We thank all reviewers and AC for the time and constructive comments. Below we address the main concerns, and will fix other issues/typos and cite suggested references in the revised paper. Code will be released upon acceptance.

R2 / R3 / R4: Sensitivity analysis on γ: Evaluated on VOC07, CASD with $\gamma = 0.05, 0.075, 0.1, 0.15, 0.2$ gets mAP $54.1\%, 54.7\%, 55.3\%, 55.0\%, 55.0\%$ respectively, demonstrating CASD is robust to γ.

R1: Advantages of attention-based method: In Line 41-55 and Fig 1, we motivate our attention-based method by visualizing the features activated by WSOD networks. We observe that failures in WSOD are closely associated with the flawed features. Non-attention-based WSOD methods typically improve pseudo-labels and loss functions, instead of the features. In contrast, our method explicitly uses the structural info in attention maps, and promotes discriminative and consistent features on whole objects (Fig 1). Our method is orthogonal and applicable to non-attention-based WSOD.

Visualization of layer-wise attention maps: Fig. 3 (b) visualizes the attention maps of 2nd/3rd/4th conv blocks.

R2: Data splits: We follow the data splits in [7,8,10,11,13,34,42]: For VOC 2007 and 2012, we train on the train+val set and evaluate on the test set. For COCO, we train on the train set and evaluate on the val set.

Clarification on hyper-parameter selection: We agree that over-tuning hyper-parameters does not improve WSOD research. However, as long as hyper-parameters are database-independent, certain choices of networks, parameters and settings represent reasonable assumptions required to solve the ill-posed WSOD problem. Examples of the above hyper-parameters include those in selective search, object proposal, score aggregation, NMS, scaling and flipping. CASD inherits above hyper-parameters from PCL and OICR and fixes them on VOC07, VOC12 and COCO.

For the remaining CASD-specific hyper-parameters (Inverted Attention, choice of layers, max operation, loss weight γ), we follow the routine of [7,8,10,11,13,34,42], and only tune CASD parameters on VOC07 with the VOC07 test set, and then keep them fixed for VOC12 and COCO. The state-of-the-art results on VOC12 and COCO demonstrate that the VOC07 hyper-parameters of CASD transfer well and are practical for new databases.

We share the reviewer's concern that some WSOL methods in [A] lack clear validation. However, our work and [7,8,10,11,13,34,42] all select hyper-parameters on VOC07 and fix them for VOC12 and COCO. The VOC07 results, although over-tuned on the test set, are still valuable for comparing performance "upper bounds" among WSOD methods. The VOC12 and COCO results are NOT tuned on their test sets.

Performance bounded by the proposal method: We agree that using RPN in WSOD should beat Selective Search (SS). However, CASD is a training module independent of proposal methods. We adopt the same SS as most WSOD literatures only for fair comparisons. We will add both RPN-based CASD and Faster-RCNN results in the revised paper.

R3: IW-CASD without classification score aggregation: Without score aggregation and IA, IW-CASD achieves 51.4% mAP, clearly improving the baseline by 2.5% mAP. Adding score aggregation brings 1.2% mAP improvement.

Clarification on k and α: The vanilla OICR [10] has 3 OICR branches ($k = 3$), and suggests the larger k the better results. We only use $k = 2$ in all our experiments due to our GPU limitations. We reimplemented the OICR loss in Pytorch by ourselves. On VOC07, the OICR baseline at $\alpha = 0.1$ achieves 48.9% which is slightly superior to 48.3% using the adaptive weighting policy in vanilla OICR. Thus we keep the fixed weight for the OICR loss.

Clarification on method: (1). For L_{IW} and L_{LW}, the gradients are not back-propagated to the comprehensive attention maps A_r^{IW} and A_r^{LW}. **(2).** In Eq. (4), A^{IW} should be A_r^{IW}. **(3).** Following Fast-RCNN, we define $L_{reg}^k = \sum_{r=1,\ldots,G^k} smooth_{L1}(t_r, \hat{t}_r)/G^k$, G^k is the total number of positive proposals in k-th branch. t_r and $\hat{t}_r$ are tuples including location offsets and sizes of r-th predicted and gt bounding-box. **(4).** We will adopt the notation of $A_r = max(A_r^{IW}, A_r^{LW})$ in revised paper. Thanks for the suggestion.

R4: We sincerely thank the detailed feedback. We will fix the typos, inconsistency, add details and improve writing.
Different γs for L_{IW} and L_{LW}: We conduct the following ablation study on VOC07. Fixing $\gamma_{IW} = 0.1$, CASD gets $55.0\%, 55.5\%, 55.3\%, 54.8\%, 54.8\%$ when $\gamma_{LW} = 0.05, 0.075, 0.1, 0.15, 0.2$ respectively. Fixing $\gamma_{LW} = 0.1$, CASD gets $54.3\%, 54.6\%, 55.3\%, 55.1\%, 54.9\%$ when $\gamma_{IW} = 0.05, 0.075, 0.1, 0.15, 0.2$ respectively. At $L_{IW} = 0.1$ and $L_{LW} = 0.075$, CASD achieves 55.5% which is only 0.2% better than 55.3% (the best performance of single γ). Thus we conclude that single γ is a good trade-off between performance and hyper-parameter tuning.

Evidence for consistency and completeness: First, the attention maps in Fig. 1 b) compare the results of baseline OICR and CASD, qualitatively demonstrating CASD gets more consistent and complete object features. More examples will be added in the revised paper. Second, on the horizontal flipped VOC07 test set, CASD achieves 56.5% mAP which is similar to 56.8% of the unflipped test set. This indicates that CASD is consistent w.r.t flipping.

Result with Grad-CAM: The layer-wise CASD with Grad-CAM gets 52.0% mAP on VOC07. This is slightly worse than 52.6% mAP of layer-wise CASD with channel-wise average pooling+sigmoid. The latter is computationally more efficient and adopted in our paper.