Discriminative Co-Saliency and Background Mining Transformer for Co-Salient Object Detection (Supplementary Materials)

Long Li[1] Junwei Han[1] Ni Zhang[1] Nian Liu[2†]
Salman Khan[2,3] Hisham Cholakkal[2] Rao Muhammad Anwer[2] Fahad Shahbaz Khan[2,4]
[1] Northwestern Polytechnical University [2] Mohamed bin Zayed University of Artificial Intelligence
[3]Australian National University [4]CVL, Linköping University

Abstract

There are four aspects included in this supplementary material. 1) The supplementary diagrams for R2R and CoT2T. 2) Some visual samples to certify the effectiveness of R2R. 3) Ablation study to verify the necessity of iteratively deploying the R2R, CtP2T, CoT2T, and TGFR. 4) More qualitative comparison examples between DMT and other state-of-the-art models and more visual analysis examples for verifying the effectiveness of the CtP2T and TGFR modules. 5) The limitation discussion of our model.

1. Diagrams

Due to the space limitation of the main paper, we provide the R2R and CoT2T diagrams in this supplementary material.

The diagram of R2R is shown in Figure 1. Specifically, we first obtain region-level query, *i.e.* $\mathrm{R}_1(\boldsymbol{F}_j^d)$, and key and value, *i.e.* $\mathrm{R}_2(\boldsymbol{F}_j^d)$, for each image feature using the R_1 and R_2 operations, respectively. Then, we use the basic transformer operation on these region representations and achieve inter-image correlated region-level features. Finally, we upsample these region features via R_1^{-1} and sum them with the original image features to propagate the aggregated inter-image information.

The diagram of CoT2T is shown in Figure 2. Specifically, we first collect all co-saliency tokens from the CtP2P module and transform their information to a group token via the basic transformer operation for consensus information aggregation. Then, the consensus cues contained in the group token are distributed to all co-saliency tokens via another basic transformer operation. In this way, we model the consensus patterns at the token level, as a supplement to the region-level consensus modeling in the R2R correlation.

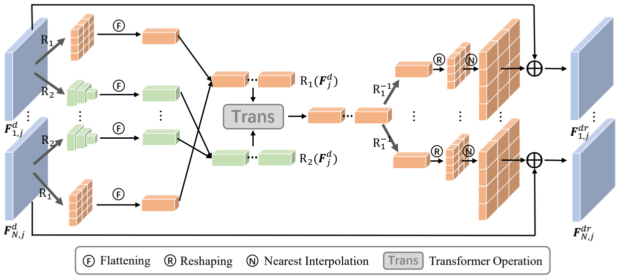

Figure 1. **Diagram of the R2R module.** It consists of region-level features extraction via R_1 and R_2, the transformer operation on the extracted region-level features, and the residual connection between the processed region features and the original features.

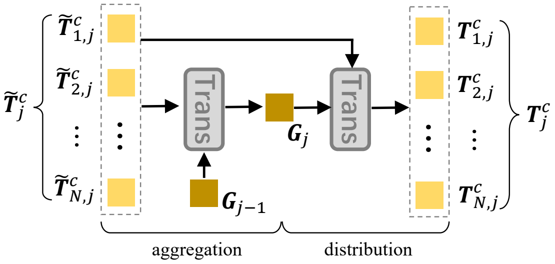

Figure 2. **Diagram of the CoT2T module.** We first aggregate the consensus information from all co-saliency tokens to the group token. Then, the aggregated information is distributed to all co-saliency tokens.

2. R2R Analysis

We provide some visual samples in Figure 3 to certify the effectiveness of the R2R correlation. It can be seen that the R2R-enhanced features can distinguish the co-salient objects from the complex backgrounds to guarantee accurate detection and precise segmentation of the co-salient objects. As a result, the model with R2R can integrally segment the challenging co-salient objects and exclude distracting objects, while the model without R2R can not achieve this. This indicates that the inter-image correlation modeled by R2R can improve the consensus representation learning ability for the segmentation features and enhance their cor-

[†]Corresponding author: liunian228@gmail.com.

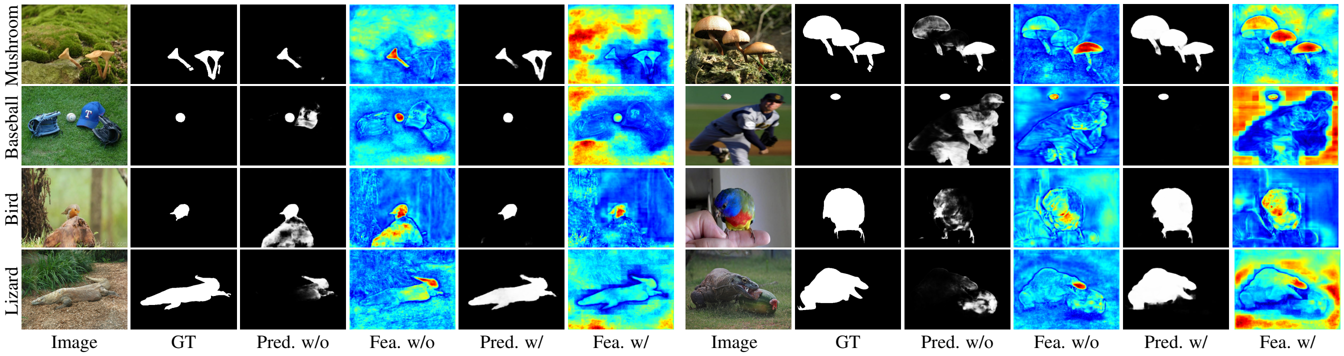

Figure 3. **Visualization of some feature maps (Fea.) and predictions (Pred.) of the models with (w/) or without (w/o) using R2R.**

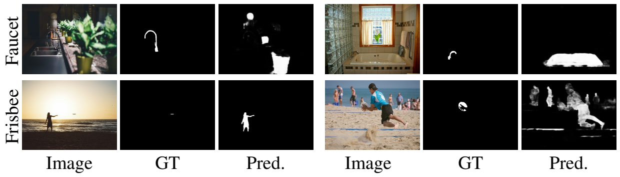

Figure 4. **Some failed predictions.**

responding details.

3. Ablation Study on Iterative Learning

In this paper, we iteratively deploy our proposed components, *i.e*. R2R, CtP2T, CoT2T, and TGFR, synchronizing with the decoder layers of the original FPN architecture. To verify the necessity of this deployment strategy, we conduct an ablation study in Table 1 by gradually increasing the iteration steps from 1 to 5. we can observe that the performance generally improves when increasing the number of iterations. This proves that iteratively using these components can gradually enhance the representation ability of the detection tokens and the segmentation features.

Table 1. **Comparison of the model performance using different iteration steps for deploying the four components.**

Iterations	CoCA [1]			
	S_m ↑	E_ξ ↑	maxF↑	MAE ↓
1	0.7131	0.7944	0.5967	0.1103
2	0.7175	0.7929	0.6020	0.1056
3	0.7199	0.7937	0.6136	0.1133
4	0.7192	0.8056	0.6190	0.1153
5	0.7246	0.8001	0.6190	0.1084

4. More Qualitative Comparisons and Visual Analysis

We provide more qualitative comparisons with the state-of-the-art models in Figure 5. Our model not only detects the co-salient objects more accurately but also performs high-quality segmentation for them, such as cups with handles and small mushrooms with stems.

We also provide more visual analysis examples for the effectiveness demonstration of CtP2T and TGFR in Figure 6 and Figure 7, respectively. They evidently demonstrate that 1) CtP2T can help pay more attention to the channels that can differentiate co-salient objects from backgrounds and suppress the channels that confuse these two regions; 2) TGFR can help obtain more discriminative segmentation features for distinguishing co-saliency objects from distractors.

5. Limitations

We also report some failure predictions in Figure 4. DMT can't deal with some extreme challenging cases, *i.e*. a group containing many images in which co-salient objects are very small while background regions are very complex at the same time. It could be because our model constructs the co-saliency tokens from the highest-level CNN features, which may lose the accurate features of some small objects due to the pooling layers in the backbone CNN. Coupled with the complexity of the background region, it is very difficult for our model to detect these small co-salient objects.

Acknowledgments: This work was supported in part by Key-Area Research and Development Program of Guangdong Province (No.2021B0101200001).

References

[1] Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng. Gradient-induced co-saliency detection. In *ECCV*, pages 455–472, 2020. 2

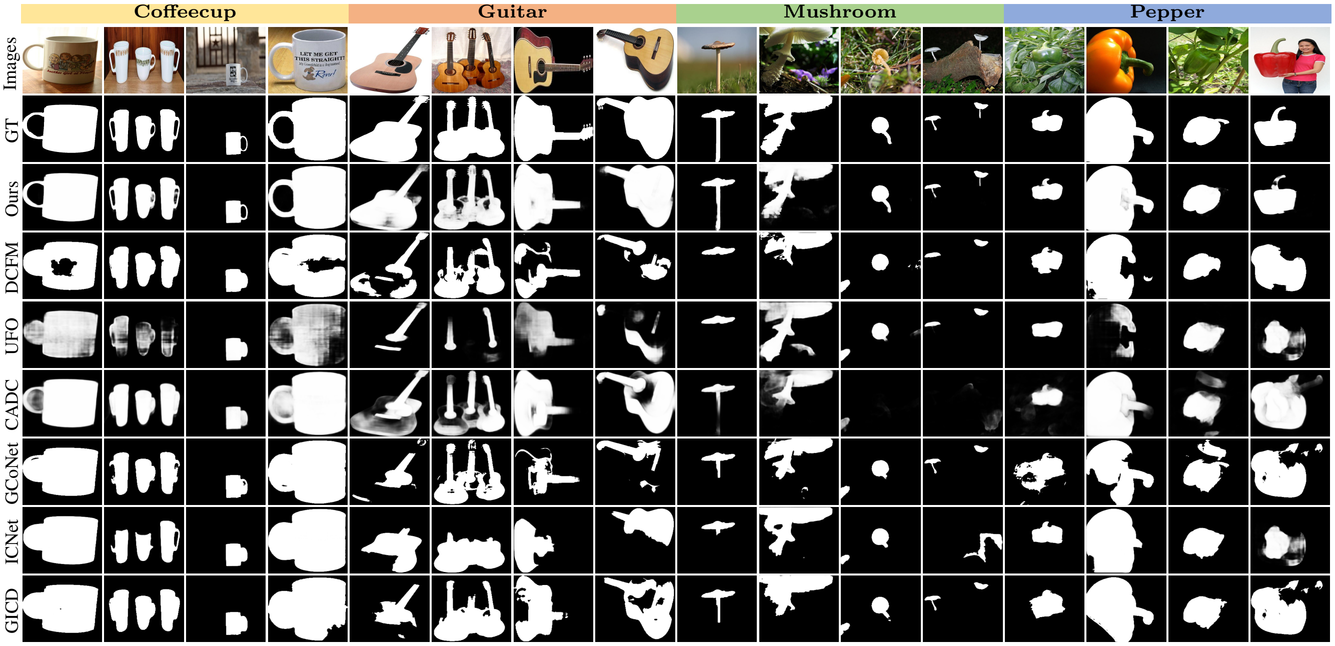

Figure 5. **Qualitative comparisons of our model with other state-of-the-art methods.**

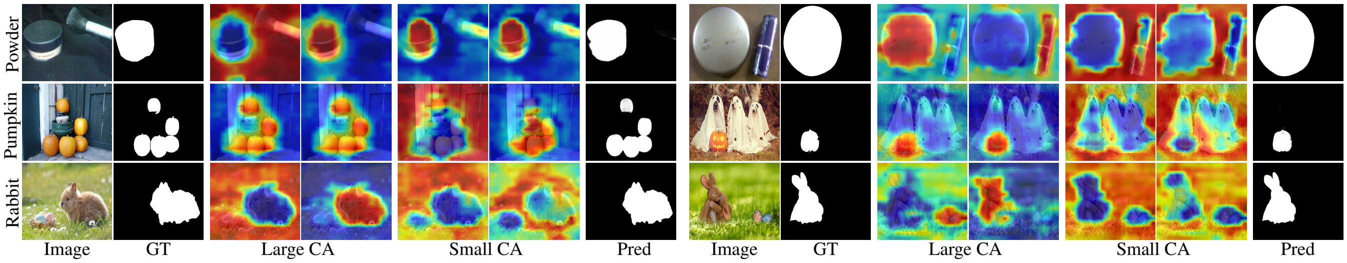

Figure 6. **Visual comparison among the channels with different channel attention weights in CtP2T.** We visualize some feature maps in $V_m^*(\boldsymbol{F})$ for the channels with large and small channel attention (CA) in CtP2T. We visualize two channels for large and small CA, respectively.

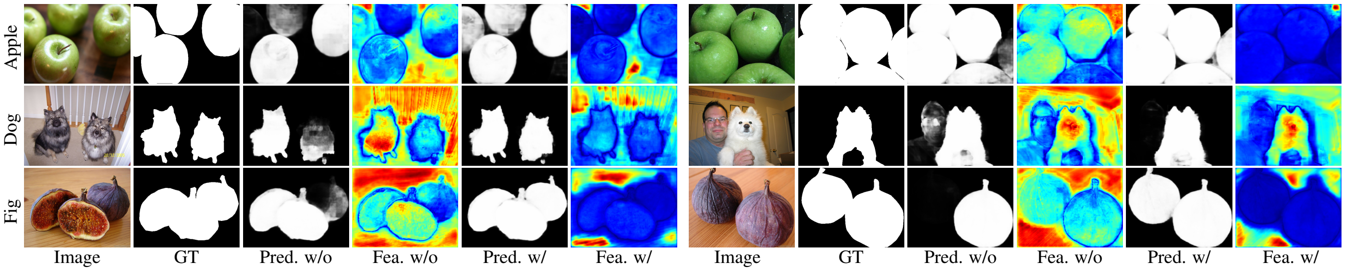

Figure 7. **Visualization of some feature maps (Fea.) and predictions (Pred.) of the models with (w/) or without (w/o) using TGFR.**