Causal Class Activation Map for Weakly-Supervised Semantic Segmentation

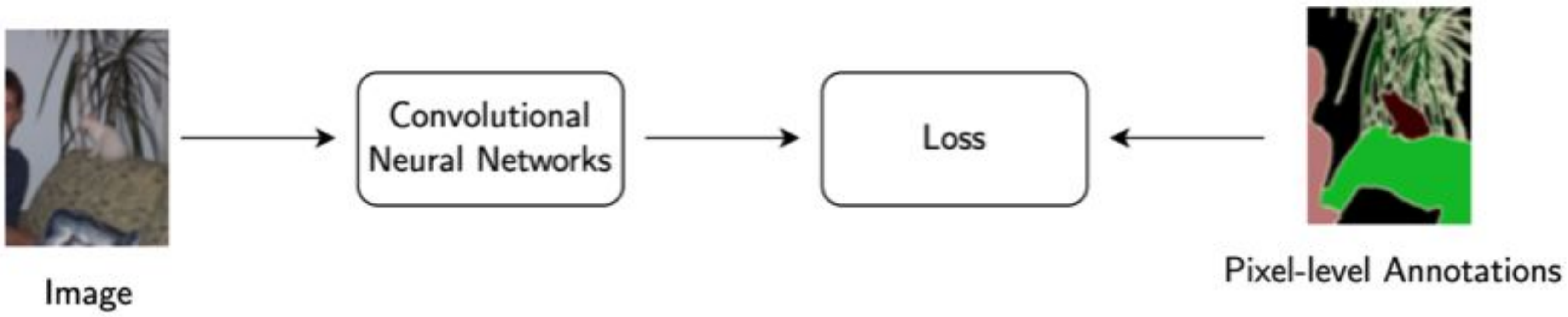
Code

Paper

Yiping Wang
https://yipingwang.ca
2022.08.05

I am available for a PhD position!

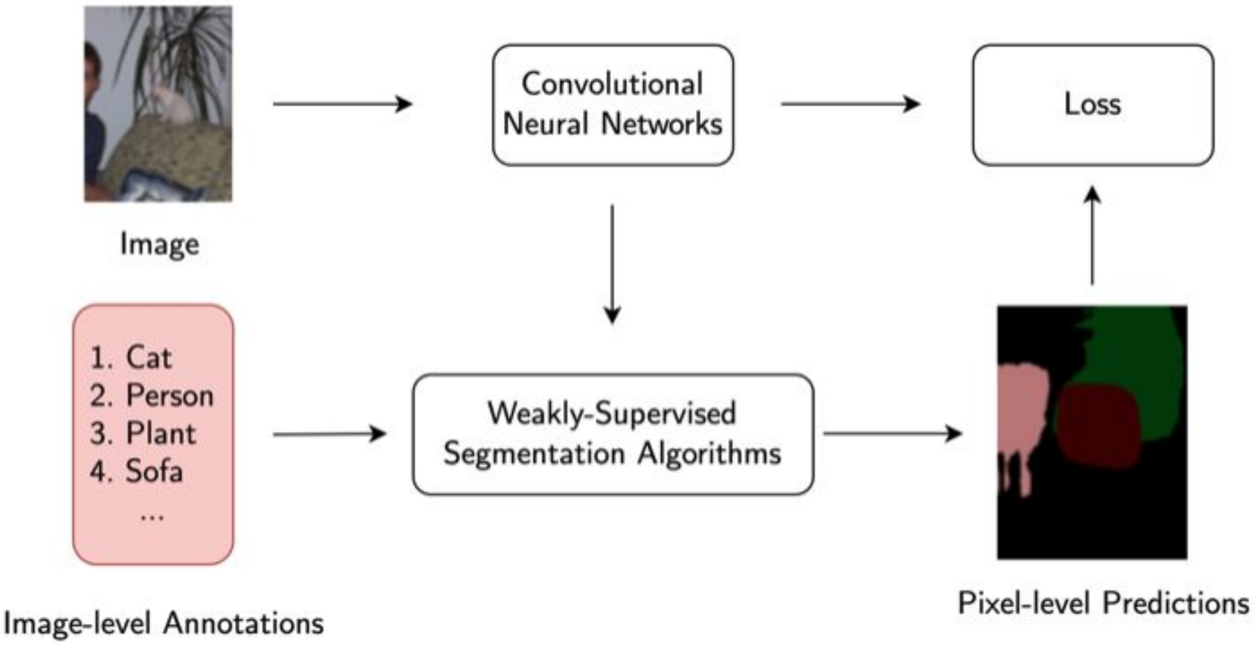
Semantic Segmentation



Supervision: pixel-level labels

Goal: pixel-level semantic segmentation, i.e., classifying each pixel to a class

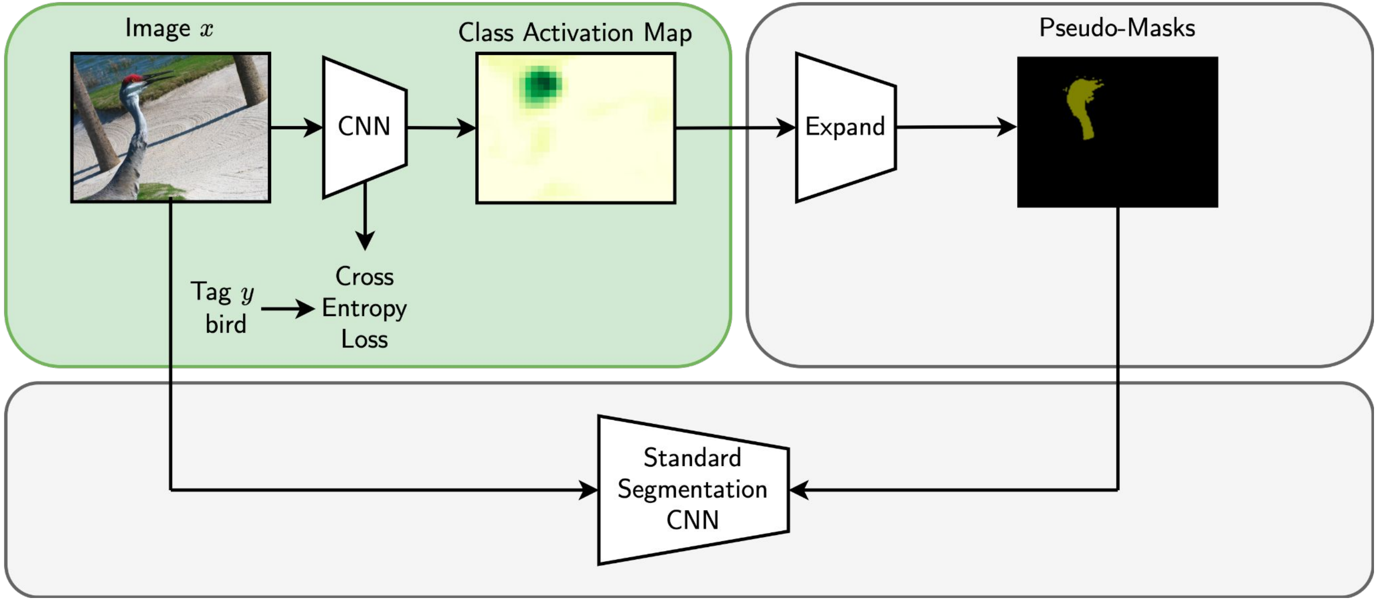
Weakly-Supervised Semantic Segmentation



Supervision: only image-level labels

Goal: pixel-level semantic segmentation, i.e., classifying each pixel to a class

Popular Pipeline



Supervision: only image-level labels

Goal: pixel-level semantic segmentation, i.e., classifying each pixel to a class

Class Activation Map

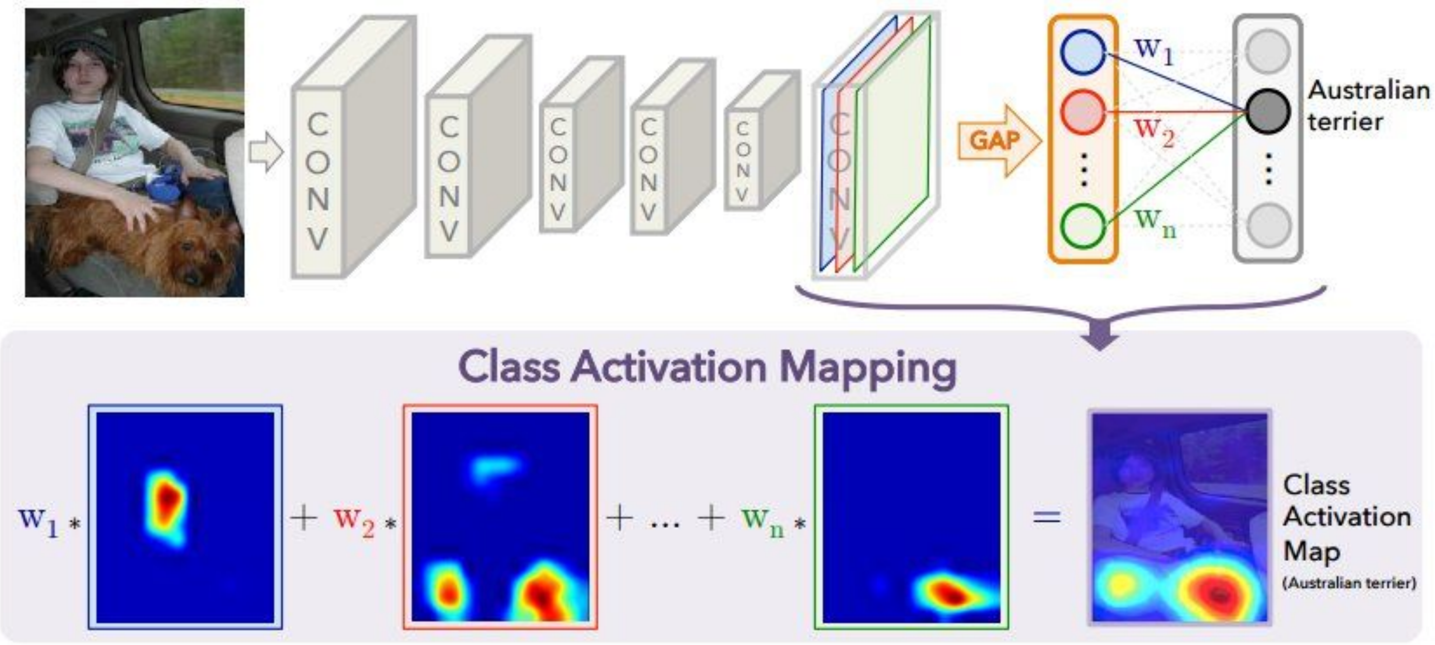


Figure Credit: B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. CVPR'16

CAMs in Out-Of-Distribution dataset

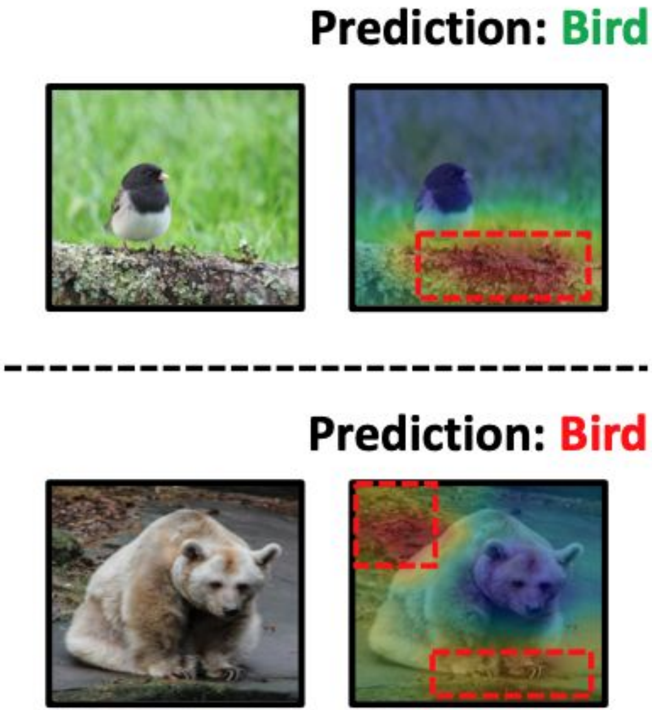
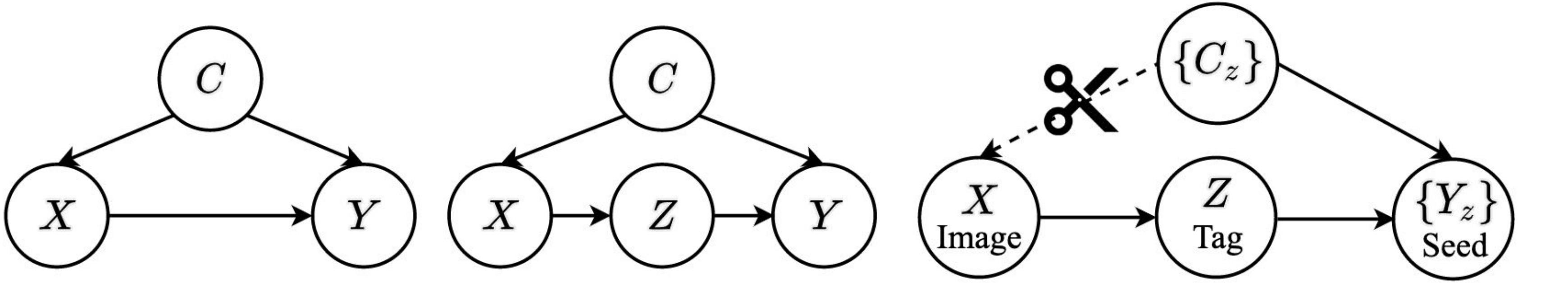


Figure Credit: T. Wang, C. Zhou, Q. Sun, H. Zhang et al., Causal Attention for Unbiased Visual Recognition, ICCV'21.

Front-Door Adjustment



(a) Back-Door (b) Front-Door (c) Structural Causal Model for Causal CAM

x denotes images with shape 3 * H * W, z denotes image-level label, y denotes CAM with shape H * W

Assumptions

$$P(Y=y|do(X=x)) = \sum_{z} P(Z=z|X=x) \sum_{x'} P(Y=y|X=x', Z=z) P(X=x') \qquad (3)$$

x denotes images with shape 3 * H * W, z denotes image-level label, y denotes CAM with shape H * W

Assumptions

$$P(Y=y|do(X=x)) = \sum_{z} P(Z=z|X=x) \sum_{x'} P(Y=y|X=x', Z=z) P(X=x') \quad (3)$$

$$P(Y|do(X)) = \sum_{z} P(Y_z|do(X)) = \sum_{z} \underbrace{\overbrace{P(Z|X=x)}^{\text{Prob. for } z} \overbrace{\sum_{x_z \in X_z} P(Y|X=x_z, Z=z) P(X=x_z)}^{\text{Global CAM for } z \text{ over training set}}}_{P(Y_z|do(X))\text{: Class-specific adjusted map for } z \text{ of } x}$$

x denotes images with shape 3 * H * W, z denotes image-level label, y denotes CAM with shape H * W

Assumptions

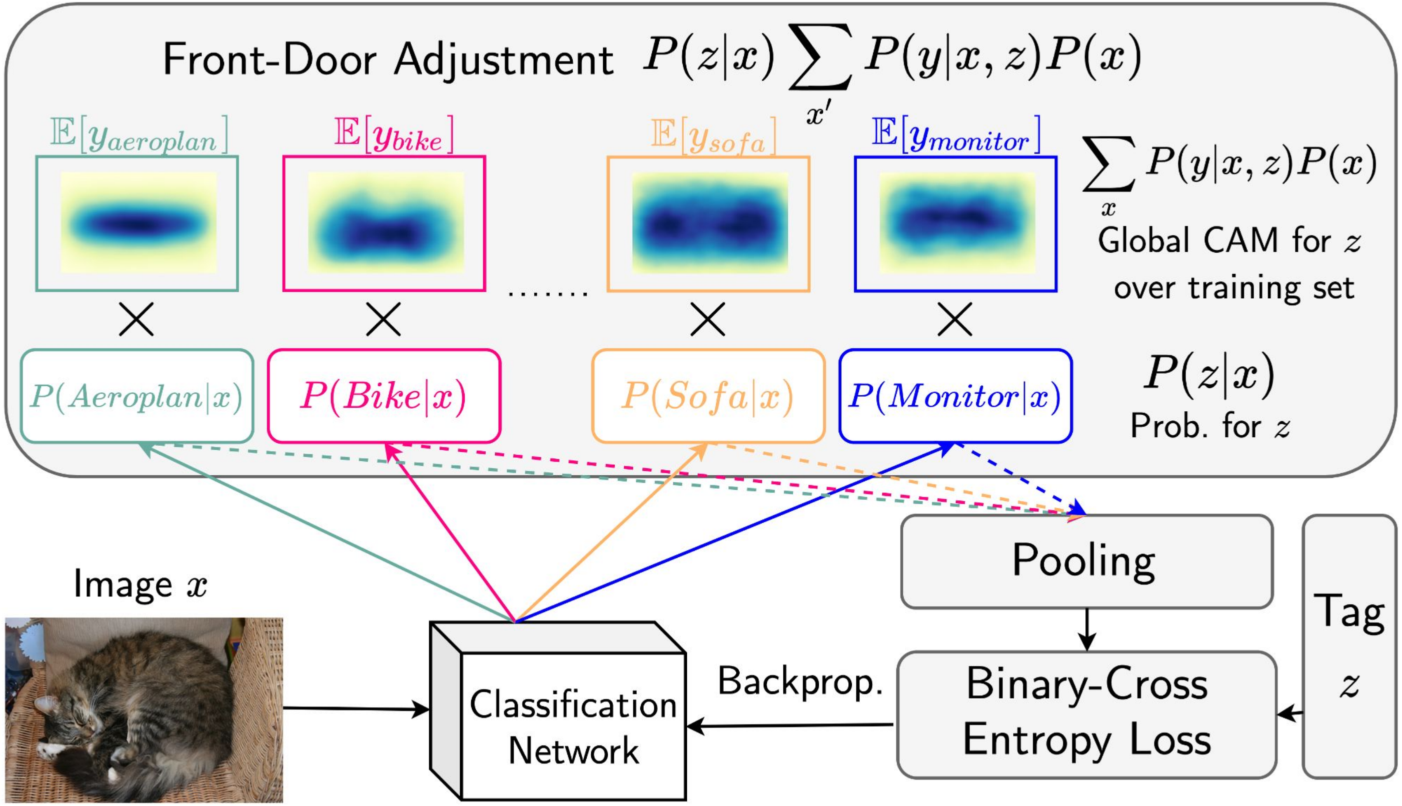
$$P(Y=y|do(X=x)) = \sum_z P(Z=z|X=x) \sum_{x'} P(Y=y|X=x', Z=z) P(X=x') \qquad (3)$$

$$P(Y|do(X)) = \sum_z P(Y_z|do(X)) = \sum_z \underbrace{\overbrace{P(Z|X=x)}^{\text{Prob. for } z} \overbrace{\sum_{x_z \in X_z} P(Y|X=x_z, Z=z) P(X=x_z)}^{\text{Global CAM for } z \text{ over training set}}}_{P(Y_z|do(X)):\ \text{Class-specific adjusted map for } z \text{ of } x}$$

- $P(Z=z|X)$: the probability of an image x for class z can be computed by the classifier.
- $P(X=x_z)$: assuming that each training sample is equiprobable, the probability of an image x of class z occurs is approximately $\frac{1}{N_z}$.
- $P(Y=y_z|X=x_z, Z=z)$: the probability distribution for the localization $y_z \in \mathbb{R}^{1\times H\times W}$ can be computed by Eq. 1 with a trained classifier.

x denotes images with shape 3 * H * W, z denotes image-level label, y denotes CAM with shape H * W

Training

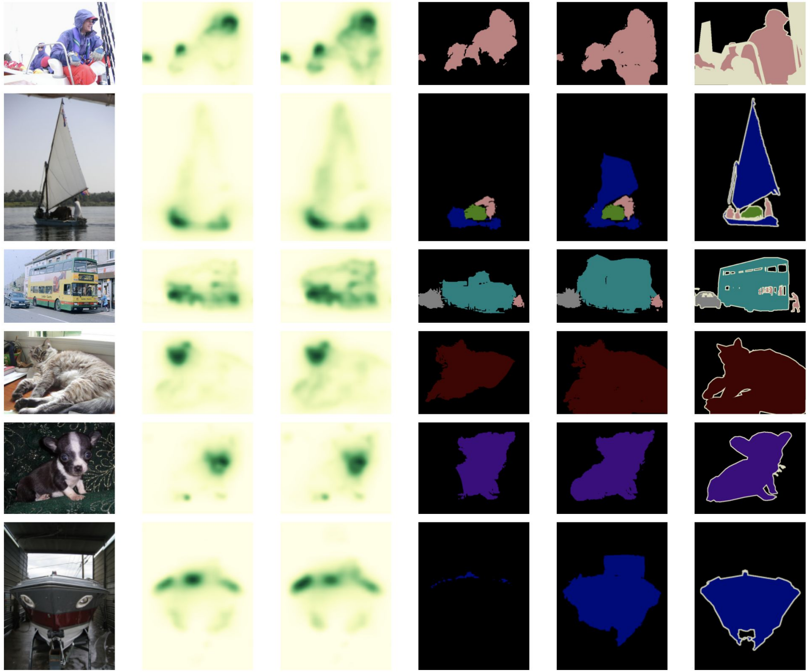


Implementation – one liner!

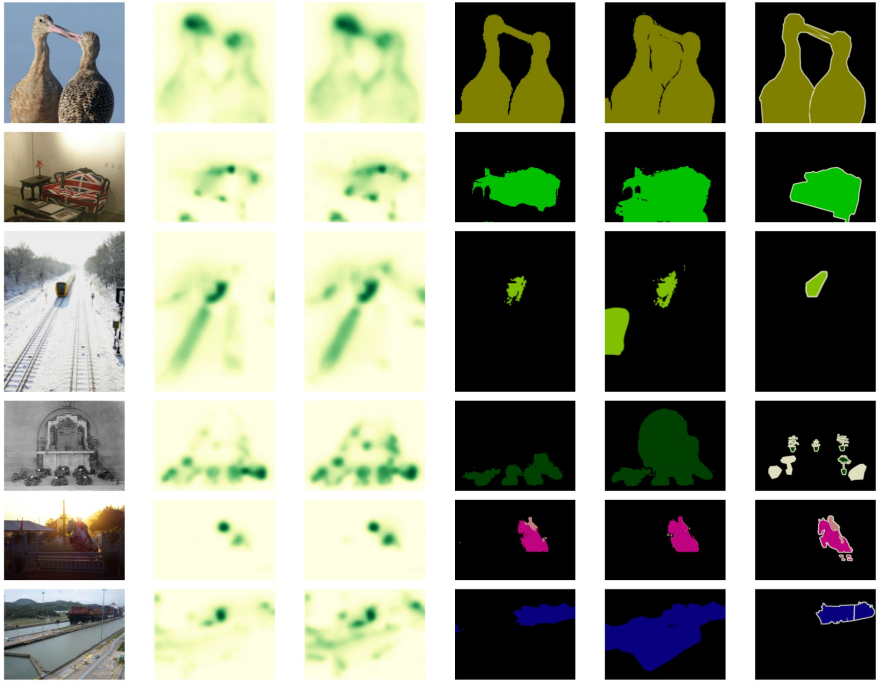
```
# Classification forward pass
x = classification_model(images) # x.shape == B * 20
# Multiply with Global CAM
# x.shape == B * 20 * H * W
# x = x.unsqueeze(2).unsqueeze(2) * global_cam
# Mean Pooling
# x = torch.mean(x, dim=(2, 3)) # x.shape == B * 20
# Ours
x = torch.mean(x.unsqueeze(2).unsqueeze(2) *
               global_cam, dim=(2, 3))
# Loss
loss = torch.nn.BCELoss()(x, labels)
```

Qualitative

Success cases



Failure cases



(a)Original Image (b)CAM (c)Causal CAM (d)Pseudo-Masy by CAM (e)Pseudo-Mask by Causal CAM (f) GT

Quantitative

Method	Type	Backbone	Seed	Pseudo-Mask	`val`	`test`
CAM[72]$_{CVPR'16}$	/	ResNet50	48.3	65.9	63.5	64.8
CONTA[70]$_{NeurIPS'20}$	$\mathcal{A},\mathcal{C}$	ResNet50	48.8	67.9	65.3	66.1
CONTA+SEAM [64]	$\mathcal{A},\mathcal{C}$	ResNet38	56.2	65.4	66.1	66.7
C^{2}AM (Ours)	$\mathcal{C}$	ResNet50	52.1	69.6	67.5	67.7
AdvCAM[36]$_{CVPR'21}$	$\mathcal{I}$	ResNet50	55.6	69.9	68.1	68.0
ReCAM[11]$_{CVPR'22}$	$\mathcal{A}$	ResNet50	54.8	70.8	68.7	68.5
RCA[73]$_{CVPR'22}$	$\mathcal{M}$	ResNet38	/	74.1	72.2	72.8

Thanks!

VECTOR INSTITUTE

UNIVERSITY OF WATERLOO