指导教师：杨涛　　提交时间：2015.3.29

The task of Digital Image Processing

数字图像处理

School of Computer Science

No: 1

姓名：田慧媛

学号：2012302553

班号：10011207

告诉我你看到了什么，我会告诉你它在哪里

Jia Xu[1] Alexander G. Schwing[2] Raquel Urtasun[2,3]

[1]University of Wisconsin - Madison [2]University of Toronto [3]TTI Chicago

jiaxu@cs.wisc.edu {aschwing, urtasun}@cs.toronto.edu

摘要

我们解决弱标记语义分割问题，在这个问题中图像标签编码哪些类在目前在场景中是的唯一的标注来源。在没有像素标签可用，甚至不在训练时间的情况下这是一个极其困难的问题。在本文中，我们认为这一问题可以被形式化为一个学习潜在的结构的预测框架的实例，在那里图形化的模型对存在和不存在的类以及超像素的语义标签分配进行编码。因此，我们能够利用标准的具有良好的理论性能的算法。我们使用具有挑战性的 SIFT-流量数据集证明我们的方法的有效性，并显示每个类平均精度超出最高水准 7%。

1.介绍

传统的语义分割方法要求大量标记于像素水平的训练图像的集合。尽管存在像亚马逊土耳其机器人(MTurk)那样的众包系统，浓密地标记图像仍然是一个非常昂贵的过程，特别是因为多标注通常被用来标记每个图像。此外，质量控制过程经常要对批注进行审查。

在这里，我们对利用弱标注来降低标记成本感兴趣。尤其是，我们利用捕获在当前场景中的类的图像标记作为我们唯一标注的来源 （举例说明见图 1）。这是一个有趣的设置，比如标签要么可以很容易地从大部分联机照片集合得到，要么可以很容易地以比标注语义分割更低的成本得到。然而，这项任务非常具有挑战性，由于语义标签的超像素分配是未知的，所以即使是在训练时间，外观模型不能被训练。

图 1. 我们的方法要训练以在场景中存在的类的形式来进行标签，并学会了一种分割模型，即使没有任何基于像素的标注可用。

几种方法被用来研究这个设置。在早期的工作，Verbeek 和 Triggs [29] 提出了潜伏面模型，采用概率潜在语义分析 (PLSA）模拟每个作为潜在的有限混合模型类的图像，也被称为方面。作者用马尔可夫随机场 （MRF）扩展这种方法来捕获空间关系。这种模式由 Vezhnevets et al.[30，31，32]一系列的文件进一步扩展，例如，若要利用多个图像之间的信息。然而，由此产生的优化问题非常复杂而且不光滑，这使我们的学习成了一个很困难任务。因此，几种启发式算法被采用来使这一问题在计算上更加容易。

在本文中，我们表明这这一问题可以被形式化为一个学习潜在的结构的预测框架的实例，在那里图形化的模型对存在和不存在的类以及超像素的语义标签的分配进行编码。作为结果，我们能够利用已经为更一般的设置开发地具有良好理论性能的算法。根据我们的模型，不同程度的监督可以通过指定哪个变量是潜在的或者是可观察的来简单地表达，而无需更改已学习的和推理的算法。我们使用具有挑战性的 SIFT-流量数据集[14]证明我们的方法的有效性，并显示每个类平均精度超出最高水准 7%。在剩下的论述里，我们首先审查有关的工作。然后，我们提出我们弱标签分割框架，其次是实验评价和结论。我们的代码公开在 http://pages.cs.wisc.edu/ jiaxu/projects/weak-label-seg/。

2.相关工作

许多不同的技术被提出来解决像素标签在训练时可用的全面监督的设置。其中最成功的技术是基于自底向上的区域提取方法，然后重新排列以获得最终的分割[8，4]。另一种流行的方法是，制定作为推理的（条件）的马尔科夫随机分割字段 [3]，特别是在寻求一个完全整体的场景解释 [34，12]时。

注释器提供哪些对物体/类在场景中的信息是相对容易的。然而，仔细描画所有可见的对象的轮廓是很繁琐的。因此，标注花的时间与费用可以通过利用图像标签大大降低，特别是当这些标注可以很容易从很多的图像集合中获得时。

不过，由于全面监督任务很具有挑战性，所以在弱标记设置方面只有小部分工作完成。其中第一个学习只给定图像标签分割模型的方法，是利用潜在方面模型的[29]几个外观描述符和图像位置来学习概率潜在语义分析（PLSA）模型。“方面模型”源于著名的用于文档分类和涉及像素分类标签的“方面”的主题模型。因为这些模型不会捕获通常在图像中的空间的二维关系，所以 PLSA 在一元马尔可夫随机场中起重要作用。概括的介绍在一系列论文[30，31，32]里。对比潜在方面模型，这些新方法利用了不同的图像之间标签的相关性。然而，它们导致非凸、 非光滑的复杂的最优化问题问题变得很难优化。

监督不力的另一种形式是 2D 边界盒。切割以及切割的扩展广泛用于交互式图形/背景分割 [2，20]。这些方法学习了高斯混合模型的前景和背景，而且二进制 MRF 被用来对外观和光滑度的编码进行分割。被采用的能源是分模块化的，这使得通过图割的精确推理是可能的。笔画是提供弱标注的另一个流行方式，通常使用在有人类来更改错误的决策圈内。在[18]中，可变形的基于部件的模型 [7]和潜在的结构的支持向量机用来开发边界盒形式的弱标签。最近，[5] 表明，在开发 3D 信息和 3D 边界盒形式的弱标志时，可以实现人类标签性能。

一个相关的问题是协同分割，它对分割同时出现在一组图像[21]的对象感兴趣。大多数以前的方法专注于出现在所有图像[33，16]单一前景对象的设置。通过分析多前景对象[17]的子空间结构的多个对象，使用子模块最优化[11]的贪婪算法或者通过光谱判别聚类[10]的分组图像地区，此设置已扩展到分割多物体。

与我们最相关的工作是[32]，在这个工作中来自标签的弱标记分割问题的推导，使用了在节点表示像素级别的语义类的条件随机场（CRF），一元可能性编码外观，成对可能性编码平滑度。他们的主要贡献是一个学习外观模型和 CRF 权重的三步算法。特别是，在每次更新 CRF 权重之后，在给定当前的模型找到像素标签和给定估计标签的情况下，更新外观模型之间出现交替的最优化迭代。作者将特征权值优化看成是一种每个可能的权重向量定义一个不同的模型的模型选择过程。采用的最优化标准达到了预期的结果，它是通过将数据分成两个分别满足它们预期的分区来计算。因为价值函数是不可微分的，他们使用贝叶斯最优化来选择下一套参数。这使得学习变得极为困难和计算代价昂贵。

与此相反的是，在本文中我们展示来自弱标记的数据分割类存在问题可以表述为学习带有潜变量的结构化预测框架。因此，很好研究的算法，如隐藏条件随机场（HCRFs）[19]或潜在结构化支持向量机（LSSVMs）[35]及其高效的扩展[23]是可以利用的。这样就可以使更简单的最优化问题可以通过拥有良好的理论保证算法的来实现优化。

3.弱标记语义分割

在本文中，我们研究弱监督如何用于实现语义分割。尤其是，我们专注于描述哪些类在目前图像而且监督是由一系列标记提供的场景。为了实现这一目标，我们把这个问题构造为学习一个可以对每一个存在与不存在的类以及超像素语义类进行编码的图形化模型。

3.1 标记语义的分割

更加正式地，令 $y_i \in \{0,1\}$ 代表一个描述 i 是否在图像中的随机变量，$i \in \{1,...C\}$ 表示语义类，$h_j \in \{1,...,C\}$ 为使用 j 这一系列超像素来表示语义标注的随机变量，x 是图像证据。我们定义 $\mathbf{h} = (h_1,...,h_N)$ 为一个的图像中所有超像素的细分变量集合，$\mathbf{y} = (y_1,...,y_C)$ 为表示所有类的存在与不存在的二元变量集合。注意我们假设 h 没有任何训练实例而且只有 y 被标记。采用上述的假设，我们定义在知道 x 的情况下给定的组态（y，h）的可能性为

$$p_\epsilon(\mathbf{y},\mathbf{h}\mid x)=\frac{1}{Z_\epsilon(w)}\exp\frac{w^\top\phi(\mathbf{y},\mathbf{h},x)}{\epsilon},$$

其中 $Z_\epsilon(w)$ 是规格化常量也被称为配分函数。注意权重 w 是模型的参数，ϵ 是温度参数。

我们定义可能性 $\phi(\mathbf{y},\mathbf{h},x)$ 是一元术语编码标记的可能性，一元可能性编码外观模型分割的可能性和成对可能性确保两种类型的变量之间的兼容性可能性的总和。因此

$$w^\top\phi(\mathbf{y},\mathbf{h},x)=\sum_i w_i^{presT}\phi^{pres}(x,y_i)+\sum_j w_j^{apT}\phi^{ap}(x,h_j)+\sum_{i,j} w_{i,j}^{coT}\phi^{co}(y_i,h_j). \quad (1)$$

图 2 显示了引进概率模型的对依赖关系进行编码的图形化模型，用灰色节点描述观察到的变量。我们注意到，这种体系结构类似于整体模型的[34]中的前两个图层 ，但我们使用不同的可能性并实现在弱标记语义分割设置。现在，我们讨论关于可能性的更多的细节。

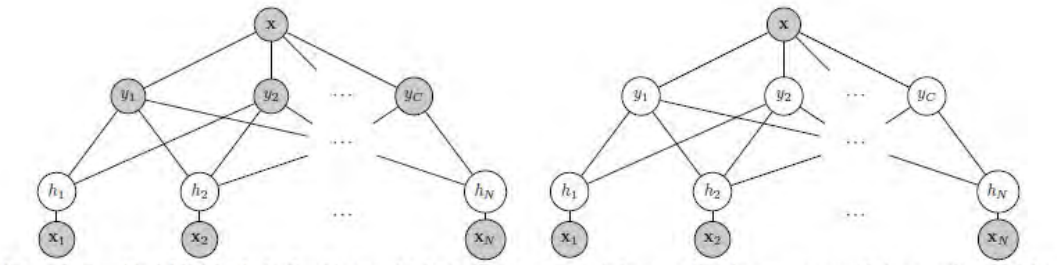

Figure 2. **Graphical Model:** (Left) Graphical model for learning as well as inference when the tags are provided at test time. (Right) Graphical model for inference when the tags are not provided at test time.

外观模型：我们利用[27]的超像素特征，其中包括结构/SIFT、 颜色、 形状、 位置和 GIST。这导致我们可以使用 PCA 将一个 1690 维特征向量减少到 100 维向量。为了实现最终的功能，我们追加超像素位置(即中点的 y 坐标)来形

成最后的特征。注意我们学习不同类的权重集合，其适合一个 $101 \cdot C$ 维的特征向量。

存在/不存在：我们构建一个二维向量对每个类的存在进行编码。在训练中，这种可能性建立于地面实况，如果 i 不存在，$i.e., \phi^{pres}(y_i, x) = [1; -1]$，如果 i 存在，那么 $\phi^{pres}(y_i, x) = [-1; 1]$。在测试的时候，当此信息是隐藏的，这种可能性来自图像水平标记分类器。我们建议读者参阅实验部分来获取更多关于该预报器的具体形式的更多详细信息。注意，通常情况下同时在训练和测试时间使用一个预报器，然而，我们发现 oracle 预测器在训练时使用能得到更好的效果。我们假设这是因为在这样的设置中监督是非常薄弱的。

兼容性： 兼容性支持存在变量类和超像素之间的一致性，这样的信息是传播所有分割的方法。尤其是，它惩罚被推断为不存在的超像素类的配置。因此

$$\phi^{comp}(y_i, h_j) = \begin{cases} -\eta & \text{if } y_i = 0 \text{ and } h_j = i \\ 0 & \text{otherwise} \end{cases}$$

其中 η 是一个大数字 （在我们的实验中是10^5的数量级）。

3.2.在弱标记的设置学习

在学习期间，我们感兴趣的是线性功能组合的估计，比如分布方程式(1)能够区分变量 y 和 h 代表的 '好' 和 '坏'。我们通过给定数据样本的训练定义'好'。对比包含充分标记配置（y,h）的训练样本的全面监督的设置，可用的数据只是部分被标记。尤其是，训练集 D 由|D| 图像标记对（y, x）组成，即

$$\mathcal{D} = \{(\mathbf{y}, x)_i\}_{i=1}^{|\mathcal{D}|}.$$

在学习期间，亏损函数$\ell(\hat{\mathbf{y}}, \mathbf{y})$一般被包含进去使算法更完善

$$p_\epsilon^\ell(\hat{\mathbf{y}}, \hat{\mathbf{h}}, x) = \frac{1}{Z_\epsilon^\ell(w)} \exp \frac{w^\top \phi(\hat{\mathbf{y}}, \hat{\mathbf{h}}, x) + \ell(\hat{\mathbf{y}}, \mathbf{y})}{\epsilon}.$$

直觉使问题很难更好的概括，因此，我们想要去寻找一个权重向量 w，这使消极的（损失增加）培训数据 D 边际对数后和对 w 先验分布的正则项的总和最小化 . 最后的程序如下所示：

$$\min_{w} \frac{1}{2}\|w\|_2^2 - \sum_{(\mathbf{y},x)\in\mathcal{D}} \epsilon \ln \sum_{\hat{\mathbf{h}}} p_\epsilon^\ell(\mathbf{y},\hat{\mathbf{h}} \mid x). \qquad (2)$$

请注意，我们在忽视了不能被观察到的超像素变量 h 来获得的观测数据的可能性（即类标签）。

上述计划概括了几个著名设置。让 $\epsilon=0$，我们得到结构的带有如[35]介绍的潜变量的支持向量机，当设置ϵ =1，产生隐藏随机场[19]。 在分别假设时$\epsilon=$1 和ϵ =0 时，在充分观测数据的情况下[13]的条件下我们获得条件随机域框架或结构化支持向量机的[25、28]。

弱标记的设置在解决一般的图形化模型的时候难度更大。除了 h 和 y 以指数大小集求和外，其他的挑战来自于配分函数的式(2)给定的物体的非凸性。然而，我们注意到式（2）给出的程序代价函数与这个不同，每一个对于 w. 都是凸面的。我们发现可以利用期望最大化（EM）泛化的凹凸程序(CCCP)[36]，尽可能简化式（2）。

CCCP 是一种迭代方法。我们在每次迭代过程中使目前迭代的 w 的凹面部分线性化，剩下的凸面目标通过线性项更新权重向量 w 增强。重要的是，这种方法能保证收敛到平稳点[24]。为了使凹面部分线性化，我们需要计算关于分布的非观测的变量 h 的特征向量$\phi(\mathbf{y},\mathbf{h},x)$ 的期望 。更正式地说这种期望被定义为

$$E_{p(\hat{\mathbf{h}}|x)}\left[\phi(\mathbf{y},\hat{\mathbf{h}},x)\right] = \sum_{\hat{\mathbf{h}}} p(\hat{\mathbf{h}} \mid x)\phi(\mathbf{y},\hat{\mathbf{h}},x).$$

鉴于这种期望，我们使用改良的经验均值实现了全面监督目标。请注意，这个求导很自然地推导出了一个我们为了获得期望首先需计算的非观测变量 h 分布的两步方法，在此之前使用此信息来解决完全监督式学习问题。该过程如图 3 所示。

Structured prediction with latent variables
Iterate between

1. The latent variable prediction problem:

$$\forall x \text{ compute } \quad E_{p(\hat{\mathbf{h}}|x)}\left[\phi(\mathbf{y}, \hat{\mathbf{h}}, x)\right]$$

2. Solving the parameter update task

$$\min_{w} \frac{1}{2}\|w\|_2^2 + \sum_{(\mathbf{y},x)\in\mathcal{D}} \left(\epsilon \ln Z_\epsilon^\ell(w) - w^\top E_{p(\hat{\mathbf{h}}|x)}\left[\phi(\mathbf{y}, \hat{\mathbf{h}}, x)\right]\right)$$

Figure 3. Latent Structured Prediction via CCCP

第一步需要着重注意的是，在我们的图形模型中我们可以琐细地解决给定双分裂模型弱标记分割任务的“潜变量的预测问题”。假设地面实况标记 y 是已知的（见图 2），该模型分解成一元的超像素，而且 $p(\hat{\mathbf{h}} \mid x)$ 分布的推理可以有效而精确地完成。第二步我们需要解决完全监督式学习任务。我们向读者推荐[23]去寻求最优化代价函数的有效途径。

3.3.亏损函数

存在类以及像素方位标注的分布遵循幂律分布（即许多类很少发生）。为了把这个考虑进去我们推导出采用在图像一级存在类的统计数字的损失函数。因为分割度量标准是每类的平均精度，我们的损失认为在类中出现的很少的错误更重要。尤其是，对于每个类，我们对包含此类的许多训练图像进行计数，然后使频率向量 t 规格化使其总和为 1。损失函数 $\ell(\hat{\mathbf{y}}, \mathbf{y})$ 被分解成一元关系，即

$$\ell(\hat{\mathbf{y}}, \mathbf{y}) = \sum_{i\in\{1,\ldots,C\}} \ell_i(\hat{y}_i, y_i)$$

$$\ell_i(\hat{y}_i, y_i) = \begin{cases} \frac{1}{t_i} & \text{if } y_i \neq \hat{y}_i \text{ and } y_i = 0 \\ t_i & \text{if } y_i \neq \hat{y}_i \text{ and } y_i = 1 \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

其中 y_i 是地面实况标签，$\hat{y}_i$ 是关于 i 的预测。注意我们的亏损函数只是定义在训练中可见的场景中存在的变量 y 上。

3.4.推理

最小能量或最高概率$p(\mathbf{y},\mathbf{h}\mid x)$，又称为最大后验概率（MAP）的配置，可以通过解决下面的问题求解

$$(\mathbf{y}^*,\mathbf{h}^*)=\arg\max_{\mathbf{y},\mathbf{h}} w^\top\phi(\mathbf{y},\mathbf{h},x) \qquad (4)$$

图像 x 已给出。这是一个 NP 难的任务，因为最优化等价于一个整数线性规划。幸运的是，线性规划（LP）松弛已被证明非常有效。我们利用图形化的模型结构使用一种消息传递方法。尤其是，我们使用分布式凸面置信传播 (dcBP) [22]其具有收敛性保证。请注意，这并不适合其他信息传递算法如多圈置信传播。

4.实验评价

我们的实验利用的是 SIFT 流分割的数据集 [14]，其中包含 2688 种图像和 C = 33 类。这个数据集非常具有挑战性，由于含有大量类（4.43 类每图像）而且他们的频率按幂律分布。如表 2 的第一行所示，几个“背景”类像天空，海洋和树是非常常见的，而“对象”类如人、 公交车和太阳是非常罕见的。我们使用[14]提供的标准数据集分割（2488 个训练图像和 200 个测试图像）。

沿着[32]我们报告每类平均精度作为我们的度量标准。这个度量认为每个类是同等重要的而不取决于他们的频率。我们使用了超度量的等高线图[1]构建我们的超像素，即使只有小数目的超像素被使用仍然可以很好尊重边界限。在我们的实验中，我们设置边界概率阈值为 0.14，其结果是平均每个图像有 19 个分割。

在我们的实验中，我们利用两个设置。在第一个实例中，我们遵循标准的弱标记设置，训练只给出了图像水平标记而没有给出任何像素级标注。在此设置中学习对应于图 2（左）的图形化模型，而推理显示在图 2（右）上。在第二个设置我们假设在训练和测试时间都给出了标签，结果是图 2（左）的图形化模型同时刻画了学习和推理过程。在标签一应俱全时采用图像集合是很自然的设置。

我们第一次实验只在训练利用标签。我们利用深度学习的图像标记分类器构建测试时间$\phi^{press}(x,y_i)$。尤其是，我们首先从预先在 ImageNet [6]训练的卷积神经网络（CNN）的第二层到最后一层提取每个图像 4096 维特征向量 。我

们使用公开执行[9]来计算特征以及每类的线性 SVM 形成最终的可能性。我们引用此设置为“我们的（CNN-Tag）”。

与最高水平的比较：表 1 比较了我们的方法和弱标记最高水平的方法。仅供参考，当像素标签在训练时可得（完全标记设置），我们也使用了最先进的技术。我们想要强调我们的方法显著优于（高出 7%） 弱标记的所有方法。此外，我们在全面监督也超越了由[14]中 Liu 等人提出的方法。每类的类速率由表 2 提供。我们发现我们的方法适合于有非常的独特而一致的外观，如沙滩、 阳光、 楼梯。我们忽视了几个类，如巴士、 人行道、 鸟，很大程度上是由于不同的外观和小训练设置大小。

Method	Supervision	Per-Class (%)
Tighe et al. [26]	full	39.2
Tighe et al. [27]	full	30.1
Liu et al. [14]	full	24
Vezhnevets et al. [31]	weak	14
Vezhnevets et al. [32]	weak	21
Ours (CNN-Tag)	weak	**27.9**
Ours (Truth-Tag)	weak	**44.7**

Table 1. Comparison to state-of-the-art on the SIFT-flow dataset. We outperformed the state-of-the-art in the weakly supervised setting by 7%.

图像标记预测质量：我们的 CNN 标记预测器预测标记的精度为 93.7%，其测量混淆矩阵对角线的平均数。表 2 的最后一行显示了标记预测器对于每个类的性能。有趣的是，标记预测误差与分割错误并没有关系，例如，人行道和鸟的标记预测精度很高，但是两个类的分割精度都很低。

%	sky	tree	building	mountain	road	car	sidewalk	sea	window	person	plant	rock	river	grass	door	field	sign	streetlight	sand	fence	pole	bridge	boat	awning	staircase	sun	balcony	crosswalk	bus	bird	averg.
Tag freq.	85.4	50.1	45.8	37.9	31.7	23.8	17.0	14.1	13.4	12.7	12.4	10.2	9.8	9.3	9.3	8.9	8.1	8.1	5.8	5.5	3.5	3.4	3.2	2.9	2.4	2.2	1.8	1.4	1.2	0.3	
CNN-Tag	5.8	25.1	18.5	38.0	9.1	6.8	1.6	13.5	8.7	16.7	15.2	60.3	63.2	53.0	48.6	76.7	38.7	26.3	91.1	81.2	40.8	20.1	77.3	56.3	81.8	100.0	43.9	0.5	61.3	0.0	**27.9**
Truth-Tag	12.3	27.9	23.3	33.0	10.0	14.2	4.5	18.8	10.8	22.0	37.1	83.0	64.6	63.1	49.3	81.4	41.7	22.0	87.6	81.3	36.9	39.5	74.9	44.5	79.5	100.0	37.6	23.7	58.5	58.5	**44.7**
CNN-ILT	93.0	81.5	86.5	82.5	91.0	94.5	90.0	97.5	93.0	89.5	86.0	91.0	92.5	89.5	93.5	92.5	91.5	93.0	95.5	93.5	95.5	94.5	96.5	95.0	98.0	100.0	99.0	99.0	98.0	99.0	**93.7**

Table 2. **Accuracy for each class:** First row shows tag frequency (percentage of images) for each class. Rows 2 and 3 show segmentation accuracy for each class when a CNN tag predictor or the ground truth tags are used respectively. The last row shows the accuracy of our image tag predictor for each class.

定性结果： 图 4 和图 5 分别显示成功和失败案例。典型的失败情形是创建超像素以及解决不同实例有很不同的外观的情形下的分割，例如角度的变化。

训练和测试给出的标签： 在我们的第二设置中，培训和测试时都给出了标记。注意这里的训练过程与以前的设置相同。然而，在测试时我们图像水平类的可能性是根据观察到的地面实况标记建立的。我们把这个设置表示为“我们的(Truth-Tag)”。如表 1 所示，我们几乎能使每级精度比以前的设置增加一倍。出人意料地是，虽然不需要任何标记在像素级的例子，我们的表现优于所有充分标记的方法。图 6 描述了此设置的定性结果。当给出了图像级别标记时，我们的方法能找出更具挑战性的类，例如建筑物。

测试时给出的局部标签： 当只提供了一个子集的标记时，我们进一步评价我们模型。对于每次运行，我们随机抽取一小部分的地面实况（GT）标签，并利用我们 CNN 标记分类器预测通过剩下的。结合的可能性被存储到我们的推理型。我们使用四个不同的样本比率{0.1,0.2,0.3,0.5}进行我们的实验。对于每个设置，我们重复我们的程序 10 次而且报表平均值和标准偏差。如图 7 所示，当给出了更多的 GT 标记时我们方法性能逐渐提高。

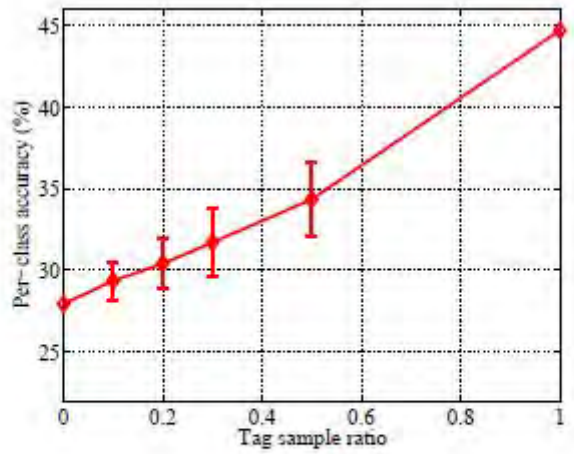

Figure 7. Per-Class accuracy as a function of the percentage of ground-truth tags available at test time.

5.结论

我们提出了语义分割方法，这种方法在没有像素级标记可用时能够开发图像标记形式的弱标记。我们已经展示了这一问题可以被正式化为带有潜在变量的图形化模型的结构化预测。与现有方法不同，这使得我们可以利用具有很好的理论保证的标准算法。我们证明了我们的方法的有效性和而且比最高水平有 7%的改进。我们对这个问题新奇的看法可以用于合并其他类型的监督而无需更改学习或

推理算法。在未来，我们计划利用其他如场景类型或边界盒的标注以及其他形式的如积极学习[15]的学习来进一步减少监督的需要。

鸣谢：感谢 Sanja Fidler 和 VikasSingh 对我们有帮助的论述。这项工作是由 NSF RI 1116584 和 ONR-N00014-13-1-0721 提供资金的。

参考文献

[1] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik. Contour Detection and Hierarchical Image Segmentation. PAMI,2011.

[2] Y. Boykov and M.-P. Jolly. Interactive Graph Cuts for Optimal Boundary and Region Segmentation of Objects in N-D Images. In Proc. ICCV, 2001.

[3] G. Cardinal, X. Boix, J. van de Weijer, A. D. Bagdanov, J. Serrat, and J. Gonzalez. Harmony Potentials for Joint Classification and Segmentation. In Proc. CVPR, 2010.

[4] J. Carreira, F. Li, and C. Sminchisescu. Object Recognition by Sequential Figure-Ground Ranking. IJCV, 2011.

[5] L. C. Chen, S. Fidler, A. Yuille, and R. Urtasun. Beat the MTurkers: Automatic Image Labeling from Weak 3D Supervision. In Proc. CVPR, 2014.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In Proc. CVPR, 2009.

[7] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained partbased models. PAMI, 2010.

[8] C. Gu, J. J. Lim, P. Arbelaez, and J. Malik. Recognition using region. In Proc. CVPR, 2009.

[9] Y. Jia. Caffe: An open source convolutional architecture for fast feature embedding. http://caffe.berkeleyvision.org/, 2013.

[10] A. Joulin, F. Bach, and J. Ponce. Multi-class cosegmentation. In Proc. CVPR, 2012.

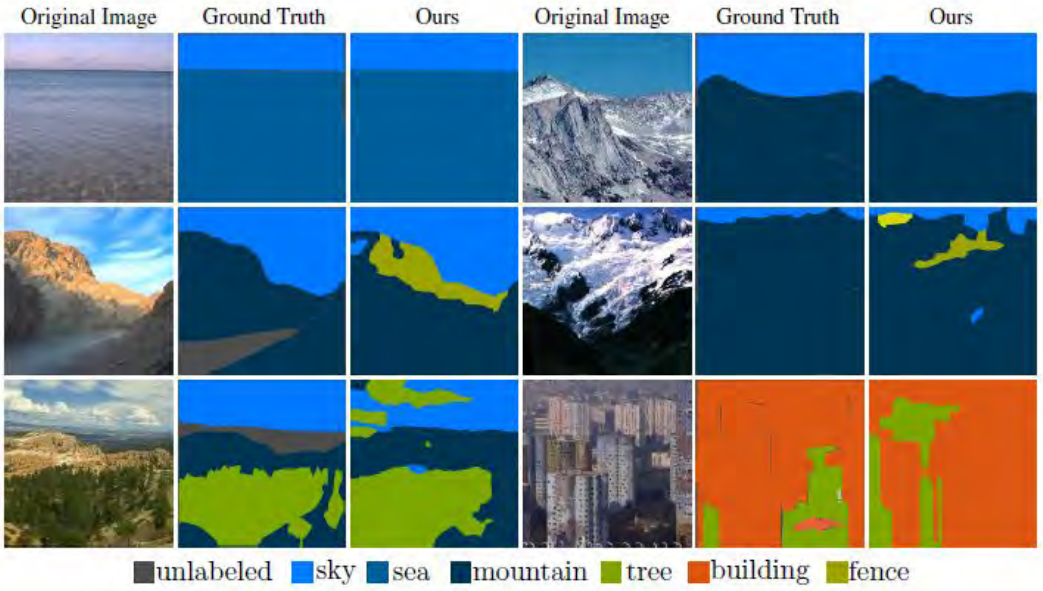

Figure 4. Sample results when tags are predicted at test time using a convolutional net. **Best viewed in color.**

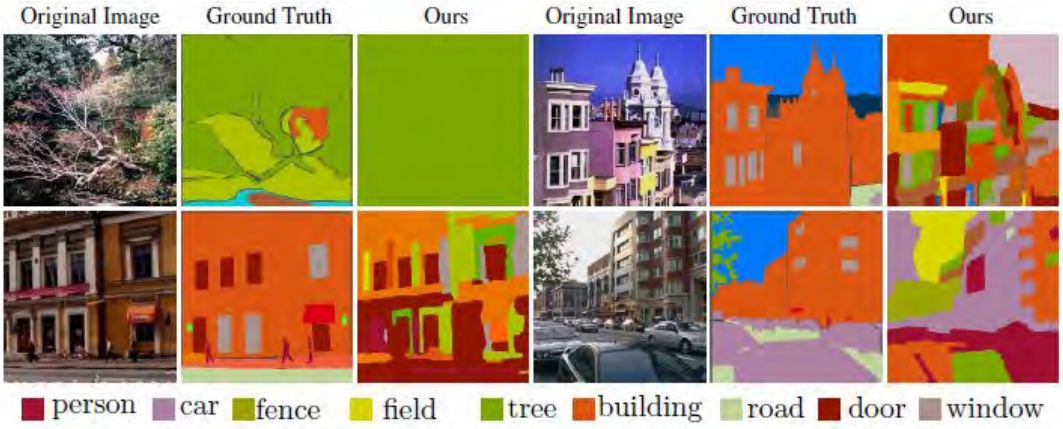

Figure 5. Failure cases when tags are predicted using a convolutional net at test time. **Best viewed in color.**

[11] G. Kim and E. P. Xing. On multiple foreground cosegmentation. In Proc. CVPR, 2012.

[12] L. Ladick´y, C. Russell, P. Kohli, and P. H. S. Torr. Graph Cut based Inference with Co-occurrence Statistics. In Proc. ECCV, 2010.

[13] J. Lafferty, A. McCallum, and F. Pereira. Conditional Random Fields: Probabilistic Models for segmenting and labeling sequence data. In Proc. ICML, 2001.

[14] C. Liu, J. Yuen, and A. Torralba. Nonparametric Scene Parsing via Label Transfer. PAMI, 2011.

[15] W. Luo, A. Schwing, and R. Urtasun. Latent structured active learning. In NIPS, 2013.

[16] L. Mukherjee, V. Singh, and J. Peng. Scale invariant cosegmentation for image groups. In Proc. CVPR, 2011.

[17] L. Mukherjee, V. Singh, J. Xu, and M. D. Collins. Analyzing the subspace structure of related images: Concurrent segmentation of image sets. In Proc. ECCV, 2012.

[18] M. Pandey and S. Lazebnik. Scene Recognition and Weakly Supervised Object Localization with Deformable Part-Based Models. In Proc. ICCV, 2011.

[19] A. Quattoni, S.Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden-state Conditional Random Fields. PAMI, 2007.

[20] C. Rother, V. Kolmogorov, and A. Blake. "GrabCut": inter active foreground extraction using iterated graph cuts. Siggraph, 2004.

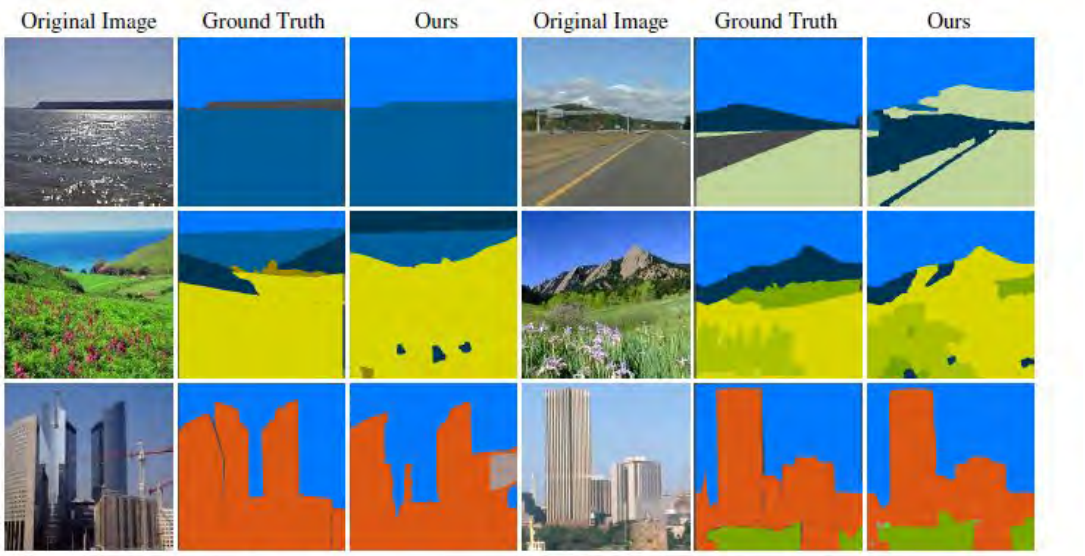

Figure 6. Sample results when ground truth tags are given at test time. **Best viewed in color.**

[21] C. Rother, T. P. Minka, A. Blake, and V. Kolmogorov. Cosegmentation of Image Pairs by Histogram Matching - Incorporating a Global Constraint into MRFs. In Proc. CVPR, 2006.

[22] A. G. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Distributed Message Passing for Large Scale Graphical Models. In Proc. CVPR, 2011.

[23] A. G. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Efficient Structured Prediction with Latent Variables for General Graphical Models. In Proc. ICML, 2012.

[24] B. Sriperumbudur and G. Lanckriet. On the convergence of the concave-convex

procedure. In Proc. NIPS, 2009.

[25] B. Taskar, C. Guestrin, and D. Koller. Max-Margin Markov Networks. In Proc. NIPS, 2003.

[26] J. Tighe and S. Lazebnik. Finding Things: Image Parsing with Regions and Per-Exemplar Detectors. In Proc. CVPR, 2013.

[27] J. Tighe and S. Lazebnik. Superparsing - Scalable Nonparametric Image Parsing with Superpixels. IJCV, 2013.

[28] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large Margin Methods for Structured and Interdependent Output Variables. JMLR, 2005.

[29] J. Verbeek and B. Triggs. Region classification with Markov field aspect models. In Proc. CVPR, 2007.

[30] A. Vezhnevets and J. M. Buhmann. Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In Proc. CVPR, 2010.

[31] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly supervised semantic segmentation with a multi image model. In Proc. ICCV, 2011.

[32] A. Vezhnevets, V. Ferrari, and J. M. Buhmann. Weakly Supervised Structured Output Learning for Semantic Segmentation. In Proc. CVPR, 2012.

[33] S. Vicente, V. Kolmogorov, and C. Rother. Cosegmentation Revisited: Models and Optimization. In Proc. ECCV, 2010.

[34] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In Proc. CVPR, 2012.

[35] C.-N. Yu and T. Joachims. Learning structural SVMs with latent variables. In Proc. ICML, 2009. 2,

[36] A. L. Yuille and A. Rangarajan. The Concave-Convex Procedure (CCCP). Neural Computation, 2003.