**A Network Structure to Explicitly Reduce Confusion Errors in Semantic Segmentation**

arXiv:1808.00313v1 [cs.CV] 1 Aug 2018

**一种明显减少语义分割中混淆错误的网络结构**

Qichuan Geng1*,*2, Xinyu Huang1, Zhong Zhou2, and Ruigang Yang1

1 Baidu Research, Beijing, China, 100193

gengqichuan,huangxinyu01,yangruigang @baidu.com

*{ }*

2 School of Computer Science and Engineering, Beihang University, Beijing, China, 100083

[zz@buaa.edu.cn](mailto:zz@buaa.edu.cn)

**Abstract.** 在现实世界中普遍存在的混淆类通常会降低许多与视觉相关的应用程序（如对象检测，分类和分段）的性能。混淆错误不仅由类似的视觉模式引起，而且在我们设计的模型训练期间也被各种因素放大，例如编码过程中的特征分辨率降低或数据分布不均衡。 近年来，已经提出了大量基于深度学习的网络结构来处理这些个体因素并改善网络性能。然而，据我们所知，语义图像分割中的任何现有工作都没有被设计用于明确地解决混淆错误。 在本文中，我们提出了一种新颖的通用网络结构，以更直接的方式减少混淆错误，并将网络应用于语义分割。 我们的网络结构有两个主要的贡献：1）我们根据有区别的混淆组合成具有异构输出空间的子网。 每个子网的培训可以区分组内的混淆类，而不会影响组外的不相关类。 2）我们提出了一种改进的交叉熵损失函数，该函数最大化分配给正确类的概率，并同时惩罚分配给混淆类的概率。 我们的网络结构是一般结构，可以很容易地适应任何其他网络，以进一步减少混淆错误。在没有对特征编码器和后处理步骤进行任何更改的情况下，我们的实验证明了Cityscapes [12]和PASCAL VOC数据集[14]上不同基线模型的一致且显着的改进（例如，比ResNet-101高出3.05％[18] 和ResNet-38相比1.30％[43]。

**Keywords:** Semantic Segmentation *·* Confusion Errors Deep Network

关键词：语义分割 混淆错误 深度网络

# Introduction

在现实世界中很容易找到具有相似视觉模式的混乱类。例如，在街景中，道路和人行道可以具有接近的颜色外观。 在动作视频剪辑中，手拍和拳击可以共享共同的移动模式[27]。一般而言，即使对于人类来说，也不可能完全避免混淆错误。 但是，在我们为各种视觉相关任务设计的模型中，混乱错误可以在整个训练过程中传播和放大。 在这项工作中，我们主要关注基于深度学习技术的语义分割任务。 扩展可以很容易地进行其他类似的任务，如目标检测和图像分类。在语义分割中，最近提出了大量的网络结构来处理可能产生混淆错误的个别因素。 这些因素包括不平衡的数据分布和特征编码过程中降低的分辨率。 重新采样和重新加权（即成本敏感）策略[4,16,21,19,2]通常用于处理不平衡的类别分布。 但是，由于一些负面因素（如过度拟合的风险）和增加更多少数民族类别的数据集复杂性增加，绩效可能并不总是得到改善。 例如，在Cityscapes数据集中，超过90％的注释来自六个主要类别，包括道路，建筑和植被。 不到10％的注释来自其余13个少数民族。

1. 为了解决由特征编码器引起的分辨率降低问题，基于全卷积神经网络（FCN）提出了大量的深度网络[37,28]。图像金字塔可以输入到同一模型中，特征图最后融合在一起[15,13,​​32,26,8,6]。可以添加多级解码器来恢复特征映射的细节[28,30,34,1,25,33,20,41]。可以应用新的卷积层和汇集层来捕获多尺度的上下文信息[6,7,9,46]。语义分割的当前趋势表明大内核（例如，[31]中的15x15）或具有不同速率的多个迂回卷积（例如，[7,9]中的（6,12,18））以捕获更丰富的上下文信息。但是，仍然不清楚应选择多少不同的速率或内核大小，以及所选择的速率或内核大小是否适合覆盖大范围的对象大小。在街景中，物体尺寸可以显着不同，例如，经过的卡车可以具有相同的图像高度，而交通灯可以仅具有几十个像素。因此，即使[31]中提到的大内核也可能不足以覆盖像卡车这样的对象并捕获足够的上下文信息。

在本文中，我们提出了一种新颖的网络结构，旨在明确减少语义混淆错误。 与现有方法相比，我们的网络结构能够以更直接的方式处理所有因素。 此外，该结构是通用的，并且可以容易地集成到任何现有网络中以进一步改善性能。 具体来说，我们提出的网络结构主要有以下两个贡献。

* + 我们提出了一种使用集成多个子网构建异构输出空间的方法。 这些子网是基于归一化混淆矩阵推断出的判别混淆群构建的。 每个子网旨在扩大每个混淆组中混淆类之间的距离，而不会影响组外的不相关类。
  + 我们提出了一种改进的多类交叉熵损失，它考虑了正确和不正确的标签。 通过为不正确的标签添加新术语，通常由混乱的类引起的错误否定和误报将直接受到惩罚。 基于混淆矩阵的重新加权也适用于新的损失，以进一步加强反馈。

在第2节中，我们介绍了语义图像分割中的相关工作。 第3节详细描述了我们的网络结构和新的损失函数。 本节给出了基于信息论的损失函数分析。 在第4节中，我们提供了一组关于Cityscapes和Pascal VOC数据集的实验。 我们在第5节给出结论。

# Related Work 相关工作

如第1节所述，在网络训练期间，各种因素可能会加剧分类混乱。 我们将因子分为两类：不平衡的数据分布和降低的特征分辨率(imbalanced data distribution and reduced feature resolution)。 在本节中，我们主要描述了处理这两个类别的相关工作。

**Imbalanced Data Distribution.** 一种方法是对少数类进行过采样和/或对多数类进行欠采样。由于这种策略改变了数据分布，过采样可能导致过度拟合，欠采样可能会消除可能的有价值信息。已经提出SMOTE及其变体[4,16,21]以通过生成新的非复制示例来避免过度拟合。另一个方向，即重新加权方法，在不改变数据分布的情况下对少数民族类别施加了额外的惩罚。例如，逆频率和中频重新加权[5,29,44,45,13]已应用于语义分割工作中。在[40]中，建议在线硬件示例挖掘（OHEM）自动选择用于训练基于区域的ConvNet探测器的硬实例。黄等人。 [19]制定了一种新的五元组抽样方法和大规模不平衡分类的三重标题丢失。损失最大池化层[2]定义了一个新的损耗函数，它从像素级加权函数中获得最高损失。在Cityscapes数据集中，这种损失函数可以获得许多少数类别（在一个少数类别上的表现，“卡车”，在不明原因下降级）的性能增益。

**Reduced Feature Resolution.**在特征编码期间，特征图的分辨率逐渐减小，以便捕获对输入图像变换不太敏感的长距离信息。但是，在此编码过程中逐渐压缩或丢失上下文的细节。龙等人。在[28]中提出了用于语义分割的完全卷积网络（FCN），它将完全连接的层转换为卷积层，以便直接生成空间标签图。基于这种结构，已经提出了许多深层网络。改善FCN主要有三个方向。 1）可以添加解码器以逐步恢复上下文细节。 DeconvNet和SegNet [30,1]应用反汇集层来构建类似玻璃的网络以上采样特征映射。 2）扩张卷积（dilated convolutions），也称为空洞卷积（atrous convolutions）[6]，可用于在编码过程中生成具有更高分辨率的特征映射。由于GPU内存有限和其他原因，我们仍需要对许多网络中的特征映射（通常为8）进行下采样。 3）最近的工作重点是捕获多尺度的背景信息。在[31]中，彭等人。将全局卷积网络（GCN）集成到不同级别的特征映射中，并应用解卷积操作来恢复高分辨率标签映射。 GCN中使用的大核（例如，1515）显着扩大了有效的感受野。陈等人。提出了[7]中的atrous空间金字塔汇集（ASPP）模块，它将不稳定的卷积运算与不同的灾害率并行排列，以获得多尺度的背景信息。该模块进一步与[9]中的简单解码器模块相结合。 4）其他上下文模块（例如，条件随机场（CRF））也可以用作后处理步骤或与深度网络联合训练[23,47,26,36]。

**Batch Normalization(批量标准化).** 批量标准化层是语义分割中常用的层。由于最近发现了这一层的重要性，我们简要介绍了该领域的工作，尽管它与我们的贡献没有直接关系。批量标准化参数已添加到ASPP模块中，并且在培训期间发现很重要[7,9]。该策略是计算具有较大批量大小和较小特征图（例如，16个下采样率）的批量标准化参数，并使用较大的特征图（例如，8个下采样率）冻结参数。在[3]中，已经提出了就地激活批量标准化（INPLACE-ABN）来减少训练记忆，从而可以增加批量大小并且批量标准化的统计数据可以更准确。这种新颖的批量标准化层可以将ResNet-38模型的性能从78.08％提高到79.40％，而无需对网络进行其他修改。目前，INPLACE-ABN在Cityscapes基准测试中排名第一。

在本节中，我们描述了先进的深度网络，它们极大地提高了语义图像分割的性能。 到目前为止，据我们所知，这一领域的现有工作都没有探讨如本文所提出的明确减少混淆错误。

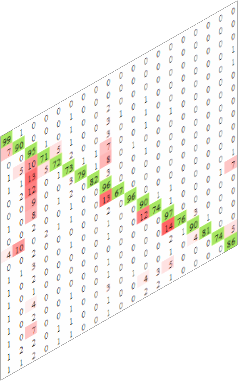
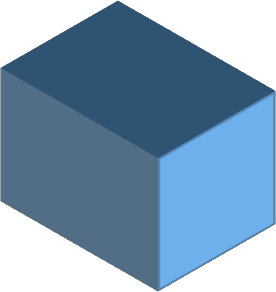
# Our Approach 我们的方法

整体网络结构如图1所示。我们的子网和损失层可以轻松集成到大多数现有网络结构中。我们首先将原始网络分为两部分。主要部分用作特征编码器。一个或多个卷积层和相关的批量标准化或激活层用作我们的第一个子网（即图1中的子网0）。剩余子网（M-1）的数量由混淆组的数量决定。这个数字一般不会太大，例如包含19个类的Cityscapes数据集中的三个。此外，如果我们有一个非常复杂的数据集，我们可以使用阈值来减少数量，以便选定的子网专注于更混乱的组。每个子网都针对每个混淆组进行单独培训。训练之后，异构输出分数被转换并融合在一起以获得最终概率或分数。请注意，可以针对特定的混淆组调整每个子网的结构。例如，我们可以使用更多或更少的卷积层，不同的黑色损失率，以及来自不同ResNet块的连接特征映射。

## Discriminative Confusing Groups 判别混淆组

从集成学习的角度来看，我们可以将网络中的子网融合和[7,9]中的ASPP模块视为集合分类器。 每个子网都可以被视为我们网络中的分类器，并且ASPP模块中的批量规范化之后的每个自发卷积操作也可以被视为分类器。主要区别在于ASPP模块将所有类考虑在内，尽管模块使用不同的动态速率。目前尚不清楚各个分类器是否足够多样化以及整体中应包括多少组件分类器。

…



Confusion Matrix

Dim = C0

Subnet 0

Encoder

Dim = C1

Subnet 1

Dim = C2

Subnet 2

Feature Map

Output stride:8

Dim = CM

Subnet M

Loss M

Loss 2

Loss 1

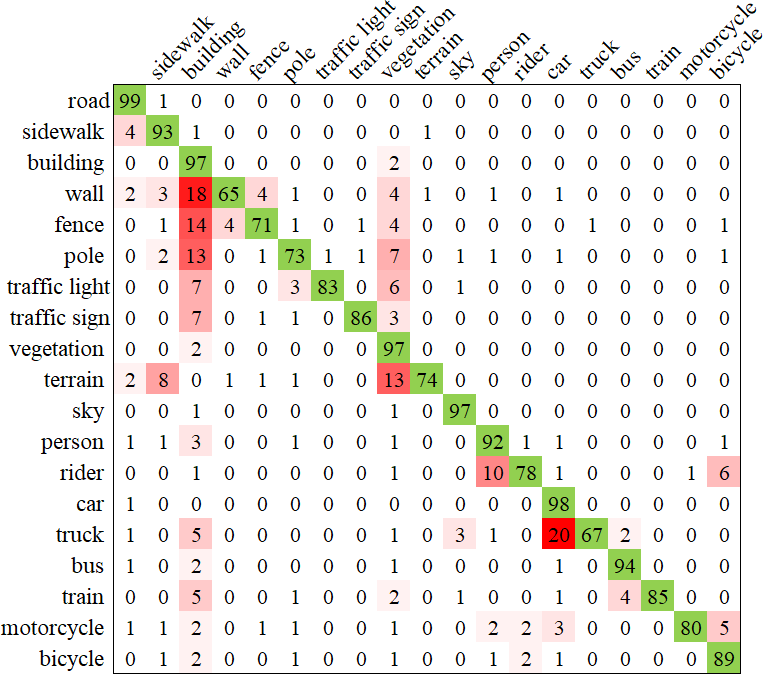
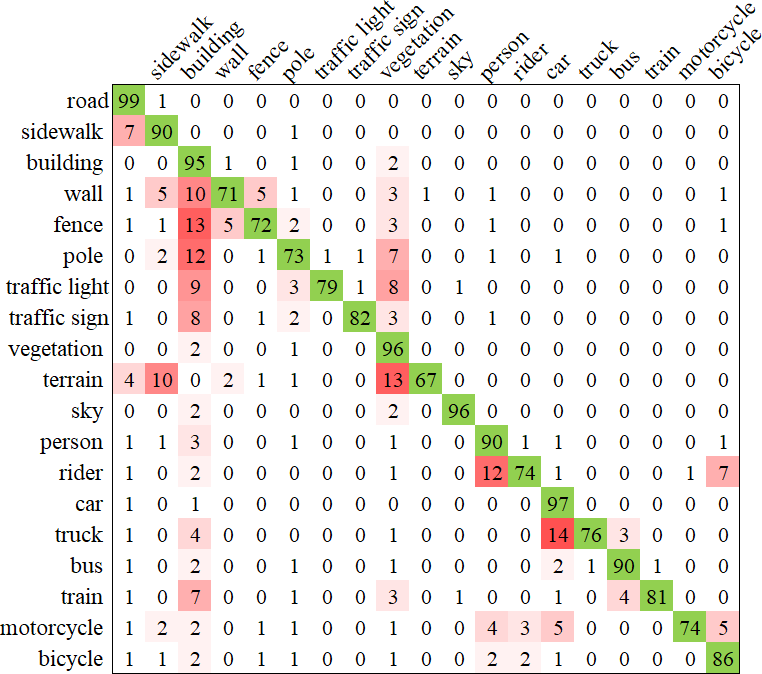
Loss 0



Fusion

**Fig. 1.**Overview of our proposed network structure that contains *M* subnets.

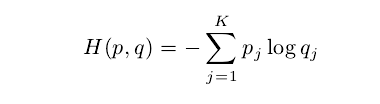
在图2中，我们比较了针对Cityscapes数据集从ResNet-101和ResNet-38的预训练模型计算的归一化混淆矩阵。 很容易发现计算的矩阵具有非常相似的模式，例如“墙”类与两个网络中的“建筑”类密切相关，尽管错误分类率分别为18％和10％。 鉴于这种模式，我们可以将所有类别划分为几个有区别的混淆组，其中组间混淆错误非常小并且可以忽略。 结果，歧视性混淆群体的数量可能是实际且适当的整体大小或整体基数。



**Fig. 2.**Two confusion matrices computed based on the pre-trained models of ResNet-101 (*right*) and ResNet-38 (*left*). The matrices share a very similar pattern that indicates the indepedent confusing groups.

## Improved Cross-Entropy Loss 改进的交叉熵损失

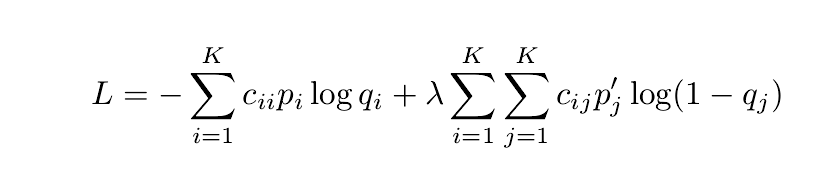
在信息论中，地面实际分布p（即分类中的单热标签）与估计分布q之间的交叉熵由下式给出：

(1)

*K* 表示分类的数量，是softmax分类器的输出。 *φ* 代表网络、 *x*表示输入图像，他也可以解释为与分配给正确类的概率相关的损失，而不考虑正确的类和其余类之间的关系，尤其是混淆类。 为了减少对另一个错误类的混淆，我们还需要减去分配给不正确类的概率。最终, 我们可以得到一个新的损失

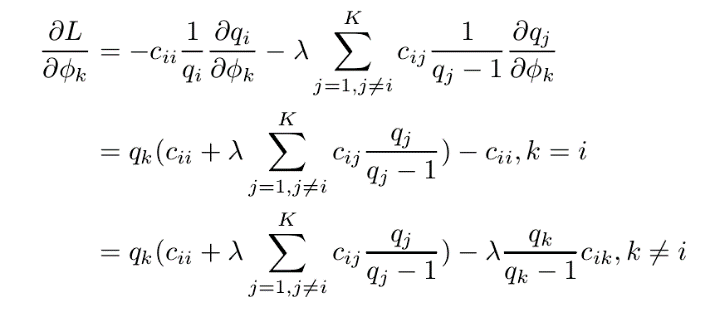
*L* = *H*(*p, q*) *− H*(1 *− p,* 1 *− q*) (2)

在上面的等式中，我们平等地对待正确的类和剩余的类，其中仍然没有考虑混淆类。然后, 我们使用权重矩阵对新损失进行加权 (*C* = [*cij*]*K×K*) ，这可以从归一化的混淆矩阵计算出来。 然后将等式2转换成：



(3)

其中 *p′*= *p*-1并且*λ*被用作平衡正确与错误分类的损失，损失函数（3）的关于x的导数 *φ*(*x*) 如下：



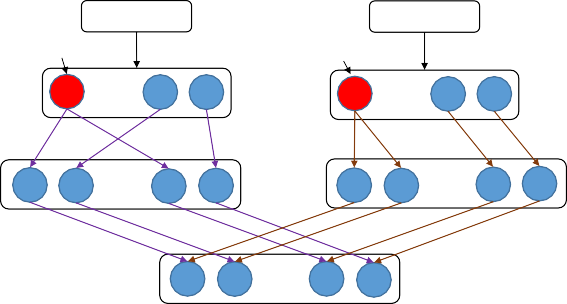
(4)

## Fusion of Heterogeneous Output Spaces 异构输出空间的融合

由于子网属于异构输出空间（表示为源输出空间），我们还需要找到一种方法将它们首先转换为包含所有标签的目标输出空间，然后我们可以应用传统的集合方法，如求和规则或产品规则[22]。 有许多方法可以进行转换，例如回归模型[38,39]和神经网络[11]。I在[39]中，相似性保持原理指出，对于源输出空间中的每对类(*v*1*, v*2) ，存在相似性指标 *π*使得*π*(*v*1*, v*2) = *π*(*f* (*v*1)*, f* (*v*2)), 其中 *f* 代表从源输出空间到目标输出空间的变换函数。

在我们网络中的每个子网中，我们有属于相应混淆组的类和一个包含组外所有类的“其他”类。 由于源输出空间具有一定的重叠区域，我们不需要在[39,38]中应用回归模型。 在这里，我们可以应用如图3所示的简单方法。对于组中的混淆类，源输出空间和目标输出空间之间存在一对一的映射。 对于“其他”类，我们构建从源到目标输出空间中组外的所有剩余类的一对多映射。 不难证明这种简单的变换满足相似性保持原理(*similarity preserving principle*)。

Transformation



Subnet 1

… …

Others 1

Subnet M

Source Output 1 Others M Source Output M

… … … … … …

… …

… …

… …

Target Output 1

Target Output M

Target Output

… …

Fusion

**Fig. 3.**The transformation from source output space to target output space that satisfies *similarity preserving principle* in [[39].](#_bookmark58)

也可以设计另一个网络来学习转换，这可能是我们未来的工作之一。 在实验中，我们证明，即使使用我们的简单转换和融合，我们仍然可以实现对基线模型的一致改进。

# Experiments 实验

我们使用官方MXNET工具[10]评估Cityscapes数据集[12]和扩展PAS-CAL VOC数据集[14]的网络结构。基于平均交叉结合（mIoU）评估性能。在训练期间，我们使用标准SGD [24]，动量为0.9，重量衰减为0.0005。初始学习率为0.002，并以线性时间表更新。在训练期间应用数据增加，例如平均牵引力，随机裁剪和随机左右翻转。

## Cityscapes （数据集）

包含在50个城市收集的24,998个街景.有分辨率2048x1024的5,000张图像被精确注释，剩余的19,998张被粗略注释.5,000个精细注释图像进一步分为列车，验证和测试集，分别具有2,975,500和1,525个图像。 19个语义对象类用于评估。选择两个基线模型ResNet-101和ResNet-38用于Cityscapes数据集上的实验。删除原始ResNet-101的最后1000路分类层。 通过改变第3块和第4块的卷积步长，语义分段任务的特征步幅从32减少到8. ResNet-101在ImageNet [35]上进行了预训练，并在Cityscapes上进行了100个时期的微调。在[43]中，使用ResNet-101在Cityscapes验证数据集上报告了73.63％的mIoU。 我们的ResNet-101获得74.70％，高出1.03％。 ResNet-38基线模型是[43]中最初发布的模型，其中mIoU为78.08％。

我们将所有类划分为三个混淆组，并构建三个额外的子网。 表1给出了这些令人困惑的群体的详细信息。每个子网中的“其他”类包含与相应组中的混淆类无关的所有剩余类。 图4显示了所有子网的结构。 在子网训练期间，特征编码器中的所有参数都是固定的。

**Table 1.**Class names of three confusing groups on Cityscapes dataset.

|  |  |  |
| --- | --- | --- |
| subnet | nClass | Class Name |
| subnet 0 | 19 | all the 19 classes |
| subnet 1 | 7 | others1, building, wall, fence, pole, traffic light, traffic sign |
| subnet 2 | 5 | others2, car, truck, bus, train |
| subnet 3 | 5 | others3, person, rider, motorcycle, bicycle |

**Improvement (提升)on ResNet-101. 表**[3](#_bookmark14)  显示了使用子网的实验和改进的交叉熵损失。在不使用我们提出的新损失函数的情况下，我们将mIoU提高了0.85％。我们的融合方法的改进主要集中在混淆类。 这些混乱群体的平均收益分别为0.45％，1.57％和1.70％。为了评估新的损失，我们仅通过移除子网1到3来仅仅只使用子网0.我们可以通过λ= 5 将mIoU提高到76.21％。当三个子网和新损失堆叠在一起时，我们能够获得77.75％的mIoU，比基线提高3.05％。 图8给出了一些混淆类（例如，墙，杆，骑手，卡车和围栏）的视觉结果示例。

Feature Map

Subnet 0 Subnet 1 Subnet 2 Subnet 3

3×3×512，Dilation 12 3×3×512，Dilation 12 3×3×512，Dilation 12 3×3×512，Dilation 12

3×3×512，Dilation 12

3×3×512，Dilation 12 3×3×512，Dilation 12

3×3×19，Dilation 12

3×3×7，Dilation 18

3×3×5，Dilation 12

3×3×5，Dilation 12

Heterogeneous Transformation + Fusion

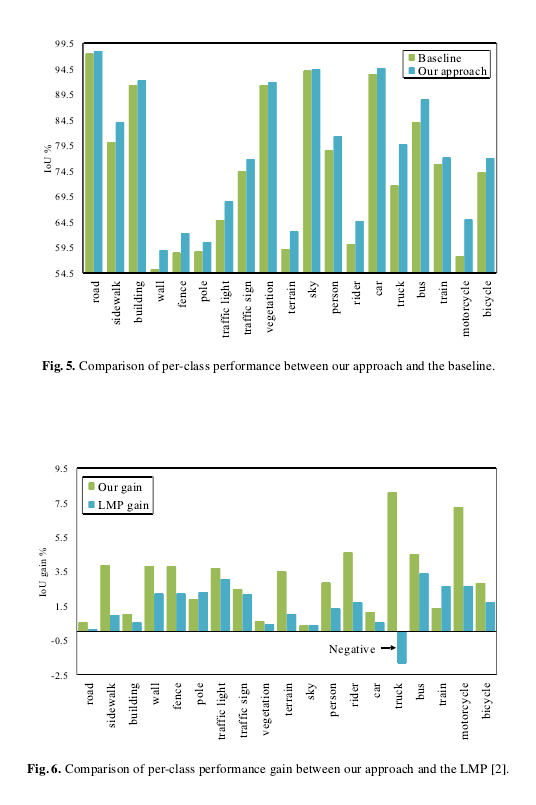
Output Scores

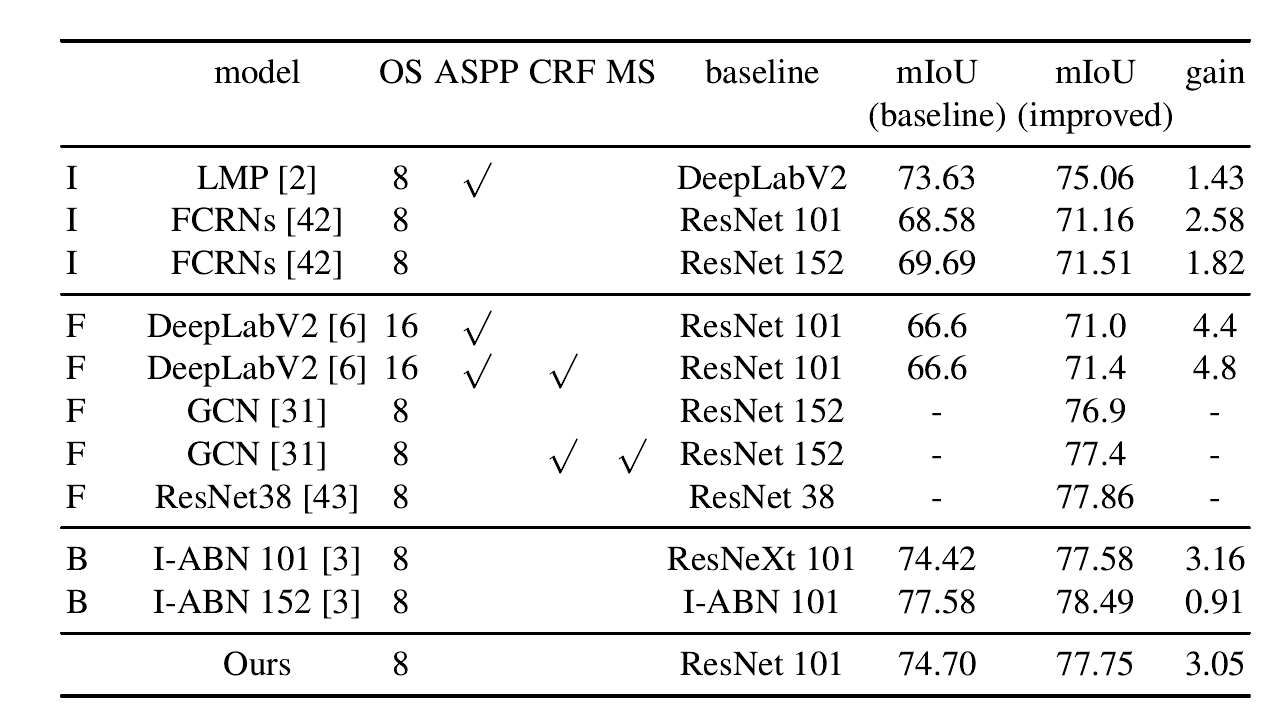
**Fig. 4.**The structure and shape information of four subnets used for Cityscapes dataset.

**Per-class Performance on ResNet-101（**ResNet-101上的每级性能**）.** 在图5中，我们展示了我们的方法与ResNet-101基线的每级性能。 我们发现，与基线相比，18类的IoU值得到了极大的改善。 某些容易混淆的分类，如人行道，墙壁，围栏，骑手，卡车，公共汽车和摩托车，IoU增加超过3.5％。 该结果表明，我们的子网和改进的交叉熵损失可有效减少混淆错误。 在图6中，我们进一步将每类IoU增益与LMP的每类IoU增益进行比较[2]。 我们发现大多数类的IoU增益大于LMP的增益。一个可能的原因是LMP主要用于减少不平衡的数据问题。然而，混淆错误可能来自其他因素，不仅限于不平衡的数据分布。 因此，这表明明确处理混淆错误可能更有益和有效。

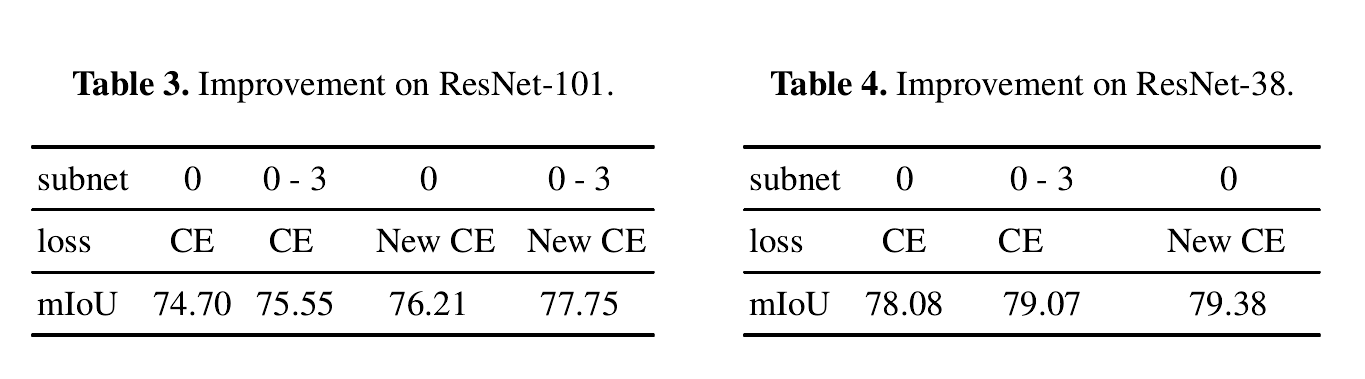
**Comparison with Other Approaches（与其它方法相比较）。**这里，我们将我们的方法和其他方法进行了比较。 为了显示包括基线模型、参数、mIoU收益的综合比较，我们主要选择近年来已发表文章中报道的方法。 我们基于作者要求解决的问题将这些方法大致划分为三个不同的类别，即“I”表示不平衡数据，“F”表示特征提取（例如，多尺度上下文信息），“B” 改进批量标准化。 表2显示了比较。 由于已对表中的所有方法进行了数据扩充，因此表中未显示此选项。

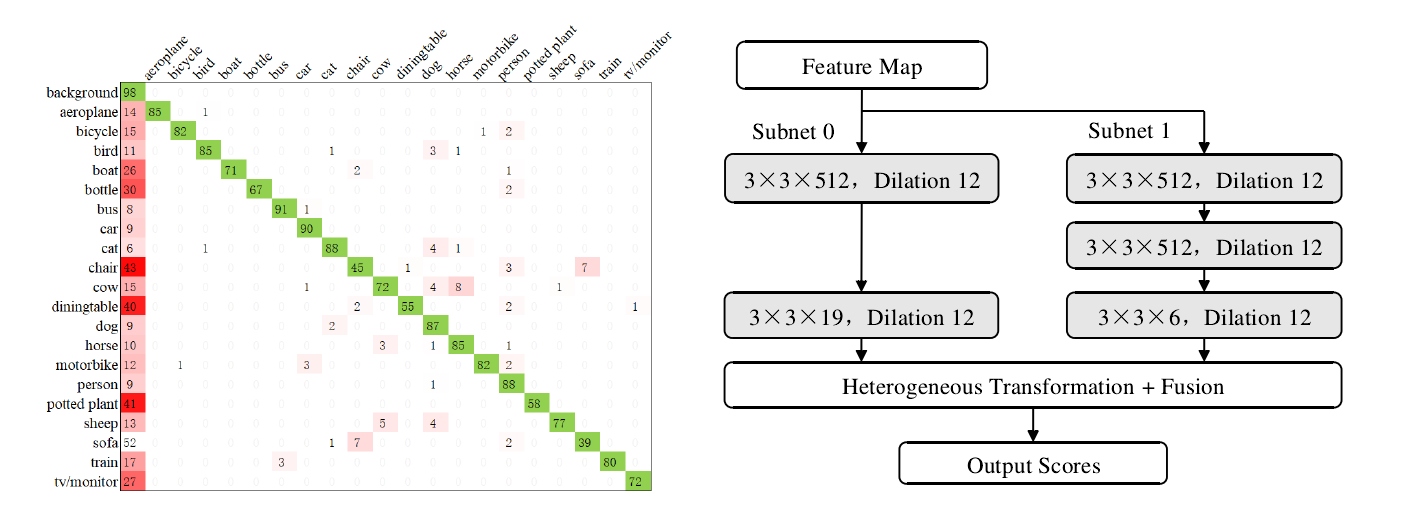
虽然表中的模型旨在解决不同的问题，但我们改进的mIoU和mIoU增益与近年来提出的现有方法相当。 请注意，I-ABN [3]目前在Cityscapes基准测试中以mIoU 82.0％排名第一。 如果我们只考虑修改的批量标准化层，本文报道的改进的mIoU（77.58％）分别接近基线ResNeXt 101和Resnet 101的改进的mIoU（77.75％）。

基于此表，我们发现，为了获得Cityscapes的最佳分割性能，有必要将多种技术集成在一起，例如ASPP，I-ABN和LMP。 这告诉我们网络结构应该通用且灵活，以适应其他不同的结构。 我们的方法很通用，可以很容易地与许多现有工作相结合，以进一步提高性能。

**Table 2.** **与城市景观验证集的其他方法进行综合比较。 第一栏是指作者要求解决的问题类别。 OS表示输出步幅，MS表示测试中使用的多尺度输入图像。 √表示该技术应用于模型中，白色空间表示未在模型中使用，“ - ”表示该文章中未报告该技术或数字。**

当然，我们希望将我们的方法应用于表2中列出的现有最先进的算法。但是，经过仔细检查后，这些算法的发布版本不足以允许修改，通常只发布测试模型。 因此，我们选择评估我们在ResNet-38发布模型上的方法（mIoU为78.08％，略高于本文报道的77.86％）。 在不使用新损失的情况下，我们将mIoU提高了0.99％。 随着新的损失和子网0，我们可以将mIoU提高到79.38％，比ResNet-38的发布型号提高1.30





**Fig. 7.**(*left*) The confusion matrix for PASCAL VOC dataset. (*right*) The structure and shape information of our subnets used for PASCAL VOC dataset.

## PASCAL VOC 2012

PASCAL VOC 2012 有1,464张图像用于训练，1,449张图像用于验证，1,456张图像用于测试。 包括“背景”类在内的21个对象类被标记。 我们还使用语义边界数据集[17]作为辅助数据集，产生了10,582个用于训练的图像。

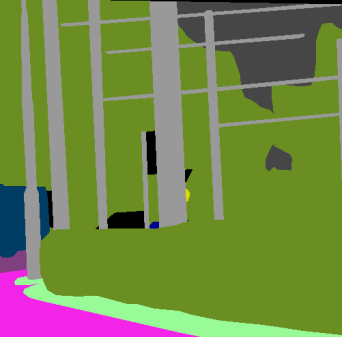
选择ResNet-101 进行PASCAL VOC数据集的实验。ResNet-101的结构与用于评估Cityscapes数据集的结构相同。同样，ResNet-101在ImageNet上进行了预训练，并在VOC数据集上进行了80个时期的微调。在[43]中，在PASCAL VOC验证数据集上报告了ResNet-101的75.35％mIoU。我们的ResNet-101获得75.43％，略高。

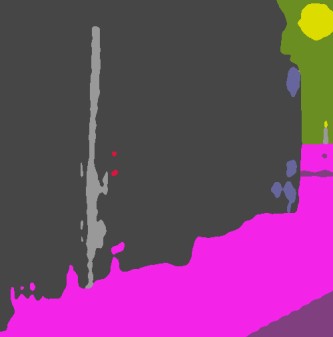
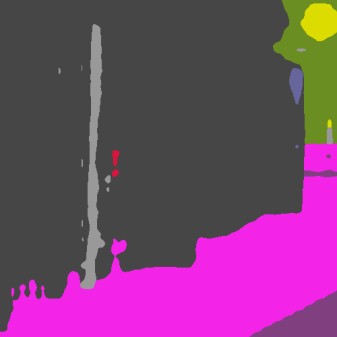
基于图7左侧所示的混淆矩阵，我们发现许多类与“背景”类混淆。 因此，对于这个令人困惑的组（即，其他，背景，椅子，餐桌，盆栽植物和沙发）仅添加一个子网。 我们网络的结构如图7右侧所示。

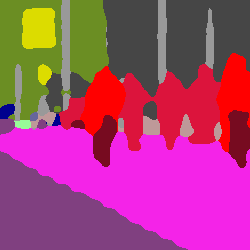
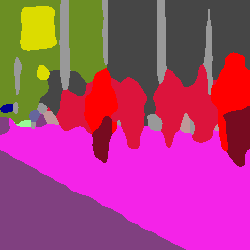
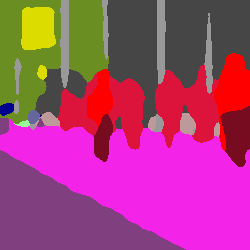
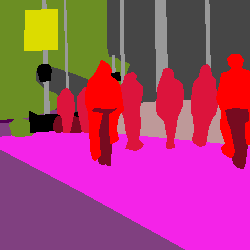
**Improvement（提升） on ResNet-101.** 表5显示了使用子网和改进的交叉熵损失的实验。 使用子网1时，mIoU增加到75.51％。 当应用改进的交叉熵损失时，mIoU进一步增加到76.91％。 一些视觉结果如图9所示。

**Table 5.**Improvement on PASCAL VOC dataset using ResNet-101.

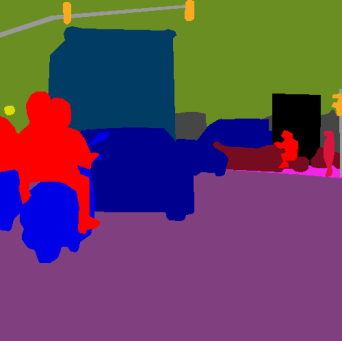
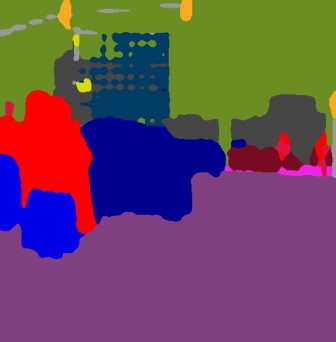
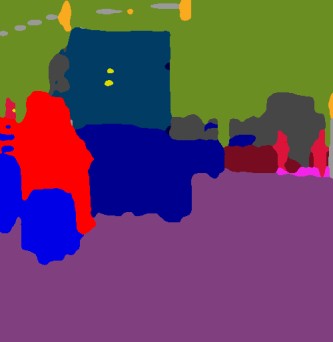
|  |  |  |
| --- | --- | --- |
| subnet | 0 0 and 1 | 0 and 1 |
| loss | CE CE | New CE |
| mIoU | 75.43 76.51 | 76.91 |



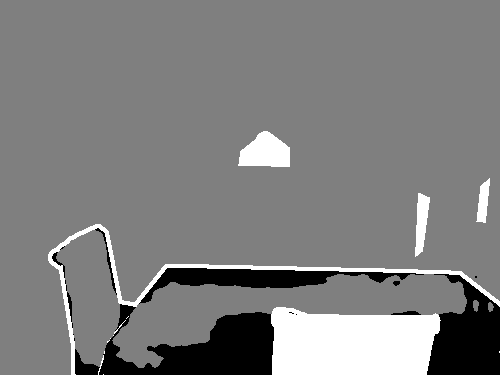
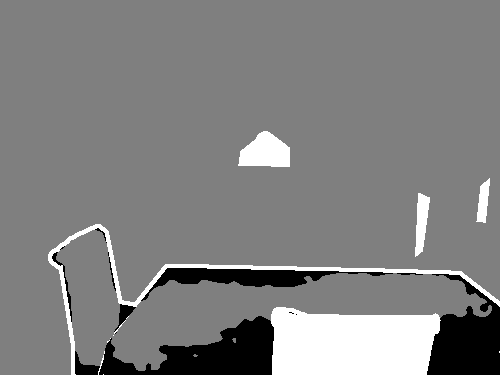
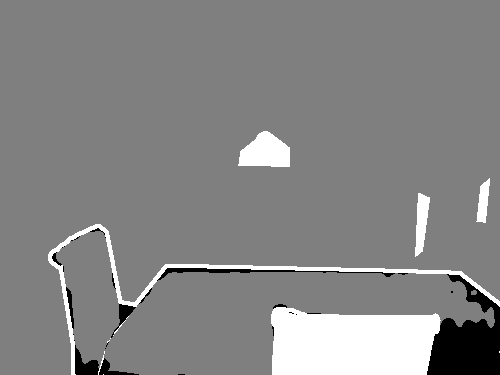
    

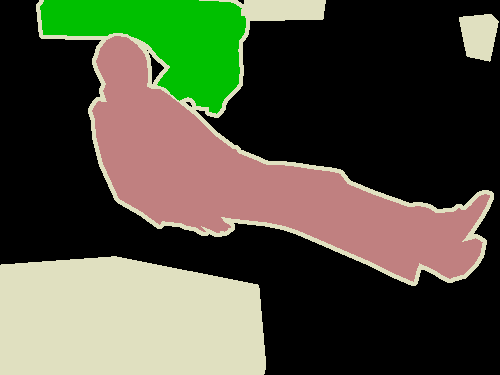
    

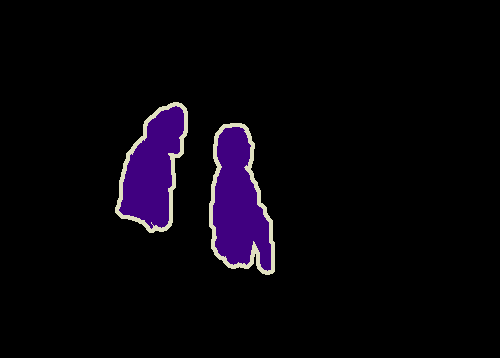
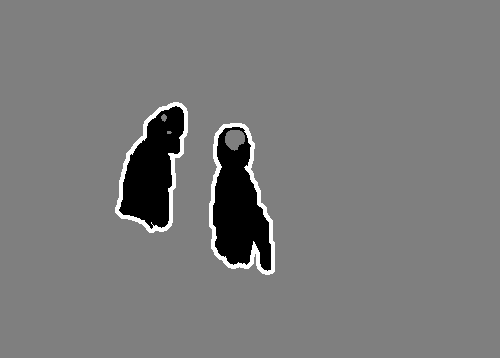
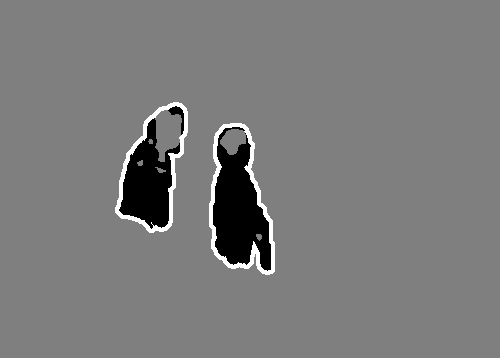
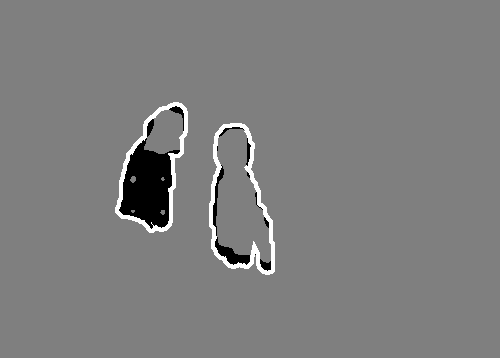
    

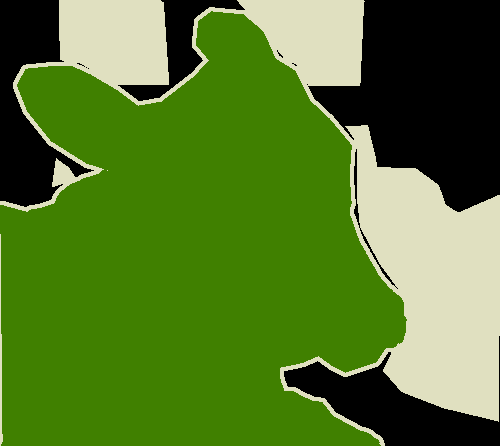


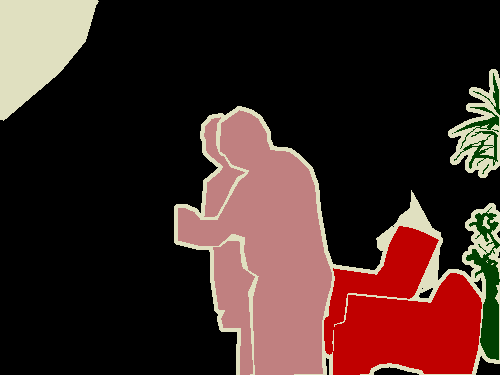
**Fig. 8.**Examples of semantic segmentation results on Cityscapes (cropped for visualization pur- pose). For every row, we list the cropped input image, ground truth label, label estimated from the baseline, label estimated from baseline + subnets, and label estimated from baseline + sub- nets + improved loss. Confusing classes for each column are (from left to right): (building, pole), (building, pole, fence), (rider, person), (truck, bus), (truck, building), and (fence, building).

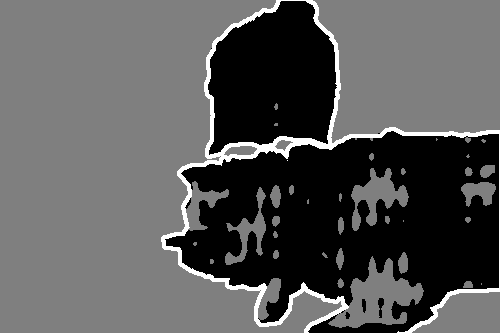
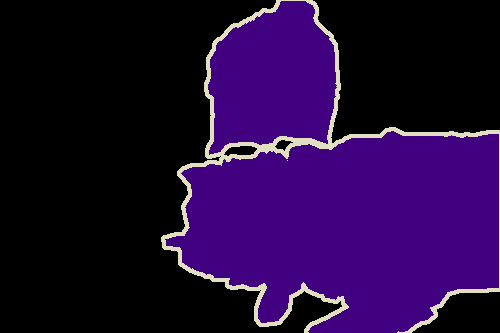
    

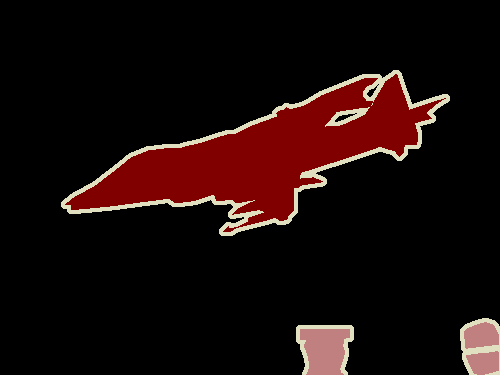
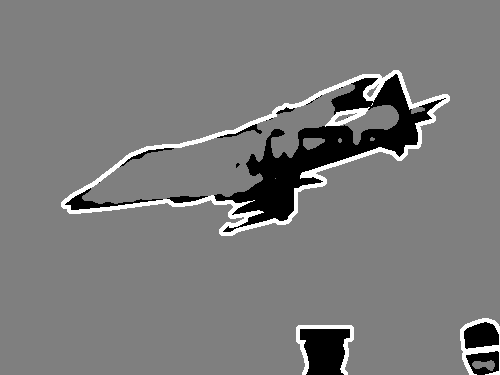
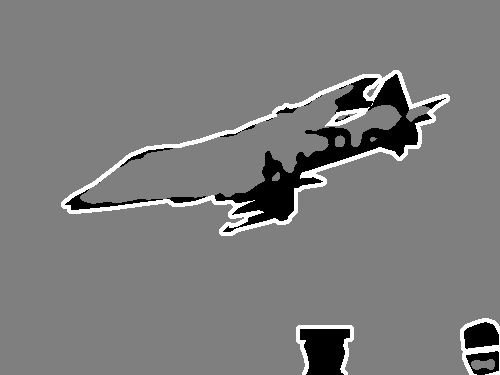
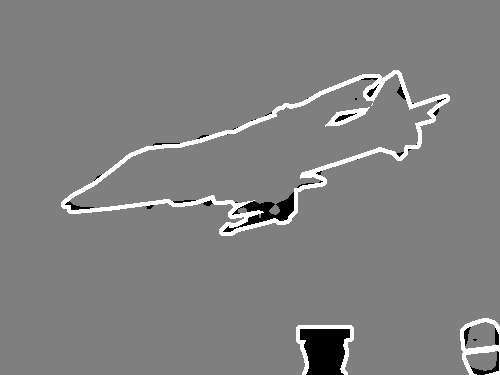


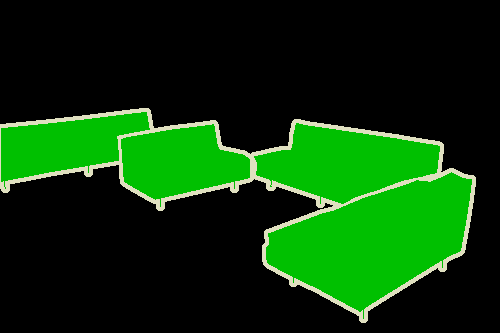
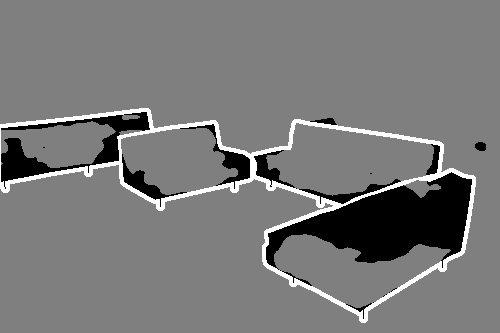
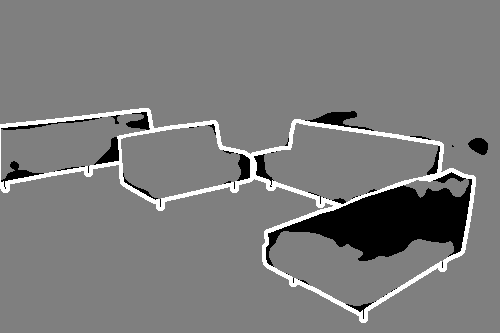
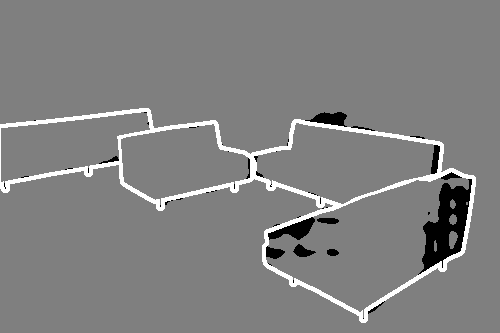
    





**Fig. 9.**Examples of semantic segmentation results on PASCAL VOC dataset. For every row, we list input image, ground truth label, error map for baseline, error map for baseline + subnets, error map for baseline + improved loss. Error maps in the last three rows represent the label inconsistencies, in which black color means labels are different and gray color means labels are the same.

# Conclusion

在本文中，我们提出了一种新的网络结构，以减少可能来自不同因素的语义混淆错误。大多数现有作品仅旨在处理个别影响因素，而我们的方法处理混淆错误更加直接。我们的方法包括两个主要组成部分：

1）从归一化混淆矩阵估计的具有判别混淆群的异质输出的子网集合；

2）改进的交叉熵损失，新术语惩罚通常由混乱的类引起的假阴性和误报；

我们的实验表明，这两个组件都是有效的，并且可以改善不同基线模型和具有不同复杂性的数据集的分割性能。更重要的是，我们的方法是通用且灵活的，可以很容易地适应大多数现有的网络结构。

# References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine in- telligence **39**(12), 2481–2495 (2017)
2. Bulo, S.R., Neuhold, G., Kontschieder, P.: Loss maxpooling for semantic image segmenta- tion. CVPR), July **7** (2017)
3. Bul o`, S.R., Porzi, L., Kontschieder, P.: In-place activated batchnorm for memory-optimized training of dnns. arXiv preprint arXiv:1712.02616 (2017)
4. Bunkhumpornpat, C., Sinapiromsaran, K., Lursinsap, C.: Safe-level-smote: Safe-level- synthetic minority over-sampling technique for handling the class imbalanced problem. In: Pacific-Asia conference on knowledge discovery and data mining. pp. 475–482. Springer (2009)
5. Caesar, H., Uijlings, J., Ferrari, V.: Joint calibration for semantic segmentation. arXiv preprint arXiv:1507.01581 (2015)
6. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic im- age segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Transactions on Pattern Analysis and Machine Intelligence **PP**(99), 1–1 (2016)
7. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for se- mantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
8. Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to scale: Scale-aware semantic image segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3640–3649 (2016)
9. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. arXiv preprint arXiv:1802.02611 (2018)
10. Chen, T., Li, M., Li, Y., Lin, M., Wang, N., Wang, M., Xiao, T., Xu, B., Zhang, C., Zhang, Z.: Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems. arXiv preprint arXiv:1512.01274 (2015)
11. Ciss e´, M., Al-Shedivat, M., Bengio, S.: Adios: Architectures deep in output space. In: Inter- national Conference on Machine Learning. pp. 2770–2779 (2016)
12. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213– 3223 (2016)
13. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2650–2658 (2015)
14. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**(1), 98–136 (2015)
15. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Learning hierarchical features for scene labeling. IEEE transactions on pattern analysis and machine intelligence **35**(8), 1915–1929 (2013)
16. Han, H., Wang, W.Y., Mao, B.H.: Borderline-smote: a new over-sampling method in imbal- anced data sets learning. In: International Conference on Intelligent Computing. pp. 878– 887. Springer (2005)
17. Hariharan, B., Arbel a´ez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: Computer Vision (ICCV), 2011 IEEE International Conference on. pp. 991– 998. IEEE (2011)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Pro- ceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. Huang, C., Li, Y., Change Loy, C., Tang, X.: Learning deep representation for imbalanced classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5375–5384 (2016)
20. Islam, M.A., Rochan, M., Bruce, N.D., Wang, Y.: Gated feedback refinement network for dense image labeling. In: 2017 IEEE Conference on Computer Vision and Pattern Recogni- tion (CVPR). pp. 4877–4885. IEEE (2017)
21. Jeatrakul, P., Wong, K.W., Fung, C.C.: Classification of imbalanced data by combining the complementary neural network and smote algorithm. In: International Conference on Neural Information Processing. pp. 152–159. Springer (2010)
22. Kittler, J.: Combining classifiers: A theoretical framework. Pattern analysis and Applications

**1**(1), 18–27 (1998)

1. Kr a¨henbu¨hl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: Advances in neural information processing systems. pp. 109–117 (2011) 24.Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional

neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)

1. Lin, G., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks with iden- tity mappings for high-resolution semantic segmentation. arXiv preprint arXiv:1611.06612 (2016)
2. Lin, G., Shen, C., Van Den Hengel, A., Reid, I.: Efficient piecewise training of deep struc- tured models for semantic segmentation. In: Proceedings of the IEEE Conference on Com- puter Vision and Pattern Recognition. pp. 3194–3203 (2016)
3. Liu, W., Tsang, I.W., M u¨ller, K.R.: An easy-to-hard learning paradigm for multiple classes and multiple labels. The Journal of Machine Learning Research **18**(1), 3300–3337 (2017)
4. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)
5. Mostajabi, M., Yadollahpour, P., Shakhnarovich, G.: Feedforward semantic segmentation with zoom-out features. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3376–3385 (2015)
6. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1520–1528 (2015)
7. Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J.: Large kernel matters–improve semantic seg- mentation by global convolutional network. arXiv preprint arXiv:1703.02719 (2017)
8. Pinheiro, P., Collobert, R.: Recurrent convolutional neural networks for scene labeling. In: International conference on machine learning. pp. 82–90 (2014)
9. Pohlen, T., Hermans, A., Mathias, M., Leibe, B.: Full-resolution residual networks for se- mantic segmentation in street scenes. arXiv preprint (2017)
10. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical im- age segmentation. In: International Conference on Medical image computing and computer- assisted intervention. pp. 234–241. Springer (2015)
11. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. Interna- tional Journal of Computer Vision **115**(3), 211–252 (2015)
12. Schwing, A.G., Urtasun, R.: Fully connected deep structured networks. arXiv preprint arXiv:1503.02351 (2015)
13. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: Inte- grated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229 (2013)
14. Shi, X., Liu, Q., Fan, W., Philip, S.Y., Zhu, R.: Transfer learning on heterogenous feature spaces via spectral transformation. In: Data Mining (ICDM), 2010 IEEE 10th International Conference on. pp. 1049–1054. IEEE (2010)
15. Shi, X., Liu, Q., Fan, W., Yang, Q., Yu, P.S.: Predictive modeling with heterogeneous sources. In: Proceedings of the 2010 SIAM International Conference on Data Mining. pp. 814–825. SIAM (2010)
16. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 761–769 (2016)
17. Wojna, Z., Ferrari, V., Guadarrama, S., Silberman, N., Chen, L.C., Fathi, A., Uijlings, J.: The devil is in the decoder. arXiv preprint arXiv:1707.05847 (2017)
18. Wu, Z., Shen, C., Hengel, A.v.d.: High-performance semantic segmentation using very deep fully convolutional networks. arXiv preprint arXiv:1604.04339 (2016)
19. Wu, Z., Shen, C., Hengel, A.v.d.: Wider or deeper: Revisiting the resnet model for visual recognition. arXiv preprint arXiv:1611.10080 (2016)
20. Xu, J., Schwing, A.G., Urtasun, R.: Tell me what you see and i will show you where it is. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3190–3197 (2014)
21. Xu, J., Schwing, A.G., Urtasun, R.: Learning to segment under various forms of weak super- vision. In: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on. pp. 3781–3790. IEEE (2015)
22. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 2881–2890 (2017)
23. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.: Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1529–1537 (2015)