Causal Class Activation Maps for Weakly-Supervised Semantic Segmentation

Yiping Wang
yiping.wang@uwaterloo.ca

Cheriton School of Computer Science
University of Waterloo
Waterloo, Ontario, Canada

Abstract

Weakly-supervised semantic segmentation (WSSS) has emerged in recent years due to its appealing requirements for training data, i.e., with only image-level labels available as supervision. Most existing WSSS methods exploit the class activation maps (CAMs) as the seeds and generate the pseudo-pixel-level ground truth to train a segmentation network. In this work, we introduce a causal inference framework to ameliorate the quality of CAMs, conducing to the performance raise of existing WSSS algorithms that rely on CAMs. Our motivation is to deconfound a set of class-specific latent *confounders* in a dataset, which are the potential cause of low-quality and poorly-localized CAMs. Due to the unobservable nature of the confounders, we present the utilization of *front-door adjustment* for causal intervention to deconfound a classification neural network, without presuming and estimating the confounders explicitly. Our proposed algorithm, Causal CAM (C^2AM), outperformed the prior causal framework for WSSS [59] by a large margin, without any additional parameters, network architecture modification, or manipulation of images, and only needs to add one more line of code in a standard classifier training loop. Furthermore, we provide an optimization interpretation of the front-door adjustment for training a classifier to explain the improvements by C^2AM. We evaluated C^2AM on PASCAL VOC 2012 and achieved mIoU 69.6% of pseudo-mask generation on the training set, and mIoU 67.5% and 67.7% on validation and test set after training DeepLabV2 on the pseudo-masks. Our implementation and model weights for reproducibility are released at https://github.com/yiping-wang/c2am/

1 Introduction

Semantic segmentation is the task of classifying each pixel of an image into its corresponding semantic class [30]. It is a core and fundamental building block for many visual computing applications, such as scene understanding [22] and biomedical image analysis [19]. Previously, training deep learning models for semantic segmentation requires pixel-level annotations [8, 39, 40, 48], which are expensive and laborious to obtain, e.g., annotating a 500×500 natural image with pixel-level ground truth for an object can easily take ten times longer than creating a bounding box around it [13]. In contrast, image-level labels are the easiest and cheapest to collect, which merely take almost one second per object-category [42]. With the massive amount of image data available nowadays, weakly-supervised learning for semantic segmentation has been gaining attention for its "weak" requirements for labels of training data, and "weak" emphasizes the cheaper labelling cost at image-level [4, 53].

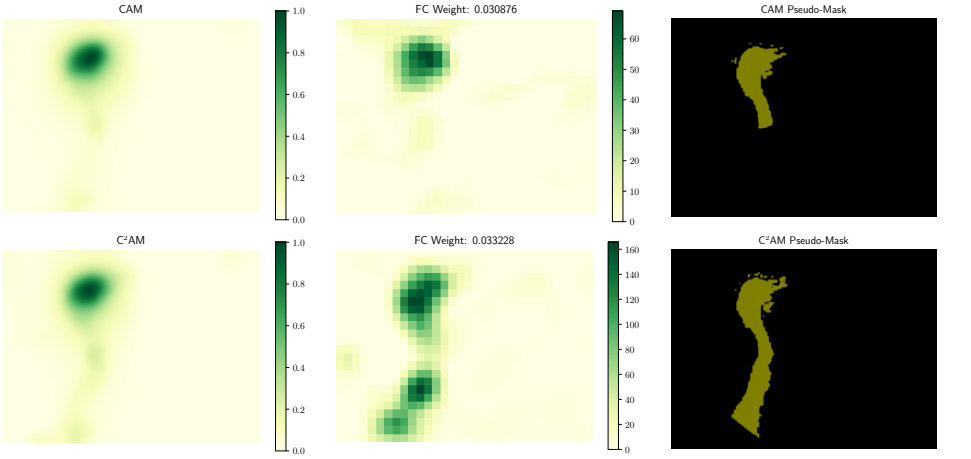


Figure 1: Visual inspections of the Class Activation Map (CAM) and Causal Class Activation Map (C^{2}AM). C^{2}AM highlights more regions of the object of interest than CAM, which ameliorates the quality of pseudo-masks. The middle column shows one of the feature maps and its fully-connected layer weight.

Class Activation Maps (CAMs) [62, 71] has been a popular and powerful starting point, or *seed*, for most weakly-supervised semantic segmentation (WSSS) algorithms [1, 2, 4, 10, 31, 63]. CAMs localize the most discriminative, albeit coarse and incomplete, regions of a semantic class in an image as the *seed* areas that are further exploited and expanded to obtain the pixel-level pseudo-masks [24, 65], which are treated as the pseudo-ground-truth for training a standard pixel-level supervised semantic segmentation algorithm [4, 70].

Despite the successful applications of CAMs for WSSS, CAMs are generated from a classification network, and a classification network does not always learn the *causal features* that are robust in any confounding context, e.g., the foreground object features are invariant in any different background context [59]. This happens as classification does not necessitate a precise localization of the objects, the network could take advantage of the spurious correlation in the confounding context as long as it benefits the prediction when the training and testing data are i.i.d. Unfortunately, enlarging the scale of the datasets won't alleviate this bias [52], as such biases are embedded in the nature of data, as indicated in Zipf's law [47]. Indeed, "yellow banana" occurs more often than "green banana" in reality. Thus, such spurious correlations and bias learned in a classification network could pose an inferior quality of CAMs, and lead to poor performance of WSSS algorithms that are based on CAMs. The fundamental solution to learn the robust causal features is by *causal intervention* [59, 69]. In this paper, we propose an explainable causal inference framework to adjust the confounding variables in the classification network by *front-door adjustment* [45, 57] to generate high-quality CAMs for any existing WSSS that requires them.

Our proposed method, Causal CAM (C^{2}AM), is a simple yet effective algorithm for enhancing CAM quality, which conduces to better pseudo-masks for training a standard pixel-level semantic segmentation network. It requires no extra parameters, modification of the classification network architecture, or manipulating of images. Further, the implementation of C^{2}AM is straightforward, it only requires one additional line of code inside the standard training loop of the classification network. The schematic of C^{2}AM is presented in Figure 3.

2 Related Works

Weakly-supervised learning for semantic segmentation have three main sources of supervision: image-level tags [1, 2, 4, 6, 7, 9, 10, 14, 15, 23, 24, 27, 31, 33, 36, 43, 44, 46, 49, 55, 64, 68], image scribbles [38, 56, 57, 58, 66], and bounding boxes [11, 26, 29, 32, 34, 37, 54]. Arguably, training models for semantic segmentation with only image-level tags supervision is more challenging than with other weak supervisions [4], such as scribbles or bounding boxes, which presume localization cues are available. In this work, we focus on WSSS with image-level supervision. Pinheiro *et al*. [46] formulated WSSS as a Multiple Instance Learning [25] problem, and then aggregated the feature maps to class scores by the `LogExpSum` operator, and refined the pixel-level segmentation with image-level priors. Kolesnikov *et al*. [29] demonstrated the idea of exploiting and expanding the localization cues from CAMs [71] to constrain the segmentation to conform with the boundaries of objects. Ahn *et al*. [2] introduced the Inter-pixel Relation Network (IRN) to estimate a class-agnostic instance map and pairwise semantic affinities for additional information complementary to CAMs, which were employed together to accurately propagate the seeds within the object boundaries.

The consensus of CAMs [71] is their limitations of attaining only the most discriminative and class-specific evidence. Intuitively, if high-quality CAMs could be generated, all existing WSSS algorithms rely on them could enjoy a performance improvement, without altering the algorithms. Thus, besides the vanilla CAM [71], a line of work for *Seed Generation*, focuses on producing and refining high-quality CAMs. Li *et al*. [36] applies CAM on original images to generate the masked images and minimizes the model cross-entropy loss to force the model to attain features in the rest regions in the next training iterations. Jo *et al*. [27] proposed a method that minimizes the differences between the features from different regions of patches and the entire image. Lee *et al*. [35] proposed AdvCAM which formulated the improvement of the quality of CAM as an adversarial attack problem to manipulate the images for extending the discriminative regions of a target object. Chen *et al*. [10] argued that the binary cross-entropy (BCE) loss widely in CAM is the crux of unsatisfactory pseudo-mask generation, and Chen *et al*. proposed reactivating the converged CAM trained with BCE, by using softmax cross-entropy loss to train another FC layer, to explicitly enforces class exclusion within multi-labelled images. In contrast to prior works, C^2AM requires no additional parameters, network architecture changes, or manipulation of the images.

Causal Inference [17, 45] and Causal Representation Learning [61, 62], have been explored in the Computer Vision community to remove the spurious bias and modularize reusable features [6, 50, 59, 61, 67, 69]. Causal features, which are robust in any confounding context, are not automatically learned by deep learning models. This is because the confounders could mislead the attention (e.g., CAM) to capture spurious correlations that benefit the prediction when the training and testing data are i.i.d. [59]. Zhang *et al*. [69] proposed a structural causal model [45] to perform causal interventions to mitigate the context confounder by *back-door adjustment* [45], and also introduced the Context Adjustment (CONTA) framework which could be integrated with other WSSS algorithms. C^2AM is formulated in a causal framework, to encourage the generation of accurate and complete CAM, by *front-door adjustment* [45, 67]. C^2AM outperformed CONTA for WSSS by a large margin. To the best of our knowledge, we are the first to apply the front-door adjustment for enhancing CAM quality to raise the performance of WSSS algorithms.

3 Causal Class Activation Maps

3.1 Class Activation Map

To generate CAM [71], the first step is to train a multi-label classification network with global average pooling (GAP) layer followed by a FC prediction layer, and minimizing the binary-cross entropy (BCE) loss. Once the model converges, the CAM of class z in an image $\mathbf{x}$ can be extracted by

$$\mathrm{CAM}_z(\mathbf{x}) = \frac{\mathrm{ReLU}(\mathcal{A}_z)}{\max(\mathrm{ReLU}(\mathcal{A}_z))}, \ \mathcal{A}_z = \mathbf{w}_z^T f(\mathbf{x}) \tag{1}$$

where $\mathbf{w}_z$ denotes the FC weights corresponds to the z-class, and $f(\mathbf{x})$ denotes the feature map of $\mathbf{x}$ before GAP. For instance, for a ResNet50 [21] trained on PASCAL VOC 2012 dataset for multi-label classification, $f(\mathbf{x}) \in \mathbb{R}^{2048\times 32\times 32}$ and $\mathbf{w} \in \mathbb{R}^{20\times 2048}$.

3.2 Structural Causal Model

Our motivation is that if the quality of CAMs can be enhanced, other WSSS algorithms (such as IRN [2]) employ CAMs should expect a performance boost. In Section 3.2, we detail the Structural Causal Model (SCM) described in Figure 2(c). In Section 3.3, we introduce our method of applying front-door adjustment for a pre-trained classification network. In Section 3.4, we justify mathematically the improvement caused by C^{2}AM from the optimization point of view. A short introduction to the structural causal model, back-door adjustment, and front-door adjustment is given in Appendix C.

To analyze the causality between image $x \in \mathbb{R}^{3\times H\times W}$, image-level tag $z \in \mathbb{R}$, pixel-level localization $y \in \mathbb{R}^{1\times H\times W}$, and a set of class-specific latent confounders $\{C_z\}$, we present a SCM as illustrated in Figure 2(c). Here, confounders can be any factors that trick the classifier to attend spurious localization via $P(Y|X)$, such as context [53, 67, 69, 70], content and style [41]. Prior works [41, 59, 69] require identification and estimation of the confounders explicitly due to the demand from back-door adjustment. Conversely, the front-door adjustment does not necessitate the knowledge of the confounders [45, 67]. C^{2}AM does not pinpoint the confounders directly, as C^{2}AM utilized front-door adjustment for deconfouding.

Moreover, we argue it is necessary and essential to presume the existence of class-specific confounder C_z, since a different class of foreground objects has distinct properties and the confounder could create spurious correlation differently for each class. For instance, suppose the confounder is "context", the confounder for `horse` might associate `horse` with `person` as both objects co-occur frequently, while the confounder for `aeroplane` might correlate `aeroplane` with the unlabeled background object, such as `cloud`.

$\mathbf{C_z \to X}$: This edge indicates the data generation process of an image x by the class-specific confounder C_z, such as content and style [16, 20, 28, 41, 60], as well as context [53, 67, 69, 70]. The content could include various kinds of objects but still belong to the same semantic class, such as the dogs in a dataset might have different species, but they are considered as `dog`, and the style contains colours, lighting conditions and camera lens characteristics. To stimulate all possible combinations of the data generation factors, we propose to use causal intervention for this link, to pursue the true causality from image x to localization cue y.

$\mathbf{C_z \to Y_z}$: This link emphasizes the attentions Y of a classifier are the effect of the class-specific confounder C_z. For a classification task, the confounder C_z might help learn a better association between image x and its label z, especially when training and test set are i.i.d. [59,

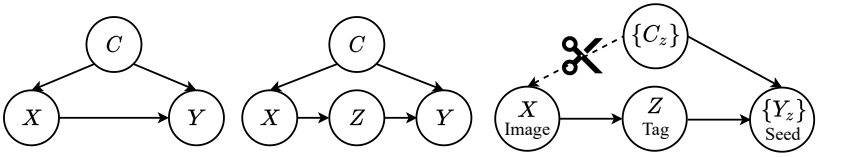


Figure 2: Structural Causal Models (SCMs) for illustrating the fundamental (a) Back-Door model and (b) Front-Door model (c) presents our proposed SCM to analyze the generation of the localization Y_z, in which the image X, image-level tag Z, pixel-level localization Y_z, and class-specific confounder C_z can be formulated in a front-door model. The "scissor" in (c) denotes causal intervention. See Section 3.2 and Appendix C for details.

62]. For instance, one commonly assumed confounder, context [53, 67, 69, 70], introduces the non-causal features via $P(Y|X)$, e.g., `bird` co-occurs frequently with `tree`, and $P(Y|X)$ might mistakenly focus on `tree` features instead of `bird`.

$X \rightarrow Z$: This link indicates that the image-level tag Z is an effect of image X. As the image-level tag Z is determined and annotated by the dataset collector, and the objects in a dataset are not annotated exhaustively.

$Z \rightarrow Y_z$: This edge emphasizes that a weak localization cue Y is an effect of the image-level tag z, as the computation of CAM from the trained classifier in Eq. 1 requires an image-level label. Hence, the image-level tag Z is a *mediator* [45] that helps us estimate the causal effect of image X on localization Y_z for a class z.

3.3 Front-Door Adjustment for a Classification Neural Net

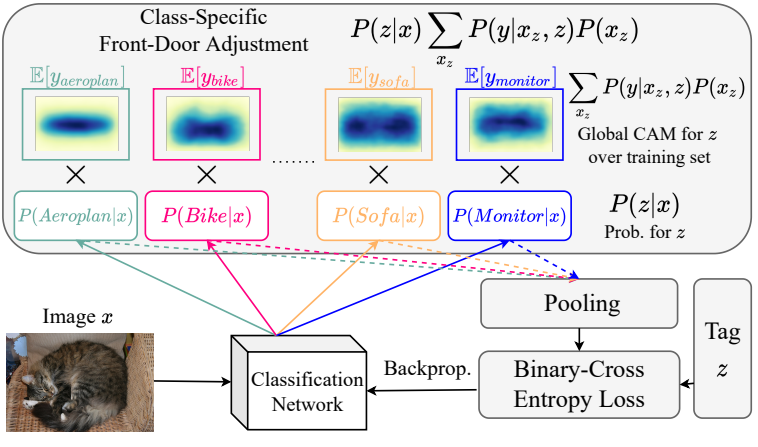


Figure 3: Overview of the proposed method of applying front-door adjustment for a classification network. See details in Section 3.3.

The overview of our approach is illustrated in Figure 3. For a target multi-label dataset, a standard classification neural network is first trained with BCE loss. Afterwards, a class-specific CAM of each image can be extracted by Eq. 1. The probability of an image $x \in \mathbb{R}^{3\times H\times W}$ belonging to a class z predicted by the classification network is denoted as $P(z|x)$.

Further, the distribution of a CAM for an image $x \in \mathbb{R}^{3\times H\times W}$ and a class z is denoted as $P(y|x,z)$, where $y \in \mathbb{R}^{1\times H\times W}$. In Section 3.2, we argued the assumption of the class-specific confounder C_z. Thus, to perform classification, we utilize the class-specific adjusted map for z, denoted as $P(Y_z|do(X)) = P(Z|X=x)\sum_{x_z\in X_z} P(Y|X=x_z,Z=z)P(X=x_z)$

$$P(Y|do(X)) = \sum_z P(Y_z|do(X)) = \sum_z \overbrace{P(Z|X=x)}^{\text{Prob. for } z} \overbrace{\sum_{x_z\in X_z} P(Y|X=x_z,Z=z)P(X=x_z)}^{\text{Global CAM for } z \text{ over training set}} \tag{2}$$

$P(Y_z|do(X))$: Class-specific adjusted map for z of x

The derivation of Eq. 2 is shown in Appendix C. To calculate $P(Y_z|do(X))$ in Eq. 2, it needs the prior knowledge of $P(X=x_z)$, which is the probability of an image x belongs to class z occurring. Inspired by Amrani *et al.* [3], we assume the training samples are equiprobable, i.e., $P(X=x_z)$ is a uniform distribution. $P(Y_z|do(X)) \in \mathbb{R}^{1\times H\times W}$ now can be computed with the following available quantities:

- $P(Z=z|X)$: 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.

The semantic meaning of $\sum_{x_z\in X_z} P(Y|X=x_z,Z=z)P(X=x_z)$ in Eq. 2, is the expectation of localization y_z of the entire training images for a class z. We term this quantity as *Global CAM*. Note, Global CAM only needs to be calculated once with the pre-trained classifier. Therefore, the computation overhead is negligible. Global CAM can be treated as a *prior* in the training set for the probability of the object for class z occur in each pixel. Visualizations of Global CAMs for all classes are shown in Figure 10 in Appendix. To train the classifier, we employ the Multiple Instance Learning technique [16], by pooling the adjusted attention map in Eq. 2 into a score $s_z \in \mathbb{R}$ for class z

$$s_z = \text{Pooling}(P(Z=z|X)\sum_{x_z\in X_z} P(Y=y_z|X=x_z,Z=z)P(X=x_z)) \tag{3}$$

Thus, s_z in Eq. 3 is treated as the prediction score to train the classification network to minimize the BCE loss. After the training is converged, the enhanced CAMs can be produced by Eq. 1 with the front-door adjusted classifier. The implementation simply requires one more line of code for a classifier training loop. See Appendix B for details.

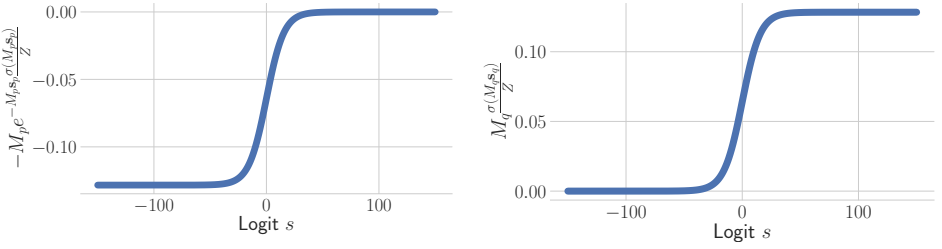
3.4 Optimization Interpretation

The BCE loss for multi-label classification is defined as

$$\mathcal{L}_{bce} = -\frac{1}{Z}\sum_i \mathbf{z}_i \log\sigma(\mathbf{s}_i) + (1-\mathbf{z}_i)\log(1-\sigma(\mathbf{s}_i)) \tag{4}$$

where Z denotes the number of classes, $\mathbf{z}$ denotes the ground-truth label, and $\mathbf{s}$ denotes the logit. The gradient of $\mathcal{L}_{bce}$ w.r.t. logit $\mathbf{s}$ can be derived as

$$\nabla_{\mathbf{s}}\mathcal{L}_{bce} = \frac{\sigma(\mathbf{s})-\mathbf{z}}{Z} \tag{5}$$



(a) (b)

Figure 4: For illustration purpose, suppose the logit s by Eq. 6 ranges from -150 to 150, and the visualizations of (a) the $\nabla_s \mathcal{L}_{bce}$ w.r.t. positive `aeroplane` label (Eq. 7), and (b) the $\nabla_s \mathcal{L}_{bce}$ w.r.t. to other negative labels (Eq. 8). The plot contains only M_z for class `aeroplan` in the PASCAL VOC 2012 dataset. See Appendix A for a complete plot.

Suppose the Pooling in Eq. 3 is a Global Average Pooling (GAP) operator. Rewrite Eq. 2 as

$$\begin{aligned} s_z &= \text{GAP}(P(Z=z|X=x) \sum_{x_z \in X_z} P(Y|X=x_z, Z=z) P(X=x_z)) \\ &= P(Z=z|X=x)\text{GAP}(\sum_{x_z \in X_z} P(Y|X=x_z, Z=z) P(X=x_z)) \\ &= P(Z=z|X=x) M_z \end{aligned} \tag{6}$$

where M_z is a constant computed by the GAP operator on the Global CAM for class z. Essentially, the logits in Eq. 4 and Eq. 5 are multiplied by a constant M_z.

For positive class p, $\mathbf{z}_p = 1$, the gradient of $\mathcal{L}_{bce}$ w.r.t. class p is

$$\nabla_{\mathbf{s}_p} \mathcal{L}_{bce} = -\frac{1}{Z} \nabla_{\mathbf{s}_p} (\mathbf{z}_p \log \sigma(M_p \mathbf{s}_p)) = -\frac{1}{Z} \frac{M_p}{e^{M_p \mathbf{s}_p} + 1} = -M_p e^{-M_p \mathbf{s}_p} \frac{\sigma(M_p \mathbf{s}_p)}{Z} \tag{7}$$

For negative class q, $\mathbf{z}_q = 0$, the gradient of $\mathcal{L}_{bce}$ w.r.t. class q is

$$\nabla_{\mathbf{s}_q} \mathcal{L}_{bce} = -\frac{1}{Z} \nabla_{\mathbf{s}_q} ((1-\mathbf{z}_q) \log(1 - \sigma(M_q \mathbf{s}_q))) = \frac{1}{Z} \frac{M_q e^{M_q \mathbf{s}_q}}{e^{M_q \mathbf{s}_q} + 1} = M_q \frac{\sigma(M_q \mathbf{s}_q)}{Z} \tag{8}$$

The visualizations for $\nabla_{\mathbf{s}_p} \mathcal{L}_{bce}$ w.r.t. positive `aeroplane` label (Eq. 7) and $\nabla_{\mathbf{s}_q} \mathcal{L}_{bce}$ w.r.t. other negative labels (Eq. 8) for the logit s from -150 to 150 are presented in Figure 4. From Figure 4(a), it can be deduced that for positive labels and positive logits, the gradient is nearly 0, which indicates that for correct predictions, the performance of the classifier is not impacted. However, for positive labels and negative logits, the gradient is non-zero and negative, which means that the gradient continues to push the weights of the classifier to minimize the BCE loss. Similar analysis can also be applied to Figure 4(b).

Hence, the front-door adjustment applied in Section 3.3 fine-tunes the classifier with class-specific prior M_z. In Figure 1, for the i^{th} feature map F_i and its corresponding fully-connected layer weight w_i, F_i from the adjusted classifier shows higher responses and highlights more regions than the unadjusted classifier. Essentially, as M_z are in the range of 0.04 to 0.2 (see Figure A in Appendix), the network learns to adjust the numeric value of the feature map, to counter the effect of the M_z. By inspecting the kernels in each convolutional layer, we found most of the weights in the adjusted classifier increased by 0.0001 to 0.001. Thus, the accumulated adjustments of the weights cause the enhancement of the numeric

value of the feature maps. Based on this observation, we argue that the feature maps can be expanded and enhanced by simply multiplying the logits by a constant below 1. We test this idea in Section 4.2.1.

4 Experimental Results

In Section 4.1, we introduce the dataset, evaluation metric, and state-of-the-art algorithms. In Section 4.2, the effectiveness of C^2AM is demonstrated both quantitatively and qualitatively, and the comparisons with the state-of-the-arts are reported. For reproducibility, the random seed is fixed as 0 for all experiments.

4.1 Settings

PASCAL VOC 2012 [12] is a commonly used dataset for evaluating semantic segmentation algorithms, it contains 20 foreground object categories and 1 background class. Following the conventional practice in related works [2, 10, 69], the training set is augmented with additional data proposed by Hariharan *et al*. [18]. In total, there are 10,582 images in the training set, 1,499 images in the validation set, and 1456 images in the test set.

For training our method, a ResNet50 [21] is pre-trained on the PASCAL VOC 2012 for the multi-label classification task. Afterwards, we generate the Global CAM for each class, and the classifier is then trained by the front-door adjustment as shown in Figure 3.

4.2 Results

4.2.1 Quantitative evaluation

Quantitative evaluations are shown in Table 1. Three types of masks are evaluated. First, seed area masks are produced by CAMs. Second, pseudo-masks constructed by IRN [2] based on CAM seed area. Third, segmentation masks are predicted by DeepLabV2 trained on the pseudo-masks. The standard quantitative evaluation metric, mean Intersection over Union (mIoU), was computed against the ground-truth pixel-level masks. Moreover, we compared C^2AM with three CAM generation algorithms, vanilla CAM [71], AdvCAM [35], ReCAM [10], and one causal inference algorithm, CONTA [69]. Specifically, AdvCAM [35] requires successive manipulation of images in every iteration, ReCAM [10] and CONTA [69] require additional network parameters. C^2AM does not require any additional parameters, network architecture changes, or manipulation of images.

As shown in Table 1, C^2AM outperforms the vanilla CAM (+3.7%) and the prior causal framework CONTA (+1.7%) by a large margin in the Pseudo-Mask generation section, which also improved the mIoU on validation and test set of the PASCAL VOC 2012 dataset predicted by a standard DeepLabV2 trained on pseudo-masks. Nevertheless, C^2AM didn't outperform the state-of-the-art, such as AdvCAM [35], ReCAM [10] and RCA [72]. We analyze the reasons in the qualitative evaluation in Section 4.2.2. Interestingly, we test the idea of multiplying a constant, such as 0.2, to the classification logits, during training the classification network, and it also outperforms the vanilla CAM for all three types of masks.

Method	Type	Backbone	Seed	Pseudo-Mask	val	test
CAM[71]$_{CVPR'16}$	/	ResNet50	48.3	65.9	63.5	64.8
CONTA[69]$_{NeurIPS'20}$	$\mathcal{A},\mathcal{C}$	ResNet50	48.8	67.9	65.3	66.1
CONTA+SEAM [63]	$\mathcal{A},\mathcal{C}$	ResNet38	56.2	65.4	66.1	66.7
CAM+Constant	/	ResNet50	51.3	69.4	67.2	67.4
C^{2}AM (Ours)	$\mathcal{C}$	ResNet50	52.1	69.6	67.5	67.7
AdvCAM[55]$_{CVPR'21}$	$\mathcal{I}$	ResNet50	55.6	69.9	68.1	68.0
ReCAM[10]$_{CVPR'22}$	$\mathcal{A}$	ResNet50	54.8	70.8	68.7	68.5
RCA[72]$_{CVPR'22}$	$\mathcal{M}$	ResNet38	/	74.1	72.2	72.8

Table 1: Quantitative comparison with state-of-the-arts in mIoU (%) on the PASCAL VOC 2012 dataset. IRN [2] is the default algorithm to produce the Pseudo-Masks on the CAM seeds generated by various algorithms. The results of the prior causal framework for WSSS, CONTA, include both IRN and SEAM [63]. $\mathcal{A}$ denotes using additional parameters. $\mathcal{C}$ denotes employing causal inference. $\mathcal{I}$ denotes manipulating of images. $\mathcal{M}$ denotes the utilization of a memory bank. "CAM+Constant" denotes multiplying the classification logit by a constant (0.2) during training. DeepLabV2 [3] is trained on the pseudo-masks, and the mIoUs of its segmentation prediction on the validation and test sets are reported. Our hyperparameter settings can be found in Appendix D.

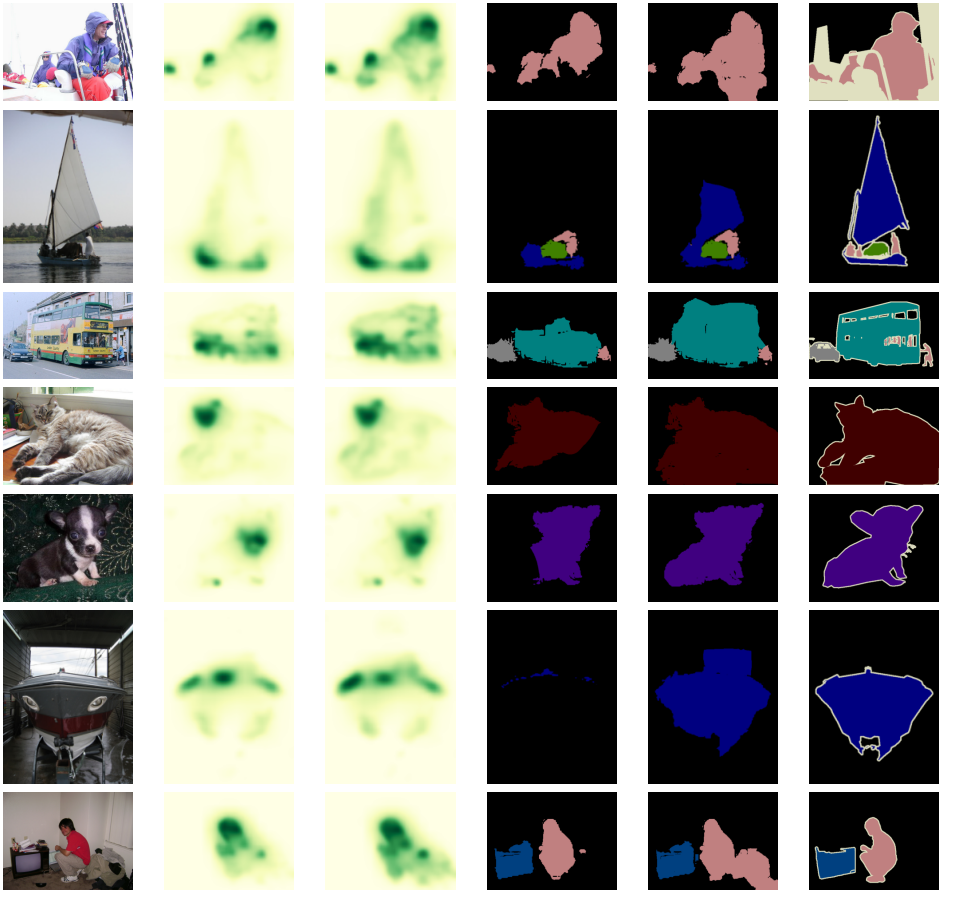
4.2.2 Qualitative evaluation

Qualitative evaluations are shown in Figure 5. More qualitative evaluations are shown in Figure 12 in Appendix. In most cases, C^{2}AM does enhance the quality of the CAM by covering more parts of the objects, which causes the improvement of pseudo-masks. However, as shown in the last row in Figure 5, it is noticeable that C^{2}AM tends to over-localize the object, which causes the pseudo-mask contains over-segmentation areas. This is the main reason that we didn't outperform the state-of-the-arts. More failure cases are shown in Figure 13 in Appendix. The predictions of DeepLabV2 trained on the C^{2}AM pseudo-masks on the test set are shown in Figure 11 in Appendix.

As we use the Global CAM in our method described in Section 3.3, we qualitatively show C^{2}AM is also shift-invariant w.r.t. the input images, as illustrated in Figure 6. We randomly crop and resize the input images to make the object of interest appear in various locations in the images. C^{2}AM is able to detect and localize the object. Thus, we believe the Global CAM can be thought of as a training dataset knowledge to help the classification network to expand its feature map, which caused the enhancement of class activation maps.

5 Conclusions

In this work, we aim to ameliorate the quality of CAMs, which intuitively should conduce to the performance of any existing WSSS algorithms that utilize CAMs. We formulate the generation of CAM in a front-door model from causality, and we quantitatively and qualitatively demonstrated the simplicity and effectiveness of this method. C^{2}AM outperforms the vanilla CAM and prior causal framework CONTA, while didn't reach the state-of-the-arts. Further research is required to alleviate the over-localization issue, such as applying regularization.



(a) (b) (c) (d) (e) (f)

Figure 5: Qualitative evaluation of C^{2}AM and its pseudo-mask generated by IRN [2] on the PASCAL VOC 2012 training set. (a) original images. (b) CAM [38]. (c) C^{2}AM. (d) pseudo-masks generated by IRN with CAMs. (e) pseudo-masks generated by IRN with C^{2}AM. (f) pixel-level ground truth. C^{2}AM highlights more regions of objects than CAM, which conduces to a better quality of pseudo-masks. However, C^{2}AM tends to produce over-localized objects, which causes the IRN to over-segment the objects. The last row shows a failure case due to over-activation. See more qualitative results in Appendix F.

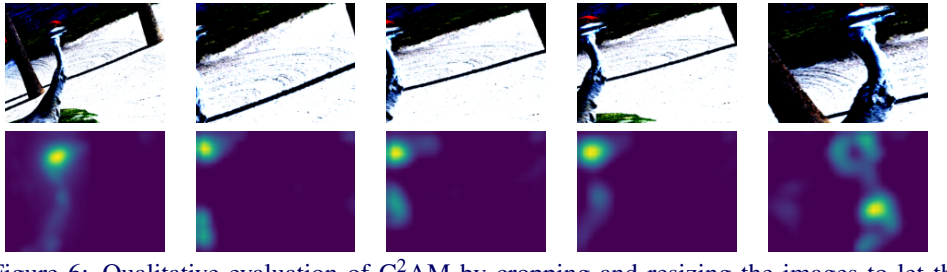


Figure 6: Qualitative evaluation of C^{2}AM by cropping and resizing the images to let the object of interest to appear in various locations of the images. C^{2}AM is able to detect the objects despite it utilizes the Global CAM information.

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[2] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[3] Elad Amrani and Alex Bronstein. Self-supervised classification network. *arXiv preprint arXiv:2103.10994*, 2021.

[4] Nikita Araslanov and Stefan Roth. Single-stage semantic segmentation from image labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[5] Yuval Atzmon, Felix Kreuk, Uri Shalit, and Gal Chechik. A causal view of compositional zero-shot recognition. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1462–1473. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1010cedf85f6a7e24b087e63235dc12e-Paper.pdf.

[6] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[7] Arslan Chaudhry, Puneet K. Dokania, and Philip H. S. Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. In *Proceedings of the 28th British Machine Vision Conference (BMVC)*, July 2017.

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4):834–848, 2018. doi: 10.1109/TPAMI.2017.2699184.

[9] Qi Chen, Lingxiao Yang, Jianhuang Lai, and Xiaohua Xie. Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation. *arXiv preprint arXiv:2203.02909*, 2022.

[10] Zhaozheng Chen, Tan Wang, Xiongwei Wu, Xian-Sheng Hua, Hanwang Zhang, and Qianru Sun. Class re-activation maps for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2203.00962*, 2022.

[11] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2): 303–338, June 2010.

[13] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, Jan 2015. ISSN 1573-1405. doi: 10.1007/s11263-014-0733-5. URL https://doi.org/10.1007/s11263-014-0733-5.

[14] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[15] Ruochen Fan, Qibin Hou, Ming-Ming Cheng, Gang Yu, Ralph R. Martin, and Shi-Min Hu. Associating inter-image salient instances for weakly supervised semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[16] Leon Gatys, Alexander Ecker, and Matthias Bethge. A neural algorithm of artistic style. *arXiv*, 08 2015. doi: 10.1167/16.12.326.

[17] Tomas Geffner, Javier Antoran, Adam Foster, Wenbo Gong, Chao Ma, Emre Kiciman, Amit Sharma, Angus Lamb, Martin Kukla, Nick Pawlowski, et al. Deep end-to-end causal inference. *arXiv preprint arXiv:2202.02195*, 2022.

[18] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *2011 International Conference on Computer Vision*, pages 991–998, 2011. doi: 10.1109/ICCV.2011.6126343.

[19] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18–31, 2017. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2016.05.004. URL https://www.sciencedirect.com/science/article/pii/S1361841516300330.

[20] Mohammad Havaei, Ximeng Mao, Yiping Wang, and Qicheng Lao. Conditional generation of medical images via disentangled adversarial inference. *Medical Image Analysis*, 72:102106, 2021. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2021.102106. URL https://www.sciencedirect.com/science/article/pii/S1361841521001523.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

[22] Markus Hofmarcher, Thomas Unterthiner, José Arjona-Medina, Günter Klambauer, Sepp Hochreiter, and Bernhard Nessler. *Visual Scene Understanding for Autonomous Driving Using Semantic Segmentation*, pages 285–296. Springer International Publishing, Cham, 2019. ISBN 978-3-030-28954-6. doi: 10.1007/978-3-030-28954-6_15. URL https://doi.org/10.1007/978-3-030-28954-6_15.

[23] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances*

in Neural Information Processing Systems, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/c042f4db68f23406c6cecf84a7ebb0fe-Paper.pdf.

[24] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[25] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

[26] Zongliang Ji and Olga Veksler. Weakly supervised semantic segmentation: From box to tag and back. In *Proceedings of the 32nd British Machine Vision Conference (BMVC)*, November 2021.

[27] Sanghyun Jo and In-Jae Yu. Puzzle-cam: Improved localization via matching partial and full features. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 639–643, 2021. doi: 10.1109/ICIP42928.2021.9506058.

[28] Hadi Kazemi, Seyed Mehdi Iranmanesh, and Nasser Nasrabadi. Style and content disentanglement in generative adversarial networks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 848–856, 2019. doi: 10.1109/WACV.2019.00095.

[29] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[30] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollar. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[31] Alexander Kolesnikov and Christoph H. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 695–711, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46493-0.

[32] Viveka Kulharia, Siddhartha Chandra, Amit Agrawal, Philip Torr, and Ambrish Tyagi. Box2seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 290–308, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58583-9.

[33] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[34] Jungbeom Lee, Jooyoung Choi, Jisoo Mok, and Sungroh Yoon. Reducing information bottleneck for weakly supervised semantic segmentation. *Advances in Neural Information Processing Systems*, 34, 2021.

[35] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4071–4080, June 2021.

[36] Kunpeng Li, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Yun Fu. Tell me where to look: Guided attention inference network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[37] Qizhu Li, Anurag Arnab, and Philip H.S. Torr. Weakly- and semi-supervised panoptic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[38] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[39] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[40] Shervin Minaee, Yuri Y. Boykov, Fatih Porikli, Antonio J Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. doi: 10.1109/TPAMI.2021.3059968.

[41] Jovana Mitrovic, Brian McWilliams, Jacob C Walker, Lars Holger Buesing, and Charles Blundell. Representation learning via invariant causal mechanisms. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=9p2ekP904Rs.

[42] Dim P. Papadopoulos, Alasdair D. F. Clarke, Frank Keller, and Vittorio Ferrari. Training object class detectors from eye tracking data. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 361–376, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

[43] George Papandreou, Liang-Chieh Chen, Kevin P. Murphy, and Alan L. Yuille. Weakly- and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[44] Deepak Pathak, Philipp Krahenbuhl, and Trevor Darrell. Constrained convolutional neural networks for weakly supervised segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[45] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009. ISBN 052189560X.

[46] Pedro O. Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[47] William J Reed. The pareto, zipf and other power laws. *Economics Letters*, 74 (1):15–19, 2001. ISSN 0165-1765. doi: https://doi.org/10.1016/S0165-1765(01)00524-9. URL https://www.sciencedirect.com/science/article/pii/S0165176501005249.

[48] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

[49] Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. *arXiv preprint arXiv:2203.02664*, 2022.

[50] Axel Sauer and Andreas Geiger. Counterfactual generative networks. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=BXewfAYMmJw.

[51] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

[52] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[53] Feifei Shao, Yawei Luo, Li Zhang, Lu Ye, Siliang Tang, Yi Yang, and Jun Xiao. Improving weakly supervised object localization via causal intervention. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3321–3329, 2021.

[54] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[55] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 347–365, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58536-5.

[56] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[57] Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On regularized losses for weakly-supervised cnn segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[58] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[59] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3091–3100, October 2021.

[60] Wei Wang, Junyu Gao, and Changsheng Xu. *Weakly-Supervised Video Object Grounding via Stable Context Learning*, page 760–768. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450386517. URL https://doi.org/10.1145/3474085.3475245.

[61] Wei Wang, Junyu Gao, and Changsheng Xu. *Weakly-Supervised Video Object Grounding via Stable Context Learning*, page 760–768. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450386517. URL https://doi.org/10.1145/3474085.3475245.

[62] Yixin Wang and Michael I Jordan. Desiderata for representation learning: A causal perspective. *arXiv preprint arXiv:2109.03795*, 2021.

[63] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[64] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[65] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S. Huang. Revisiting dilated convolution: A simple approach for weakly- and semi-supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[66] Jia Xu, Alexander G. Schwing, and Raquel Urtasun. Learning to segment under various forms of weak supervision. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3781–3790, 2015. doi: 10.1109/CVPR.2015.7299002.

[67] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9847–9857, June 2021.

[68] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[69] Dong Zhang, Hanwang Zhang, Jinhui Tang, Xian-Sheng Hua, and Qianru Sun. Causal intervention for weakly-supervised semantic segmentation. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 655–666. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/07211688a0869d995947a8fb11b215d6-Paper.pdf.

[70] Fei Zhang, Chaochen Gu, Chenyue Zhang, and Yuchao Dai. Complementary patch for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7242–7251, October 2021.

[71] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[72] Tianfei Zhou, Meijie Zhang, Fang Zhao, and Jianwu Li. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. *arXiv preprint arXiv:2203.09653*, 2022.

A Figures

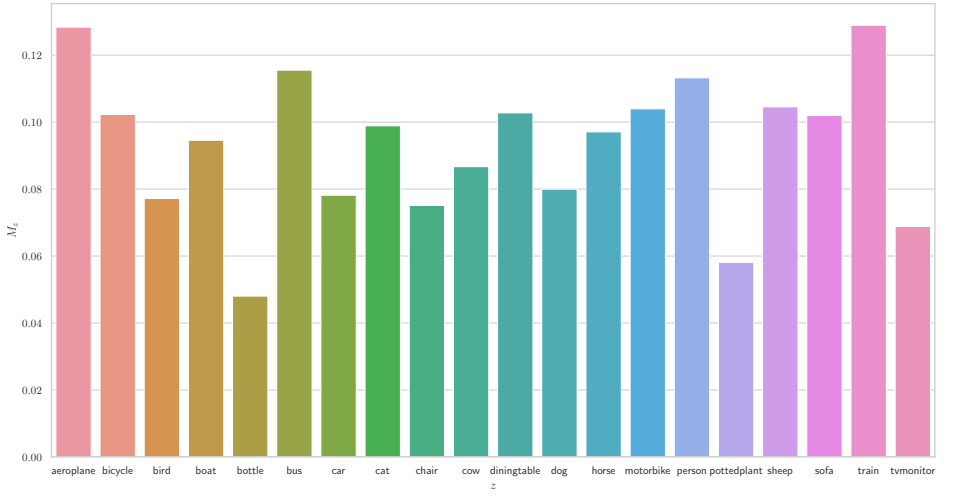


Figure 7: Visualizations for the constants M_z computed by Eq. 6 for all classes in PASCAL VOC 2012.

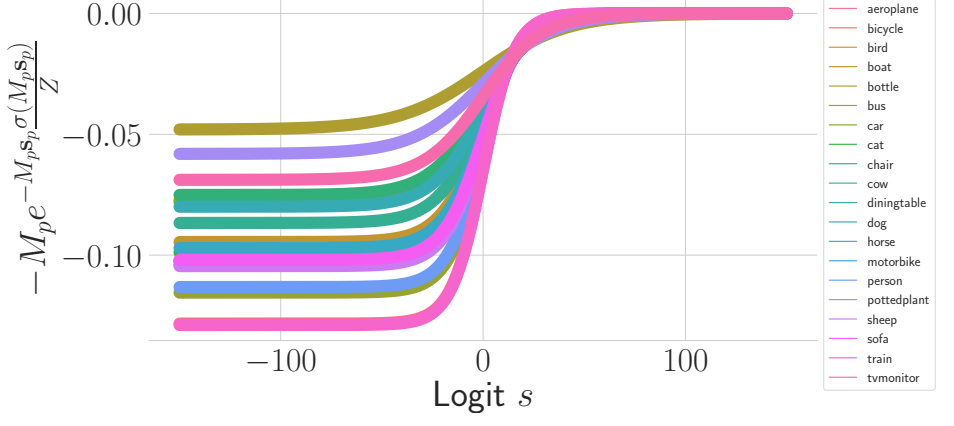


Figure 8: Visualizations for logit s ranges from -150 to 150 on the gradient $-M_z e^{-M_z s} \frac{\sigma(M_z s)}{Z}$ from Eq. 7 for all M_z for every class in PASCAL VOC 2012.

B C^{2}AM Implementation

After the classification network has been trained on a multi-label dataset, such as PASCAL VOC 2012, the Global CAM mentioned in Section 3.3 can be computed efficiently with multi-process parallel computing. It takes about three minutes to compute Global CAM in our experiment. After the Global CAM has been computed, it is stored as a constant without updating it. Thus, the computation cost is insignificant.

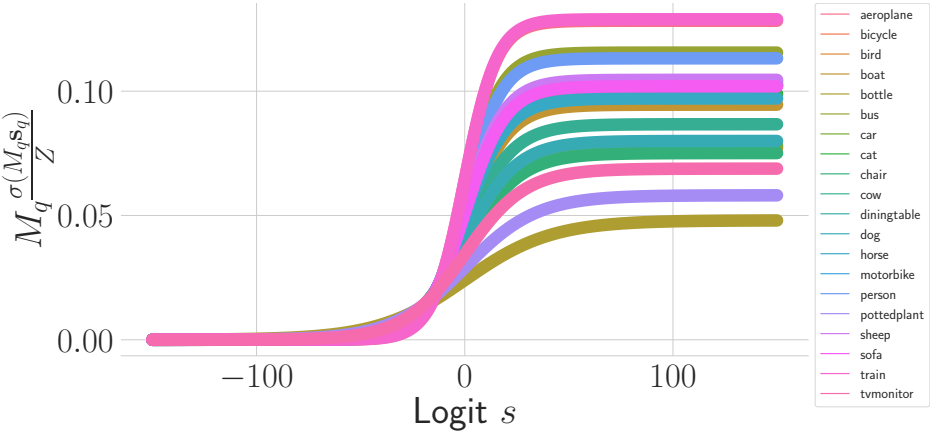


Figure 9: Visualizations for logit s ranges from -150 to 150 on the gradient $M_z \frac{\sigma(M_z s)}{Z}$ from Eq. 8 for all M_z for every class in PASCAL VOC 2012.

To apply the front-door adjustment mentioned in Section 3.3, one only needs to update the classification network training loop with one more line of code, as demonstrated in the following Python code snippet.

```
# 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)
```

C Causality

Structural Causal Model In causality [45], the concept of *structural causal model*, or SCM, is a directed acyclic graph to describe the causal relationships. SCM consists of a set of nodes representing the variables, and a set of edges between the nodes denote the causal directions. For instance, $X \to Y$ denotes X is the cause of Y. If a variable is the common cause of two variables, it is called a *confounder* [45]. In Figure 2(a), C is the cause of both X and Y, it is a confounder that will induce spurious correlations between X and Y to disturb the identification of the causal effect between them, due to the *back-door path* [45].

Back-Door Adjustment A back-door path [45], is defined as any path between X and Y that contains an arrow pointing into X, which can be identified by the Back-Door Criterion [45]. In Figure 2(a), $X \leftarrow C \to Y$ is a back-door path, and C satisfies the back-door criterion for X and Y. If a set of variables C satisifis the back-door criterion for X and Y, the causal effect

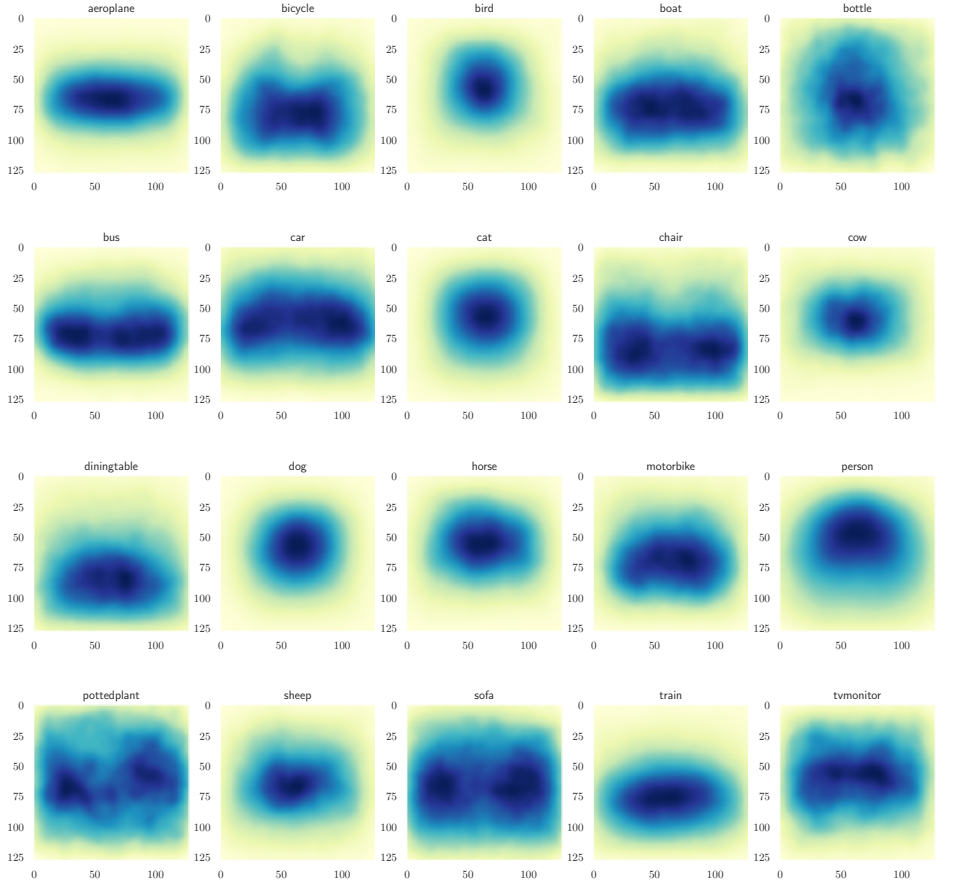


Figure 10: Visualizations of the Global CAM for each class z.

of X on Y can be computed by the back-door adjustment formula [45]

$$P(Y=y|do(X=x)) = \sum_{c} P(Y=y|X=x,C=c)P(C=c) \tag{9}$$

where $do(\cdot)$ denotes the interventional operation [45]. From Eq. 9, notice that applying back-door adjustment necessities prior knowledge of the confounders. However, the confounders are latent and unobservable. To apply the back-door adjustment formula, prior works assumed explicitly the nature of the confounders and attempted to estimate them, such as context, content and style [41, 59, 69]. Nevertheless, the *front-door adjustment* [45, 67] does not require any knowledge of the confounders and can also compute the causal effect between X and Y is a front-door SCM as illustrated in Figure 2(b).

Front-Door Adjustment The front-door adjustment calculates $P(Y|do(X))$ in the front-door path $X \rightarrow Z \rightarrow Y$ (see Figure 2(b)) by chaining together two partially causal effects $P(Z|do(X))$ and $P(Y|do(Z))$ by back-door adjustment, to obtain the overall effect of X on Y. Front-Door Criterion [45] can be applied to identify the variables that need to be adjusted.

The causal effect of X on Y is then identifiable by the front-door adjustment formula [45]

$$P(Y=y|do(X=x)) = \sum_z P(Z=z|X=x) \sum_{x'} P(Y=y|X=x',Z=z)P(X=x') \tag{10}$$

Eq. 10 is a fundamental causal inference technique for deconfounding the latent and unobserved confounder C [45]. We derive Eq. 10 by the following the *do*-calculus rules [45].

Given an arbitrary causal directed acyclic graph $\mathcal{G}$, assume that we have at least 4 node sets X, Y, Z, and W. Let $\mathcal{G}_{\overline{X}}$ denote the intervened causal graph where all incoming arrows to X are removed, and let $\mathcal{G}_{\underline{X}}$ denote the intervened causal graph where are outgoing arrows from X are removed. Let $\mathcal{G}_{\overline{Z(W)}}$ denote a graph which blocks all incoming arrows to nodes in Z that aren't ancestors of nodes in W. For any interventional distribution compatible with $\mathcal{G}$, we have the following rules:

Rule 1. Insertion/deletion of observations We can ignore an observation of a quantity if it does not influence the outcome by any path.

$$P(y|do(x),z,w) = P(y|do(x),w), \quad \text{if}(Y \perp Z|X,W)_{\mathcal{G}_{\overline{X}}} \tag{11}$$

Rule 2. Action/observation exchange Observations and interventions are equivalent when the causal effect of a variable on the outcome only influences the outcome by the directed paths.

$$P(y|do(x),do(z),w) = P(y|do(x),z,w), \quad \text{if}(Y \perp Z|X,W)_{\mathcal{G}_{\overline{X}\underline{Z}}} \tag{12}$$

Rule 3. Insertion/deletion of actions We can ignore an intervention if it does not influence the outcome by any path.

$$P(y|do(x),do(z),w) = P(y|do(x),w), \quad \text{if}(Y \perp Z|X,W)_{\mathcal{G}_{\overline{XZ(W)}}} \tag{13}$$

Thus, by applying the above rules, we have

$$\begin{aligned}
P(y|do(x)) &= \sum_z P(y|do(x),z)P(z|do(x)) \quad \text{(Total Probability)} \\
&= \sum_z P(y|do(x),do(z))P(z|x) \quad \text{(Apply Rule 2 Twice)} \\
&= \sum_{z,u} P(y|do(x),do(z),u)P(u|do(x),do(z))P(z|x) \quad \text{(Total Probability)} \\
&= \sum_{z,u} P(y|do(z),u)P(u|do(z))P(z|x) \quad \text{(Apply Rule 3)} \\
&= \sum_{z,u} P(y|do(z))P(z|x) \quad \text{(Marginalization)} \\
&= \sum_z P(z|x) \sum_{x'} P(y|do(z),x')P(x'|do(z)) \quad \text{(Total Probability)} \\
&= \sum_z P(z|x) \sum_{x'} P(y|z,x')P(x'|do(z)) \quad \text{(Apply Rule 2)} \\
&= \sum_z P(z|x) \sum_{x'} P(y|z,x')P(x') \quad \text{(Apply Rule 3)}
\end{aligned} \tag{14}$$

D Hyperparameter Settings

Our main implementation and hyperparameters follow the prior works [2, 10]. The hyperparameters for the continual training of the classifier are reported in Table 2. The hyperparameters for IRN [2] are reported in Table 3. The hyperparameters for DeepLabV2 [3] are reported in Table 4.

Hyperparameter	Value
Learning rate	0.0005
Number of epochs	4
Batch size	16
Image crop size	512
Weight decay	0.0001
Optimizer	SGD

Table 2: Hyperparameters for continual training of the pre-trained classification network on the PASCAL VOC 2012 dataset, with front-door adjustment described in Section 3.3.

Hyperparameter	Value
Learning rate	0.1
Number of epochs	3
Batch size	32
Image crop size	512
Weight decay	0.0001
Optimizer	SGD
Threshold for foreground	0.38
Threshold for background	0.10
Threshold for semantic segmentation background	0.21

Table 3: Hyperparameters for the training of IRN [2] on the produced CAMs by C^2AM to generate the pseudo-masks on the training set of PASCAL VOC 2012.

Hyperparameter	Value
Learning rate	$2.5e^{-4}$
Number of iterations	30000
Batch size	10
Image crop size	321
Weight decay	$5.0e-4$
Image scales	0.5, 0.75, 1, 1.25, 1.5, 1.75, 2

Table 4: Hyperparameters for the training of DeepLabV2 [3] on the pseudo-masks on the training set of PASCAL VOC 2012.

E PASCAL VOC 2012 Test Set Results

The predictions from DeepLabV2 [8] which trained on pseudo-masks by C^2AM and IRN [2] on the test set of PASCAL VOC 2012 are shown in Figure 11.

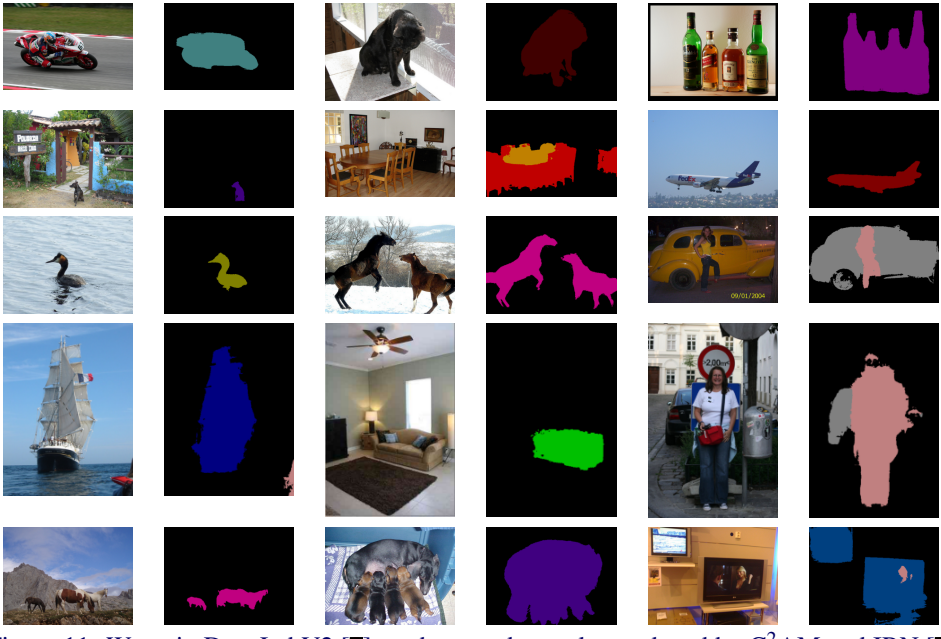
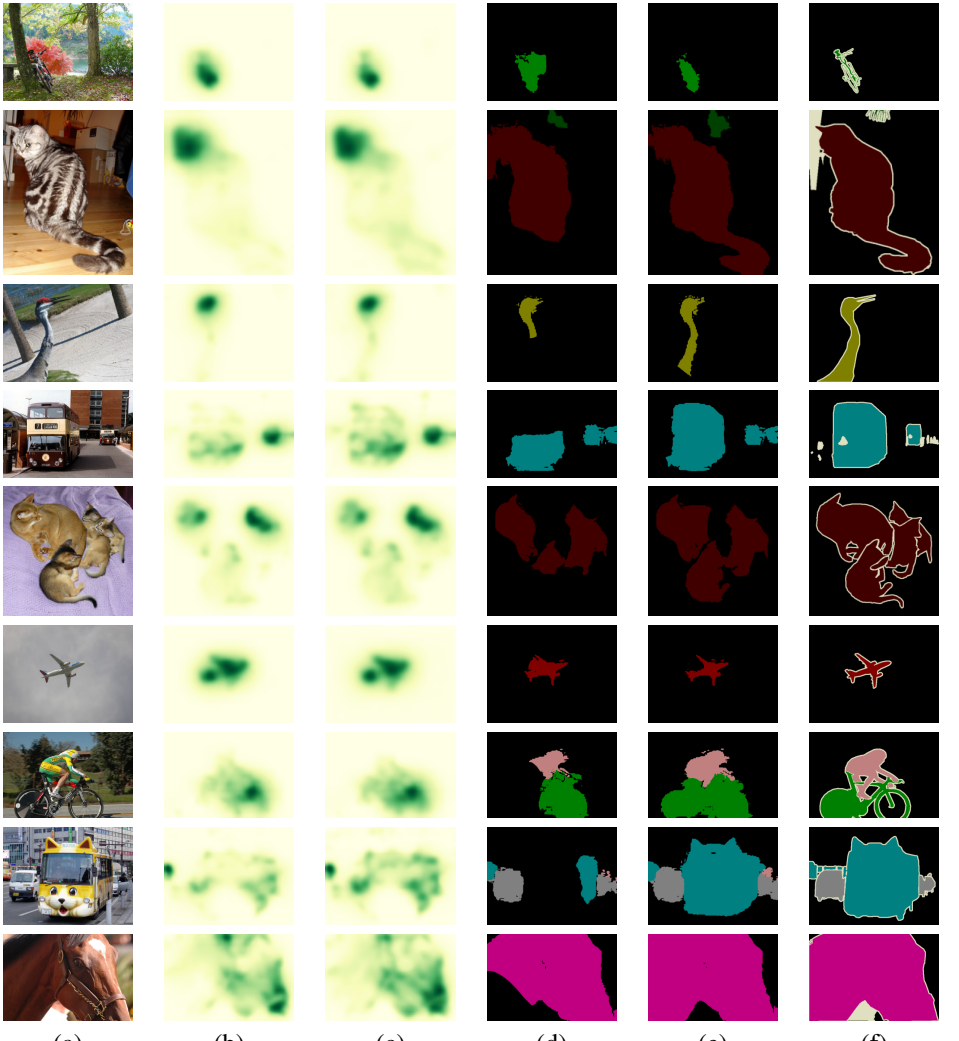


Figure 11: We train DeepLabV2 [8] on the pseudo-masks produced by C^2AM and IRN [2], and the above figures illustrate some sample results after applying the trained DeepLabV2 on the test set.

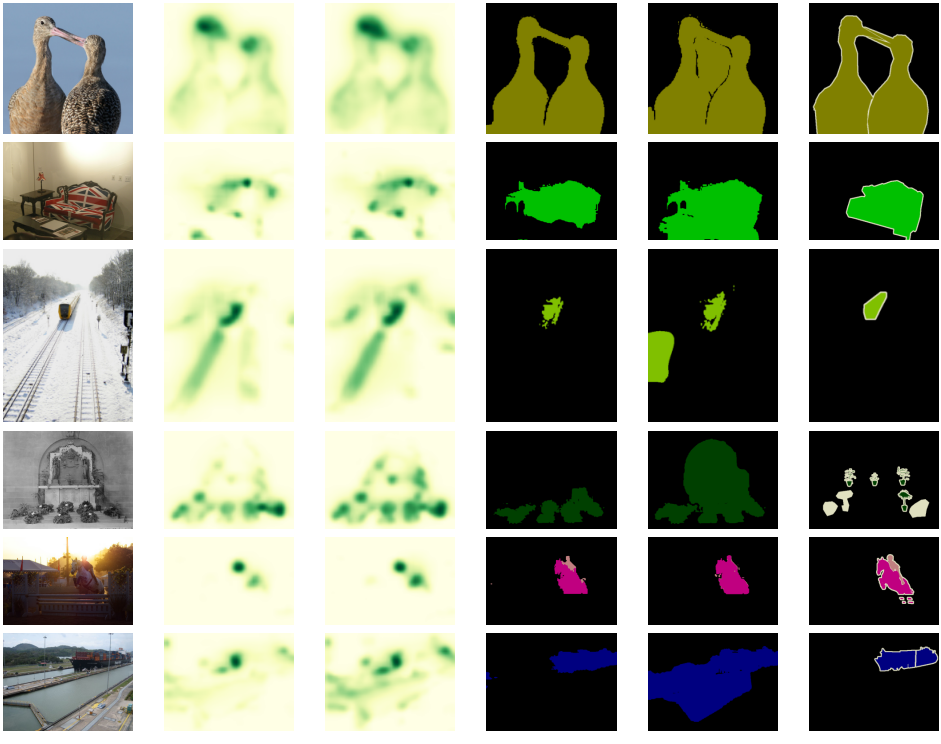
F Additional Qualitative Results

Additional qualitative evaluation of C^2AM and the pseudo-masks are shown in Figure 12. Also, more failure cases are shown in Figure 13.



(a) (b) (c) (d) (e) (f)

Figure 12: More qualitative evaluation of C^{2}AM and its pseudo-mask generated by IRN [2] on the PASCAL VOC 2012 training set. (a) original images. (b) CAM [33]. (c) C^{2}AM. (d) pseudo-masks generated by IRN with CAMs. (e) pseudo-masks generated by IRN with C^{2}AM. (f) pixel-level ground truth.



(a) (b) (c) (d) (e) (f)

Figure 13: More failure cases of C^{2}AM and its pseudo-mask generated by IRN [2] on the PASCAL VOC 2012 training set. (a) original images. (b) CAM [58]. (c) C^{2}AM. (d) pseudo-masks generated by IRN with CAMs. (e) pseudo-masks generated by IRN with C^{2}AM. (f) pixel-level ground truth.