Supplementary Materials:

Visual Consensus Prompting for Co-Salient Object Detection

Additional supplementary materials are provided to complement the necessary experimental details and further demonstrate the effectiveness and efficiency of our VCP. The materials mainly include: A. Additional Experimental Details. B. Ablation Analysis on Architecture Configuration and Hyperparameters. C. Some alternative schemes of VCP with more efficient parameters. D. Generalization verification on the bi-modal RGB-D Co-salient object detection (CoSOD).

A. Additional Experimental Details.

Datasets. All experiments are evaluated based on the most commonly used three datasets for CoSOD tasks: CoCA [19], CoSOD3k [2], and CoSal2015 [15]. The CoCA [19] dataset consists of 80 categories with 1295 images. Each image in the CoCA dataset contains extraneous salient objects apart from co-salient objects, and there are many data categories in this dataset that are not encountered in the training data. Therefore, the CoCA dataset is the most challenging and best reflects the robustness and generalization of each method. Cosal2015 [15] contains 2015 images of 50 categories, and each group has one or more challenging issues such as complex environments, object appearance variations, occlusion, and background clutters. CoSOD3k [2] is the largest CoSOD evaluation dataset so far, which contains 160 groups and totally 3,316 images.

The three primary training datasets widely used in CoSOD tasks are the DUT-class [19], COCO-9k [8], and COCO-SEG [11] datasets, and we denote them as D, C, and S, respectively. The DUTS-class [19] dataset consists of 291 categories with 8,250 images. The COCO-9k [8] dataset consists of 65 categories with 9,213 images, and the COCO-SEG dataset contains 200k images of 78 categories. The combinations of training datasets primarily used by existing methods include D+C [7], [21], [17], [6], as well as D+S [13], [14], [20]. In the comparative experiments, we conduct experiments using both combinations of training data to ensure fairness of comparison. In all ablation experiments, we use the smaller datasets DUT-class and COCO-9k as the training set.

Quantitative Results. Table 1 presents a quantitative comparison of our VCP with the most representative works in the past three years across multiple metrics. Compared to 12 full fine-tuning CoSOD methods, our VCP demonstrates a significant performance advantage on three commonly used benchmark test sets. Particularly on the CoCA dataset, which is the most challenging and best reflects the essence of the CoSOD task, our method outperforms the second-best by 5.6% and 6.8% in terms of S_m and F_m metrics, respectively. Compared to the prompt-based tuning method EVP [9], our VCP exhibits overwhelmingly superior performance. Comprehensive experiments demonstrate the promising prospects of introducing prompt learning for CoSOD tasks and validate the powerful capability of our proposed concise and parameter-efficient VCP in addressing this task. The Precision-Recall and F_m-Threshold curves shown in Fig. **S1** further demonstrate the effectiveness of our method compared to 12 state-of-the-art methods.

Table S1: Ablation on the tuning stages, i.e., tuning different Transformer blocks of the foundation model (SegFormer).

Tuning Stage	Tunable Param. (M)	COCA			CoSOD3k			CoSal2015		
		$S_m\uparrow$	$F_m^{\max}\uparrow$	MAE$\downarrow$	$S_m\uparrow$	$F_m^{\max}\uparrow$	MAE$\downarrow$	$S_m\uparrow$	$F_m^{\max}\uparrow$	MAE$\downarrow$
Stage-1	1.82	.683	.532	.123	.841	.821	.066	.883	.881	.057
Stage-1,2	2.22	.699	.559	.104	.847	.828	.064	.887	.891	.055
Stage-1,2,3	3.34	.723	.594	.097	.858	.836	.060	.901	.903	.047
Stage-1,2,3,4	4.94	**.774**	**.680**	**.069**	**.874**	**.868**	**.049**	**.911**	**.920**	**.037**

Qualitative Results. Fig. 5 illustrates the visual comparisons between our VCP and seven representative works across four selected scenarios. "Normal" denotes scenarios with relatively simple backgrounds and minimal interference from non-co-salient objects. "Distraction" indicates scenarios where there are numerous interfering salient objects apart from the co-salient objects. "Tiny Object" indicates that the co-salient objects in the image have a smaller pixel ratio. "Variation" represents scenarios where co-salient objects within specific groups exhibit various presentation styles. Through the visual comparisons in Fig. 5, it can be observed that our VCP achieves more effective localization and segmentation performance even when facing challenging cases such as multiple salient object interferences, small or diverse co-salient objects. The performance

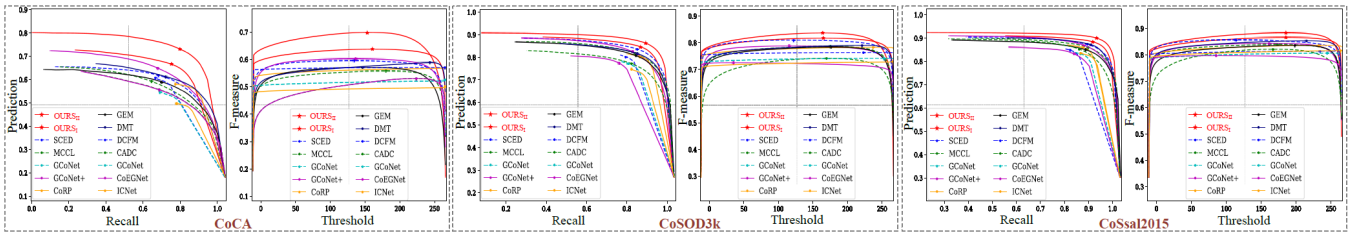

Fig. S1: Precision-Recall and F_m-Threshold curves are compared between our VCP and 12 SOTA methods on three test sets.

Table S2: Ablation on the parameter scale factor r.

Scale-r	Tunable Param. (M)	COCA $S_m\uparrow$	COCA $F_m^{\max}\uparrow$	COCA MAE$\downarrow$	CoSOD3k $S_m\uparrow$	CoSOD3k $F_m^{\max}\uparrow$	CoSOD3k MAE$\downarrow$	CoSal2015 $S_m\uparrow$	CoSal2015 $F_m^{\max}\uparrow$	CoSal2015 MAE$\downarrow$
r=64	2.71	.715	.582	.099	.858	.842	.055	.896	.901	.046
r=32	2.76	.740	.627	.088	.865	.853	.055	.905	.912	.044
r=16	2.89	.759	.656	.081	.872	.862	.053	.911	.918	.039
r=8	3.34	.769	.665	**.069**	.872	.864	.051	**.913**	0.918	**.035**
r=4	4.94	**.774**	**.680**	**.069**	**.874**	**.868**	**.049**	.911	**.920**	.037

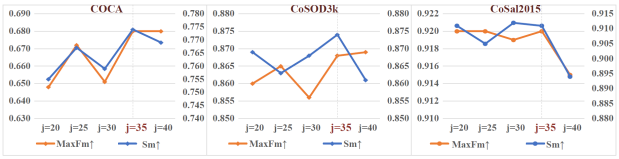

Fig. S2: Ablation on the predefined Saliency Seeds (SS) quantity j in our CPG. Bold indicates the default settings in our VCP.

gains are attributed to our task-specific VCP, which adequately stimulates the semantic representations in the foundation model.

B. Ablation Analysis on Architecture Configuration and Hyperparameters.

Effectiveness of the tuning stages. Further experiments are conducted to validate the impact of tuning different stages of the foundation model on the final performance. As shown in Table **S1**, the improvement in model performance becomes increasingly evident as the prompt addition stage deepens.

Effectiveness of the parameter scale factor r. Table **S2** presents the ablation analysis of the scale factor used when performing downsampling with MLP on frozen embeddings. The scale factor significantly impacts the tunable parameters of the model, and while $r = 8$ strikes a balance between parameters and performance, $r = 4$ is ultimately chosen considering the model's performance.

Effectiveness of hyperparameters in CPG: The main insight of our CPG is to enforce effective tunable parameters to focus on and mine co-salient representations within intra-group frozen embeddings, thereby generating task-specific consensus prompts P_{Co}.

We achieve our CPG by predefining j learnable saliency seeds, clustering representative prototypes of salient objects

Table S3: Ablation on the number of representative consensus seeds selected in our CPG (top-k).

Top-k	Tunable Param. (M)	COCA $S_m\uparrow$	COCA $F_m^{\max}\uparrow$	COCA MAE$\downarrow$	CoSOD3k $S_m\uparrow$	CoSOD3k $F_m^{\max}\uparrow$	CoSOD3k MAE$\downarrow$	CoSal2015 $S_m\uparrow$	CoSal2015 $F_m^{\max}\uparrow$	CoSal2015 MAE$\downarrow$
k=16	4.94	.760	.659	.085	.873	.867	.055	.908	.916	.046
k=24	4.94	.768	.669	.075	.871	.862	.055	**.912**	**.922**	.041
k=**32**	4.94	**.774**	**.680**	**.069**	.874	.868	**.049**	.911	.920	**.037**
k=40	4.95	.752	.655	.079	**.876**	**.870**	.052	.906	.917	.044
k=48	4.96	.754	.644	.076	.869	.861	.052	.907	.916	.040

Table S4: Ablation experiments are conducted based on different unified mapping dimensions d of the Segformer prediction head. "OOM" indicates out-of-memory.

Segformer Head	Tunable Param. (M)	COCA $S_m\uparrow$	COCA $F_m^{\max}\uparrow$	COCA MAE$\downarrow$	CoSOD3k $S_m\uparrow$	CoSOD3k $F_m^{\max}\uparrow$	CoSOD3k MAE$\downarrow$	CoSal2015 $S_m\uparrow$	CoSal2015 $F_m^{\max}\uparrow$	CoSal2015 MAE$\downarrow$
d =**768**	7.83	OOM			OOM			OOM		
d =384	4.98	OOM			OOM			OOM		
d =128	3.84	.747	.656	.101	.872	.861	.054	.909	.916	.038

Table S5: Ablation regarding the preset parameter d in PH.

PH-d	Tunable Param. (M)	COCA $S_m\uparrow$	COCA $F_m^{\max}\uparrow$	COCA MAE$\downarrow$	CoSOD3k $S_m\uparrow$	CoSOD3k $F_m^{\max}\uparrow$	CoSOD3k MAE$\downarrow$	CoSal2015 $S_m\uparrow$	CoSal2015 $F_m^{\max}\uparrow$	CoSal2015 MAE$\downarrow$
d=64	4.53	.762	.658	.075	.870	.862	.051	.907	.914	.040
d=96	4.74	.766	.661	.080	.871	.865	.058	.909	**.921**	.051
d=**128**	4.94	**.774**	**.680**	**.069**	.874	**.868**	.049	**.911**	.920	**.037**
d=192	5.39	.757	.647	.082	**.878**	.867	**.048**	.910	.913	**.037**
d=256	5.87	.749	.633	.084	.871	.852	.051	.909	.912	**.037**

to generate saliency estimation maps $\{M^s\}_{s=1}^4$ for each individual images. As shown in Fig. **S2**, different numbers of saliency seeds result in certain fluctuations in the model's tunable parameters, but all ensure competitive performance gains. Considering both the effectiveness of performance and the efficiency of tunable parameters, $j = 35$ is selected.

The generated saliency estimation maps are utilized to focus on pixel embeddings to form consensus seeds $Co_{seed} \in \mathbb{R}^{NL_s \times C_r}$. Finally, we select the top-k most relevant consensus seeds to form representative consensus $Co_{seed}^{rep} \in \mathbb{R}^{k \times C_r}$. We further conduct experiments to validate the effectiveness of the predefined saliency seeds and the number of selected representative consensus seeds (top-k). As shown in Table **S3**, the quantity of representative consensus seeds has a minimal impact on the overall number of tunable parameters in the model, but it does introduce

Table S6: Some alternative schemes of VCP exhibit more efficient parameters and competitive performance. r represents the parameter scale factor (default $r = 4$), d represents the unified mapping dimension in PH (default $d = 128$), and "Share MLP" indicates the adoption of stage-shared MLP in CPD (default as adaptive tuning).

Combination	Train Dataset	Tunable Param. (M)	Model Size (MB)	COCA			CoSOD3k			CoSal2015		
				$S_m\uparrow$	$F_m^{\max}\uparrow$	MAE$\downarrow$	$S_m\uparrow$	$F_m^{\max}\uparrow$	MAE$\downarrow$	$S_m\uparrow$	$F_m^{\max}\uparrow$	MAE$\downarrow$
EVP	S+D	3.70	14.10	0.686	0.546	0.126	0.839	0.813	0.076	0.876	0.874	0.068
VCP($r = 8 + d = 96$)	S+D	3.13	12.08	0.814	0.745	0.054	**0.895**	0.892	**0.043**	0.922	0.933	0.034
VCP($r = 8$+Share MLP)	S+D	3.24	12.47	0.817	**0.753**	**0.053**	0.889	0.887	0.047	0.918	0.926	0.035
VCP($r = 4$)	S+D	4.94	19.02	**0.819**	0.752	0.054	**0.895**	**0.893**	**0.043**	**0.927**	**0.941**	**0.030**
VCP($r = 8 + d = 96$)	C+D	3.13	12.08	0.770	0.672	0.079	0.873	0.869	0.050	0.911	0.919	0.038
VCP($r = 8$+Share MLP)	C+D	3.24	12.47	0.767	0.665	0.072	0.870	0.862	0.051	0.909	0.918	0.040
VCP($r = 8$)	C+D	3.34	12.89	0.769	0.665	0.069	0.872	0.864	0.051	0.913	0.918	0.035
VCP($r = 4$)	C+D	4.94	19.02	0.774	0.680	0.069	0.874	0.868	0.049	0.911	0.920	0.037

Table S7: Comprehensive comparison with 5 cutting-edge methods and ablation analysis of the used backbone.

Datasets	Metrics	CoRP TPAMI$_{23}$	**OURS**$_I$ —	EVP CVPR$_{23}$	GEM CVPR$_{23}$	MCCL AAAI$_{23}$	SCED ACMM$_{23}$	**OURS**$_{II}$ —	**OURS**$_{II}$ —
CoCA	$S_m\uparrow$	0.732	0.774	0.686	0.726	0.713	0.741	**0.819**	0.805
	$F_m^{\max}\uparrow$	0.619	0.680	0.546	0.599	0.584	0.629	**0.752**	0.731
	$MAE\downarrow$	0.093	0.069	0.126	0.095	0.097	0.084	**0.054**	0.059
	$E_m\uparrow$	0.773	0.813	0.708	0.767	0.764	0.804	**0.830**	0.823
Training Dataset		C+D	C+D	S+D	S+D	S+D	S+D	S+D	S+D
Foundation Model		PVTv2	SegF	SegF	T2T-ViT	PVTv2	CLIP-ViT	SegF	PVTv2
Tunable Param. (M)		—	4.9	3.7	52.3	27.1	156.7	4.9	4.95
Model Size (MB)		—	19	14.1	199.7	104.5	1750	19	19

some performance fluctuations. For the comprehensive effectiveness of the model's performance, $k = 32$ is determined as the default setting.

Effectiveness of hyperparameters in PH: As shown in Table **S4**, segformer's original prediction head up-projects multi-scale features onto a unified high dimensionality (d=768) and employs multiple fully connected layers for prediction, resulting in a higher number of tunable parameters (3.15M).

Hence, a more concise prediction head (PH) is designed (1.49M). Our PH initially projects multi-scale features to a lower dimensionality ($d = 128$). Subsequently, the deepest features are processed with ASPP and utilized to guide FPN-like decoding and input to a linear classifier for category prediction. We conducted ablation experiments on the set unified dimensionality d. As shown in Table **S5**, the hyperparameter d is not the key factor affecting the overall parameter count of the model. It can be observed that $d = 96$ demonstrates certain performance competitiveness while reducing a certain number of tunable parameters. Considering both the effectiveness of performance and the efficiency of tunable parameters, $d = 128$ is selected.

Comparison and analysis regarding the foundation models used. For fairness in comparison, we use foundation models of comparable scale to existing methods (Table **S7**). Additionally, without optimizing hyperparameters, we conduct ablation experiments on an additional baseline model, PVTv2 [12]. Through comparison, it is observed that using foundation models of similar or even identical scale, our proposed parameter-efficient method achieves a substantial performance improvement.

C. Parameter-efficient combination schemes

We conduct additional experiments to validate the potential of parameter efficiency while ensuring competitive performance gains. Through extensive analysis of ablation experiments, we find that three main predefined parameters significantly impact the overall parameter scale: the parameter scale factor r in CPG, the unified mapping dimension d in PH, and the fine-tuning mode in CPD. Based on the performance from the conducted ablation experiments as a reference, and considering the competitive scale of tunable parameters compared to existing prompt tuning methods (3.7M), some combinations of VCP schemes are implemented.

As shown in Table **S6**, our supplementary three lightweight schemes (with a minimum of only 3.13M tunable parameters) exhibit significantly better parameter efficiency and overwhelming performance effectiveness compared to the state-of-the-art method EVP (with 3.7M tunable parameters). These lightweight schemes achieve competitive performance while maintaining a reduced parameter footprint, making them strong alternatives to our default settings.

D. Generalization verification on the RGB-D CoSOD

To validate the high generalization capability of the proposed VCP for CoSOD-related tasks, we transfer VCP to address the bimodal RGB-D CoSOD task. For fairness of comparison, we retrain VCP with the same training data as existing methods and conduct tests on three benchmark datasets. Specifically, DUT-class (consists of 291 categories with 8,250 images) [19] is used as the training set, and cor-

Table S8: Generalization verification on the RGB-D CoSOD task. Our VCP is retrained with the same training set (DUT-class [19]) as existing RGB-D CoSOD methods and conducts testing on three benchmark datasets (CoSal1k [18], CoSal150 [16] and CoSal183 [5]).

Datasets	Metrics	HSCS [1] TMM$_{19}$	CBCS [4] TIP$_{22}$	CoEG [2] TPAMI$_{21}$	GICD ECCV$_{20}$ [19]	GICD$_{+D}$	ICNet NeurIPS$_{20}$ [6]	ICNet$_{+D}$	GCoN CVPR$_{21}$ [3]	GCoN$_{+D}$	CADC ICCV$_{21}$ [17]	CADC$_{+D}$	CTNet TIP$_{22}$ [18]	OURS
CoSal1k	$S_m\uparrow$	0.478	0.529	0.811	0.793	0.756	0.840	0.788	0.807	0.810	0.850	0.841	0.875	**0.893**
	$E_m^{\max}\uparrow$	0.640	0.651	0.866	0.846	0.816	0.895	0.844	0.864	0.861	0.892	0.885	0.913	**0.933**
	$F_m^{\max}\uparrow$	0.428	0.505	0.799	0.783	0.746	0.846	0.767	0.814	0.815	0.837	0.832	0.865	**0.896**
	MAE$\downarrow$	0.242	0.230	0.085	0.090	0.102	0.073	0.093	0.083	0.080	0.077	0.090	0.063	**0.045**
	$E_m\uparrow$	0.512	0.520	0.815	0.842	0.809	0.891	0.839	0.856	0.855	0.861	0.835	0.882	**0.921**
	$F_m\uparrow$	0.293	0.354	0.770	0.778	0.743	0.836	0.754	0.809	0.810	0.802	0.779	0.834	**0.881**
CoSal150	$S_m\uparrow$	0.661	0.576	0.851	0.821	0.812	0.888	0.856	0.821	0.867	0.895	0.895	0.910	**0.920**
	$E_m^{\max}\uparrow$	0.884	0.726	0.890	0.878	0.880	0.939	0.920	0.876	0.924	0.933	0.759	0.944	**0.953**
	$F_m^{\max}\uparrow$	0.836	0.608	0.870	0.846	0.831	0.906	0.867	0.834	0.878	0.893	0.905	0.912	**0.924**
	MAE$\downarrow$	0.163	0.211	0.079	0.077	0.089	0.050	0.069	0.080	0.059	0.062	0.068	0.053	**0.039**
	$E_m\uparrow$	0.605	0.550	0.873	0.874	0.876	0.935	0.910	0.871	0.921	0.904	0.892	0.920	**0.944**
	$F_m\uparrow$	0.590	0.443	0.850	0.840	0.876	0.896	0.844	0.828	0.869	0.858	0.848	0.880	**0.905**
CoSal183	$S_m\uparrow$	0.704	0.620	0.728	0.645	0.661	0.763	0.719	0.707	0.708	0.709	0.649	0.764	**0.795**
	$E_m^{\max}\uparrow$	0.790	0.748	0.764	0.668	0.687	0.768	0.757	0.726	0.719	0.770	0.759	0.837	**0.873**
	$F_m^{\max}\uparrow$	0.603	0.486	0.558	0.442	0.448	0.617	0.548	0.553	0.556	0.609	0.592	0.654	**0.699**
	MAE$\downarrow$	0.079	0.088	0.073	0.111	0.100	0.066	0.073	0.088	0.094	0.110	0.121	0.068	**0.051**
	$E_m\uparrow$	0.710	0.636	0.742	0.655	0.673	0.751	0.742	0.703	0.705	0.680	0.668	0.752	**0.790**
	$F_m\uparrow$	0.529	0.399	0.535	0.433	0.438	0.600	0.533	0.531	0.537	0.524	0.505	0.583	**0.630**
Model Size (MB)		—	—	69	533	591	70	163	542	600	1499	1596	570	**19.02**
Running Time (FPS)		0.12	1	0.43	55.56	40	76.9	71.4	62.5	41.32	31.25	19.23	17.54	**40.4**

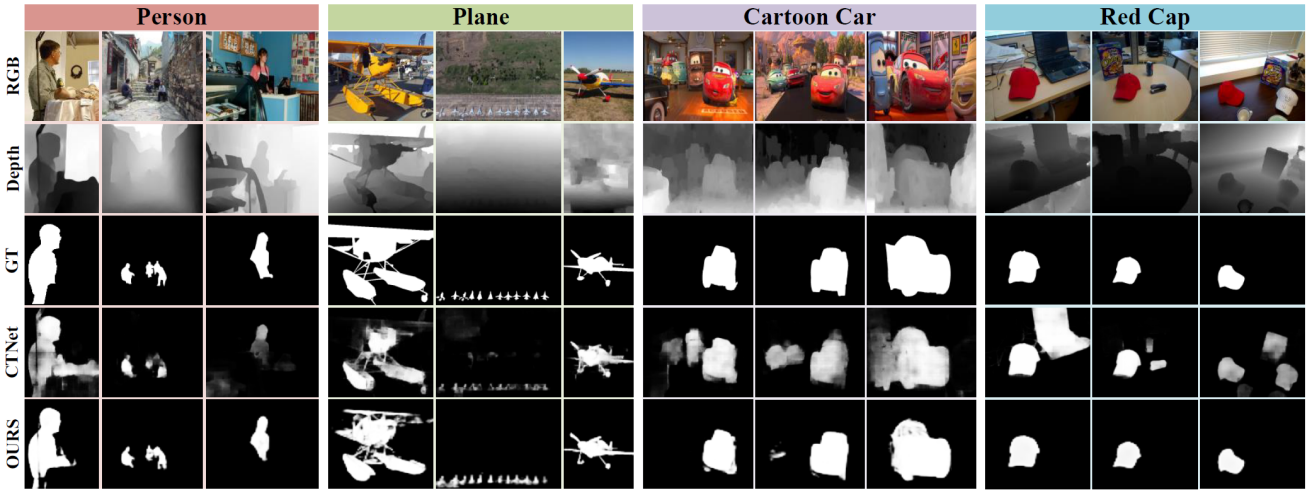

Fig. S3: Visual comparison between our VCP and the most representative RGB-D CoSOD method CTNet [18].

responding depth data is generated using depth estimation algorithms [10]. The three widely used RGB-D CoSOD test sets include: CoSal1k [18] (with 106 groups, totaling 1000 image pairs), CoSal150 [16] (with 21 groups, totaling 150 image pairs), and CoSal183 [5] (with 16 groups, totaling 183 image pairs).

To provide more convincing evidence of the effectiveness of the proposed VCP on this task, we refrain from performing any model fine-tuning or post-processing. Furthermore, unlike existing methods that employ dual-stream architectures and specialized cross-modal fusion modules to enhance performance, we utilize a simple single-stream architecture and employ the most basic early fusion strategy by simply adding the cross-modal images as inputs. Comprehensive experiments demonstrate that our simplified version of VCP still achieves remarkable performance on RGB-D CoSOD tasks. As shown in Fig. **S3**, in scenarios with some interference and insufficient depth effectiveness, our VCP still achieves superior segmentation performance. Table **S8** presents the quantitative comparison results with state-of-the-art RGB-D CoSOD methods on three benchmark datasets, providing strong evidence of the high generalization capability of our VCP for related group-based segmentation tasks.

References

[1] Runmin Cong, Jianjun Lei, Huazhu Fu, Qingming Huang, Xiaochun Cao, and Nam Ling. Hscs: Hierarchical sparsity based co-saliency detection for rgbd images. *IEEE Trans. Multimedia*, 21(7):1660–1671, 2018. 4

[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 Trans. Pattern Anal. Mach. Intell.*, 44(8):4339–4354, 2021. 1, 4

[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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 12288–12298, 2021. 4

[4] Huazhu Fu, Xiaochun Cao, and Zhuowen Tu. Cluster-based co-saliency detection. *IEEE Trans. Image Process.*, 22(10): 3766–3778, 2013. 4

[5] Huazhu Fu, Dong Xu, Stephen Lin, and Jiang Liu. Object-based rgbd image co-segmentation with mutex constraint. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 4428–4436, 2015. 4

[6] Wen-Da Jin, Jun Xu, Ming-Ming Cheng, Yi Zhang, and Wei Guo. Icnet: Intra-saliency correlation network for co-saliency detection. In *Proc. Adv. Neural Inform. Process. Syst.*, pages 18749–18759, 2020. 1, 4

[7] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 7247–7256, 2023. 1

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proc. Eur. Conf. Comput. Vision*, pages 740–755. Springer, 2014. 1

[9] Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 19434–19445, 2023. 1

[10] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proc. IEEE/CVF Int. Conf. Comput. Vision*, pages 12179–12188, 2021. 4

[11] Chong Wang, Zheng-Jun Zha, Dong Liu, and Hongtao Xie. Robust deep co-saliency detection with group semantic. In *Proc. AAAI Conf. Artif. Intell.*, pages 8917–8924, 2019. 1

[12] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Comput. Vis. Media*, 8(3):415–424, 2022. 3

[13] Yang Wu, Huihui Song, Bo Liu, Kaihua Zhang, and Dong Liu. Co-salient object detection with uncertainty-aware group exchange-masking. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 19639–19648, 2023. 1

[14] Peiran Xu and Yadong Mu. Co-salient object detection with semantic-level consensus extraction and dispersion. In *Proc. ACM Int. Conf. Multimedia*, pages 2744–2755, 2023. 1

[15] Dingwen Zhang, Junwei Han, Chao Li, and Jingdong Wang. Co-saliency detection via looking deep and wide. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 2994–3002, 2015. 1

[16] Dingwen Zhang, Junwei Han, Chao Li, Jingdong Wang, and Xuelong Li. Detection of co-salient objects by looking deep and wide. *Int. J. Comput. Vis.*, 120:215–232, 2016. 4

[17] Ni Zhang, Junwei Han, Nian Liu, and Ling Shao. Summarize and search: Learning consensus-aware dynamic convolution for co-saliency detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pages 4167–4176, 2021. 1, 4

[18] Ni Zhang, Junwei Han, and Nian Liu. Learning implicit class knowledge for rgb-d co-salient object detection with transformers. *IEEE Trans. Image Process.*, 31:4556–4570, 2022. 4

[19] Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng. Gradient-induced co-saliency detection. In *Proc. Eur. Conf. Comput. Vision*, pages 455–472. Springer, 2020. 1, 3, 4

[20] Peng Zheng, Jie Qin, Shuo Wang, Tian-Zhu Xiang, and Huan Xiong. Memory-aided contrastive consensus learning for co-salient object detection. In *Proc. AAAI Conf. Artif. Intell.*, pages 3687–3695, 2023. 1

[21] Ziyue Zhu, Zhao Zhang, Zheng Lin, Xing Sun, and Ming-Ming Cheng. Co-salient object detection with co-representation purification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(7):8193–8205, 2023. 1