Semantic Soft Segmentation Supplementary Material

YAĞIZ AKSOY, MIT CSAIL and ETH Zürich

TAE-HYUN OH, MIT CSAIL

SYLVAIN PARIS, Adobe Research

MARC POLLEFEYS, ETH Zürich and Microsoft

WOJCIECH MATUSIK, MIT CSAIL

语义软分割补充材料

原文链接：http://people.inf.ethz.ch/aksoyy/papers/TOG18-sss-supp.pdf

译文出处：www.github.com/hyifan/TranslatePapers

为了补充主文档，我们在第 1 节中提供了我们的特征向量估计的细节，以及图 3-5 中的附加结果和比较。

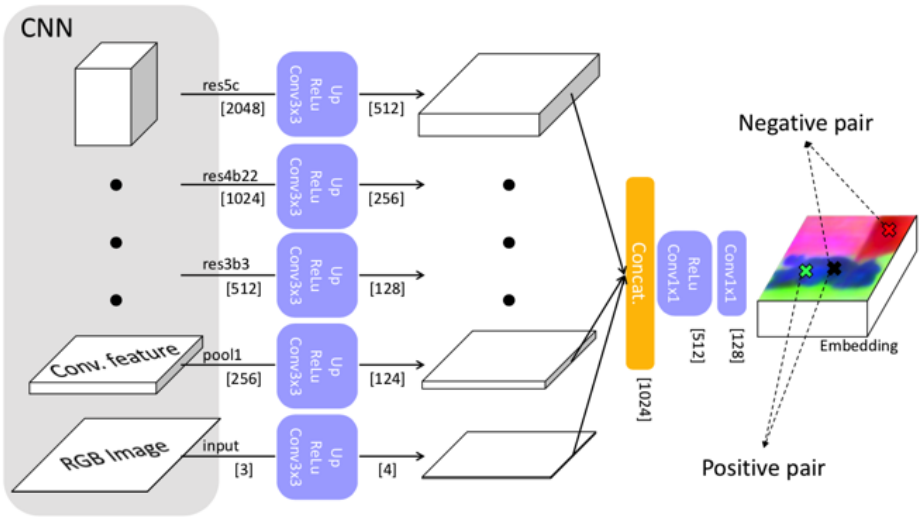

图 1

我们的网络架构。我们提取了基本卷积神经网络的中间表示特征（使用 ResNet-101 的 DeepLap 变量）。该特征经 3×3 卷积压缩，再经 ReLu 和双线性上采样，与输入分辨率相同。连接的特征被送入随后的 1×1 卷积。除此之外，我们还以端到端的方式应用了基于采样的度量学习。我们将特征尺寸表示为[#]。

1 生成语义特征描述符

我们首先为每个输入图像计算一组每个像素的语义特征。原则上，生成这些特征的网络可以很容易地

被替换，以在语义分割取得进展的同时改进结果，或者改变语义对象的定义，例如提供细粒度或基于实例的语义分割场景。

我们训练了一个级联度量学习的深度卷积神经网络，以生成属于彼此相距很远的同一个对象类的相似特征。网络输出 d=128 维的每像素语义特征。为了简单起见，我们对每像素 p 表示一个语义特征向量 $f_p \in R^d$。

我们的特征提取器的基础网络基于 DeepLab-ResNet-101。DeepLab 模型基于使用 atrous 卷积和 atrous 的空间金字塔池的 ResNet-101 的完全卷积变体。在 DeepLab-ResNet-101 中，res4b22 层是最常用的作为通用特征的输出，在原始图像恢复的十六分之一是 2048 维。由于我们的目的是提取每个像素的具有合理的对象轮廓和边界的特征，直接利用多尺度的上下文信息比在较高层使用压缩特征（如 res5b_relu)是更有利的。我们对架构进行了修改 ,以考虑更低和更高级别的特性。我们使用了由 Bertasius 等人驱动的特征链接，但我们保持一个简单的表示，以避免大内存瓶颈。我们对 input、pool1、res3b3、res4b22 和 res5c 层进行特征提取，然后用 ReLu 进行 3×3 卷积，将中间特征尺寸分别从 3、256、512、1024、2048 压缩到 4、124、128、256、512，共 1024 维。接着我们使用双线形上采用对输入图像的分辨率进行采样，然后是两个 1×1 的卷积层，这些卷积层逐渐将 1024 个特征尺寸减小到 512，最后达到 d=128。这个过程的最终输出定义了我们每像素的语义特征f_p。我们的架构如图 1 所示。值得指出的是，我们的架构是完全卷积的，允许它应用于任何分辨率的输入。当 Bertasius 等人和 Hariharan 等人在没有重新训练的情况下利用预先培训过的网络，我们对整个网络进行了微调以达到我们的目的。

为了训练整个网络，我们使用像素特征之间的 L2 距离作为度量标准来度量语义相似度。现在我们将在像素级描述我们的损失函数。给定一个像素 p 的查询向量（query vector）f_p，我们使用正向量将查询拉到正 1 和负向量，一次从查询中得到正和负示例[Hoffer 和 Ailon 2015]。由于我们致力于输入图像分辨率，为了方便地利用更多的数据和更高效的计算，我们使用了 N 对损失（N-pair loss），并进行了轻微的修改。N 对损失通过硬负数据挖掘（hard negative data-mining）方式公式化和交叉熵方式来提高数据效率，通过三重损失（triplet loss）的损失平衡（loss-balancing）来缓解收敛速度慢的问题。当 N 对损失是基于内积度量定义的，我们用 L2 距离来代替它。因此，我们的损失定义如下：

$$L_m = \frac{1}{|P|}\sum_{p,q\in P} I[l_p = l_q] log\left(\left(1+exp\left(\|f_p - f_q\|\right)\right)/2\right) + I[l_p \neq l_q] log\left(\left(1+exp\left(-\|f_p - f_q\|\right)\right)/2\right) \quad (1)$$

其中 P 表示采样像素集；$\|\cdot\|$ 为 L2 标准（我们将其除以 d 进行规范化）；I[·] 为指示函数，当语句为真则返回 1，否则返回 0；l_p为像素 p 的语义标签。

在（1）中，对于正对，如$l_p = l_q$，相应的术语$log\left(\left(1+exp\left(\|f_p - f_q\|\right)\right)/2\right)$值接近 0。共轭关系适用于（1）中第二项中的负对。由于我们只使用这个提示，无论两个像素是否属于同一类别，在训练过程中都不使用特定的对象类别信息。因此，我们的方法是一种类不可知论方法。这一事实并不妨碍我们语义软

分割的总体目标，因为我们的目标是创建覆盖语义对象的软段，而不是对图像中的对象进行分类。这也使我们能够考虑语义的多样性，而不局限于用户选择的类。

我们构造了一组采样像素 p ,如下所示。在培训过程中 ,我们将单个图像作为一个小批量传送到网络，并获得所有像素的特性。给定一个输入图像及其对应的语义的真值标签，我们首先随机抽取P_{inst}个实例，然后针对每个实例随机抽取每个实例标签遮罩内的P_{pix}个像素，使每组像素的数量均衡。我们对所选样本最小化（1），对每个图像重复采样 10 次，从中积累梯度，并立即更新。我们设置$P_{inst} = 3$和$P_{pix} = 1000$。我们可以通过使用$\mathrm{D} = \mathrm{F}1_d 1_d{}^T + \left(\mathrm{V}1_d 1_d{}^T\right)^T - 2VV^T$很容易地以矩阵形式计算（1），其中 D 是特征向量之间包含 L2 距离的矩阵，1_d是数值为 1 的行向量，F 包含样本像素的特征向量：

$$\mathrm{F} = \left[f_{1,1} \cdots f_{1,P_{pix}}, f_{2,1} \cdots f_{2,P_{pix}}, f_{P_{inst},P_{pix}}\right]^T \quad (2)$$

我们使用 COCO-Stuff 训练集对我们的网络进行训练，其中有 182 个对象和具有实例级注释的 stuff 个类别。在 MS-COCO（80 个类别）的语义分割任务中，我们用预先训练的权重初始化了基础 DeepLab 部分，其余部分用 Xavier 初始化。我们将基础部分的学习率设置为 5×10^{-4}，其余部分设置为 5×10^{-3}，以补偿随机初始化。我们使用动量为 0.9 的随机梯度下降，Chen 等人建议多个学习率衰减为 0.9，权重衰减为 5×10^{-4}。最后两个阶段的 1×1 卷积也使用概率为 0.5 来结束。我们训练了 60k 次迭代 ,在 Nvidia Titan x Pascal GPU 上大约需要不到一天的时间。

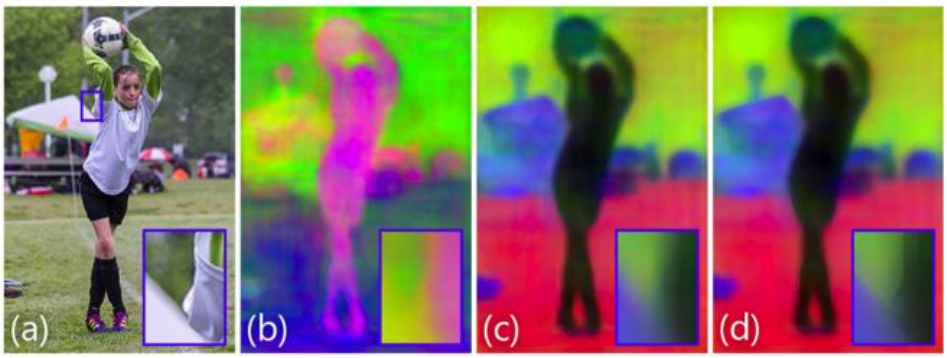

图 2

我们首先为给定的图像(a)生成每像素 128 维的特征向量。128 维到 3 维的随机投影如(b)所示。我们使用每幅图像的主成分分析(c)，将特征的维数减少到 3。为了使特征向量与图像边缘对齐，我们首先用引导滤波器对 128 个维度进行滤波，然后应用降维(d)。

1.1 预处理

128 维特征向量f_p具有足够的容量来表示现实世界语义类别的巨大多样性。然而，对于一个给定的图像，由于场景中存在的对象类别数量天然地有限，因此特征向量的有效维数要小得多。根据这一事实，为了使图的构造（如本文所述）更易于处理，更不容易进行参数调整，我们使用每幅图像的主成分分析将特征向量的维数减少到三个。

语义硬分割（semantic hard segmentation）的一个主要缺点是它的边界的不确定性。这个事实也很好地反映在生成的特征向量中，如图 2 所示。为了在图形中插入语义信息时，计算更有效的亲和度，我们在输入图像的指导下，利用引导滤波对特征向量进行正则化。这使得特性与图像中的硬边界更加一致，如图 2 所示。在降维之前，我们对所有 128 维都进行这种过滤。最后，我们将低维特征归一化到[0,1]范围内，得到三维特征向量f_q用于亲和性计算。

图 3

我们将我们的结果与 Zhao 等人（PSPNet）、He 等人（Mask R-CNN）和光谱抠图（spectral matting）一起展示。分段被覆盖到输入图像的灰度版本上，以便更好地评估分段边界。注意 PSPNet 和 Mask R-CNN 在物体边界周围的不精确性，以及延伸到物体外的光谱抠图的软段。

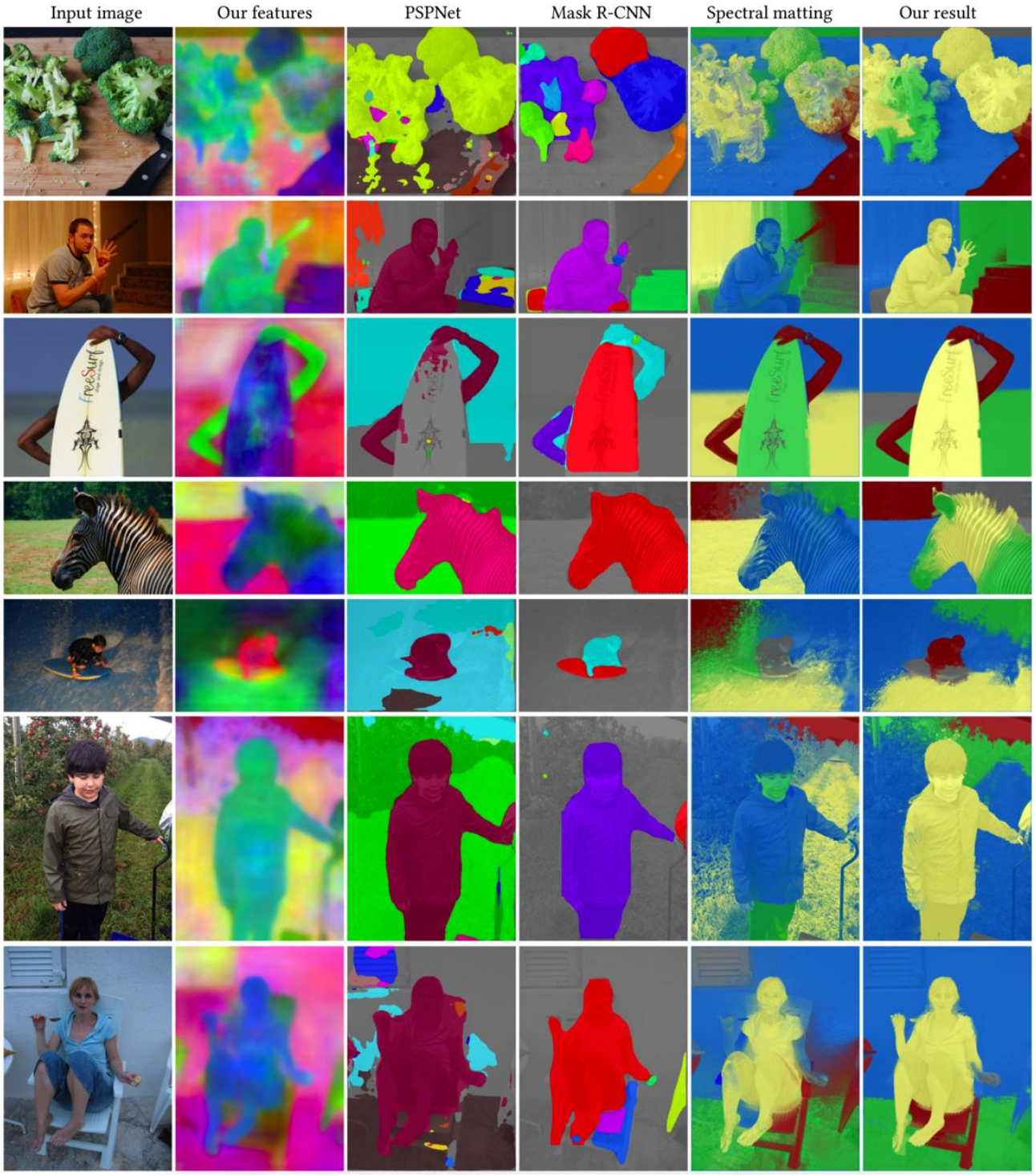

图 4

我们将我们的结果与 Zhao 等人（PSPNet）、He 等人（Mask R-CNN）和光谱抠图（spectral matting）一起展示。分段被覆盖到输入图像的灰度版本上，以便更好地评估分段边界。注意 PSPNet 和 Mask R-CNN 在物体边界周围的不精确性，以及延伸到物体外的光谱抠图的软段。

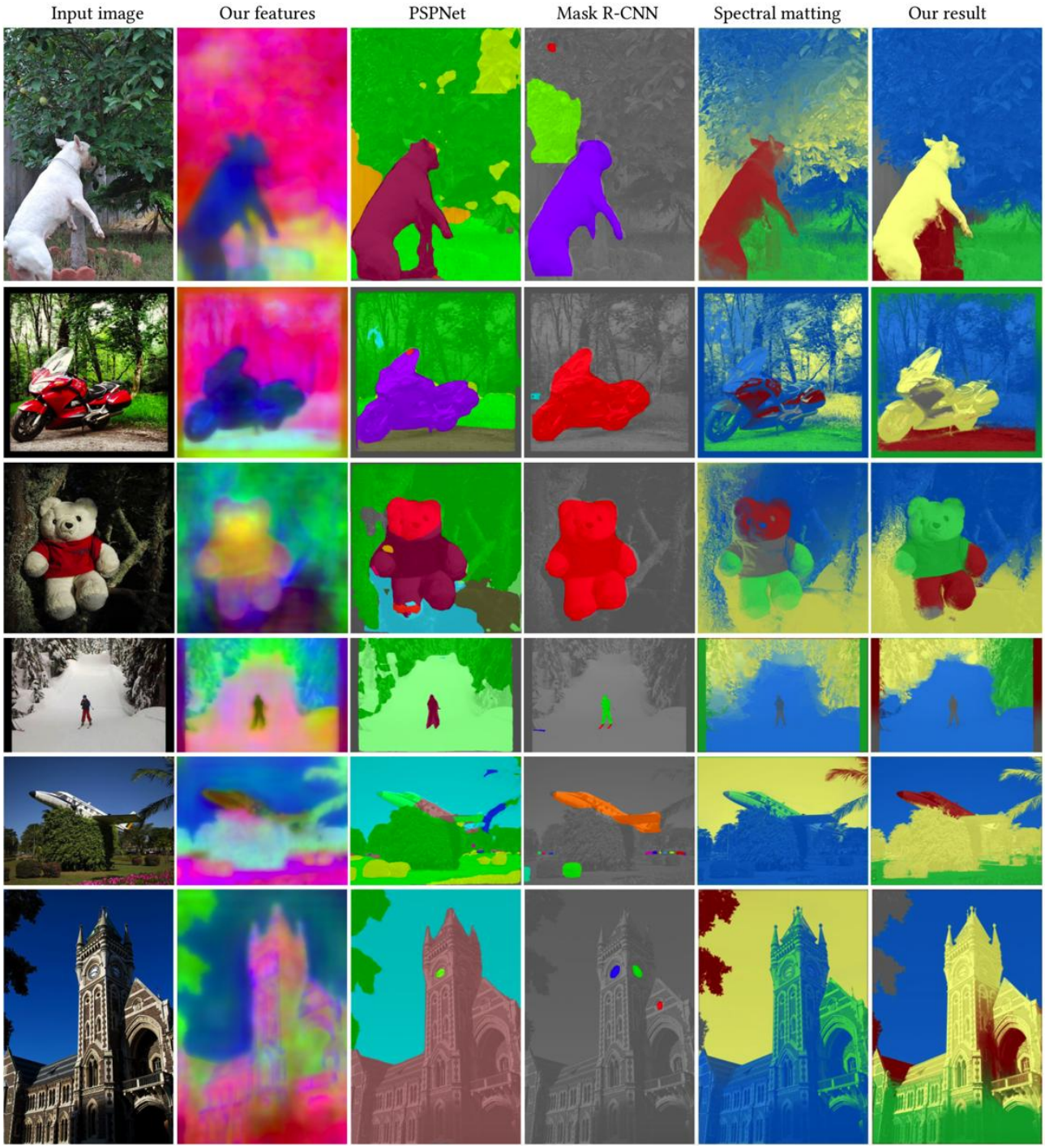

图 5

我们将我们的结果与 Zhao 等人（PSPNet）、He 等人（Mask R-CNN）和光谱抠图（spectral matting）一起展示。分段被覆盖到输入图像的灰度版本上，以便更好地评估分段边界。注意 PSPNet 和 Mask R-CNN 在物体边界周围的不精确性，以及延伸到物体外的光谱抠图的软段。