SDDF: Specificity-Driven Dynamic Focusing for Open-Vocabulary Camouflaged Object Detection

Supplementary Material

In this supplementary material, we provide additional implementation details and extensive experimental results to further elucidate our proposed method. The supplementary material is organized as follows:

- In Section 1, we present a comprehensive description of the dataset construction process, including details on data collection, annotation protocols, and novel class overlap analysis.
- In Section 2, we delineate the complete experimental pipeline, baselines' text pipeline, and provide a full listing of hyperparameter configurations and numerical stability designs.
- In Section 3, we provide additional quantitative results, including comparison with state-of-the-art COD methods and analysis across specific camouflage difficulty levels.
- In Section 4, we conduct extended ablation studies examining the influence of SVD configurations and the SFGLU convolutional layers.
- In Section 5, we perform detailed qualitative comparisons through visualization of detection bounding boxes and feature maps.

1. Dataset Information and Analysis

1.1. OVCOD-D Construction and Quality Assurance

The newly constructed OVCOD-D benchmark comprises 6,469 training samples and 3,957 test samples, into which 83 high-quality images of red imported fire ant (RIFA) nests have been incorporated. All RIFA nest images were collected in natural field environments, thereby addressing a gap in existing datasets concerning camouflaged insect-nest targets.

For a reproducible construction pipeline, we established a standardized three-stage pipeline: (a) Instance masks were converted to axis-aligned YOLO-format boxes. Coordinates were normalized with strict boundary clamping to ensure validity. (b) We filtered COD10K [3], retaining only the subset of camouflaged ones. (c) Label spaces were unified via case-insensitive matching and modifier removal, yielding an 87-category vocabulary.

For generating fine-grained sub-descriptions of target objects, we employed the following prompt:

> The image contains a {display_name}. **YOU MUST GENERATE EXACTLY THREE SEPARATE SENTENCES.**
> **FIRST SENTENCE:** Describe the background in **ONE VERY CONCISE English sentence** containing **AT LEAST THREE ADJECTIVES**, without mentioning the {display_name}.
> **SECOND SENTENCE:** Only describe the {display_name} in **ANOTHER VERY CONCISE English sentence** with **AT LEAST THREE PHYSICAL FEATURES** that distinguish it from its surroundings. **DO NOT MENTION** "background," "environment," "surroundings," or any context-related words.
> **THIRD SENTENCE:** Describe the physical relationship between the {display_name} and its background in **ONE VERY CONCISE English sentence**.

Subsequently, we obtain fine-grained object descriptions for each category and process them accordingly. Through a series of text preprocessing steps, including the removal of punctuation, conversion to lowercase, tokenization, and stop-word filtering. We extract semantically meaningful specificity-describing terms. For each species, we then construct a term-frequency statistics profile and retain the top 30 highest-frequency terms to form a species-specific vocabulary repository. We generate over 20 fine-grained sub-descriptions per class (4–8 valid tokens on average), with CLIP [10] native tokenization and the CLIP text encoder for text encoding. Each sub-description incorporates both the original scientific name of the species and its plural form. Furthermore, we select the top six pairs of high-frequency terms and assemble them into structured phrases, thereby ensuring that at least 20 semantically rich descriptive phrases are produced for each species.

For Text Quality Assurance, as shown in Table A1, five sub-descriptions are randomly sampled per category (or all if fewer than five are available). Positive and negative samples are identified based on three pre-defined metrics to calculate the precision rate. A category is deemed "qualified" if over 50% of its sampled sub-descriptions are identified as positive.

Table A1. Human evaluation of generated textual descriptions.

Metric	Accuracy	Relevance	Hallucination-free	Qualified category
Rate (%)	98.15	99.77	92.61	100.00

1.2. Novel Class Overlap with Pretraining Data

In OVCOD-D benchmark, 26 out of 47 novel classes have no overlap with pre-training data. Nevertheless, the domain

gap induced by high camouflage prevents model from relying on pre-learned features.

2. Implementation Setup

We conduct all experiments using PyTorch 2.1.2 with CUDA 11.8 on 8 RTX 4090 GPUs. In the data processing pipeline, we disable mosaic augmentation to better balance the difficulty of the camouflaged object detection task. We adopt an 80-epoch training schedule with a lr of 2×10^{-4}.

2.1. Text Pipeline and Fusion

For sub-description text fusion in the baseline methods, we adopt a simple yet effective strategy of summing the text feature vectors followed by normalization:

$$\mathbf{t}_c^{\text{fused}} = \frac{\sum_{k=1}^{K} \mathbf{t}_{c,k}}{\left\| \sum_{k=1}^{K} \mathbf{t}_{c,k} \right\|_2}. \tag{A1}$$

$t_{c,k}$ denotes the embedding vector of the k-th sub-description text in category c, and t_c^{fused} represents the fused sub-description text for category c. Empirically, this operation preserves sub-description information more effectively and yields improved detection performance.

YOLO-World [1] and DOSOD [6] leverage CLIP to extract sub-description vectors, then the fused category embeddings subsequently passed to a Vision-Language PAN and an MLP adapter, respectively. Similarly, YOLOE [12] utilizes MobileCLIP [11] to generate fused embeddings for its auxiliary network. In contrast, GLIP-T [8], GLIPv2-T [14], and Grounding DINO-T [9] employ BERT [2] for text extraction, integrating the fused results into the GLIP fusion encoder or the GroundingDINO neck. To ensure experimental consistency, all models are trained with an identical random seed and a learning rate (lr) of 2×10^{-4}.

In terms of training configuration, we set the `lr_mult` of the MLP-based text adapter in SDDF to 0.01, while assigning an `lr_mult` of 0.5 to both the backbone and PAN. We also fix the random seed to the same value across all experiments to ensure fair and reproducible comparisons.

2.2. Numerical Stability of SF-GLU

We ensure stable convergence via Epsilon Clamping ($\epsilon = 10^{-8}$), Argmax Gradient Detach, and Sigmoid-bounded gain $(1, 1+\alpha)$ with $\alpha = 1$.

3. Additional Quantitative Results

3.1. Comparison with State-of-the-Art COD Methods

Mainstream COD methods primarily focus on segmentation tasks. To bridge the gap between segmentation-based COD and bbox detection, we train State-of-the-Art (SOTA) models on the OVCOD-D under open-set setting. During inference, detection results are derived from the axis-aligned bounding boxes of predicted and ground-truth masks, with localization performance evaluated using the AP shown in Table A2.

Table A2. Comparison with SOTA COD methods on OVCOD-D.

Method	AP	AP_{50}	AP_{75}
SINet-V2 [4]	40.2	69.3	39.4
FSPNet [7]	47.9	76.2	49.4
CamoFormer [13]	55.6	80.2	59.0
HDPNet [5]	56.3	**81.5**	59.6
SDDF-L (Ours)	**56.4**	76.4	**60.7**

3.2. Analysis across Specific Camouflage Difficulty Levels

To balance the impact of target scale on visual similarity, we define a difficulty metric by weighting the target-image CLIP cosine similarity (0.6) and a scale coefficient $1 - \mathrm{A}_{\mathrm{bbox}}/\mathrm{A}_{\mathrm{img}}$ (0.4), $\mathrm{A}_{\mathrm{bbox}}$ and $\mathrm{A}_{\mathrm{img}}$ denote target and image areas, respectively. Table A3 presents results across three quantile-split levels: Mild, Moderate, and Severe.

Table A3. Multi-level camouflability performance of SDDF-S on OVCOD-D.

Method	AP_{mild}	AP_{mod}	AP_{severe}
SDDF-S	64.2	49.8	39.0

4. Extended Ablation Studies

4.1. Ablation Study on SVD Settings

SVD is performed on zero-centered matrices $\mathrm{T} \in \mathbb{R}^{K \times D}$. Retained PCs $\in [3, 10]$. The core function of SVD lies in its ability to eliminate semantic redundancy within sub-descriptions through orthogonal decomposition, thereby rendering different sub-descriptions more semantically independent in the latent space. To quantitatively evaluate this effect, we designed three comparative groups and conducted ablation studies atop the DOSOD baseline, systematically investigating the impact of varying the number of retained maximum principal components in SVD on model detection performance.

As show in Table A4, employing ten principal components significantly improves model performance, achieving a 2.0 AP increase over the baseline. Reducing the number of components to five leads to substantial semantic information loss, resulting in performance degradation compared to the ten-component configuration. Conversely, increasing the number to fifteen retains excessive redundancy among

sub-descriptions, also yielding diminished performance relative to the optimal ten-component setting. Thus, ten principal components strike the ideal balance between preserving semantic integrity and eliminating redundancy.

Table A4. Ablation study on different maximum number of components obtainable from SVD on the OVCOD-D dataset. All model variants adopt the fusion approach defined in Equation A1. The coverage auxiliary loss and SF-GLU are disabled.

SVD Configuration	AP	AP_{50}	AP_{75}
w/o SVD	44.8	67.4	46.8
w/ SVD (5)	46.5	**69.8**	48.3
w/ SVD (10)	**46.8**	69.5	**49.2**
w/ SVD (15)	46.1	69.2	48.4

4.2. Ablation Study on SF-GLU Convolutional Layers

SF-GLU uses a single-layer 1×1 convolution to expand text-visual similarity to the same dimension as visual features. Table A5 shows that increasing depth degrades performance, confirming that single-layer adaptation is effective.

Table A5. Ablation study on the number of convolutional layers.

Conv Layers	AP	AP_{50}	AP_{75}
1	**48.1**	**70.7**	**50.3**
2	47.1	69.6	49.8
4	46.0	68.7	48.6

5. Additional Qualitative Results

5.1. Qualitative Examples

As evidenced by the visualizations in the Figure A1, the proposed SDDF method significantly outperforms the other two comparative models in precisely localizing camouflaged objects, successfully detecting regions that exhibit high visual similarity with their surrounding environments—regions that often prove challenging for competing approaches.

Furthermore, when dealing with small and medium sized targets (e.g., Owl and Tiger in the fourth and third columns of Figure A1, respectively), SDDF exhibits markedly stronger robustness to interference, effectively avoiding the common pitfall of misclassifying background elements as foreground objects. In complex scenes containing multiple targets (e.g., Bird in the sixth column of Figure A1), SDDF also demonstrates superior localization accuracy and clearer separation among instances, yielding more reliable and precise bounding-box predictions overall.

5.2. Attention Improvement of SF-GLU

To investigate the effectiveness of the proposed SF-GLU module in promoting feature focus on camouflaged objects, we visualize heatmap representations of the response differences in feature maps before and after this module.

As demonstrated in the Figure A2, regions exhibiting high activation (indicated in red and yellow) are precisely concentrated on the camouflaged targets, whereas background regions (marked in blue) show significantly suppressed responses. This highly targeted activation pattern clearly illustrates that, by leveraging specificity-oriented sub-descriptions, the SF-GLU module successfully directs the model's attention toward regions that exhibit strong semantic alignment with fine-grained descriptors, thereby enabling effective discrimination between camouflaged objects and complex, cluttered backgrounds.

For instance, in the "Mouse" example, the high-response areas are tightly localized to critical parts of the animal, such as the head and body, further underscoring the module's capability to achieve precise and semantically meaningful feature focusing even under severe camouflage conditions.

References

[1] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024. 2

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 2

[3] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2777–2787, 2020. 1

[4] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6024–6042, 2022. 2

[5] Jinpeng He, Biyuan Liu, and Huaixin Chen. Hdpnet: Hourglass vision transformer with dual-path feature pyramid for camouflaged object detection. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8638–8647. IEEE, 2025. 2

[6] Yonghao He, Hu Su, Haiyong Yu, Cong Yang, Wei Sui, Cong Wang, and Song Liu. A light-weight framework for open-set object detection with decoupled feature alignment in joint space. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13590–13596. IEEE, 2025. 2

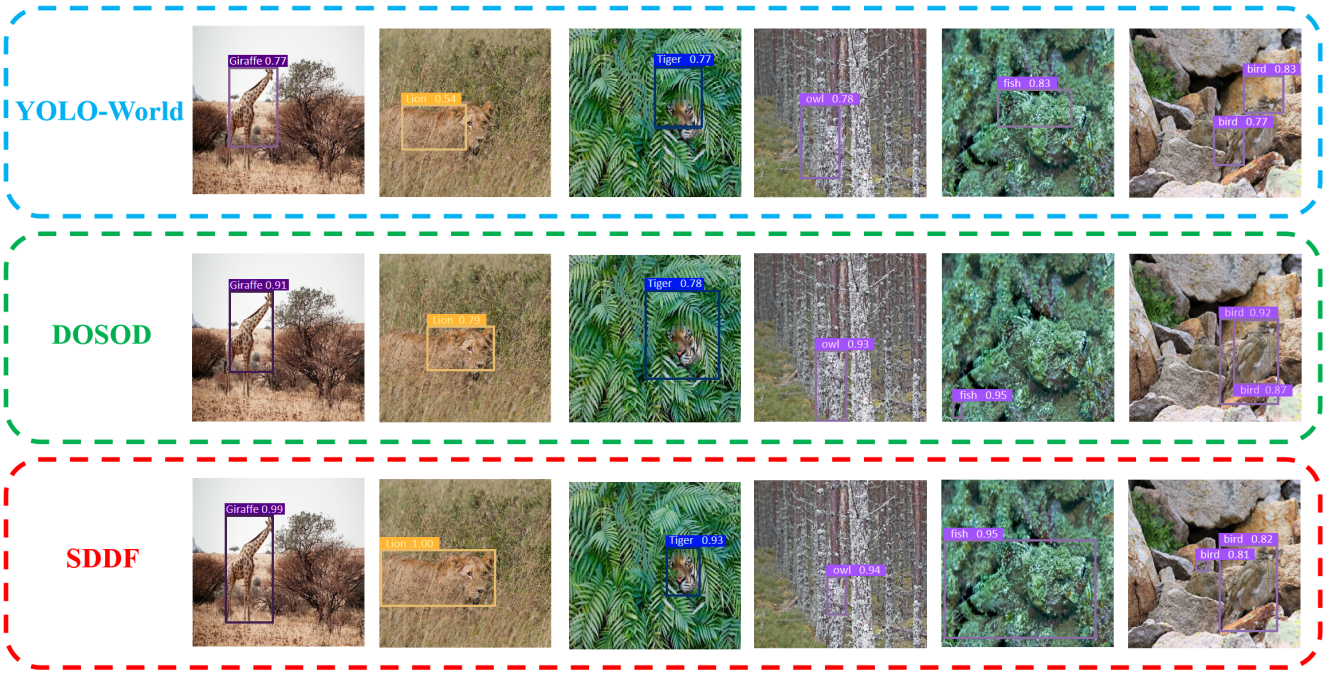

Figure A1. Quantitative comparison is conducted via visualization of the detection bounding boxes.

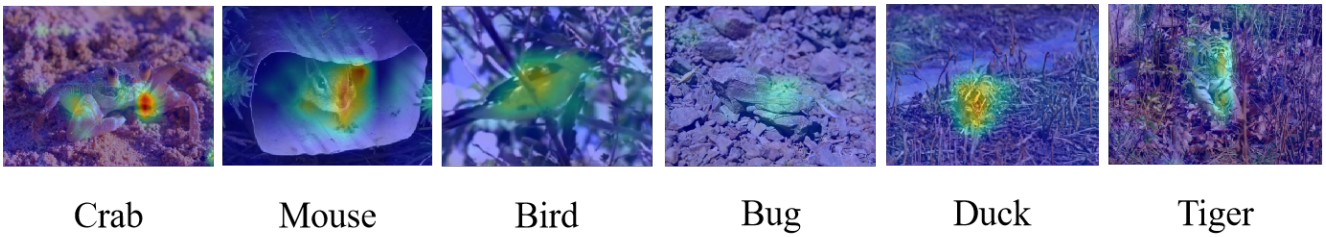

Figure A2. Quantitative comparisons are further conducted by visualizing heatmap representations of the response differences in feature maps with and without the proposed SF-GLU. Specifically, these feature maps are extracted from the PAN outputs.

[7] Zhou Huang, Hang Dai, Tian-Zhu Xiang, Shuo Wang, Huai-Xin Chen, Jie Qin, and Huan Xiong. Feature shrinkage pyramid for camouflaged object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5557–5566, 2023. 2

[8] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2

[9] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *Proceedings of the European Conference on Computer Vision*, pages 38–55. Springer, 2024. 2

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1

[11] Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Raviteja Vemulapalli, and Oncel Tuzel. Mobileclip: Fast image-text models through multi-modal reinforced training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15963–15974, 2024. 2

[12] Ao Wang, Lihao Liu, Hui Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yoloe: Real-time seeing anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24591–24602, 2025. 2

[13] Bowen Yin, Xuying Zhang, Deng-Ping Fan, Shaohui Jiao, Ming-Ming Cheng, Luc Van Gool, and Qibin Hou. Camoformer: Masked separable attention for camouflaged object

detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[14] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. Glipv2: Unifying localization and vision-language understanding. In *Advances in Neural Information Processing Systems*, 2022. 2