TF-SSD: A Strong Pipeline via Synergic Mask Filter for Training-free Co-salient Object Detection

Supplementary Material

A. Overview

This supplementary material provides additional details to complement our main paper. We elaborate on our implementation details in Sec. B, present more quantitative and qualitative results in Sec. C and Sec. D, and discuss the limitations of our method in Sec. E.

B. Additional implementation details

To handle challenging cases, several mechanisms are necessary to ensure robustness in practice. Due to space limitations in the main paper, we describe these details below.

B.1. Fallback mechanism in ISF

Our Intra-image Saliency Filter (ISF) relies on DINO's attention maps to identify salient objects from SAM [3] proposals. However, in some challenging scenarios, QMG may fail to segment salient masks, where all masks yield low saliency scores $s^{sal}_{n,t}$. To ensure robustness, we employ a fallback mechanism that directly generates masks from attention maps to avoid these bad cases.

In particular, if the maximum saliency score $\max_t(s^{sal}_{n,t})$ is lower than a threshold τ_{fb}, the attention map $\mathcal{A}_n$ is binarized to obtain a mask $m^{fb}_{n,t}$ that captures the salient region, formulated as:

$$m^{fb}_{n,t}(x,y) = \begin{cases} 1 & \text{if } A_n(x,y) > \tau_{attn}, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where τ_{attn} is the binarization threshold for $\mathcal{A}_n$. This single fallback mask $m^{fb}_{n,t}$ will replace the low-quality masks. This operation ensures that each image retains at least one salient candidate mask for subsequent processing. In our implementation, we set $\tau_{fb} = 0.05$, and τ_{attn} is dynamically set to a threshold that retains the top 50% of the highest attention values for each map $\mathcal{A}_n$.

B.2. Merging of multiple co-salient objects in IPS

Our Inter-image Prototype Selector (IPS) is designed to identify the target co-salient object in each image. However, individual images may contain multiple instances of the co-salient object, and SAM often segments them separately. To obtain accurate results for CoSOD, we have to merge these separate masks into a single one.

To address such cases, we employ a dual-verification mechanism with two thresholds: a semantic similarity threshold τ_{sem} and a consistency difference threshold τ_{diff}. For each image I_n with a selected mask $\hat{m}_n$, we examine the remaining masks in $\mathcal{M}^{salient}_n$ to identify extra co-salient objects. A candidate mask m_j is considered an extra co-salient object if it satisfies the following two conditions:

1) **Semantic similarity:** The candidate mask is semantically similar to the primary mask, measured by pairwise similarity $\langle p_{\hat{m}_n}, p_j \rangle \geq \tau_{sem}$.
2) **Cross-image consistency:** The candidate mask exhibits comparable cross-image consistency, measured by $|S^{co}_j - S^{co}_{\hat{m}_n}| < \tau_{diff}$.

The first condition ensures semantic coherence within the image, while the second verifies that the candidate mask represents an object that appears consistently across the group. Verified masks are merged with the primary mask to form the final segmentation mask. In our implementation, τ_{sem} is dynamically set as the top 80% of intra-image pairwise similarities, and τ_{diff} is set to 0.1.

C. Additional quantitative results

C.1. Different SAM area ratios

The area ratio threshold τ_{area} is a critical parameter in our QMG that filters out excessively small masks (Eq. 2 in the main paper). To evaluate its impact on CoSOD performance, we conduct ablation experiments with four different threshold values: 0.002, 0.005, 0.01, and 0.02 on the CoCA [2] dataset. We employ two metrics: (1) **F-measure**: the final CoSOD performance obtained using the complete TF-SSD framework; (2) **Oracle F**: the upper bound performance computed using selected oracle masks from QMG.

As shown in Tab. 1, when the threshold is too low (0.005), excessive trivial masks are retained in the candidate set, introducing noise that degrades the subsequent ISF and IPS. Meanwhile, the Oracle F score reflects the quality of the candidate pool. When the threshold is too high (0.02), some valid co-salient objects are incorrectly filtered out, which limits both the quality of the candidate pool and the final performance. $\tau_{area} = 0.01$ strikes an optimal balance, which effectively removes trivial masks while preserving diverse co-salient object candidates.

C.2. Performance using other DINO backbones

Our main experiments use DINO ViT-B/8 [1] as the backbone for feature extraction in ISF. To provide a broader performance assessment, we compare it with two other notable backbones: (1) ViT-B/16 [1] from the same DINO framework with a larger patch size; (2) ViT-B/14 [4] from the

Table 1. Ablation study on different area ratio thresholds τ_{area} on the CoCA dataset.

τ_{area}	$F_\beta^{\max}$	Oracle $F_\beta^{\max}$
0.002	0.545	0.696
0.005	0.601	0.722
0.01	**0.686**	**0.768**
0.02	0.653	0.737

more recent DINOv2 framework. We conduct experiments on the CoCA dataset, which highlights the trade-offs between different DINO versions.

As shown in Tab. 2, DINO ViT-B/8 significantly outperforms other configurations across all three metrics. A smaller patch size enables our model to capture finer spatial details, which is crucial for feature matching and prototype selection in our ISF and IPS. In contrast, although DINOv2 ViT-B/14 utilizes more advanced pretraining strategies, its larger patch size limits the perception of fine-grained information. This result demonstrates that a smaller patch size is more suitable for CoSOD that require precise localization and fine-grained feature alignment.

Table 2. Performance comparison of different DINO backbones on the CoCA dataset.

Backbone	$F_\beta^{\max}$	$E_\xi^{\max}$	S_α
DINO ViT-B/16	0.556	0.742	0.697
DINOv2 ViT-B/14	0.607	0.783	0.708
DINO ViT-B/8	**0.686**	**0.815**	**0.763**

C.3. Ablation on hyperparameter settings

We conduct additional ablation studies to evaluate the impact of hyperparameter settings.

Quality score weights. The balanced quality score $S_{n,t}^{ba}$ (Eq. 5 in the main paper) combines IoU prediction confidence and area preference through weights α and β, where $\alpha + \beta = 1$ to ensure normalized weighting between the two terms. Tab. 3 illustrates the performance with different weight settings on the CoCA dataset.

Table 3. Ablation on quality score weights on the CoCA dataset.

α	β	$F_\beta^{\max}$	$E_\xi^{\max}$	S_α
0.0	1.0	0.473	0.702	0.611
0.3	0.7	0.591	0.737	0.651
0.5	0.5	0.658	0.795	0.744
0.7	**0.3**	**0.686**	**0.815**	**0.763**
1.0	0.0	0.641	0.782	0.735

Ideal size range. The ideal size range $[r_{min}, r_{max}]$ (Eq. 4 in the main paper) defines the appropriate size for co-salient objects. Tab. 4 presents results for different range settings. The range $[0.15, 0.7]$ performs best, suggesting that co-salient objects typically occupy 15%-70% of the image area. More restrictive ranges (*e.g.* [0.2, 0.5]) limit the number of valid masks, while much wider ranges (*e.g.* [0.05, 0.85]) fail to effectively filter out background and large masks.

Table 4. Ablation on ideal size range on the CoCA dataset.

r_{min}	r_{max}	$F_\beta^{\max}$	$E_\xi^{\max}$	S_α
0.05	0.85	0.661	0.798	0.747
0.1	0.6	0.678	0.808	0.756
0.15	**0.7**	**0.686**	**0.815**	**0.763**
0.2	0.5	0.619	0.741	0.669

Table 5. Ablation on overlap threshold on the CoCA dataset.

τ_{con}	$F_\beta^{\max}$	$E_\xi^{\max}$	S_α
0.5	0.652	0.791	0.741
0.7	0.673	0.805	0.754
0.85	**0.686**	**0.815**	**0.763**
0.95	0.679	0.796	0.756

Overlap threshold. The overlap threshold τ_{con} controls how aggressively overlapping masks are filtered. As shown in Tab. 5, $\tau_{con} = 0.85$ achieves the optimal performance. Lower thresholds (*e.g.*, 0.5) retain too many redundant masks, while higher thresholds (*e.g.*, 0.95) will incorrectly remove valid masks that partially overlap with larger objects. They both lead to the reduced diversity of candidate masks for subsequent processing.

D. Additional qualitative results

In this section, we present visualizations of several key stages in our TF-SSD framework.

D.1. Visualization of SAM proposals

To demonstrate SAM's segmentation capability and the effectiveness of our QMG module, Fig. 1 visualizes the top-5 mask proposals ranked by our quality score $S_{n,t}^{ba}$ (Eq. 5 in the main paper). The masks with green borders represent our final predictions after the complete TF-SSD pipeline. It's observed that SAM can generate mask proposals that closely align with the GT, and our quality score $S_{n,t}^{ba}$ effectively ranks these masks at the top positions. It validates the effectiveness of SAM for mask generation and confirms that QMG can successfully identify promising candidate masks for subsequent processing.

Figure 1. Visualization of SAM mask proposals for "Accordion" and "bicycle" image groups. For each group, we show the original images (row 1), ground truth(GT) masks (row 2), and the top-5 mask proposals ranked by quality score $S^{ba}_{n,t}$ (rows 3-7). Green borders indicate the final predictions selected by our TF-SSD pipeline.

D.2. Visualization of DINO attention maps

To illustrate the saliency-aware capability of DINO's attention mechanism, Fig. 2 visualizes the attention maps $\mathcal{A}_n$ of DINO's CLS-token. The attention maps can highlight the salient object regions that closely align with the GT masks. These observations validate the effectiveness of leveraging DINO attention maps to select salient mask proposals from the refined candidate set derived from our QMG.

E. Limitations

While our TF-SSD framework achieves strong performance on most CoSOD benchmarks, it still exhibits limitations in detecting small co-salient objects. As illustrated in Fig. 3, the "moon" category contains predominantly small objects that occupy a relatively small portion of the image. Due to SAM's initial area-based filtering mechanism in our QMG, masks with small area ratios ($r^{area}_{n,t} < \tau_{area}$) are typically denoted as trivial objects that are filtered out. In addition, small objects also tend to receive lower quality scores $S^{ba}_{n,t}$ from DINO's attention maps, which leads to wrong identification for co-salient objects.

Our future work concentrates on addressing these bad cases in challenging scenarios. We will explore incorporating semantic cues into an adaptive threshold that is modulated according to contexts of the image. We argue that it can identify the small salient objects from the trivial ones.

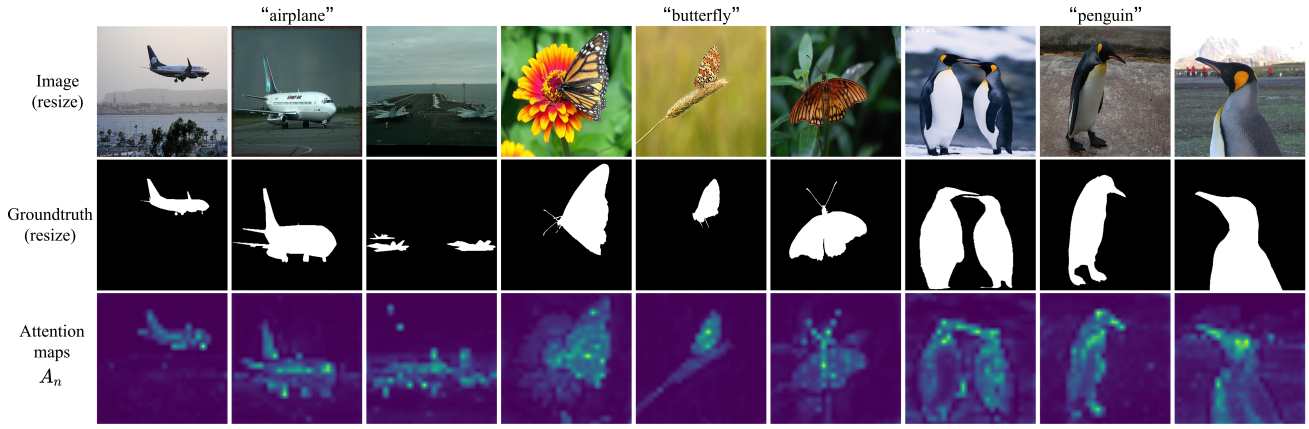

Figure 2. Visualization of DINO attention maps for three categories: ”airplane”, ”butterfly”, and ”penguin”. Row 1: Resized input images. Row 2: Resized GT masks. Row 3: DINO attention maps $\mathcal{A}_n$ that naturally highlight salient objects.

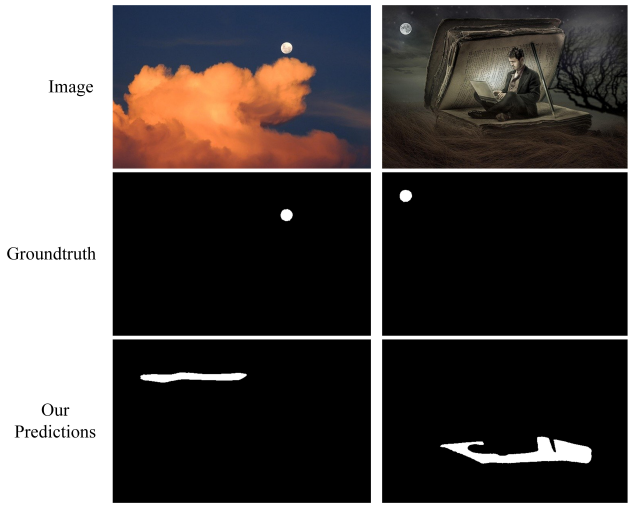

Figure 3. Limitation on small object detection. Visualizations of failed cases from the ”moon” category.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 1

[2] Deng-Ping Fan, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Huazhu Fu, and Ming-Ming Cheng. Taking a deeper look at co-salient object detection. In *CVPR*, pages 2919–2929, 2020. 1

[3] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 1

[4] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1