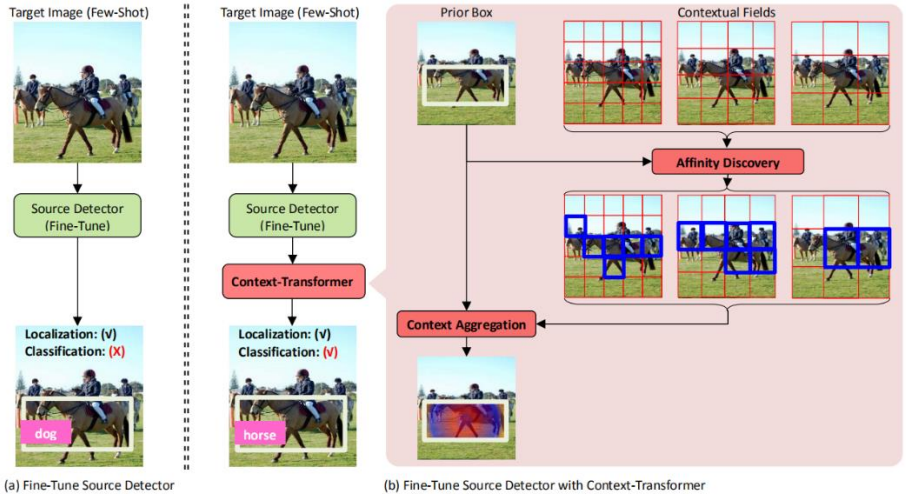
Context-Transformer: Tackling Object Confusion for Few-Shot Detection
少样本检测是一个现实又有挑战的工作，目前很多方法是预训练一个检测器，微调到新的样本，但是存在问题：由于缺少样本多样性，容易失败。 为了解决这一问题，提出了一个新的上下文迁移在一个简洁的深度迁移框架内。 具体来说，上下文迁移可以有效地利用源域对象知识作为指导，自动利用目标域中只有几个训练图像的上下文。 随后，它可以自适应地集成这些关系线索，以提高检测器的鉴别能力，以减少对象混淆。
迁移学习的方法，定位准确，分类出现混淆。原因是，检测由 bbox 回归来定位，object 和 background 分类器进行分类，因为 bbox 和类别无关，所以可以直接使用；分类器与类别有关，需要在目标域上进行调整。

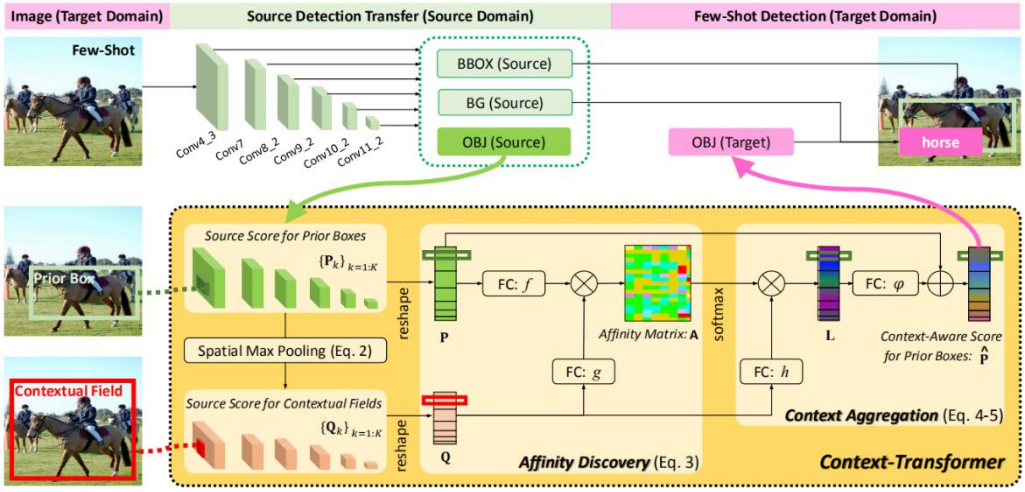


(a) Fine-Tune Source Detector (b) Fine-Tune Source Detector with Context-Transformer

它可以自动地从手头上的几幅图像中利用上下文，并仔细地整合这些不同的线索来概括检测。在很少的监督场景下，尝试探索环境中不同的线索（我们在本文中称之为上下文字段），以澄清对象混淆。 affinity discovery 和 context aggregation.关联发现和上下文聚合。 对于目标域图像，首先根据检测器中默认的先验框（也称为锚框）构造一组上下文字段。然后，它自适应地利用先验框和上下文字段之间的关系；最后上下文聚合模块利用指导等关系，并将关键上下文集中到每个先前的框中。

Source Detection Transfer

在迁移学习的框架下，首先 Cs 个类别的数据集上预训练，目标域有 Ct 个类别，然后有 N shot，backbone：SSD；该体系结构中的多尺度空间接收字段提供了丰富的上下文。
用 BBOX 进行定位.由于它在不同的类别之间共享，源域 BBOX 可以在目标域中重用。
BG 也可以复用，因为是二分类，显示是背景还是目标。
OBJ：我们建议保留源域 OBJ，并在其之上添加一个新的目标域 OBJ。 主要原因是，源域 OBJ 中预测分数的维数往往较高，小于卷积层中特征通道的数目。 当在源域 OBJ 上添加一个新的目标域 OBJ 时，我们将引入更少的额外参数，从而减轻其影响。



Source-Domain Object Knowledge of Prior Boxes

先前的框是具有不同纵横比的默认锚盒。由于分类是在这些框的表示上执行的，关联发现首先为先前的框构建了一组上下文字段。随后，在源域对象知识的指导下，利用目标域图像中的先验框和上下文字段之间的关系。我们将目标域图像输入预先训练的 SSD 样式检测器，并从源 OBJ(Softmax 之前)提取分数张量。

$$\mathbf{P}_k \in \mathbb{R}^{H_k \times W_k \times (M_k \times C_s)}, \quad k = 1, ..., K,$$

源域分类器的得分往往提供了有关目标域对象类别的丰富语义知识

Contextual Field Construction via Pooling.

设计不能太复杂，因为标注数据很少。人类经常检查稀疏的上下文字段，而不是关注图像中的每一个微小细节。建议在先前的框p_k上执行空间池（例如，最大池）。因此，我们得到了一组上下文字段的分数张量$Q_k \in R^{U_k \times V_k \times M_k \times C_s}$。

$$\mathbf{Q}_k \in \mathbb{R}^{U_k \times V_k \times (M_k \times C_s)} \text{ for a set of contextual fields,}$$

$$\mathbf{Q}_k = SpatialPool(\mathbf{P}_k), \quad k = 1, ..., K,$$

Affifinity Discovery

为了方便起见，我们将分数张量$P_{1:K}$和$Q_{1:K}$分别重塑为矩阵 P∈$R^{D_p * C_s}$和 Q∈$R^{D_q * C_s}$，其中每一行P(或Q)都是指先验框(或上下文字段)的源域分数向量。此外，$D_p = \sum_{k=1}^{K} H_k \times W_k \times M_k$和$D_q = \sum_{k=1}^{K} U_k \times V_k \times M_k$分别是目标域图像中先前框和上下文字段的总数。为了简单起见在嵌入空间中对广泛使用的点积核进行 P 和 Q 的比较.. 因此，我们得到了一个关联矩阵A∈$R^{D_p \times D_q}$之间的先验框和上下文字段，

$$\mathbf{A} = f(\mathbf{P}) \times g(\mathbf{Q})^{\top},$$

关联发现允许一个前框从各种纵横比、位置和空间尺度自动识别其重要的上下文字段。 这种多样化的关系可以进行判别，减少注释稀缺引起的对象混淆的线索

Context Aggregation

$$\mathbf{L}(i,:) = softmax(\mathbf{A}(i,:)) \times h(\mathbf{Q}),$$

它指示每个上下文字段对于第一个先验框有多重要。我们用它来总结所有的背景
最后，将加权上下文矩阵 L∈$R^{D_p \times C_s}$聚合到原始分数矩阵 P 中，得到先验框 P 的上下文感知分数矩阵

$$\widehat{\mathbf{P}} = \mathbf{P} + \varphi(\mathbf{L}).$$

最后，我们将$\hat{P}$引入目标域 OBJ，

$$\widehat{\mathbf{Y}} = softmax(\widehat{\mathbf{P}} \times \Theta),$$

是分类的目标域评分矩阵。目标 OBJ 是在不同的纵横比和空间尺度之间共享的，具有共同的参数矩阵 $\Theta \in R^{C_s \times C_t}$。 一个原因是，每个先验框都结合了不同纵横比和空间尺度的重要上下文场，即每一行$\hat{P}$已经成为一个多尺度的评分向量。 因此，没有必要在每个单独的尺度上分配一个独占的 OBJ。 更重要的是，目标域很少。在这种情况下，共享 OBJ 可以有效地减少过度拟合.

数据集：
VOC，COCO 上预训练，并减去了和 VOC 重合的类

VOC2007 (5-Shot Case)	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	avg
Prototype (Snell et al. 2017)	50.0	55.0	23.7	**26.1**	8.9	54.2	71.2	41.6	29.8	23.6	34.0	30.7	46.3	53.3	60.2	**21.9**	37.3	24.3	50.2	**54.4**	39.8
Imprinted (Qi et al. 2018)	49.1	54.8	26.0	23.5	14.7	53.0	71.2	53.0	30.4	21.0	34.0	28.6	48.1	**56.4**	63.5	21.4	39.2	32.9	45.1	50.8	40.9
Non-local (Wang et al. 2018)	51.9	58.1	25.3	**26.1**	8.5	49.7	71.9	55.3	**32.3**	20.1	31.9	**32.1**	44.7	55.8	63.7	16.2	41.7	**33.2**	52.4	49.9	41.0
Our Context-Transformer	**55.4**	**59.1**	**28.6**	23.9	**15.9**	**58.3**	**74.5**	**57.1**	31.4	**26.0**	**38.1**	31.7	**55.8**	56.1	**64.1**	18.1	**45.8**	33.0	**53.2**	49.9	**43.8**