[Abstract 2](#_Toc48195109)

[Introduction 3](#_Toc1781366261)

[The Single Shot Detector（SSD） 5](#_Toc2071970706)

[1、Model 5](#_Toc985389590)

[2、Training 8](#_Toc1856055244)

[Experimental Results 14](#_Toc936992899)

SSD

# Abstract

这篇文章在既保证速度，又要保证精度的情况下，提出了 SSD 物体检测模型，与现在流行的检测模型一样，将检测过程整个成一个 single deep neural network。便于训练与优化，同时提高检测速度。

SSD 将输出一系列离散化（discretization） 的 bounding boxes，这些 bounding boxes 是在不同层次（layers） 上的 feature maps 上生成的，并且有着不同的 aspect ratio（长宽比）。

**在 prediction 阶段：**

要计算出每一个default box 中的物体，其属于每个类别的可能性，即 score得分。如对于 PASCAL VOC 数据集，总共有 20 类，那么得出每一个 bounding box 中物体属于这 20 个类别的每一种的可能性。

同时，要对这些 bounding boxes 的 shape 进行微调，以使得其符合物体的外接矩形。

还有就是，为了处理相同物体的不同尺寸的情况，SSD 结合了不同分辨率的 feature maps 的 predictions。

相对于那些需要 object proposals 的检测模型，本文的 SSD 方法完全取消了 proposals generation、pixel resampling（像素重采样） 或者 feature resampling 这些阶段。这样使得 SSD 更容易去优化训练，也更容易地将检测模型融合进系统之中。

在 PASCAL VOC、MS COCO、ILSVRC 数据集上的实验显示，SSD 在保证精度的同时，其速度要比用 region proposals 的方法要快很多。

SSD 相比较于其他单结构模型（YOLO），SSD 取得更高的精度，即是是在输入图像较小的情况下。如输入 300×300大小的 PASCAL VOC 2007 test 图像，在 Titan X 上，SSD 以 58 帧的速率，同时取得了 72.1%的 mAP。如果输入的图像是 500×500，SSD 则取得了 75.1% 的 mAP，比目前最 state-of-art 的 Faster R-CNN 要好很多。

# Introduction

现在流行的 state-of-art 的检测系统大致都是如下步骤，先生成一些假设的 bounding boxes，然后在这些 bounding boxes 中提取特征，之后再经过一个分类器，来判断里面是不是物体，是什么物体。

这类pipeline 自从 IJCV 2013, Selective Search for Object Recognition 开始，到如今在 PASCAL VOC、MS COCO、ILSVRC 数据集上取得领先的基于 Faster R-CNN 的 ResNet 。但这类方法对于嵌入式系统，所需要的计算时间太久了，不足以实时的进行检测。当然也有很多工作是朝着实时检测迈进，但目前为止，都是牺牲检测精度来换取时间。

本文提出的实时检测方法，消除了中间的 bounding boxes、pixel or feature resampling 的过程。虽然本文不是第一篇这样做的文章（YOLO），但是本文做了一些提升性的工作，既保证了速度，也保证了检测精度。

这里面有一句非常关键的话，基本概括了本文的核心思想：

Our improvements include using a small convolutional filter to predict object categories（类别） and offsets in bounding box locations, using separate（分离） predictors (filters) for different aspect ratio detections, and applying these filters to multiple feature maps from the later stages of a network in order to perform detection at multiple scales.

本文的主要贡献总结如下：

（1）提出了新的物体检测方法：SSD，比原先最快的 YOLO: You Only Look Once 方法，还要快，还要精确。保证速度的同时，其结果的 mAP 可与使用 region proposals 技术的方法（如 Faster R-CNN）相媲美。

（2）SSD 方法的核心就是 predict object（物体），以及其归属类别的 score（得分）；同时，在 feature map 上使用小的卷积核，去predict 一系列 bounding boxes 的 box offsets。

（3）本文中为了得到高精度的检测结果，在不同层次的 feature maps 上去 predict object、box offsets，同时，还得到不同 aspect ratio 的 predictions。

（4）本文的这些改进设计，能够在当输入分辨率较低的图像时，保证检测的精度。同时，这个整体 end-to-end 的设计，训练也变得简单。在检测速度、检测精度之间取得较好的 trade-off。

（5）本文提出的模型（model）在不同的数据集上，如 PASCAL VOC、MS COCO、ILSVRC， 都进行了测试。在检测时间（timing）、检测精度（accuracy）上，均与目前物体检测领域 state-of-art 的检测方法进行了比较。

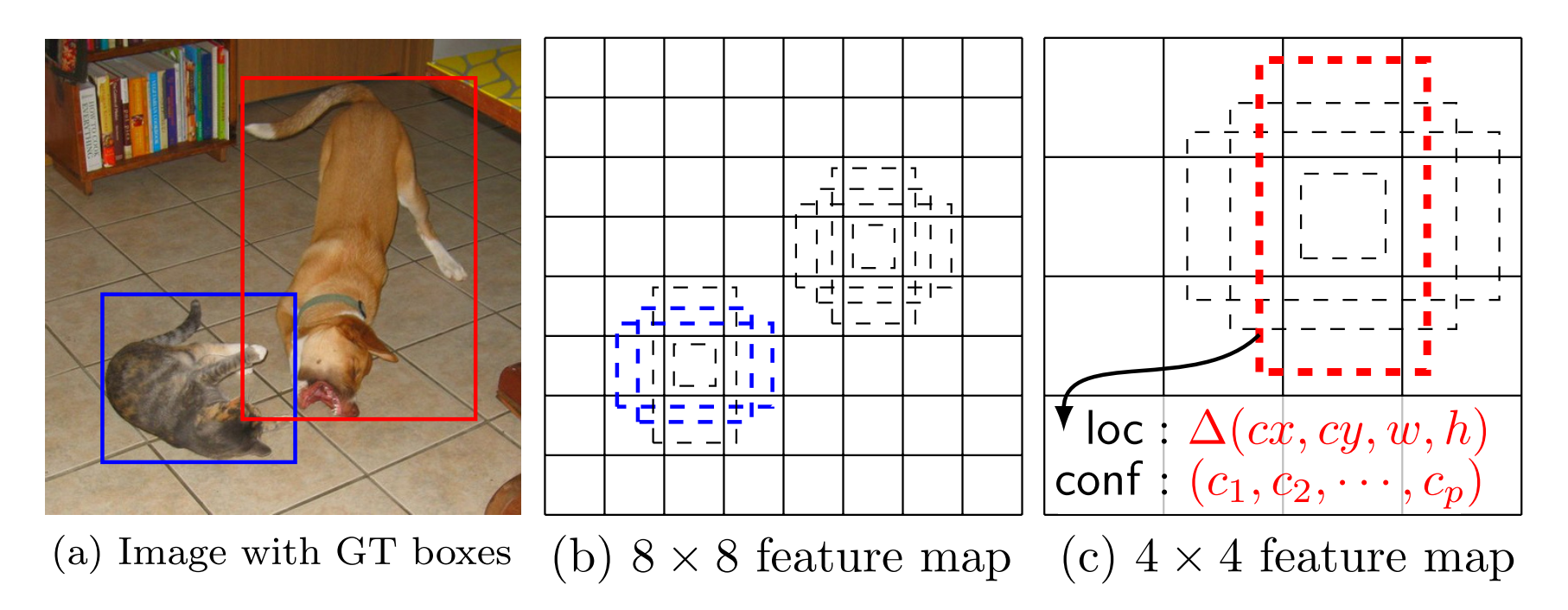
# The Single Shot Detector（SSD）

这部分详细讲解了 SSD 物体检测框架，以及 SSD 的训练方法。

这里，先弄清楚下文所说的 default box 以及 feature map cell 是什么。看下图：

（1）feature map cell 就是将 feature map 切分成 8×8或者 4 ×4之后的一个个 格子；

（2）而 default box 就是每一个格子上，一系列固定大小的 box， 即图中虚线所形成的一系列 boxes。



## 1、Model

SSD 是基于一个前向传播 CNN 网络，产生一系列固定大小（fixed-size） 的 bounding boxes，以及每一个 box 中包含物体实例的可能性，即 score。之后，进行一个非极大值抑制（Non-maximum suppression） 得到最终的 predictions。

SSD 模型的最开始部分，本文称作 base network，是用于图像分类的标准架构。在 base network 之后，本文添加了额外辅助的网络结构：

（1）Multi-scale feature maps for detection

在基础网络结构后，添加了额外的卷积层，这些卷积层的大小是逐层递减的，可以在多尺度下进行 predictions。

（2）Convolutional predictors for detection

每一个添加的特征层（或者在基础网络结构中的特征层），可以使用一系列 convolutional filters，去产生一系列固定大小的 predictions，具体见 Fig.2。对于一个大小为 m×n，具有 p 通道的特征层，使用的 convolutional filters 就是 3×3×p 的 kernels。产生的 predictions，那么就是归属类别的一个得分，要么就是相对于 default box coordinate 的 shape offsets。

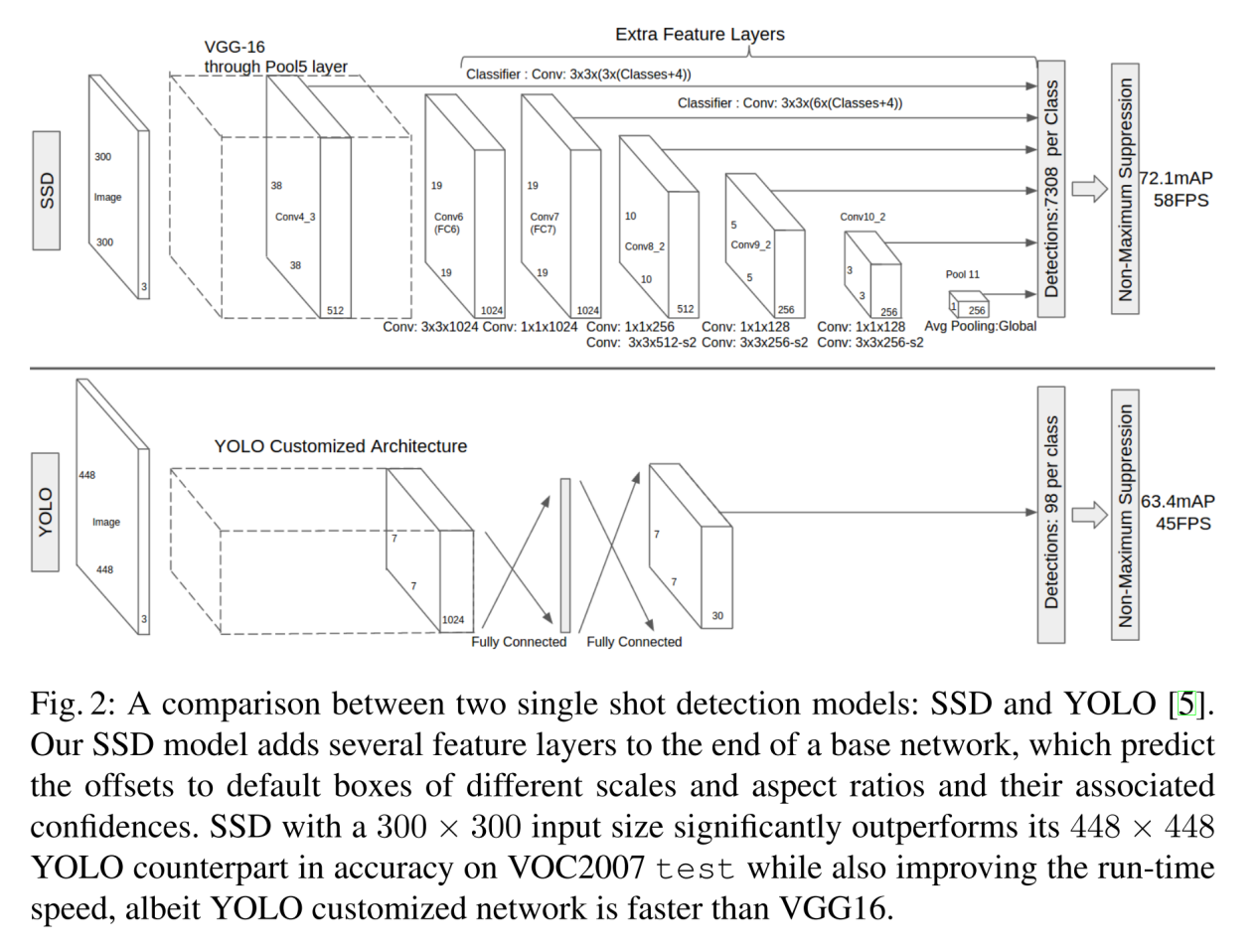
在每一个 m×n 的特征图位置上，使用上面的 3×3 的 kernel，会产生一个输出值。bounding box offset 值是输出的 default box 与此时 feature map location 之间的相对距离（YOLO 架构则是用一个全连接层来代替这里的卷积层）。

（3）Default boxes and aspect ratios

每一个 box 相对于与其对应的 feature map cell 的位置是固定的。 在每一个 feature map cell 中，我们要 predict 得到的 box 与 default box 之间的 offsets，以及每一个 box 中包含物体的 score（每一个类别概率都要计算出）。

因此，对于一个位置上的 k 个boxes 中的每一个 box，我们需要计算出 c 个类，每一个类的 score，还有这个 box 相对于 它的默认 box 的 4 个偏移值（offsets）。于是，在 feature map 中的每一个 feature map cell 上，就需要有 (c+4)×k 个 filters。对于一张 m×n 大小的 feature map，即会产生 (c+4)×k×m×n 个输出结果。

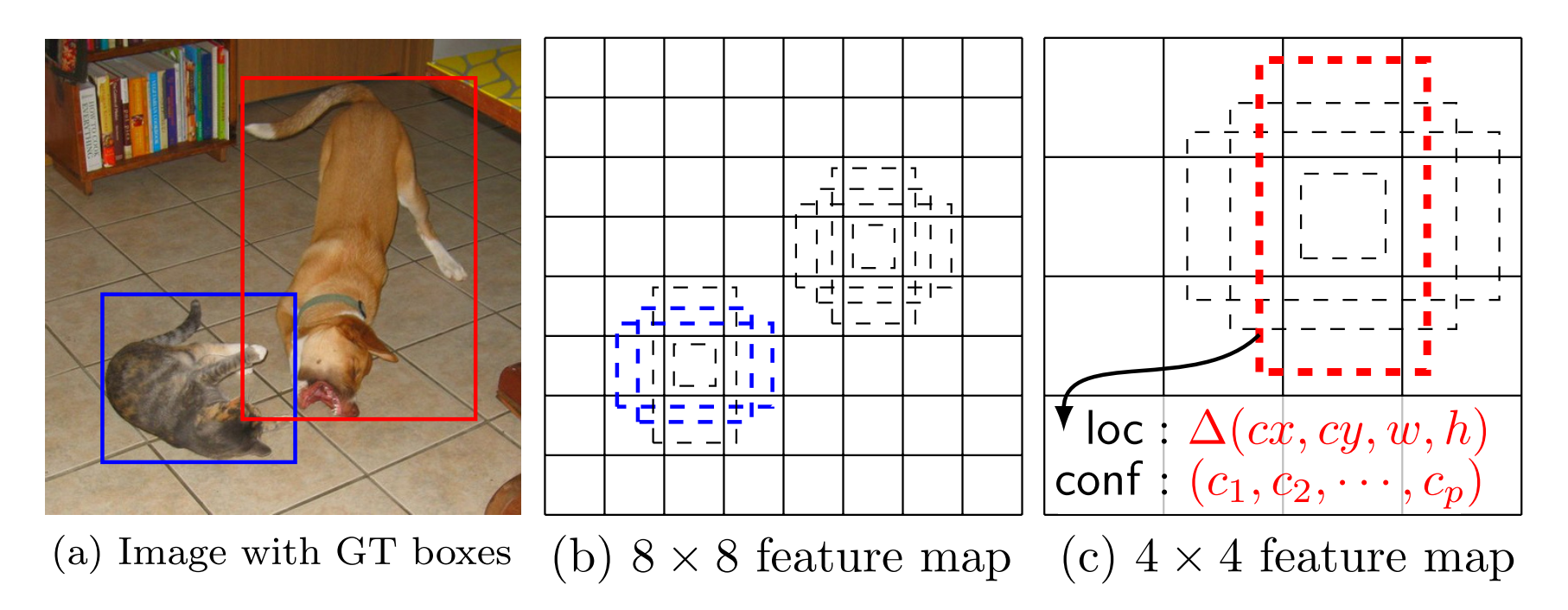
这里的 default box 很类似于 Faster R-CNN 中的 Anchor boxes，关于这里的 Anchor boxes，详细的参见原论文。但是又不同于 Faster R-CNN 中的，本文中的 Anchor boxes 用在了不同分辨率的 feature maps 上。



## 2、Training

在训练时，本文的SSD 与那些用 region proposals + pooling 方法的区别是，SSD 训练图像中的 groundtruth 需要赋予到那些固定输出的 boxes 上。在前面也已经提到了，SSD 输出的是事先定义好的，一系列固定大小的 bounding boxes。

如下图中，狗狗的 groundtruth 是红色的 bounding boxes，但进行 label 标注的时候，要将红色的 groundtruth box 赋予 图（c）中一系列固定输出的 boxes 中的一个，即 图（c）中的红色虚线框。



事实上，文章中指出，像这样定义的 groundtruth boxes 不止在本文中用到。在 YOLO 中，在 Faster R-CNN 中的 region proposal 阶段，以及在 MultiBox 中，都用到了。

当这种将训练图像中的 groundtruth 与固定输出的 boxes 对应之后，就可以 end-to-end 的进行 loss function 的计算以及 back-propagation 的计算更新了。

训练中会遇到一些问题：

（1）选择一系列 default boxes

（2）选择上文中提到的 scales 的问题

（3）hard negative mining

（4）数据增广的策略

下面会谈本文的解决这些问题的方式，分为以下下面的几个部分。

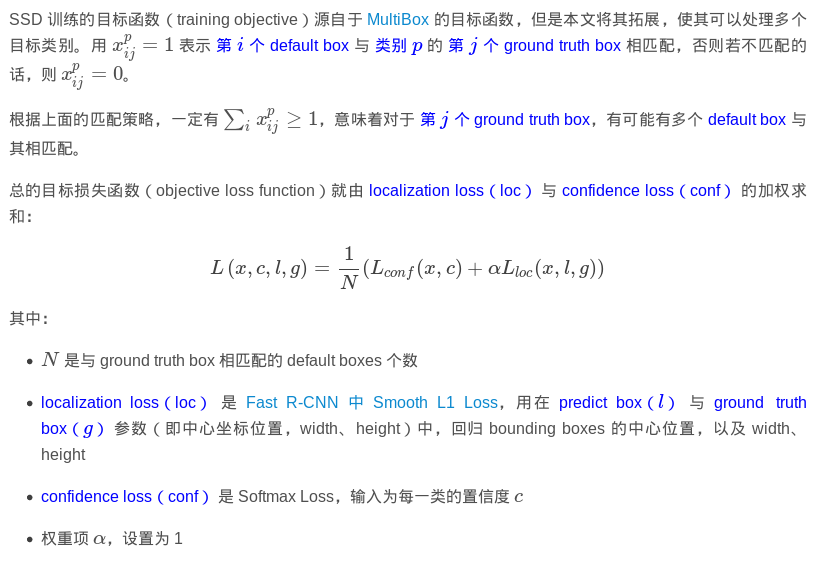
（1）Matching strategy：

如何将 groundtruth boxes 与 default boxes 进行配对，以组成 label 呢？

在开始的时候，用 MultiBox 中的 best jaccard overlap 来匹配每一个 ground truth box 与 default box，这样就能保证每一个 groundtruth box 与唯一的一个 default box 对应起来。

但是又不同于 MultiBox ，本文之后又将 default box 与任何的 groundtruth box 配对，只要两者之间的 jaccard overlap 大于一个阈值，这里本文的阈值为 0.5。

（2）Training objective



（3）Choosing scales and aspect ratios for default boxes

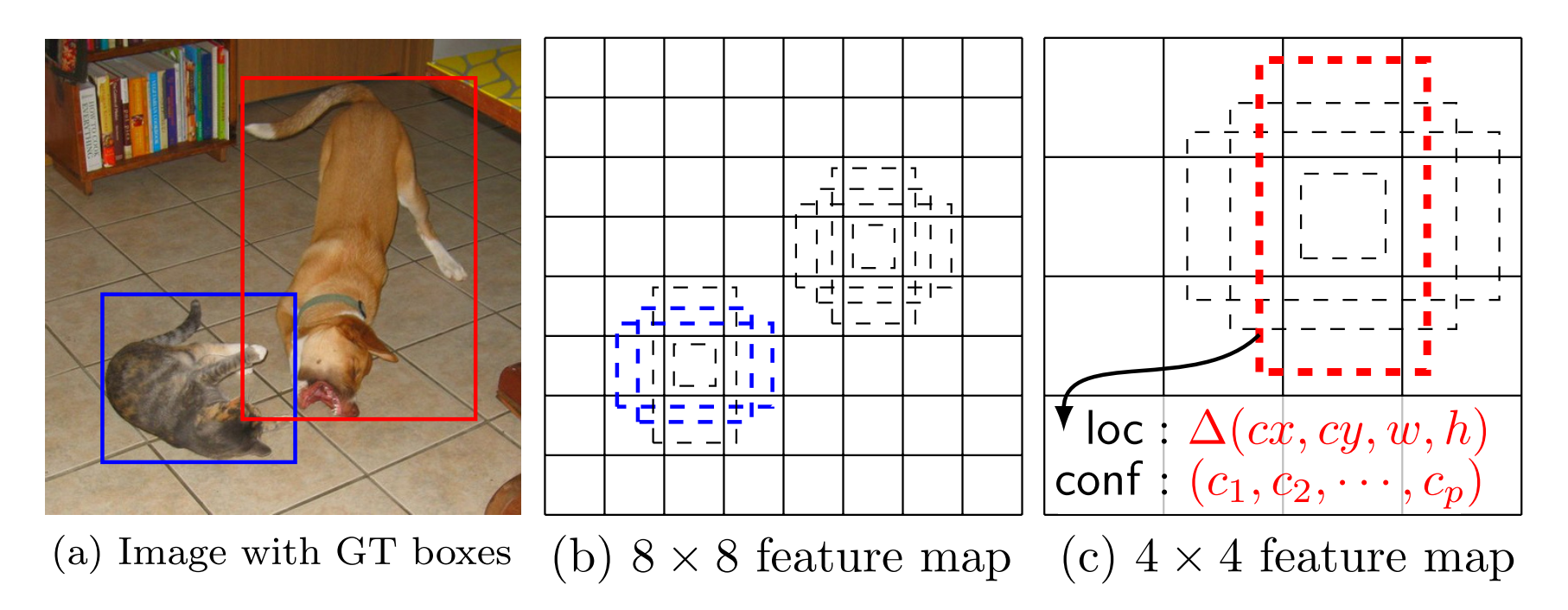
大部分 CNN 网络在越深的层，feature map 的尺寸（size）会越来越小。这样做不仅仅是为了减少计算与内存的需求，还有个好处就是，最后提取的 feature map 就会有某种程度上的平移与尺度不变性。

同时为了处理不同尺度的物体，一些文章，如 ICLR 2014, Overfeat: Integrated recognition, localization and detection using convolutional networks，还有 ECCV 2014, Spatial pyramid pooling in deep convolutional networks for visual recognition，他们将图像转换成不同的尺度，将这些图像独立的通过 CNN 网络处理，再将这些不同尺度的图像结果进行综合。

但是其实，如果使用同一个网络中的、不同层上的 feature maps，也可以达到相同的效果，同时在所有物体尺度中共享参数。

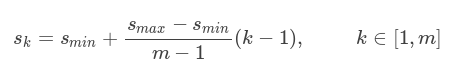
之前的工作，如 CVPR 2015, Fully convolutional networks for semantic segmentation，还有 CVPR 2015, Hypercolumns for object segmentation and fine-grained localization 就用了 CNN 前面的 layers，来提高图像分割的效果，因为越底层的 layers，保留的图像细节越多。文章 ICLR 2016, ParseNet: Looking wider to see better 也证明了以上的想法是可行的。

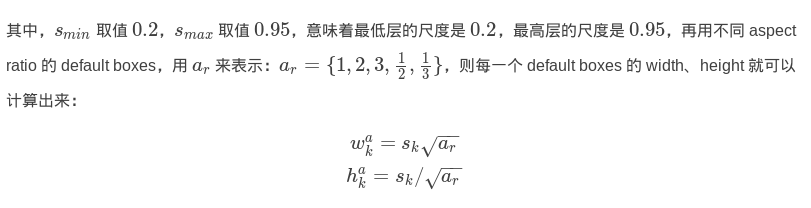
因此，本文同时使用 lower feature maps、upper feature maps 来 predict detections。下图展示了本文中使用的两种不同尺度的 feature map，8×8的feature map，以及 4×4的 feature map：

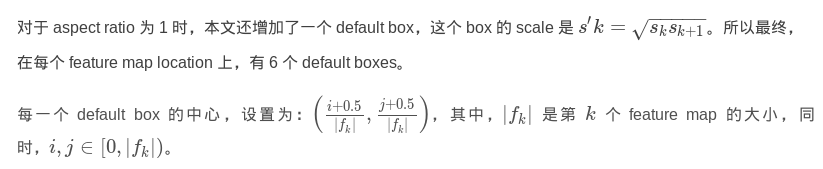


一般来说，一个 CNN 网络中不同的 layers 有着不同尺寸的 感受野（receptive fields）。这里的感受野，指的是输出的 feature map 上的一个节点，其对应输入图像上尺寸的大小。

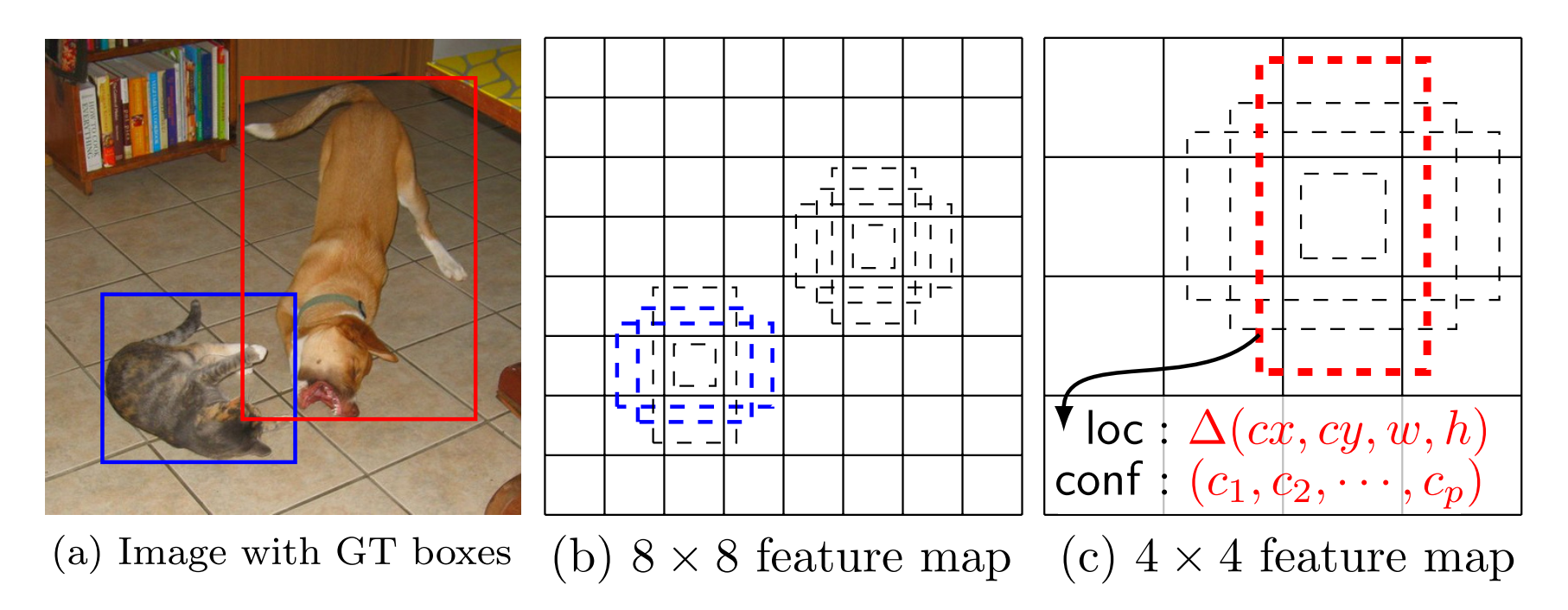
所幸的是，SSD 结构中，default boxes 不必要与每一层 layer 的 receptive fields 对应。本文的设计中，feature map 中特定的位置，来负责图像中特定的区域，以及物体特定的尺寸。假如我们用 m 个 feature maps 来做 predictions，每一个 feature map 中 default box 的尺寸大小计算如下：







在结合 feature maps 上，所有 不同尺度、不同 aspect ratios 的 default boxes，它们预测的 predictions 之后。可以想见，我们有许多个 predictions，包含了物体的不同尺寸、形状。如下图，狗狗的 ground truth box 与 4×4feature map 中的红色 box 吻合，所以其余的 boxes 都看作负样本。



（4）Hard negative mining

在生成一系列的 predictions 之后，会产生很多个符合 ground truth box 的 predictions boxes，但同时，不符合 ground truth boxes 也很多，而且这个 negative boxes，远多于 positive boxes。这会造成 negative boxes、positive boxes 之间的不均衡。训练时难以收敛。

因此，本文采取，先将每一个物体位置上对应 predictions（default boxes）是 negative 的 boxes 进行排序，按照 default boxes 的 confidence 的大小。选择最高的几个，保证最后 negatives、positives 的比例在 3:1。

本文通过实验发现，这样的比例可以更快的优化，训练也更稳定。

（5）Data augmentation

本文同时对训练数据做了 data augmentation，数据增广。关于数据增广，推荐一篇文章：Must Know Tips/Tricks in Deep Neural Networks，其中的section 1 就讲了data augmentation 技术。

每一张训练图像，随机的进行如下几种选择：

A）使用原始的图像

B）采样一个 patch，与物体之间最小的 jaccard overlap 为：0.1，0.3，0.5，0.7 与 0.9

C）随机的采样一个 patch

采样的 patch 是原始图像大小比例是 [0.1，1]，aspect ratio 在 12 与 2之间。

当 groundtruth box 的 中心（center）在采样的 patch 中时，我们保留重叠部分。

在这些采样步骤之后，每一个采样的 patch 被 resize 到固定的大小，并且以 0.5的概率随机的水平翻转（horizontally flipped）

# Experimental Results