论文分享：
Causal Intervention for Weakly-Supervised Semantic Segmentation

Bowen XU
EECS, Peking University
Dec. 6th, 2020

CAUSALITY IN CLASSIFICATION TASKS

Grass, Yes!

Dog, Yes!

Dataset#1

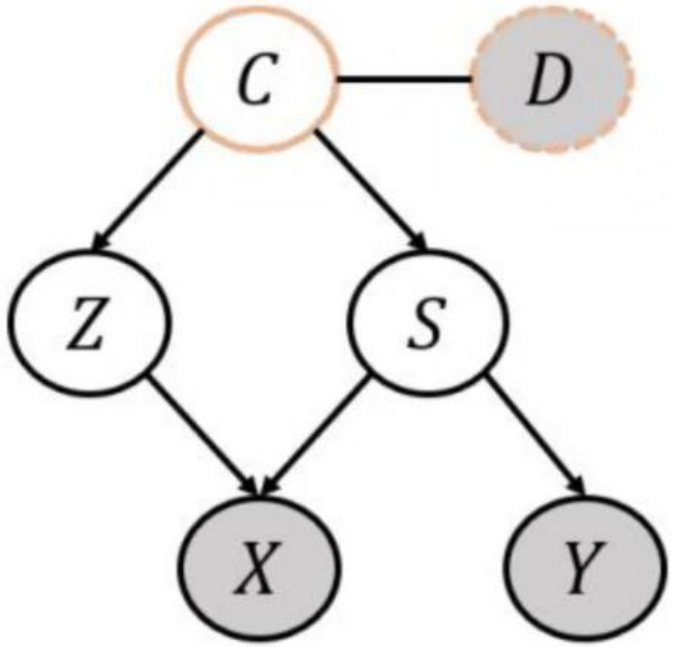
Grass, No!

Dog, Yes!

Dataset#2

MOTIVATION FOR LACIM

- Current supervised learning can learn spurious correlation.
  - X: picture
  - Y: label
  - Z: background(grass)
  - S: foreground(dog)
  - C: domain
  - D: index variable



Sun, Xinwei, et al. "Latent Causal Invariant Model." arXiv preprint arXiv:2011.02203 (2020).

基础信息

作者及论文信息

- 一作：张冬
  - 南京理工大学在读博士
  - 师从唐金辉教授
- 论文被NeurIPS 2020收录

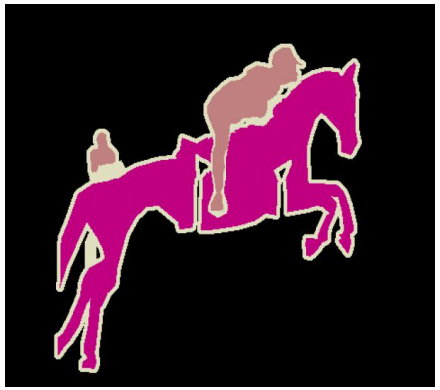
题目解析

- Weakly-Supervised Semantic Segmentation (WSSS) 弱监督语义分割
- 语义分割(Semantic segmentation)的目的：
  - 将每个图像像素分类为相应的语义类。
  - 常应用于自动驾驶和医学成像。
- 像素级别标签十分昂贵。例如：
  - 一个500×500的日常生活图像要花费1.5人·时。
- 弱监督语义分割：
  - 所谓“弱”，是指在实例级别甚至图像级别打标签，其代价更低。

Supervised:

Data(image)

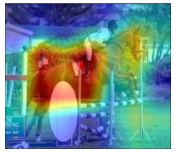
Label (ground-truth/mask)



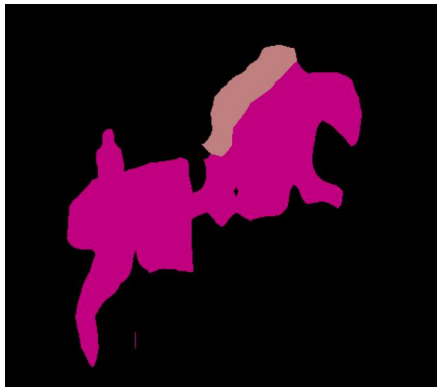
Weakly-Supervised:

Label

“hourse”
“man”
……

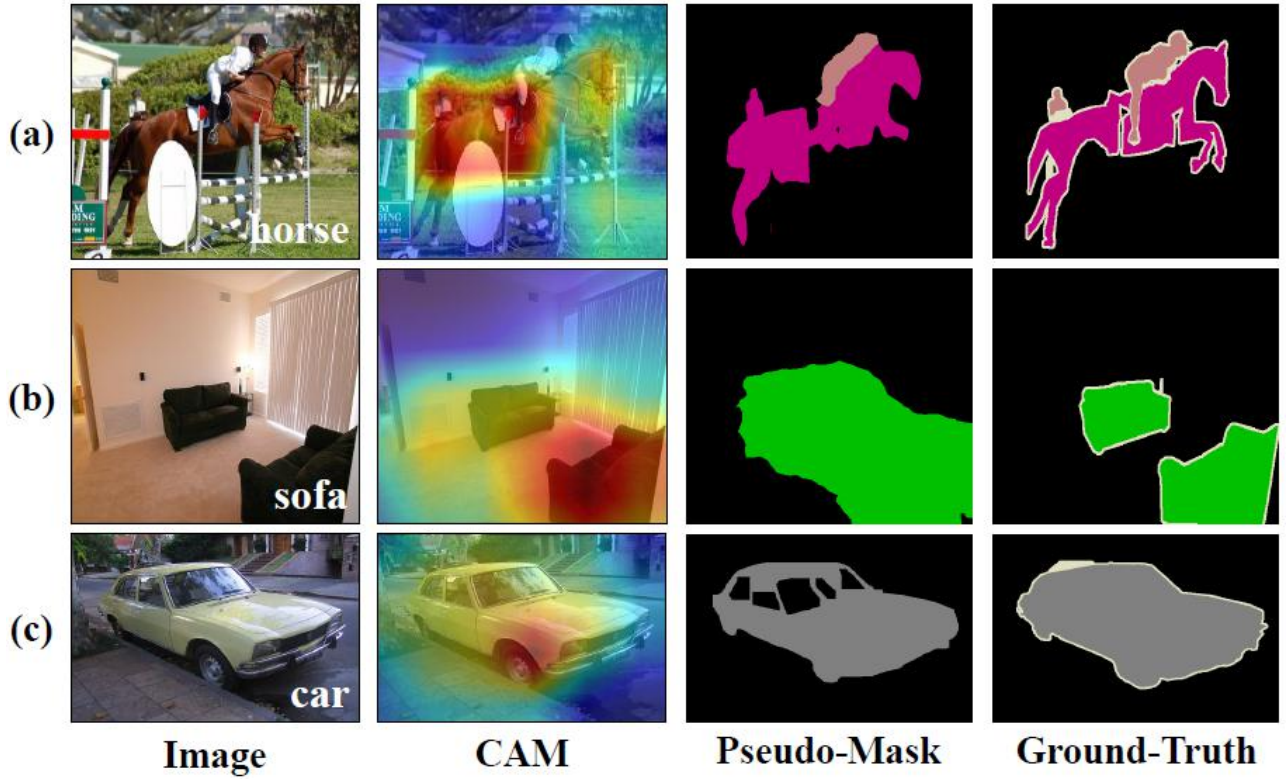


Pseudo-label (pseudo-mask)



论文背景

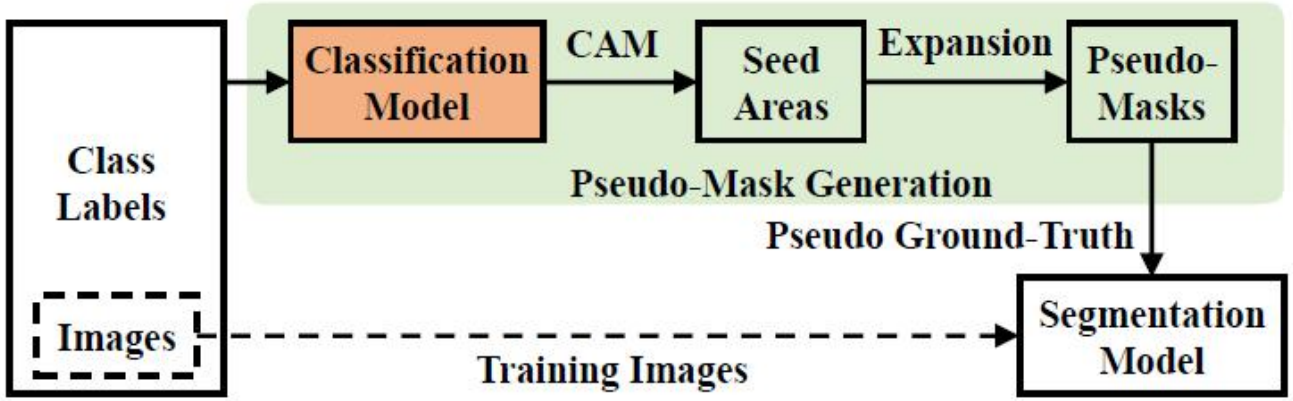
- 语义分割：自动驾驶/医学成像；人工标记成本高昂。
- 弱监督语义分割中存在虚假关联问题。



相关工作

- **弱监督语义分割 WSSS**

  - 主流的图像级类标签的WSSS流程，主要包括pseudo-mask生成和分割模型训练两个步骤。
  - 如下图所示，生成pseudo-mask的主要流程：
    1. 多标签图像分类模型获取图像的类响应激活图(Class Activation Map, CAM)作为种子区域(Seed Area)；
    2. 通过计算像素之间的语义相似性对种子区域进行扩张(Expansion)得到图像的伪标签(Pseudo-Mask)；
    3. 使用伪标签训练一个全监督的语义分割模型，并在训练好的模型上对验证/测试集进行预测。
  - 更精确的pseudo-mask → 训练更好的分割模型。



WSSS的一般流程

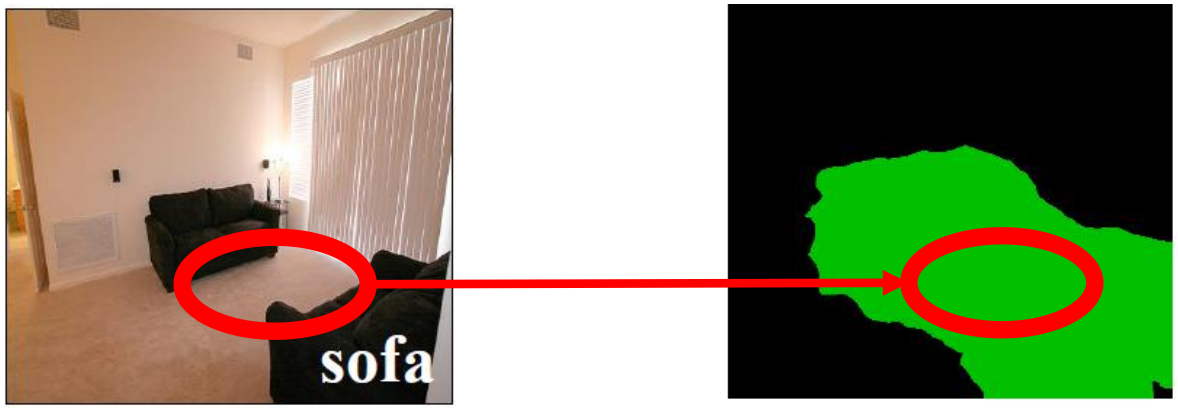
相关工作

- **视觉上下文 Visual Context**

  - 视觉上下文对识别至关重要。
  - 大多数WSSS模型通过在膨胀卷积的帮助下扩大感受野，隐式地在骨干网络中使用上下文。
  - 在本文中，作者提出利用因果干预，明确地使用了上下文：上下文调整(Context Adjustment, CONTA)。

解决的科学问题

- 视觉分类模型做的是计算$P(Y|X)$，即给定图像判断标签。
- 沙发总是在地板上，因此在判断一张图像是否是沙发时，
- **往往会把地板（像素）与“沙发”（标签）建立关联，**
- 由此产生错误的seed areas，导致错误的分割区域pseudo-mask。

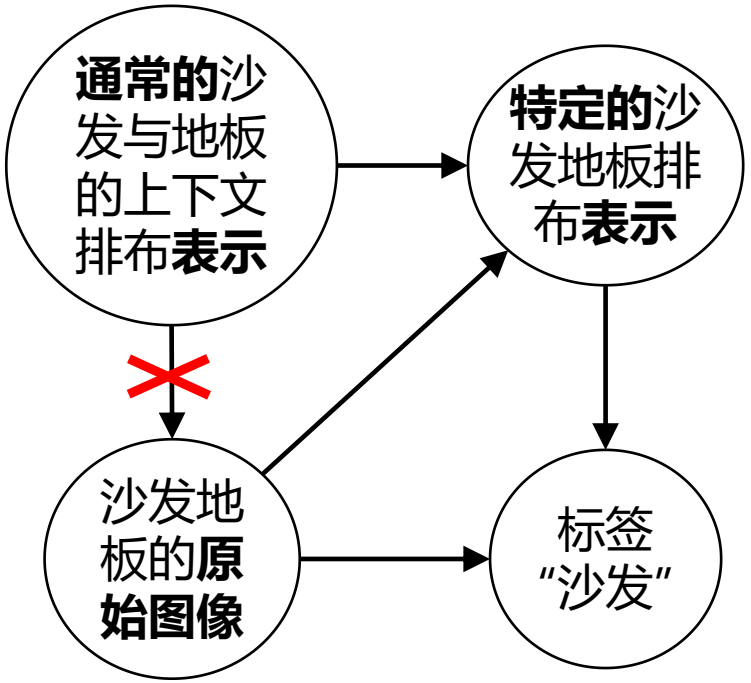


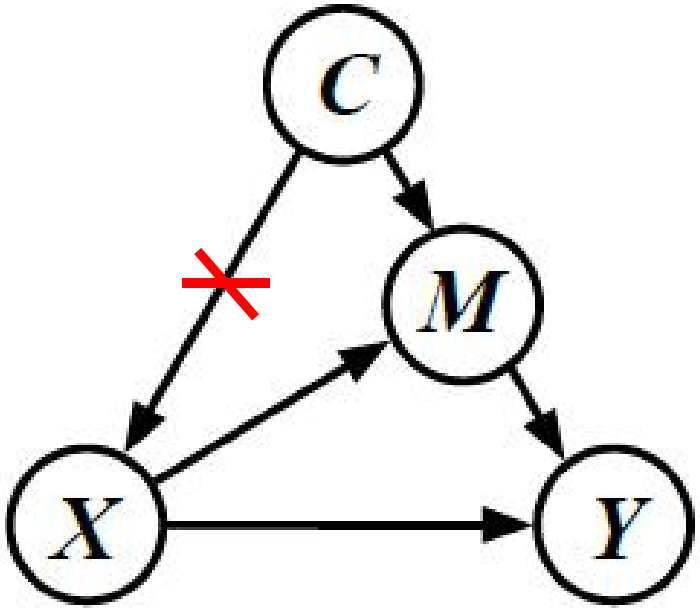
原始图像　　传统方法生成的pseudo-mask

1. 弱监督语义分割中的因果关系是什么？
2. 如何借助因果干预的手段消除像素与标签之间的虚假关联？

解决的科学问题

- 上下文先验作为混杂因子：沙发与地板总是一起出现，因此地板也被当作标签“沙发”的因（后门路径$X \leftarrow C \rightarrow M \rightarrow Y$）。
- 如果能使沙发处在任何上下文下，即$do(X)$，可得到$P(Y|do(X))$，则可消除X中“地板”像素与“沙发”标签的虚假关联。
- 知道哪些像素是标签的真正的因，获得更好的pseudo-mask，帮助分割模型训练取得更好效果。



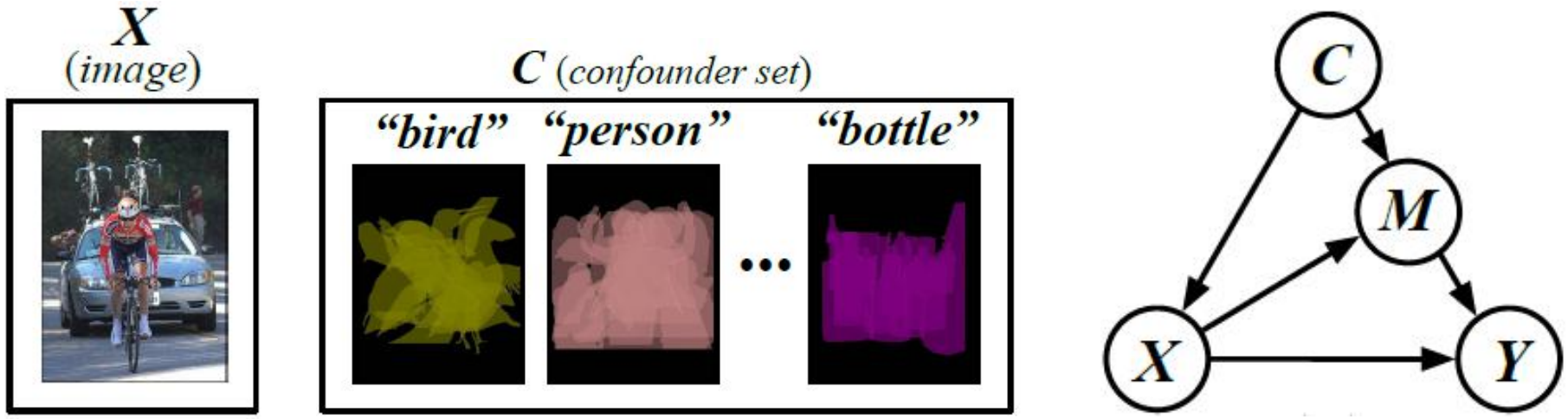


解决问题的思路

- **因果图的含义及结点的表征**

$C \rightarrow X$:

- “上下文先验”在计算机视觉中的一般的含义：视觉场景中物体之间的关系。
- C告诉我们图像中一般哪里放“车”，哪里放“马路”，哪里放“楼房”。
- 尽管为$C \rightarrow X$构建一个生成模型对复杂场景而言是极富挑战性的，我们在因果介入中避免了这一过程。

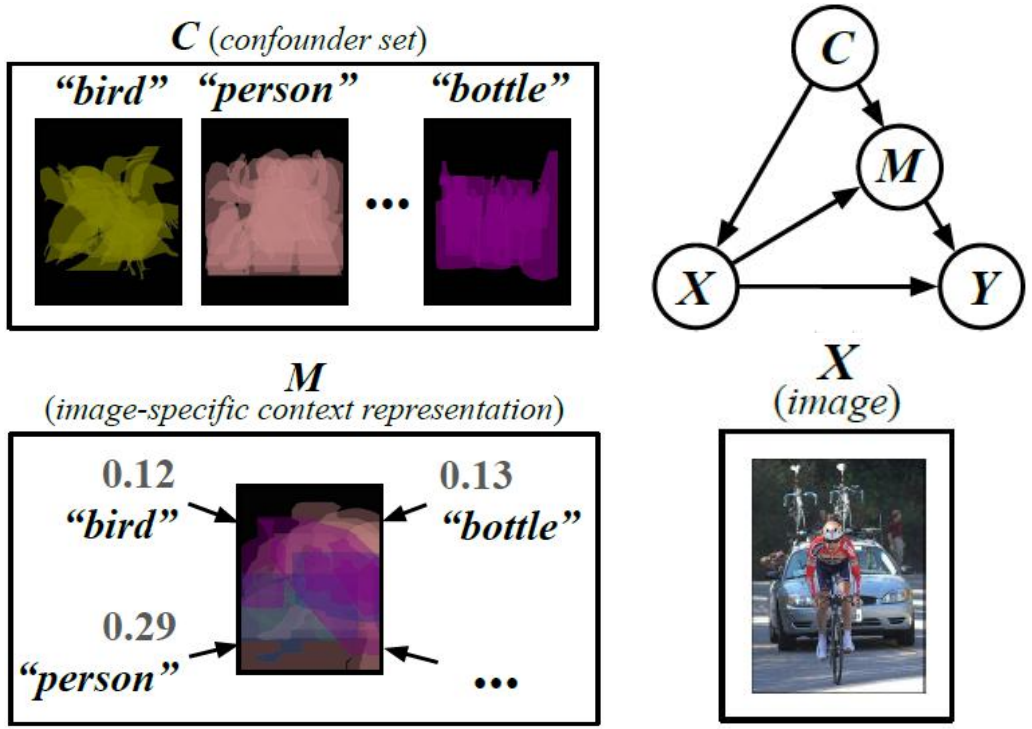


解决问题的思路

- **因果图的含义及结点的表征**

$C \rightarrow M \leftarrow X$：

- M是特定图像的表征，由C中的上下文模板和特定X得到。
- 例如，一张车的图像的模板是典型场景（背景）中“车”（前景）的典型形状和典型位置。

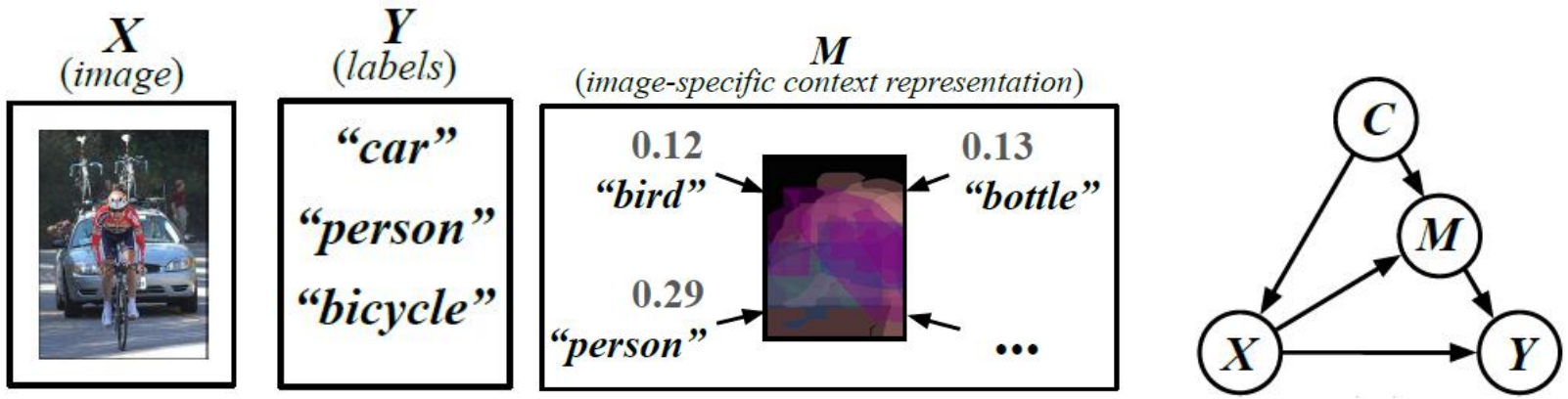


解决问题的思路

- **因果图的含义及结点的表征**

$X \rightarrow Y \leftarrow M$：

- 通用的C不能直接影响图像的标签Y。除了$X \rightarrow Y$，Y也受到特定X调制的M的影响。
- $M \rightarrow Y$表示一种显而易见的因果关系：图像的上下文构造影响了图像的标签。
- 即使我们不显式地将M作为分类模型的输入，$M \rightarrow Y$也依然存在：
  - 当训练图像时，视觉上下文在CNN的高层中会涌现出来，这本质上是作为现代视觉检测（例如Fast R-CNN和SSD）中高度依赖于上下文的骨干(backbone)网络的特征图。



解决问题的思路

- 如何进行因果干预

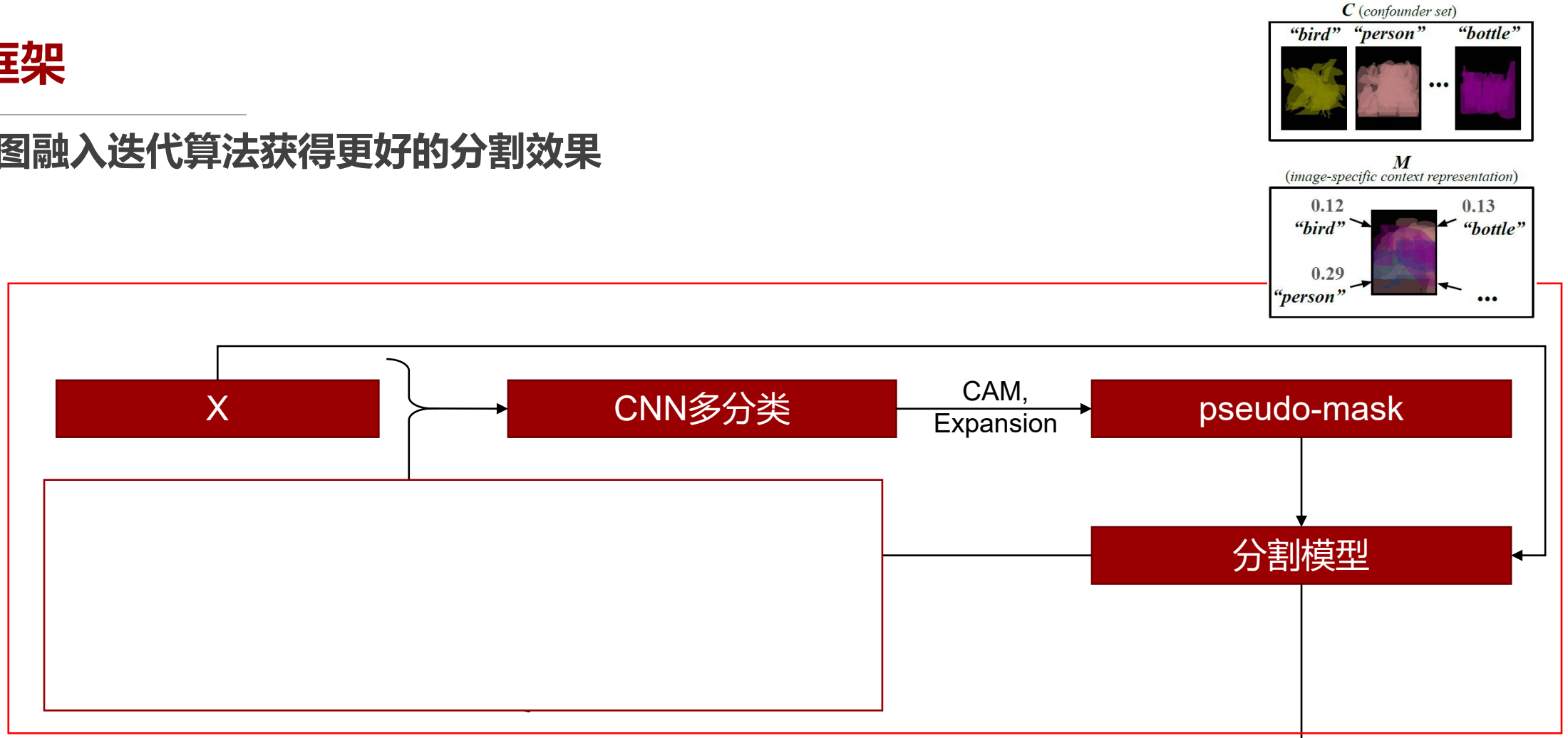
  - 由于“物理”干预——收集在任何在上下文下的对象——是不可能的，因此我们用后门调整(backdoor adjustment)来“虚拟地”实现$P(Y|do(X))$。
  - 关键思路： C分成若干块$C = \{c_1, c_2, \dots, c_n\}$

$$P(Y|do(X)) = \sum_c P(Y|X, M = f(X, c))P(c)$$

  - 其中n是数据集中类的大小，且$c \in \mathbb{R}^{h \times w}$对应的是第$i$类图像的尺寸为$h \times w$的平均mask。令$P(c) = {}^1/_n$。
  - 在传统的pseudo-mask生成过程中，学习CNN分类模型，以最大化$P(Y|X)$，存在标签与像素的虚假关联；而此处，加入因果干预后，最大化$P(Y|do(X))$，避免了虚假关联。

算法框架

- 因果图融入迭代算法获得更好的分割效果



$$P(Y|do(X)) = \sum_{c} P(Y|X, M = f(X, c))P(c)$$

算法框架

- 因果图融入迭代算法获得更好的分割效果

$$P(Y|do(X)) = \sum_c P(Y|X, M = f(X,c))P(c) \approx P(Y|X, M = \sum_c f(X,c)P(c))$$

$$P(\mathrm{Y}|do(X)) = \mathrm{cnn}\left(X, M_t = \sum_{i=1}^{n} softmax\left(\frac{\left(\boldsymbol{W}_1 \mathrm{seg}(X)\right)^T (\boldsymbol{W}_2 c_i)}{\sqrt{n}}\right) c_i P(c_i)\right)$$

用神经网络拟合概率分布P。

$$s_i = \mathrm{cnn}\left(X, M_t; \theta_t^i\right)$$

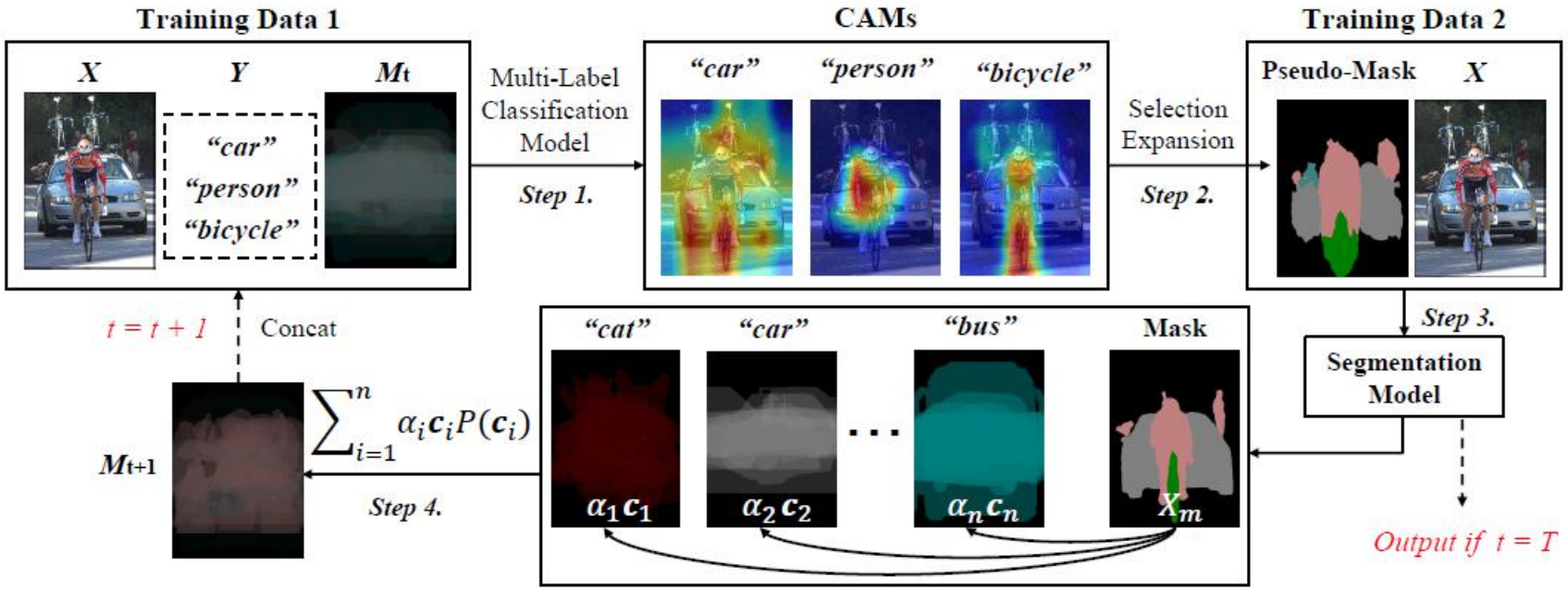
$$P_{joint}(Y|do(X); \Theta_t) = \prod_{i=1}^{n}\left[\mathbb{I}_{i\in Y}\frac{1}{1+\exp(-s_i)} + \mathbb{I}_{i\notin Y}\frac{1}{1+\exp(s_i)}\right]$$

优化目标：

$$\underset{X_m=\mathrm{seg}(X)}{\mathrm{argmax}}\ P_{joint}(Y|do(X); \Theta_t)$$

详细介绍

- **Step 1.** 图像分类。初始时 M_t 为空集，通过最大化 $P(Y|do(X))$ 学习卷积神经网络的分类模型。
- **Step 2.** Pseudo-Mask生成。按照传统方法生成pseudo-masks。
- **Step 3.** 分割模型训练。用pseudo-mask作为标签训练分割模型。
- **Step 4.** 计算M_{t+1}。



详细介绍

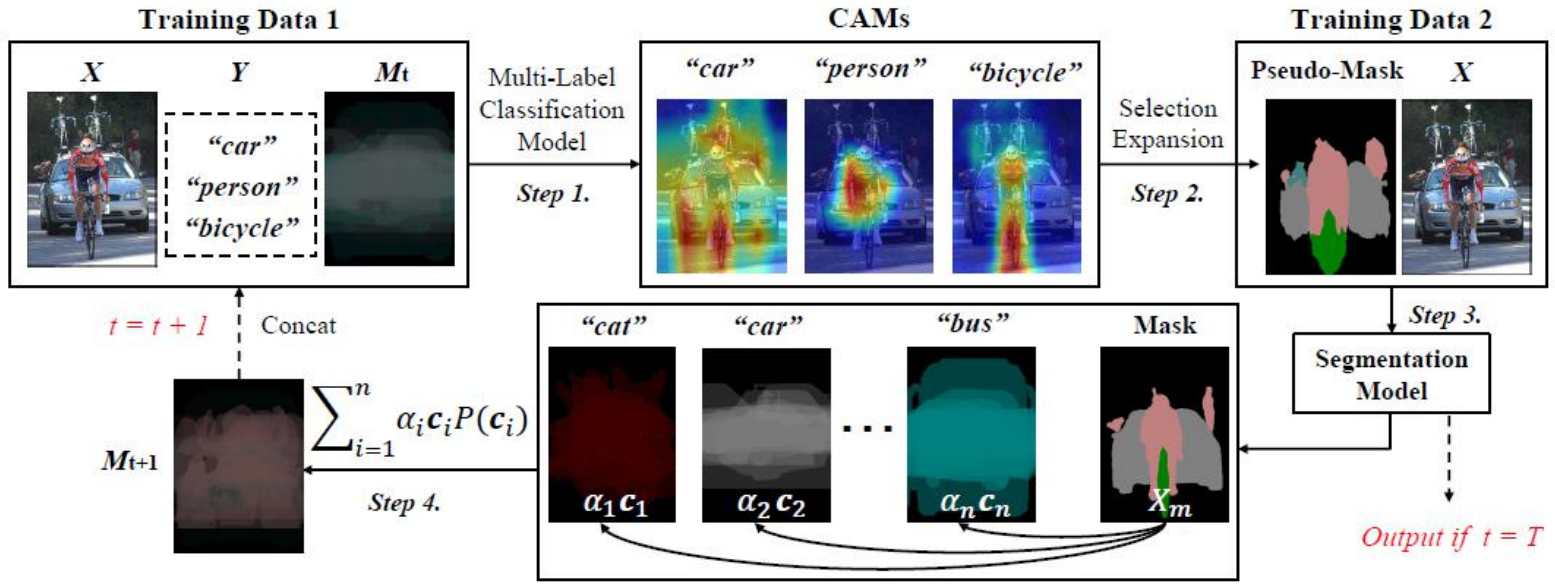
- **Step 1. 图像分类**

$$s_i = f\left(X, M_t; \theta_t^i\right)$$

f是由类间权重共享的卷积网络组成。通道方向上拼接的特征图$[X, M_t]$，后接特定类的全连接网络（最后一层基于全局平均池化）。

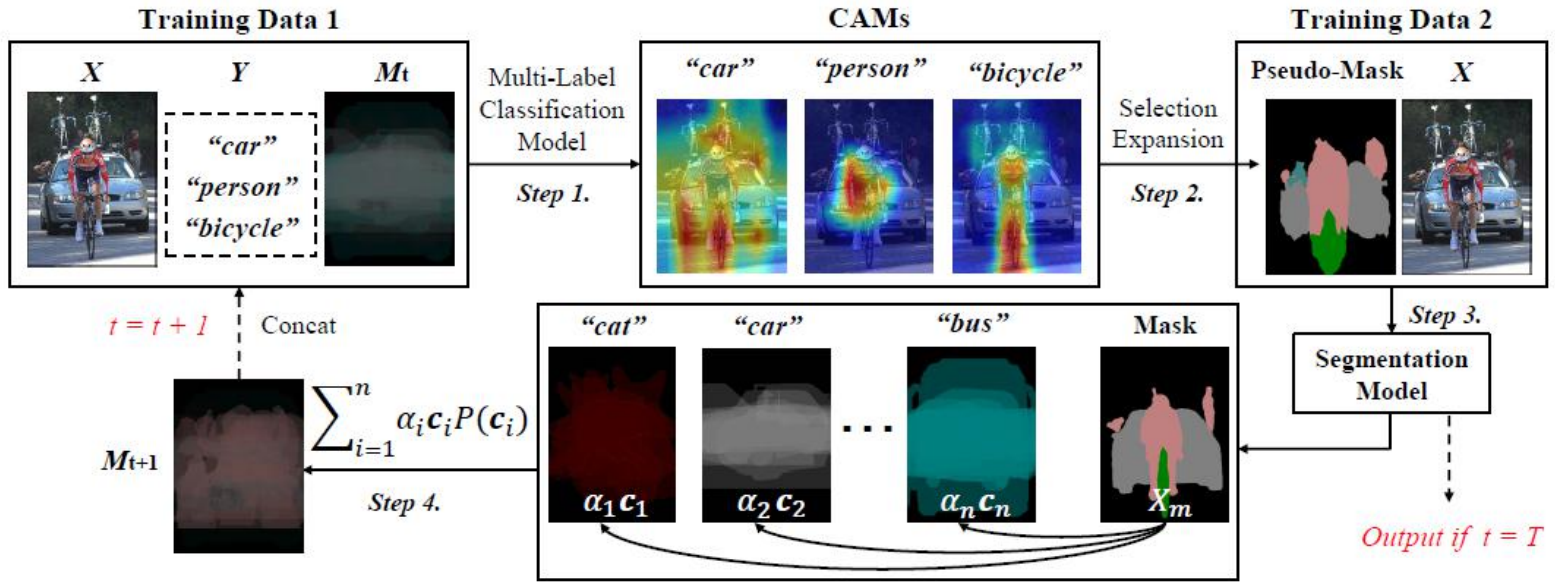
$$P(Y|do(X); \Theta_t) = \prod_{i=1}^{n}\left[\mathbb{I}_{i\in Y}\frac{1}{1+\exp(-s_i)} + \mathbb{I}_{i\notin Y}\frac{1}{1+\exp(s_i)}\right]$$

其中Θ_t是神经网络的参数，$\mathbb{I}$是0/1指示器。上式是所有n个类的联合概率，鼓励$i \in Y$而惩罚$i \notin Y$。



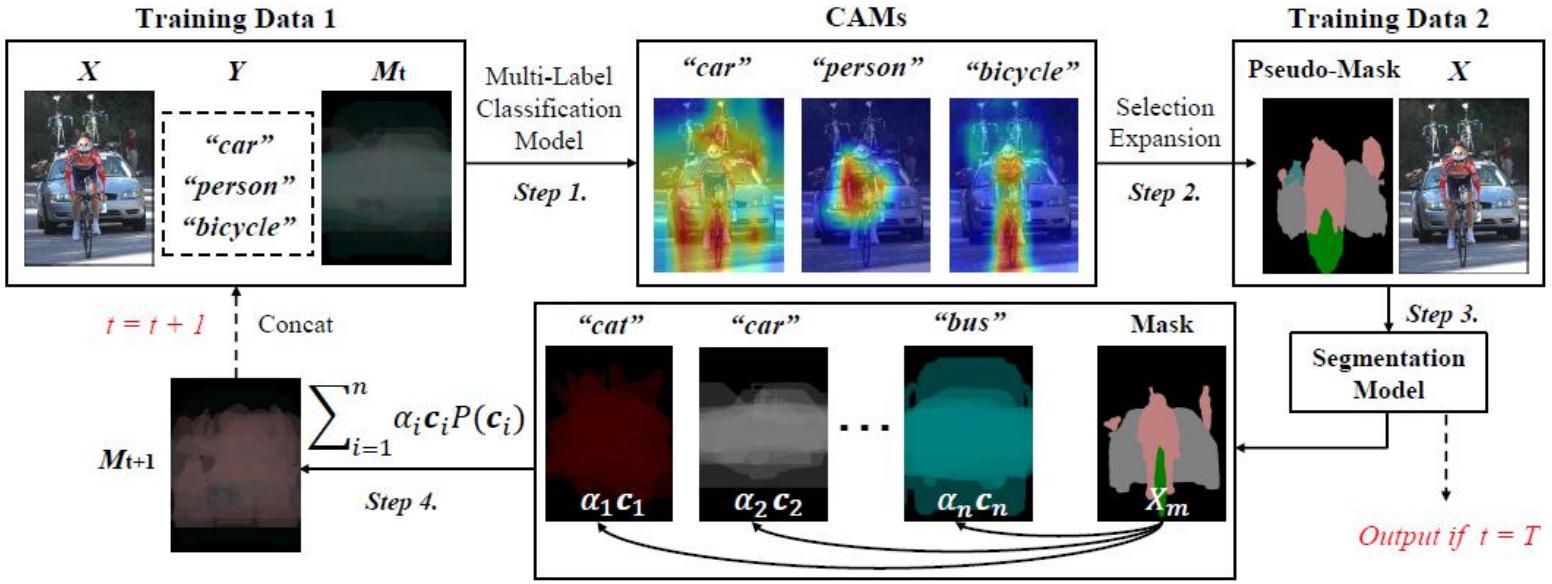
详细介绍

- **Step 2. Pseudo-Mask生成**
  - 对于每一幅图像，我们可以使用上面训练好的分类器计算出一组特定类的CAMs。
  - 然后，我们遵循传统的两个后处理步骤:
    - 1)为种子区域选择激活的CAM区域（服从某个阈值）；
    - 2) 我们将它们扩展为最终的pseudo-masks。



详细介绍

- **Step 3. 分割模型训练**
  - 每个pseudo-masks作为pseudo ground-truth用于训练任何标准的监督语义分割模型。如果$t = T$则模型收敛，否则，它的分割mask可以看作是pseudo-mask平滑的一个额外的后处理步骤。为了和其他WSSS公平比较，采用经典的DeepLab-v2作为监督语义分割模型。如果采用更高级的方法，性能应该还会提升。



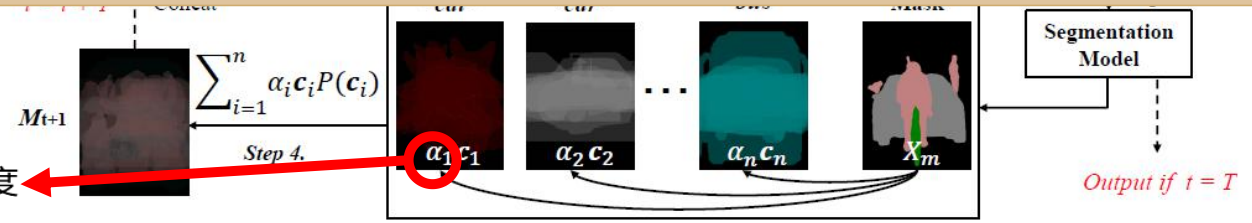
详细介绍

- **Step 4. 计算M_{t+1}**

  - 从分割模型中收集每张图片的预测的mask X_m。每一个特定类的混淆因素集合C中的c是每一个对应类的X_m的平均mask，并且reshape为$hw \times 1$的向量。
  - 对所有的n个类，分别计算网络前向传播的开销是很昂贵的。幸运的是，在实际假设下，我们可以采用标准化加权几何平均值(Normalized Weighted Geometric Mean)将外面的求和$\sum_c P(\cdot)$移动到特征的层次$\sum_c P(Y|X, M = f(X, c))P(c) \approx P(Y|X, M = \sum_c f(X, c)P(c))$，因此我们只需要前向传播一次，我们有：

$$M_{t+1} = \sum_{i=1}^{n} \alpha_i c_i P(c_i), \qquad \alpha_i = softmax\left(\frac{(\boldsymbol{W}_1 X_m)^T (\boldsymbol{W}_2 c_i)}{\sqrt{n}}\right)$$

作者回复：如果不把Σ放在P的内部的话，那对于具有N个类别的数据集，则需要forward N次，尤其当N的数量比较大的时候这个计算量就会很大。这点我建议你看一下我们组另外一个同学的论文[Yang X, Zhang H, Cai J. Deconfounded image captioning: A causal retrospect[J]. arXiv preprint arXiv:2003.03923, 2020.]，他在这个论文里面对这个有比较详细的说明。



当前mask X_m与上下文先验的相似度

实验

语义分割评价指标：

- mIoU (Mean Intersection over Union, 均交并比)

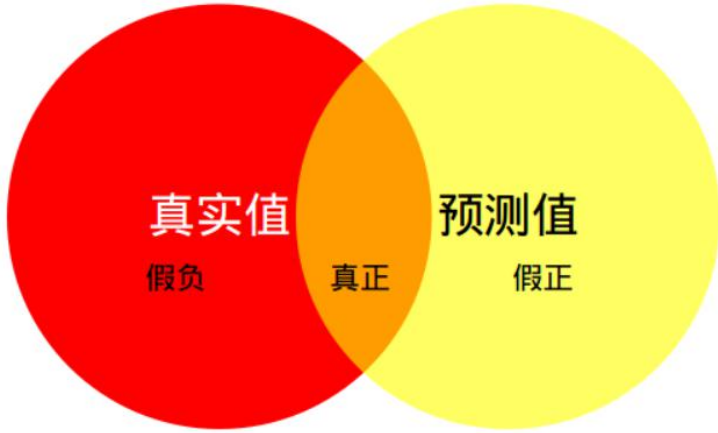
数据集：

- PASCAL VOC 2012：21个类，数据量训练集1464，验证集1449，测试集1456。训练时用了额外的10582张图片。
- MS-COCO：81个类，数据量训练集80k，验证集40k。

对比模型：

- 种子区域生成模型 SEAM
- 种子区域扩张模型 IRNet, DSRG, SEC

直观理解mIoU



$$mIoU = \frac{1}{n}\sum_{i=1}^{n}\frac{P \cap G}{P \cup G}$$

P： Prediction 预测值
G： Ground Truth 真实值
n： 类别数量

实验结果

在现有的方法上加入CONTA后，效果均有提升

Method	Backbone	CAM	Pseudo-Mask	Seg. Mask
SEC [26]	VGG-16	46.5	53.4	50.7
+ CONTA	VGG-16	47.9$_{+1.4}$	55.7$_{+2.3}$	53.2$_{+2.5}$
SEAM* [63]	ResNet-38	55.1	63.1	64.3
+ CONTA	ResNet-38	**56.2**$_{+1.1}$	65.4$_{+2.3}$	**66.1**$_{+1.8}$
IRNet* [1]	ResNet-50	48.3	65.9	63.0
+ CONTA	ResNet-50	48.8$_{+0.5}$	**67.9**$_{+2.0}$	65.3$_{+2.3}$
DSRG [22]	ResNet-101	47.3	62.7	61.4
+ CONTA	ResNet-101	48.0$_{+0.7}$	64.0$_{+1.3}$	62.8$_{+1.4}$

Table 2: Different baselines+CONTA on PASCAL VOC 2012 [14] dataset in mIoU (%). "*" denotes our re-implemented results. "Seg. Mask" refers to the segmentation mask on the *val* set.

实验结果

在现有的方法上加入CONTA后，取得了SOTA的表现。

Method	Backbone	*val*	*test*
AffinityNet [2]	ResNet-38	61.7	63.7
RRM [73]	ResNet-38	62.6	62.9
SSDD [52]	ResNet-38	**64.9**	65.5
SEAM [63]	ResNet-38	64.5	**65.7**
IRNet [1]	ResNet-50	63.5	64.8
IRNet+CONTA	ResNet-50	65.3	66.1
SEAM+CONTA	ResNet-38	**66.1**	**66.7**

(a) PASCAL VOC 2012 [14].

Method	Backbone	*val*
BFBP [50]	VGG-16	20.4
SEC [26]	VGG-16	22.4
SEAM* [63]	ResNet-38	31.9
IRNet* [1]	ResNet-50	**32.6**
SEC+CONTA	VGG-16	23.7
SEAM+CONTA	ResNet-38	32.8
IRNet+CONTA	ResNet-50	**33.4**

(b) MS-COCO [35].

Table 3: Comparison with state-of-the-arts in mIoU (%). "*" denotes our re-implemented results. The **best** and **second best** performance under each set are marked with corresponding formats.

实验结果

加入CONTA后，边界分割更加精细了。

但也存在失败情况。作者的解释：“一个可能的解释是，分割mask直接从8倍降采样的特征图获得，所以一些复杂轮廓的对象不能准确地划定。”

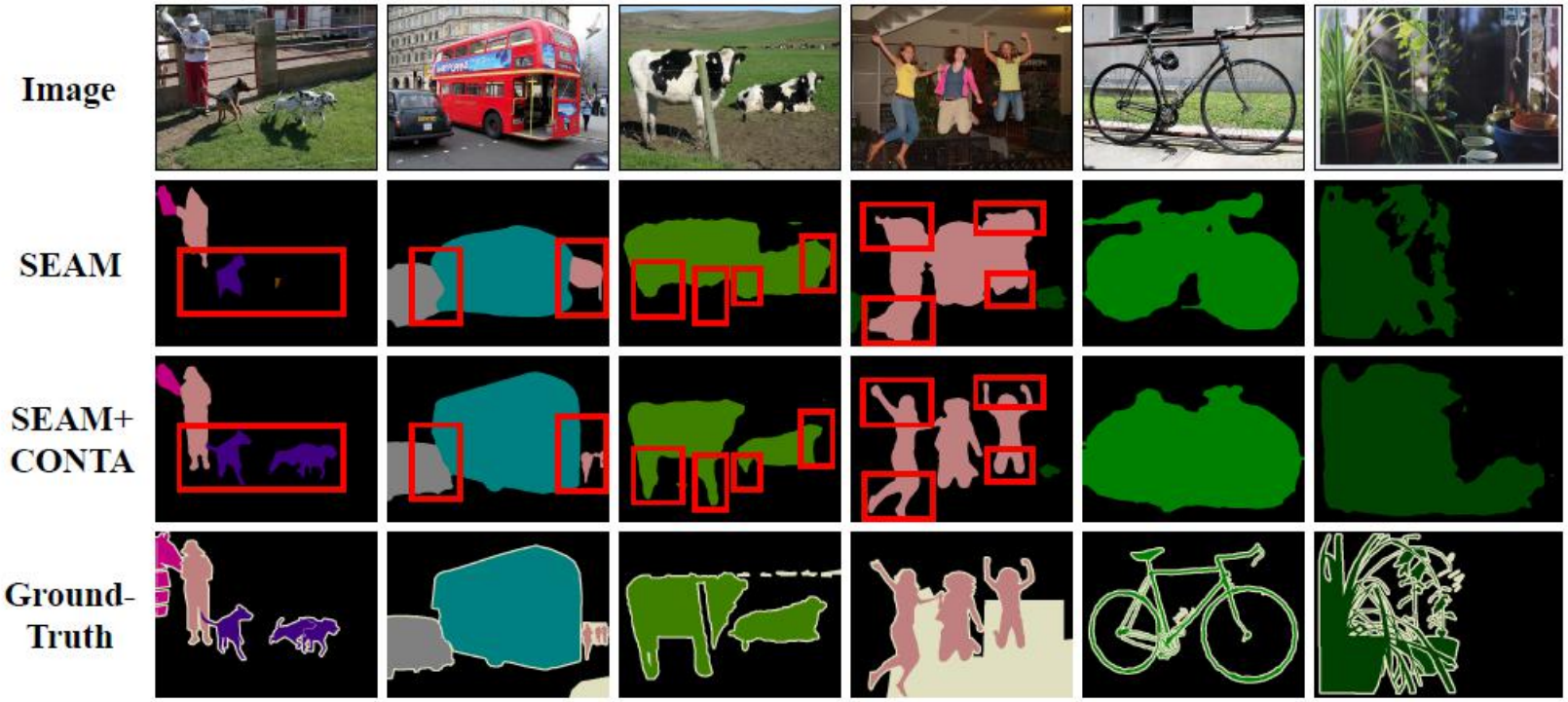
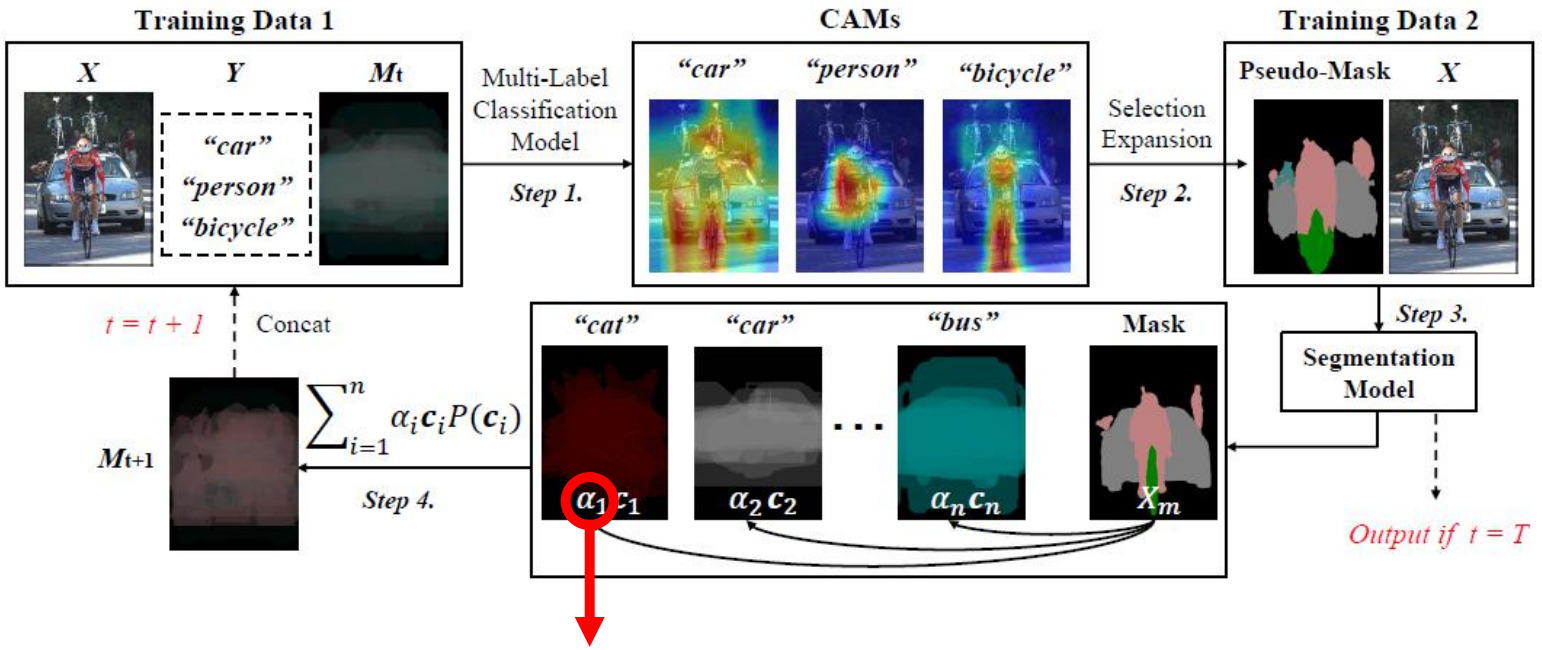


Figure 6: Visualization of segmentation masks, the last two columns show two failure cases (dataset: PASCAL VOC 2012 [14]). The red rectangle highlights the better areas for SEAM+CONTA.

讨论

本文的启发：因果图与神经网络的结合范式。结构方程可以间接地用神经网络表示，因果图中的结点可以用张量表示。建立出合理的因果图，这种范式可以迁移到其他问题上。

用某个类的所有mask的平均作为该类的上下文先验?
用相似度作为权重对上下文先验做线性组合作为特定图像的上下文表示?

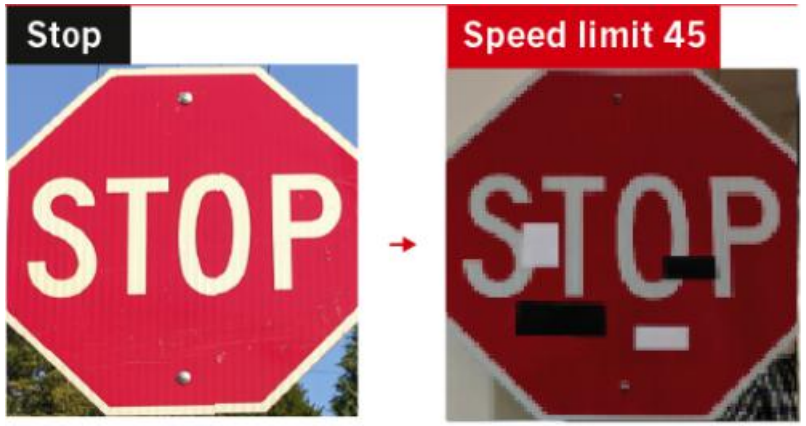


当前mask X_m与上下文先验的相似度

未来方向

- 作者也指出了，混杂因子的集合的近似是有问题的。
- **未来方向**：
  - 1) 开发更先进的混杂因子集合的发现方法；
  - 2) 将专家知识融入混杂因子
- 对样本对抗是否有帮助？

加入噪声后“狗”识别成“鸵鸟”



加入色块和调整色调之后，“停止”标志识别为“限速45”

文献/资源列表

- 论文获取：https://papers.nips.cc/paper/2020/hash/07211688a0869d995947a8fb11b215d6-Abstract.html
- 源码获取：https://github.com/ZHANGDONG-NJUST/CONTA
- 数据集下载：
  - Pascal VOC Dataset
    http://host.robots.ox.ac.uk/pascal/VOC/voc2012/#devkit
    （镜像）https://pjreddie.com/projects/pascal-voc-dataset-mirror/
  - MS-COCO (2014) https://cocodataset.org/
- 作者的知乎文章介绍：https://zhuanlan.zhihu.com/p/260967655
- CAM (Classification Activation Map, 类响应激活图)：https://arxiv.org/pdf/1512.04150.pdf
- Deconfounded image captioning: A causal retrospect: https://arxiv.org/pdf/2003.03923.pdf

The End

Thanks