**图像处理**

**文献阅读**

**姓名：臧亚宁 学号：202318019420011**

**Title: Attention to Scale: Scale-aware Semantic Image Segmentation**

本文提出了一种注意力机制，在全卷积神经网络（FCNs）中引入多尺度特征，并且在语义图像分割任务上取得了很好的性能。传统的提取多尺度特征的方法是将多个尺寸的输入图像输入到共享的深度网络中，然后将生成的特征合并用于图像分割。本文中提出的注意力机制，它可以在每个像素位置上对多尺度特征进行软加权，并学习到权重参数。本文将这个注意力模型与最先进的语义图像分割模型相结合，并进行联合训练，同时使用多尺度输入图像。实验证明，所提出的注意力模型不仅优于平均池化和最大池化，而且可以用于直观地可视化不同位置和尺度处特征的重要性。此外，本文还展示了在合并多尺度特征时，对每个尺度的输出进行额外监督对于取得更优的性能是很重要的。

# 网络背景

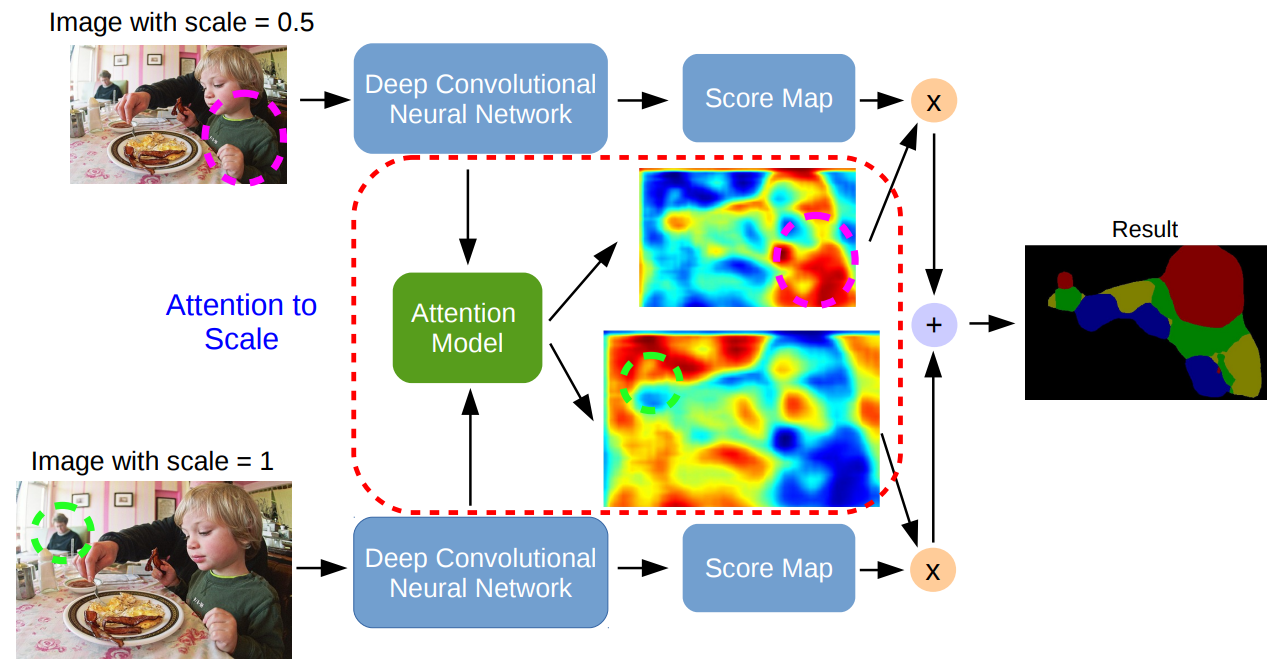
最近，基于全卷积网络（FCNs）的各种方法在几个语义分割基准测试中展示了非常好的结果。在这些模型中，成功进行语义分割的关键要素之一是使用多尺度特征。这其中主要有两种类型的多尺度特征模型。  
 第一种类型，称之为skip-net，结合了FCN（全卷积网络）中间层的特征。由于感受野尺寸逐渐增大，skip-net中的特征具有多尺度的特点。在训练过程中，skip-net通常采用两步策略。首先，它训练深层网络的主干部分，然后在多尺度特征提取阶段进行固定或微调。这种策略的问题在于训练过程不理想（即分类器训练和特征提取是分开的），而且训练时间通常较长（例如三到五天）。

第二种类型，称之为共享网络（share-net），将输入图像调整为几个尺度，并将每个尺度通过共享的深度网络进行处理。然后，它融合了多尺度特征然后计算最终的预测。共享网络不需要上述提到的两步训练过程。

|  |  |
| --- | --- |
|  |  |

图‑1：skip-net模型（左图）和share-net模型（右图）

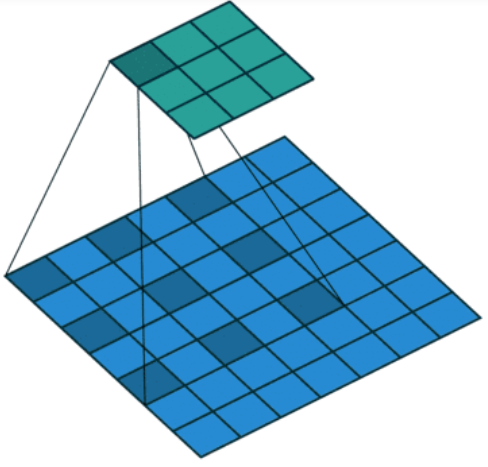
本文中将语义分割模型Deeplab v1调整为一个了共享网络，并使用软注意力模型来实现在不同比例的平均池化和最大池化，如图1-2所示。本文所提出的注意力模型根据图像中呈现的物体的比例对多尺度特征进行加权（例如，模型学习对粗糙尺度上的大物体的特征进行权重放大）。对于每个比例尺度，注意力模型输出一个权重图，然后对所有比例尺度上的FCN生成的得分图进行加权求和以进行分类。



图‑2：模型结构

# 网络实施细节

本文用的基础模型为DeepLab，DeepLab采用了VGG-16网络。将原始VGG-16网络的最后几个全连接层被转换为卷积层（例如，最后一层使用尺寸为1×1的空间卷积核）。将原始的VGG-16网络中的最后两个步长为2的池化层修改为步长为1的池化层，此时原始VGG-16下采样倍率从32变为8，再通过空洞卷积（如图‑1）对特征图做采样扩大感知，以恢复到原始图像的分辨率。此时得到DeepLab 的其中一个变体DeepLab-LargeFOV。本文主要关注DeepLab-LargeFOV。



图‑1：空洞卷积示意图

本文中，根据share-net，假设输入图像为多个不同尺度。每个尺度经过DeepLab进行处理，并产生尺度s的得分图，记为，其中表示遍历所有空间位置，。，其中C是感兴趣类别的数量。通过双线性插值将的得分图调整为与分辨率最大的尺度相同的分辨率。我们将定义为所有尺度处得分图的加权和。

权重定义为

其中表示由注意力模型生成的的得分图在位置和尺度上的值决定，权重反映了位置和尺度上特征的重要性。因此，注意力模型决定不同位置和尺度上特征需要付出多少注意力。它进一步使我们能够通过可视化权重来可视化每个尺度上的注意力。

注意力模型对每个尺度和位置都计算了一个软权重，它使得模型能够进行反向传播。因此，本文能够同时实现注意力模型和FCN（即DeepLab）部分的端到端训练。

本文中使用经过逐像素标注的图像进行训练。最终模型的输出结果是将所有尺度合并后通过一个softmax函数得到的。并且使用随机梯度下降的方法优化平均交叉熵损失函数。网络参数使用经过ImageNet预训练的VGG-16模型进行初始化。

本文除了对最终输出引入的监督之外，还为每个尺度的FCN增加了额外的监督。这样做的动机是希望将鉴别性特征（在池化或注意模型之后）合并到最终的分类器输出中。使用鉴别性特征训练的鉴别性分类器在分类任务中表现更好。本文不是向中间层添加额外的监督，而是为每个尺度的DeepLab的最终输出注入额外的监督，以使要合并的特征训练得更有区分度。具体来说，总损失函数包含个交叉熵损失函数（一表示用于最终输出的交叉熵函数，表示每个尺度的交叉熵损失函数），每个损失函数的权重为1。在训练过程中，根据输出分辨率适当地对ground truth进行下采样。

# 实验分析

本文从三个方面进行实验：多尺度输入（从一个尺度到三个尺度，其中），不同的方法（平均池化、最大池化或注意力模型）来合并多尺度特征，以及使用或不使用额外监督进行训练。并且在三个数据集上实验本文的方法，包括PASCAL-Person-Part ，PASCAL VOC 2012 和MS-COCO 2014的一个子集。  
 网络架构：本文的网络框架使用公开的模型DeepLab-LargeFOV，它修改了VGG-16 net的最后几个全连接层。

训练：本文使用小batch对SGD进行训练。将batch大小设置为30张图像，初始学习率为0.001（最后一个分类器层为0.01）。在经过2000次迭代后，学习率乘以0.1。使用0.9的动量和0.0005的权重衰减。在NVIDIA Tesla K40 GPU上训练网络需要大约21小时。在训练过程中，本文的模型使用了所有缩放后的输入并进行联合训练。因此，总的训练时间是一个普通的DeepLab-LargeFOV的两倍。一个PASCAL图像的平均推理时间为350毫秒。

评估指标：用IOU指标来评估效果。

本文在PASCAL VOC 2010数据集的语义部分分割上进行了实验。本文专注于该数据集中的人体部分，其中包含更多的训练数据和大规模的变化。具体来说，该数据集包含了每个人的详细部位注释，包括眼睛、鼻子等。将这些注释合并为头部、躯干、上臂/下臂和上腿/下腿，从而得到六个人物部位类别和一个背景类别。本文只使用包含人物的图像进行训练（1716张图像）和验证（1817张图像）。  
 改进DeepLab：在图3-1中展示了使用DeepLab-LargeFOV时的结果。发现使用两个输入尺度比使用一个输入尺度要好，并且与使用平均池化或注意力模型结合使用三个输入尺度相比稍微更好。当合并三个尺度输入时，要合并的特征必须具有足够的区分性，否则直接融合会降低性能。另一方面，最大池化似乎对这种效应具有鲁棒性。无论使用多少个尺度，我们的注意模型都比平均池化和最大池化产生更好的结果。本文进一步在图3-2中展示了最大池化和本文的注意力模型产生的权重图，清楚地显示出本文的注意力模型在不同尺度学习到了更好的权重图。此外，本文表明为每个尺度的FCN引入额外的监督明显提高了性能（见w/E-Supv列），无论采用何种融合方案。结果表明，为合并多尺度特征增加额外的监督对性能至关重要。最后，将本文中的提出的方法与DeepLab-MSc-LargeFOV进行比较，后者利用中间层的特征进行分类。本文的最佳模型（56.39%）比DeepLab-MSc-LargeFOV（53.72%）提高了2.67%的性能。

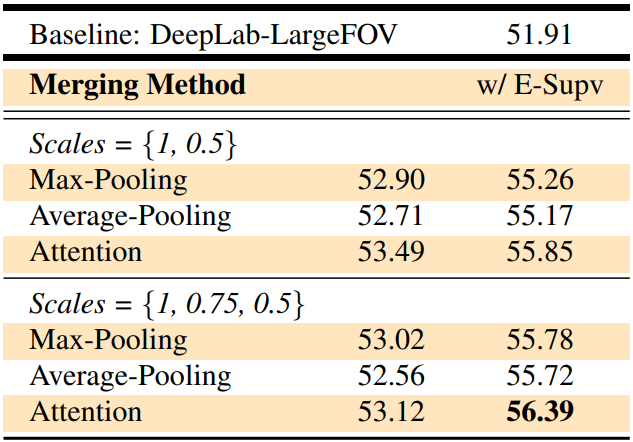


图3‑1：在PASCAL-Person-Part数据集上的结果，w/E-Supv表示引入额外监督

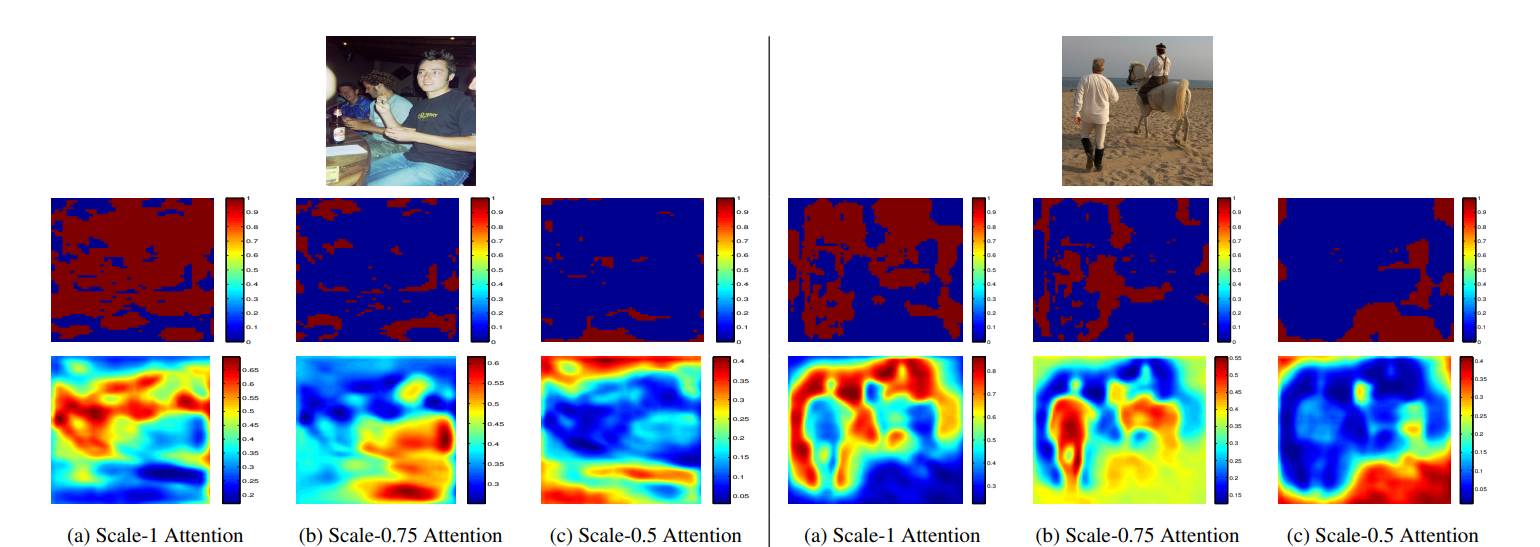


图3‑2：最大池化产生的权重图（第2行）和注意力模型产生的权重图（第3行）。注意，我们的注意力模型学习到了更好的不同尺度权重图的能力。（a）Scale-1 attention（即尺度的权重图）捕捉到了小尺度的物体。（b）Scale-0.75 attention通常关注中等尺度的物体。（c）Scale-0.5 attention强调背景上下文信息。

在本数据集中的所有实验中，本文的注意力模型采用卷积化的fc7特征作为输入，并采用由两层组成的FCN（第一层具有512个3×3的滤波器，第二层具有S个1×1的滤波器，其中S是所采用的尺度数量）。本文发现只使用一个层的注意力模型，将第一层的卷积核大小改为1×1，以及改变第一层滤波器的数量。性能变化不大；性能下降范围为0.1%到0.4%。此外，本文使用fc8作为注意力模型的特征会导致较差的性能（下降约0.5%），fc6和fc7也有类似的结果。本文还尝试了增加一个尺度（总共四个尺度：），然而，性能下降了0.5%。其中尺度为s=0.25产生的得分图太小，无法提供有用的信息。

本文提出的注意力模型不仅能够获得更好的性能，还能产生更可解释的权重图。具体来说，注意力模型学习到的尺度1注意力（即尺度s = 1时学习到的权重图）通常聚焦于小尺度的物体，尺度0.75的注意力集中在中等尺度的物体上，而尺度0.5的注意力通常会对大尺度的物体或背景赋予更大的权重，因为当图像缩小到原始分辨率的一半时，更容易捕捉到最大尺度的物体或背景的上下文信息。但是在图3-3的底部，展示了两个效果不好的示例。出现这些的原因是由于人体姿势的特殊性或者服装与人体部位之间的混淆。第一个问题可能通过获取更多数据来解决，而第二个问题则比较难解决，因为人体部位大部分时候被衣物遮盖住了。

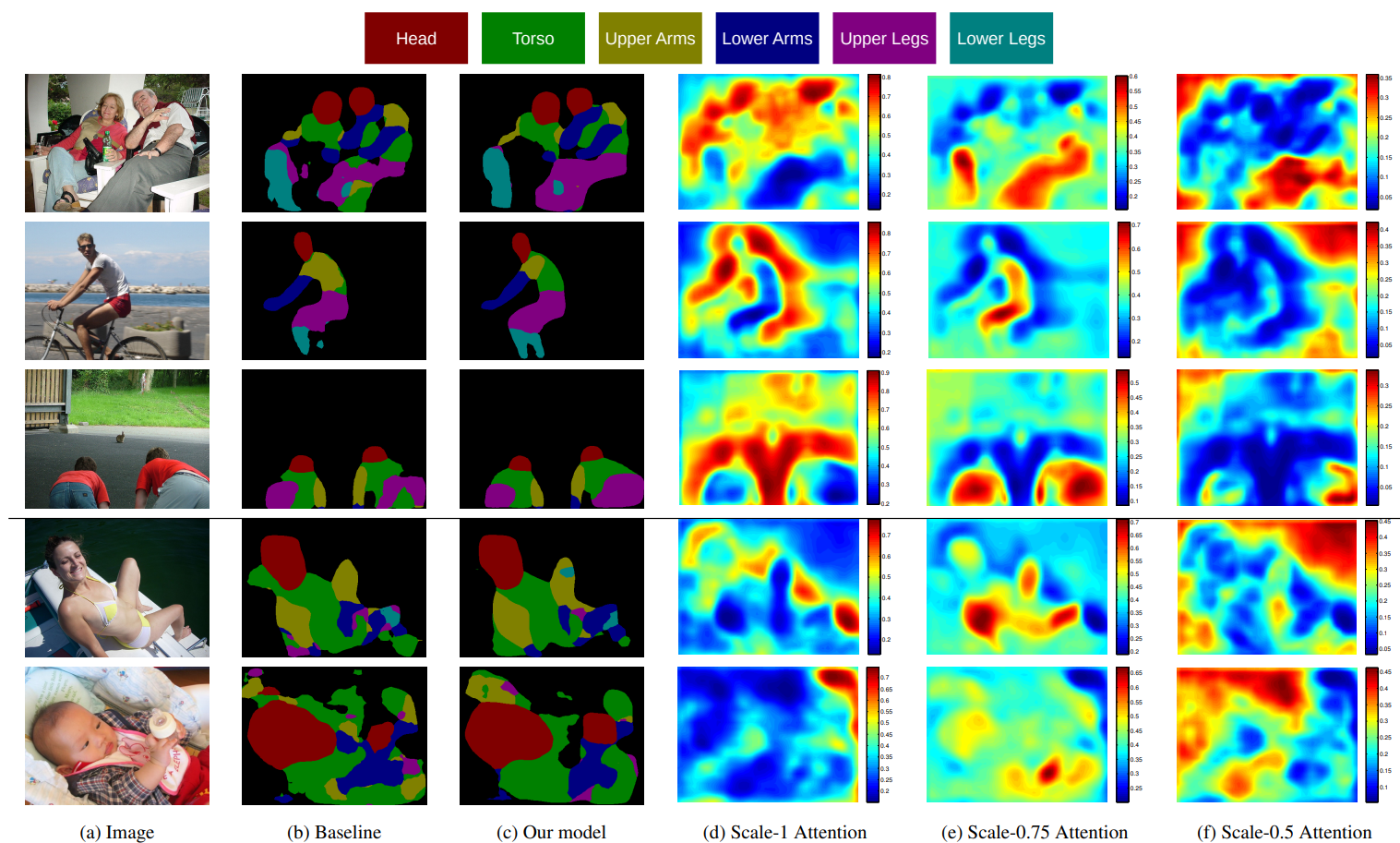


图3‑3：在PASCAL-Person-Part验证集上的结果。以DeepLab-LargeFOV单尺度输入为基准。本文模型使用了三个尺度输入、注意力模型和额外的监督。尺度-1的注意力捕捉小尺度的部分，尺度-0.75的注意力捕捉中等尺度的躯干和腿部，而尺度-0.5的注意力则关注大尺度的腿部和背景。底部两行显示了失败的例子。

在图像分割检测实验中，采用HRNetV2作为网络框架，softmax作为损失函数，平均分级类别交并比作为验证指标。同时，在本实验中采用三个数据集进行实验，分别为两个场景解析数据集PASCAL-Context、Cityscapes和一个人体解析数据集LIP，并且在训练过程中采用了数据增强策略、随机梯度下降优化器、动态调整学习率策略。实验在Cityscapes测试集上的结果如图3‑2所示，在PASCAL-Context测试集上的结果如图3‑3所示，在LIP测试集上的结果如图3‑4所示。

PASCAL VOC 2012数据集包含20个前景对象类别和一个背景类别。按照相同的实验标准，并且本文对原始训练集进行了扩充。

本文尝试了仅在ImageNet上进行预训练的情况。以DeepLab-LargeFOV和DeepLab-MSc-LargeFOV为基础修改的模型的在验证集上的性能分别为62.28%和64.39%。如图3-4所示。本文发现了与PASCAL-Person-Part数据集相似的实验结果：

(1)使用两种输入尺度优于单一输入尺度。

(2)为了将三个输入尺度合并并获得更好的性能，尤其是对于平均池化和提出的注意力模型，需要增加额外的监督。

(3) 在使用三个输入尺度、注意力模型和额外监督的情况下，获得最佳性能（相对于DeepLabLargeFOV基线改进了6.8%），其性能比DeepLab-MSc-LargeFOV（64.39%）提高了4.69%。

本文在图3-5中还展示了最佳模型在测试集上的结果。首先，注意力模型相比平均池化提高了1%，与在验证集上的结果一致。然后，本文提出的模型与DeepLab-LargeFOV和DeepLab-MSc-LargeFOV 进行比较。发现，本文提出的模型比DeepLab-LargeFOV提高了6.4%，比DeepLab-MSc LargeFOV提高了4.5%。最后，本文提出的模型与另外两种方法ParseNet 和TTI zoomout v2 进行比较。ParseNet将图像级特征作为全局上下文信息进行整合。可以将ParseNet视为利用多尺度特征的特殊情况，其中整个图像由图像级特征进行总结。TTI zoomout v2也利用了不同空间尺度的特征。如表中所示，本文提出的模型优于它们。

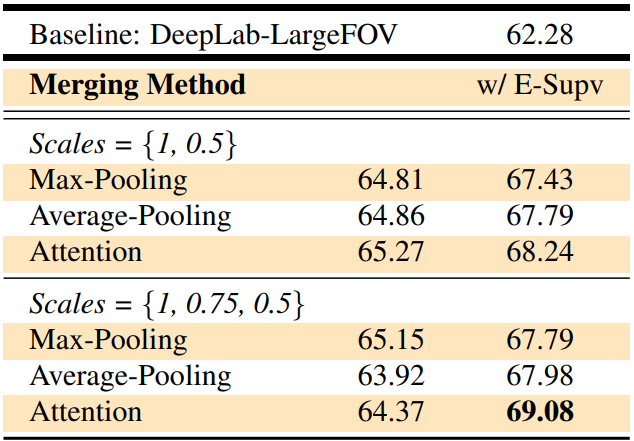


图3‑4：在PASCAL VOC 2012数据集上的结果

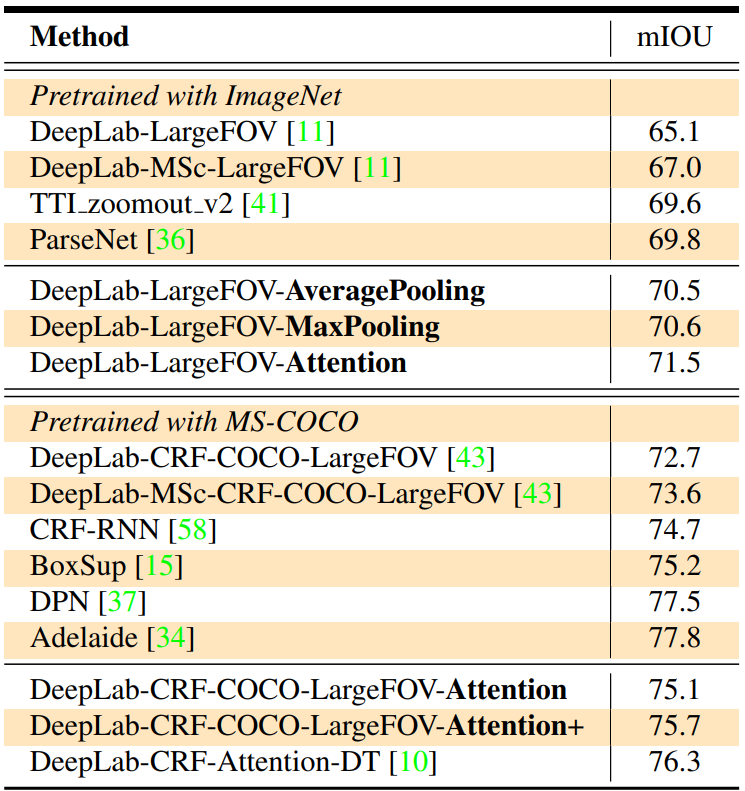


图3‑5：在PASCAL VOC 2012测试集上的IOU

本文还做了使用MS-COCO进行预训练的实验：使用基准模型DeepLab-LargeFOV在MS-COCO 2014数据集上进行了预训练。如图3-6所示，本文再次观察到类似的实验结果，本文提出的最优的模型仍然比DeepLab-LargeFOV基准模型提高了3.84%。我们还在图3-5的底部报告了在测试集上的最优模型。为了与测试集上报告的DeepLab变体进行公平比较，我们采用了全连接CRF作为后处理。从表中可以看出，我们的模型达到了75.1%的性能。在分别使用DeepLab-CRF-LargeFOV和DeepLab-MSc-CRF-LargeFOV情况下，性能分别提高了2.4%和1.5%。本文在训练过程中还采用数据增强技术，通过随机缩放输入图像（从0.6到1.4），使模型额外提升了0.6%的性能。

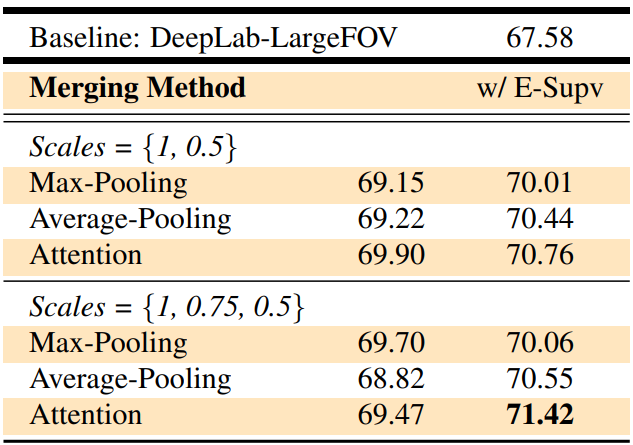


图3‑6：在PASCAL VOC 2012数据集上的结果

在MS-COCO 2014数据集上实验，此数据集包含80个前景对象类别和一个背景类别。训练集约有80,000张图像，验证集有40,000张图像。本文从训练集中随机选择了10,000张图像，并从验证集中选择了1,500张图像。

对DeepLab的改进：除了观察到与以前相似的结果外，还发现基于DeepLab-LargeFOV的模型由于MS-COCO数据集的难度（如大尺度对象的变化和更多的对象类别）而只有较低平均IOU 31.22%，如图3-7所示。然而，采用多尺度输入、注意力模型和额外的监督仍然可以比DeepLab-LargeFOV基准提高4.6%，比DeepLab-MSc-LargeFOV提高4.17%（31.61%）。还发现采用平均池化和注意力模型作为合并方法的结果非常相似。本文认为许多小物体类别（如叉子、鼠标和牙刷）的预测准确率极低，因此降低了改进的效果。

但是小物体的识别是一个非常困难的问题。在图3-8中展示了对于人类别的性能，因为人类别在该数据集中频繁出现并且以不同的尺度出现。如表中所示，本文提出的模型取得了一定的改进。

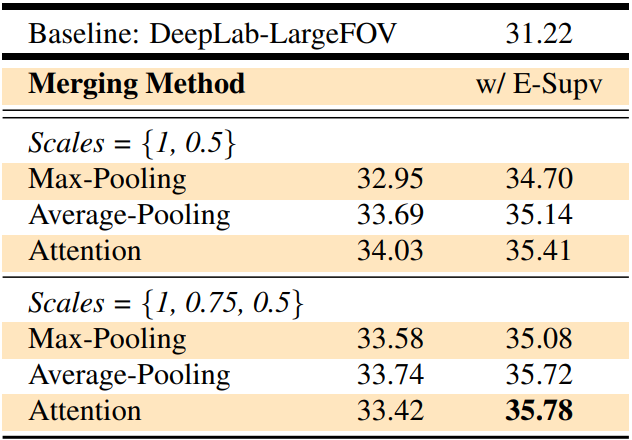


图3‑7：在MS-COCO的子数据集上的结果，基于 DeepLab-LargeFOV模型

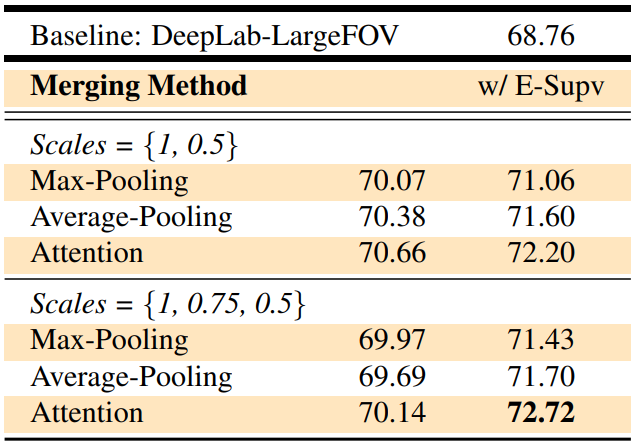


图3‑7：在MS-COCO的子数据集上对person类别的结果，基于 DeepLab-LargeFOV模型

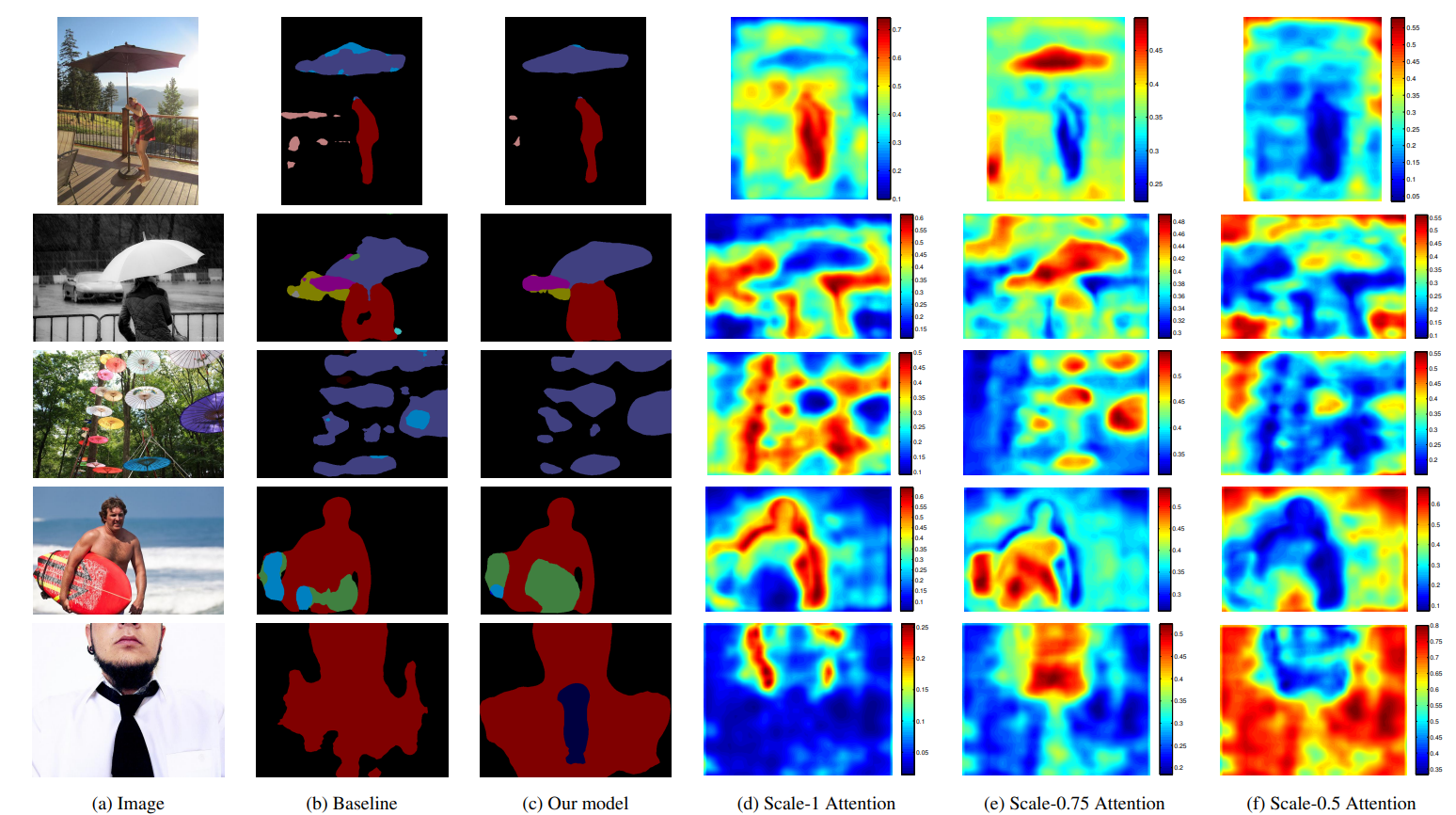


图3‑8：在MS-COCO 2014的数据集上的结果，基于DeepLab-LargeFOV模型，采用多尺度输入、注意力模型和额外的监督，尺度1的注意力集中在小比例的人（红色）和雨伞（紫色），尺度0.75的注意力集中在中等比例的雨伞和头部上，而比例0.5的注意力则捕捉到大比例的物体。

# 实验结论与展望

这篇论文将最先进的模型（即DeepLab-LargeFOV）改进为能够利用多尺度输入的方法，以进行语义分割。通过在三个数据集上进行实验，研究结果表明：（1）使用多尺度输入比使用单一尺度输入能够获得更好的性能。 （2）将多尺度特征与提出的注意力模型相结合，不仅可以改善平均池化或最大池化的基准模型的性能，还能够对不同位置和尺度上的特征进行诊断性可视化。 （3）通过对每个尺度的网络的最终输出添加额外的监督，可以获得优秀的性能。