International Journal of Computer Vision manuscript No.
(will be inserted by the editor)

基于不同深监督的语义边缘检测

Yun Liu[1] · **Ming-Ming Cheng**[1] · **Deng-Ping Fan**[1] · **Le Zhang**[2] · **Jia-Wang Bian**[3] · **Dacheng Tao**[4]

Received: date / Accepted: date

摘要 语义边缘检测（Semantic Edge Detection, SED）旨在同时提取边缘及其类别信息，在语义分割、拟物性采样和物体识别等领域具有广泛应用。SED 需要实现两个不同的监督目标，即检测出详细的边缘信息并识别其高级语义。我们分析了这种分散的监督目标是如何阻碍最新的 SED 方法有效地利用深监督来改善结果的。在本文中，我们提出了一种新颖的全卷积神经网络，它在多任务框架中使用不同深监督（Diverse Deep Supervision, DDS），其中低层旨在生成与类别无关的边缘，而高层则负责检测类别感知的语义边缘。为了克服分散监督的挑战，本文引入了一种新的信息转换器，其有效性在包括 SBD、Cityscapes 和 PASCAL VOC2012 在内的几个流行的基准数据集中得到了广泛评测。源代码将在论文被接收后公布。

关键词 语义边缘检测，不同的深监督，信息转换器

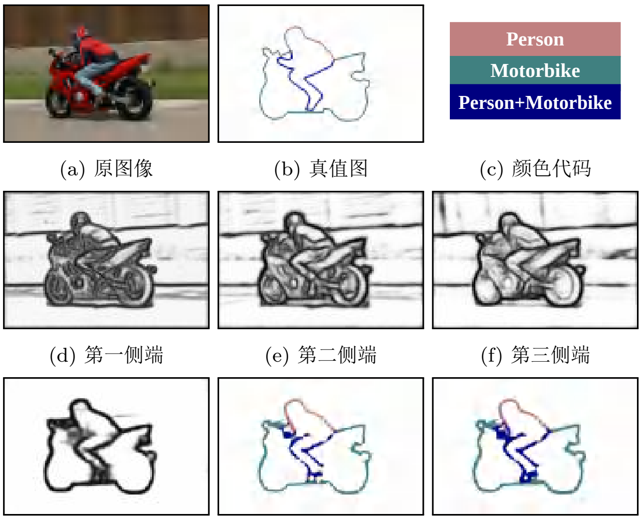
This research was supported by NSFC (NO. 61572264), the national youth talent support program, Tianjin Natural Science Foundation for Distinguished Young Scholars (NO. 17JCJQJC43700), Tianjin key S&T Projects on new generation AI, and Huawei Innovation Research Program.

[1] Y. Liu, M.M. Cheng and D.P. Fan are with College of Computer Science, Nankai University. M.M. Cheng is the corresponding author (cmm@nankai.edu.cn).
[2] L. Zhang is with Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR).
[3] J.W. Bian is with the School of Computer Science, University of Adelaide
[4] D. Tao is with the School of Information Technologies, University of Sydney.

1 引言

传统的边缘检测的目的是检测自然图像中的边缘和物体边界。它是与类别无关的，即不需要识别物体类别。传统的边缘检测可以看作是逐像素的二分类问题，其目的是将每个像素分类为边缘像素或非边缘像素。在本文中，我们考虑更现实的场景，即语义边缘检测（Semantic Edge Detection, SED），它同时实现了图像中的边缘检测和边缘类别识别。由于 SED 具有广泛的应用，如拟物性采样(Bertasius et al. 2015b)、遮挡和深度推理(Amer et al. 2015)、三维重建(Shan et al. 2014)、物体检测(Ferrari et al. 2008, 2010)和基于图像的定位(Ramalingam et al. 2010)等，使得其 (Bertasius et al. 2015b; Hariharan et al. 2011; Maninis et al. 2017; Yu et al. 2017) 成为计算机视觉中的一个热门研究话题。

在过去的几年中，深度卷积神经网络毫无争议地成为用于类别无关的边缘检测的基本方法 (Hu et al. 2018; Liu et al. 2019, 2017; Xie & Tu 2015, 2017)，已经实现了接近人类水平的性能。但是，基于深度学习的类别感知的 SED（可以同时检测视觉上显著的边缘并识别其类别）仍没有得到如此广泛的探索。Hariharan 等人(2011)率先结合一般的物体检测器与自底向上的边缘来识别语义边缘。Yang 等人(2016)提出了一个全卷积的编码-解码网络，用于检测物体边界，但不识别其特定类别。最近，CASENet (Yu et al. 2017)引入了跳层结构来利用低层特征来丰富顶层的类别感知的边缘响应，从而显著地改善了以前的方法。然而，CASENet 仅对第五侧端（Side）和最终融合的分类添加了监督，并认为没有必要对



(a) 原图像 (b) 真值图 (c) 颜色代码
(d) 第一侧端 (e) 第二侧端 (f) 第三侧端
(g) 第四侧端 (h) 第五侧端 (i) DDS

图 1 所提出的 DDS 算法的一个例子。(a) 为 SBD 数据集(Hariharan et al. 2011)中的原图像。(b)-(c) 展示了其语义边缘的真值图和相应的颜色代码。(d)-(g) 显示了第一侧端 ∼ 第四侧端的与类别无关的边缘。(h)-(i) 分别显示了第五侧端和 DDS（DDS-R）输出的语义边缘。

低层（第一侧端 ∼ 第四侧端）进行深监督。在设计中，CASENet 只使用了第一侧端 ∼ 第三侧端中的特征图，而不添加深监督。

SED 中的分散监督悖论. SED 需要实现两个不同的监督目标：i) 通过捕获图像区域之间的不连续性来定位精确的细节边缘（主要使用底层特征）；ii) 通过总结不同物体类别的外观变化来识别抽象的高层语义。这种分散监督悖论已经阻碍了将深监督成功应用于最新的 SED 方法 CASENet (Yu et al. 2017)，而深监督的有效性已经在许多其他计算机视觉任务，如图像分类(Szegedy et al. 2015)、物体检测(Lin et al. 2020)、视觉跟踪(Wang et al. 2015)、以及类别无关的边缘检测(Liu et al. 2017; Xie & Tu 2017) 中得到了证明。为了避免分散监督悖论，基于 CASENet 的方法 (Acuna et al. 2019; Hu et al. 2019; Yu et al. 2017, 2018) 仅融合了第一侧端 ∼ 第三侧端中的特征，并且没有对其使用深监督。

在本文中，我们提出了一个使用不同深监督（Diverse Deep Supervision, DDS）的方法，该方法对高层和底层的特征学习使用具有不同损失函数的深监督，如图 2(b) 所示。尽管将高层卷积特征用于语义分类、将底层卷积特征用于非语义边缘细化这一做法是直观且直接的，但是像 CASENet (Yu et al. 2017) 一样直接这么做的话，其性能会比直接学习语义边缘而不加深监督或类别无关的边缘指导还要低。在(Yu et al. 2017)中，在尝试了各种添加深监督的方法均未成功后，作者们认为对较低层网络层使用深监督是没有必要的。如图 2(b) 所示，我们提出了一个信息转换器单元，可以将骨干网络的特征转换为不同的表示形式，分别用于训练类别无关的边缘或语义边缘。如果没有信息转换器，那么底层卷积特征 (第一侧端 ∼ 第四侧端) 和高层卷积特征 (第五侧端) 将分别针对类别无关的边缘和语义边缘进行优化，这很难通过第四侧端和第五侧端之间的简单卷积来转换。通过引入信息转换器单元，可以针对不同目标对单个骨干网络的表征进行有效地端对端地学习。DDS 的一个例子如图 1 所示。神经网络的低层可以帮助第五侧端获得细节信息，从而使最终融合的语义边缘（图 1(i)）比第五侧端（图 1(h)）的要更加平滑。

总而言之，我们的主要贡献是：

– 分析 SED 领域中的分散监督悖论，以及为什么它阻碍了先进的 SED 方法 (Yu et al. 2017)使用深监督来提高结果（第 3 节）；
– 提出一种新的基于不同深监督（DDS）的 SED 方法，该方法使用信息转换器来避免在分散监督下学习强大的骨干特征所固有的困难（第 4 节）；
– 提供详细的消融实验来进一步理解所提出的方法（第 5.2 节）。

我们在 SBD、Cityscapes 和 PASCAL VOC2012 数据集上对所提出的方法进行了广泛的评测，我们的方法达到了新的最佳性能。在 Cityscapes 数据集上，所提出的 DDS 算法在*最佳数据集尺度*（Optimal Dataset Scale, ODS）下的 *类别平均的最大 F-measure*（Mean Maximum F-measure）为 79.3%，而之前的最佳性能仅为 75.0% (Yu et al. 2018)。

2 相关工作

对有关此主题的大量文献进行详尽的综述超出了本文的范围。相反，我们首先总结了解决传统的类别无关的边缘检测问题的最重要的研究方向，然后讨论了基于深度学习的方法、语义边缘检测（SED）和深监督技术。

传统的类别无关的边缘检测. 传统的边缘检测通过设计各种滤波器（例如，Sobel (Sobel 1970) 和 Canny (Canny 1986)）或复杂的模型 (Mafi et al. 2018; Shui

& Wang 2017)来检测局部邻域中具有最高梯度的像素 (Hardie & Boncelet 1995; Henstock & Chelberg 1996; Trahanias & Venetsanopoulos 1993)。据我们所知，Konishi 等人(2003)提出了首个数据驱动的边缘检测器，与之前的基于模型的方法不同，该检测器将边缘检测看作统计推理。(Martin et al. 2004)中使用了由亮度、颜色和纹理组成的 Pb 特征，以获得每个边界点的后验概率。通过从多尺度计算局部信息并使用频谱聚类将它们全局化，(Arbeláez et al. 2011)将 Pb 特征进一步扩展为 gPb 特征。(Lim et al. 2013)从手绘草图中学习 Sketch Token 特征来进行轮廓检测，而(Dollár & Zitnick 2015)中采用随机决策森林来学习边缘小块的局部结构，达到了非深度学习方法中的较高的结果。

基于深度学习的类别无关的边缘检测. 最近，机器学习在许多计算机视觉任务中取得了前所未有的成就。其根本思想是深度学习，它利用具有许多隐藏层的神经网络来从原始数据中学习复杂的特征表示 (Chan et al. 2015; Liu et al. 2018; Tang et al. 2017; ?)。因此，基于深度学习的方法也被广泛应用于边缘检测 (Deng et al. 2018; Wang et al. 2019; Yang et al. 2017)。Ganin 等人(2014)通过使用字典学习和最近邻算法将深度神经网络应用于边缘检测。DeepEdge (Bertasius et al. 2015a)首先提取候选轮廓点，然后对这些候选点进行分类。与 DeepEdge 中使用的 Canny (Canny 1986)不同，HFL (Bertasius et al. 2015b)使用 SE (Dollár & Zitnick 2015)生成候选边缘点。与必须为每个候选点处理输入小块的 DeepEdge 相比，HFL 仅需把图像输入到网络一次，因此在计算上更加灵活。DeepContour (Shen et al. 2015)将边缘数据划分为多个子类，并使用不同的模型参数来拟合每个子类。Xie 等人(2015; 2017) 利用深监督网络构建了一个用于“图像-图像”预测的全卷积神经网络。他们所提出的深度模型（称为 HED）融合了来自低层和高层的卷积层信息。Kokkinos (2016)提出了一些重新训练 HED 的训练策略。Liu 等人(2019; 2017) 提出了首个实时的边缘检测器，并在著名的 BSDS500 数据集 (Arbeláez et al. 2011)上取得了比人眼更高的 F-measure。

语义边缘检测. 凭借其强大的语义表征学习能力，基于深度神经网络的边缘检测器倾向于在物体边界位置产生高响应（例如图 1(d)-(g)）。这启发了同时检测边缘像素并分类（将其与一个或多个物体类别相关联）的研究。这种所谓的类别感知的边缘检测对包括物体识别、立体视觉、语义分割和拟物性采样在内的多种视觉任务非常有益。

Hariharan 等人(2011)率先提出了一种将通用对象检测器与自下而上的轮廓组合以检测语义边缘的方法。Yang 等人(2016)提出了一种用于物体轮廓检测的全卷积的编码-解码网络。HFL (Bertasius et al. 2015b)生成与类别无关的二值边缘，并使用深度语义分割网络将类别标签分配给所有边界点。Maninis 等人(2017) 将其卷积的定向的边界（Convolutional Oriented Boundaries, COB）与由空洞卷积(Yu & Koltun 2016)生成的语义分割相结合来获得语义边缘。(Khoreva et al. 2016)引入了一种弱监督学习策略，这种策略无需任何特定于物体的标注，仅用边界框标注就足以产生高质量的物体边界。

Yu 等人(2017)提出了一种新的网络 CASENet，它将 SED 的性能推向了新的高度。在其体系结构中，底层特征仅用于增强高层的分类。经过几次失败的实验后，他们认为，对于 SED，在较低的侧端进行深监督是没有必要的。最近，Yu 等人(2018)提出了一种新的训练方法 SEAL，来训练 CASENet (Yu et al. 2017)。这种方法可以同时对齐真值图边缘并学习语义边缘检测器。但是，由于巨大的 CPU 计算荷载，SEAL 的训练非常耗时。比如，尽管我们使用了功能强大的 CPU （Intel Xeon(R) CPU E5-2683 v3 @ 2.00GHz × 56），在 SBD 数据集上训练 CASENet 仍需要 16 天。Hu 等人(2019)提出了一种新的动态特征融合（Dynamic Feature Fusion, DFF）策略，可以在多尺度的卷积神经网络特征的融合中自适应地为不同的输入图像和位置分配不同的融合权重。Acuna 等人(2019)专注于语义细边缘的对齐学习（Semantic Thinning Edge Alignment Learning, STEAL)。他们提出了一个简单的新层以及损失函数来训练 CASENet (Yu et al. 2017)，使他们可以学习清晰而精确的语义边界。但是，以上所有方法都具有与 CASENet 相同的分散监督悖论。在这个工作中，我们旨在解决 SED 的深度卷积神经网络设计中存在的这一悖论，所以我们的方法与以前的方法兼容，如 SEAL (Yu et al. 2018)、DFF (Hu et al. 2019)和 STEAL (Acuna et al. 2019)等。

深监督. 事实证明，深监督在许多计算机视觉和学习任务中都是有效的，如图像分类(Lee et al. 2015;

Szegedy et al. 2015)、物体检测(Lin et al. 2017, 2020; Liu et al. 2016)、视觉跟踪(Wang et al. 2015)、类别无关的边缘检测 (Liu et al. 2017; Xie & Tu 2017)、显著性物体检测(Hou et al. 2019)等。从理论上讲，深度网络的较低层可以学习有判别力的特征，使得在较高层进行的分类/回归会更加容易。实际中，可以使用深监督来显式地影响隐藏层权重/过滤器的更新过程，使其更偏向于高区分度的特征图。但是，由于上述所提到的分散监督，对于 SED 任务，直接在低层添加类别无关的边缘的深监督可能不是最优的。在下面几节，我们将先分析 SED 中的分散监督，然后介绍一种具有成功的不同深监督的新型语义边缘检测器。

3 SED 中的分散监督悖论

在阐述所提出的方法之前，我们首先分析基于深度学习的 SED 中的分散监督悖论。

3.1 SED 的典型深度模型

为了介绍之前在 SED 中使用深监督的尝试，不失一般性地，我们以一个典型的深度模型 CASENet (Yu et al. 2017)为例。如图 2(a) 所示，这个典型的模型是基于著名的骨干网络 ResNet (He et al. 2016)的。它在第一侧端 ∼ 第三侧端中的每一侧端后都连接一个 1×1 卷积层，来生成单通道特征图 $F^{(m)}$。顶部的第五侧端也连接了一个 1×1 卷积层，来输出具有 K 个通道的类激活图 $A^{(5)} = \{A_1^{(5)}, A_2^{(5)}, \cdots, A_K^{(5)}\}$，其中 K 为类别数。然后进行共享拼接（Shared Concatenation)，即将底部特征 $F^{(m)}$ 进行复制后分别与类激活图的每一个通道进行拼接：

$$A^f = \{F^{(1)}, F^{(2)}, F^{(3)}, A_1^{(5)}, \cdots, F^{(1)}, F^{(2)}, F^{(3)}, A_K^{(5)}\}. \tag{1}$$

接下来，在 A^f 上用一个具有 K 组的 1×1 分组卷积来生成带有 K 个通道的语义边缘图，其中第 k 个通道代表第 k 个类别的边缘图。其他 SED 模型(Hu et al. 2019; Yu et al. 2018) 都具有类似的网络设计。

3.2 讨论

以前的 SED 模型 (Bertasius et al. 2015b; Hu et al. 2019; Yu et al. 2017, 2018) 仅对第五侧端和最终的融合激活进行监督。在(Yu et al. 2017)中，作者尝试了几种深监督的体系结构。他们首先将所有的第一侧端 ∼ 第五侧端都单独用于 SED，每一侧端都连接一个分类损失函数。然而，评测结果甚至比在第五侧端上直接应用 1×1 卷积获得语义边缘的基本体系结构还要差。众所周知，神经网络的较低层包含底层、语义较少的特征（例如局部边缘），这些特征不适合进行语义分类，因为语义类别识别需要抽象的高层特征，这些高层特征出现在神经网络的顶层。所以，他们在底部会获得较差的分类结果。毫不奇怪，对于 SED 任务，仅仅简单地在每个低层和顶层后连接一个分类损失函数和深监督将会导致性能的明显下降。

Yu 等人(2017)也曾尝试对 CASENet 中第一侧端 ∼ 第三侧端添加二值边缘的深监督，但发现与第五侧端的语义分类存在分歧。在顶部的语义边缘的监督下，网络的顶层会学习抽象的高层语义，这些语义信息可以概括关于物体类别的不同外观变化。由于深度卷积神经网络的顶层具有表征能力，而低层是顶层的基础，所以低层通过反向传播的方式被监督以服务顶层去获得高层语义。相反，对低层使用类别无关的边缘进行监督的话，那么低层就被训练来专注于边缘和非边缘之间的区别，而不是用于语义分类的视觉表示。这将会导致低层中的冲突，因此无法为参数更新提供有判别性的梯度信号。

值得注意的是，CASENet 中没有使用第四侧端。我们认为，通过将整个 *res4* 模块作为低层和顶层之间的缓冲单元是一个直观的用于缓解信息冲突的方法。实际上，当将第四侧端添加到 CASENet（详见第 5.2 节）时，得到的新模型（*CASENet+S4*）的 F-measure 为 70.9%，而原始的 CASENet 模型为 71.4%。这刚好证实了我们关于 *res4* 模块的缓冲作用的假设。此外，每个侧端之后都连接一个经典的 1×1 卷积层 (Xie & Tu 2017; Yu et al. 2017)太简单而无法缓冲冲突。因此，我们提出了信息转换器单元，以解决分散监督的冲突。

4 方法

直观上，通过对低层和顶层使用不同但“适当的”的真值图监督，学习到的不同层次的中间表示可能会包含互补信息，但是，直接添加深监督却似乎没有作用。在本节中，我们提出了一个新的网络架构，用于 SED 低层和顶层互补信息的学习。

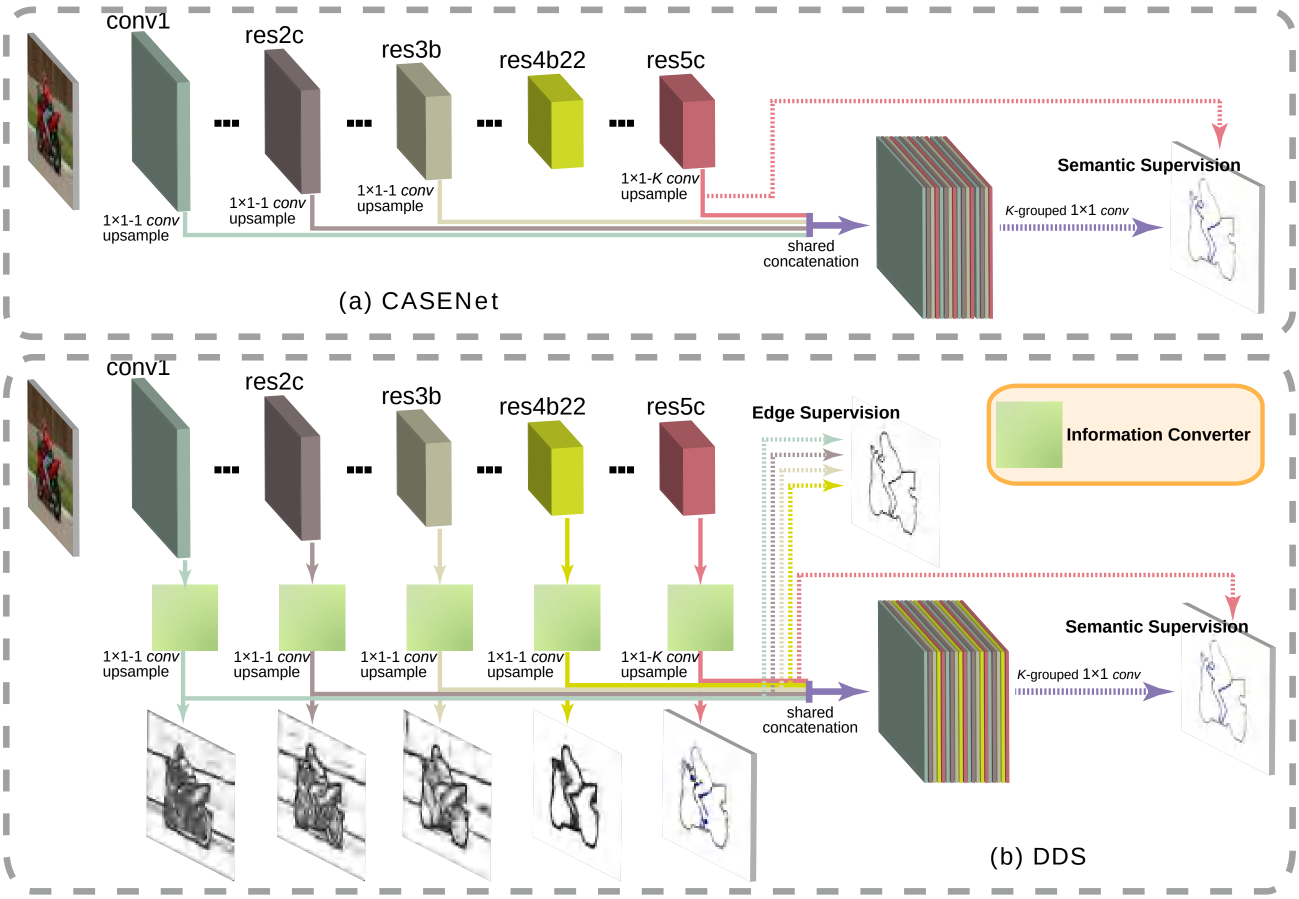


图 2 两种 SED 模型之间的比较：CASENet (Yu et al. 2017)和所提出的 DDS。CASENet 仅对第五侧端的激活添加了监督，并且认为在其体系结构中不需要深监督。但是，我们提出的 DDS 网络为所有侧端的激活都添加了深监督。这里需要注意，信息转换器对于避免类别无关的和语义的边缘检测之间的分散监督至关重要。

4.1 所提出的 DDS 算法

基于以上讨论，我们假设神经网络的低层可能并不直接对 SED 有利。但是，我们仍然认为低层编码了与顶层（第五侧端）互补的细节信息。我们相信通过对网络架构进行适当的重新设计，低层可以用于类别无关的边缘检测来提高由顶层生成的语义边缘的定位精度。为此，我们设计了一种新的信息转换器，以帮助底层特征的学习并从较高层生成一致的梯度信号。这是至关重要的，因为它们可以直接影响隐藏层权重/过滤器的更新过程，从而为 SED 生成具有较高区分性的特征图。

图 2(b) 中展示了我们所提出的网络体系结构。我们遵循 CASENet 来使用 ResNet (He et al. 2016)作为骨干网络。在第一侧端 ∼ 第四侧端中的每个信息转换器（第 4.2 节）之后，我们都连接一个具有单个输出通道的 1×1 卷积层来产生边缘响应图，并使用双线性插值将这些预测图上采样到原图像大小。这些侧输出由与类别无关的二值边缘来监督。我们在第五侧端上使用一个具有 K 个通道的 1×1 卷积来获得语义边缘，其中每个通道代表一个类别的二值边缘图。我们对其使用与第一侧端 ∼ 第四侧端相同的上采样操作。语义边缘被用于监督第五侧端的训练。

我们将从第一侧端 ∼ 第四侧端生成的二值边缘图表示为 $E=\{E^{(1)},E^{(2)},E^{(3)},E^{(4)}\}$。从第五侧端生成的语义边缘图仍然由 $A^{(5)}$ 表示。然后使用共享拼接来获得堆叠的边缘激活图：

$$E^f=\{E,A_1^{(5)},E,A_2^{(5)},E,A_3^{(5)},\cdots,E,A_K^{(5)}\}. \tag{2}$$

注意，E^f 是堆叠的边缘激活图，而在 CASENet 中的 A^f 是堆叠的特征图。最后，我们在 E^f 上连接一个具有 K 组的 1×1 分组卷积来生成融合的语义边缘，并由语义边缘的真值图进行监督。正如 HED(Xie & Tu 2017)中所示，1×1 卷积很好地融合了低层和顶层的边缘。

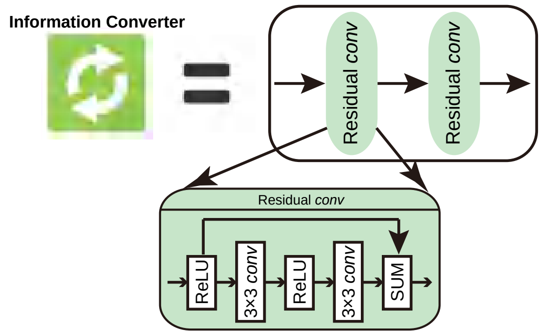


图 3 所提出的信息转换器单元的示意图（图 2 中的橙色框）。

4.2 信息转换器

通过以上分析，提升 SED 的关键是信息转换器的存在。在本文中，我们尝试设计一个简单的信息转换器来验证我们的假设。最近，残差网络已被证明比直连的网络更易于优化(He et al. 2016)，残差学习操作通常是通过短连接和逐元素相加来实现的。我们在图 3 中描述了一个残差卷积模块，它由四个交替连接的 ReLU 和卷积层组成，并且第一个 ReLU 层的输出被加到最后一个卷积层的输出中。我们提出的信息转换器结合了两个残差模块，并连接到 DDS 网络的每一个侧端上，以此将学习到的表征转换为更合适的形式。该操作有望避免由不同损失函数的差异性所带来的冲突。

顶层的语义边缘的监督将产生用于学习语义特征的梯度信号，而低层的类别无关的边缘的监督将产生类别无关的梯度信号。正如第 3 节所说的，如果直接使用分散监督，那么通过反向传播，这些相互矛盾的梯度信号会使骨干网络混乱。所提出的信息转换器可以通过将这些冲突信号转换为适当的表示形式来起到缓冲作用，因此这些转换器可以避免将梯度从侧端的监督引向骨干网络。通过这种方式，骨干网络将接收到一致的更新信号，并朝着相同的目标进行优化；此外，低层和顶层的不同任务将由信息转换器执行。需要注意的是，本文主要阐述信息转换器存在的重要性，而不是其特定形式，所以我们仅采用了一个简单的设计。在实验部分，我们将证明信息转换器的不同设计取得了相似的性能。

我们提出的网络可以成功地结合低层的细节信息和顶层的语义信息。我们的实验结果表明，该算法解决了由不同深监督引发的冲突问题。与 CASENet 不同，我们可以很好地优化第五侧端上的语义分类而没有任何分歧。由于从低层得到的二值边缘有助于第五侧端弥补细节，因此，最终融合的语义边缘可以实现更好的定位质量。

我们使用单像素宽度的二值边缘来监督第一侧端 ∼ 第四侧端，使用较粗的语义边界来监督第五侧端和最终融合的边缘。如果一个像素属于任一类别的语义边界，就将其视为二值边缘。和 CASENet(Yu et al. 2017)一样，我们通过在真值的语义分割中寻找一个像素与其相邻像素之间的差异来获得较粗的语义边界。如果至少存在一个带有标签 k'（$k' \neq k$））的邻接像素，则带有标签 k 的像素被视为类别 k 的边界。

4.3 多任务损失函数

在我们的多任务学习框架中使用了两种不同的损失函数，即类别无关的边缘检测损失和类别感知的边缘检测损失。我们使用 W 表示网络中的所有层的参数。假设图像 I 具有相应的二值边缘图 $Y = \{y_i : i = 1, 2, \cdots, |I|\}$。第一侧端 ∼ 第四侧端的加权的 Sigmoid 交叉熵损失函数可以表示为

$$\begin{aligned} L^{(m)}_{side}(W) = -\sum_{i \in I} [\beta \cdot (1 - y_i) \cdot log(1 - P(E^{(m)}_i; W)) \\ + (1 - \beta) \cdot y_i \cdot log(P(E^{(m)}_i; W))], \\ (m = 1, \cdots, 4), \end{aligned} \tag{3}$$

其中，$\beta = |Y^+|/|Y|$ 且 $1 - \beta = |Y^-|/|Y|$。Y^+ 和 Y^- 分别表示边缘和非边缘的真值标签集。$E^{(m)}_i$ 是在第 m 侧端和像素 i 处产生的激活值。$P(\cdot)$ 是标准的 Sigmoid 函数。

对于图像 I，假设其语义边缘的真值标签集为 $\{\bar{Y}^1, \bar{Y}^2, \cdots, \bar{Y}^K\}$，其中 $\bar{Y}^k = \{\bar{y}^k_i : i = 1, 2, \cdots, |I|\}$ 是第 k 个类别的二值边缘图。需要注意的是，每个像素可以属于多个类别的边界。我们将第五侧端的加权的多标签损失函数定义为

$$\begin{aligned} L^{(5)}_{side}(W) = -\sum_{k}\sum_{i \in I} [\beta \cdot (1 - \bar{y}^k_i) \cdot log(1 - P(A^{(5)}_{k,i}; W)) \\ + (1 - \beta) \cdot \bar{y}^k_i \cdot log(P(A^{(5)}_{k,i}; W))], \end{aligned} \tag{4}$$

其中，$A^{(5)}_{k,i}$ 是第五侧端在像素 i 处属于第 k 类的激活值。类似地，融合的语义激活图的损失函数 $L_{fuse}(W)$

可以定义为

$$L_{fuse}(W) = -\sum_{k}\sum_{i\in I}[\beta \cdot (1-\bar{y}_i^k) \cdot log(1-P(E_{k,i}^f;W)) + (1-\beta) \cdot \bar{y}_i^k \cdot log(P(E_{k,i}^f;W))], \tag{5}$$

其中，E^f 是公式 (2) 中的融合的激活图。最终，总的损失函数被定义为

$$L(W) = \sum_{m=1,\cdots,5} L_{side}^{(m)}(W) + L_{fuse}(W). \tag{6}$$

通过此损失函数,我们可以使用随机梯度下降(Stochastic Gradient Descent, SGD)来优化所有参数。我们将使用加权的损失 $L(W)$ 训练的 DDS 表示为 DDS-R。

最近，Yu 等人(2018)提出了同时对齐和学习语义边缘。他们发现，使用他们的对齐训练策略，未加权的（常规的）Sigmoid 交叉熵损失的表现优于加权的损失。由于 CPU 上的计算量很大，他们的方法在训练网络时非常耗时（对于 SBD 数据集(Hariharan et al. 2011)，使用 28 核的 CPU 和一个 NVIDIA TITAN Xp GPU 训练，要花费 16 天以上的时间）。在训练之前，我们使用他们的方法（SEAL）仅将真值边缘对齐一次，然后使用未加权的 Sigmoid 交叉熵损失来训练对齐过的边缘。第一侧端 $\sim$ 第四侧端的损失函数可以表示为

$$L'^{(m)}_{side}(W) = -\sum_{i\in I}[(1-y_i) \cdot log(1-P(E_i^{(m)};W)) + y_i \cdot log(P(E_i^{(m)};W))], \quad (m=1,\cdots,4). \tag{7}$$

第五侧端的未加权的多标签损失函数为

$$L'^{(5)}_{side}(W) = -\sum_{k}\sum_{i\in I}[(1-\bar{y}_i^k) \cdot log(1-P(A_{k,i}^{(5)};W)) + \bar{y}_i^k \cdot log(P(A_{k,i}^{(5)};W))]. \tag{8}$$

类似地，$L'_{fuse}(W)$ 可以定义为

$$L'_{fuse}(W) = -\sum_{k}\sum_{i\in I}[(1-\bar{y}_i^k) \cdot log(1-P(E_{k,i}^f;W)) + \bar{y}_i^k \cdot log(P(E_{k,i}^f;W))]. \tag{9}$$

总损失是各侧端损失的总和：

$$L'(W) = \sum_{m=1,\cdots,5} L'^{(m)}_{side}(W) + L'_{fuse}(W). \tag{10}$$

我们将使用未加权的损失 $L'(W)$ 训练的 DDS 表示为 DDS-U。

4.4 实现细节

我们使用著名的深度学习框架 Caffe (Jia et al. 2014)来实现我们的算法。所提出的网络基于 ResNet (He et al. 2016)。我们遵循 CASENet (Yu et al. 2017)将第一和第五卷积块的步幅从 2 改为 1，并使用空洞算法来使感受野与原 ResNet 保持相同。我们还遵循 CASENet 将卷积块在 COCO 数据集(Lin et al. 2014) 上进行预训练。该网络通过随机梯度下降（SGD）进行了优化。每个 SGD 迭代均随机地选择 10 张图像，并从每个图像中裁剪出 352×352 的小块。权重衰减（Weight Decay）和动量（Momentum）分别设置为 0.0005 和 0.9。我们使用“poly”学习率策略，即当前学习率等于初始学习率乘以 $(1-curr\_iter/max\_iter)^{power}$，其中，$power$ 设置为 0.9。在 SBD (Hariharan et al. 2011)和 Cityscapes (Cordts et al. 2016)数据集上，我们分别运行 25k/80k 次 SGD 迭代 (max_iter)。对于 DDS-R 的训练，在 SBD 和 Cityscapes 数据集上的初始学习率分别设置为 5e-7/2.5e-7。对于 DDS-U 的训练，训练开始时的损失会非常大。因此，对于 SBD 和 Cityscapes，我们首先以 1e-8 的固定学习率对网络进行 3k 次迭代预训练，然后使用 1e-7 的初始学习率以上述的相同设置继续训练。我们使用在 SBD 上训练的模型来测试 PASCAL VOC2012 数据集，而无需重新训练。侧端的上采样操作使用反卷积层来实现，反卷积层的参数被固定为双线性插值核。所有实验均使用一块 NVIDIA TITAN Xp GPU 进行。

5 实验

5.1 实验设置

数据集. 我们分别在 SBD (Hariharan et al. 2011)、Cityscapes (Cordts et al. 2016)和 PASCAL VOC2012 (Everingham et al. 2012) 数据集上评测我们的方法。SBD 数据集(Hariharan et al. 2011)包含 11355 张图像和相应的具有 Pascal VOC 20 类标签的语义边缘图。它被划分为 8498 张训练图像和 2857 张测试图像。像(Yu et al. 2017)一样，我们使用训练集来训练所提出的网络并使用测试集进行评测。Cityscapes 数

表 1 当使用(Hariharan et al. 2011)中的原始基准时，DDS-R/DDS-U 以及消融实验在 SBD 数据集(Hariharan et al. 2011)上的 ODS F-measure（%）。每列的最佳性能以粗体突出显示。

Methods	aer.	bike	bird	boat	bot.	bus	car	cat	cha.	cow	tab.	dog	hor.	mot.	per.	pot.	she.	sofa	train	tv	mean
Softmax	74.0	64.1	64.8	52.5	52.1	73.2	68.1	73.2	43.1	56.2	37.3	67.4	68.4	67.6	76.7	42.7	64.3	37.5	64.6	56.3	60.2
Basic	82.5	74.2	80.2	62.3	68.0	80.8	74.3	82.9	52.9	73.1	46.1	79.6	78.9	76.0	80.4	52.4	75.4	48.6	75.8	68.0	70.6
DSN	81.6	75.6	78.4	61.3	67.6	82.3	74.6	82.6	52.4	71.9	45.9	79.2	78.3	76.2	80.1	51.9	74.9	48.0	76.5	66.8	70.3
CASENet+S4	84.1	76.4	80.7	63.7	70.3	81.3	73.4	79.4	56.9	70.7	47.6	77.5	81.0	74.5	79.9	54.5	74.8	48.3	72.6	69.4	70.9
DDS\Convt	83.3	77.1	81.7	63.6	70.6	81.2	73.9	79.5	56.8	71.9	48.0	78.3	81.2	75.2	79.7	54.3	76.8	48.9	75.1	68.7	71.3
DDS\DeSup	82.5	77.4	81.5	62.4	70.8	81.6	73.8	80.5	56.9	72.4	46.6	77.9	80.1	73.4	79.9	54.8	76.6	47.5	73.3	67.8	70.9
CASENet	83.3	76.0	80.7	63.4	69.2	81.3	74.9	**83.2**	54.3	74.8	46.4	80.3	80.2	76.6	80.8	53.3	77.2	50.1	75.9	66.8	71.4
DDS-R	85.4	78.3	83.3	65.6	71.4	83.0	75.5	81.3	59.1	75.7	50.7	80.2	82.7	77.0	81.6	58.2	79.5	50.2	76.5	71.2	73.3
DDS-U	**87.2**	**79.7**	**84.7**	**68.3**	**73.0**	**83.7**	**76.7**	82.3	**60.4**	**79.4**	**50.9**	**81.2**	**83.6**	**78.3**	**82.0**	**60.1**	**82.7**	**51.2**	**78.0**	**72.7**	**74.8**

表 2 在 SBD 数据集(Hariharan et al. 2011)上，关于信息转换器设计的消融研究。结果是使用(Hariharan et al. 2011)中的原始基准的 ODS F-measure（%）。每列的最佳性能以粗体突出显示。

Methods	aer.	bike	bird	boat	bot.	bus	car	cat	cha.	cow	tab.	dog	hor.	mot.	per.	pot.	she.	sofa	train	tv	mean
1 *conv* unit	85.2	78.1	82.8	**66.0**	**71.8**	83.2	**75.6**	80.9	58.7	75.5	49.8	79.9	82.4	76.6	81.2	57.5	79.2	49.9	76.2	71.2	73.1
3 *conv* unit	**85.8**	78.7	83.5	**66.0**	**71.8**	**83.6**	75.4	**81.4**	58.9	**76.9**	49.5	**80.4**	**83.0**	76.7	**81.7**	**58.3**	**80.2**	**51.3**	76.0	**71.5**	**73.5**
w/o residual	85.3	**79.0**	**83.7**	65.5	70.9	**83.6**	75.2	81.1	58.6	75.5	49.9	79.3	82.3	76.8	81.3	57.7	79.3	50.6	**76.6**	70.9	73.1
DDS-R	85.4	78.3	83.3	65.6	71.4	83.0	75.5	81.3	**59.1**	75.7	**50.7**	80.2	82.7	**77.0**	81.6	58.2	79.5	50.2	76.5	71.2	73.3

表 3 SBD 数据集上(Hariharan et al. 2011) 与类别无关的评测结果。结果是使用(Hariharan et al. 2011)中原始基准的 ODS F-measure（%）。

Methods	DSN	CASENet	**DDS-R**
ODS	76.6	76.4	**79.3**

据集(Cordts et al. 2016)是一个大规模的语义分割数据集，包含来自 50 个不同城市的街道场景中记录的立体视频序列。它由 5000 张图像组成，分为 2975 张训练图像、500 张验证图像和 1525 张测试图像。由于测试集是一个针对语义分割和场景理解的在线竞赛，所以它的真值图尚未发布。因此，我们使用训练集进行训练，使用验证集进行测试。语义分割数据集 PASCAL VOC2012 (Everingham et al. 2012) 由 1464 张训练、1449 张验证和 1456 张测试图像组成，这些图像与 SBD 数据集具有相同的 20 个类别。由于与 Cityscapes 相同的原因，测试集的语义标签尚未发布。我们生成一个移除了 SBD 训练图像的新验证集，从而产生 904 个验证图像。我们使用这个新验证集和在 SBD 训练集上训练的模型来测试各种方法的泛化性。

评测指标. 为了进行性能评测，我们采用了几种标准的基准并使用它们原论文中推荐的参数设置。我们首先使用(Hariharan et al. 2011)中的基准来评测每个类别的准确率-召回率曲线。对于所有数据集，我们都使用匹配距离公差为 0.02 的默认设置。我们报告每个类别在最佳数据集尺度（ODS）下的最大 F-measure（F_m）和所有类别的平均最大 F-measure。

然后，我们遵循(Yu et al. 2018)来使用比(Hariharan et al. 2011)中的基准更严格的规则，以评测语义边缘。(Yu et al. 2018)的真值图是实例敏感的边缘，这与(Hariharan et al. 2011)中使用实例不敏感的边缘不同。此外，(Hariharan et al. 2011)默认在匹配之前将预测的边缘变细。(Yu et al. 2018)进一步提出将原始预测与未变细的真值图进行匹配，这种模式和以前的常规模式分别称为“Raw”和“Thin”。在本文中，我们为(Yu et al. 2018)中的基准计算了“Thin”和“Raw”的结果。我们遵循(Yu et al. 2018)为原始的 SBD 数据集(Hariharan et al. 2011)设置匹配距离公差为 0.02，为重新标注的 SBD 数据集(Yu et al. 2018)设置匹配距离公差为 0.0075，为 Cityscapes 数据集(Cordts et al. 2016)设置为 0.0035，VOC2012 数据集(Everingham et al. 2012)则为 0.02。对于 SBD 和 VOC2012 数据集，将忽略 5 个像素宽度的图像边界；而对于 Cityscapes 数据集，则不这么做。

我们遵循(Yu et al. 2018)为实例敏感和实例不敏感的边缘生成“Thin”和“Raw”的真值图。所产生的边缘可被看成语义分割中的语义物体或介质的边界。像之前的研究 (Acuna et al. 2019; Hu et al. 2019; Yu et al. 2017, 2018) 一样，我们将 Cityscapes 数据集的真值图和预测的边缘图缩小到原尺寸的一半以加快评测速度。

对之前的方法的评测. 为了与基准方法进行性能比较，我们使用作者发布的默认代码和预训练模型来生成边缘预测。请注意，根据(Hariharan et al. 2011)中的指标，我们在 Cityscapes 数据集上得到

了与 CASENet (Yu et al. 2017) 不同的评测结果（更好），这是因为 CASENet 使用了不正确的真值边缘下采样策略，而我们按照 SEAL (Yu et al. 2018)进行下采样。具体来说，CASENet 直接对真值的细边缘进行下采样使其分辨率变成原图像的一半，这将导致不连续的真值边缘。但是，我们根据 SEAL，首先对真值的语义分割图进行下采样，然后从下采样后的分割图中生成边缘。

5.2 消融实验

在将所提出的 DDS 算法与最新的方法进行比较之前，我们先在 SBD 数据集(Yu et al. 2018)上进行消融实验，以从各个角度研究提出的 DDS 算法。为此，我们提出了六个 DDS 变体：

- ***Softmax***，仅使用顶层（第五侧端）并添加具有 21 类的 softmax 损失函数，于是每个类别的真值边缘不会重叠，所以每个像素都有一个特定的类别标签。
- ***Basic***，使用顶层（第五侧端）进行多标签分类，这意味着我们直接在 *res5c* 上连接损失函数 $L^{(5)}_{side}(W)$ 来训练检测器。
- ***DSN***，直接使用深监督的网络体系结构，其中骨干网络的每一侧端都连接一个具有 K 个输出通道的 1×1 卷积层来实现 SED，并融合了来自各个侧端的激活图以生成最终的语义边缘。
- ***CASENet+S4***，类似于 CASENet，但同时考虑了第四侧端，方法是将其连接到一个 1×1 卷积层来生成一个单通道的特征图，而 CASENet 仅使用第一侧端 ~ 第三侧端和第五侧端。
- ***DDS\Convt***，移除了 DDS 中的信息转换器，从而使得深监督被直接添加在每一侧端之后。
- ***DDS\DeSup***，移除了 DDS 的第一侧端 ~ 第四侧端的深监督，但保留了信息转换器。

所有这些变体都使用加权的损失函数（即公式 (6)，除了 *Softmax*）和原始的 SBD 数据进行训练，以进行公平比较。

我们在 SBD 数据集上使用(Hariharan et al. 2011)中的基准来评测这些变体以及原始的 DDS(Yu et al. 2017)和 CASENet(Hariharan et al. 2011)。评测结果如表 1 所示。可以看到，*Softmax* 的性能明显下降。因为神经网络预测的语义边缘通常很粗且与其他类别重叠，所以为每个像素分配单个标签是不合适的。因此，我们在公式 (4) 和 公式 (5) 中使用多标签损失。*Basic* 网络在 ODS F-measure 上达到了 70.6%，比 *DSN* 高 0.3%。这进一步验证了我们在第 3 节中提出的假设，即底层特征对于语义分类没有足够的判别力。此外，*CASENet+S4* 的性能优于 *DSN*，表明底层卷积特征更适合于二值边缘检测。而且，*CASENet+S4* 的 F-measure 低于原本的 CASENet。

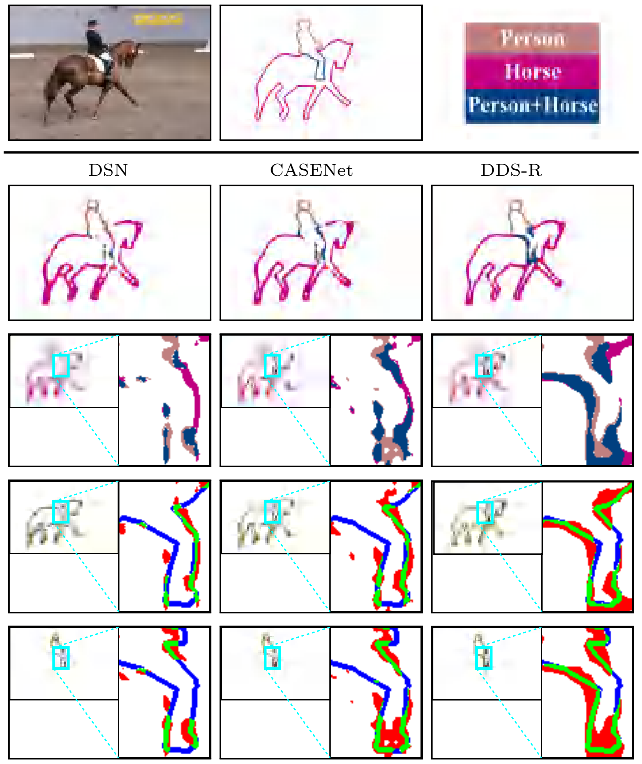


图 4 DSN、CASENet 和 DDS-R 的定性比较。第一行：原始图像、真值图和类别颜色代码。该图像来自 SBD 数据集(Hariharan et al. 2011)。第二行：通过不同方法预测的语义边缘。第三行：预测边缘的一个放大的区域。第四行：预测的马的边界。最后一行：预测的人的边界。绿色，红色，白色和蓝色像素分别代表在阈值 0.5 时的真阳性、假阳性、真阴性和假阴性。

为什么 DDS 效果好? 从 *DDS\DeSup* 到 DDS-R 的改进表明，DDS 的成功并不是因为引入了更多的参数（卷积层），而是因为深监督与信息转换器之间的协调。相反，添加更多卷积层但没有深监督可能会使网络收敛变得更加困难。将 *DDS\Convt* 与 CASENet 的结果进行比较时，我们得出的结论与(Yu et al. 2017)一致，即直接在低层添加二值边缘监督没有任何价值。

表 4 DDS-R/DDS-U 和其他方法在 SBD 数据集(Hariharan et al. 2011) 上的 ODS F-measure（%）。每列的最佳性能以粗体突出显示。

Methods	aer.	bike	bird	boat	bot.	bus	car	cat	cha.	cow	tab.	dog	hor.	mot.	per.	pot.	she.	sofa	train	tv	mean
With the evaluation metric in (Hariharan et al. 2011)																					
InvDet	41.5	46.7	15.6	17.1	36.5	42.6	40.3	22.7	18.9	26.9	12.5	18.2	35.4	29.4	48.2	13.9	26.9	11.1	21.9	31.4	27.9
HFL-FC8	71.6	59.6	68.0	54.1	57.2	68.0	58.8	69.3	43.3	65.8	33.3	67.9	67.5	62.2	69.0	43.8	68.5	33.9	57.7	54.8	58.7
HFL-CRF	73.9	61.4	74.6	57.2	58.8	70.4	61.6	71.9	46.5	72.3	36.2	71.1	73.0	68.1	70.3	44.4	73.2	42.6	62.4	60.1	62.5
BNF	76.7	60.5	75.9	60.7	63.1	68.4	62.0	74.3	54.1	76.0	42.9	71.9	76.1	68.3	70.5	53.7	79.6	**51.9**	60.7	60.9	65.4
WS	65.9	54.1	63.6	47.9	47.0	60.4	50.9	56.5	40.4	56.0	30.0	57.5	58.0	57.4	59.5	39.0	64.2	35.4	51.0	42.4	51.9
DilConv	83.7	71.8	78.8	65.5	66.3	82.6	73.0	77.3	47.3	76.8	37.2	78.4	79.4	75.2	73.8	46.2	79.5	46.6	76.4	63.8	69.0
DSN	81.6	75.6	78.4	61.3	67.6	82.3	74.6	82.6	52.4	71.9	45.9	79.2	78.3	76.2	80.1	51.9	74.9	48.0	76.5	66.8	70.3
COB	84.2	72.3	81.0	64.2	68.8	81.7	71.5	79.4	55.2	79.1	40.8	79.9	80.4	75.6	77.3	54.4	**82.8**	51.7	72.1	62.4	70.7
CASENet	83.3	76.0	80.7	63.4	69.2	81.3	74.9	**83.2**	54.3	74.8	46.4	80.3	80.2	76.6	80.8	53.3	77.2	50.1	75.9	66.8	71.4
SEAL	85.2	77.7	83.4	66.3	70.6	82.4	75.2	82.3	58.5	76.5	50.4	80.9	82.2	76.8	**82.2**	57.1	78.9	50.4	75.8	70.1	73.1
DDS-R	85.4	78.3	83.3	65.6	71.4	83.0	75.5	81.3	59.1	75.7	50.7	80.2	82.7	77.0	81.6	58.2	79.5	50.2	76.5	71.2	73.3
DDS-U	**87.2**	**79.7**	**84.7**	**68.3**	**73.0**	**83.7**	**76.7**	82.3	**60.4**	**79.4**	**50.9**	**81.2**	**83.6**	**78.3**	82.0	**60.1**	82.7	51.2	**78.0**	**72.7**	**74.8**
With the "Thin" evaluation metric in (Yu et al. 2018)																					
CASENet	83.6	75.3	82.3	63.1	70.5	83.5	76.5	82.6	56.8	76.3	47.5	80.8	80.9	75.6	80.7	54.1	77.7	52.3	77.9	68.0	72.3
SEAL	84.5	76.5	83.7	64.9	71.7	83.8	78.1	85.0	58.8	76.6	50.9	82.4	82.2	77.1	83.0	55.1	78.4	54.4	79.3	69.6	73.8
STEAL	85.2	77.3	84.0	65.9	71.1	85.3	77.5	83.8	59.2	76.4	50.0	81.9	82.2	77.3	81.7	55.7	79.5	52.3	79.2	69.8	73.8
DDS-R	85.6	77.1	82.8	64.0	73.5	85.4	78.8	84.4	57.7	77.6	51.9	81.2	82.4	77.1	82.5	56.3	79.5	54.5	80.3	70.4	74.1
DDS-U	86.5	78.4	84.4	67.0	74.3	85.8	80.2	**85.9**	60.4	**80.8**	**53.9**	83.0	84.4	**78.8**	**83.9**	**58.7**	**81.9**	**56.0**	**82.1**	**73.0**	**76.0**
DFF	86.5	79.5	85.5	**69.0**	73.9	86.1	80.3	85.3	58.5	80.1	47.3	82.5	**85.7**	78.5	83.4	57.9	81.2	53.0	81.4	71.6	75.4
DDS-R	**86.7**	**79.6**	**85.6**	68.4	**74.5**	**86.5**	**81.1**	**85.9**	**60.5**	79.3	53.5	**83.2**	85.2	**78.8**	**83.9**	58.4	80.8	54.4	81.8	72.2	**76.0**
With the "Raw" evaluation metric in (Yu et al. 2018)																					
CASENet	71.8	60.2	72.6	49.5	59.3	73.3	65.2	70.8	51.9	64.9	41.2	67.9	72.5	64.1	71.2	44.0	71.7	45.7	65.4	55.8	62.0
SEAL	81.1	69.6	81.7	60.6	68.0	80.5	75.1	80.7	57.0	73.1	48.1	78.2	80.3	72.1	79.8	50.0	78.2	51.8	74.6	65.0	70.3
STEAL	77.2	66.2	78.9	56.8	63.2	77.8	71.9	75.3	55.0	69.4	43.8	73.1	76.9	69.8	75.5	48.3	76.2	47.7	70.4	60.5	66.7
DDS-R	80.5	68.2	78.6	56.4	67.6	80.9	72.7	77.6	55.4	70.9	47.0	74.9	77.5	70.0	77.4	50.9	75.7	50.7	74.5	65.5	68.6
DDS-U	**83.8**	**71.8**	**82.1**	**61.7**	**70.4**	**82.9**	**76.9**	**80.8**	**58.5**	**77.1**	**49.9**	**77.8**	**81.5**	**73.5**	**81.0**	**52.9**	**81.3**	**53.0**	**76.3**	**69.1**	**72.1**
DFF	77.6	65.7	79.3	57.2	65.5	78.5	72.0	76.2	53.7	71.9	42.5	72.0	77.0	68.8	75.1	50.6	76.6	46.9	71.9	63.6	67.1
DDS-R	79.2	67.6	77.7	58.7	65.9	81.0	72.9	76.6	55.8	70.3	47.6	74.0	76.9	68.8	76.5	52.5	77.0	48.8	72.8	65.7	68.3

表 5 DDS-R/DDS-U 和其他方法在重新标注的 SBD 数据集(Hariharan et al. 2011) 上的 ODS F-measure（%）。每列的最佳性能以粗体突出显示。

Methods	aer.	bike	bird	boat	bot.	bus	car	cat	cha.	cow	tab.	dog	hor.	mot.	per.	pot.	she.	sofa	train	tv	mean
With the evaluation metric in (Hariharan et al. 2011)																					
DSN	83.8	73.6	76.0	61.4	69.2	84.2	74.8	82.0	53.5	73.7	45.3	81.9	79.9	73.0	83.5	55.0	77.2	51.9	80.6	66.7	71.4
CASENet	84.8	72.8	77.9	62.6	70.9	83.5	73.4	81.7	54.7	75.6	44.8	82.6	82.0	74.0	83.0	53.5	77.8	51.7	78.7	63.8	71.5
SEAL	85.5	74.9	80.9	64.7	70.4	85.9	76.5	84.3	58.3	74.2	47.7	84.0	82.4	76.1	**85.7**	59.1	80.1	54.0	81.1	67.1	73.7
DDS-R	86.6	76.4	79.7	65.7	72.7	**86.0**	77.3	83.4	58.5	77.5	51.7	83.4	82.6	76.5	84.9	59.6	80.4	55.2	81.5	69.6	74.5
DDS-U	**88.2**	**77.1**	**82.4**	**67.9**	**73.0**	85.6	**79.2**	**85.2**	**60.6**	**80.5**	**53.2**	**84.2**	**84.0**	**77.5**	85.5	**62.9**	**83.2**	**56.8**	**82.4**	**71.7**	**76.1**
With the "Thin" evaluation metric in (Yu et al. 2018)																					
CASENet	74.5	59.7	73.4	48.0	67.1	78.6	67.3	76.2	47.5	69.7	36.2	75.7	72.7	61.3	74.8	42.6	71.8	48.9	71.7	54.9	63.6
SEAL	78.0	65.8	76.6	52.4	68.6	80.0	70.4	79.4	50.0	72.8	41.4	78.1	75.0	65.5	78.5	49.4	73.3	52.2	73.9	58.1	67.0
STEAL	77.1	63.6	76.2	51.1	68.0	80.4	70.0	76.8	49.4	71.9	40.4	78.1	74.7	64.5	75.7	45.4	73.5	47.5	73.5	58.7	65.8
DDS-R	79.7	65.2	74.6	51.8	71.9	81.3	72.5	79.4	49.2	75.1	43.9	77.8	75.3	65.2	78.9	51.1	74.9	54.1	75.1	61.7	67.9
DDS-U	**81.4**	67.6	77.8	**55.7**	70.9	82.0	74.5	**81.2**	**52.1**	76.5	**47.2**	**79.6**	77.3	**68.1**	**80.2**	**53.4**	**78.5**	**56.1**	76.6	63.9	**70.0**
DFF	78.6	66.2	77.9	53.2	**72.3**	81.3	73.3	79.0	50.7	**76.8**	38.7	77.2	**78.6**	65.2	77.9	49.4	76.1	49.7	74.7	62.9	68.0
DDS-R	78.8	**68.0**	**78.3**	55.0	71.9	**82.4**	**74.6**	80.5	52.0	74.0	42.0	78.3	77.1	66.1	78.5	49.3	77.5	49.3	**76.9**	**64.8**	68.8
With the "Raw" evaluation metric in (Yu et al. 2018)																					
CASENet	65.8	51.5	65.0	43.1	57.5	68.1	58.2	66.0	45.4	59.8	32.9	64.2	65.8	52.6	65.7	40.9	65.0	42.9	61.4	47.8	56.0
SEAL	75.3	60.5	75.1	51.2	65.4	76.1	67.9	75.9	49.7	69.5	39.9	**74.8**	72.7	62.1	74.2	48.4	72.3	49.3	70.6	56.7	64.4
STEAL	70.9	55.9	71.6	47.6	61.5	72.6	64.6	70.2	47.5	67.4	37.3	70.6	69.4	59.1	69.2	44.3	69.1	42.6	67.7	53.5	60.6
DDS-R	75.6	61.1	71.0	49.5	67.7	76.1	67.2	74.2	48.8	69.1	40.4	72.5	71.7	60.4	73.4	49.6	70.6	49.5	71.9	59.4	64.0
DDS-U	**78.4**	**62.7**	**75.6**	**53.4**	**67.8**	**78.5**	**71.4**	**77.4**	**51.3**	**72.8**	**44.5**	74.7	**74.8**	**64.3**	**76.3**	**51.9**	**77.3**	**51.9**	**73.7**	**62.9**	**67.1**
DFF	72.3	58.4	73.4	48.7	65.4	74.8	66.4	72.5	47.8	70.1	34.7	69.2	71.5	58.7	70.2	47.5	71.2	43.7	69.5	59.1	62.3
DDS-R	74.2	61.2	71.3	51.9	65.5	77.3	68.0	73.8	50.0	66.0	39.4	70.8	70.5	58.9	71.8	49.0	72.6	44.7	71.6	62.2	63.5

关于所提出的 DDS 的讨论. 直观地，在低层和顶层使用不同但“适当的”的真值图进行监督可以增强不同层中的特征学习。基于此，学习到的不同层的中间表示将倾向于包含补充信息。但是，在我们的情况下，由于损失函数公式 (6) 中的差异性，可能会产生具有较弱判别性的梯度信号，使得直接在低层添加类别无关的边缘的深监督是无效的。相反，我们表明，通过适当的体系结构的重新设计，我们可以利用

深监督来显著地提高性能。所提出的方法中的信息转换器在指导低层进行类别无关边缘检测时起关键作用。这样，来自低层的底层边缘会编码更多的细节，从而有助于高层更好地定位语义边缘。它们还用来从较高层生成一致的梯度信号。这是必不可少的，因为它们可以直接影响隐藏层权重/过滤器的更新过程，从而为实现正确的 SED 产生具有较高区分性的特征图。

我们所提出的 DDS-R/DDS-U 相对于 *CASENet+S4* 和 *DDS\Convt* 的显著性能提升证明了我们的设计的重要性，即在不同侧端的信息格式转换之后使用深监督。我们还注意到，通过使用未加权的损失函数和对齐的边缘(Yu et al. 2018)，DDS-U 的性能比 DDS-R 更好。

关于信息转换器设计的讨论. 本文主要讨论并解决了 SED 中的分散监督悖论，其核心是信息转换器的存在，而不是它的具体形式。因此，我们设计了一个简单的信息转换器，它由两个连续的残差卷积单元组成。这里，我们对这个设计进行了消融实验，结果展示在表 2 中。我们实验了三种不同的转换器设计：i) 仅有 1 个卷积单元；ii) 具有 3 个卷积单元；iii) 卷积单元中没有残差连接的设计（直连的卷积单元）。可以观察到，具有 3 个残差卷积单元的信息转换器获得了最佳性能，但它仅比具有 2 个残差卷积单元的信息转换器要好一点而已。为了在模型复杂度和性能之间进行权衡，我们使用 2 个残差卷积单位作为默认设置。

对边缘定位的改善. 为了证明引入的信息转换器是否确实改善了语义边缘的定位，我们忽略语义标签，并对所提出的 DDS 和以前的方法进行了类别无关的评测。给定一张输入图像，SED 方法为每个类别生成边缘概率图。为了对一张图像生成类别无关的边缘图，在每个像素点上，我们将所有类别中最大的边缘概率视为该像素处的类别无关的边缘的概率。对于真值图的每个像素，如果有任何类别在该像素上具有边缘，则将该像素视为与类别无关的边缘像素。然后，我们使用(Hariharan et al. 2011)中的标准基准进行评测。从表 3 中，我们发现 DDS 可以显著提高边缘定位的准确性，这表明在网络低层实施类别无关的边缘监督非常有利于边缘定位。在探索了 DDS 的多种变体并确定了该方法的有效性之后，我们总结了我们的方法达到的结果，并将其与几种最新的方法进行了比较。

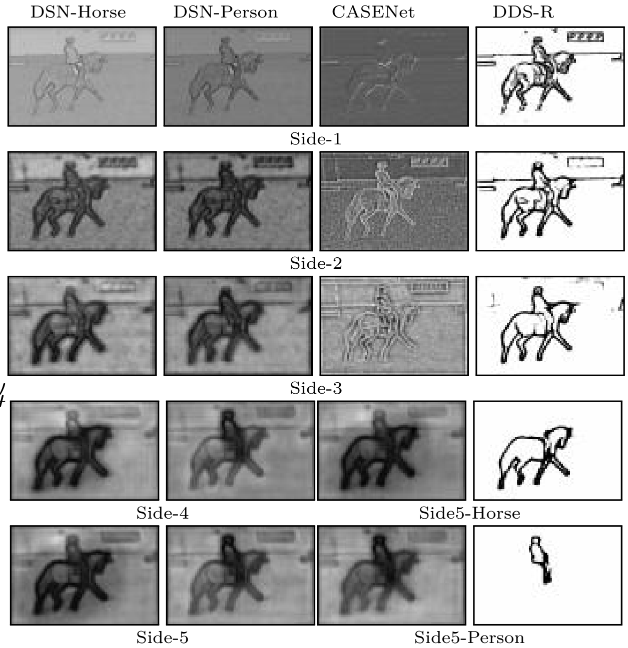


图 5 图 4 中输入图像的侧端激活图。前两列分别展示 DSN 对马和人的侧端分类激活。最后两列分别展示 CASENet 和所提出的 DDS-R 的第一侧端 ~ 第三侧端的特征以及第五侧端的分类激活。这些图像是通过将激活标准化为 $[0, 255]$ 获得的。请注意，所有激活均直接输出而不做任何非线性化（如 Sigmoid 函数）。

表 6 SBD 数据集(Hariharan et al. 2011)上每张图像的平均运行时间。

Methods	DSN	CASENet	SEAL	DDS
Time (s)	0.171	0.166	0.166	0.175

5.3 SBD 上的评测

我们在 SBD 数据集(Hariharan et al. 2011)上，将 DDS-R/DDS-U 与最新的方法进行比较，包括 InvDet (Hariharan et al. 2011)、HFL-FC8 (Bertasius et al. 2015b)、HFL-CRF (Bertasius et al. 2015b)、BNF (Bertasius et al. 2016)、WS (Khoreva et al. 2016)、DilConv (Yu & Koltun 2016)、DSN (Yu et al. 2017)、COB (Maninis et al. 2017)、CASENet (Yu et al. 2017)、SEAL (Yu et al. 2018)、STEAL (Acuna et al. 2019)和 DFF (Hu et al. 2019)。由于 DFF 与 CASENet 具有相同的分散监督悖论，因此我们也将 DDS-R 集成到 DFF 中以证明 DDS-R 的可扩展性。对于基于 DFF 的 DDS-R，我们采用与原始 DFF 相同的代码实现和训练策略。

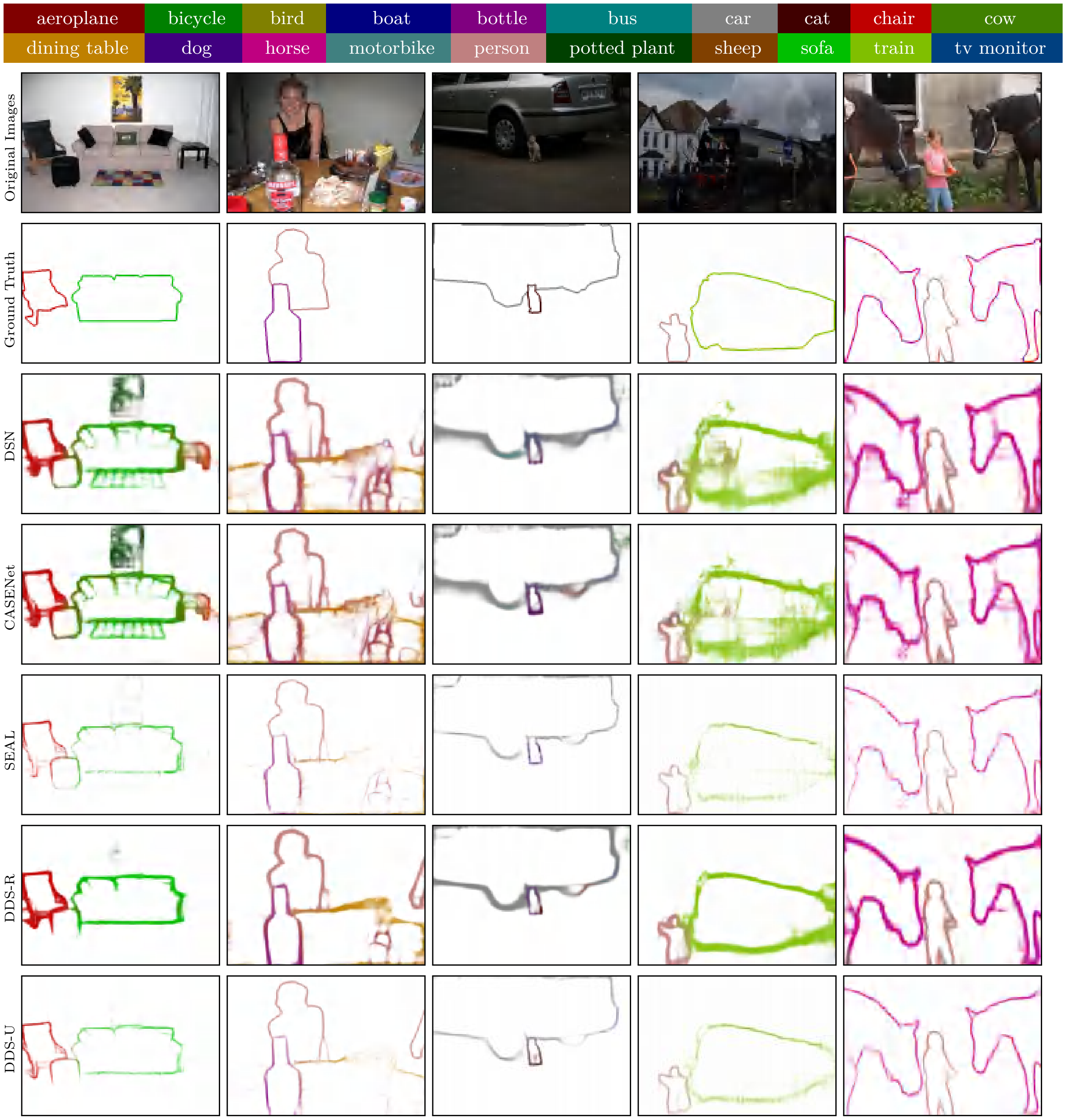


图 6 SBD 数据集(Hariharan et al. 2011)中的一些示例。从上到下依次是：颜色代码、原图像、真值图、DSN、CASENet (Yu et al. 2017)、SEAL (Yu et al. 2018)、所提出的 DDS-R 和 DDS-U。我们遵循(Yu et al. 2018)中的颜色编码协议。

结果如表 4 所示。DDS-U 在所有方法中达到了最佳性能。就(Hariharan et al. 2011)中的指标而言，所提出的 DDS-U 在 ODS F-measure 上比 SEAL 高 1.7%，比 CASENet 高 3.4%，所以，DDS-U 达到了最新的性能。我们还观察到，DDS-R 也可以改善 DFF (Hu et al. 2019)的性能。因此，提出的 DDS 可以看作是提高 SED 的一般性的思路。从 CASENet 到 DDS 的提升也大于 STEAL 的提升。此外，InvDet (Hariharan et al. 2011)是一种基于非深度学习的方法，与其他常规方法相比，它展示出具有竞争力的结果。COB (Maninis et al. 2017)是一种先进的类别无关的边缘检测方法，将其与 DilConv (Yu & Koltun 2016)的语义分割相结合可得到高性能的语义边缘检测器。COB 优于 DilConv 反映了它的融合算法的有效性。CASENet 和 DDS-R/DDS-U 都优于 COB 说明了直接学习语义边缘的重要性，因为将二值边

表 7 DDS-R/DDS-U 和其他方法在 Cityscapes 数据集 (Cordts et al. 2016) 上的 ODS F-measure（%）。每列的最佳性能以粗体突出显示。

Methods	road	sid.	bui.	wall	fen.	pole	light	sign	veg.	ter.	sky	per.	rider	car	tru.	bus	tra.	mot.	bike	mean
With the evaluation metric in (Hariharan et al. 2011)																				
DSN	87.8	82.5	83.2	55.2	57.5	81.4	75.9	78.9	86.6	66.1	82.3	87.9	76.2	91.0	55.4	73.2	53.9	61.6	85.4	74.8
CASENet	87.2	82.2	83.0	53.7	57.9	82.9	78.7	79.2	86.0	65.8	82.7	88.0	77.1	90.3	50.6	72.1	56.1	63.5	85.3	74.9
PSPNet	58.7	79.9	73.0	58.4	59.8	79.3	75.3	75.5	76.7	66.0	70.2	80.1	74.6	84.2	**63.1**	76.6	**70.3**	64.5	76.1	71.7
DeepLabv3+	39.2	32.8	39.5	9.0	7.0	25.2	12.5	19.6	34.6	10.2	23.6	22.7	12.0	22.4	2.3	11.1	9.5	6.0	14.0	18.6
SEAL	88.1	84.5	83.4	55.3	57.2	83.6	78.6	79.7	87.3	69.0	83.5	86.8	77.8	87.2	54.5	73.1	49.0	61.8	85.3	75.0
DDS-R	**90.5**	84.2	86.2	57.7	61.4	85.1	83.8	80.4	88.5	67.6	**88.2**	89.9	80.1	**91.8**	58.6	76.3	56.2	68.8	**87.3**	78.0
DDS-U	90.3	**85.3**	**86.7**	**58.8**	**61.5**	**86.9**	**84.7**	**83.0**	**89.3**	**69.8**	**88.2**	**90.3**	**80.5**	91.7	62.5	**77.4**	61.5	**70.5**	**87.3**	**79.3**
With the “Thin” evaluation metric in (Yu et al. 2018)																				
CASENet	86.2	74.9	74.5	47.6	46.5	72.8	70.0	73.3	79.3	57.0	86.5	80.4	66.8	88.3	49.3	64.6	47.8	55.8	71.9	68.1
SEAL	87.6	77.5	75.9	47.6	46.3	75.5	71.2	75.4	80.9	60.1	87.4	81.5	68.9	88.9	50.2	67.8	44.1	52.7	73.0	69.1
STEAL	87.8	77.2	76.4	49.5	49.2	74.9	73.2	76.3	80.8	58.9	86.8	80.2	69.0	83.2	52.1	67.7	53.2	55.8	72.8	69.7
DDS-R	86.1	76.5	76.1	49.8	49.9	74.6	76.4	76.8	80.4	58.9	87.2	83.5	70.7	89.6	52.9	71.5	50.4	61.8	74.4	70.9
DDS-U	89.2	79.2	79.0	51.9	52.9	77.5	79.4	80.3	82.6	61.4	88.8	85.0	74.1	91.1	59.0	**76.0**	**55.7**	63.6	76.3	73.8
DFF	89.4	**80.1**	79.6	51.3	**54.5**	81.3	81.3	**81.2**	83.6	**62.9**	89.0	85.4	75.8	91.6	54.9	73.9	51.9	64.3	76.4	74.1
DDS-R	**89.7**	79.4	**80.4**	**52.1**	53.0	**82.4**	**81.9**	80.9	**83.9**	62.0	**89.4**	**86.0**	**77.8**	**92.3**	**59.8**	74.8	55.3	**64.4**	**77.4**	**74.9**
With the “Raw” evaluation metric in (Yu et al. 2018)																				
CASENet	66.8	64.6	66.8	39.4	40.6	71.7	64.2	65.1	71.1	50.2	80.3	73.1	58.6	77.0	42.0	53.2	39.1	46.1	62.2	59.6
SEAL	**84.4**	73.5	72.7	**43.4**	**43.2**	76.1	68.5	69.8	77.2	**57.5**	**85.3**	77.6	63.6	84.9	**48.6**	61.9	**41.2**	49.0	66.7	65.5
STEAL	75.8	68.5	69.8	34.9	36.1	73.4	66.7	67.7	73.5	49.7	78.7	72.9	59.1	76.5	35.3	52.8	37.7	43.8	63.7	59.8
DDS-R	73.3	65.9	70.9	33.2	37.4	76.8	70.1	70.2	74.6	50.4	80.6	77.9	62.6	82.5	37.1	55.0	32.0	49.4	66.1	61.4
DDS-U	83.5	**74.2**	76.0	37.5	40.7	79.5	75.6	75.3	**79.3**	55.7	**85.3**	81.1	67.1	**87.9**	44.6	**63.4**	40.4	52.3	70.0	**66.8**
DFF	72.8	68.3	72.6	37.2	42.2	79.6	75.0	73.9	75.3	51.4	80.8	78.6	69.4	83.0	44.1	56.7	38.4	52.0	68.8	64.2
DDS-R	80.8	70.8	**76.4**	38.9	41.1	**80.0**	**78.2**	**76.3**	79.2	53.2	82.5	**81.8**	**72.2**	86.2	44.8	59.5	37.6	**55.7**	**71.3**	66.7

表 8 DDS-R/DDS-U 和其他方法在 VOC2012 数据集(Everingham et al. 2012)上的 ODS F-measure（%）。每列的最佳性能以粗体突出显示。

Methods	aer.	bike	bird	boat	bot.	bus	car	cat	cha.	cow	tab.	dog	hor.	mot.	per.	pot.	she.	sofa	train	tv	mean
With the evaluation metric in (Hariharan et al. 2011)																					
DSN	83.5	**60.5**	81.8	58.0	66.4	82.7	69.9	83.0	49.7	78.6	50.8	78.4	74.7	74.1	82.0	55.0	79.9	55.2	78.3	68.6	70.5
CASENet	84.6	60.1	82.7	59.2	68.1	84.3	69.9	83.5	51.9	81.2	50.4	80.4	76.7	74.4	81.9	55.8	82.0	54.9	77.8	67.0	71.3
SEAL	85.2	60.0	84.4	**61.8**	70.3	85.5	71.7	83.7	53.8	82.1	50.1	**81.4**	76.8	75.4	83.7	59.1	80.9	54.4	78.7	72.2	72.6
DDS-R	86.3	58.2	86.0	60.2	71.6	85.2	72.6	83.0	53.0	82.1	54.0	79.4	77.8	74.9	83.5	57.3	81.7	53.6	79.7	71.0	72.6
DDS-U	**87.1**	60.0	**86.6**	60.8	**72.6**	**87.0**	**73.2**	**85.3**	**56.5**	**83.9**	**55.8**	80.3	**79.6**	**75.9**	**84.5**	**61.7**	**85.1**	**57.0**	**80.5**	**74.0**	**74.4**
With the “Thin” evaluation metric in (Yu et al. 2018)																					
CASENet	80.7	55.0	81.1	57.8	67.7	78.9	67.9	78.5	51.6	76.6	43.9	76.8	74.0	70.0	78.8	54.7	78.7	52.8	75.4	67.4	68.4
SEAL	83.3	**57.5**	82.9	**60.1**	69.2	82.1	69.5	80.5	53.6	78.4	46.8	78.2	76.0	72.1	81.6	57.8	79.1	54.0	76.2	69.0	70.4
STEAL	82.5	54.9	82.7	57.0	70.2	80.3	69.8	80.0	51.6	76.6	42.8	78.0	74.9	71.6	79.3	55.8	78.7	49.2	76.8	69.6	69.1
DFF	85.2	55.0	84.0	59.0	70.0	82.8	70.0	79.5	53.2	**81.4**	46.2	**80.1**	**79.8**	72.8	80.6	58.0	**82.4**	52.9	**79.4**	70.7	71.2
DDS-R	83.7	56.4	81.2	57.8	69.7	**83.3**	69.8	80.0	53.1	77.6	48.3	77.1	74.8	73.5	80.9	57.1	79.7	53.9	77.6	68.6	70.2
DDS-U	**85.6**	57.4	**85.3**	59.7	**71.8**	**83.3**	**71.2**	**82.0**	**55.0**	80.3	**53.4**	78.8	77.0	**74.1**	**82.7**	**61.9**	**82.4**	**55.3**	78.1	**72.6**	**72.4**
With the “Raw” evaluation metric in (Yu et al. 2018)																					
CASENet	69.7	58.5	71.0	47.0	54.8	69.7	60.6	67.5	48.1	64.4	38.2	66.6	66.3	61.1	68.9	46.8	70.2	47.1	65.0	57.7	60.0
SEAL	79.8	64.3	79.1	**55.0**	63.4	78.3	66.8	75.5	**52.7**	74.3	44.3	**77.0**	**73.4**	68.2	76.6	52.2	76.6	50.9	73.5	66.1	67.4
STEAL	75.6	61.3	75.2	48.9	58.2	72.2	65.9	71.9	48.8	67.6	38.3	72.5	68.2	65.1	72.3	49.6	73.7	44.9	69.8	62.0	63.1
DFF	76.9	61.0	76.6	51.1	59.8	75.3	63.8	72.0	49.3	72.3	40.0	71.8	71.0	64.0	71.0	49.9	72.3	46.3	72.2	64.6	64.1
DDS-R	78.7	63.9	77.5	53.2	62.8	77.9	65.1	74.8	51.9	69.2	44.4	73.4	70.5	66.8	75.1	54.1	74.4	50.2	75.2	65.4	66.2
DDS-U	**81.6**	**65.9**	**79.7**	54.8	**65.5**	**79.4**	**68.9**	**77.1**	52.6	**74.5**	**49.7**	76.5	**73.4**	**69.9**	**78.1**	**55.6**	**78.7**	**51.8**	**75.8**	**68.6**	**68.9**

缘和语义分割进行组合用于 SED 是不够的。DSN、CASENet 和 DDS 的平均运行时间如表 6 所示。DDS 可以生成最先进的语义边缘，尽管速度略有下降。

Yu 等人(2018)发现，有一些 SBD 标签存在噪音，因此他们重新标注了测试集中的 1059 张图像，从而形成了一个新的测试集。我们在此新数据集上，将我们的方法与 DSN (Yu et al. 2017)、CASENet (Yu et al. 2017)、SEAL (Yu et al. 2018)和 DFF (Hu et al. 2019)进行了比较。在所有评测指标上，DDS 都可以提高 CASENet 和 DFF 的性能。具体而言，就(Hariharan et al. 2011)中的指标而言，DDS-R 和 DDS-U 的 ODS F-measures 分别比最近的 SEAL (Yu et al. 2018)高 0.8% 和 2.4%。需要注意的是，SEAL 使用新的训练策略对 CASENet 进行了再训练，即，同时进行对齐和学习。在相同的训练策略

下，基于(Hariharan et al. 2011)中的指标，在 ODS F-measure 上，DDS-R 比 CASENet 高了 3.0%。

为了更好地可视化边缘预测的结果，我们在图 4 中展示了一个示例。我们还在图 5 中展示了侧端激活的归一化图像。所有激活都在 Sigmoid 非线性化之前获得。为了简化图的布置，我们没有展示 DDS-R 的第四侧端的激活。从第一侧端到第三侧端，可以看到 DDS-R 的特征图比 DSN 和 CASENet 清晰得多。DDS-R 可以发现清晰的类别无关的边缘，而 DSN 和 CASENet 却存在噪音激活。例如，在 CASENet 中，如果不对第一侧端 ~ 第三侧端添加深监督，那么几乎找不到边缘激活。对于分类激活，DDS-R 可以清楚地区分人与马，而 DSN 和 CASENet 却不能。因此，信息转换器还有助于更好地优化第五侧端来进行类别相关的分类，这进一步验证了所提出的 DDS 体系结构的可行性。

图 6 中展示了更多定性比较示例图。与其他检测器相比，DDS-R/DDS-U 可以产生更加清晰和平滑的边缘。有趣的是，在第二列中，大多数检测器可以识别出带有缺失标注的物体的边界，例如，被遮挡的餐桌和人的手臂。在第三列中，DDS-R/DDS-U 可以在小猫的边界处产生强响应，而其他检测器仅具有弱响应。这表明 DDS 在检测小物体方面更强大。我们还发现 DDS-U 和 SEAL 可以生成更细的边缘，这表明使用常规的的未加权的 Sigmoid 交叉熵损失和细化的真值图边缘进行训练有助于准确定位细边界。

5.4 Cityscapes 上的评测

Cityscapes 数据集(Cordts et al. 2016)比 SBD (Hariharan et al. 2011)更具挑战性。Cityscapes 数据集中的图像是在更复杂的场景中捕获的，通常是在不同城市的城市街道场景中。每个图像中有更多的物体，尤其是重叠的物体。因此，Cityscapes 对于测试语义边缘检测器可能更有说服力。我们不仅将 DDS 与 5 个语义边缘检测器进行比较，即 DSN (Yu et al. 2017)、CASENet (Yu et al. 2017)、SEAL (Yu et al. 2018)、STEAL (Acuna et al. 2019) 和 DFF (Hu et al. 2019)，还有两个先进的语义分割模型，即 PSPNet (Zhao et al. 2017)和 DeepLabv3+ (Chen et al. 2018)。我们提取 PSPNet 和 DeepLabv3+ 的语义分割边界，以生成它们对应的单个像素宽度的语义边缘(Hariharan et al. 2011)。评测结果如表 7 所示。DDS-R 和 DDS-U 均明显优于其他方法。PSPNet (Zhao et al. 2017)在 SED 方面具有竞争力，但性能不如边缘检测器。尽管在语义分割方面，DeepLabv3+ (Chen et al. 2018)比 PSPNet (Zhao et al. 2017)性能更好，但是 DeepLabv3+ 在 SED 方面的表现却远差于其他方法。这表明语义分割不能总是产生可靠的边界，因此有必要进一步进行 SED 的研究。在具有相同损失函数的情况下，根据(Hariharan et al. 2011)中的指标，DDS-R 的 ODS F-measure 比 CASENet 高 3.1%，而 DDS-U 则比 SEAL 高 4.3%。图 7 中显示了一些定性比较。我们可以看到 DDS-R/DDS-U 产生了更平滑和更清晰的边缘。

5.5 PASCAL VOC2012 上的评测

VOC2012 (Everingham et al. 2012)包含与 SBD 数据集(Hariharan et al. 2011) 中相同的 20 个物体类别。对于 VOC2012 的验证集，我们移除了 SBD 训练集中出现的图像，从而生成了一个包含 904 张图像的新验证集。因此，在生成的新验证集和 SBD 训练集之间没有重叠。我们使用新的 VOC2012 验证集，并使用在 SBD 上训练的模型来评测一些最近的模型。这样，我们可以测试各种方法的普适性。但是，VOC2012 的原始标注在每个物体边界附近都留下了一个未标注的细区域，从而影响评测。取而代之的是，我们使用(Yang et al. 2016)中的方法，采用稠密的 CRF 模型(Krähenbühl & Koltun 2011)用相邻的物体标签填充不确定区域。我们进一步根据(Hariharan et al. 2011)生成单个像素宽度的语义边缘。据我们所知，这是在 VOC2012 数据集上评测 SED 的首个研究。评测结果如表 8 所示。和预期的一样，基于 DDS 的方法可达到最佳性能，这表明 DDS 网络具有良好的普适性。

6 总结

在本文中，我们研究了 SED 问题。先前的方法认为深监督对于 SED 是不必要的 (Hu et al. 2019; Yu et al. 2017, 2018)。我们证明这是错误的，通过对体系结构进行适当地重新设计，可以对网络进行深监督来改善检测结果。我们方法的核心是引入信息转换器，它在帮助为高层的类别感知边缘和低层的类

图 7 Cityscapes 数据集(Hariharan et al. 2011)中的一些示例。从上到下依次是：颜色代码、原图像、真值图、DSN、CASENet (Yu et al. 2017)、SEAL (Yu et al. 2018)、所提出的 DDS-R 和 DDS-U。我们遵循(Yu et al. 2018)中的颜色编码协议。可以看出，DDS 产生的边缘更加平滑和清晰。

别无关边缘生成一致的梯度信号方面起着核心作用。DDS 在包括 SBD (Hariharan et al. 2011)、Cityscapes (Cordts et al. 2016) 和 PASCAL VOC2012 (Everingham et al. 2012) 在内的几个流行数据集上均达到了最佳检测性能。我们利用深监督来训练深度网络的思想为 SED 以及其他高级任务（如语义分割(Chen et al. 2016; Maninis et al. 2017)、物体检测(Ferrari et al. 2008; Maninis et al. 2017)和实例分割(Hayder et al. 2017; Kirillov et al. 2017)）更好的利用深度网络中的丰富的特征层次结构开辟了新的方向。

未来工作. 除了类别无关的边缘检测和 SED，计算机视觉(Zamir et al. 2018) 中通常还存在相关任务，例如分割和显著性检测、物体检测和关键点检测、边缘检测和骨架提取。建立多任务网络以解决相关任务是在实际应用中节省计算资源的一种好方法(Hou et al. 2018)。但是，如本文所示，不同任务之间的分散监督通常会阻碍此目标。从这个角度来看，所提出的 DDS 为多任务学习提供了新的视角。将来，我们计划将信息转换器的思想用于更多的相关的任务。

参考文献

Acuna, D., Kar, A., & Fidler, S. (2019). Devil is in the edges: Learning semantic boundaries from noisy annotations. In *IEEE Conf. Comput. Vis. Pattern Recog.* (pp. 11075–11083).

Amer, M. R., Yousefi, S., Raich, R., & Todorovic, S. (2015). Monocular extraction of 2.1 d sketch using constrained convex optimization. *Int. J. Comput. Vis.*, *112*(1), pp. 23–42.

Arbeláez, P., Maire, M., Fowlkes, C., & Malik, J. (2011). Contour detection and hierarchical image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, *33*(5), pp. 898–916.

Bertasius, G., Shi, J., & Torresani, L. (2015a). DeepEdge: A multi-scale bifurcated deep network for top-down contour detection. In *IEEE Conf. Comput. Vis. Pattern Recog.* (pp. 4380–4389).

Bertasius, G., Shi, J., & Torresani, L. (2015b). High-for-low and low-for-high: Efficient boundary detection from deep object features and its applications to high-level vision. In *Int. Conf. Comput. Vis.* (pp. 504–512).

Bertasius, G., Shi, J., & Torresani, L. (2016). Semantic segmentation with boundary neural fields. In *IEEE Conf. Comput. Vis. Pattern Recog.* (pp. 3602–3610).

Canny, J. (1986). A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, *8*(6), pp. 679–698.

Chan, T.-H., Jia, K., Gao, S., Lu, J., Zeng, Z., & Ma, Y. (2015). PCANet: A simple deep learning baseline for image classification? *IEEE Trans. Image Process.*, *24*(12), pp. 5017–5032.

Chen, L.-C., Barron, J. T., Papandreou, G., Murphy, K., & Yuille, A. L. (2016). Semantic image segmentation with task-specific edge detection using CNNs and a discriminatively trained domain transform. In *IEEE Conf. Comput. Vis. Pattern Recog.* (pp. 4545–4554).

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Eur. Conf. Comput. Vis.* (pp. 833–851).

Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., et al. (2016). The cityscapes dataset for semantic urban scene understanding. In *IEEE Conf. Comput. Vis. Pattern Recog.* (pp. 3213–3223).

Deng, R., Shen, C., Liu, S., Wang, H., & Liu, X. (2018). Learning to predict crisp boundaries. In *Eur. Conf. Comput. Vis.* (pp. 570–586).

Dollár, P., & Zitnick, C. L. (2015). Fast edge detection using structured forests. *IEEE Trans. Pattern Anal. Mach. Intell.*, *37*(8), pp. 1558–1570.

Everingham, M., Van Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2012). The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.

Ferrari, V., Fevrier, L., Jurie, F., & Schmid, C. (2008). Groups of adjacent contour segments for object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, *30*(1), pp. 36–51.

Ferrari, V., Jurie, F., & Schmid, C. (2010). From images to shape models for object detection. *Int. J. Comput. Vis.*, *87*(3), pp. 284–303.

Ganin, Y., & Lempitsky, V. (2014). N^4-Fields: Neural network nearest neighbor fields for image transforms. In *Asian Conf. Comput. Vis.* (pp. 536–551).

Hardie, R. C., & Boncelet, C. G. (1995). Gradient-based edge detection using nonlinear edge enhancing prefilters. *IEEE Trans. Image Process.*, *4*(11), pp. 1572–1577.

Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., & Malik, J. (2011). Semantic contours from inverse detectors. In *Int. Conf. Comput. Vis.* (pp. 991–998).

Hayder, Z., He, X., & Salzmann, M. (2017). Boundary-aware instance segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.* (pp. 5696–5704).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recog.* (pp. 770–778).

Henstock, P. V., & Chelberg, D. M. (1996). Automatic gradient threshold determination for edge detection. *IEEE Trans. Image Process.*, *5*(5), pp. 784–787.

Hou, Q., Cheng, M.-M., Hu, X., Borji, A., Tu, Z., & Torr, P. (2019). Deeply supervised salient object detection with short connections. *IEEE Trans. Pattern Anal. Mach. Intell.*, *41*(4), pp. 815–828.

Hou, Q., Liu, J., Cheng, M.-M., Borji, A., & Torr, P. H. (2018). Three birds one stone: A unified framework for salient object segmentation, edge detection and skeleton extraction. *arXiv preprint arXiv:1803.09860*.

Hu, X., Liu, Y., Wang, K., & Ren, B. (2018). Learning hybrid convolutional features for edge detection. *Neurocomputing*.

Hu, Y., Chen, Y., Li, X., & Feng, J. (2019). Dynamic feature fusion for semantic edge detection. In *Int. Joint Conf. Artif. Intell.* (pp. 782–788).

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., et al. (2014). Caffe: Convolutional architecture for fast feature embedding. In *ACM Int. Conf. Multimedia.* (pp. 675–678).

Khoreva, A., Benenson, R., Omran, M., Hein, M., & Schiele, B. (2016). Weakly supervised object boundaries. In *IEEE Conf. Comput. Vis. Pattern Recog.* (pp. 183–192).

Kirillov, A., Levinkov, E., Andres, B., Savchynskyy, B., & Rother, C. (2017). Instancecut: from edges to instances with multicut. In *IEEE Conf. Comput. Vis. Pattern Recog.* (pp. 5008–5017).

Kokkinos, I. (2016). Pushing the boundaries of boundary detection using deep learning. In *Int. Conf. Learn. Represent.* (pp. 1–12).

Konishi, S., Yuille, A. L., Coughlan, J. M., & Zhu, S. C. (2003). Statistical edge detection: Learning and evaluating edge cues. *IEEE Trans. Pattern Anal. Mach. Intell.*, *25*(1), pp. 57–74.

Krähenbühl, P., & Koltun, V. (2011). Efficient inference in fully connected CRFs with gaussian edge potentials. In *Adv. Neural Inform. Process. Syst.* (pp. 109–117).

Lee, C.-Y., Xie, S., Gallagher, P., Zhang, Z., & Tu, Z. (2015). Deeply-supervised nets. In *Artificial Intelligence and Statistics.* (pp. 562–570).

Lim, J. J., Zitnick, C. L., & Dollár, P. (2013). Sketch tokens: A learned mid-level representation for contour and object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.* (pp. 3158–3165).

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.* (pp. 2117–2125).

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2020). Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, *42*(2), pp. 318–327.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft COCO: Common objects in context. In *Eur. Conf. Comput. Vis.* (pp. 740–755).

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., et al. (2016). SSD: Single shot multibox detector. In *Eur. Conf. Comput. Vis.* (pp. 21–37).

Liu, Y., Cheng, M.-M., Hu, X., Bian, J.-W., Zhang, L., Bai, X., et al. (2019). Richer convolutional features for edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, *41*(8), pp. 1939–1946.

Liu, Y., Cheng, M.-M., Hu, X., Wang, K., & Bai, X. (2017). Richer convolutional features for edge detection. In *IEEE Conf. Comput. Vis. Pattern Recog.* (pp. 3000–3009).

Liu, Y., Jiang, P.-T., Petrosyan, V., Li, S.-J., Bian, J., Zhang, L., et al. (2018). DEL: deep embedding learning for efficient image segmentation. In *Int. Joint Conf. Artif. Intell.* (pp. 864–870).

Mafi, M., Rajaei, H., Cabrerizo, M., & Adjouadi, M. (2018). A robust edge detection approach in the presence of high impulse noise intensity through switching adaptive median and fixed weighted mean filtering. *IEEE Trans. Image Process.*, *27*(11), pp. 5475–5490.

Maninis, K.-K., Pont-Tuset, J., Arbelaez, P., & Van Gool, L. (2017). Convolutional oriented boundaries: From image segmentation to high-level tasks. *IEEE Trans. Pattern Anal. Mach. Intell.*, *40*(4), pp. 819–833.

Martin, D. R., Fowlkes, C. C., & Malik, J. (2004). Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Trans. Pattern Anal. Mach. Intell.*, *26*(5), pp. 530–549.

Ramalingam, S., Bouaziz, S., Sturm, P., & Brand, M. (2010). Skyline2gps: Localization in urban canyons using omni-skylines. In *IEEE\RSJ Int. Conf. Intell. Robot. Syst.* (pp. 3816–3823).

Shan, Q., Curless, B., Furukawa, Y., Hernandez, C., & Seitz, S. M. (2014). Occluding contours for multi-view stereo. In *IEEE Conf. Comput. Vis. Pattern Recog.* (pp. 4002–4009).

Shen, W., Wang, X., Wang, Y., Bai, X., & Zhang, Z. (2015). DeepContour: A deep convolutional feature learned by positive-sharing loss for contour detection. In *IEEE Conf. Comput. Vis. Pattern Recog.* (pp. 3982–3991).

Shui, P.-L., & Wang, F.-P. (2017). Anti-impulse-noise edge detection via anisotropic morphological directional derivatives. *IEEE Trans. Image Process.*, *26*(10), pp. 4962–4977.

Sobel, I. (1970). Camera models and machine perception. *Technical report*, Stanford Univ Calif Dept of Computer Science.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. (2015). Going deeper with convolutions. In *IEEE Conf. Comput. Vis. Pattern Recog.* (pp. 1–9).

Tang, P., Wang, X., Feng, B., & Liu, W. (2017). Learning multi-instance deep discriminative patterns for image classification. *IEEE Trans. Image Process.*, *26*(7), pp. 3385–3396.

Trahanias, P. E., & Venetsanopoulos, A. N. (1993). Color edge detection using vector order statistics. *IEEE Trans. Image Process.*, *2*(2), pp. 259–264.

Wang, L., Ouyang, W., Wang, X., & Lu, H. (2015). Visual tracking with fully convolutional networks. In *Int. Conf. Comput. Vis.* (pp. 3119–3127).

Wang, Y., Zhao, X., Li, Y., & Huang, K. (2019). Deep crisp boundaries: From boundaries to higher-level tasks. *IEEE Trans. Image Process.*, *28*(3), pp. 1285–1298.

Xie, S., & Tu, Z. (2015). Holistically-nested edge detection. In *Int. Conf. Comput. Vis.* (pp. 1395–1403).

Xie, S., & Tu, Z. (2017). Holistically-nested edge detection. *Int. J. Comput. Vis.*, *125*(1-3), pp. 3–18.

Yang, J., Price, B., Cohen, S., Lee, H., & Yang, M.-H. (2016). Object contour detection with a fully convolutional encoder-decoder network. In *IEEE Conf. Comput. Vis. Pattern Recog.* (pp. 193–202).

Yang, W., Feng, J., Yang, J., Zhao, F., Liu, J., Guo, Z., et al. (2017). Deep edge guided recurrent residual learning for image super-resolution. *IEEE Trans. Image Process.*, *26*(12), pp. 5895–5907.

Yu, F., & Koltun, V. (2016). Multi-scale context aggregation by dilated convolutions. In *Int. Conf. Learn. Represent.* (pp. 1–13).

Yu, Z., Feng, C., Liu, M.-Y., & Ramalingam, S. (2017). CASENet: Deep category-aware semantic edge detection. In *IEEE Conf. Comput. Vis. Pattern Recog.* (pp. 5964–5973).

Yu, Z., Liu, W., Zou, Y., Feng, C., Ramalingam, S., Kumar, B., et al. (2018). Simultaneous edge alignment and learning. In *Eur. Conf. Comput. Vis.* (pp. 400–417).

Zamir, A. R., Sax, A., Shen, W., Guibas, L., Malik, J., & Savarese, S. (2018). Taskonomy: Disentangling task transfer learning. In *IEEE Conf. Comput. Vis. Pattern Recog.* (pp. 3712–3722).

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *IEEE Conf. Comput. Vis. Pattern Recog.* (pp. 2881–2890).