软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.000000]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于注意力机制及类别层次结构的弱监督目标定位方法*

冯 迅, 杨 健, 周涛, 宫 辰,

(南京理工大学计算机科学与工程学院模式计算与应用实验室 南京 210094)

(南京理工大学计算机科学与工程学院高维信息智能感知与系统教育部重点实验室 南京 210094)

(南京理工大学计算机科学与工程学院江苏省社会安全图像与视频理解重点实验室 南京 210094)

通讯作者: 宫辰, E-mail: chen.gong@njust.edu.cn

摘 要: 弱监督目标定位是指仅利用图像级的类别标注信息来训练目标定位器，而不需要使用精确的目标位置标注信息来进行算法训练。当前的一些方法往往只能定位出目标对象中最具鉴别性的部分而无法准确地标识出完整的目标对象，或者易受背景无关信息干扰从而导致定位结果不精确。为了解决上述问题，本文提出了一种基于注意力机制和类别层次结构的弱监督目标定位方法。该方法通过对卷积神经网络的注意力图进行均值分割提取更完整的目标区域。进一步，通过类别层次结构网络实现对背景区域注意力的削弱，从而提高对感兴趣目标的定位精度。基于多个网络结构和公共数据集上的大量实验结果表明，相比目前已有的弱监督定位方法，本文提出的方法在多个评价指标下均能够获得更好的定位效果。

关键词: 弱监督目标定位;网络注意力;背景干扰;层次结构网络;卷积神经网络

中图法分类号: TP391

Attention Mechanism and Categorical Hierarchy Based Weakly Supervised Object Localization

FENG Xun YANG Jian ZHOU Tao GONG Chen

(PCA Lab, the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Laboratory of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094)

Abstract: Weakly supervised object localization aims to decide object locations in a given image by only using the image-level labels instead of location annotations. Some current methods can only identify the most discriminative region of the target object, and are incapable of covering the complete object, or can easily be misled by irrelevant background information, which leads to an inaccurate object location. In this paper, we propose a weakly supervised object localization algorithm which employs attention mechanism and categorical hierarchy. The proposed method extracts a more complete object area by performing mean segmentation on the attention map of the convolutional neural network. In addition, we use the category hierarchy network to weaken the attention caused by background area, leading to more accurate object location results. Extensive experimental results on multiple public datasets demonstrate the effectiveness of the proposed method over other weakly supervised object localization methods under various evaluation metrics.

Key words: weakly supervised object localization; network attention; background interference; hierarchical network; convolutional neural network.

* 基金项目: 国家自然科学基金项目面上项目(61973162，62172228)、中国科协青年人才托举工程(2018QNRC001)
收稿时间: 0000-00-00

1 引言

弱监督目标定位是指仅使用图像级标签(存在或不存在某类目标)作为监督信息进行定位算法训练，并实现目标定位功能。相比于全监督方法[1-5]，弱监督方法不需要昂贵的边界框标注信息，从而可以大大降低人力成本。得益于其轻量级的标注需求，近年来引起了各国学者对弱监督目标定位越来越多的关注。

目前的主要方法是基于类激活图(Class Activation Maps, CAM)[6]来寻找目标鉴别性区域从而进行目标定位。由于 CAM 方法在测试时很大程度上依赖于目标分类结果，因此往往仅能识别出目标对象中与预测类别相关的最具鉴别性的部分区域，如图 1 中第 2 行第 1 列所示，CAM 方法提取的鸟类激活区域仅局限于头部，因此导致定位结果不精确。最近的一些研究针对该问题提出了不同的改进方法[7-16]。这些方法大多采用通过擦除(erasing)[13]已发现的目标区域，来驱使网络发现其它鉴别性较弱的目标区域，从而定位出更完整的目标区域。尽管基于擦除策略可以帮助算法发现更多目标区域，但通常发现的目标区域较为稀疏，仍旧无法保证目标的完整性，这对定位仍是不利的。此外，对于具有“共现”现象（指一些目标类别总是与特定的背景频繁地一起出现，这种情况下背景区域对分类网络来说也具有一定的鉴别性信息）的图片，将目标最具鉴别性的区域擦除后进而寻找到的其它较弱鉴别性区域，可能包含频繁与目标一起出现的背景区域[12]，如图 1 中第 3 行第 5 列所示。这也常常带来一些不满意的目标定位结果。

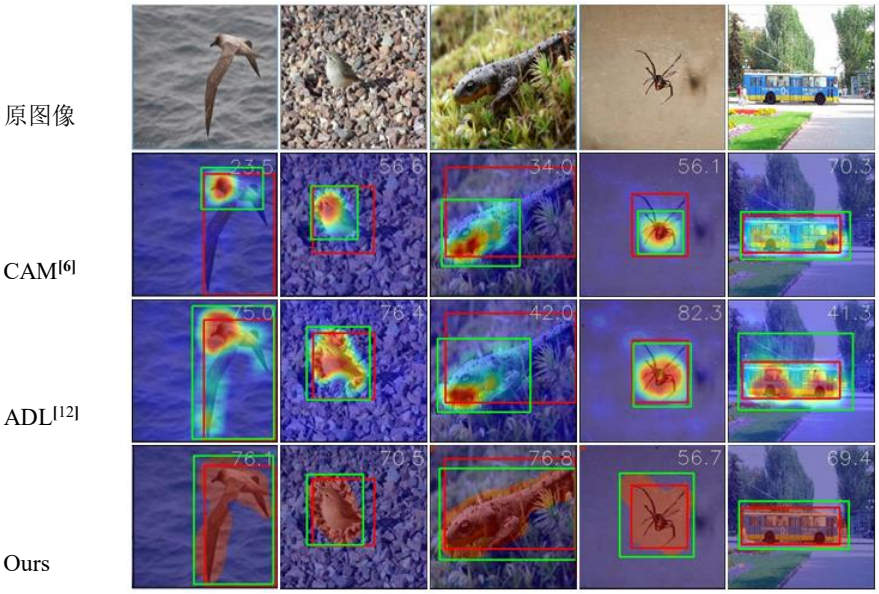

图 1　本文方法和 CAM[6]、ADL[12]的定位结果可视化，其中绿色框为算法输出的目标边界框，红色框为真实的目标边界框。从上图可以发现 CAM 方法适合小目标定位(如第 2 行 2,4 列)，对于大目标，CAM 方法获得的注意力图往往局限于目标最具鉴别性的部分区域，从而导致定位不精确(如第 2 行 1,3 列)。基于擦除策略的 ADL 方法缓解了 CAM 方法在大目标上获得的注意力图仅局限于目标部分区域的问题 (如第 3 行 1 列)，但同时也导致注意力向背景区域扩散使得目标定位结果不精确(如第 3 行 5 列)。本文提出的方法在大目标和小目标上均具有较好的定位效果(如第 4 行 1,2,3,4 列)，同时缓解了注意力向背景区域扩散的问题(如第 4 行 5 列)。

针对上述问题，本文提出了一种基于注意力机制及类别层次结构的弱监督目标定位方法。首先，可以从分类网络深层特征中提取注意力图。进一步，通过对注意力图进行阈值分割可以获得与预测类别相关的具有较高激活值的区域。与 CAM[6]方法相比，这些高激活值区域可以覆盖更多目标区域，如图 1 所示。但这种方法获得的目标区域可能包含一些背景区域，这会影响定位精度。由于分类网络在训练过程中倾向于识别图片中具有鉴别性信息的区域，因此，如果让训练图片的背景尽量相似，则可以有效减少网络对背景区域的注意力，从而提高目标区域的识别精度。然而，图片中背景的相似性一般难以直接衡量，但本文发现属于相似类别的图片往往具有相似的背景信息，如下图 2 中所示。在图 2 中，狮子和花豹属于相似的类别，这两种类别的图片背景往往都为野外环境，因此背景具有一定相似性；出租车和校车属于相似的类别，

这两种类别的图片背景往往都为道路环境，这些背景也具有一定相似性。基于该发现，本文提出了一种新颖的类别层次结构网络，该网络包含一个由多个卷积块构成的主干网络和多个并行的具有相同结构的分支网络。主干网络用来提取公共特征，分支网络用来抑制整个网络对背景区域产生的错误注意力。具体地，本文所提方法按照类别相似性对数据集中的所有类别进行自下而上的层次化归并，并形成多个根类别（例如，图 2 中将“狮子”和“花豹”归为一个根类别，将“出租车”和“校车”归为另一个根类别）。之后，再根据归并出的根类别构建一个具有类别层次特性的多分支结构网络，并将同一根类别下的相似子类别的训练图片统一放在一个独立的分支网络中进行训练。由于类别相似的图片的背景往往具有一定的相似性，因此当这些图片被放在同一分支网络中训练时，背景区域对待定位目标来说不再具有足够的鉴别信息，从而可以有效地减少网络对背景区域的注意力，进而提高对感兴趣目标的定位精度。

图 2　相似类别示例。狮子和花豹属于相似类别，它们往往具有相似的野外背景信息。出租车和校车属于相似类别，它们往往具有相似的道路背景信息

最后，本文结合伪监督目标定位方法[16]，在训练过程中，利用注意力技术从类别层次结构网络中提取目标区域来生成目标边界框，并将其作为伪监督信息直接训练一个定位网络，从而实现端到端的训练。此外，由于卷积网络对含噪声的标注具有一定容忍性，随着训练的进行可以从有噪声的标注中学习到鲁棒的模式[16]，因此，本文提出在训练过程的后半段，利用目标定位网络的输出对类别层次结构网络生成的含噪声伪监督信息进行过滤，剔除那些质量较差的伪监督信息，使得参与训练的伪监督信息的质量逐渐提高，从而进一步提高模型的定位性能。通过实验我们验证了本文方法与基于 CAM 的定位方法相比能获得更完整的目标定位结果；与基于擦除策略的定位方法相比缓解了目标注意力向背景区域扩散的问题（如图 1 所示）。综上，本文的主要贡献如下：

- 提出了一种基于注意力机制及类别层次结构的目标区域提取方法，该方法可以获得更完整的目标区域，同时有效地抑制背景无关区域对定位的干扰，从而提高目标定位精度。
- 提出了一种在训练过程中利用定位网络的预测结果对含噪声的伪监督信息进行过滤的方法，以提高参与训练的伪监督信息的质量，从而并进一步提升最终定位网络的定位能力。
- 通过在不同公共数据集和不同网络结构上的大量实验，验证了本文提出的方法比当前最优的弱监督目标定位方法取得更好的定位效果。

2　相关工作

弱监督目标定位的主要挑战在于获得完整的目标区域。由于弱监督目标定位使用图片类别标签进行分类任务训练，因此通过 CAM[6]获得的目标激活区域往往局限于目标最具鉴别性的区域，而不是目标整体。所以，如何将目标激活区域扩展到目标整体范围且同时不引入无关背景一直是弱监督目标定位的主要研究课题。当前，存在的弱监督目标定位主要基于擦除和数据增广的方法。

基于擦除的方法：Wei 等人[13]基于 CAM 技术提出对抗擦除 (Adversarial Erasing, AE)方法。首先，通过训练分类网络并使用 CAM 方法寻找到目标鉴别性区域，接下来将输入图片中该区域擦除后作为新的训

练数据，并再次训练一个新的分类网络从而寻找其它的目标区域。该过程迭代多次后训练出多个模型，从而可以识别出目标不同的区域，进而提高目标定位的完整性。Choe 等人[12]基于擦除策略提出了一种轻量级的独立模块。该模块通过对卷积特征生成注意力图，随后随机选择使用注意力图中较高值对应的区域对原图对应区域进行擦除，或使用注意力图生成重要性系数并与原特征相乘，从而可以使得网络在保持一定分类精度的前提下发现更多目标区域，提高定位精度。Zhang 等人[10]基于对抗擦除的思想提出了一种双分支网络结构。他们提出对第一个分支网络发现的目标区域进行擦除，并将擦除后的特征送入第二个分支网络，从而驱使第二个分支网络寻找到互补的目标区域。通过这种对抗互补的方式激活更多的目标区域，从而提高定位精度。Lu 等人[15]提出一种几何约束网络(Geometry Constrained Network)模型，通过训练一个目标检测器直接输出包含目标的矩形或椭圆形区域，再利用该区域对输入图片进行擦除分别获得目标区域和背景区域，最后结合一个多任务损失函数对目标检测器输出的几何形状进行约束，以实现目标定位。

基于数据增广的方法：DeVries[17]等人提出通过对输入图片进行随机遮挡的正则化方法 Cutout，使得网络可以发现其他较弱鉴别性区域。Zhang 等人[18]基于邻域风险最小化(Vicinal Risk Minimization)提出了一种将输入数据和对应标签进行混合的数据增广策略 Mixup，以提高模型的泛化性。 Yun 等人[9]基于 Cutout 和 Mixup 提出在训练图片间进行裁剪粘贴，并且同时按混合图片的面积比例来对标签进行混合的数据增广策略，从而提高了模型的泛化能力和定位能力。Singh 等人提出 HaS (Hide-and-Seek)[8]方法，通过对输入图片进行网格划分，并且在训练过程中随机擦除一些网格区域中的图片内容，使得目标中最具鉴别性区域对网络并不总是可见，从而驱使网络发现其他鉴别性较弱的目标区域。

其它方法：除了上文所述的基于“擦除”和“数据增广”的两大类方法，还有一些学者提出了一些其它方法来解决弱监督目标定位问题。例如，Zhang 等人[11]基于蒸馏思想，提出了一种分阶段的方法逐步使用网络中间层信息获得自生成指导(Self-Produced Guidance, SPG)信息，并使用该信息分离目标和背景区域，从而为分类网络提供像素间的空间相关性辅助信息，使网络能够获得更好的定位结果。Selvaraju 等人[19]提出基于梯度加权的类激活映射(Gradient-weighted Class Activation Mapping，Grad-CAM)。相比于 CAM 方法需要修改网络结构添加全局平均池化层(Global Average Pooling, GAP), Grad-CAM 可以直接应用在现有网络上，而不需要对网络结构进行改动。Chattopadhyay 等人[20]基于 Grad-CAM 进一步对权重系数进行了优化并提出 Grad-CAM++方法。该方法对模型预测具有更好的视觉解释以及定位效果。此外，Zhu 等人[21]通过计算基于特征差异和空间距离的相异性，为网络提供了关于目标置信度的高层指导信息从而获得更好的定位效果。Zhang 等人[16]使用伪监督学习[22]提出了伪监督目标定位框架，通过使用 Wei 等人[23]提出的协同定位方法生成目标预测边界框，并将其作为伪监督信息来训练边界框(Bounding Box)回归模型，从而实现目标定位功能。本文同样使用了伪监督目标定位框架。不同于文献[16]采用的两阶段方法，需要先对训练数据集整体生成伪监督边界框信息，再基于伪监督信息进行边界框回归模型训练，本文方法在训练过程中直接将基于类别层次结构网络生成的目标预测边界框作为监督信息，用于边界框回归模型训练，从而实现端到端的网络训练。同时本文还利用了卷积神经网络可以从有噪声的数据中学习到鲁棒模式[16]，在训练过程中对伪监督信息进行了过滤，从而获得更干净的伪监督信息，并进一步提升了模型的定位精度。

3 本文所提方法

本文方法的整体结构如图 3 所示。主要由三部分组成，即“目标激活区域提取”、“层次结构分类网络”和“目标定位网络”。

在训练阶段，同时将训练图片输入层次结构分类网络和目标定位网络。其中，层次结构分类网络根据训练图片对应的根类别标签将主干网络输出的特征送入对应分支网络，并使用训练图片对应的子类别标签进行分类训练。通过在分支网络中提取对应图片的目标激活区域，可获得目标的预测边界框，进而使用该预测边界框作为目标定位网络的监督信息进行训练。

测试阶段，直接使用目标定位网络即可获得目标边界框预测。为了获得类别预测，本文按照文献[16]中的方法，使用一个在已有数据集上预训练的分类网络来获得类别预测。

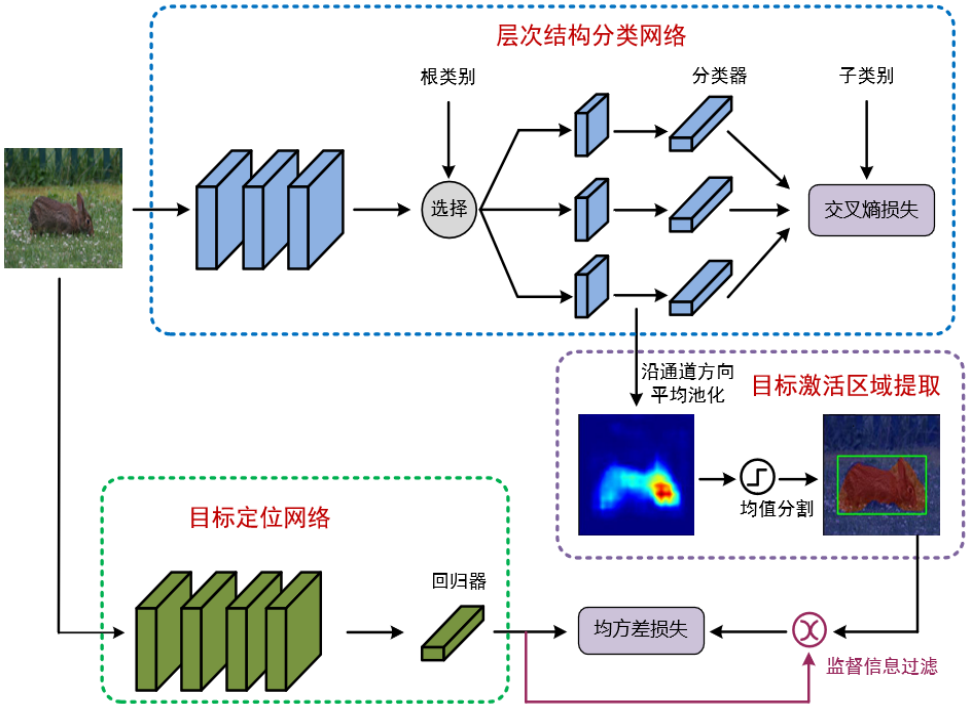

图 3　本文所提方法的整体结构。其中，“层次结构分类网络”通过将相似的类别归并到一个抽象根类别下，并单独集中在一个分支网络中进行训练，由于相似类别的图片往往具有相似的背景信息，相似的背景使得背景区域对网络不在提供足够的鉴别性信息，因此类别层次结构分类网络可以削弱算法对背景区域的注意力，从而提高定位精度。“目标激活区域提取模块”用来从分类网络特征中获得注意力图，并生成预测边界框，该预测边界框将作为“目标定位网络”的训练监督信息。在整个训练过程的后半阶段，我们利用目标定位网络学习到的更鲁棒的模式对监督信息进行过滤，提高参与训练的监督信息质量，从而进一步提高模型的定位精度。

3.1　目标激活区域提取

之前的一些研究工作表明，分类网络的深层特征实际隐含了目标的位置信息[24,25]。为了获得这种网络隐含的目标位置信息，本文对分类网络最后一层卷积输出特征$\mathbf{F} \in \mathbb{R}^{c\times h\times w}$ (其中c 表示特征通道数，$h \times w$表示特征的大小)沿通道方向执行平均池化操作，从而获得网络对输入图片的注意力图$\mathbf{A} \in \mathbb{R}^{h\times w}$。该过程表示如下：

$$\mathbf{A}_{i,j} = \frac{1}{c}\sum_{k=0}^{c} \mathbf{F}_{k,i,j} \tag{1}$$

注意力图中的像素值反映了该位置对于预测类别的重要性，因此注意力图实际表现为网络对不同位置的关注程度[26,27]。由于输入图片中的目标区域对预测结果提供了更多的鉴别信息，因此网络对目标区域的特征的激活程度更强。所以，本文认为注意力图$\mathbf{A}$隐含了目标的位置信息。

为了提取目标激活区域，本文对获得的注意力图$\mathbf{A} \in \mathbb{R}^{h\times w}$进行均值分割并过滤其中激活较弱的非目标区域，以获得更好的目标激活区域图。具体地，首先求取$\mathbf{A}$中所有值的均值，再将$\mathbf{A}$中大于等于均值的位置的值置为 1，其它置为 0，从而获得目标激活区域图$\mathbf{M} \in \mathbb{R}^{h\times w}$。该过程可以表示为：

$$\mu = \frac{1}{h\times w}\sum_{i,j} \mathbf{A}_{i,j} \tag{2}$$

$$\mathbf{M}_{i,j} = \begin{cases} 0, & \mathbf{A}_{i,j} < \mu \\ 1, & \mathbf{A}_{i,j} \geq \mu \end{cases} \tag{3}$$

为了从目标激活区域中获得目标的边界框坐标信息，本文沿用 CAM[6]中的方法，即通过使用包围框覆盖$\mathbf{M}$中的最大连通区域来获得目标边界框。

3.2 层次结构分类网络

分类网络往往倾向于关注对分类判别更具鉴别性的图片区域，因此，如果将训练图片中具有相似背景的样本集中在一起训练，那么背景区域将不再包含足够的与分类判别相关的鉴别性信息，从而可以有效减少网络对背景区域的注意力，降低背景区域被误判为前景的概率。

图片中的背景的相似性一般难以直接衡量，但本文发现类别相近的图片通常具有相似的背景，如图 2 所示，校车和出租车属于相近的类别，它们的图片通常具有相似的背景（道路背景），这为本文实现将具有一定相似背景的训练图片集中起来提供了一种间接实现方案。因此，本文提出对数据集中的所有类别进行相似性归并，从而将原始数据集中的类别归并成更高层次的少量抽象根类别。于是，原始的数据集中的类别就构成了各根类别下的子类别，且每个根类别下的这些相似子类别一般具有相似的背景。进而，本文对这些在同一抽象根类别下的多个子类别单独构建一个分支网络进行训练，从而可以抑制分支网络对背景区域的注意力，提高目标定位精度。具体地，本文使用一个具有多个卷积块的主干网络提取公共特征，之后根据归并的根类别个数创建个数相同的分支网络。这些分支网络拥有相同的结构，并均使用交叉熵损失(cross-entropy loss)对各根类别下的子类别进行分类训练。该过程如图 3 所示。类别层次结构分类网络总体损失为各分支网络损失之和，表示如下：

$$\mathcal{L}_i = -\sum_{c=1}^{C_i} y_c \log(p_c) \tag{4}$$

$$\mathcal{L}_{HN} = \sum_i \mathcal{L}_i \tag{5}$$

其中C_i表示第i个分支网络中包含的子类别数量，y_c表示真实标签，p_c表示网络预测结果。

为了将相似的类别进行归并，需要首先获得数据集中每个类别的特征表示。一个分类网络一般由卷积层、全连接层以及 softmax 层构成。假设卷积层最后一层输出特征为$\mathbf{F} \in \mathbb{R}^{k\times h\times w}$(其中$k$ 表示特征通道数，$h \times w$表示特征图的大小)，其展开为一维向量后表示为 $\overline{\boldsymbol{f}} \in \mathbb{R}^d$（这里$d = k * h * w$），全连接层权重矩阵表示为$\mathbf{W} \in \mathbb{R}^{c\times d}$（其中$c$表示类别数，这里不考虑偏置项），则对于类别$m$，全连接层输出的分数为$S_n = \sum_{i=0}^{d} \boldsymbol{\omega_m}\overline{\boldsymbol{f}}$，于是最终 softmax 层输出类别$m$的概率表示为$P_m = \frac{\exp(S_m)}{\sum_{m=0}^{c}\exp(S_m)}$。对于类别$m$的权重向量$\boldsymbol{\omega}_m \in \mathbb{R}^d$，它指示了$\overline{\boldsymbol{f}}$中每个神经元的重要性，于是对于预测类别更重要的特征便会被赋予更高的权重。本质上，$\boldsymbol{\omega}_m$ 起到了一个特征选择作用，即选择那些对类别m更具重要性的特征。换言之，$\boldsymbol{\omega_m}$可以被看作类别m的特征过滤器。因此，我们可以将$\boldsymbol{\omega}_m$作为类别m的特征表示。本文利用原始数据集训练一个分类网络，训练完成后该分类网络全连接层中的权重矩阵$\mathbf{W} \in \mathbb{R}^{c\times d}$中的每个向量$\boldsymbol{\omega}_m \in \mathbb{R}^d, m \in [0, c-1]$,分别作为类别$m$的特征表示。基于该特征表示，本文使用 K-means 方法进行对数据集中的类别进行聚类，从而实现将特征相似的类别归并到一个抽象根类别下。该过程表示如下：

$$\boldsymbol{\alpha}_i = \frac{1}{n_i}\sum_{j\in C_i} \boldsymbol{x}_j \tag{6}$$

$$\operatorname{argmin}\left(\sum_{i=1}^{K}\sum_{\boldsymbol{x}\in C_i}\|\boldsymbol{x} - \boldsymbol{\alpha}_i\|_2^2\right) \tag{7}$$

其中$\boldsymbol{x}_j$表示第j个类别的特征表示向量，C_i表示第i个簇，$\boldsymbol{\alpha}_i$表示第i个簇的中心向量，n_i表示第i个簇中的类别数量，K为聚类簇数，需要手动设置。一般而言，对于训练数据较少的数据集，聚类簇数K设置过大会导致用于每个分支网络的训练数据过少，因此此时可将聚类簇数K设置为一个比较小的数值，如$K = 2$。如果数据集比较大，则可以适当增大K。为了选择合适的聚类簇数K，本文使用轮廓系数(Silhouette Coefficient)[28]来衡量不同聚类簇数K设置下，聚类结果的相对好坏，轮廓系数值越接近 1 表示聚类效果越好。在实验部分本文分别测试了设置不同聚类簇数K的情况下，聚类结果的轮廓系数。

本文根据上述聚类方法对数据集进行相似性归并。假设通过聚类方法将原数据集划分为了K个抽象根类别，分别为$\mathrm{A}_1, \mathrm{A}_2, \ldots \mathrm{A}_k$。其中，第$i$ 个根类别A_i下存在j个子类别，分别表示为$\mathrm{A}_{i1}, \mathrm{A}_{i2}, \ldots, \mathrm{A}_{ij}$，则根据该表示可以将原始数据集中的类别分别映射成其对应的抽象根类别 A_i 和根类别下对应的子类别A_{ij}，表示为$[\mathrm{A}_i, \mathrm{A}_{i,j}]$。在训练过程中，根据输入图片对应的根类别$\mathrm{A}_i$ 选择对应的分支网络，再根据对应的子类别$\mathrm{A}_{i,j}$ 来计算交叉熵，并在该分支网络下利用 3.1 节中的方法获得目标区域。

3.3 目标定位网络

为了训练目标定位网络，本文将从类别层次结构分类网络生成的目标边界框坐标预测作为伪监督信息，对目标位置的坐标训练一个回归器。具体地，本文遵循文献[16]中的设置，使用 ResNet50[29]或 VGG16[30]的卷积层作为主干网络，构建具有两个全连接层和相应 ReLU 层的子网络作为回归器，并接入到主干网络之后构成完整的目标定位网络。假设从类别层次结构分类网络生成的预测边界框坐标表示为$\{x, y, w, h\}$，其中x, y为预测边界框的左上角坐标，w, h为预测边界框的宽和高。首先将其转换成相对值，即$x^* = \frac{x}{w_i}, y^* = \frac{y}{h_i}, w^* = \frac{w}{w_i}, h^* = \frac{h}{h_i}$，其中$w_i, h_i$分别为输入图片的宽和高。将$\{x^*, y^*, w^*, h^*\}$作为目标定位网络的监督信息，并使用均方误差损失(Mean Squared Error Loss)作为损失函数进行训练，表示如下：

$$\mathcal{L}_{LN} = \frac{1}{n}\sum_{i=1}^{n}\|y_i - y_i'\|_2^2 \tag{8}$$

其中n表示样本数量，$\boldsymbol{y}_i$表示从类别层级结构分类网络生成的，作为目标定位网络监督信息的边界框坐标向量，$\boldsymbol{y}_i'$表示目标定位网络自身预测的边界框坐标向量。

与文献[16]不同的是，本文并没有在整个训练过程中使用所有的伪监督信息。由于卷积神经网络可以容忍一些标注错误，并能从有噪声的数据中学习鲁棒的模式[16]，因此随着训练的进行，目标定位网络可以从含噪声的伪监督信息中逐渐学习到关于定位任务更好的模式。因此，在训练的后半段过程中，本文利用目标定位的输出对类别层次结构分类网络生成的伪监督坐标信息进行过滤。具体地，对于每个训练样本，本文计算目标定位网络对训练图片输出的预测边界框和类别层次结构分类网络生成的目标边界框的交并比(intersection over union, IoU)：

$$IoU = \frac{HN\_PB \cap LN\_PB}{HN\_PB \cup LN\_PB} \tag{9}$$

其中，*HN_PB* 表示训练过程中类别层次结构分类网络生成的目标预测边界框，*LN_PB* 表示训练过程中，目标定位网络预测的目标边界框。在训练的后半阶段，对于每个训练样本只有 IoU 大于 50%时，本次类别层次结构分类网络生成的预测边界框才作为目标定位网络的监督信息并参与训练。

3.4 网络总损失

如图 3 所示，本文在训练阶段同时对类别层次结构分类网络和目标定位网络进行训练，与之前方法[16]相比，本文不需要先对数据集生成伪监督信息后再训练目标定位网络，从而实现了端到端的训练。模型的总损失为类别层次结构分类网络损失(公式 5)与目标定位网络损失(公式 8)的和：

$$\mathcal{L}_{all} = \mathcal{L}_{\mathrm{HN}} + \mathcal{L}_{LN} \tag{10}$$

4 实验

为了验证本文所提方法能够获得更好的定位效果，本节在 ImageNet-1K 和 CUB-200-2011 两个数据集上进行了一系列验证和对比实验。

4.1 数据集和评价指标

本文在两个弱监督目标定位任务常用数据集 ImageNet-1K[32]和 CUB-200-2011[31]上对所提方法进行了充分的实验验证。其中，CUB-200-2011 为细粒度图像数据集，由 200 种鸟类数据组成，包含 11,788 张图片数据。其中，训练集包含 5,994 张图片，测试集包含 5,794 张图片。ImageNet-1K 数据由 1000 种类别的图片数据构成，其中包含 1,281,167 张训练图片。由于 ImageNet-1K 不包含测试集，本文遵循之前的方法使用数据集提供的 50,000 张验证集图片作为测试集进行最终模型测试，在训练阶段本文并不会使用该验证集数据。训练阶段本文不使用数据集本身提供的精确边界框注释信息，仅使用数据集提供的图片类别标签。

本文采用了弱监督目标定位普遍使用的 Top-1/Top-5 定位精度(Top-1/Top-5 Loc)[12,16]，以及已知输入图片真实类别标签情况下的定位精度(GT-Known Loc)[16]作为评价指标。其中，GT-Known Loc 表示在给定输入图片真实类别标签的情况下，模型对输入图片的预测边界框和真实边界框的交并比(intersection over union,

IoU)大于等于 50%的图片比例。Top-1 Loc 表示模型对输入图片的 Top-1 类别预测正确，并且目标预测边界框和真实边界框 IoU 大于等于 50%的图片比例。Top-5 Loc 表示模型对输入图片的类别和目标边界框的 Top-5 预测中，至少有一个类别预测正确并且目标预测边界框和真实边界框 IoU 大于等于 50%的图片比例。

4.2 实现细节

本文使用在 ImageNet-1K 数据集上预训练的网络参数作为模型初始参数，并在目标数据集上对定位网络和分类网络进行微调。

为了保证对比实验的公平性，本文采用了之前方法的数据预处理策略[10,16]。具体地，在训练阶段，本文将输入图片尺寸调整为256×256，并在调整后的图片中随机裁剪出尺寸为224×224的图片内容送入网络中进行训练。测试阶段，本文同样将输入图片尺寸调整为256×256，随后在调整后的图片中裁剪出尺寸为224×224的图片中心内容，并送入网络进行类别和位置预测。本文采用随机梯度下降算法(Stochastic Gradient Descent, SGD)对模型进行训练，初始学习率设置为 0.0005，权重衰减设置为 0.0001，动量参数设置为 0.9。在 CUB 和 ImageNet-1K 数据集上分别训练 25 和 11 轮，并每 10 和每 5 轮将学习率衰减 0.1 倍。此外，本文采用 ResNet50[29]和 VGG16[30]两种网络架构分别作为模型的主干网络。

4.3 消融实验

为了后续实验验证本文提出的类别层次结构分类网络可以有效抑制背景区域注意力，本文分别构建了不包含类别层次结构的“基线分类网络”以及“类别层次结构分类网络”，并比较二者的效果。

表 1 基于 VGG16 的网络架构

层名	输出大小	基线网络	层次结构网络
conv1	112×112	$\begin{matrix}3\times3,64\\3\times3,64\end{matrix}$	$\begin{matrix}3\times3,64\\3\times3,64\end{matrix}$
conv2	56×56	$\begin{matrix}3\times3,128\\3\times3,128\end{matrix}$	$\begin{matrix}3\times3,128\\3\times3,128\end{matrix}$
conv3	28×28	$\begin{matrix}3\times3,256\\3\times3,256\\3\times3,256\end{matrix}$	$\begin{matrix}3\times3,256\\3\times3,256\\3\times3,256\end{matrix}$
conv4	28×28	$\begin{matrix}3\times3,512\\3\times3,512\\3\times3,512\end{matrix}$	$\begin{matrix}3\times3,512\\3\times3,512\\3\times3,512\end{matrix}$
conv5	28×28	$\begin{matrix}3\times3,512\\3\times3,512\\3\times3,512\end{matrix}$	$\begin{Bmatrix}\begin{bmatrix}3\times3,512\\3\times3,512\\3\times3,512\end{bmatrix}\\ \cdots\cdots \\ \begin{bmatrix}3\times3,512\\3\times3,512\\3\times3,512\end{bmatrix}\end{Bmatrix}$
	28×28	$\begin{matrix}3\times3,1024\\3\times3,1024\end{matrix}$	$3\times3,1024$
	1×1	average pool, fc, softmax	

表 2 基于 ResNet50 的网络架构

层名	输出大小	基线网络	层次结构网络
conv1	112×112	7×7,64, stride 2	
layer1	56×56	3×3 max pool, stride 2	
		$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$
layer2	28×28	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$
layer3	28×28	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$
layer4	28×28	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{Bmatrix}\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3\\ \cdots\cdots \\ \begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3\end{Bmatrix}$
	1×1	average pool, fc, softmax	

具体地，基于 VGG16 的基线分类网络，本文删除 VGG16 网络的全连接层，之后添加 2 个大小为3×3、步长为 1、填充为 1、具有 1024 个输出通道的卷积层，以及一个全局平均池化层(global average pooling)和具有 1000(CUB 数据集上设为 200)个输出单元的全连接层。此外，本文遵循之前方法[10]的设置，将前端卷积块中的 conv4，conv5 两个卷积块后的池化操作步长设置为 1，从而使得最终特征输出大小为28×28；基于 VGG16 的类别层次结构分类网络，本文使用 VGG16 网络的 conv1-conv4 作为主干网络，随后的每个分支网络由 VGG16 网络的 conv5 与一个大小为3×3、步长为 1、填充为 1、具有 1024 个输出通道的卷积层，以及对应全连接层实现。

基于 ResNet50 的基线分类网络，本文同样将 ResNet50 网络的 layer3、layer4 的卷积步长设置为 1，从而获得尺寸为28×28的输出特征；基于 ResNet50 的类别层次结构分类网络，本文使用 ResNet50 网络的 layer1-layer3 作为主干网络，随后的每个分支网络由 ResNet50 网络的 layer4 与对应的全连接层实现。表 1、表 2 分别列出了基于 Vgg16 和 ResNet50 的基线分类网络和类别层次结构分类网络的具体参数。

首先，本文使用没有类别层次结构的基线分类网络，来测试本文所提的注意力图均值分割方法与 CAM[6]方法的定位效果，本文复现了 CAM 方法在不同网络和数据集上的定位效果，实验结果如表 3 所示。

表 3 注意力图均值分割方法和 CAM 方法的定位精度比较

	CUB			ImageNet-1K		
方法	Top1 Loc	Top5 Loc	GT- Known Loc	Top1 Loc	Top5 Loc	GT- Known Loc
VGG(ours)	**62.58**	**78.17**	**83.29**	**46.51**	**57.6**	**61.09**
VGG-CAM[6]	31.58	41.23	44.17	40.34	50.53	54.92
ResNet50(ours)	**75.09**	**88.44**	**91.01**	**53.2**	**62.75**	**65.35**
ResNet50-CAM[6]	53.21	62.88	64.72	48.22	57.3	59.92

实验结果表明基于注意力图分割方法在 CUB 和 ImageNet-1K 两个数据集上的定位能力均明显好于 CAM 方法。本文对两种方法的定位效果进行了可视化展示，如图 1 所示。通过图 1 可以发现 CAM 方法获得的目标激活区域往往集中于对分类判别最具鉴别性的部分区域，这导致 CAM 方法在小目标上定位结果较好，在大目标上定位结果不精确。而基于注意力图分割的方法对大目标和小目标均可以获得更完整的目标激活区域，因此具有更好的定位效果。本文认为这是由于基于注意力方法通过对网络最后一层卷积特征输出沿通道方向执行平均池化操作，可以获得网络对输入图片的空间注意力图，这种注意力图反映了与预测类别相关的语义部分，而不仅仅是目标最具鉴别性的部分，因此可以找到更完整的目标区域，有利于提高定位精度。

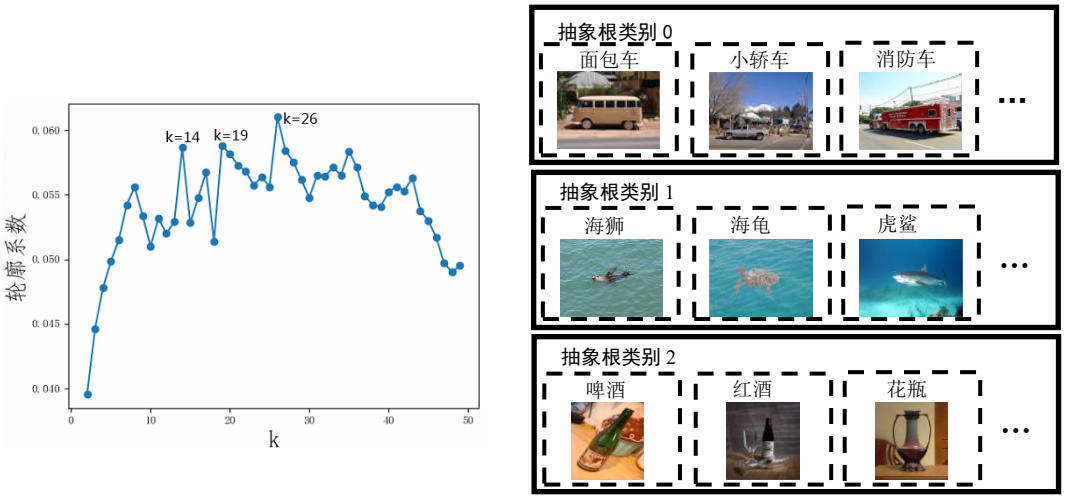

图 4　使用 K-means 对 ImageNet-1K 数据集中的原始类别进行聚类，左图为在不同聚类簇数K设置下聚类结果的轮廓系数。右图为选择 K=14 时聚类结果的一些示例，如第一行面包车、小轿车、消防车等类别被归并到一个抽象根类下。正如算法设计所期望的，这些相似类别的图片背景具有一定的相似性。

进一步，本文对类别层次结构有利于抑制背景区域注意力进行实验验证。首先，我们使用 2.2 节中描述的方法对数据集中的所有类别进行相似性归并。对于 CUB 数据集，由于其训练数据较少，使用 K-means 进行类别聚类时，如果簇数K设置过大会导致分支网络训练数据过少，因此本文仅设置聚类簇数$K = 2$。对于 ImageNet-1K 数据集，本文测试了$K \in [2,50]$情况下聚类结果的轮廓系数(Silhouette Coefficient)，如图 4 中左边所示。考虑到计算资源本文选取了聚类簇数$K = 14$来实施聚类，并在图 4 右边展示使用$K = 14$进行

聚类后获得的根类别中的三个根类的一些示例图片。可以发现相似的类别被很好地归并到了一个抽象根类别下，并且这些类别的图片背景具有一定的相似性。例如，第一行显示面包车、小轿车、消防车等类别被归并到抽象根类别 0 下，并且这些类别的图片往往具有相似的道路背景；第三行显示了啤酒、红酒、花瓶等类别被归并到抽象根类别 2 下，并且这些类别的图片往往具有相似的桌面背景。

基于上述聚类结果本文构建了类别层次结构分类网络，并分别测试了基线分类网络和类别层次结构分类网络下基于注意力图分割获得的目标定位精度(这里仅考虑已知目标类别情况的下的定位精度，以便更准确地衡量模型的定位能力)。实验结果如表 4 所示：

表 4　不同方法的 GT- Known Loc 指标比较

方法	backbone	CUB	ImageNet-1K
基线分类网络	VGG16	83.2	61.09
类别层次结构分类网络		83.05	**64.59**
基线分类网络	ResNet50	91.01	65.35
类别层次结构分类网络		90.42	**68.66**

从表 4 实验结果可以发现，在 ImageNet-1K 数据集上，基于 VGG16 和 ResNet50 两种网络架构，使用类别层次结构分类网络实现的定位精度相比基线分类网络均有明显提升，在 VGG16 网络架构下相比基线分类网络提升了 3.5%，在 ResNet50 网络架构下相比基线分类网络提升了 3.3%。以上实验结果表明，类别层次结构分类网络通过将每个根类别中的训练图片集中在单个分支网络中，实现了将背景具有一定相似性的图片集中在一个分支网络中进行训练，从而抑制网络对背景区域注意力，并提升目标定位效果。在细粒度数据集 CUB 上，由于所有训练数据均属于鸟类，训练图片的背景已经具有相似性，因此类别层次结构分类网络相比基线分类网络的定位精度基本保持不变，另一方面，ImageNet-1K 数据集中包含了各种背景，而表 4 实验数据显示在 CUB 数据集上的定位效果明显好于 ImageNet-1K 数据集，这也支持了本文的观点，即将具有相似背景的训练数据集中在一起更有利于网络抑制背景区域注意力，提高对目标区域的定位精度。

最后，本文选择使用基于 ResNet50 的类别层次结构分类网络来生成目标边界框伪监督信息，并将其作为目标定位网络的监督信息进行训练，从而构建本文的整个完整系统。为了验证本文在 3.3 节中提出的伪监督信息过滤策略的有效性，本文测试了整个训练过程中使用全部类别层次结构分类网络生成的伪监督信息训练出来的目标定位网络性能，以及训练后半段仅使用过滤后的伪监督信息训练出来的目标定位网络性能。表 5 列出了在 ImageNet-1K 数据集上，各个模块在提升模型定位精度上的效果。从表 5 可以发现类别层次结构分类网络本身可以达到的定位精度为 68.66，而将类别层次结构分类网络对训练数据生成的预测边界框作为伪监督信息，并训练一个目标定位网络时，目标定位网络的定位精度可以提高到 69.38，这验证了卷积网络可以从噪声标签中学习到一些鲁棒的信息。最后，通过在训练过程中对这些伪监督进行过滤，本文将模型的定位精度提升到了 70.5。

表 5　ImageNet-1K 数据集上本文方法各个模块对定位精度的提升效果

方法	GT- Known Loc
基于注意力的基线分类网络	65.35
+ 使用类别层次结构	68.66
+ 结合目标定位网络进行伪监督训练	69.38
+ 训练时对伪监督信息过滤	**70.5**

4.4 层次结构效果展示

为了更直观地展示类别层次结构分类网络抑制背景区域注意力的能力，本文对使用类别层次结构分类网络和基线分类网络获得的目标激活区域进行了可视化展示，如图 5 所示，从图 5 可以发现，类别层次结构分类网络相比基线分类网络可以消除大量非目标区域注意力。这验证了本文所提的类别层次结构网络在

抑制网络对背景无关区域注意力的有效性。类别层次结构分类网络通过将相近的类别归并在同一个抽象根类别下，而相近的类别的图片往往具有相似的背景信息(如图 4)，背景的相似性使得背景区域对网络的预测不再具有足够的鉴别性信息，从而实现削弱网络对背景区域的注意力，提高定位精度。此外，从图 5 也可以发现本文方法对于输入图片中存在多个目标实例的情况，也能获得较好的目标激活区域。

图 5　基线分类网络和类别层次结构分类网络下获取的目标激活区域对比，第 1、2、3 列分别展示了类别层次结构分类网络消除了目标对象区域以外的大量背景干扰(第 1、2 列目标对象为“狗”，第 3 列目标对象为“牛”)。第 4、5 列展示了本文方法对于目标存在多实例的情况也能获得较好的目标激活区域，并且类别层次结构分类网络相比基线分类网络可以获得精度更高的目标激活区域。

4.5　与现有方法比较

本节对本文提出的方法和 CutMix[9]、ACoL[10]、SPG[11]、ADL[12]、GC-Net[15]、DA-Net[33]、DDT-PSOL[16]方法在 CUB 和 ImageNet-1K 数据集上的定位效果进行了比较，本文分别测试了基于 VGG16 和 ResNet50 的定位网络可以实现的目标定位精度。从实验数据来看，本文方法在不同网络架构和不同数据集上均取得了更好的定位效果。

表 6　CUB 数据集上本文方法和其它方法的性能比较

方法	backbone	Top1 Loc	Top5 Loc	GT-Known Loc
ACoL[10]	VGG16	45.92	56.51	-
ADL[12]		52.36	-	-
CutMix[9]		52.53	-	-
DA-Net[33]		52.52	61.96	67.7
GC-Net[15]		63.24	75.54	81.1
DDT-PSOL[16]		66.3	84.05	-
Ours		**69.26**	**85.83**	**91.2**
ADL[12]	ResNet50	62.29	-	-
CutMix[9]		54.81	-	-
DDT-PSOL[16]		70.68	86.64	-
Ours		**74.73**	**87.21**	**90.3**

表 6 展示了 CUB 数据集上本文方法和其它方法的定位性能数值比较。由表中的数据可知，基于 VGG16 网络架构，本文方法相比当前最优的弱监督目标定位方法，在 Top-1/Top-5 Loc 指标上分别提高了 2.9%和 1.7%。基于 ResNet50 网络架构，本文方法相比当前最优的弱监督目标定位方法，在 Top-1/Top-5 Loc 指标上分别提高了 4%和 0.57%。

表 7 展示了在 ImageNet-1K 数据集上本文方法和其他方法的定位性能数值比较。由表中数据可知，基于 VGG16 网络架构，本文方法相比当前最优的弱监督目标定位方法，在 Top-1/Top-5 Loc 指标上分别提高了 2.8%和 4.4%。基于 ResNet50 网络架构，本文方法与当前最优的弱监督目标定位方法相比，在 Top-1/Top-5 Loc 指标上分别提高了 2.2%和 3.3%。

表 7 ImageNet-1K 数据集上本文方法和其它方法的性能比较

方法	Top1 Loc	Top5 Loc	GT- Known Loc
SPG-Inception[11]	48.6	60	64.69
DA-Net-Inception[33]	47.53	58.28	-
GC-Net-Inception[15]	49.06	58.09	-
ACoL-VGG[10]	45.83	59.43	62.96
ADL-VGG[12]	44.92	-	-
CutMix-VGG[9]	43.45	-	-
DDT-PSOL-VGG[16]	50.89	60.9	64.03
Ours-VGG	**53.7**	**65.3**	**69.35**
ADL-ResNet50[12]	48.53	-	-
CutMix-ResNet50[9]	47.25	-	-
DDT-PSOL-ResNet50[16]	53.98	63.08	65.44
Ours-ResNet50	**56.18**	**66.42**	**70.5**

以上实验结果表明，与当前弱监督目标定位方法相比，本文方法在不同的网络架构以及不同的公共数据集上均能获得更好的目标定位精度。此外，我们对本文方法和 CAM[6]、ADL[12]方法的定位效果进行了可视化对比，如图 1 所示。通过图 1 可以发现，CAM 方法在小目标上具有较好定位效果，在大目标定位效果较差，往往不能定位出目标完整区域。ADL 方法缓解了 CAM 方法在大目标上定位结果不完整的问题，但由于使用擦除操作，使得寻找到的较弱鉴别性区域包含了背景区域，因此导致定位结果不精确。而本文方法通过对类别层次结构分类网络中提取的空间注意力图进行均值分割，可以获得更完整、精确的定位结果。

5 总结

本文提出了一种基于注意力机制及类别层次结构的弱监督目标定位方法。与当前弱监督目标定位方法相比，本文提出的方法可以获得更完整的目标激活区域，并且可以有效地削弱网络对背景区域的注意力，从而获得更精确的目标定位结果。通过大量的实验比较，验证了本文方法在目标定位上的有效性和泛化性，并且和当前典型弱监督目标定位方法相比，本文方法在多个目标定位评价指标上均具有优越性。在未来的工作中，我们将进一步探索更好的背景相似性度量方法，并通过这种度量方法对具有相似背景的训练数据进行聚类，以进一步提高目标定位精度。

References:

[1] Sermanet Pierre, Eigen David, Zhang Xiang, Mathieu Michaël, Fergus Robert and LeCun Yann. OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks. arXiv:1312.6229, 2014

[2] Redmon Joseph, Divvala Kumar Santosh, et al. You Only Look Once: Unified, Real-Time Object Detection. arXiv:1506.02640,2016

[3] Ren Shaoqing, He Kaiming, Girshick B. Ross and Sun Jian. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Trans Pattern Analysis and Machine Intelligence, 2017, pp. 1137-1149

[4] Liu W , Anguelov D , Erhan D, et al. SSD: Single Shot MultiBox Detector. arXiv:1512.02325, 2016

[5] Pei W, Xu YM, Zhu YY, Wang PQ, Lu MY, Li F. The target detection method of aerial photography images with improved SSD. Ruan Jian Xue Bao/Journal of Software, 2019,30(3):738−758 (in Chinese). http://www.jos.org.cn/1000-9825/5695.htm

[6] Zhou Bolei, Khosla Aditya, et al. Learning Deep Features for Discriminative Localization. EEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, 2016, pp. 2921-2929

[7] Li Y, Liu Y, Liu GJ, Guo MZ. Weakly supervised image semantic segmentation method based on object location cues. Ruan Jian Xue Bao/Journal of Software, 2020,31(11):3640−3656 (in Chinese). http://www.jos.org.cn/1000-9825/5828.htm

[8] Singh Kumar Krishna and Lee Jae Yong.Hide-and-Seek: Forcing a Network to be Meticulous for Weakly-supervised Object and Action Localization. arXiv:1704.04232, 2017

[9] Yun Sangdoo, Han Dongyoon, Oh Joon Seong, Chun Sanghyuk, Choe Junsuk and Yoo Youngjoon "CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. IEEE/CVF International Conference on Computer Vision, Seoul, Korea (South), 2019, pp. 6022-6031

[10] Zhang Xiaolin, Wei Yunchao, et al. Adversarial Complementary Learning for Weakly Supervised Object Localization. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 1325-1334

[11] Zhang Xiaolin, Wei Yunchao, Kang Guoliang, Yang Yi and Huang Thomas. Self-produced Guidance for Weakly-supervised Object Localization. arXiv:1807.08902, 2018

[12] Choe Junsuk and Shim Hyunjung. Attention-Based Dropout Layer for Weakly Supervised Object Localization. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 2214-2223

[13] Wei Yunchao, Feng Jiashi, et al. Object Region Mining with Adversarial Erasing: A Simple Classification to Semantic Segmentation Approach. IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, 2017, pp. 6488-6496

[14] Wonho Bae, Noh Junhyug. Rethinking Class Activation Mapping for weakly supervised object localization. European Conference on Computer Vision. Glasgow, Scotland, 2020

[15] Lu Weizeng, Jia Xi, et al. Geometry Constrained Weakly Supervised Object Localization. arXiv:2007.09727, 2020

[16] Zhang Chenlin, Cao Yunhao and Wu Jianxin. Rethinking the Route Towards Weakly Supervised Object Localization. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 2020, pp. 13457-13466

[17] Devries Terrance and Taylor W. Graham. Improved Regularization of Convolutional Neural Networks with Cutout. arXiv:1708.04552, 2017

[18] Zhang Hongyi, Cissé Moustapha, Dauphin N. Yann and Lopez-Paz David. mixup: Beyond Empirical Risk Minimization. arXiv:1710.09412, 2018

[19] Selvaraju R. Ramprasaath, Cogswell Michael, Das Abhishek, Vedantam Ramakrishna, Parikh Devi and Batra Dhruv. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision, no. 2 (2020): 336-359

[20] Chattopadhyay Aditya, Sarkar Anirban, Howlader Prantik and Balasubramanian N. Vineeth. Grad-CAM++: Generalized Gradient-based Visual Explanations for Deep Convolutional Networks. workshop on applications of computer vision, 2018

[21] Zhu Yi, Zhou Yanzhao, Ye Qixiang, Qiu Qiang and Jiao Jianbin. Soft Proposal Networks for Weakly Supervised Object Localization. IEEE International Conference on Computer Vision, Venice, 2017, pp. 1859-1868

[22] Lee, D.-H. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In Workshop on Challenges in Representation Learning, ICML 2013

[23] Wei Xiushen, Zhang Chenlin, Wu Jianxin, Shen Chunhua, and Zhou Zhi-Hua. Unsupervised object discovery and co-localization by deep descriptor transformation. Pattern Recognition, 88:113–126, 2019

[24] Zeiler, M. and Fergus, R. Visualizing and understanding convolutional networks. European Conference on Computer Vision., Zürich, Switzerland, 2014

[25] Zhou Bolei, Khosla Aditya, Lapedriza Àgata, Oliva Aude and Torralba Antonio.Object detectors emerge in deep scene cnns. International Conference on Learning Representations, 2015

[26] Woo Sanghyun, Park Jongchan, et al. CBAM: Convolutional Block Attention Module. European Conference on Computer Vision., Munich, Germany, 2018

[27] Park Jongchan, Woo Sanghyun, et al. BAM: Bottleneck Attention Module. arXiv:1807.06514, 2018

[28] Rousseeuw P.J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. Journal of Computational and Applied Mathematics 20(1987):53-65

[29] He Kaiming, Zhang Xiangyu, Ren Shaoqing and Sun Jian. Deep Residual Learning for Image Recognition, IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016 pp. 770-778

[30] Simonyan Karen and Zisserman Andrew. Very Deep Convolutional Networks for Large-Scale Image Recognition. International Conference on Learning Representations, 2015

[31] Wah Catherine, Branson Steve, Welinder Peter, Perona Pietro,and Belongie Serge. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy,Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3):211–252, 2015

[33] Xue Haolan, Liu Chang, Wan Fang, Jiao Jianbin, Ji Xiangyang and Ye Qixiang. DANet: Divergent Activation for Weakly Supervised Object Localization. IEEE International Conference on Computer Vision, Seoul, Korea (South), 2019, pp. 6588-6597

附中文参考文献:

[5] 裴伟,许晏铭,朱永英,王鹏乾,鲁明羽,李飞.改进的 SSD 航拍目标检测方法.软件学报,2019,30(3):738−758. http://www.jos.org.cn/1000-9825/5695.htm

[7] 李阳,刘扬,刘国军,郭茂祖.基于对象位置线索的弱监督图像语义分割方法.软件学报,2020,31(11):3640−3656. http://www.jos.org.cn/1000-9825/5828.htm