Supplementary: Unsupervised and semi-supervised co-salient object detection via segmentation frequency statistics

Souradeep Chakraborty[1,2†], Shujon Naha[2,3], Muhammet Bastan[2], Amit Kumar K C[2], Dimitris Samaras[1]
[1]Stony Brook University [2]Visual Search & AR, Amazon [3]Indiana University
{souchakrabor,samaras}@cs.stonybrook.edu, {mbastan, amitkrkc}@amazon.com, snaha@iu.edu

1. Additional qualitative results

1.1. Comparison of CoSOD predictions

In Fig. 1 we present additional results of co-salient object detection using the proposed models and the other baselines.

In the first image group in Fig. 1 we show the CoSOD predictions on the *eggplant* category from the CoCA dataset. While our US-CoSOD model detects the salient objects well, it fails to accurately segment the eggplant. Similarly, both the DCFM and our SS-CoSOD model trained with 1/4 labels fail to accurately detect the eggplant instances. Our SS-CoSOD model when trained with all labels predicts CoSOD segmentations most closely resembling the ground truth.

In the *zebra* image group (selected from the CoSOD3k dataset), we observe that the segmentation maps obtained from our SS-CoSOD model trained with all labels most closely resemble the ground truth. The DCFM model trained with all labels suffers from overestimating the co-saliency of certain image regions e.g. in columns 1 and 2 and produces incomplete segmentations in columns 3 and 4. While the DCFM model trained with all labels segments the zebra in column 1, the segmentation prediction fails to preserve the shape. In column 2, all models except our SS-CoSOD model trained with all labels detect the giraffes as being co-salient along with the zebras.

In the third image group, we compare the segmentation results on the *penguin* group from the Cosal2015 dataset. Overall, our SS-CoSOD model trained with all labels produces more accurate co-salient object segmentations compared to the other baselines. In column 1, our SS-CoSOD models trained with 1/4 labels and with all labels well segment the penguin. In the last column of this group, we show an instance where all the models including our SS-CoSOD fail to distinguish the penguin from the seal. This could be due to the fact that the seal has similar visual features as the penguin, which makes it difficult for the models to distinguish between the two categories. Training on more fine-grained categories might help our model resolve this ambiguity.

1.2. Comparison of unsupervised CoSOD predictions

In Fig. 2 we present additional results comparing the self-attention (SA) maps from DINO (DI), the pseudo co-salient ground truth masks - our DINO+STEGO model (DI+ST), and predictions from our US-CoSOD model.

In column 2 of row block 1 (the *teddy bear* image group), we observe that the most frequent unsupervised semantic clusters representing the teddy bear are colored light green and pink. Our US-CoSOD model effectively eliminates the inaccurate segmentation of the child (that carries the teddy bear) produced by the DINO SA and the DI+ST models. In rows 2 and 3 of this group, US-CoSOD rectifies the inaccurate segmentation masks obtained from the DI+ST model. In row 4, the teddy bear segmentation from both the DI+ST and US-CoSOD models is quite accurate.

In row block 2 (the *hourglass* image group), we observe that blue and dark blue colored unsupervised semantic clusters mainly constitute the hourglass object. In row 2 of this group, although the SA map from DINO highlights both the person and the hourglass to be salient, the segmentation predictions from the DI+ST and US-CoSOD models correctly show only the hourglass to be co-salient, which is due to the fact that the co-occurrence frequency of the unsupervised semantic cluster denoting the hourglass object is sufficiently high compared to that for the person. Our US-CoSOD model further improves the segmentations predicted by the DI+ST model.

In Fig. 3 we present qualitative results comparing our method with the different unsupervised methods for single-image segmentation and co-segmentation tasks. We observe that our US-CoSOD model has better segmentation predictions compared to the SegSwap [6], DVFDVD [1], and the TokenCut [8] models for two image groups - *hour glass* and *teddy bear* from the CoCA dataset.

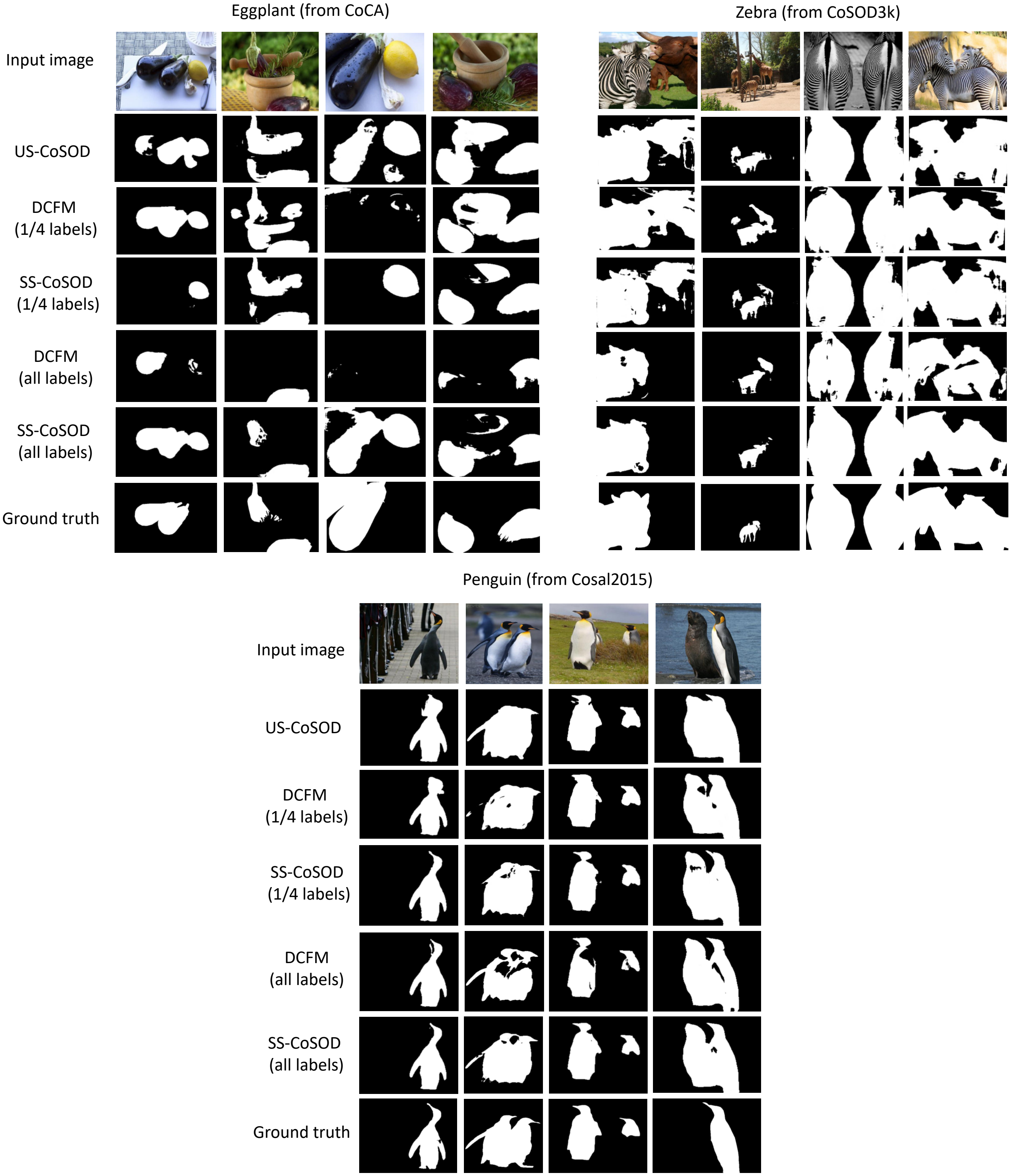


Figure 1. Additional qualitative comparisons of our model with different baselines on three image groups selected each from CoCA, CoSOD3k and Cosal2015. Our SS-CoSOD model trained with all labels produces the most accurate segmentation mask compared to the other baselines.

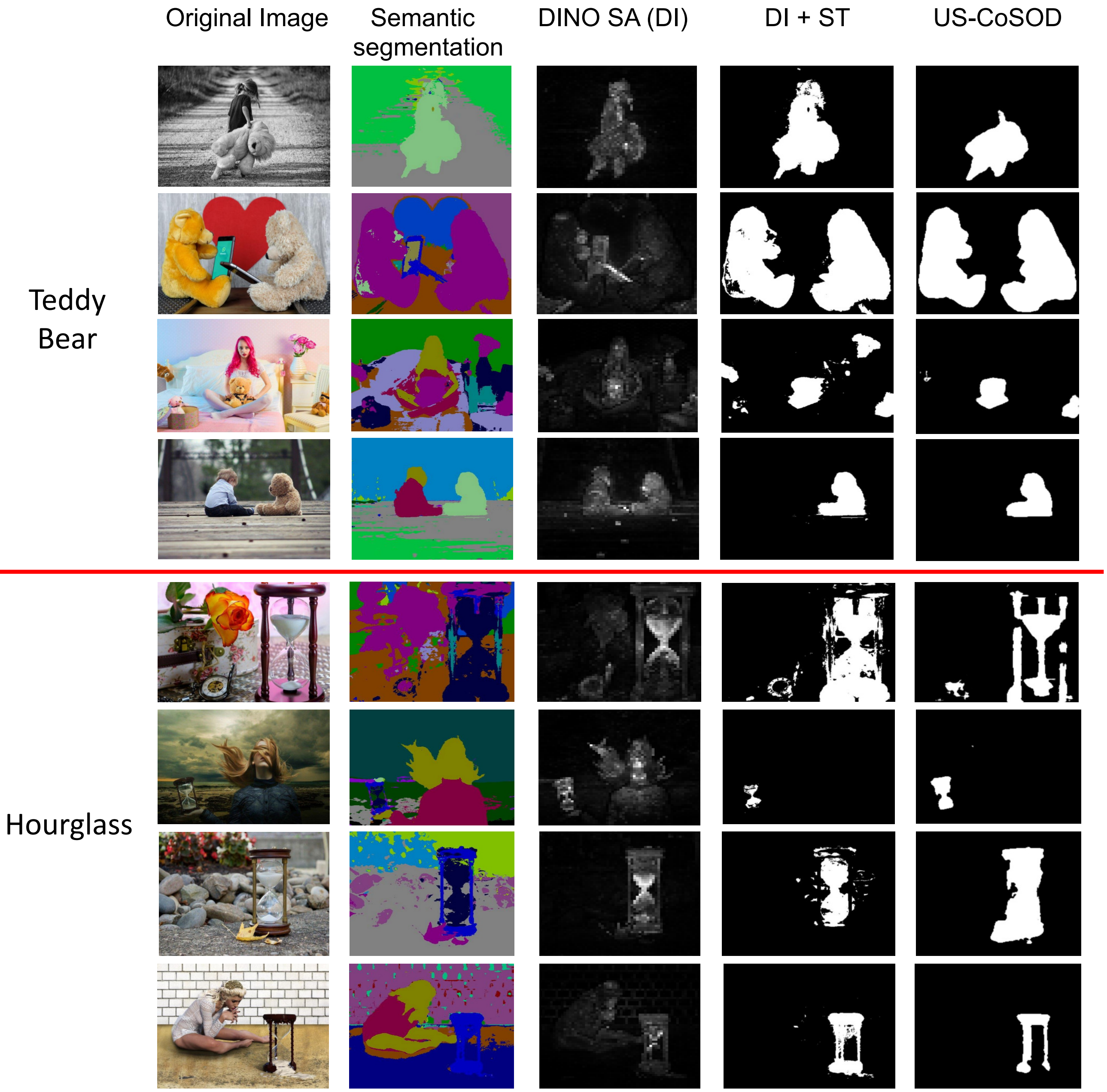
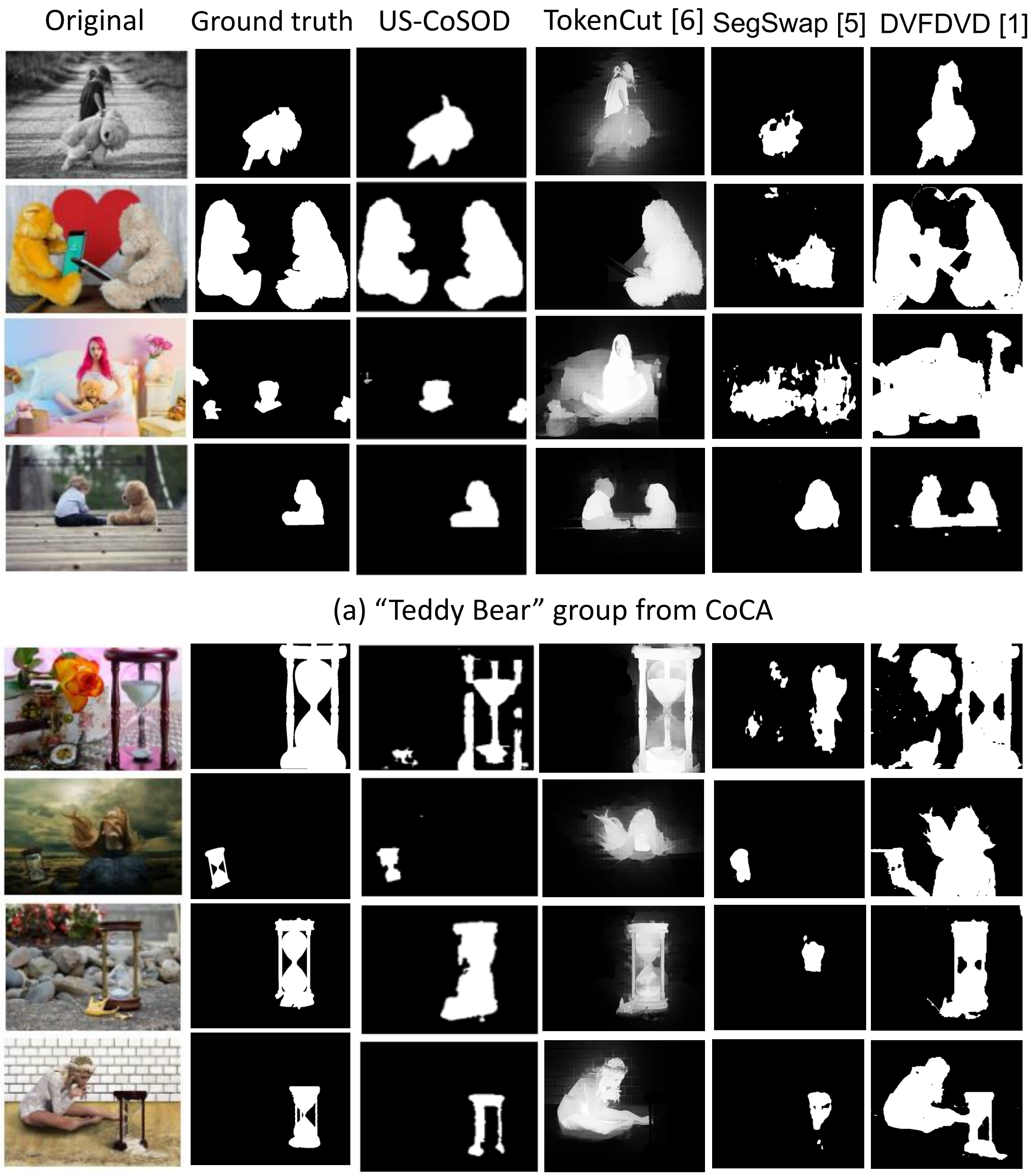


Figure 2. Additional qualitative comparisons of the DINO self-attention maps (DI), pseudo ground truth co-saliency masks from our DI+ST model, and predictions from our US-CoSOD model.

1.3. Confidence Estimation Network predictions

In Fig. 4 we show the ground truth and the predicted confidence scores from our Confidence Estimation Network (CEN) module using 1/2 and 1/8 labels for training. *GT* denotes the max. F1-score of the predictions obtained from the pretrained f_{PT} model (see Fig. 3 in the main paper) on the unlabeled data and *Pred* denotes the F1-score predicted by our trained CEN module. We observe that the confidence scores vary in proportion to the image complexity in terms of the image contents. In particular, we observe that the ground truth confidence score is high when the co-salient object is more salient and has a clear demarcation with respect to the scene background, while the ground truth confidence score is low when the image scene is more cluttered (e.g. for the *train* class in row 2 and the *banana* class in row 4) or the co-salient objects are out-of-distribution (e.g. for the *boat* class in row 3, the boat is on the land). Also, we observe that our CEN module is able to predict the ground truth confidence score well. Therefore, the CEN model ef-



(a) “Teddy Bear” group from CoCA

(b) “Hour glass” group from CoCA

Figure 3. Qualitative comparisons of prediction results from our unsupervised CoSOD model, US-CoSOD vs. corresponding segmentations from existing unsupervised segmentation models. TokenCut [8] is a single-image segmentation method and SegSwap [6] and DVFDVD [1] are unsupervised co-segmentation methods.

fectively suppresses the error propagation during training caused due to inaccurate confidence estimation on images that are difficult for the CoSOD task.

2. Additional quantitative results

In Tab. 1, we provide additional results of the performance evaluation of our US-CoSOD model compared to Tab. 1 in the main paper. In particular, we additionally show the prediction performance of US-CoSOD when trained on a set of 50K images (with 50 images per class) along with the baselines presented in Tab. 1 in the main paper. US-CoSOD when trained on 150 images per class produces the best performance. Training US-CoSOD using 50 images per class (for each of the 1000 ImageNet classes) leads to inferior performance due to limited training data. On the other hand, training the model using 450 images per class reduces

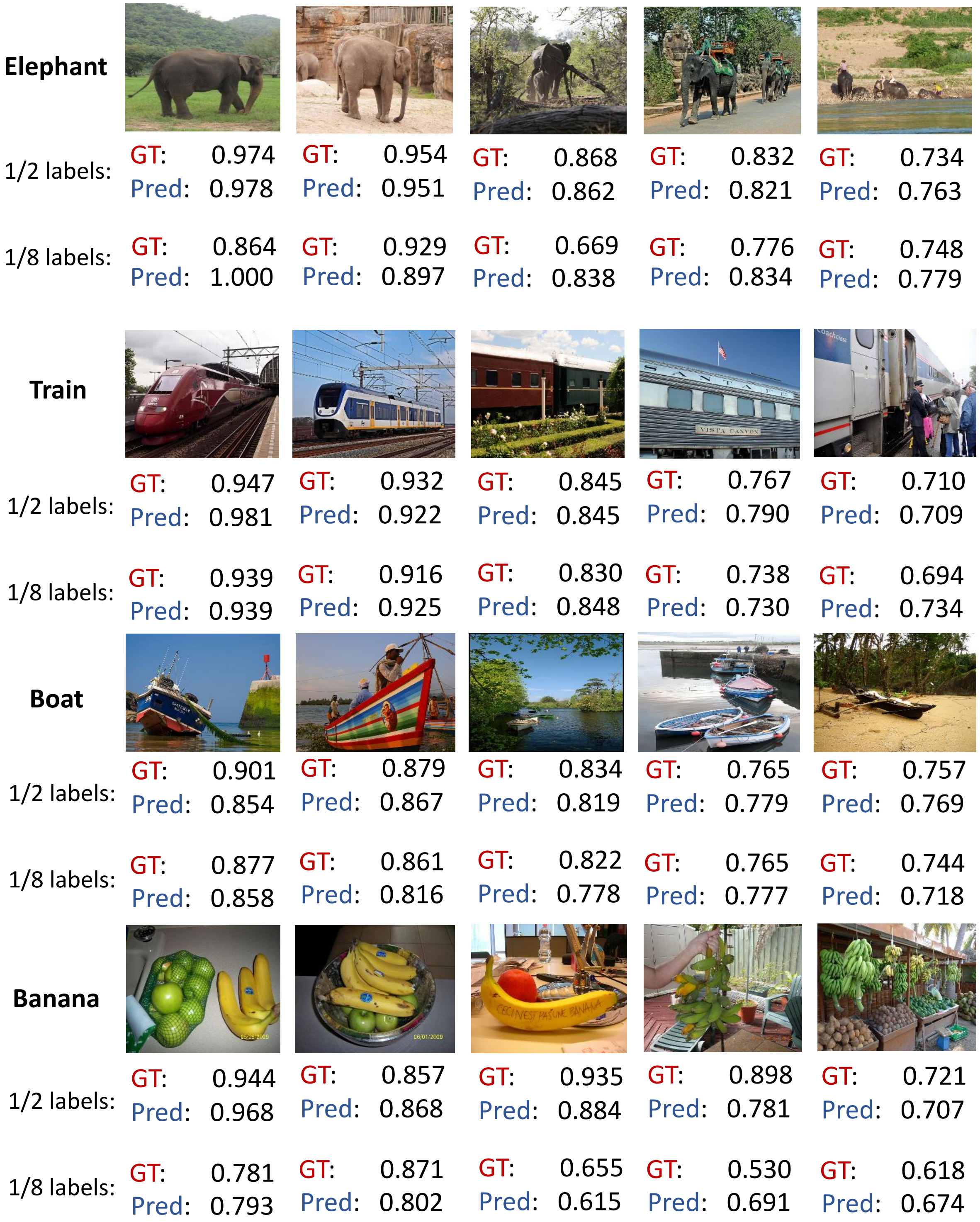


Figure 4. Depiction of the ground truth and the predicted confidence scores from our Confidence Estimation Network (CEN) module using 1/2 and 1/8 labels for training. *GT* denotes the max. F1-score of the predictions obtained from the pretrained f_{PT} model (see Fig. 3 in the main paper) on the unlabeled data and *Pred* denotes the F1-score predicted by our trained CEN module. We observe that the confidence scores vary in proportion to the image complexity in terms of the image contents. Also, we observe that our CEN module is able to predict the ground truth confidence score well.

Table 1. Performance evaluation of our US-CoSOD model: we show the prediction performance of US-CoSOD when trained on a set of 50K images (with 50 images per class) along with the other baselines presented in Table 1 in the main paper. US-CoSOD when trained on 150 images per class produces the best performance.

	CoCA				Cosal2015				CoSOD3k			
Method	MAE↓	F_β^{max} ↑	E_ϕ^{max} ↑	S_α ↑	MAE↓	F_β^{max} ↑	E_ϕ^{max} ↑	S_α ↑	MAE↓	F_β^{max} ↑	E_ϕ^{max} ↑	S_α ↑
DINO (DI)	0.214	0.372	0.572	0.540	0.154	0.659	0.753	0.688	0.146	0.624	0.749	0.679
STEGO (ST)	0.235	0.353	0.555	0.523	0.164	0.618	0.717	0.676	0.204	0.543	0.660	0.615
TokenCut [8] (CVPR 2022)	0.167	0.467	0.704	0.627	0.139	0.805	0.857	0.793	0.151	0.720	0.811	0.744
DVFDVD [1] (ECCVW 2022)	0.223	0.422	0.592	0.581	0.092	0.777	0.842	0.809	0.104	0.722	0.819	0.773
SegSwap [6] (CVPRW 2022)	0.165	0.422	0.666	0.567	0.178	0.618	0.720	0.632	0.177	0.560	0.705	0.608
Ours (DI+ST)	0.165	0.461	0.676	0.610	0.112	0.760	0.823	0.767	0.124	0.684	0.793	0.724
Ours (US-CoSOD-COCO9213)	0.140	0.498	0.702	0.641	0.090	0.792	0.852	0.806	0.095	0.735	0.832	0.772
Ours (US-CoSOD-ImgNet50)	0.141	0.516	0.703	0.648	0.076	0.823	0.876	0.827	0.092	0.752	0.841	0.783
Ours (US-CoSOD-ImgNet150)	**0.116**	**0.546**	**0.743**	**0.672**	**0.070**	**0.845**	**0.886**	**0.840**	**0.076**	**0.779**	**0.861**	**0.801**
Ours (US-CoSOD-ImgNet450)	0.127	0.543	0.726	0.666	0.071	0.844	0.884	0.842	0.079	0.775	0.854	0.800

Table 2. Performance comparison of the different versions of our unsupervised and semi-supervised models. In column 1, we indicate the fraction of labeled data for training, followed by the actual number of images.

		CoCA				Cosal2015				CoSOD3k			
Split	Method	MAE↓	F_β^{max} ↑	E_ϕ^{max} ↑	S_α ↑	MAE↓	F_β^{max} ↑	E_ϕ^{max} ↑	S_α ↑	MAE↓	F_β^{max} ↑	E_ϕ^{max} ↑	S_α ↑
1/16 (576)	DCFM [10] (CVPR 22)	0.119	0.485	0.725	0.636	0.088	0.780	0.847	0.786	0.088	0.716	0.827	0.753
	US-CoSOD+DCFM	0.108	0.557	0.754	0.683	0.068	0.854	0.888	0.846	0.076	0.783	0.857	0.801
	SS-CoSOD-DJ (w/o CEN)	0.107	0.485	0.728	0.635	0.094	0.771	0.834	0.771	0.089	0.709	0.817	0.742
	SS-CoSOD-DJ (w/ CEN)	0.115	0.488	0.730	0.639	0.086	0.782	0.847	0.787	0.086	0.717	0.828	0.755
	SS-CoSOD	0.113	0.492	0.733	0.641	0.085	0.788	0.850	0.792	0.084	0.721	0.830	0.758
	US-CoSOD+SS-CoSOD	0.111	0.554	0.751	0.681	**0.066**	**0.855**	**0.890**	**0.849**	0.075	0.783	0.858	**0.803**
	SS-CoSOD with ImgNet	**0.098**	**0.562**	**0.757**	**0.684**	0.072	0.837	0.880	0.828	**0.068**	**0.784**	**0.865**	0.800
1/8 (1152)	DCFM [10] (CVPR 22)	0.110	0.493	0.731	0.639	0.096	0.780	0.839	0.779	0.096	0.727	0.818	0.746
	US-CoSOD+DCFM	0.111	0.558	0.754	0.683	**0.067**	0.857	**0.890**	**0.847**	0.076	0.785	0.857	0.801
	SS-CoSOD-DJ (w/o CEN)	0.103	0.497	0.732	0.641	0.096	0.777	0.835	0.777	0.094	0.727	0.816	0.744
	SS-CoSOD-DJ (w/ CEN)	0.116	0.499	0.735	0.644	0.085	0.793	0.854	0.800	0.087	0.740	0.834	0.767
	SS-CoSOD	0.114	0.500	0.736	0.645	0.084	0.795	0.856	0.802	0.086	0.740	0.835	0.767
	US-CoSOD+SS-CoSOD	0.108	0.558	0.755	0.683	0.068	**0.857**	0.888	0.845	0.076	0.785	0.856	0.799
	SS-CoSOD with ImgNet	**0.097**	**0.560**	**0.755**	**0.685**	0.068	0.845	0.884	0.838	**0.068**	**0.791**	**0.871**	**0.808**
1/4 (2303)	DCFM [10] (CVPR 22)	0.107	0.547	0.758	0.672	0.073	0.829	0.880	0.824	0.075	0.775	0.862	0.794
	US-CoSOD+DCFM	0.109	0.569	0.758	0.685	0.069	0.855	0.888	0.844	0.077	0.783	0.854	0.797
	SS-CoSOD-DJ (w/o CEN)	0.097	0.552	0.763	0.678	0.076	0.828	0.874	0.818	0.075	0.776	0.859	0.790
	SS-CoSOD-DJ (w/ CEN)	0.096	0.560	0.764	0.685	0.069	0.839	0.885	0.831	0.069	0.784	0.867	0.802
	SS-CoSOD	0.097	0.562	0.765	0.686	0.068	0.841	0.886	0.833	0.068	0.785	0.868	0.803
	US-CoSOD+SS-CoSOD	0.107	0.566	0.757	0.686	0.066	**0.858**	0.891	**0.848**	0.073	0.787	0.859	0.803
	SS-CoSOD with ImgNet	**0.091**	**0.581**	**0.772**	**0.698**	**0.066**	0.851	**0.891**	0.841	**0.064**	**0.799**	**0.875**	**0.812**
1/2 (4607)	DCFM [10] (CVPR 22)	0.101	0.566	0.764	0.690	0.065	0.845	0.889	0.838	0.070	0.792	0.870	0.807
	US-CoSOD+DCFM	0.105	0.569	0.760	0.688	0.068	0.856	0.889	0.843	0.074	0.793	0.862	0.804
	SS-CoSOD-DJ (w/o CEN)	0.092	0.572	0.771	0.694	0.068	0.846	0.885	0.834	0.071	0.791	0.865	0.802
	SS-CoSOD-DJ (w/ CEN)	0.090	0.578	0.772	0.699	0.062	0.851	0.892	0.843	0.067	0.795	0.870	0.810
	SS-CoSOD	0.088	0.582	0.773	0.700	0.062	0.854	0.892	0.843	0.066	0.797	0.872	0.809
	US-CoSOD+SS-CoSOD	0.110	0.563	0.755	0.686	0.064	0.858	0.894	0.850	0.072	0.794	0.866	0.810
	SS-CoSOD with ImgNet	**0.088**	**0.590**	**0.775**	**0.705**	**0.062**	**0.861**	**0.896**	**0.850**	**0.063**	**0.804**	**0.876**	**0.817**
Full (9213)	DCFM [10] (CVPR 22)	**0.085**	**0.598**	**0.783**	**0.710**	0.067	0.856	0.892	0.838	0.067	0.805	0.874	0.810
	US-CoSOD+DCFM	0.102	0.573	0.764	0.692	0.068	0.860	0.890	0.845	0.077	0.791	0.856	0.799
	SS-CoSOD with ImgNet	0.091	0.591	0.778	0.707	**0.061**	**0.865**	**0.901**	**0.852**	**0.062**	**0.809**	**0.882**	**0.821**

segmentation accuracy. This could be because adding more difficult unlabeled images to the training set may lead to erroneous training due to the inaccurate pseudo ground truth masks generated by the DI+ST model, using which US-CoSOD is trained.

Quantitative evaluation of SS-CoSOD In Tab. 2, we show a more detailed version of Tab. 2 in the main paper. Here, we additionally show the prediction results with 1/8 labeled data.

Comparison with SOTA In Tab. 3, we compare the performance of our model with other state-of-the-art models on the 3 benchmark datasets. We outperform the state-of-the-art DCFM model [10] on the Cosal2015 and the CoSOD3k datasets, while we are comparable with this model on the CoCA dataset (DCFM predictions on CoCA being slightly more accurate). Also, we outperform other existing fully supervised CoSOD models by a significant margin.

Variant model In Tab. 4, we compare the performance of the proposed US-CoSOD model with a variant version of the model where we divide the overlap area, O_i^j between the DINO mask DM_i and the STEGO segmentation mask SM_i^j (for class c^j) by the area occupied by the STEGO mask $Ar(SM_i^j)$. The proposed version of the US-CoSOD model performs better than the variant version over all three test datasets using all four evaluation metrics.

Performance on challenging categories In Tab. 5, we report the average F-measure score on the categories over which DINO (a pre-trained component) scored lesser than the average DINO F-measure score over the test dataset. Specifically, for a given test dataset, categories that had an F-measure score lower than the threshold value, $F_{th}^{\beta} = \frac{1}{n}\sum_{i=1}^{n} F^{\beta}(SA_i)$ (here SA_i denotes the DINO self-attention map of image I_i and n = total number of test images) were considered for this experiment. As observed, our US-CoSOD outperforms the pre-trained DINO and DINO+STEGO models by a significant margin on such difficult categories.

CEN backbone In Tab. 6, we compared the confidence estimation error (Mean Squared Error) of different backbone networks for our CEN module on the unlabeled dataset. As we observe, ResNet50 trained using DINO provides us the least Mean Squared Error loss across all data splits. We attribute the lower accuracy of MobileNetV2 to its lower feature representation power, and that of the ViTB and ViTS models to the fact that such transformer models fail to outperform convolutional models (*e.g.* ResNet50) when less data is available for training (in the different label splits).

3. Additional implementation details

We randomly split the data in the COCO9213 dataset into the labeled and the unlabeled sets (i.e. 1/16, 1/8, 1/4, 1/2 labels) for training the fully supervised DCFM and our semi-supervised SS-CoSOD models.

The inputs are resized to 224 × 224 for both training and inference. We use Adam [4] as our optimizer to train our models. The total training time is around 5 hours for US-CoSOD and around 8 hours for SS-CoSOD using ImageNet-1K. All experiments are run on a single NVIDIA Tesla V100 SXM2 GPU.

For the unsupervised model (US-CoSOD) and the supervised pre-training on labeled data in stage 1 in Fig. 3 in the main paper, we set the learning rate is set as 10^{-5} for feature extractor and 10^{-4} for other parts, and the weight decay is set as 10^{-4}, following [10]. Training these models take around 200 epochs using 1/2 and full labels, and around 100 epochs using 1/4, 1/8, and 1/16 labels.

For our semi-supervised approach (SS-CoSOD), we fine-tune the pre-trained model (from stage 1 in Fig. 3 in the main paper) using the learning rate is set as 10^{-7} for feature extractor and 10^{-6} for other parts, and the weight decay is set as 10^{-6}.

For training our Confidence Estimation Network, we randomly divided the labeled data into training (80%) and validation (20%) sets. We used the Adam optimizer for training with initial learning rate = 2×10^{-4} with a weight decay = 10^{-4}. The step of the learning rate scheduler is set as 7. We used a batch size of 32 to train this model.

Table 3. Comparison of our model with other state-of-the-art models on 3 benchmarks. We achieve state-of-the-art performance on the test datasets.

	CoCA				Cosal2015				CoSOD3k			
Method	MAE↓	$F_\beta^{max}\uparrow$	$E_\phi^{max}\uparrow$	$S_\alpha\uparrow$	MAE↓	$F_\beta^{max}\uparrow$	$E_\phi^{max}\uparrow$	$S_\alpha\uparrow$	MAE↓	$F_\beta^{max}\uparrow$	$E_\phi^{max}\uparrow$	$S_\alpha\uparrow$
GCAGC [11] (CVPR20)	0.111	0.523	0.754	0.669	0.085	0.813	0.866	0.817	0.100	0.740	0.816	0.785
CoEGNet [2] (TPAMI21)	0.106	0.493	0.717	0.612	0.077	0.832	0.882	0.836	0.092	0.736	0.825	0.762
GICD [12] (ECCV20)	0.126	0.513	0.715	0.658	0.071	0.844	0.887	0.844	0.079	0.770	0.848	0.797
GCoNet [3] (CVPR21)	0.105	0.544	0.760	0.673	0.068	0.847	0.887	0.845	0.071	0.777	0.860	0.802
DCFM [10] (CVPR22)	**0.085**	0.598	**0.783**	**0.710**	0.067	0.856	0.892	0.838	0.067	0.805	0.874	0.810
CoRP [13] (TPAMI23)	-	0.551	0.715	0.686	-	**0.885**	**0.913**	**0.875**	-	0.798	0.862	0.820
UFO [7] (TMM23)	0.095	0.571	**0.782**	**0.697**	0.064	0.865	0.906	0.860	0.073	0.797	0.874	0.819
GEM [9] (CVPR23)	0.095	0.599	**0.808**	**0.726**	0.053	0.882	0.933	0.885	0.061	0.829	0.911	0.853
DMT [5] (CVPR23)	0.108	**0.619**	**0.800**	**0.725**	0.0454	0.905	0.936	0.897	0.063	0.835	0.895	0.851
Ours (SS-CoSOD with ImgNet)	0.091	0.591	0.778	0.707	**0.061**	0.865	0.900	0.852	**0.062**	**0.809**	**0.882**	**0.821**

Table 4. Comparison of our US-CoSOD model with a variant version that normalizes the overlap area between the DINO SA mask and STEGO segmentation mask by the STEGO segmentation mask area.

Dataset	Method	MAE↓	$F_\beta^{max}\uparrow$	$E_\phi^{max}\uparrow$	$S_\alpha\uparrow$
CoCA	US-CoSOD (Variant)	0.131	0.410	0.650	0.590
	US-CoSOD (Proposed)	**0.165**	**0.461**	**0.676**	**0.610**
Cosal2015	US-CoSOD (Variant)	0.143	0.613	0.713	0.681
	US-CoSOD (Proposed)	**0.112**	**0.760**	**0.823**	**0.767**
CoSOD3k	US-CoSOD (Variant)	0.127	0.579	0.714	0.666
	US-CoSOD (Proposed)	**0.124**	**0.684**	**0.793**	**0.724**

Table 5. Average F-measure of the baselines over the categories on which the categorical F-measure scores of DINO are lower than its average F-measure on the test dataset.

Model	CoCA	Cosal2015	CoSOD3k
DINO (DI)	0.269	0.598	0.452
DINO+STEGO (DI+ST)	0.331	0.654	0.529
US-CoSOD	**0.408**	**0.738**	**0.577**

Table 6. Comparison of the confidence estimation error (Mean Squared Error) of different backbone networks for our CEN module on the unlabeled dataset.

Model	1/16 (576)	1/4 (2303)	1/2 (4607)
MobileNetV2 (3.4M)	0.210	0.171	0.168
DINO (ViTS8) (22.2M)	0.207	0.177	0.174
DINO (ViTB8) (86M)	0.208	0.176	0.170
DINO (ResNet50) (25.6M)	**0.204**	**0.166**	**0.160**

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *ECCVW What is Motion For?*, 2022. 1, 4, 6

[2] Deng-Ping Fan, Tengpeng Li, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Ming-Ming Cheng, Huazhu Fu, and Jianbing Shen. Re-thinking co-salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 8

[3] Qi Fan, Deng-Ping Fan, Huazhu Fu, Chi-Keung Tang, Ling Shao, and Yu-Wing Tai. Group collaborative learning for co-salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12288–12298, 2021. 8

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 7

[5] Long Li, Junwei Han, Ni Zhang, Nian Liu, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, and Fahad Shahbaz Khan. Discriminative co-saliency and background mining transformer for co-salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7247–7256, 2023. 8

[6] Xi Shen, Alexei A Efros, Armand Joulin, and Mathieu Aubry. Learning co-segmentation by segment swapping for retrieval and discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5082–5092, 2022. 1, 4, 6

[7] Yukun Su, Jingliang Deng, Ruizhou Sun, Guosheng Lin, Hanjing Su, and Qingyao Wu. A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection. *IEEE Transactions on Multimedia*, 2023. 8

[8] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14543–14553, 2022. 1, 4, 6

[9] Yang Wu, Huihui Song, Bo Liu, Kaihua Zhang, and Dong Liu. Co-salient object detection with uncertainty-aware group exchange-masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19639–19648, 2023. 8

[10] Siyue Yu, Jimin Xiao, Bingfeng Zhang, and Eng Gee Lim. Democracy does matter: Comprehensive feature mining for co-salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 979–988, 2022. 6, 7, 8

[11] Kaihua Zhang, Tengpeng Li, Shiwen Shen, Bo Liu, Jin Chen, and Qingshan Liu. Adaptive graph convolutional network with attention graph clustering for co-saliency detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9050–9059, 2020. 8

[12] Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng. Gradient-induced co-saliency detection. In *European Conference on Computer Vision*, pages 455–472. Springer, 2020. 8

[13] Ziyue Zhu, Zhao Zhang, Zheng Lin, Xing Sun, and Ming-Ming Cheng. Co-salient object detection with co-representation purification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 8