**基于在线难样本学习的图像语义分割网络**

王振扬1,2 邓志东1,2 王诗瑶1,2

1（清华大学计算机科学与技术学院 北京 100084）

2（智能控制与系统国家重点实验室（清华大学）北京 100084）

1（[crazycry2010@gmail.com](mailto:crazycry2010@gmail.com), [michael@tsinghua.edu.cn](mailto:michael@tsinghua.edu.cn), [sy-wang14@mails.tsinghua.edu.cn](mailto:wangshiyao14@mails.tsinghua.edu.cn)）

#### Semantic Segmentation with Online Hard Example Mining

Wang Zhenyang1, Deng Zhidong1,2 , and Wang Shiyao1,2

1（*College of Computer Science and Technology, Tsinghua University, Beijing 100084*）

2 （*State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science (Tsinghua University) Beijing, 100084*）

**Abstract** Scene labeling, based on semantic segmentation, is a fundamental topic in computer vision. The goal is to assign each pixel in the image a category label. Convolutional neural networks, especially the fully convolutional neural networks, have attracted increasing attention on semantic segmentation due to the powerful capabilities of extracting hierarchical features. Since it is required to learn to make dense predictions for each pixel, a simple network is hardly to obtain considerable performances on different scenes. In this paper, we propose a novel semantic segmentation network called HMNet, which aims to pay more attention to the hard examples. The network has three branches, where the first branch produces coarse output predictions, and the second branch selects the hard examples which will be fed to the last branch. All above branches focus on their own objectives and collaboratively learn to predict from coarse to fine inference. Since the semantic segmentation dataset contains a large number of relatively easy samples and some hard ones, HMNet is encouraged to select these hard examples to make further predictions which is help to improve the final performance. In order to evaluate predicting performance of the proposed HMNet, we conduct experiments on two public datasets including Sift Flow and Stanford Background Dataset. We show that the three branches can be trained in an end-to-end manner and the experimental results show that compared to all existing models, our HMNet consistently yields the best performance, with accuracy of 91.6% and 89.7%, respectively.

**Key words** semantic segmentation; online hard example mining; convolutional neural network

摘要 图像语义分割需要正确识别并标注出图像中每一个像素所属的类别，是计算机视觉的重要任务之一。卷积神经网络，特别是全卷机神经网络的提出，推动了语义分割的发展。语义分割网络需要预测大量密集的语义标签，通过单一网络结构很难准确地预测出不同场景的像素类别。本文基于错误样本挖掘的思想，提出了一种新的语义分割网络结构HMNet，由三支不同的分割网络模型构成。HMNet在对测试图像进行语义分割的同时，预测出分割出错的难像素样本，最后针对这些可能的难样本，用另一个能力更强的分割网络进行二次分割。语义分割数据集中包含了大量的简单像素样本和少量的难样本，在训练过程中将难样本筛选出来进行二次训练，可以有效地提高网络的训练效率，最终提高语义分割结果的正确率。在SIFT Flow和Stanford Background Dataset两个语义数据集上，对HMNet进行了实验验证，分别取得了91.6%和89.7%的正确率。

关键词语义分割；难样本挖掘；卷积神经网络

中图法分类号 TP391

语义分割，又可称为场景标注，是计算机视觉领域重要的研究任务之一。语义分割的目标是识别并标注出图像中每一个像素所属的类别。这需要完全理解整张图像的语义信息，不仅要预测图像中物体的标签，还需要在像素级别标注出物体位置。语义分割在机器人环境感知和无人驾驶领域有很强的应用需求。

近年来，深度学习的出现与发展，图像分类[1-3]、语音识别[4-5]、视觉物体识别和物体检测[6-8]等任务上均取得了成功的应用，不断刷新了多个公开测试数据集的记录。在图像语义分割任务，同样出现了很多基于卷积神经网络的方法。早期的研究尝试直接将用于视觉物体识别的卷积神经网络拓展到语义分割任务，尽管也取得了较好的分割结果，但是预测结果比较粗糙，无法正确分割出物体的边缘。这主要是因为分类网络的最后几层通常是全连接层，而全连接将二维特征图压缩为一维的特征向量，丢失了空间信息。 全卷积神经网络的出现，有效地解决了传统分类网络中空间信息丢失的问题。全卷积神经网络丢弃了全连接层，采用卷积层代替全连接层进行特征提取，是目前语义分割任务最为通用的特征提取框架。

为了解决语义分割中边缘与细节模糊的问题，条件随机场、端到端的条件随机场、马尔科夫随机场、高斯条件随机场等后端处理方法被相继提出。但是卷积网络结构的研究突破，尤其是批正则化[9]（Batch Normalization）和残差网络[10]（ResNet）的提出，优化了图像特征提取的过程，可以有效地抑制边缘与细节模糊的问题。

对于语义分割任务，另一个具有挑战的问题是同时预测大量且密集的像素语义标签。而在每张图像中，不同类别的像素分布往往是不均衡的，背景像素与前景像素就是最为极端一个例子。此外，对于每张训练图像，通常包含了大量的简单像素样本（例如绝大部分背景像素、物体的中心部分），和少量的难像素样本（例如物体的边缘与细节）。针对图像语义分割中各类别像素样本分布不均和难易样本分布不均的问题，对不同类别和难易的像素样本进行重采样，是一个很容易想到的解决方式。本文基于难样本挖掘（Hard Negative Mining）的思想，提出了一种新的语义分割网络结构HMNet。

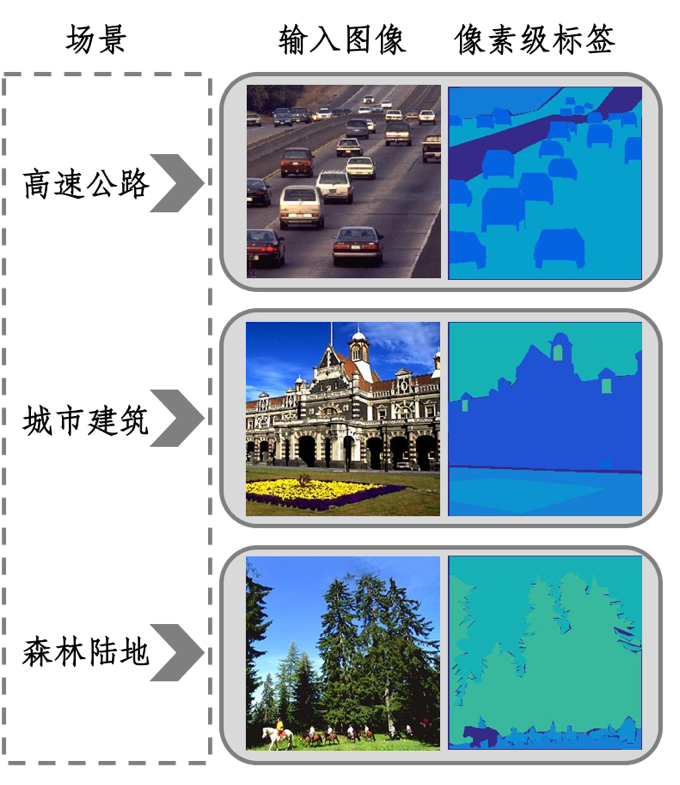
事实上，难样本挖掘并不是一个新概念，早在1994年，Sung和Poggio就在人脸检测算法中提出了自展（Bootstrapping）的概念。两者的主要思想都是要搜集被分类器误判的样本，在进一步迭代学习或训练的过程中，通过改变难易样本的分布来着重处理这些难样本。借鉴与之相似的思想，HMNet由三支网络构成，第一支网络是一个简单语义分割网络，可以预

Fig. 1 Examples of semantic segmentation.

图1 语义分割/场景标注举例

测出绝大多数简单像素样本的语义类别；与此同时，HMNet引入第二支网络对第一支网络预测结果的正确性进行预测，对于第一支网络分类正确的那些像素样本，优先相信第一支网络的预测结果；对于第一支网络分类错误的像素样本，用HMNet的第三支网络进行二次强化学习。

对于图像语义分割任务，每张训练图像均包含了大量且密集的识别目标，采用单一的分割网络难以同时满足所有的优化目标。因此HMNet采用多支的网络结构，多个分割网络融合的方式来优化最终的分割结果。对于第一支网络判别错误的像素样本，即在线学习到的难像素样本，HMNet采用另外一支结构更为复杂的分割网络来有针对地学习难像素样本的分类，最终采用难易两支分割网络融合后的概率图作为整张图像的语义预测结果。

本文提出了一种基于在线难样本学习的语义分割网络模型HMNet，与传统的语义分割网络相比，HMNet具有如下特点：

1) 采用在线难样本学习的方法，提高网络训练效率和分割的正确率。

2) 提出多分支的网络结构，分别用于学习难易样本的语义分割和图像中难易样本的概率分布。

3) 采用多支语义分割网络融合的方式，来提高密集的语义像素分割结果的正确率。

4) 在Sift Flow[11]和Stanford Background Dataset[12]两个公开数据集上验证了HMNet结构的有效性。

## 1 相关工作

场景标注是对输入图像的每一个像素预测其所属的语义类别，例如道路，水，海洋等。而图像的全局以及局部区域特征则对标注的准确性有着较大的影响。在近几年的研究工作中，该任务主要可分为基于传统手工设计特征以及基于深度神经网络特征的研究方法。其中，基于深度神经网络特征的方法中，以全卷积神经网络为代表，取得了尤为重要的性能发展。

传统方法 在近几年的研究中，场景标注问题获得了较多的解决方案。许多方法依赖于MRF，CRF或其他类型的图形模型来保证标签的一致性，并且考虑上下文信息[12]，[13]，[14]，[15]，[16]，[17]，[18]。 大多数方法依赖于预分割成超像素或其他分段候选，并且从个体分段和从相邻分段的各种组合提取特征和类别。图形模型推理出了覆盖图像的最一致的段集合。

深度神经网络 在近几年的研究中，卷积神经网络以其强大的特征提取能力以及对局部视觉特征的鲁棒辨识能力，被大量应用在图像的分割任务中。较多研究工作采用卷积神经网络对输入图像每一个局部区域的像素进行直接概率建模。[19] 最先采用卷积神经网络进行场景标注。作者提出了一种多尺度卷积神经网络，对于不同尺度的图像块进行学习并提取对应的特征向量。文中结果表明，网络通过学习能够比传统手工设计的特征有更好的分割效果，将纹理，形状以及领域信息隐式包含于所学的特征表达中。除此之外，该网络也具有对室内场景RGB-D图像进行分割的能力[20]。为了能够保证视觉特征一致性以及保持较高的预测准确率，[21]提出了对于一定范围内像素的类别标注相关性进行建模。文中利用循环卷积神经网络对于复杂的空间关系进行建模，每一次网络输出的特征映射图都会作用在前一次的输出结果中，该方并在较少的时间代价下取得了较高的预测精度。[22-24]训通过对不同图像块的采样，训练了多个参数卷积神经网络，大量地减少了训练时间，加快了训练的过程。他们发现，基于图像块的卷积神经网络会受到局部区域的影响。对于该问题的解决方法，[22]利用了全局置信度。由于有限的背景以及相关信息，一些相似的相似的像素并不能很好地被区分开，因此作者引入了全局场景约束，利用全局的场景信息对于当前预测值进行约束和推断。[23-24]采用了循环神经网络，对于特征在相邻空间维度上的所提取的CNN特征进行了一致性建模，即利用CNN来进行局部区域的特征提取，并利用二维循环神经网络对于像素间的长程相关性进行学习。该模型能够将图像全局的上下文信息自动嵌入到局部特征表示中以增加特征的表达能力，从而有效地提高了标注的准确率。

与此同时，研究者们尝试用预先训练的深度卷积神经网络对于物体进行语义分割。[25]采用一个网络对于局部以及近处的物体进行特征提取，并利用AlexNet[26] 对于远距离以及全局特征进行提取。两者特征的聚合向量则作为最终的特征表达。他们利用了图像和标签空间中的统计结构，而没有设置明确的结构化预测机制，从而避免了复杂和昂贵的推断，并在语义分割的任务上取得了非常好的效果。与以上方法有着明显区别的为[27]中提出的全卷积神经网络，该模型对于原始图像以及密度的类别标签进行了直接训练以及预测。与现有方法不同之处在于，文中在现有的深度分类网络上进行了改进，使用图像分类作为监督预训练，但采用全卷积网络来学习从整个图像输入到像素级别标签的映射。该全卷积神经网络的卷基层参数是由在ImageNet上训练的模型参数作为初始化，而反卷积层的参数则是通过上采样到标签维度来进行徐诶。[28]也利用了预训练的深度卷积神经网络来对像素级别的标签进行学习。考虑到边界对齐的不完整性，原本的网络并不能对于对象边缘进行完整地分割，他们进一步使用全连接的条件随机场（CRF）来克服网络原本的显示，从而提高预测性能。为了更准确地感知场景，增加场景的先验信息则显得尤为重要。[29]认为当前基于FCN的模型的主要问题是缺乏使用全局场景类别来选择合适的策略。[29] 则介绍了金字塔池模块和金字塔场景解析网络（PSPNet），基于不同区域进行上下文聚合从而对全局上下文信息进行特征表示。 文中的全局先验表达在场景解析任务上能够有效地提升模型性能，而与此同时，PSPNet为像素级预测任务提供了一个优越的框架。

**2**基于在线难样本学习的语义分割网络

这节提出了基于在线样本学习的语义分割网络HMNet。其中，2.1详细分析了机器学习领域难样本挖掘的通用实现方式，以及HMNet如何结合语义分割任务的特点，将难样本挖掘的思想引入到端到端的语义分割网络的学习过程；2.2详细描述了HMNet的网络结构；2.3简要阐述了HMNet的实现细节。

**2. 1**在线难样本学习

难样本挖掘是机器学习在训练过程中常用的训练技巧之一。对于SVM分类器，难样本挖掘的步骤可以定义为：

1）在一段时间内，固定训练模型，用这个固定的模型筛选出难样本，并以一定的难易比率加进当前活跃的训练集；

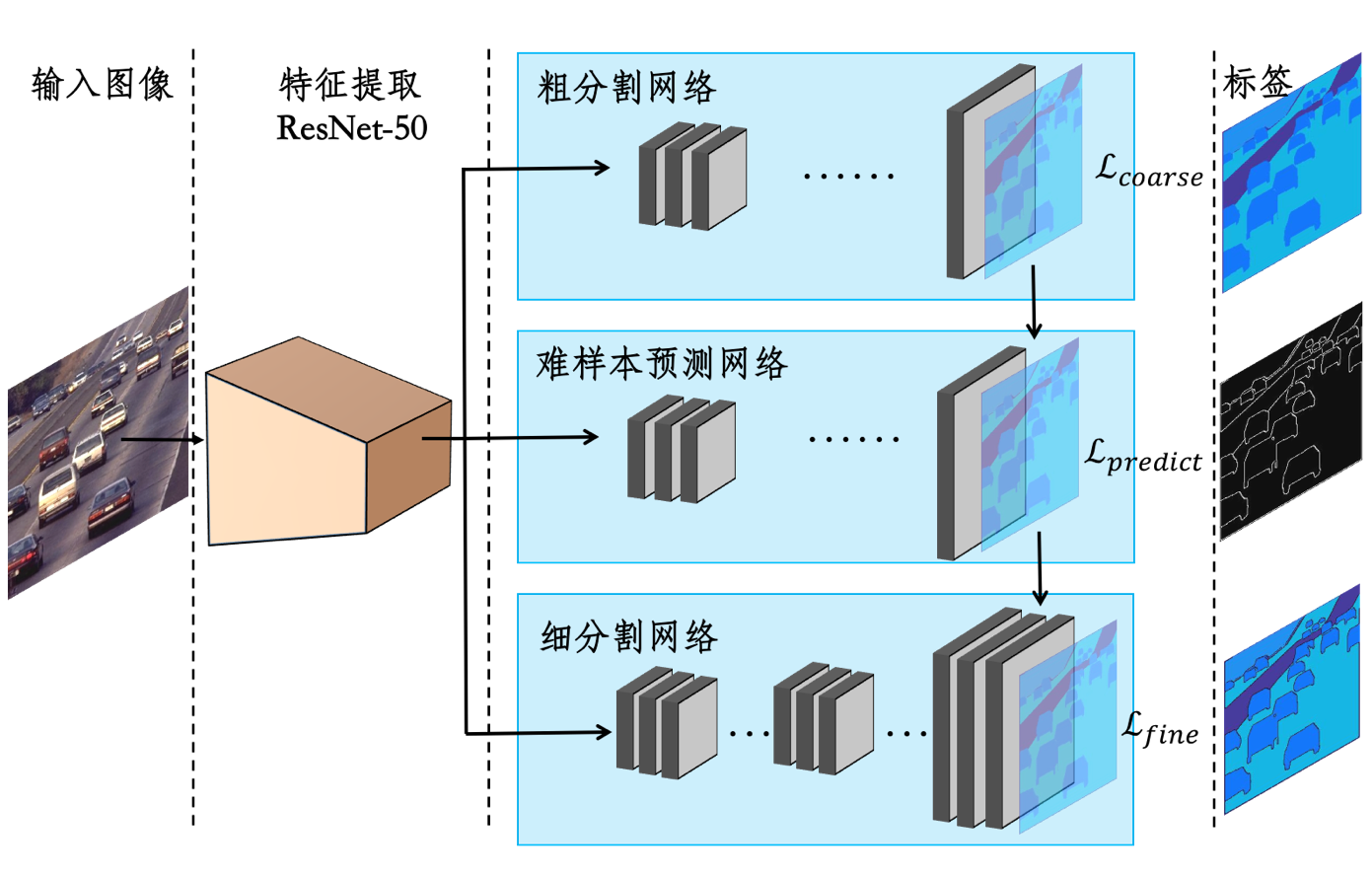


Fig. 2 Semantic segmentation with online hard example, HMNet.

图2 基于在线难样本学习的语义分割网络结构HMNet

2）在接下来的另外一段时间内，固定当前活跃训练集，在这个固定的训练集上，重新训练SVM模型。

实际训练过程中步骤1）从全体训练样本中检索判别错误的难样本，加入当前活跃训练集，直到当前训练集达到样本上限。步骤2）使用当前训练集训练SVM分类器，直到SVM重新收敛并趋于稳定。这两个步骤反复迭代，直到当前训练集包含了SVM的所有支持向量。步骤1）的引入会增加整体训练时间。

步骤1）的难样本挖掘与步骤2）的训练过程是串行进行的，每一轮迭代过程中都会消耗额外的训练时间。本节提出的在线难样本学习，将上述两个串行步骤合并在一个并行的分割网络结果中同步进行。每张训练图像均包含大量的训练像素样本，从这些像素样本中，我们可以采用启发式的策略，筛选出一部分难像素样本，使图像语义分割和难样本预测同步进行。这种难样本预测网络的提出，大大简化了步骤1）中先固定模型再筛选样本的限制。

难样本预测网络的处理过程如下，如图2所示。对于任意一张训练图像，首先通过一个特征提取网络来提取图像特征，本文采用具有50层深度的残差网络Resnet-50作为特征提取网络。其次，采用一个简单的语义分割网络（粗语义分割网络）对图像进行初步的语义分割，本文采用的是一个具有2层卷积结构的全卷机网络。语义分割之后，将会生成一张预测语义标签图，与真实的标注结果（Ground Truth）对比，可以得到粗语义分割网络分割正确与错误的标签。用生成的正确/错误标签去监督训练一个可以预测粗语义分割网络分割结果正确性的2分类判别网络（难样本预测网络），该网络可以预测出输入图像中较难分类的难像素样本。粗语义分割网络和难样本预测网络共用相同的输入图像特征，粗语义分割网络对分割结果进行预测，难样本预测网络对粗语义分割网络预测结果的正确性进行预测。两支网络并行的结构，可以在一个前传与反传过程中同时进行参数的更新和难样本的在线学习。

**2. 2** HMNet网络结构

HMNet网络结构如图2所示，HMNet由三支网络构成，分别是粗语义分割网络、难样本预测网络和细语义分割网络，三支网络共用了同一个特征提取网络。下面本节将分别从特征提取网络和三个分支网络来介绍HMNet的网络结构。

特征提取网络，本文采用深度为50的残差网络Resnet-50[10]对输入图像进行特征提取。对于图像语义分割任务，多尺度的图像特征有助于结合图像的上下文

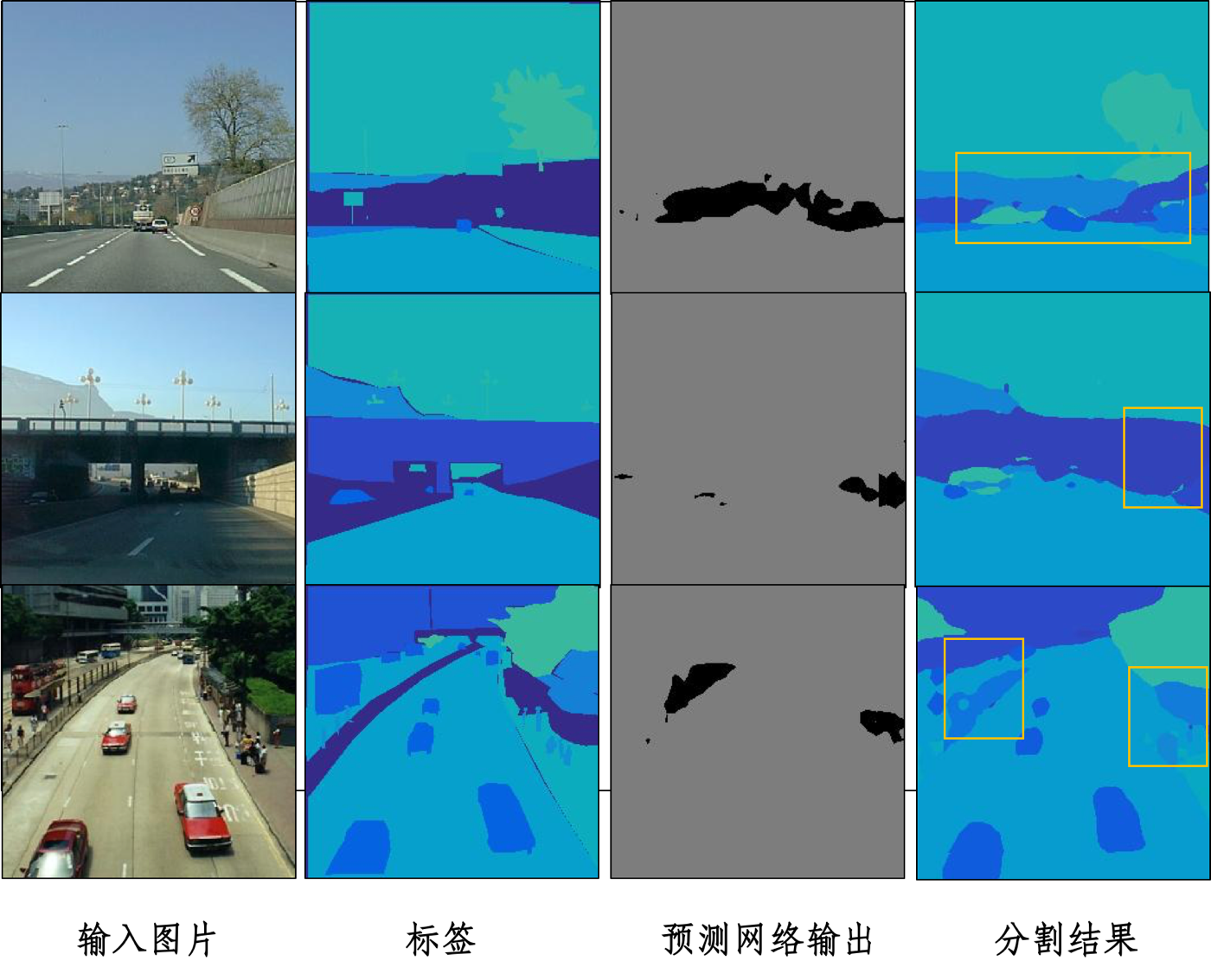


Fig. 3 The visualization predicting result of HMNet.

图3 SIFT Flow数据集上，HMNet可视化结果

特征，对像素进行分类。残差网络采用递归的网络设计方式，其展开式中包含不同范围感受野的图像特征。因此采用深度残差网络进行图像特征提取，可以有效地融合图像上下文特征。此外，为了减小特征空间信息的损失，借鉴于膨胀全卷机网络[27]（Dilated FCN）结构设计过程，本文采用具有2048维特征通道和3x3感受野大小的卷积层来替代Resnet-50的全连接层，并且保持特征图的分辨率是原始输入图像的1/8大小。

粗语义分割网络，主要用于对训练图像进行初步的语义分割。本文采用具有2个卷积结构的全卷机网络来实现初步的语义预测，此时预测网络的输出特征图大小依然是原始输入图像的1/8。随后本文采用双线性差值对输出特征进行上采样，恢复成与原始输入图像分别率相同的概率预测图。最后采用softmax函数构造粗语义分割网络的损失函数。难样本预测网络，主要用于检索训练图像中的难像素样本，该网络与粗语义分割网络共享相同的输入特征，且具有相同的网络结构。与粗语义分割网络不同的是，粗语义分割网络是一个二分类网络，仅用于预测每个像素位置分割结果的正确性。在训练过程中，粗语义分割网络可以给出一个语义预测结果，结合训练图像的真实语义标签，即可以得到预测结果正确与否的二分类标签。该标签即可作为难样本预测网络的监督标签，使难样本预测网络可以预测出训练图像每个像素分类的正确性与难易程度。

细样本分割网络，针对难样本预测网络筛选出的难像素样本，采用更为复杂的网络结构，对其进行二次分类。对于语义分割任务，通过粗语义分割网络很难正确处理所有的语义标签，因此HMNet引入第三支细样本分割网络对难分类的复杂像素样本进行单独处理。本文采用具有金字塔池化的语义分割网络PSPNet[29]作为细样本分割网络，该结构通过对特征图进行金字塔池化，来提取多尺度的图像特征，增加特征维度，提高网络的分割能力。HMNet网络语义分割的结果是粗样本分割网络和细样本分割网络结果的融合。两支网络以难样本预测网络的预测概率为权重，加权平均的结果即为预测的语义分割结果。

**2. 3** 实现方式

难样本挖掘的实现方式有很不同的方法，最简单直接的方法是通过修改网络的损失函数来实现，本文也采用这种方式实现。在网络训练过程中，三支网络分别对应3个不同的损失函数，整个网络的损失函数是三个损失函数的加权和。粗语义分割网络是一个传统的语义分割网络，其损失函数是所有像素点的softmax损失函数和。难样本预测网络可以看做是一个二分类语义分割网络，唯一的难点在于监督标签的获取，监督标签需要用到第一支粗语义分割网络的预测结果。第三支网络用于难像素样本的语义分割，而整张图像的预测结果是以难样本预测网络作为加权权重，粗细语义分割网络预测结果的加权平均值。

为了更直观地得到最终的语义分割结果，实现端到端的训练，本文对第三支网络的函数进行了优化。直接以粗细样本语义分割网络的加权结果和训练图像的真实语义标签作为softmax的输入，来构造一个端到端的语义分割模型。

**3**实验与结果

**3. 1** 训练参数配置

基于深度学习平台caffe[30]，本文实现了HMNet网络结构，采用批量随机梯度下降法端到端地训练整个网络。与大多数语义分割方法类似，HMNet在Hengshuang等人PSPNet[29]工作的基础上，对网络参数进行精调（fine-tune）。网络的初始学习率为 1e-4，以此学习率对网络训练了3k轮，之后将学习率降低为原来的1/10再训练1k轮，最后用1e-6的学习率对网络进行最后1k轮的精调。网络的动态因子（Momentum）和权重衰减（Weight Decay）分别设置为0.9和0.0001。

语义分割模型经常采用数据增广的方法来抑制过拟合问题，通过对训练图像进行各种形式的变换来扩大样本容量，提高语义分割网络模型的泛化能力。本文也对输入图像进行了数据增广，先对每张输入图像进行1-2倍的随机放大，再从放大图像中随机截取233x233像素大小的子图用于网络的训练。

采用更大的输入图像和更多的批训练样本，有助于网络性能的提升。本文限于GPU计算能力和显存大小的限制，输入图像大小限制为233x233像素，训练过程中每批仅处理4张图像。

**3. 2** 度量标准

像素级语义分割通常采用两种准确率衡量方法：像素平均准确率（Pixel Accuracy）和类别平均准确率(Class Accuracy)。像素平均准确率是测试数据集中正确分类的像素个数占像素总数的百分比，通常通过交并比（intersection-over-union，IoU）来度量。类别平均准确率是各类别像素分类正确率的平均值。

**3. 3** 数据集

本文在两个语义分割数据集对HMNet网络进行了实验测试与验证，这两个数据集分别是SIFT Flow[11]和Stranford Background[12]。SIFT Flow数据集包含2688张256x256像素大小的彩色图像，其中包含2488张训练样本和200张测试样本。该数据集一共定义了33种相互独立的语义类别，而且各像素类别的分布十分不均衡。

**Table 1 The Results on SIFT Flow with semantics.**

**表1 SIFT Flow数据集的分割结果**

|  |  |  |
| --- | --- | --- |
| *Methods* | *Pixel acc.* | *Class acc.* |
| Liu et al.[11] | 76.7 | - |
| Tighe et al. SVM[31] | 75.6 | 41.4 |
| Tighe et al. SVM+MRF[32] | 78.6 | 39.2 |
| Farabet et al. natural[32] | 72.3 | 50.8 |
| Farabet et al. balanced[19] | 78.5 | 29.6 |
| Pinheiro et al.[19] | 77.7 | 29.8 |
| Liang et al. RCNN[21] | 84.3 | 41.0 |
| Shelhamer et al. FCN-8s[27] | 85.9 | **53.9** |
| **HMNet** | **91.6** | 52.5 |

Stanford Background数据集包含了715张彩色图像，图像尺寸大小不一，但均未超过320x240像素。沿用以往的研究测试方法，本文采用5倍交叉验证的方法对数据集进行划分，每次划分出572张图像作为训练样本，剩余的143张作为测试样本。Stanford Background数据集包含了8种语义类别，而且像素的分布比SIFT Flow数据集更加平衡。

**Table 2 The Results on Stanford background with semantics.**

**表2 Stanford background数据集的分割结果**

|  |  |  |
| --- | --- | --- |
| *Methods* | *Pixel acc.* | *Class acc.* |
| Gould et al.[12] | 76.4 | - |
| Tighe and Lazebnik[31] | 77.5 | - |
| Socher et al.[33] | 78.1 | - |
| Eigen and Fergus[34] | 75.3 | 66.5 |
| Singh and Kosecka[35] | 74.1 | 62.2 |
| Lempitsky et al.[18] | 81.9 | 72.4 |
| Liang et al. RCNN[36] | 83.1 | 74.8 |
| **HMNet** | **89.7** | **75.4** |

**3. 2** 实验和结果

首先，在SIFT Flow数据集上我们对HMNet网络进行了实验验证。实验结果如图3所示，其中第一列是测试图像，第二列图像的语义标签，第三列是HMNet网络预测的难像素样本，第四列是网络的预测数据结果。由图3的第三列与第四列可以看出，HMNet网络确实可以较为准确地预测出难像素样本，在图3的第四类用已用黄色矩形框标出。

此外，在SIFT Flow数据集上，我们将HMNet的最终分割结果与其他分割方法进行了对比，实验结果如表1所示。HMNet取得了91.6%的像素平均正确率，取得了该数据集最好的分割结果。

为了验证在线难样本学习语义分割网络的通用性，采用相同的网络结构与配置参数，我们在另一个数据集Stanford Background上，对HMNet进行了实验测试，并与其它方法进行了对比，实验结果如表2所示。在Stanford Background数据集上，HMNet取得了89.7%的像素平均正确率和75.4%类平均正确率，同样取得了该数据集的最好分割结果，部分预测结果如图4所示，可以明显看出预测结果的边缘信息。

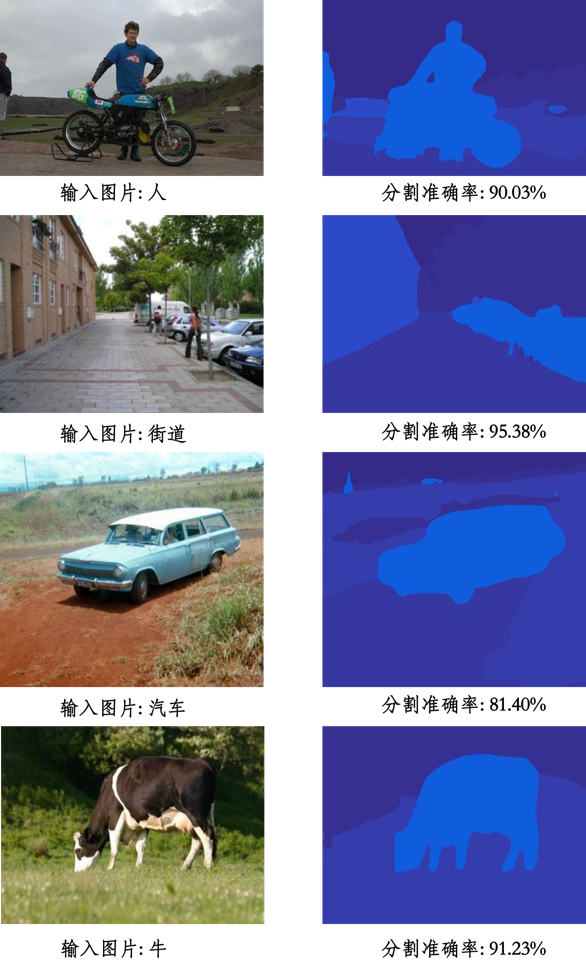


Fig. 4 The predict results on Stanford Background dataset.

图4 Stanford Background数据集预测结果

**4** 总结

本文提出了一种基于在线难样本挖掘的语义分割网络结构HMNet。在对图像进行语义预测的同时，对图像中识别困难的像素样本进行在线预测，最后针对这些难像素样本进行二次识别分类，从而提高整个网络分割的正确率。在SIFT Flow和Stanford Background两个数据集上，HMNet分别以91.6%和89.7%的测试正确率取得了两个分割数据集的最好分割结果。

**参 考 文 献**

1. Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.
2. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.
3. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 1-9.
4. Zeiler M D, Ranzato M, Monga R, et al. On rectified linear units for speech processing[C]//Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on. IEEE, 2013: 3517-3521.
5. Graves A, Mohamed A, Hinton G. Speech recognition with deep recurrent neural networks[C]//Acoustics, speech and signal processing (icassp), 2013 ieee international conference on. IEEE, 2013: 6645-6649.
6. Sermanet P, Eigen D, Zhang X, et al. Overfeat: Integrated recognition, localization and detection using convolutional networks[J]. arXiv preprint arXiv:1312.6229, 2013.
7. Girshick R, Donahue J, Darrell T, et al. Region-based convolutional networks for accurate object detection and segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2016, 38(1): 142-158.
8. He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[C]//European Conference on Computer Vision. Springer International Publishing, 2014: 346-361.
9. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[J]. arXiv preprint arXiv:1502.03167, 2015.
10. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 770-778.
11. Liu C, Yuen J, Torralba A. Sift flow: Dense correspondence across scenes and its applications[J]. IEEE transactions on pattern analysis and machine intelligence, 2011, 33(5): 978-994.
12. Gould S, Fulton R, Koller D. Decomposing a scene into geometric and semantically consistent regions[C]//Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 2009: 1-8.
13. He X, Zemel R S. Learning hybrid models for image annotation with partially labeled data[C]//Advances in Neural Information Processing Systems. 2009: 625-632.
14. Russell C, Kohli P, Torr P H S. Associative hierarchical crfs for object class image segmentation[C]//Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 2009: 739-746.
15. Kumar M P, Koller D. Efficiently selecting regions for scene understanding[C]//Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010: 3217-3224.
16. Mathieu M, Henaff M, LeCun Y. Fast training of convolutional networks through ffts[J]. arXiv preprint arXiv:1312.5851, 2013.
17. Springenberg J T, Riedmiller M. Improving deep neural networks with probabilistic maxout units[J]. arXiv preprint arXiv:1312.6116, 2013.
18. Lempitsky V, Vedaldi A, Zisserman A. Pylon model for semantic segmentation[C]//Advances in neural information processing systems. 2011: 1485-1493.
19. Farabet C, Couprie C, Najman L, et al. Learning hierarchical features for scene labeling[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(8): 1915-1929.
20. Couprie C, Farabet C, Najman L, et al. Indoor semantic segmentation using depth information[J]. arXiv preprint arXiv:1301.3572, 2013.
21. Pinheiro P H O, Collobert R. Recurrent Convolutional Neural Networks for Scene Labeling[C]//ICML. 2014: 82-90.
22. Shuai B, Wang G, Zuo Z, et al. Integrating parametric and non-parametric models for scene labeling[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 4249-4258.
23. Shuai B, Zuo Z, Wang G. Quaddirectional 2d-recurrent neural networks for image labeling[J]. IEEE Signal Processing Letters, 2015, 22(11): 1990-1994.
24. Shuai B, Zuo Z, Wang B, et al. Dag-recurrent neural networks for scene labeling[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 3620-3629.
25. Mostajabi M, Yadollahpour P, Shakhnarovich G. Feedforward semantic segmentation with zoom-out features[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3376-3385.
26. Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.
27. Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3431-3440.
28. Chen L C, Papandreou G, Kokkinos I, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs[J]. arXiv preprint arXiv:1412.7062, 2014.
29. Zhao H, Shi J, Qi X, et al. Pyramid Scene Parsing Network[J]. arXiv preprint arXiv:1612.01105, 2016.
30. Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding[C]//Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014: 675-678.
31. Tighe J, Lazebnik S. Superparsing: scalable nonparametric image parsing with superpixels[C]//European conference on computer vision. Springer Berlin Heidelberg, 2010: 352-365.
32. Tighe J, Lazebnik S. Finding things: Image parsing with regions and per-exemplar detectors[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 3001-3008
33. Socher R, Lin C C, Manning C, et al. Parsing natural scenes and natural language with recursive neural networks[C]//Proceedings of the 28th international conference on machine learning (ICML-11). 2011: 129-136.
34. Eigen D, Fergus R. Nonparametric image parsing using adaptive neighbor sets[C]//Computer vision and pattern recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012: 2799-2806.
35. Singh G, Kosecka J. Nonparametric scene parsing with adaptive feature relevance and semantic context[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013: 3151-3157.
36. Liang M, Hu X. Recurrent convolutional neural network for object recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3367-3375.



**Wang Zhenyang,** born in 1988. PhD. His main research interests include computer vision, deep learning, machine learning.



**Deng Zhidong,** born in 1957. PhD, professor, PhD supervisor. His main research interests include computer vision, deep learning, machine learning.



**Wang Shiyao,** born in 1991. PhD. Her main research interests include computer vision, deep learning, machine learning.