Table 1

Model	MSRC			iCoseg			CoSal2015		
	$F_\beta\uparrow$	$S_\alpha\uparrow$	MAE↓	$F_\beta\uparrow$	$S_\alpha\uparrow$	MAE↓	$F_\beta\uparrow$	$S_\alpha\uparrow$	MAE↓
$n_{train}=5$	0.868	0.812	0.098	0.866	0.854	0.050	0.858	0.854	0.058
$n_{train}=10$	0.869	0.814	0.097	0.874	0.862	0.048	0.860	0.855	0.058
$n_{train}=15$	0.876	0.813	0.096	0.863	0.853	0.052	0.857	0.850	0.060
$n_{test}=5$	0.870	0.815	0.097	0.865	0.856	0.050	0.844	0.846	0.066
$n_{test}=10$	0.869	0.814	0.097	0.874	0.862	0.048	0.851	0.849	0.063
$n_{test}=20$	0.869	0.815	0.097	0.873	0.862	0.048	0.856	0.852	0.061
ICNet with GC	0.867	0.808	0.098	0.863	0.858	0.049	0.845	0.846	0.065

Table 2

Model	NFs	SCFs	SIVs	CFM	R	MSRC			iCoseg		
						$F_\beta\uparrow$	$S_\alpha\uparrow$	MAE↓	$F_\beta\uparrow$	$S_\alpha\uparrow$	MAE↓
1	✓					0.698	0.673	0.164	0.643	0.678	0.131
2	✓		✓			0.762	0.709	0.154	0.779	0.773	0.084
3	✓		✓	✓		0.840	0.778	0.112	0.845	0.846	0.061
4		✓				0.837	0.785	0.115	0.771	0.786	0.079
5		✓	✓			0.844	0.793	0.113	0.830	0.831	0.065
6		✓	✓	✓		0.855	0.796	0.108	0.856	0.844	0.056
7		✓	✓	✓	✓	0.869	0.814	0.097	0.874	0.862	0.048

Novelty (R1/R3/R4). The key contribution of our work is that we exploited SISMs in a better manner, and obtained discriminative intra-saliency cues for better Co-SOD performance. To this end, we directly integrated SISMs (produced by previous SOD method) as intra-saliency priors (instead of training targets) into the deep neural network to extract intra cues, and exploited the correlation information in the intra cues to obtain inter cues for detecting co-saliency. What's more, we proposed the RSCF module to further improve our ICNet on Co-SOD. Experiments demonstrated that our ICNet well leverages SISMs and becomes a new SOTA method for Co-SOD. Ablation studies also validated the effectiveness of RSCF in **Lines 269-278** of the main paper. We do not take the NMAP and CFM modules as our contributions. We have claimed our contributions more clearly in the revised submission, and thanks for the suggestions.

Performance on *MSRC* and *iCoseg* (R1/R4). In *MSRC* and *iCoseg*, 38 out of 45 image groups contain only one category of salient objects. In *CoSal2015*, 43 out of 50 image groups contain more than one category of salient objects. Thus, the evaluations on *CoSal2015* could reflect better the capability of the comparison methods on Co-SOD than those on *MSRC* and *iCoseg*. Comparing to previous methods, our ICNet obtains large improvements on *CoSal2015*, but small gains on *MSRC* and *iCoseg*. This demonstrates that our ICNet is very capable of tackling the Co-SOD problem. F_β is based on the binary saliency map and S_α measures structural similarity, they are not sensitive to the pixel-level incorrectness as MAE. Hence, our ICNet outperforms marginally over the other methods on F_β and S_α.

How the size of the image group influences our ICNet (R1/R3)? In Table 1, we summarize the results of our ICNet with a group size of "n_{train}" or "n_{test}" in the training or test stage, respectively. When n_{train} is set as 5/10/15, in the test stage we use all images in a group. When n_{test} is set as 5/10/20, in the training stage we set $n_{train}=10$. The results show that the training group size influences little our ICNet and the inconsistent group sizes in training and test phases are empirically available. Our ICNet leverages SISMs better when taking the whole group as input during tests, as increasing the test group size brings slight improvements on CoSal2015.

Related works (R1). We have introduced the suggested related works in the revision of our submission.

Ablation studies on *MSRC* and *iCoseg* (R2). Table 2 shows that our model is also effective on *MSRC* and *iCoseg*.

Models without CFM (R2). We copy the group-level vector in $\mathbb{R}^{C'}$ for $H\times W$ times to form a tensor in $\mathbb{R}^{C'\times H\times W}$ with the same spatial size as the single-image feature in $\mathbb{R}^{C\times H\times W}$, and concatenate them along the channel dimension.

Background of SISMs (R2). Inter consistency in common backgrounds is also exploited by our ICNet to well leverage SISMs. Otherwise, the results of our ICNet on F_β/S_α/MAE drop from $0.874/0.862/0.048$ to $0.864/0.855/0.051$ on *iCoseg*, from $0.860/0.855/0.058$ to $0.849/0.847/0.064$ on *CoSal2015*, respectively.

Why removing "Tree" from *MSRC* (R2)? The SISMs of the "Tree" category (in *MSRC*) produced by previous SOD methods do not contain any saliency information. This will make our ICNet fail since it is heavily based on SISMs. To show the power of our ICNet with reasonable SISMs on Co-SOD, we remove this "Tree" category. Thanks to the reviewers for this comment, which reminds us to take the failure of our ICNet on the "Tree" category as our major limitation. Hence, we have discussed this problem as a new section of "Failure Case" in the revision of our submission.

Using SISMs provided by old SOD methods (R3). Thanks for this suggestion. We trained our ICNet with the SISMs produced by Global Contrast (GC) (Global Contrast based Salient Region Detection. In *CVPR*, 2011). As shown in Table 1, our "ICNet with GC" still outperforms other methods, validating the effectiveness of our ICNet.

Why RSCF is useful (R3)? We observed via experiments that our ICNet with combined CSA map A_k and normalized feature map F_k fails to distinguish pixels with similar but different categories. This is mainly due to the inconsistency on the category dependence between A_k and F_k: A_k is category-independent and just reflects the initial co-saliency scores, while F_k is category-related and each pixel in it is a vector representing a specific category. Specifically, in our ICNet the predictions mainly depend on category-independent A_k, indicating that the category information in F_k is neglected. To tackle this inconsistency, we calculated the similarity between pairs of pixels in F_k by utilizing the category information in it, and transferred F_k into category-independent SCF. We then rearranged the channel order of SCF to obtain the RSCF, according to the co-saliency scores in A_k. These operations enable our ICNet well exploit meaningful information from the rearranged self-correlation maps in SCF. In **Lines 269-278** of the main paper, the ablation studies have validated the effectiveness of our RSCF. We have added this point to our revised submission.

Why integrating SISMs performs better (R4)? Taking SISMs as training targets forces Co-SOD models overfit to inaccurate SISMs with performance drop, though pooling/normalization layers are used to dilute the inaccuracy. By directly integrating SISMs into the encoder of the model, our ICNet avoids the problem of overfitting to inaccurate SISMs, while the inherent inaccuracy in SISMs is largely alleviated by NMAP. The effectiveness of directly integrating SISMs into our ICNet has been validated in the ablation study (**Lines 262-265**) of the main paper.