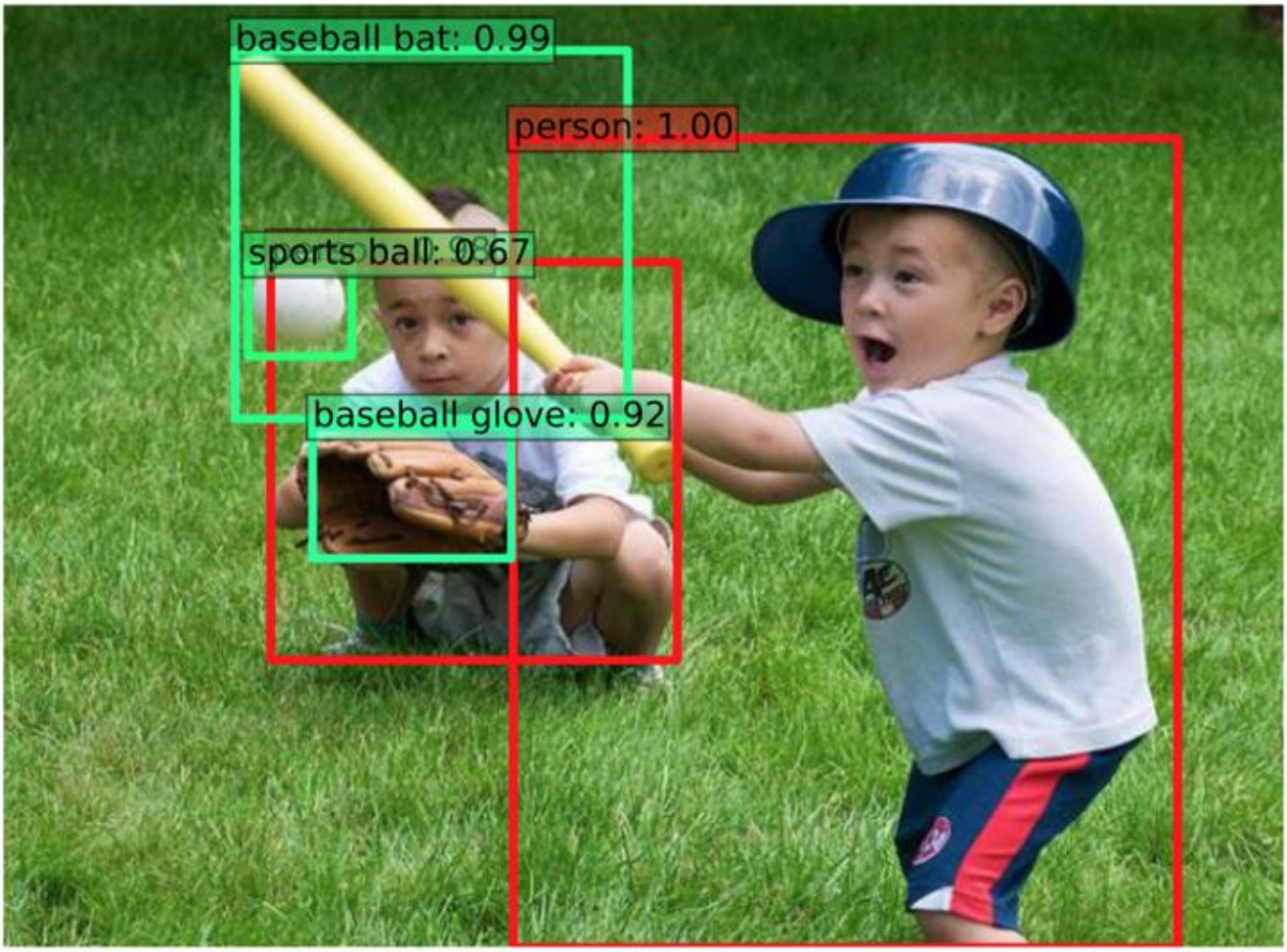
# Object as point summary

黒字：論文 赤字：コメント

## Introduction

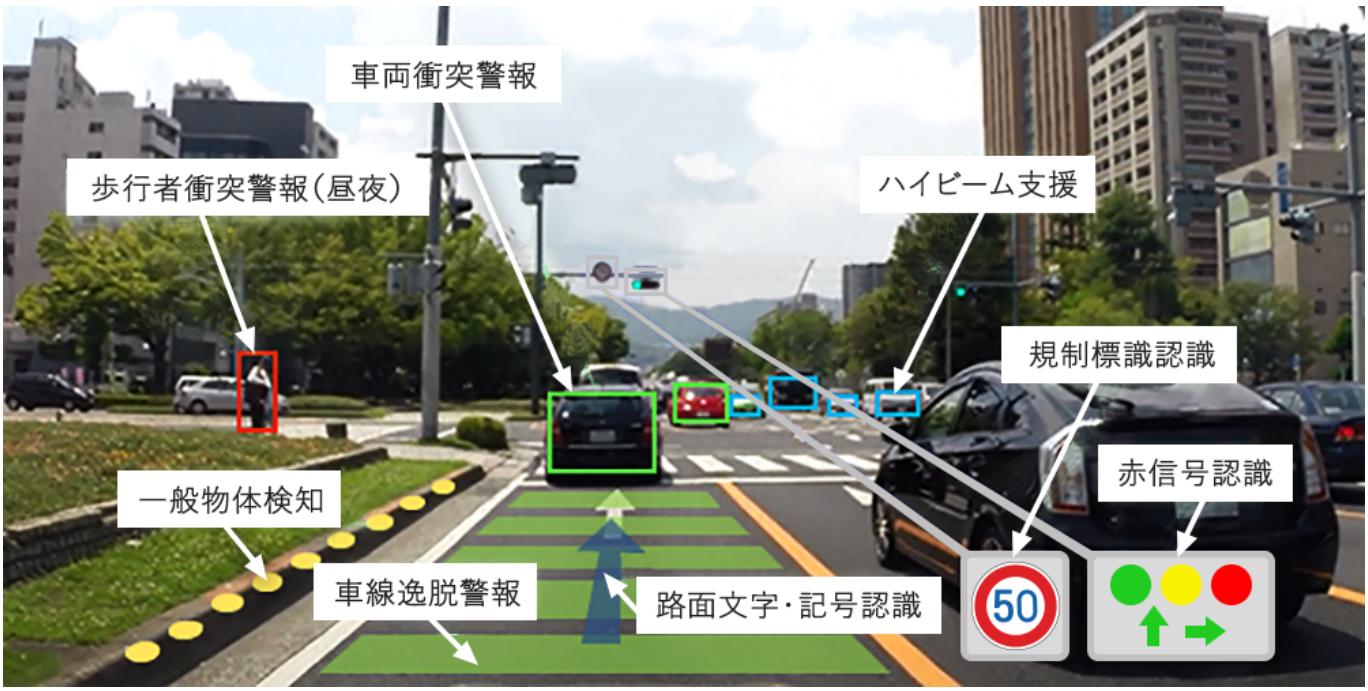
Object detection powers many vision tasks like instance segmentation (7,21,32], pose estimation [3, 15, 39), tracking [24, 27), and action recognition [5]. It has down-stream applications in surveillance [57], autonomous driving [53], and visual question answering [1]. Current object detectors represent each object through an axis-aligned bounding box that tightly encompasses the object [18,19,33, 43, 46). They then reduce object detection to image classification of an extensive number of potential object bounding boxes. For each bounding box, the classifier determines if the image content is a specific object or background. Onestage detectors [33, 43] slide a complex arrangement of possible bounding boxes, called anchors, over the image and classify them directly without specifying the box content. Two-stage detectors [18, 19, 46] recompute image features for each potential box, then classify those features. Post-processing, namely non-maxima suppression, then removes duplicated detections for the same instance by computing bounding box IoU. This post-processing is hard to differentiate and train [23], hence most current detectors are not end-to-end trainable. Nonetheless, over the past five years (19), this idea has achieved good empirical success [12,21,25,26,31,35,47,48,56,62,63]. Sliding window based object detectors are however a bit wasteful, as they need to enumerate all possible object locations and dimensions.

物体検出は，インスタンスのセグメンテーション（7,21,32），ポーズ推定（3,15,39），トラッキング（24,27），行動認識（5）など，多くのビジョンタスクをサポートしています． ▼物体検知の例




また，監視 [57]，自律運転 [53]，視覚的な質問応答 [1] などの分野でも応用されています． 現在の物体検出器は，各物体を軸に沿ったバウンディングボックスで表現し，物体を厳密に包んでいます[18,19,33,43,46] ▼

自動運転の例




そして，物体検出を，膨大な数の潜在的な物体バウンディングボックスの画像分類にまで落とし込みます．各バウンディングボックスについて，分類器は画像の内容が特定の物体か背景かを判断します．

ワンステージ検出器 [33, 43] は，アンカーと呼ばれる複雑な配置のバウンディングボックスを画像上にスライドさせ，ボックスの内容を指定せずに直接分類します． ▼1ステージ検出器 画像特徴量から直で分類

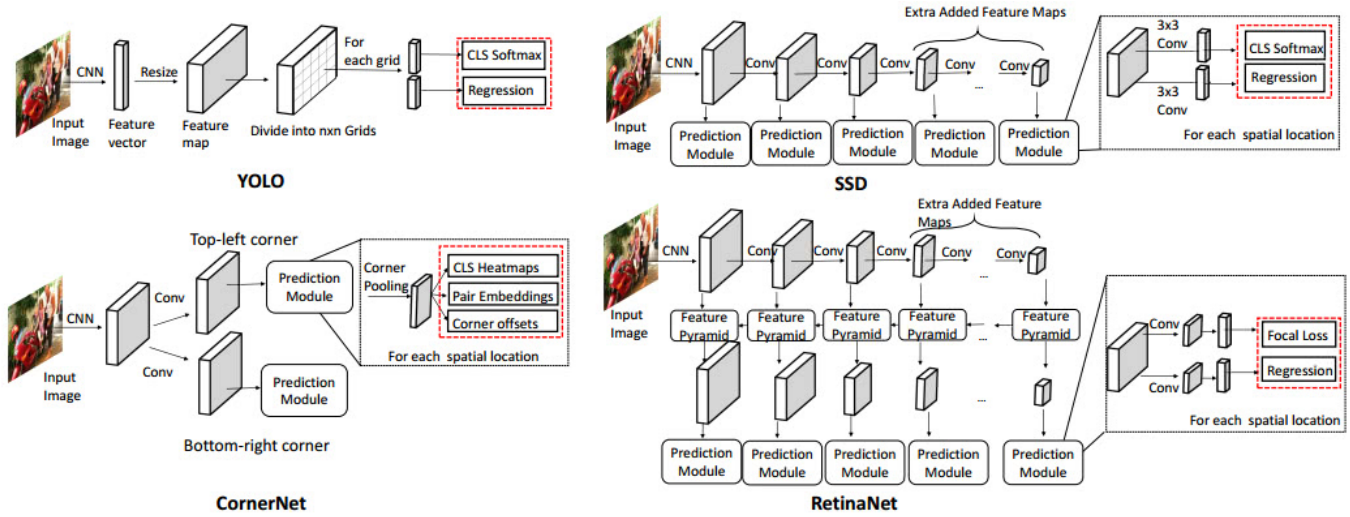



Figure 5: Overview of different one-stage detection frameworks for generic object detection. Red rectangles denotes the outputs that define the objective functions.

2段階検出器 [18, 19, 46] は，各可能性のあるボックスについて画像特徴を再計算し，それらの特徴を分類する． その後，後処理，すなわち非最大化抑制処理が行われ，バウンディングボックスIoUを計算することで，同じインスタンスについて重複した検出を除去する． ▼2ステージ検出器 画像特徴量からボックスを推定，次にそのボックスを分類

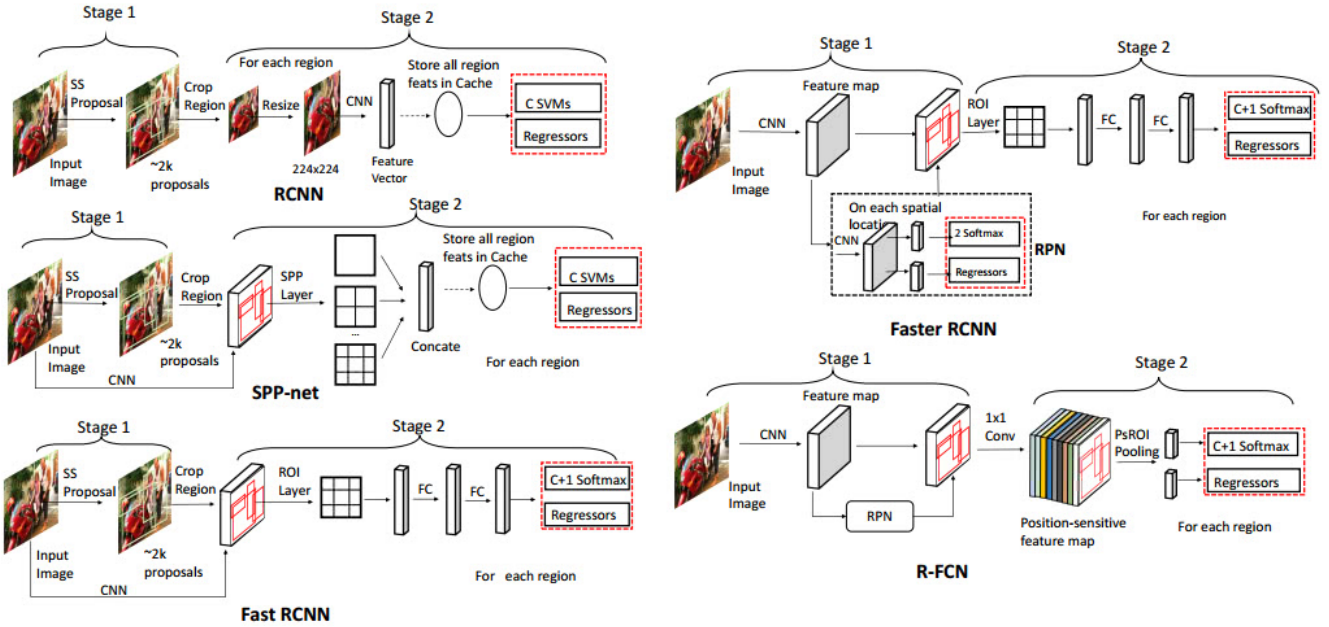



Figure 4: Overview of different two-stage detection frameworks for generic object detection. Red dotted rectangles denote the outputs that define the loss functions.

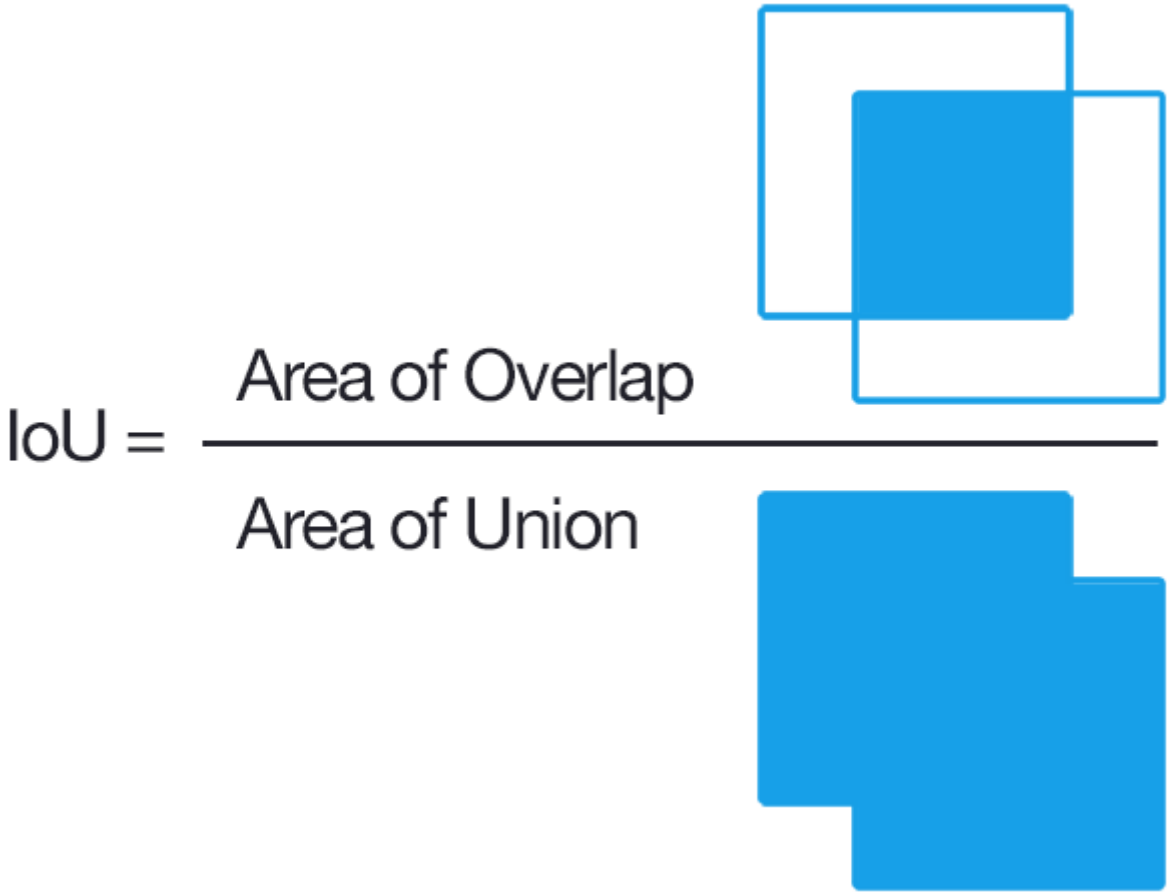
[12,21,25,26,31,35,47,48,56,62,63]． ▼非最大化抑制処理前 ![](https://i.imgur.com/SgA2ABw.jpg =x400)

▼非最大化抑制処理後 上手くやって１つとして検出 ![](https://i.imgur.com/GDCdcoU.jpg =x400)

▼IoU IoUとは，Intersection over Unionの略です． IoU値とは，画像の重なりの割合を表す値です．

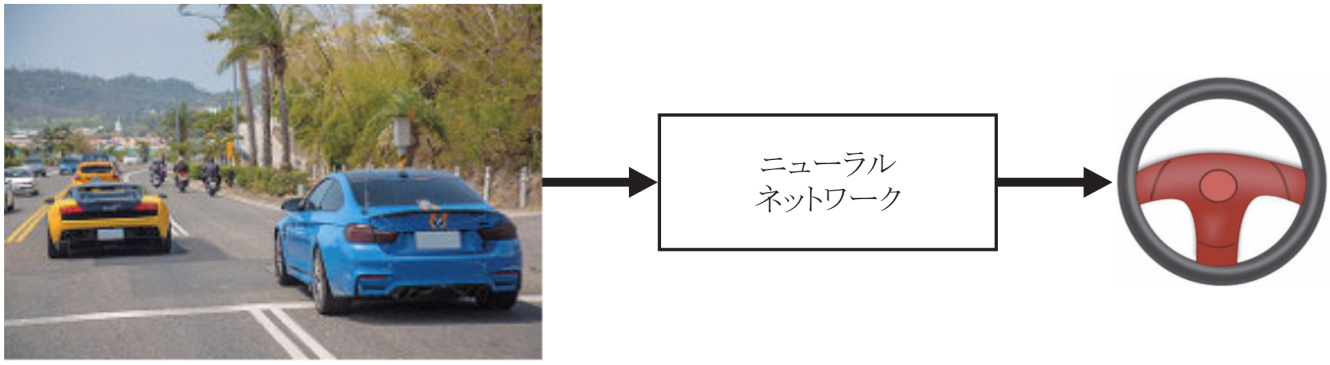
IoU値が大きいほど，画像が重なっている状態ということになります． IoU値が小さいほど，画像が重なっていない状態ということになります．

例: IoU値=0のとき，画像は全く重なっていない状態ということになります． IoU値=0.5のとき，画像は半分重なっている状態ということになります． IoU値=1.0のとき，画像は完全に重なっている状態ということに




なります.

この後処理は，区別して訓練するのが難しいため[23]，現在のほとんどの検出器は，エンドツーエンドで訓練することができません. ▼エンドツーエンド 自動運転を例に取ると，非エンドツーエンドのアプローチでは，物体認識，レーン検出，経路プランニング，ステアリング制御など，人間が設定した複数個のサブタスクを解く必要があるところ，エンドツーエンド学習では車載カメラから取得した画像から直接ステアリング操作を学習する




それにもかかわらず，過去5年間（19）で，このアイデアは経験的に良い成功を収めている しかし，スライディングウィンドウベースの物体検出器は，可能なすべての物体の位置と寸法を列挙する必要があるため，

少し無駄が多い．

In this paper, we provide a much simpler and more efficient alternative. We represent objects by a single point at their bounding box center (see Figure 2). Other properties, such as object size, dimension, 3D extent, orientation, and pose are then regressed directly from image features at the center location. Object detection is then a standard keypoint estimation problem (3,39,60]. We simply feed the input image to a fully convolutional network [37,40] that generates a heatmap. Peaks in this heatmap correspond to object centers. Image features at each peak predict the objects bounding box height and weight. The model trains using standard dense supervised learning [39,60]. Inference is a single network forward-pass, without non-maximal suppression for post-processing.

この論文では，よりシンプルで効率的な代替案を提供します．我々は，外接箱の中心にある一点でオブジェクトを表現します（図2を参照）．

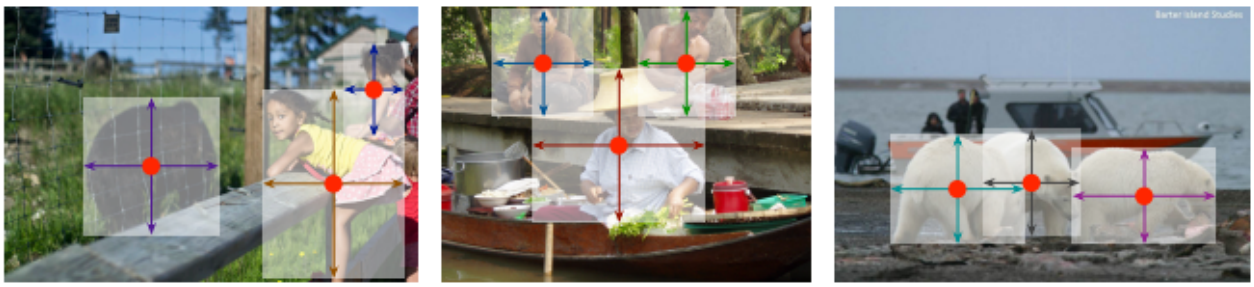


Figure 2: We model an object as the center point of its bounding box. The bounding box size and other object properties are inferred from the keypoint feature at the center. Best viewed in color.

次に，物体のサイズ，寸法，3次元の広がり，向き，ポーズなどの他の特性を，中心位置の画像特徴から直接回帰させます．物体検出は，標準的なキーポイント推定問題(3,39,60)になります．我々は単に入力画像を完全畳み込みネットワーク[37,40]に送るだけで，ヒートマップを生成します．このヒートマップのピークは物体の中心に対応しています．各ピークの画像特徴は，物体の境界ボックスの高さと重さを予測します．モデルは，標準的な密な教師付き学習[39,60]を使用して学習します．推論は単一のネットワークフォワードパスであり，後処理のための非最大抑制はありません． ！！！！後処理のための非最大抑制はありません！！！！

▶ 参考サイト

Object as Points slide

コンピュータビジョンの最新論文調査 キーポイントによる物体検出編

最近のSingle Shot系の物体検出のアーキテクチャまとめ

物体検出についての歴史まとめ

Recent Advances in Deep Learning for Object Detection - Part 1

Non-Maximum Suppressionを世界一わかりやすく解説する

エンドツーエンド深層学習のフロンティア