We thank all the anonymous reviewers for their constructive feedback. We address each comment as follows.

R1-Q1:The calculation of context prior. Sorry for the confusion. In our work, the context prior is defined as a confounder set $C = \{c_1, c_2, ..., c_n\}$, where n is the class size in dataset. Each c is the $h \times w$ average segmentation mask of the i-th class images, which is obtained from the trained segmentation model in the last round (line 159-162).

R1-Q2:Just using the predicted mask to concat. Sorry for the unclear presentation. By (**Q1**) in Table 1 of the main paper, we directly concat the predicted mask (*i.e.*, Seg.Mask in Table 1) into the backbone network. Experimental results show that just using the predicted mask without C is even worse than the baseline SEAM [46].

R1-Q3:Refine the predicted mask with CRF. We followed your suggestion to use the CRF to refine the predicted mask, then concat the refined mask into the backbone network for a new round classification. Results on the baseline SEAM show that CRF (*vs* CONTA) is only effective in the first round, *i.e.*, achieved at most 0.2 (*vs* 1.1)%, 0.3 (*vs* 2.3)% and 0.3 (*vs* 1.8)% mIoU improvements for CAM, pseudo-mask and segmentation mask, respectively.

R2-Q1:Test in the transfer-learning scenario. We followed your suggestion to train models on COCO and test on PASCAL. Results on the *val* set show that the baseline SEAM (*vs* +CONTA) can achieve 32.1 (*vs* 33.2)% mIoU.

R2-Q2:The assumption of backdoor adjustment. Its identifiability assumes that the confounder set is fully observed, *e.g.*, a ground-truth vocabulary of contexts in our visual world. Unfortunately, it is impossible in practice and thus CONTA requires an iterative "guess" of the hidden confounder. Therefore, at each iteration, we need what you suggested: "one example of horse (person) without person (horse)", or more generally, "one example of class A without B", to disentangle A and B. Fortunately, it is feasible in the PASCAL and COCO datasets. We will follow your suggestion to revisit CONTA in the Rubin's potential outcome framework in revision.

R2-Q3:Unusual context. In fact, usual context such as "cow on grass" is better to illustrate, as the object co-occurrence is more confounding in WSSS (line 43-49). We also show an unusual example of "car crashed in water" in Figure R1. We will highlight this in revision.

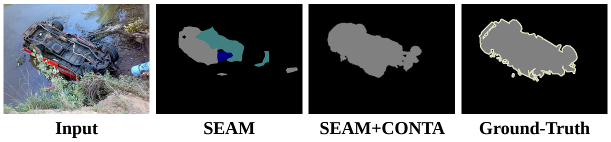

Figure R1: An unusual example: car crashed in water.

R2-Q4:The assumption and derivation in Appendix 2. We will include the assumption of NWGM approximation for self-contained purpose in revision, as in VC R-CNN [45]. For the derivation, we only derived $s_1(\cdot)$: the positive class term. Besides, we can also obtain derivation for the negative term $s_2(\cdot)$ through the similar process. We will clarify it in revision.

R2-Q5:No context can contribute to Y when training $P(Y|X)$? Sorry for the confusion. We mean that if the path $\boldsymbol{M} \rightarrow \boldsymbol{Y}$ is **NOT** existing, then no context can contribute to Y, and we can never recover the seed areas in WSSS by training $P(Y|X)$. But the reality is that we can recover the seed areas, which indicates that M is existing.

R2-Q6 and **R4-Q1:Typos/visualizations/vague statements/suggestions.** We are grateful for your constructive suggestions. We will revise our paper including typos, visualizations, and vague statements according to your suggestions.

R3-Q1:Only using C without X_m. We followed your suggestion to directly concat C corresponding to the image class into the backbone network. Experimental results show that only using C (*vs* the baseline SEAM) achieves 54.6 (*vs* 55.1)%, 62.1 (*vs* 63.1)% and 63.6 (*vs* 64.3)% mIoU on CAM, pseudo-mask and segmentation mask, respectively.

R3-Q2 and **R4-Q4:The description of Eq.3.** Sorry for the confusion. The segmentation mask $X_m \in \mathbb{R}^{hw \times n}$ denotes the logits. In our implementation of Eq.(3), X_m is first reshaped into a $hw \times 1$ vector. Therefore, the part contained in the softmax function has the shape of $1 \times n$. Following [45], $\sqrt{n}$ is used as a constant scaling factor for normalization. $\mathbf{W}_1$ and $\mathbf{W}_2$ are two learnable projection matrices. We will revise the writing of Eq.(3) in revision.

R3-Q3 and **R4-Q6:About P(c).** We are sorry for a typo here. $\sum_c$ and $P(c)$ can not be removed in Eq.(3), because each $P(c)$ corresponds to a specific entry in the confounder set $C(c)$. We will fix this typo in revision. Besides, the reason why we choose $P(c)$ to be uniform $1/n$ is that CONTA is designed to go beyond the dataset. If we use the actual class frequencies to represent $P(c)$, CONTA will be still confounded by the dataset observation.

R4-Q2:$\boldsymbol{X} \rightarrow \boldsymbol{M}$ or $\boldsymbol{M} \rightarrow \boldsymbol{X}$? In our assumption, M is the image-specific representation ($X \rightarrow M$), in the form of linear combination of context masks (Eq.3), which is certainly regularized by the context ($C \rightarrow M$). Therefore, what you think of "sampling object shapes and location" and "sampling object appearance" actually correspond to $\boldsymbol{C} \rightarrow \boldsymbol{M}$ and $\boldsymbol{C} \rightarrow \boldsymbol{X}$ in our model.

R4-Q3:Construct the confounder set. Sorry for the confusion. In our implementation, the input image is first resized into a fixed scale before feeding into the network. For example, we set 448×448 for SEAM+CONTA, and 512×512 for IRNet [1]+CONTA. Therefore, each of the entry in the confounder set follows the same scale as the input image.

R4-Q5:The projected embedding space can be any dimension? No, the projected embedding space can not be set to other dimensions, because n in $\mathbf{W}_1$ and $\mathbf{W}_2$ corresponds to the class size in dataset, which has the same size as the confounder set C. We have no reason to set n to other dimensions.

R4-Q7:Use pseudo-labels to re-train the segmentation model. We followed your suggestion to use pseudo-labels to re-train the segmentation model. Experimental results show that using pseudo-labels (*vs* the baseline SEAM) achieves 62.7 (*vs* 64.3)% mIoU. **R4-Q8:Stronger backbone can better handle context?** First, different backbones correspond to different seed generation or expansion methods. Therefore, we can not draw this conclusion. Second, this conjecture may be correct. Because the stronger backbone indeed locates object areas more accurately than a weaker one.