Singh B., Davis L. (2018) An analysis of scale invariance in object detection-SNIP. In: CVPR 7, 22, 23, 24
- [250] Singh B., Li H., Sharma A., Davis L. S. (2018) RFCN 3000 at 30fps: Decoupling detection and classification. In: CVPR 26
- [251] Singh B., Najibi M., Davis L. S. (2018) SNIPER: Efficient multiscale training. arXiv:180509300 22, 23, 24
- [252] Sivic J., Zisserman A. (2003) Video google: A text retrieval approach to object matching in videos. In: International Conference on Computer Vision (ICCV), vol 2, pp. 1470–1477 2, 4, 13
- [253] Song Han W. J. D. Huizi Mao (2016) Deep Compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. In: ICLR 26
- [254] Sun C., Shrivastava A., Singh S., Gupta A. (2017) Revisiting unreasonable effectiveness of data in deep learning era. In: ICCV, pp. 843–852 14
- [255] Sun K., Xiao B., Liu D., Wang J. (2019) Deep high resolution representation learning for human pose estimation. In: CVPR 25
- [256] Sun K., Zhao Y., Jiang B., Cheng T., Xiao B., Liu D., Mu Y., Wang X., Liu W., Wang J. (2019) High resolution representations for labeling pixels and regions. CoRR abs/1904.04514 25
- [257] Sun S., Pang J., Shi J., Yi S., Ouyang W. (2018) FishNet: A versatile backbone for image, region, and pixel level prediction. In: NIPS, pp. 754–764 14
- [258] Sun Z., Bebis G., Miller R. (2006) On road vehicle detection: A review. IEEE TPAMI 28(5):694–711 2, 3
- [259] Sung K., Poggio T. (1994) Learning and example selection for object and pattern detection. MIT AI Memo (1521) 23
- [260] Swain M., Ballard D. (1991) Color indexing. IJCV 7(1):11–32 4
- [261] Szegedy C., Toshev A., Erhan D. (2013) Deep neural networks for object detection. In: NIPS, pp. 2553–2561 9, 11
- [262] Szegedy C., Reed S., Erhan D., Anguelov D., Ioffe S. (2014) Scalable, high quality object detection. In: arXiv preprint arXiv:1412.1441 21, 22
- [263] Szegedy C., Liu W., Jia Y., Sermanet P., Reed S., Anguelov D., Erhan D., Vanhoucke V., Rabinovich A. (2015) Going deeper with convolutions. In: CVPR, pp. 1–9 2, 12, 13, 14, 15, 16, 24
- [264] Szegedy C., Vanhoucke V., Ioffe S., Shlens J., Wojna Z. (2016) Rethinking the inception architecture for computer vision. In: CVPR, pp. 2818–2826 13, 14, 24
- [265] Szegedy C., Ioffe S., Vanhoucke V., Alemi A. (2017) Inception v4, inception resnet and the impact of residual connections on learning. AAAI pp. 4278–4284 13, 14, 25
- [266] Torralba A. (2003) Contextual priming for object detection. IJCV 53(2):169–191 18
- [267] Turk M. A., Pentland A. (1991) Face recognition using eigenfaces. In: CVPR, pp. 586–591 4
- [268] Tuzel O., Porikli F., Meer P. (2006) Region covariance: A fast descriptor for detection and classification. In: ECCV, pp. 589–600 4
- [269] TychsenSmith L., Petersson L. (2017) DeNet: scalable real time object detection with directed sparse sampling. In: ICCV 13, 21, 22
- [270] TychsenSmith L., Petersson L. (2018) Improving object localization with fitness nms and bounded iou loss. In: CVPR 23
- [271] Uijlings J., van de Sande K., Gevers T., Smeulders A. (2013) Selective search for object recognition. IJCV 104(2):154–171 2, 8, 9, 20, 21
- [272] Vaillant R., Monrocq C., LeCun Y. (1994) Original approach for the localisation of objects in images. IEE Proceedings Vision, Image and Signal Processing 141(4):245–250 4
- [273] Van de Sande K., Uijlings J., Gevers T., Smeulders A. (2011) Segmentation as selective search for object recognition. In: ICCV, pp. 1879–1886 20, 21
- [274] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. (2017) Attention is all you need. In: NIPS, pp. 6000–6010 20
- [275] Vedaldi A., Gulshan V., Varma M., Zisserman A. (2009) Multiple kernels for object detection. In: ICCV, pp. 606–613 8, 20
- [276] Viola P., Jones M. (2001) Rapid object detection using a boosted cascade of simple features. In: CVPR, vol 1, pp. 1–8 2, 4, 8, 20
- [277] Wan L., Eigen D., Fergus R. (2015) End to end integration of a convolution network, deformable parts model and non-maximum suppression.

- In: CVPR, pp. 851–859 18, 25
- [278] Wang H., Wang Q., Gao M., Li P., Zuo W. (2018) Multiscale location aware kernel representation for object detection. In: CVPR 16
- [279] Wang X., Han T., Yan S. (2009) An HOG-LBP human detector with partial occlusion handling. In: International Conference on Computer Vision, pp. 32–39 2
- [280] Wang X., Shrivastava A., Gupta A. (2017) A Fast RCNN: Hard positive generation via adversary for object detection. In: CVPR 18, 22
- [281] Wang X., Cai Z., Gao D., Vasconcelos N. (2019) Towards universal object detection by domain attention. arXiv:190404402 26
- [282] Wei Y., Pan X., Qin H., Ouyang W., Yan J. (2018) Quantization mimic: Towards very tiny CNN for object detection. In: ECCV, pp. 267–283 26
- [283] Woo S., Hwang S., Kweon I. (2018) StairNet: Top down semantic aggregation for accurate one shot detection. In: WACV, pp. 1093–1102 16
- [284] Worrall D. E., Garbin S. J., Turmukhambetov D., Brostow G. J. (2017) Harmonic networks: Deep translation and rotation equivariance. In: CVPR, vol 2 18
- [285] Wu Y., He K. (2018) Group normalization. In: ECCV, pp. 3–19 25
- [286] Wu Z., Song S., Khosla A., Yu F., Zhang L., Tang X., Xiao J. (2015) 3D ShapeNets: A deep representation for volumetric shapes. In: CVPR, pp. 1912–1920 26
- [287] Wu Z., Pan S., Chen F., Long G., Zhang C., Yu P. S. (2019) A comprehensive survey on graph neural networks. arXiv preprint arXiv:190100596 5
- [288] Xia G., Bai X., Ding J., Zhu Z., Belongie S., Luo J., Dateu M., Pelillo M., Zhang L. (2018) DOTA: a large-scale dataset for object detection in aerial images. In: CVPR, pp. 3974–3983 18
- [289] Xiang Y., Mottaghi R., Savarese S. (2014) Beyond PASCAL: A benchmark for 3D object detection in the wild. In: WACV, pp. 75–82 26
- [290] Xiao R., Zhu L., Zhang H. (2003) Boosting chain learning for object detection. In: ICCV, pp. 709–715 4
- [291] Xie S., Girshick R., Dollár P., Tu Z., He K. (2017) Aggregated residual transformations for deep neural networks. In: CVPR 11, 14, 24, 25
- [292] Yang B., Yan J., Lei Z., Li S. (2016) CRAFT objects from images. In: CVPR, pp. 6043–6051 19, 21, 22, 23
- [293] Yang F., Choi W., Lin Y. (2016) Exploit all the layers: Fast and accurate CNN object detector with scale dependent pooling and cascaded rejection classifiers. In: CVPR, pp. 2129–2137 15
- [294] Yang M., Kriegman D., Ahuja N. (2002) Detecting faces in images: A survey. IEEE TPAMI 24(1):34–58 2, 3
- [295] Ye Q., Doermann D. (2015) Text detection and recognition in imagery: A survey. IEEE TPAMI 37(7):1480–1500 2, 3
- [296] Yosinski J., Clune J., Bengio Y., Lipson H. (2014) How transferable are features in deep neural networks? In: NIPS, pp. 3320–3328 14
- [297] Young T., Hazarika D., Poria S., Cambria E. (2018) Recent trends in deep learning based natural language processing. IEEE Computational Intelligence Magazine 13(3):55–75 5
- [298] Yu F., Koltun V. (2016) Multiscale context aggregation by dilated convolutions 15
- [299] Yu F., Koltun V., Funkhouser T. (2017) Dilated residual networks. In: CVPR, vol 2, p. 3 14
- [300] Yu R., Li A., Chen C., Lai J., et al. (2018) NISP: Pruning networks using neuron importance score propagation. CVPR 26
- [301] Zafeiriou S., Zhang C., Zhang Z. (2015) A survey on face detection in the wild: Past, present and future. Computer Vision and Image Understanding 138:1–24 2, 3
- [302] Zagoruyko S., Lerer A., Lin T., Pinheiro P., Gross S., Chintala S., Dollár P. (2016) A multipath network for object detection. In: BMVC 15, 20, 22
- [303] Zeiler M., Fergus R. (2014) Visualizing and understanding convolutional networks. In: ECCV, pp. 818–833 6, 13, 14, 19
- [304] Zeng X., Ouyang W., Yang B., Yan J., Wang X. (2016) Gated bidirectional cnn for object detection. In: ECCV, pp. 354–369 19, 20, 25
- [305] Zeng X., Ouyang W., Yan J., Li H., Xiao T., Wang K., Liu Y., Zhou Y., Yang B., Wang Z., Zhou H., Wang X. (2017) Crafting gbdnet for object detection. IEEE TPAMI 19, 20, 21, 25
- [306] Zhang K., Zhang Z., Li Z., Qiao Y. (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE SPL 23(10):1499–1503 2
- [307] Zhang L., Lin L., Liang X., He K. (2016) Is faster RCNN doing well for pedestrian detection? In: ECCV, pp. 443–457 2
- [308] Zhang S., Wen L., Bian X., Lei Z., Li S. (2018) Single shot refinement neural network for object detection. In: CVPR 15, 16, 17
- [309] Zhang S., Yang J., Schiele B. (2018) Occluded pedestrian detection through guided attention in CNNs. In: CVPR, pp. 2056–2063 18
- [310] Zhang X., Yang Y., Han Z., Wang H., Gao C. (2013) Object class detection: A survey. ACM Computing Surveys 46(1):10:1–10:53 1, 2, 3, 20
- [311] 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 19
- [312] Zhang X., Zhou X., Lin M., Sun J. (2018) ShuffleNet: an extremely efficient convolutional neural network for mobile devices. In: CVPR 25
- [313] Zhang Z., Geiger J., Pohjalainen J., Mousa A. E., Jin W., Schuller B. (2018) Deep learning for environmentally robust speech recognition: An overview of recent developments. ACM Trans Intell Syst Technol 9(5):49:1–49:28 5
- [314] Zhang Z., Qiao S., Xie C., Shen W., Wang B., Yuille A. (2018) Single shot object detection with enriched semantics. In: CVPR 16
- [315] 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 network. In: AAAI 15, 16, 17, 18
- [316] Zheng S., Jayasumana S., Romera-Paredes B., Vineet V., Su Z., Du D., Huang C., Torr P. (2015) Conditional random fields as recurrent neural networks. In: ICCV, pp. 1529–1537 19
- [317] Zhou B., Khosla A., Lapedriza A., Oliva A., Torralba A. (2015) Object detectors emerge in deep scene CNNs. In: ICLR 10, 14
- [318] Zhou B., Khosla A., Lapedriza A., Oliva A., Torralba A. (2016) Learning deep features for discriminative localization. In: CVPR, pp. 2921–2929 10
- [319] Zhou B., Lapedriza A., Khosla A., Oliva A., Torralba A. (2017) Places: A 10 million image database for scene recognition. IEEE Trans Pattern Analysis and Machine Intelligence 7, 14, 24
- [320] Zhou J., Cui G., Zhang Z., Yang C., Liu Z., Sun M. (2018) Graph neural networks: A review of methods and applications. arXiv preprint arXiv:181208434 5
- [321] Zhou P., Ni B., Geng C., Hu J., Xu Y. (2018) Scale transferrable object detection. In: CVPR 14, 15, 16
- [322] Zhou Y., Liu L., Shao L., Mellor M. (2016) DAVE: A unified framework for fast vehicle detection and annotation. In: ECCV, pp. 278–293 2
- [323] Zhou Y., Ye Q., Qiu Q., Jiao J. (2017) Oriented response networks. In: CVPR, pp. 4961–4970 18
- [324] Zhu X., Vondrick C., Fowlkes C., Ramanan D. (2016) Do we need more training data? IJCV 119(1):76–92 13
- [325] Zhu X., Tuia D., Mou L., Xia G., Zhang L., Xu F., Fraundorfer F. (2017) Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geoscience and Remote Sensing Magazine 5(4):8–36 5
- [326] Zhu Y., Urtasun R., Salakhutdinov R., Fidler S. (2015) SegDeepM: Exploiting segmentation and context in deep neural networks for object detection. In: CVPR, pp. 4703–4711 19, 20
- [327] 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 19, 20
- [328] Zhu Y., Zhou Y., Ye Q., Qiu Q., Jiao J. (2017) Soft proposal networks for weakly supervised object localization. In: ICCV, pp. 1841–1850 20
- [329] Zhu Z., Liang D., Zhang S., Huang X., Li B., Hu S. (2016) Traffic sign detection and classification in the wild. In: CVPR, pp. 2110–2118 2
- [330] Zitnick C., Dollár P. (2014) Edge boxes: Locating object proposals from edges. In: ECCV, pp. 391–405 20, 21
- [331] Zoph B., Le Q. (2017) Neural architecture search with reinforcement learning 26
- [332] Zoph B., Vasudevan V., Shlens J., Le Q. (2018) Learning transferable architectures for scalable image recognition. In: CVPR, pp. 8697–8710 26

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 項を参照されたい。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 trainval, "07T": VOC2007 trainval and test, "12": VOC2012 trainval, "CO": COCO 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。RP: Region Proposal, SS: Selective Search, RPN: Region Proposal Network. RCNN は “RCNN minus R” を意味し、些細な領域提案の手法を使用している。

	検出器名	RP	Backbone DCNN	入力 画像サイズ	VOC07 結果	VOC12 結果	速度 (FPS)	発表先	コード	ソース	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 と領域提案の手法を組合。従来の最先端技術から劇的な性能向上。 欠点: 順次訓練される多段階のバイオライン(外部の領域提案計算, CNN finetuning, ワーピングされた各領域提案の CNN 通過, SVM と bounding box 回帰器の訓練)。空間計算量・時間計算量の両面で訓練のコストが高い、テストが遅い。
SPPNet [99]	SS	ZFNet	任意	60.9 (07)	—	< 1	CCCV14	Caffe Python			注目点: 最初に SPP を CNN アーキテクチャに導入、量込み特徴の共有をする。性能を犠牲にするところなく RCNN の特徴を継承、訓練があまり高速化しない。Fine-tuning で SPP 層までの CONV 層を更新できない。
Fast RCNN [84]	SS	AlexNet VGG VGG16	任意	70.0 (07+12)	68.4 (VGG) (07++12)	< 1	ICCV15	Caffe Python			注目点: 最初に end-to-end (領域提案生成を無視) の検出器訓練を可能にした。RoI pooling 層を設計。SPPNet よりはるかに高速で正確。特徴のキャッシュ用のディスクストレージが不要。 欠点: 外部の領域提案計算が新たなボトルネックとなる。依然リアルタイムアリケーションは遅すぎる。
(図 5.1) ルーミー等	Faster RCNN [229]	RPN	ZFnet VGG	73.2 (VGG) (07+12)	70.4 (VGG) (07++12)	< 5	NIPS15	Caffe Matlab Python			注目点: selective search に代わる、ほとんど追加計算コスト無しに高品質な領域提案を生成する RPN を提案。RPN の参照用矩形として平行移動不変でマルチスケールの 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 よりも約 1 倍高速。VGG16 では 5 FPS でテストを実行可能。 欠点: 訓練が複雑でプロセスが合理化されていない、依然リアルタイムに達していない。
RCNN \ominus R [151]	新規	ZFNet +SPP	任意	59.7 (07)	—	< 5	BMVC15	—			注目点: selective search を静的な領域提案で置き換える。CNN のみに依存する統合された単純で高速な検出器を構築できる可能性を証明。 欠点: リアルタイムに達してない、貧弱な領域提案により精度が低下。
RFCN [50]	RPN	ResNet101	任意	80.5 (07+12)	77.6 (07++12)	< 10	NIPS16	Caffe Matlab			注目点: fully convolutional の検出ネットワーク。特殊な CONV 層のセットを使用して position sensitive score map のセットを設計。あまり精度を犠牲にするこことなく Faster RCNN より高速。 欠点: 訓練プロセスが合理化されていない、依然リアルタイムに達していない。
Mask RCNN [102]	RPN	ResNet101 ResNeXt101	任意	50.3 (ResNeXt101) (COCO での結果)	< 5	ICCV17	Caffe Matlab Python				注目点: 単純で柔軟な物体インスタンスグレンディング用フレームワーク。bounding box 予測用の既存ランチと並行する、物体クラス予測用の別のランチを追加することで Faster RCNN を拡張。Feature Pyramid Network (FPN) が活用される。傑出した性能。 欠点: リアルタイムアリケーションには及ばない。
OverFeat [239]	—	AlexNet like	任意	—	—	< 0.1	ICLR14 c++				注目点: 量込み特徴の共有。マルチスケールの画像ピラミッドでの CNN 特徴抽出。IS-LVRC2013 位置推定コンペティションによる勝利。 欠点: 順次訓練される多段階バイオライン、單一の bounding box 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ない、リアルタイムアリケーションには適すぎない。
YOLO [227]	—	GoogLeNet like	固定	66.4 (07+12)	57.9 (07++12)	< 25 (VGG)	CVPR16	DarkNet			注目点: 最初の効率的な総合検出器。領域提案プロセスを完全に削除。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高速。YOLO は 45 FPS で、Fast YOLO は 155 FPS で実行される。 欠点: 精度は最先端の検出器に遠く及ばない、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。多数の既存網絡を使用し速度と精度の両方を向上。高精度と高速さを両立。YOLO9000 は 9000 以上の物体カタログで検出可能。
SSD [175]	—	VGG16	固定	76.8 (07+12)	74.9 (07++12)	< 60	CCCV16	Caffe Python			注目点: 最初の正確かつ効率的な総合検出器。RPN と YOLO のアイデアを効果的に組み合わせ、マルチスケールの CONV 層で検出を行。YOLO より高速かつはるかに正確。50 FPS で実行可能。 欠点: 小さな物体の検出が苦手。