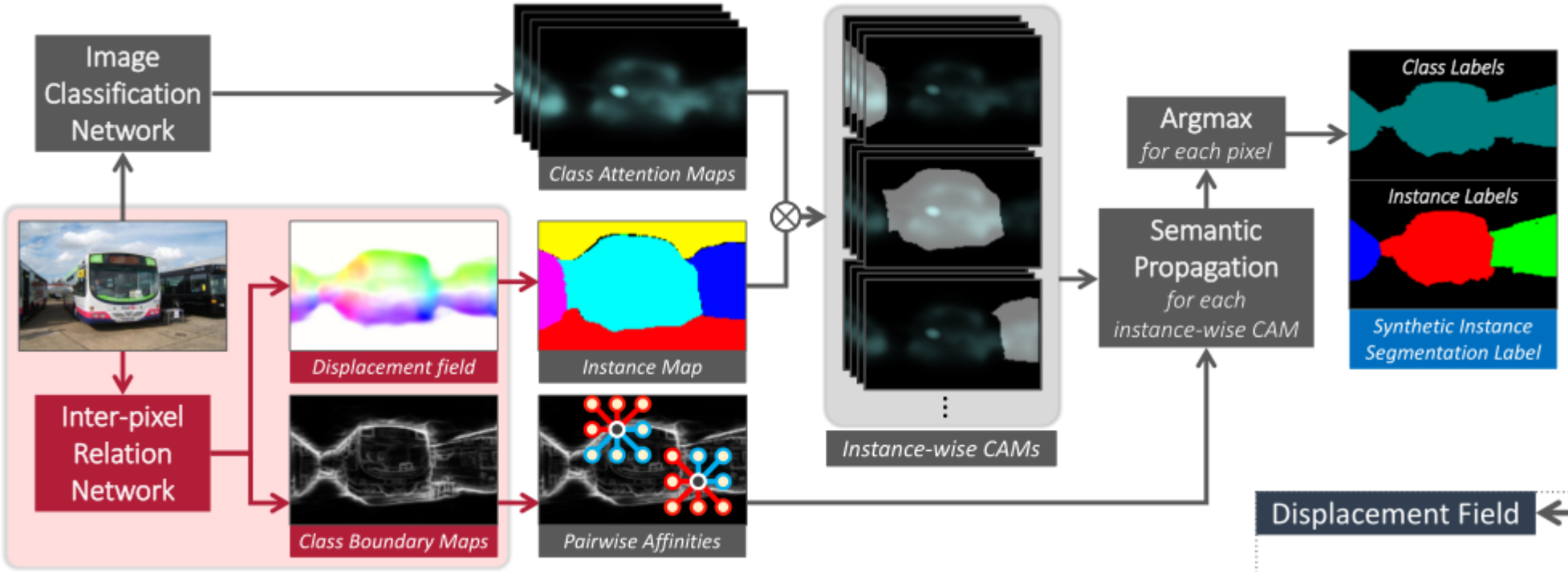
Weakly Supervised Learning of Instance Segmentation with Inter-pixel Relations
(基于像素间关系的实例分割弱监督学习)

- 生成训练图像的伪实例分割标签，用于训练完全监督的模型。为了生成伪标签，首先从图像分类模型的注意图中识别对象类的可靠种子区域，然后传播它们以发现具有精确边界的整个实例区域（估计单个实例的粗糙区域，并检测不同对象类之间的边界）。
- 为了克服CAMs的局限性，引入了像素间关系网络（IRNet），该网络用于估计与CAMs互补的两类附加信息：类不可知实例映射和成对语义相似度（类不可知实例映射和成对语义相似度。类无关实例映射是一个粗糙的实例分割掩码，没有类标签，也没有精确的边界。另一方面，一对像素之间的语义相似度是它们之间的类等价性的置信度）。
- IRNet有两个分支，分别估计实例映射和语义相似度。第一个分支预测位移向量场，其中每个像素处的2D向量指示像素所属实例的质心。通过将相同的实例标签指定给置换向量指向相同位置的像素，置换场将转换为实例贴图。第二个分支检测不同对象类之间的边界。然后，从检测到的边界计算成对语义相似度，这样，被强边界分隔的两个像素被视为具有低语义相似度的一对。
- 在PASCAL VOC 2012数据集上，我们的模型大大优于之前接受过同等监管水平的最先进模型。它甚至超越了以前基于更强监管的模型，如使用边界框标签的SDI和使用完全监管的早期模型SDS。



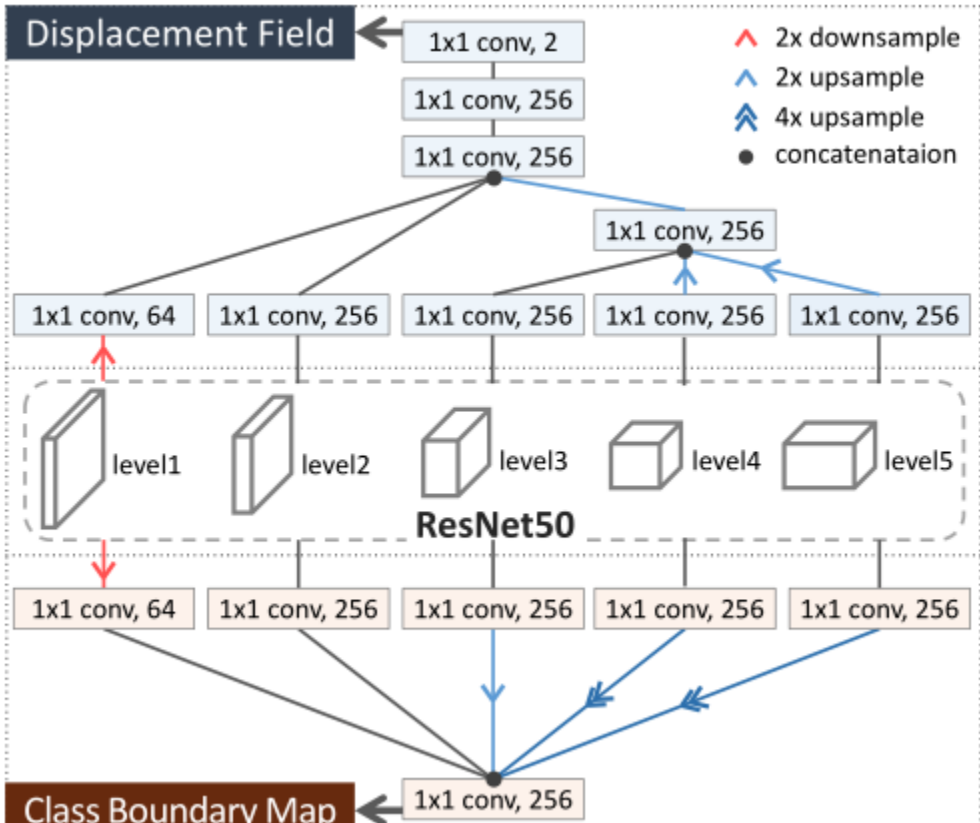
CAM在我们的框架中扮演着两个重要角色。用于定义实例的种子区域，这些种子区域稍后传播以恢复整个实例区域；学习IRNet的监督来源，通过仔细利用CAM，我们提取出可靠的像素间关系，并从中训练IRNet。

IRNet旨在提供两种类型的信息：位移向量场和类边界图，这两种信息依次用于从CAMs估计伪实例掩码。

作为输入，两个分支都从主干的所有五层获取特征映射。除最后一层外，两个分支的所有卷积层后面都是组归一化和ReLU。
位移场预测分支：首先对每个输入特征图应用1×1卷积层，如果大于该卷积层，通道数将减少到256个。在此基础上，添加了一种自上而下的路径方式，以迭代方式合并所有特征图，低分辨率特征图上采样两次，与相同分辨率的特征图串联，并通过1×1卷积层进行处理。最后，从最后拼接的特征图出发，通过三个1×1卷积层对位移场进行解码，其输出有两个通道。
边界检测分支：首先对每个输入特征映射应用1×1卷积进行降维。然后对结果进行大小调整、串联，并将其送入最后的1×1卷积层，该卷积层根据串联的特征生成类边界图。

$$M_c(\mathbf{x}) = \frac{\phi_c^\top f(\mathbf{x})}{\max_{\mathbf{x}} \phi_c^\top f(\mathbf{x})},$$

其中，f是来自CNN最后一个卷积层的特征映射，x是f上的2D坐标，φc是类别c的分类权重。不相关类别的CAM固定为零矩阵。采用ResNet50作为分类网络，并将其最后一个下采样层的步长从2减少到1，以防止CAMs进一步降低分辨率。



收集注意力score大于0.3的像素作为前景像素，小于0.05的像素作为背景像素。γ是限制成对最大距离的半径。我们进一步将P+分为P+fg和P+bg，分别是前景对和背景对。

$$\mathcal{P} = \{(i,j) \mid \|\mathbf{x}_i - \mathbf{x}_j\|_2 < \gamma, \forall i \neq j\}, \quad (2)$$

$$\mathcal{P}^+ = \{(i,j) \mid \hat{M}(\mathbf{x}_i) = \hat{M}(\mathbf{x}_j), (i,j) \in \mathcal{P}\}, \quad (3)$$

$$\mathcal{P}^- = \{(i,j) \mid \hat{M}(\mathbf{x}_i) \neq \hat{M}(\mathbf{x}_j), (i,j) \in \mathcal{P}\}, \quad (4)$$

Method	Sup.	Extra Data / Information	*val*	*test*
SEC [24]	$\mathcal{I}$	-	50.7	51.7
AffinityNet [1]	$\mathcal{I}$	-	58.7	-
PRM [50]	$\mathcal{I}$	MCG [2]	53.4	-
CrawlSeg [19]	$\mathcal{I}$	YouTube Videos	58.1	58.7
MDC [46]	$\mathcal{I}$	Ground-truth Backgrounds	60.4	60.8
DSRG [20]	$\mathcal{I}$	MSRA-B [31]	61.4	63.2
ScribbleSup [27]	$\mathcal{S}$	-	63.1	-
BoxSup [8]	$\mathcal{B}$	-	62.0	64.6
SDI [22]	$\mathcal{B}$	BSDS [33]	65.7	67.5
Upperbound	$\mathcal{F}$	-	72.3	72.5
Ours-ResNet50	$\mathcal{I}$	-	63.5	64.8

Table 4. Semantic segmentation performance on the PASCAL VOC 2012 *val* and *test* sets. The supervision type (Sup.) indicates: $\mathcal{I}$–image-level label, $\mathcal{B}$–bounding box, $\mathcal{S}$–scribble, and $\mathcal{F}$–segmentation label.

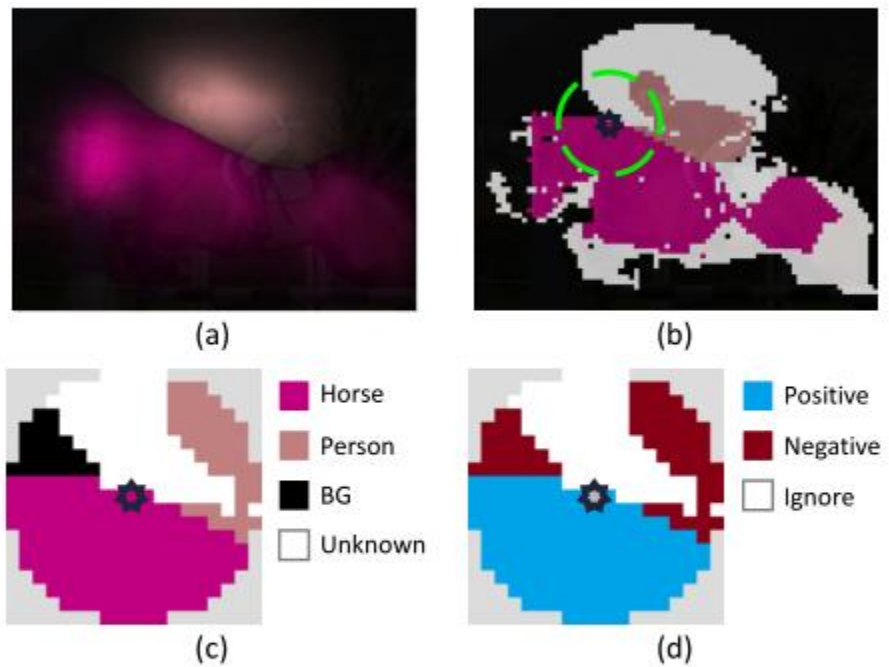


Figure 3. Visualization of our inter-pixel relation mining process. (a) CAMs. (b) Confident areas of object classes. (c) Pseudo class label map within a local neighborhood. (d) Class equivalence relations between the center and the others.

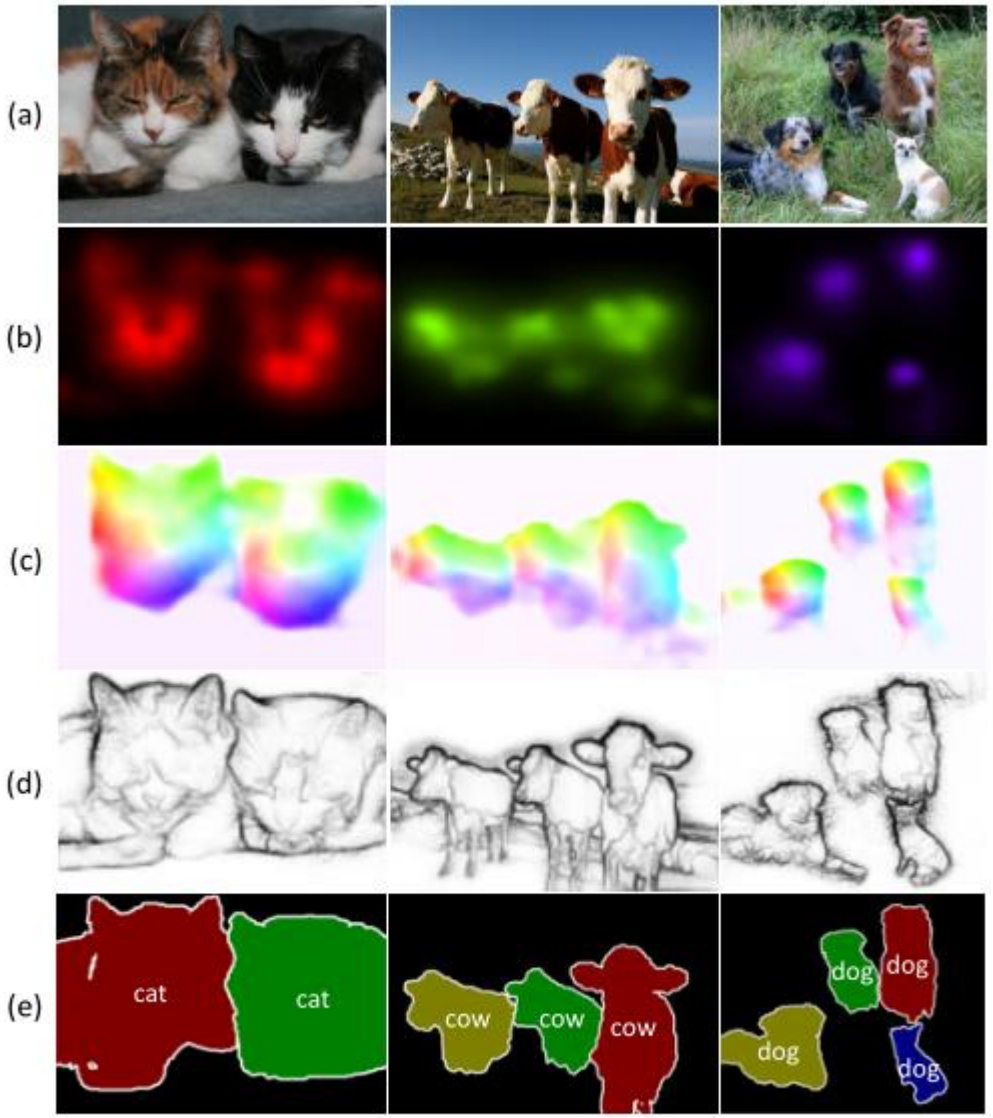


Figure 6. Examples of pseudo instance segmentation labels on the PASCAL VOC 2012 *train* set. (a) Input image. (b) CAMs. (c) Displacement field. (d) Class boundary map. (e) Pseudo labels.

Figure 7. Qualitative results of our instance segmentation model on the PASCAL VOC 2012 *val* set.

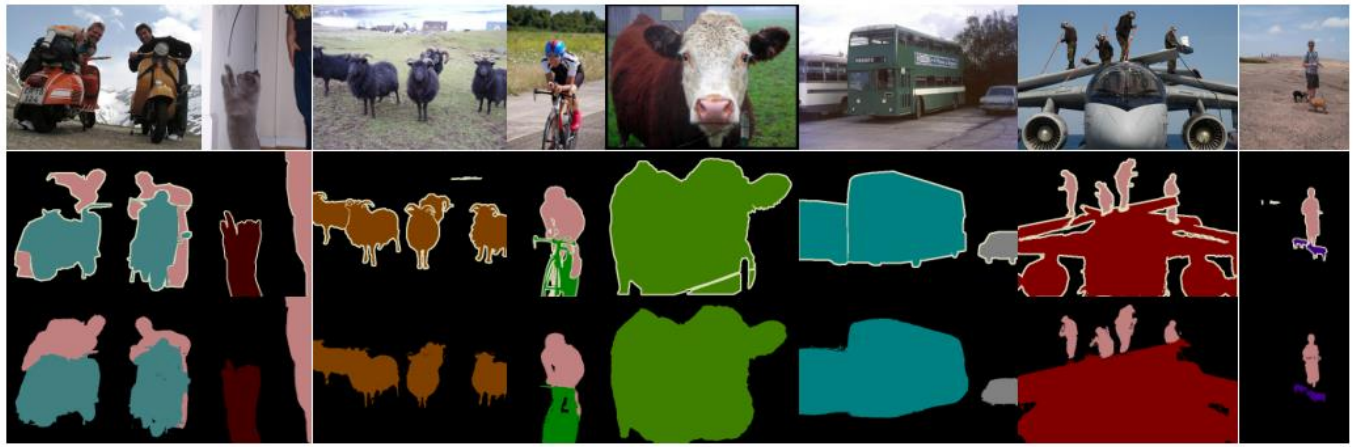
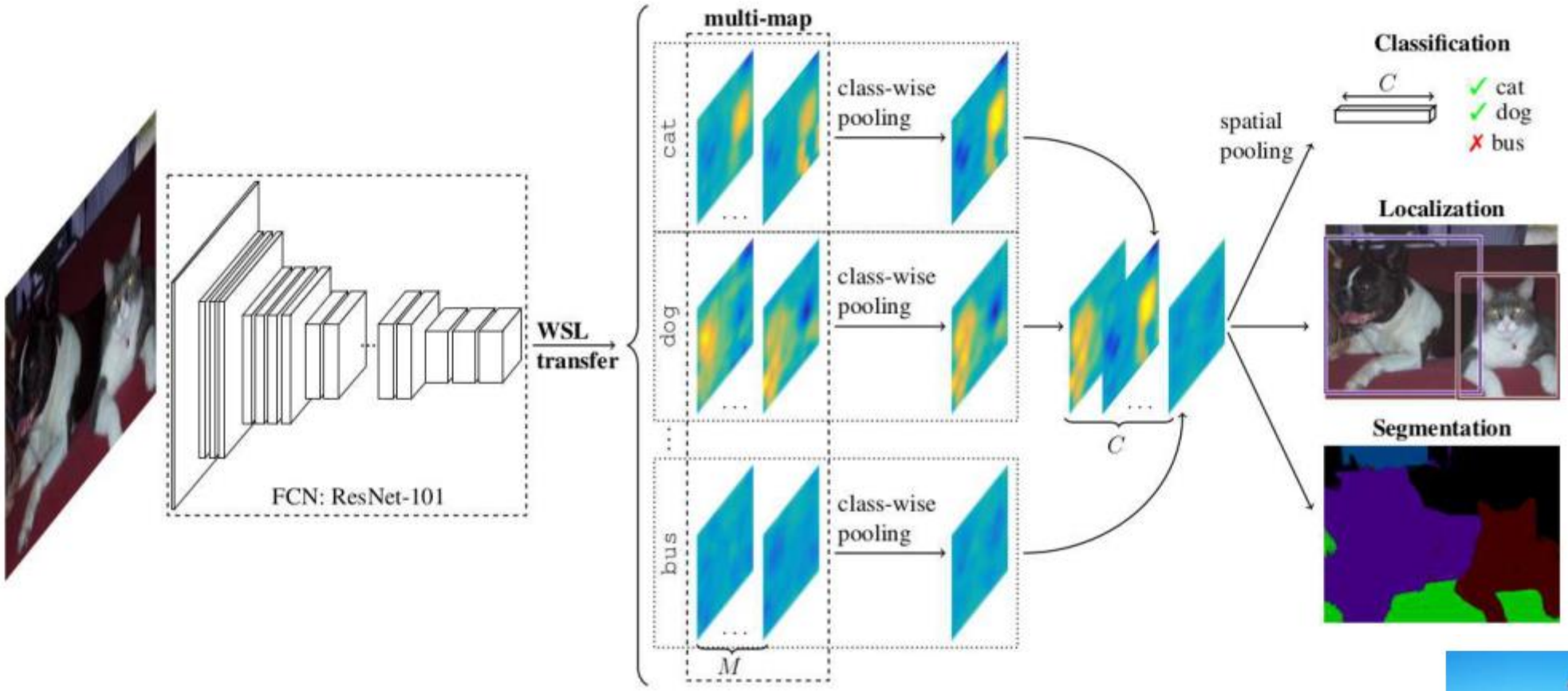


Figure 8. Qualitative results of smenatic segmentation on the PASCAL VOC 2012 *val* set. (*top*) Input images. (*middle*) Groundtruth semantic segmentaton. (*bottom*) Results of Ours-ResNet50.

WILDCAT: Weakly Supervised Learning of Deep ConvNets for Image Classification, Pointwise Localization and Segmentation (WILDCAT：用于图像分类、逐点定位和分割的深度变换的弱监督学习)

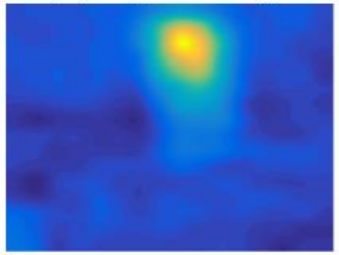
- WILDCAT，一种深度学习方法，旨在联合对齐图像区域以获得空间不变性和学习强局部特征。模型仅使用全局图像标签进行训练，并致力于三个主要的视觉识别任务：图像分类、弱监督点方向目标定位和语义分割。
- 提出了WILDCAT（深度卷积神经网络的弱监督学习），这是一种学习与类别模式相关的局部视觉特征的方法，该模型可用于图像分类以及弱监督点方向目标定位和分割。
- ①使用最新的全卷积网络（FCN）作为后端模块。②合并了一个WSL传输层，该层明确学习与互补类模式相关的多个局部特征。③解决了将空间分数聚合为全局预测的问题，这是WSL训练的一个关键问题。
- 对六个数据集上的WILDCAT模型进行了全面评估，其在分类、WSL逐点检测和分割任务方面具有优异性能。
- 一种新的弱监督学习方法，专门用于在训练期间仅使用图像级标签来学习有区别的局部视觉特征。



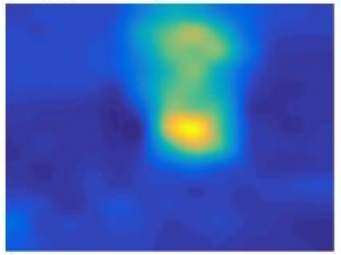
基于FCN ResNet-101，从具有良好空间分辨率的整体图像中提取局部特征。所有区域都通过WSL多地图传输层编码为多类模式。然后，使用新的空间聚合模块，将特征图单独组合，以生成特定于类别的热图，这些热图可以全局合并，以获得每个类别的单一概率。

(a) original image

(b) final predictions

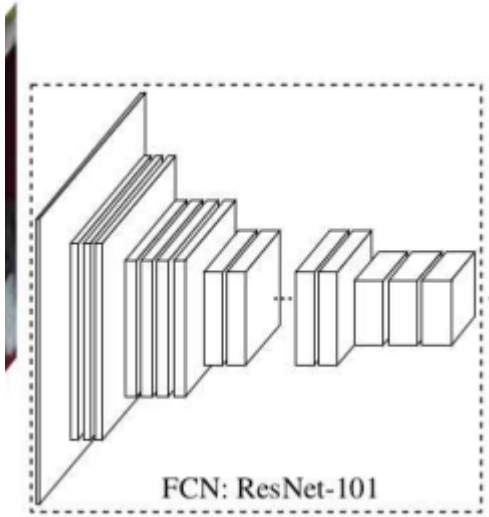


(c) *dog* heatmap 1 (head)



(d) *dog* heatmap 2 (legs)

FCN：更精细的地图保持更高的空间分辨率，并获取更具体的区域。可以在整个网络中自然地保留空间信息。它还可以在单个正向过程中计算所有区域的局部特征，而无需调整其大小。我们使用在ImageNet数据集上预先训练过的公开发布模型，并删除最后的层（全局平均池和完全连接），以替换为WSLtransfer和wildcat pooling，



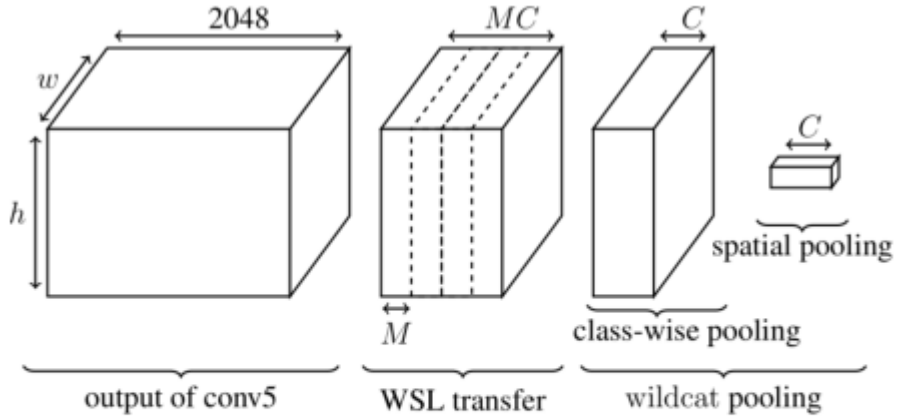


Figure 3. WILDCAT local feature encoding and pooling. Class modalities are encoded with a multi-map WSL transfer layer and pooled separately for all classes. Local features are then aggregated with a global spatial pooling to yield a single score per class.

wildcat pooling分为class-wise pooling和spatial pooling两部分

引入了一个多映射WSL传输层，该层学习多个类相关模式，通过1×1卷积将每个类编码为M个特征映射。模式以WSL方式学习，仅使用图像级标签，传输层保持空间分辨率，这是WSL的关键。

$$\begin{cases} \bar{z}^c_{i,j} = \underset{m\in\{1,\dots,M\}}{\text{Cl. Pool}} \; z^{c,m}_{i,j} & (1) \\ s^c = \underset{(i,j)\in\{1,\dots,w\}\times\{1,\dots,h\}}{\text{Sp. Pool}} \; \bar{z}^c_{i,j} & (2) \end{cases}$$

z是传输层的输出，Cl.Pool是所选的类池函数，Sp.Pool是空间聚合过程。

Method	15 Scene	MIT67
CaffeNet Places [71]	90.2	68.2
MOP CNN [25]	-	68.9
Negative parts [47]	-	77.1
GAP GoogLeNet [70]	88.3	66.6
WELDON [13]	94.3	78.0
Compact Bilinear Pooling [18]	-	76.2
ResNet-101 (*) [28]	91.9	78.0
SPLeaP [35]	-	73.5
WILDCAT	**94.4**	**84.0**

Table 3. Classification performances (multi-class accuracy) on scene datasets.

Method	VOC 2007	VOC 2012
VGG16 [56]	89.3	89.0
DeepMIL [44]	-	86.3
WELDON [13]	90.2	-
ResNet-101 (*) [28]	89.8	89.2
ProNet [58]	-	89.3
RRSVM [61]	92.9	-
SPLeaP [35]	88.0	-
WILDCAT	**95.0**	**93.4**

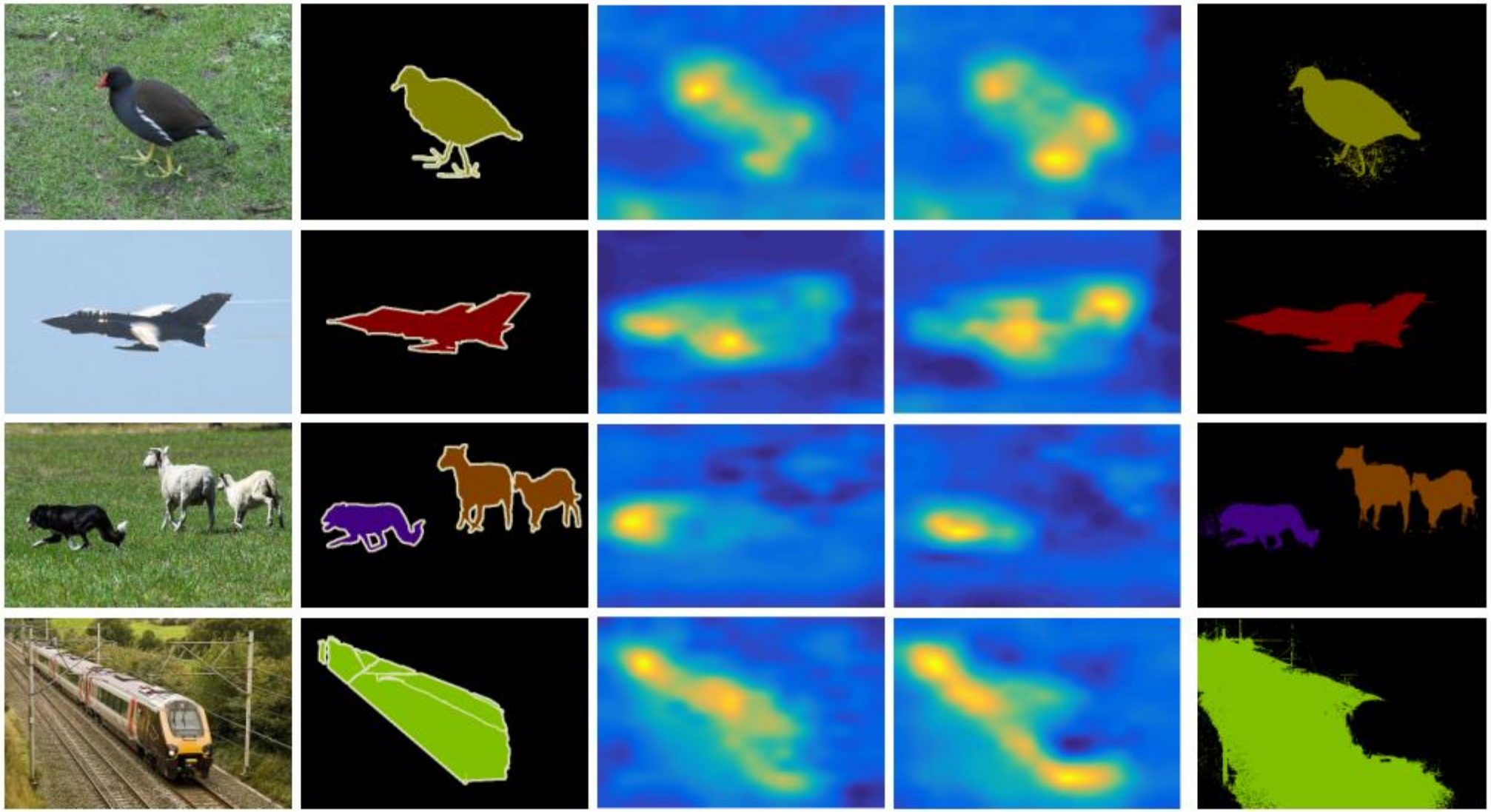
Table 2. Classification performances (MAP) on object recognition datasets. We used VOC evaluation server to evaluate on VOC 2012. (*) means that results are obtained with online code

Method	VOC 2012 Action	MS COCO
DeepMIL [44]	-	62.8
WELDON [13]	75.0	68.8
ResNet-101 (*) [28]	77.9	72.5
ProNet [58]	-	70.9
WILDCAT	**86.0**	**80.7**

Table 4. Classification performances (MAP) on context datasets.

Method	VOC 2012	MS COCO
DeepMIL [44]	74.5	41.2
ProNet [58]	77.7	46.4
WSLocalization [5]	79.7	49.2
WILDCAT	**82.9**	**53.4**

Table 8. Pointwise object localization performances (MAP) on PASCAL VOC 2012 and MS COCO.

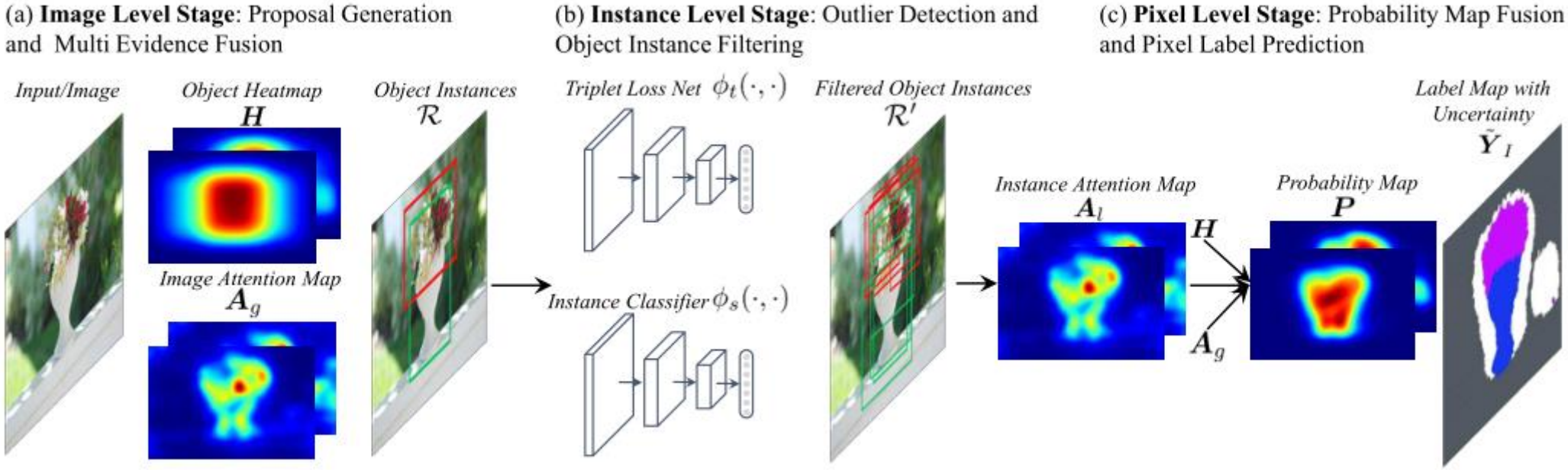


(a) original image (b) ground truth (c) heatmap1 (d) heatmap2 (e) WILDCAT prediction

Figure 6. Segmentation examples on VOC 2012. Our prediction is correct except for the train (last row) where our model aggregated rails and train regions. For objects as *bird* or *plane*, one can see how two heatmaps (heatmap1 (c) and heatmap2 (d) representing the same class: respectively *bird, aeroplane, dog* and *train*) succeed to focus on different but relevant parts of the objects.

Multi-Evidence Filtering and Fusion for Multi-Label Classification, Object Detection and Semantic Segmentation Based on Weakly Supervised Learning

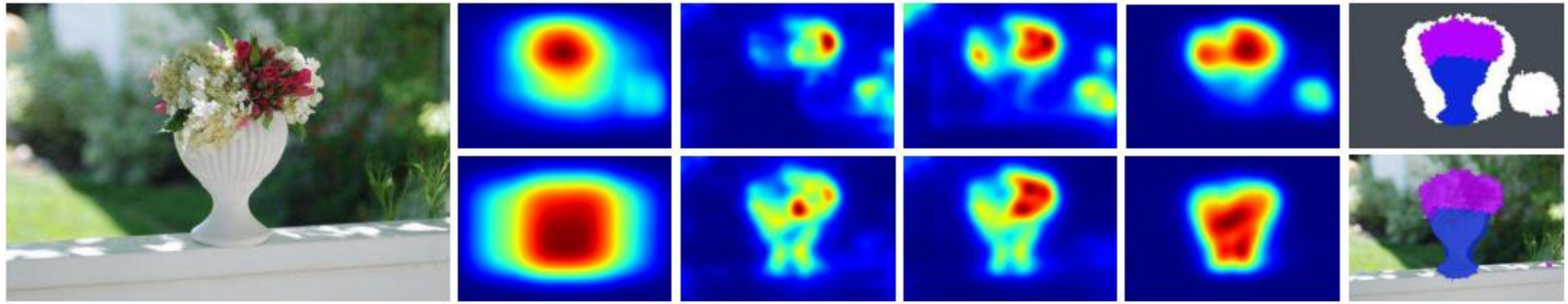
- 提出了一种新的弱监督课程学习管道，用于多标签对象识别、检测和语义分割。
- 首先获得训练图像的中间对象定位和像素标记结果，然后使用这些结果以完全监督的方式训练特定任务的深度网络。整个过程包括四个阶段，包括训练图像中的对象定位、对象实例的过滤和融合、训练图像的像素标记和特定任务的网络训练。为了在训练图像中获得干净的对象实例，我们提出了一种新的算法，用于过滤、融合和分类从多个解决方案机制中收集的对象实例。
- 实验表明，在MS-COCO、PASCAL VOC 2007和PASCAL VOC 2012上，弱监督管道在多标签图像分类和弱监督目标检测方面取得了最先进的结果，在弱监督语义分割方面取得了非常有竞争力的结果。
- ①从自下而上和自上而下的弱监督目标检测算法中收集训练图像中的目标定位结果；②结合度量学习和基于密度的聚类（metric learning and density-based clustering）来过滤检测到的对象实例。从而，获得了一组相对干净和完整的对象实例。进一步训练一个单标签对象分类器，该分类器应用于所有对象实例以获得它们的最终类标签；③融合图像级注意图、目标级注意图和目标检测热图，为每个类和每个训练图像获得一个相对干净的像素级概率图。像素级概率图用于训练全卷积网络，该网络应用于所有训练图像，以获得最终的像素级标签图；④将获得的目标实例和所有训练图像的像素级标签映射分别用于训练深度网络进行目标检测和语义分割。为了使训练图像的像素级标签映射有助于多标签图像分类，通过训练一个具有两个分支的深度网络来执行多任务学习，一个用于多标签图像分类，另一个用于像素标签。



从左到右：（a）图像级阶段：融合对象热图H和图像注意图Ag，生成实例级阶段的对象实例R，并在像素级阶段提供这两个地图进行信息融合。（b） 实例级阶段：进行基于三重损失的度量学习和基于密度的聚类进行离群点检测，并训练单标签实例分类器φs（·，·）进行实例过滤。（c） 像素级阶段：集成对象热图H、实例注意图AI和图像注意图Ag，用于不确定性像素标记。

Instance Level Stage：由于同一图像中存在多个对象类别，神经注意很难为每个类别获得准确的像素级注意图，因此我们训练了一个单标签对象实例分类网络，并在该网络中计算注意图，以获得更准确的像素级类别概率。通过度量学习和基于密度的聚类，进一步过滤图像级的融合对象实例。剩余的标记对象建议用于训练该对象实例分类器，该分类器还可用于进一步移除剩余的误报对象实例。

Pixel Level Stage：在之前的阶段中，我们已经构建了图像分类器、弱监督对象检测器和对象实例分类器。这些深层网络中的每一个都会从输入图像中产生自己的推理结果。例如，图像分类器生成全局注意图，对象检测器生成对象热图。在像素级阶段，我们仍然执行多证据过滤和融合，以整合所有这些组件网络的推理结果，获得指示每个像素处潜在对象类别的像素级概率图。来自图像分类器的全局注意图Ag对图像中的对象有充分的了解，但有时只关注最重要的对象部分。对象实例分类器具有每个单独对象的局部视图。借助实例分类器生成的特定于对象的局部注意图，我们可以避免丢失小对象。



(a) Input/Image (b) Object Heatmap (c) Image Attention (d) Instance Attention (e) Probability (e) Segmentation

Figure 4. The pixel labeling process in the pixel level stage. White pixels in the last column indicate pixels with uncertain labels.

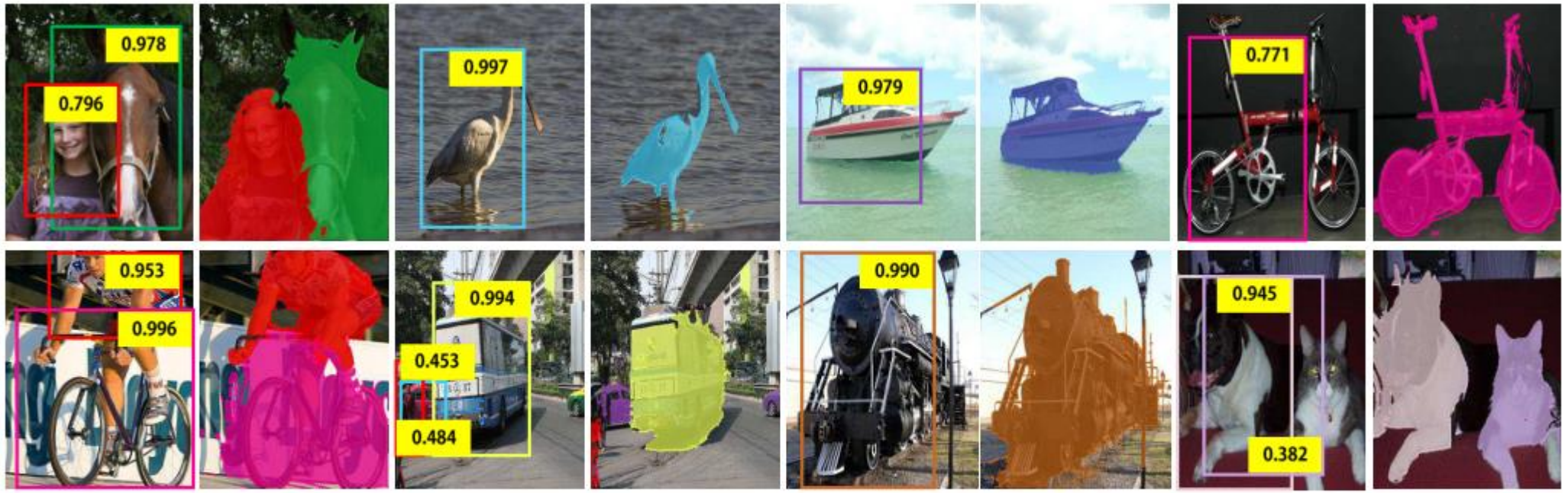


Figure 5. The detection and semantic segmentation results on Pascal VOC 2012 test set (the first row) and Pascal VOC 2007 test set (the second row). The detection results are gotten by select proposals with the highest confidence of every class. The semantic segmentation results are post-processed by CRF [22].

method	bg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mIoU
SEC[21]	83.5	56.4	28.5	64.1	23.6	46.5	70.6	58.5	71.3	**23.2**	54.0	28.0	68.1	62.1	70.0	55.0	38.4	58.0	39.9	38.4	48.3	51.7
FCL[32]	85.7	58.8	30.5	67.6	24.7	44.7	**74.8**	61.8	**73.7**	22.9	57.4	27.5	**71.3**	64.8	**72.4**	57.3	37.0	60.4	42.8	42.2	**50.6**	53.7
TP-BM[20]	83.4	62.2	26.4	**71.8**	18.2	**49.5**	66.5	**63.8**	73.4	19.0	56.6	35.7	69.3	61.3	71.7	**69.2**	39.1	66.3	**44.8**	35.9	45.5	53.8
AE-PSL[43]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	55.7
Ours+CRF	**86.6**	**72.0**	**30.6**	68.0	**44.8**	46.2	73.4	56.6	73.0	18.9	**63.3**	32.0	70.1	**72.2**	68.2	56.1	34.5	**67.5**	29.6	**60.2**	43.6	**55.6**

Table 1. Comparison among weakly supervised semantic segmentation methods on PASCAL VOC 2012 *segmentation test* set.

method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
OM+MIL+FRCNN[24]	54.5	47.4	41.3	20.8	17.7	51.9	63.5	46.1	21.8	57.1	22.1	34.4	50.5	61.8	16.2	**29.9**	40.7	15.9	55.3	40.2	39.5
HCP+DSD+OSSH3[19]	54.2	52.0	35.2	25.9	15.0	59.6	**67.9**	58.7	10.1	**67.4**	27.3	37.8	54.8	**67.3**	5.1	19.7	**52.6**	43.5	56.9	62.5	43.7
OICR-Ens+FRCNN[38]	65.5	67.2	47.2	21.6	22.1	68.0	68.5	35.9	5.7	63.1	49.5	30.3	64.7	66.1	13.0	25.6	50.0	57.1	60.2	59.0	47.0
Ours+FRCNN w/o clustering	66.7	61.8	55.3	41.8	6.7	61.2	62.5	**72.8**	12.7	46.2	40.9	**71.0**	67.3	64.7	**30.9**	16.7	42.6	56.0	65.0	26.5	48.5
Ours+FRCNN w/o uncertainty	66.8	63.4	54.5	42.2	5.8	60.5	58.3	67.8	7.8	46.1	40.3	71.0	68.2	62.6	30.7	16.5	41.1	55.2	66.8	25.2	47.5
Ours+FRCNN w/o instances	**67.7**	62.9	53.1	**44.4**	11.2	62.4	58.5	71.2	8.3	45.7	41.5	71.0	68.0	59.2	30.3	15.0	42.4	56.0	**67.2**	26.8	48.1
Ours+FRCNN	64.3	**68.0**	**56.2**	36.4	**23.1**	**68.5**	67.2	64.9	7.1	54.1	**47.0**	57.0	**69.3**	65.4	20.8	23.2	50.7	**59.6**	65.2	**57.0**	**51.2**

Table 2. Average precision (in %) of weakly supervised methods on PASCAL VOC 2007 *detection test* set.

method	F1-C	P-C	R-C	F1-O	P-O	R-O	F1-C/top3	P-C/top3	R-C/top3	F1-O/top3	P-O/top3	R-O/top3
CNN-RNN[41]	-	-	-	-	-	-	60.4	66.0	55.6	67.8	69.2	66.4
RLSD[45]	-	-	-	-	-	-	62.0	67.6	57.2	66.5	70.1	63.4
RNN-Attention[42]	-	-	-	-	-	-	67.4	79.1	58.7	72.0	84.0	63.0
ResNet101-SRN[47]	70.0	**81.2**	63.3	75.0	84.1	67.7	66.3	**85.8**	57.5	72.1	88.1	61.1
ResNet101(448×448)(baseline)	72.8	73.8	**72.9**	76.3	77.5	**75.1**	69.5	78.3	**63.7**	73.1	83.8	**64.9**
Ours	**74.9**	80.4	70.2	**78.4**	**85.2**	72.5	**70.6**	84.5	62.2	**74.7**	**89.1**	64.3