Semi-supervision, weak supervision and few shot

论文列表

- Semi-Supervised Semantic Image Segmentation with Self-correcting Networks
- Rethinking the Route Towards Weakly Supervised Object Localization
- Few-Shot Object Detection with Attention-RPN and Multi-Relation Detector

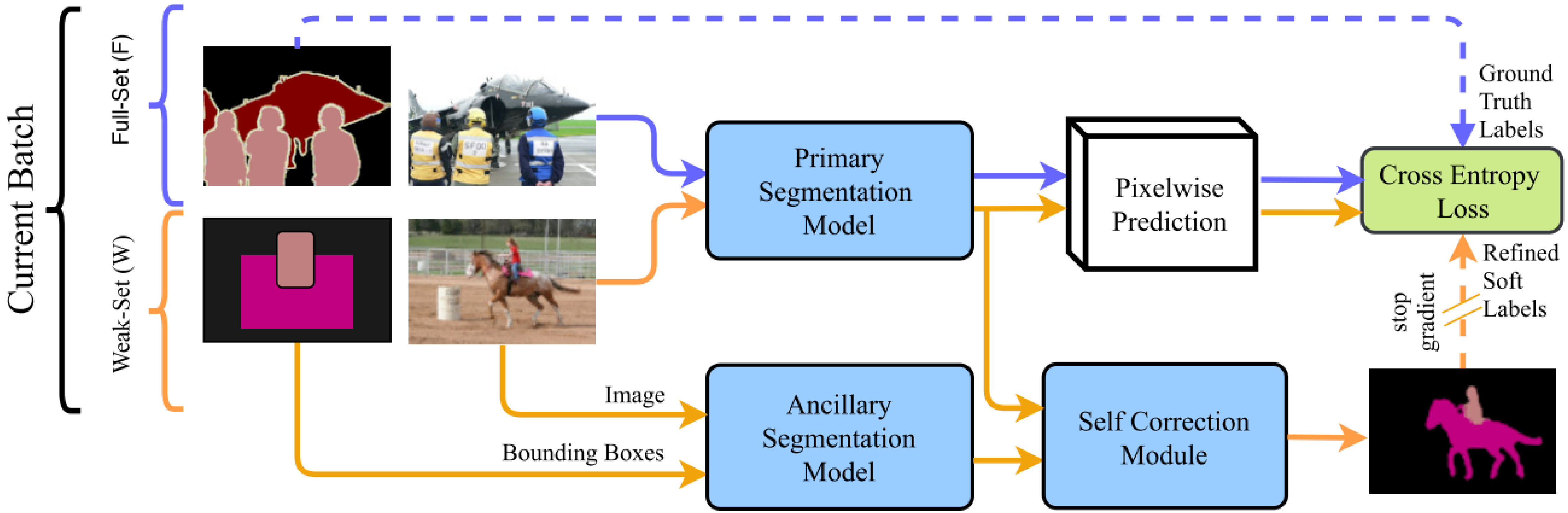
Semi-Supervised Semantic Image Segmentation with Self-correcting Networks

- 高质量大型语义分割图像数据集→费时费力
  - 比框（bounding box，bbox）标注~8×
  - 比分类（classification） 标注~78×
- 半监督或者弱监督很有意义
- 本文方法的适用场景：
  - 一小部分全监督数据集（F set）,包括分割标注和框标注
  - 一个弱监督数据集（W set）,只包括框标注
- 和主流SOTA方法的主要不同：
  - 目前方法依赖人工设计的公式来推断bbox内的分割伪标签
  - 用一个辅助CNN（ancillary CNN）代替手工制作的规则，提供弱监督集的bbox的对象的概率分割伪标签。（*关键*）
- 其他创新点：
  - 在训练过程中，使用自校正模型（self-correcting model）来校正辅助CNN的输出和初级分割模型之间的不匹配

网络结构

- **初级分割模型**（Primary segmentation model, P model）:给定一张图像，给出分割结果。也是最终测试的分割模型
- **辅助分割模型**（Ancillary segmentation model, A model）:给定一张图像和一个bbox, 给出分割结果。
  - 产生在W set上的初始伪标签，辅助P model的训练。
- **自校正模型**（Self-correction model, S model）: 矫正A model和P model在W set 上的分割结果

网络结构

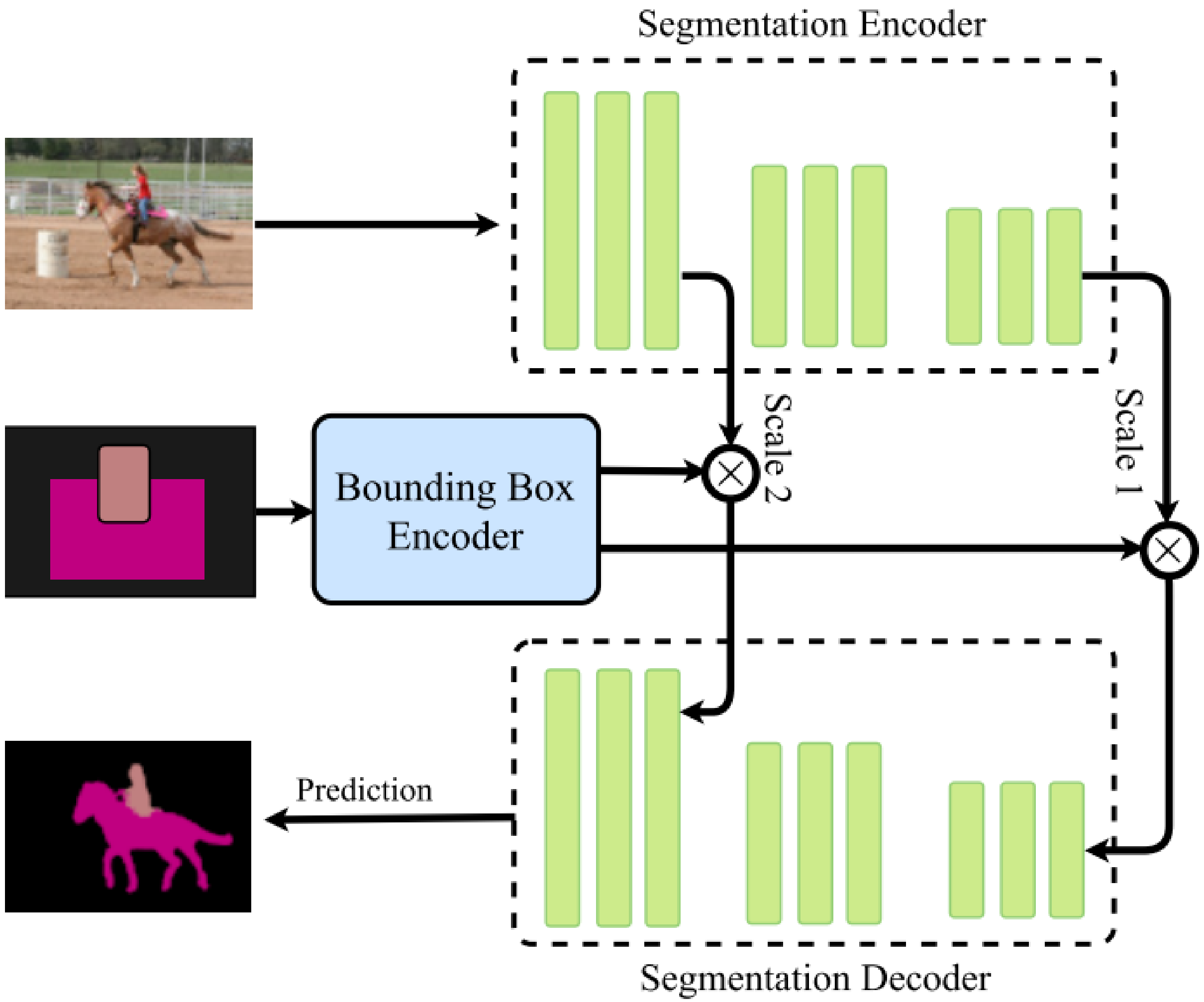


Current Batch
Full-Set (F)
Weak-Set (W)
Ground
Truth
Labels
Primary
Segmentation
Model
Pixelwise
Prediction
Cross Entropy
Loss
Refined
Soft
Labels
stop
gradient
Image
Bounding Boxes
Ancillary
Segmentation
Model
Self Correction
Module

辅助分割模型（A model）

- 弱监督训练的关键在于给定bbox，推断W set上的伪标签。
  - 现有方法主要依赖人为定义的规则（GrabCut）或者迭代思想（ EM迭代等）
  - 存在①bbox信息没有直接用来提取伪标签②可能不是最优解③存在多个bbox重叠时，分割结果会混淆
- 辅助分割模型直接输入图像和对应的bbox，得到概率分割伪标签。
  - 在F set上训练，在W set上得到W的伪标签。
  - 设计时，平行设计了两个encoder结构，分别输入原图和bbox，这样可以采用P model的encoder参数初始化其中一个encoder。

辅助分割模型（A model）

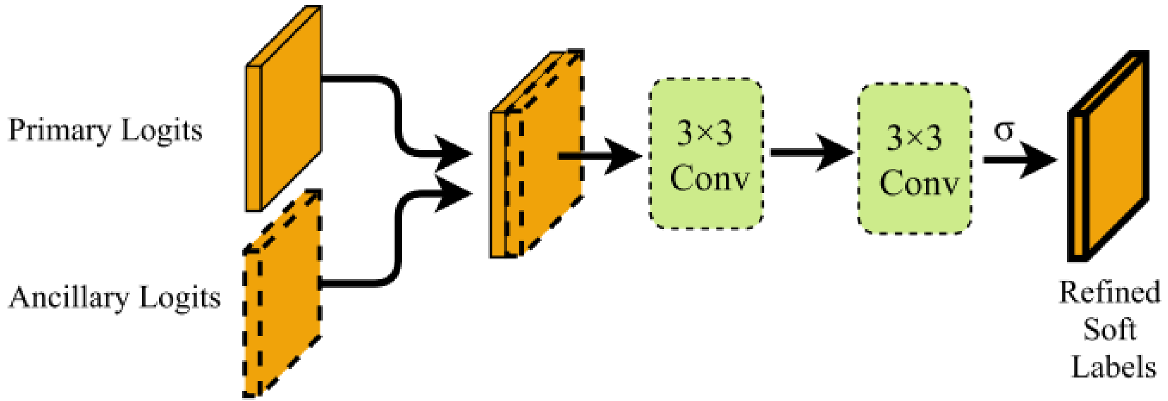


自校正模型（S model）

- 无矫正模型
  - F set 用标注， W set用A model产生的伪标签训练P model
- 线性校正模型

$$\min_{q} \mathrm{KL}(q(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{b})||p(\boldsymbol{y}|\boldsymbol{x}))+\alpha\mathrm{KL}(q(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{b})||p_{anc}(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{b}))$$

- 自校正模型
  - 一个轻量卷积网络，输入A model和P model的概率分割结果，输出监督P model在W set上训练的概率分割伪标签。
  - 在F set上训练



整体损失函数

$$\max_{\boldsymbol{\phi},\boldsymbol{\lambda}} \quad \sum_{\mathcal{F}} \log p(\boldsymbol{y}^{(f)}|\boldsymbol{x}^{(f)};\boldsymbol{\phi}) + \tag{6}$$

$$\sum_{\mathcal{W}} \sum_{\boldsymbol{y}} q_{conv}(\boldsymbol{y}|\boldsymbol{x}^{(w)},\boldsymbol{b}^{(w)};\boldsymbol{\lambda}) \log p(\boldsymbol{y}|\boldsymbol{x}^{(w)};\boldsymbol{\phi}) +$$

$$\sum_{\mathcal{F}} \log q_{conv}(\boldsymbol{y}^{(f)}|\boldsymbol{x}^{(f)},\boldsymbol{b}^{(f)};\boldsymbol{\lambda}),$$

实验

- PASCAL VOC 2012
  - 1464 training, 1449 validation, and 1456 test
- Cityscapes
  - 2975 training, 500 validation, and 1525 test images

实验

Data Split		Method	Val	Test
F	W			
1464	9118	No Self-Corr.	80.34	81.61
1464	9118	Lin. Self-Corr.	81.35	81.97
1464	9118	Conv. Self-Corr.	**82.33**	**82.72**
1464	9118	EM-fixed Ours [41]	79.25	-
10582	-	Vanilla DeepLabv3+ [9]	81.21	-
1464	9118	BoxSup-MCG [12]	63.5	-
1464	9118	EM-fixed [41]	65.1	-
1464	9118	M ∩ G+ [26]	65.8	-
1464	9118	FickleNet [30]	65.8	-
1464	9118	Song *et al.* [50]	67.5	-
10582	-	Vanilla DeepLabv1 [6]	69.8	-

Table 2: Results on **PASCAL VOC 2012 validation and test** sets. The last three rows report the performance of previous semi-supervised models with the same annotation.

# images in $\mathcal{F}$	200	400	800	1464
Ancillary Model	81.57	83.56	85.36	86.71
No Self-correction	78.75	79.19	80.39	80.34
Lin. Self-correction	**79.43**	79.59	**80.69**	81.35
Conv. Self-correction	78.29	**79.63**	80.12	**82.33**

Table 1: Ablation study of models on the **PASCAL VOC 2012 validation** set using mIOU for different sizes of $\mathcal{F}$ For the last three rows, the remaining images in the training set is used as $\mathcal{W}$, i.e. $W + F = 10582$.

实验

Data Split		Method	mIOU
F	W		
914	2061	No Self-Corr.	75.44
914	2061	Lin. Self-Correction	76.22
914	2061	Conv. Self-Correction	**79.46**
914	2061	EM-fixed [41]	74.97
2975	-	Vanilla DeepLabv3+$_{ours}$	77.49

Table 4: Results on **Cityscapes validation** set. 30% of the training examples is used as $\mathcal{F}$, and the remaining as $\mathcal{W}$.

# images in $\mathcal{F}$	200	450	914
Ancillary Model	79.4	81.19	81.89
No Self-correction	**73.69**	75.10	75.44
Lin. Self-correction	73.56	75.24	76.22
Conv. Self-correction	69.38	**77.16**	**79.46**

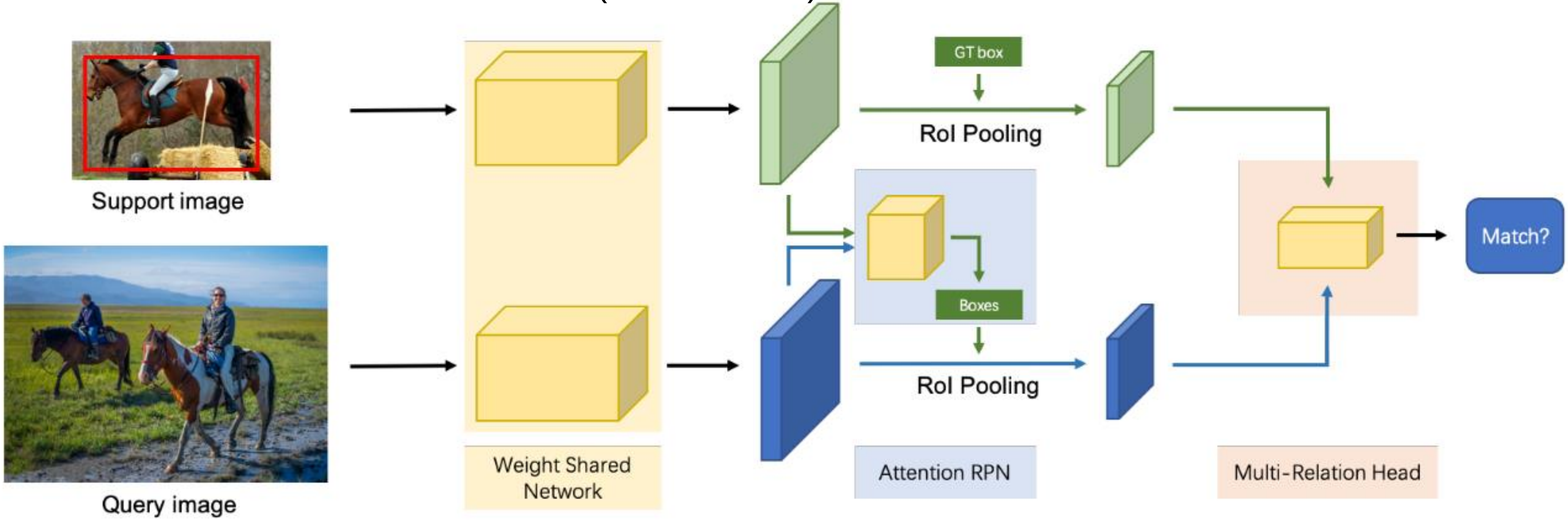
Table 3: Ablation study of our models on **Cityscapes va**
dation set using mIOU for different sizes of $\mathcal{F}$. For the l
three rows, the remaining images in the training set are us
as $\mathcal{W}$, i.e., $W + F = 2975$.

Few-Shot Object Detection with Attention-RPN and Multi-Relation Detector

- 现有的物体检测方法→大量标注数据
  - 小样本学习（few-shot）
  - 小样本学习主要集中在分类，在检测上的工作不多
- 小样本检测的任务定义
  - 给定：一个含有特定物体的支持集s_c，可能含有这一特定物体的查询集q_c
  - 任务是从查询集里找到所有支持集里的物体并用严格的边界框bbox标记
  - $N - \text{way}\ K - \text{shot}$: 支持集有$N$种物体，每种物体有$K$个样本
- 小样本检测的主要挑战：
  - 缺少专门针对小样本检测任务的数据集
  - 区域推荐网络（region proposal network，RPN）在小样本情况下很难准确给出新样本的候选框。
- 本文贡献：
  - 提出一个专门的数据集FSOD dataset
  - 深度注意力机制引入RPN中
  - Multi-Relation Detector

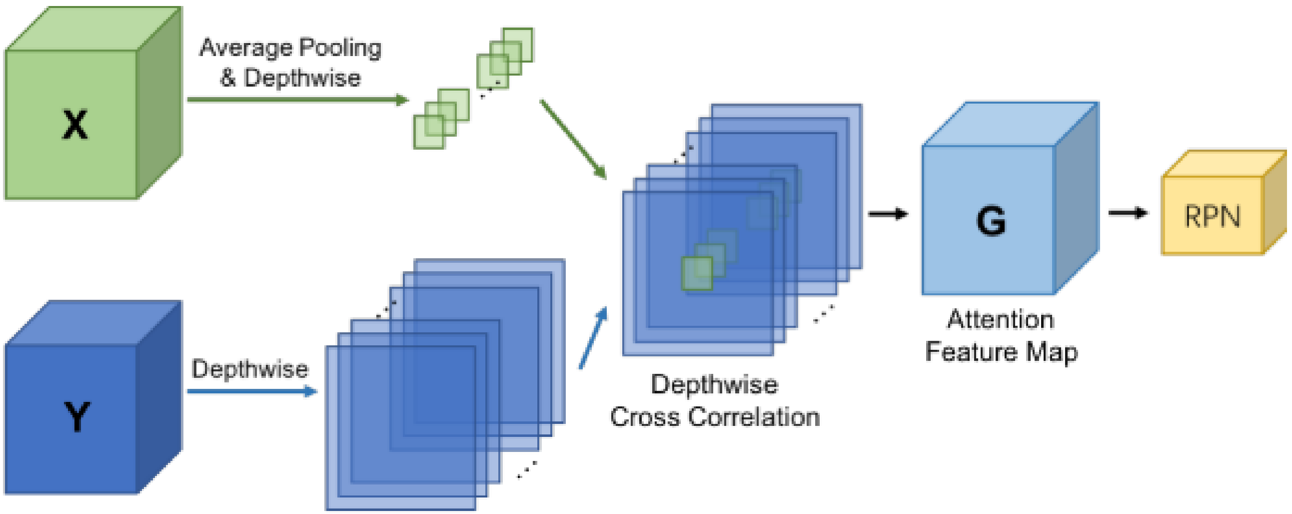
网络结构

- 基于Faster R-CNN
- 基于注意力的RPN
- Multi-Relation Detector（Match？）



基于注意力的RPN（attention-based RPN）

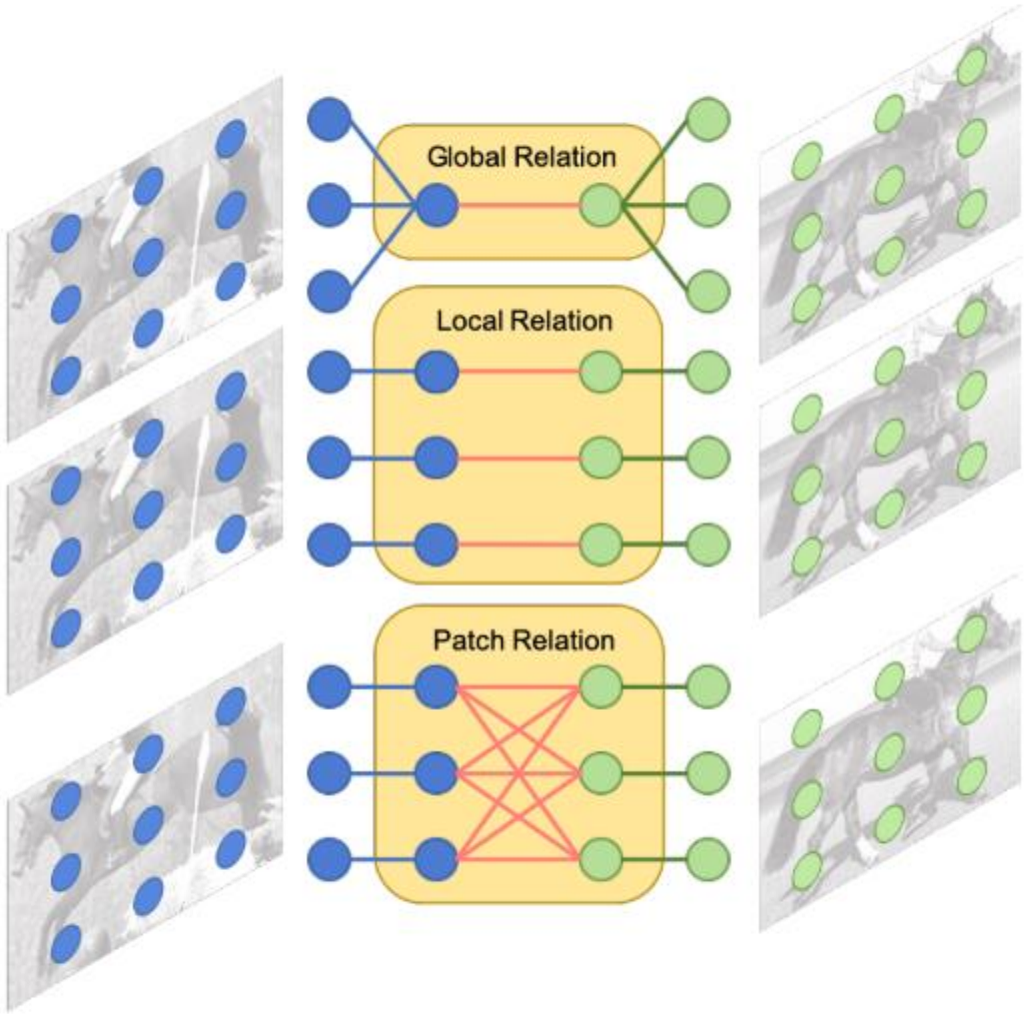
- 支持集X的特征指导查询集Y中PRN的选取
- 和种类无关，而和支持集的框选有关→能够使用与新种类的检测框提取
- X和Y做互相关



$$\mathbf{G}_{h,w,c} = \sum_{i,j} X_{i,j,c} \cdot Y_{h+i-1,w+j-1,c}, \quad i,j \in \{1,...,S\}$$

Multi-Relation Detector

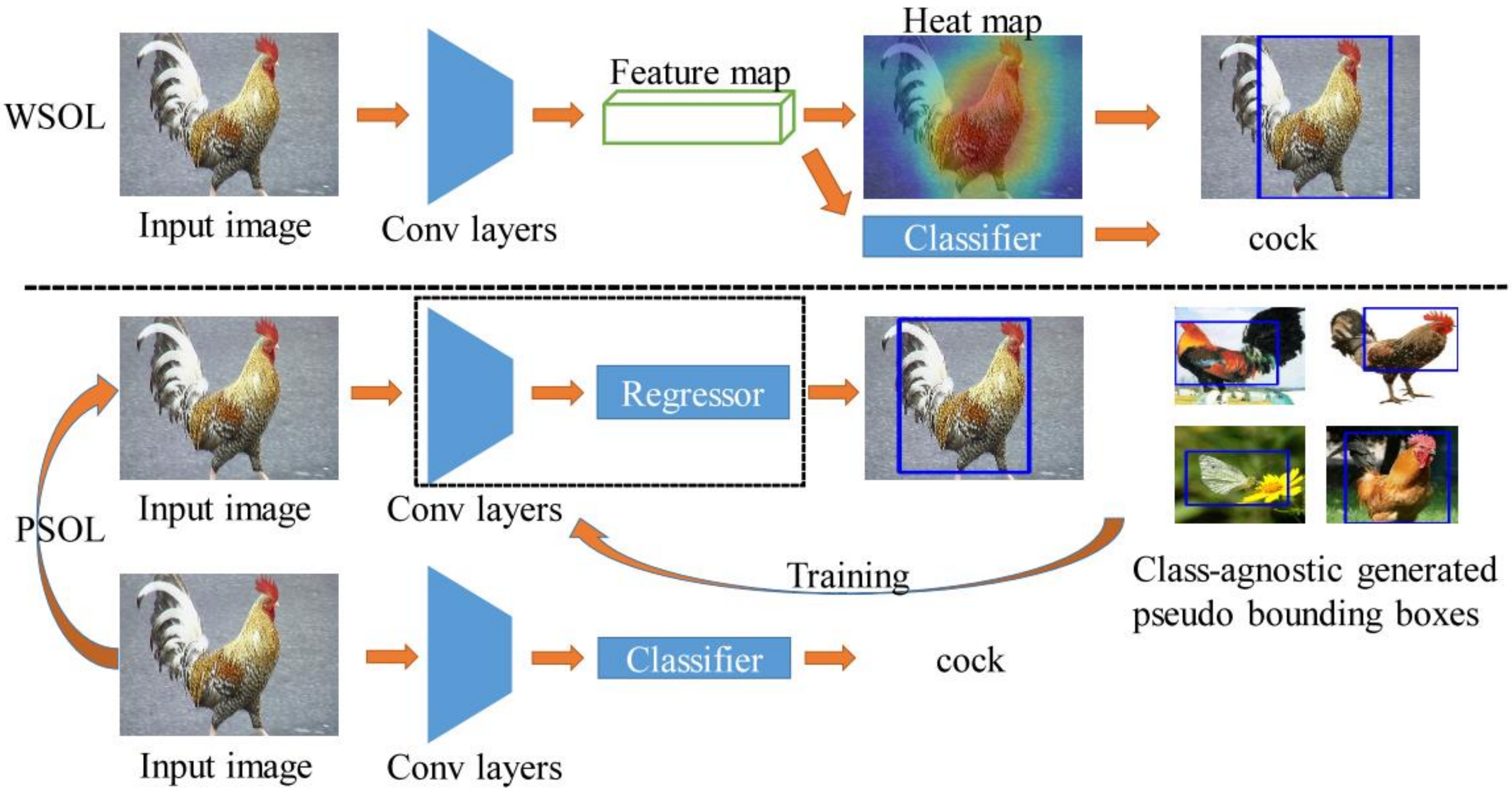
- Global Relation
- Local Relation
- Patch Relation



Rethinking the Route Towards Weakly Supervised Object Localization

- 现有的物体检测方法→大量标注数据
  - 弱监督学习（Weakly supervised object localization, WSOL）
  - 数据集只有图像级别的标注
- 主流方法的通用思路:
  - 特征图和分类权重→间接定位图片中的标注物体
- 本文的创新之处:
  - 指出WSOL应该将定位和分类两个任务分开来做，并通过实验验证了这一假设（Pseudo supervised object localization, PSOL）
  - 相比较于WSOL，PSOL的定位网络单独训练，可以适用于新的物体的检测任务而不需要fine-tuning。

主要思路


WSOL
Input image
Conv layers
Feature map
Heat map
Classifier
cock
PSOL
Input image
Conv layers
Regressor
Training
Class-agnostic generated pseudo bounding boxes
Input image
Conv layers
Classifier
cock

Bbox伪标签的产生

- WSOL methods [1][2]
- DDT[3]: n张图片构成的数据集
  - $G \in \mathbb{R}^{h \times w \times d} = \mathbb{R}^{hw \times d} = F(I)$
  - $G_{all} \in \mathbb{R}^{n \times hw \times d} = \mathbb{R}^{nhw \times d}$
  - 对G_{all}做PCA，得到最大特征值对应的特征向量P
  - $H_{i,j} = \sum_{k=1}^{d} G_{i,j,k} P_k$

[1] Junsuk Choe and Hyunjung Shim. Attention-based dropout layer for weakly supervised object localization. In CVPR, pages 2219–2228, 2019. 1, 3, 4, 5, 6

[2] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discrimi- native localization. In CVPR, pages 2921–2929, 2016. 1, 2, 3, 4, 5, 6, 7, 8

[2] Xiu-ShenWei, Chen-Lin Zhang, JianxinWu, Chunhua Shen, and Zhi-Hua Zhou. Unsupervised object discovery and co-localization by deep descriptor transformation. Pattern Recognition, 88:113–126, 2019. 1, 4, 5, 6

实验

- ImageNet-1k： 1000 classes, 1,281,197, and 50,000 validation
- CUB-200: 200 categories of birds, 5,994 training, and 5,794 testing

实验

Model	Backbone	Parameters	FLOPs	CUB-200		ImageNet-1k		
				Top-1 Loc	Top-5 Loc	Top-1 Loc	Top-5 Loc	GT-Known Loc
VGG16-CAM [30]	VGG-GAP	14.82M	15.35G	36.13	-	42.80	54.86	59.00
VGG16-ACoL [28]	VGG-GAP	45.08M	43.32G	45.92	56.51	45.83	59.43	62.96
ADL [2]	VGG-GAP	14.82M	15.35G	52.36	-	44.92	-	-
VGG16-Grad-CAM [16]	VGG16	138.36M	15.42G	-	-	43.49	53.59	-
CutMix [27]	VGG-GAP	138.36M	15.35G	52.53	-	43.45	-	-
DDT-VGG16 [26]	VGG16	138.36M	15.42G	62.30	78.15	47.31	58.23	61.41
PSOL-VGG16-Sep	VGG16	274.72M	30.83G	**66.30**	**84.05**	**50.89**	**60.90**	**64.03**
PSOL-VGG16-Joint	VGG16	140.46M	15.42G	60.07	75.35	48.83	59.00	62.1
PSOL-VGG-GAP-Sep	VGG-GAP	29.64M	30.70G	59.29	74.88	48.36	58.75	63.72
PSOL-VGG-GAP-Joint	VGG-GAP	15.08M	15.35G	58.39	72.64	47.37	58.41	62.25
SPG [29]	InceptionV3	38.45M	66.59G	46.64	57.72	48.60	60.00	64.69
ADL [2]	InceptionV3	38.45M	66.59G	53.04	-	48.71	-	-
PSOL-InceptionV3-Sep	InceptionV3	53.32M	11.42G	**65.51**	**83.44**	**54.82**	**63.25**	**65.21**
PSOL-InceptionV3-Joint	InceptionV3	29.21M	5.71G	60.32	78.98	52.76	61.10	62.83
ResNet50-CAM [30]	ResNet50	25.56M	4.10G	29.58	37.25	38.99	49.47	51.86
ADL [2]	ResNet50-SE	28.09M	6.10G	62.29	-	48.53	-	-
CutMix [27]	ResNet50	26.61M	4.10G	54.81	-	47.25	-	-
PSOL-ResNet50-Sep	ResNet50	50.12M	8.18G	**70.68**	**86.64**	**53.98**	**63.08**	**65.44**
PSOL-ResNet50-Joint	ResNet50	26.61M	4.10G	68.17	83.69	52.82	62.00	64.30
DenseNet161-CAM	DenseNet161	29.81M	7.80G	29.81	39.85	39.61	50.40	52.54
PSOL-DenseNet161-Sep	DenseNet161	56.29M	15.46G	**74.97**	**89.12**	**55.31**	**64.18**	**66.28**
PSOL-DenseNet161-Joint	DenseNet161	29.81M	7.80G	74.24	87.03	54.48	63.41	65.39

ADL:Attention-based Dropout Layer for Weakly Supervised Object Localization