Concealed Object Detection

— *Supplementary Materials*

Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, *Senior Member, IEEE*, and Ling Shao, *Fellow, IEEE*

Abstract—In this document, we provide additional materials details on, e.g., the dataset and model, metrics, quantitative results, and qualitative results, to enable a better understanding of the new concealed object detection (**COD**) task.

- **Dataset.** We present more visualizations, information on the annotation process, and statistics for our *COD10K* in Section 1.
- **Metrics.** For each metric adopted in the *manuscript*, e.g., S_α [1], E_ϕ [2], F_β^w [3], M [4], we describe the details in Section 2.
- **Results.** We provide more qualitative results in Section 3.

Index Terms—Concealed Object Detection, Camouflaged Object Detection, COD, Dataset, Benchmark.

1 Dataset

- **Taxonomy Statistics.** Our *COD10K* contains 10,000 images (5,066 concealed, 3,000 background, 1,934 non-concealed), divided into 10 super-classes, and 78 sub-classes (69 concealed, 9 non-concealed) which are collected from diverse real-world scenes (Fig. 1). The collection details can be found in our *manuscript*.
- **Attributes & Quality.** In Table 1, we summarize the attributes assigned to each image. Please refer to the *manuscript* for a detailed description of the attributes. As shown in Fig. 2, we also introduce strict data selection criteria.
- **Overall Dataset Visualization.** To reveal the holistic visual perception for each dataset, we reduce the number of dimensions of the data points to focus on only the most relevant attribute, or to cluster the color distribution. We utilize the pre-trained VGG-16 [6] model (without the top FC-layers) to map multiple images into a two-dimensional square grid within different datasets using the *t-SNE* technique [7]. For the visualizations, see Fig. 3. CAMO-COCO [8] is a medium-size dataset to fill the gap in public databases in the field. As shown in Fig. 3 (a) & (b), for other datasets, the color feature distributions are generally concentrated on the whole concealed area; in contrast, the color features of the concealed parts in CAMO-COCO are distributed in a random way. As shown in Fig. 3 (c) & (d), green, brown, white, and blue areas represent vegetation, soil, sky, and ocean, respectively. Compared with the biases of previous datasets, our *COD10K* shows better concentrated color feature distributions in both concealed and non-concealed areas.
- **Instance-Level Segmentation Visualization.** Additionally, we also provide the representative shapes of concealed animals for each sub-class of our *COD10K* in the Fig. 4, Fig. 5, Fig. 6, Fig. 7, and Fig. 8.
- **Non-Concealed Image Collection.** As noted in [9], object detection datasets always contain the object that the model needs to detect. This is a form of *data selection bias* [9]. Similarly, most previous concealed object detection datasets assume that there exists at least one concealed object in each image. This assumption does not always hold, however, as some scene do not contain any concealed objects. To avoid this issue, we collect negative samples (non-concealed images) in *COD10K*. Most non-concealed images, which have at least one normal or salient animal, are obtained from *Flickr* using the following keywords, with a valid license: *'amphibians', 'aquatic', 'flying animals', 'terrestial',* and *'mammal'*, *etc*. In addition, we use another set of keywords to obtain the background samples, which have no animals: *'coral', 'grass', 'rainforest', 'seabed', 'sandbeach', 'tree-branch', 'sky', 'vegetation',*, *'rock'*, *etc*. To provide a diverse distribution for negative samples, we further choose some images using keywords such as 'daily life', 'outdoor', and 'indoor') with salient objects from the SOC dataset [9]. For detailed statistics see Fig. 1 (e) & (f).

2 Metrics

To provide a comprehensive evaluation, we adopt two universally used metrics and two recently released metrics that have demonstrated more reliable evaluation results.

- **Mean Absolute Error (M) [4].** Generally, we evaluate the difference between the camouflage prediction (y^{Cam}) and the binary ground truth (y^{GT}), where all individuals have the same weight. In other words, all parts in the process of MAE evaluation are normalized within the range [0,1]. The MAE [4] score is defined as:

$$M = \frac{1}{W \times H} \| y^{Cam} - y^{GT} \|, \tag{1}$$

- *Deng-Ping Fan is with the CS, Nankai University, Tianjin, China, and also with the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. (E-mail: dengpfan@gmail.com)*
- *Ge-Peng Ji and Ling Shao are with the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. (E-mail: gepeng.ji@inceptioniai.org; ling.shao@inceptioniai.org)*
- *Ming-Ming Cheng is with the CS, Nankai University, Tianjin, China. (E-mail: cmm@nankai.edu.cn)*
- *A preliminary version of this work has appeared in CVPR 2020 [5].*
- *The major part of this work was done in Nankai University.*
- *Ming-Ming Cheng is the corresponding author.*

where W and H are the width and height of the image.

- **Weighted F-measure (F_β^w).** The F_β^w was proposed by Margolin *et al.* [3] to mend the existing flaws of the F-measure [3] and provide a more precise quantitative evaluation. The formulation is as follows ($\beta^2 = 1$):

$$F_\beta^w = \frac{\left(1+\beta^2\right)\text{Precision}^\omega \times \text{Recall}^\omega}{\beta^2 \times \text{Precision}^\omega + \text{Recall}^\omega}. \tag{2}$$

- **Structure-measure (S_α).** Pixel-wise evaluation metrics (*e.g.*, MAE, IoU) fail to comprehensively distinguish where the error occurs. To avoid an unsatisfactory evaluation, a novel and useful metric, which considers both the region-part and object-part, was proposed by [1]. The formulation of this structural similarity is as follows:

$$S_\alpha = \alpha * S_o + (1-\alpha) * S_r, \tag{3}$$

where α is a balance coefficient and empirically set to 0.5 as default in our experiments.

- **Enhance-measure (E_ϕ).** Previous studies on cognitive vision [10], [11] have shown that the human visual system is highly sensitive to structures (*e.g.*, global information, local details) in real scenes. Consequently, Fan *et al.* [2] took local (pixel-level) matching details and global (image-level) information into account, simultaneously, when evaluating the performance of object segmentation:

$$E_\phi = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \phi_{FM}(x,y), \tag{4}$$

where ϕ is defined as the enhanced alignment matrix.

3 Qualitative Results

In this section, we show more detection results for various challenging concealed objects, such as *spider* in Fig. 9, *moth* in Fig. 10, *sea horse* in Fig. 11, and *toad* in Fig. 12. As can be seen, for all these examples, *SINet* [5] achieves the best results, demonstrating the robustness of our framework.

References

[1] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A New Way to Evaluate Foreground Maps," in *Int. Conf. Comput. Vis.*, 2017, pp. 4548–4557. 1, 2

[2] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment Measure for Binary Foreground Map Evaluation," in *Int. Joint Conf. Artif. Intell.*, 2018, pp. 698–704. 1, 2

[3] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 248–255. 1, 2

[4] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 733–740. 1

[5] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 2777–2787. 1, 2

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learn. Represent.*, 2015. 1

[7] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008. 1, 5

[8] T.-N. Le, T. V. Nguyen, Z. Nie, M.-T. Tran, and A. Sugimoto, "Anabranch network for camouflaged object segmentation," *Comput. Vis. Image Underst.*, vol. 184, pp. 45–56, 2019. 1, 5

[9] D.-P. Fan, J.-J. Liu, S.-H. Gao, Q. Hou, A. Borji, and M.-M. Cheng, "Salient objects in clutter: Bringing salient object detection to the foreground," in *Eur. Conf. Comput. Vis.*, 2018, pp. 1597–1604. 1

[10] G. F. Luger, P. Johnson, C. Stern, J. E. Newman, and R. Yeo, *Cognitive science: The science of intelligent systems.* Academic Press, 1994. 2

[11] P. N. Johnson-Laird, *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press, 1983, no. 6. 2

TABLE 1
Attributes statistics for each sub-class in our *COD10K* dataset. From left to right: **MO** (multiple objects), **BO** (big object), **SO** (small object), **OV** (out-of-view), **OC** (ccclusions), **SC** (shape complexity), and **IB** (indefinable boundaries). Please refer to our *manuscript* for more detailed descriptions of attributes.

Sub-class	MO	BO	SO	OV	OC	SC	IB
Flying/Mockingbird			✓	✓	✓	✓	✓
Flying/Moth	✓	✓	✓	✓	✓		✓
Flying/Owl	✓	✓	✓	✓	✓	✓	✓
Flying/Owlfly		✓	✓	✓	✓	✓	✓
Other/Other	✓	✓	✓	✓	✓	✓	✓
Terrestrial/Ant	✓	✓	✓	✓	✓	✓	✓
Terrestrial/Bug	✓		✓		✓	✓	✓
Terrestrial/Cat	✓	✓	✓	✓	✓	✓	✓
Terrestrial/Caterpillar	✓	✓	✓	✓	✓	✓	✓
Terrestrial/Centipede	✓		✓		✓	✓	✓
Terrestrial/Chameleon	✓	✓	✓	✓	✓	✓	✓
Terrestrial/Cheetah	✓	✓	✓	✓	✓	✓	✓
Terrestrial/Deer	✓		✓	✓	✓	✓	✓
Terrestrial/Dog	✓		✓	✓	✓	✓	✓
Terrestrial/Duck	✓		✓	✓	✓		✓
Terrestrial/Gecko	✓		✓	✓	✓	✓	✓
Terrestrial/Giraffe	✓		✓	✓	✓	✓	✓
Terrestrial/Grouse		✓	✓		✓		✓
Terrestrial/Human	✓	✓	✓	✓	✓	✓	✓
Terrestrial/Kangaroo	✓		✓	✓	✓	✓	✓
Terrestrial/Leopard	✓		✓	✓	✓	✓	✓
Terrestrial/Lion	✓		✓		✓	✓	✓
Terrestrial/Lizard	✓	✓	✓	✓	✓	✓	✓
Terrestrial/Monkey	✓		✓	✓	✓		✓
Terrestrial/Rabbit	✓	✓	✓	✓	✓	✓	✓
Terrestrial/Reccoon			✓		✓		✓
Terrestrial/Sciuridae			✓	✓	✓	✓	✓
Terrestrial/Sheep	✓	✓	✓	✓	✓		✓
Terrestrial/Snake	✓	✓	✓	✓	✓	✓	✓
Terrestrial/Spider	✓		✓	✓	✓	✓	✓
Terrestrial/StickInsect	✓		✓	✓	✓	✓	✓
Terrestrial/Tiger	✓	✓	✓	✓	✓	✓	✓
Terrestrial/Wolf	✓		✓		✓	✓	✓
Terrestrial/Worm	✓		✓	✓	✓		✓
Amphibian/Frog	✓		✓		✓	✓	✓
Amphibian/Toad	✓	✓	✓	✓	✓	✓	✓
Aquatic/BatFish		✓	✓		✓	✓	✓
Aquatic/ClownFish	✓		✓		✓		✓
Aquatic/Crab	✓	✓	✓	✓	✓	✓	✓
Aquatic/Crocodile	✓	✓	✓	✓	✓		
Aquatic/CrocodileFish		✓		✓			✓
Aquatic/Fish	✓	✓	✓	✓	✓	✓	✓
Aquatic/Flounder	✓	✓		✓	✓	✓	✓
Aquatic/FrogFish		✓					✓
Aquatic/GhostPipefish	✓		✓	✓	✓	✓	✓
Aquatic/LeafySeaDragon	✓	✓		✓	✓	✓	✓
Aquatic/Octopus		✓	✓	✓	✓	✓	✓
Aquatic/Pagurian			✓		✓	✓	
Aquatic/Pipefish	✓	✓	✓	✓	✓	✓	✓
Aquatic/ScorpionFish	✓	✓	✓	✓	✓	✓	✓
Aquatic/SeaHorse	✓	✓	✓	✓	✓	✓	✓
Aquatic/Shrimp	✓		✓	✓	✓	✓	✓

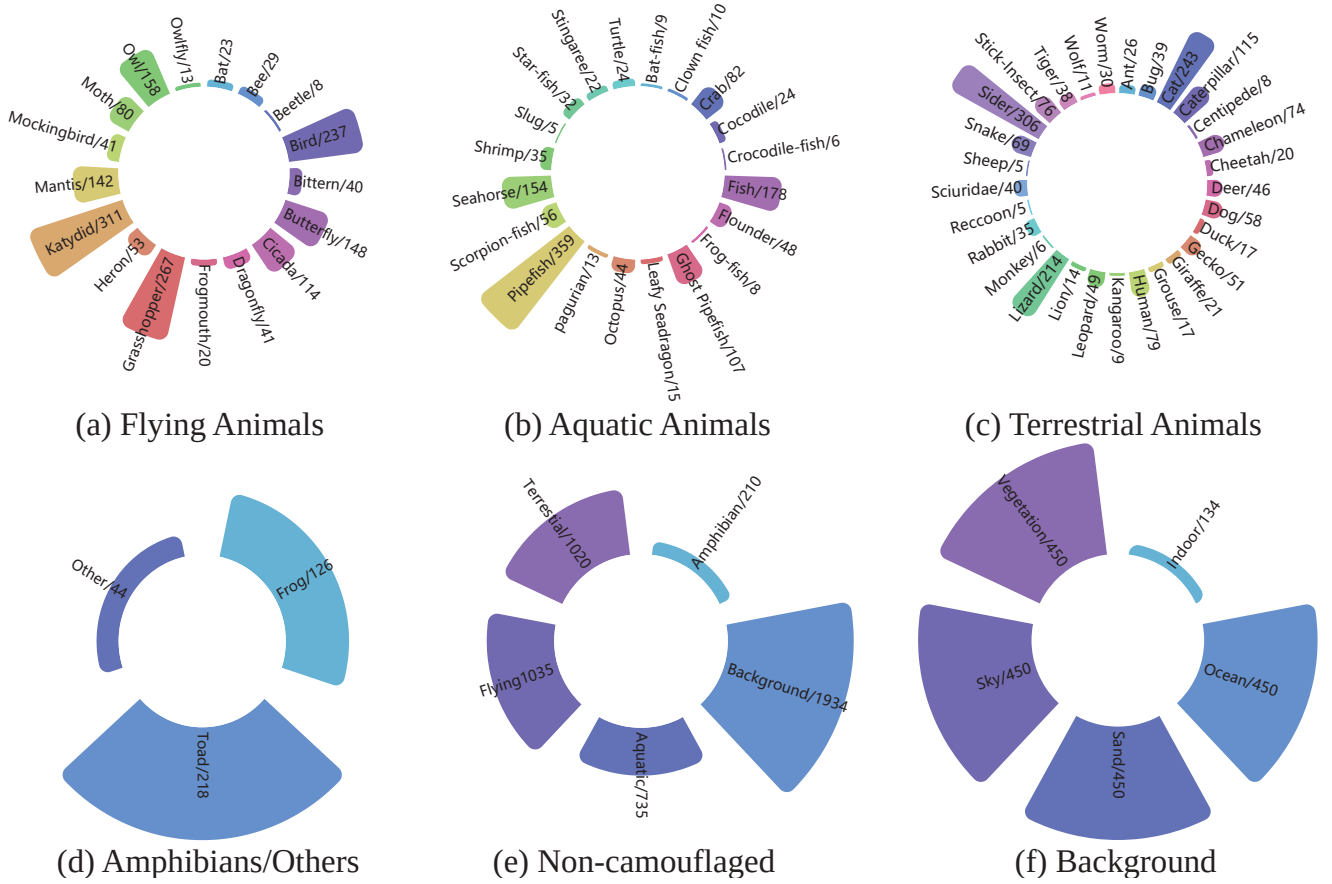

Fig. 1. **Taxonomy and statistics of super-/sub-classes in our *COD10K*.** We divide the 10K images collected into two parts: a concealed subset (five super-classes, as show in (a), (b), (c), and (d)) and a non-concealed subset (five super-classes, as show in (e)). Besides, we divide the *Background* super-class into five sub-classes, following the same ratio as the concealed subset.

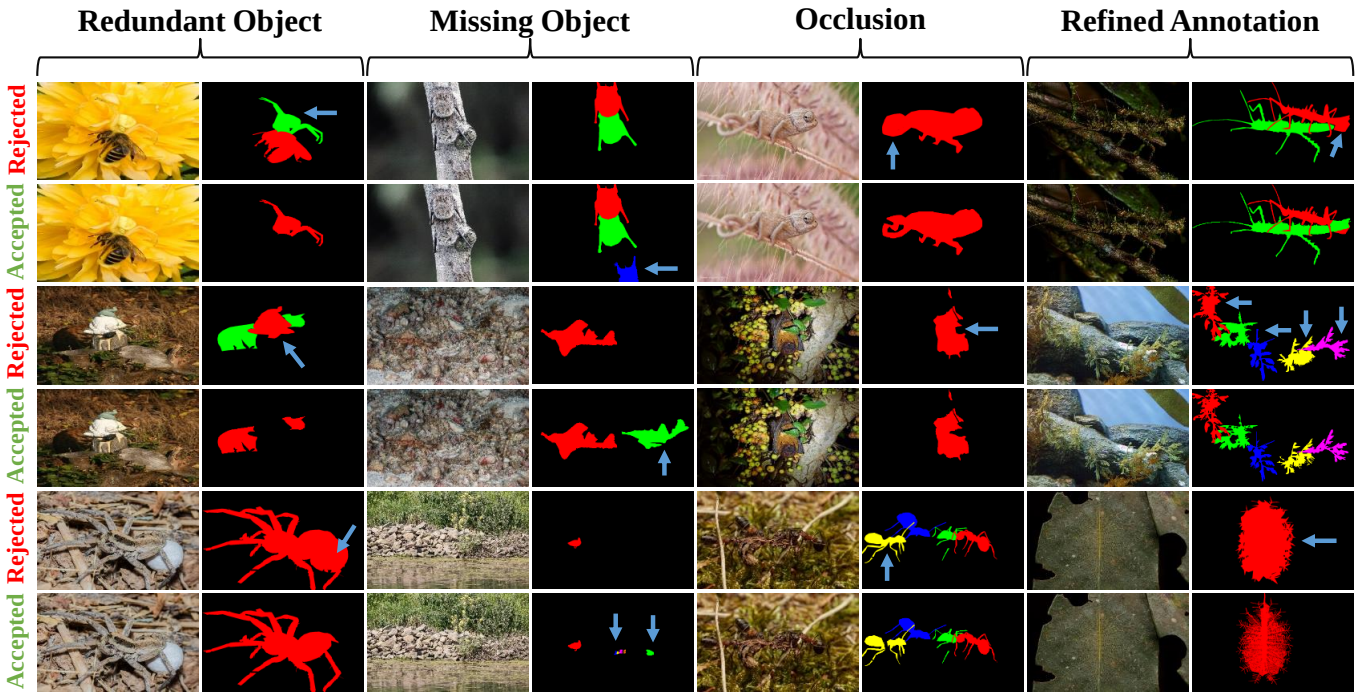

Fig. 2. **Regularized quality control during our labeling reverification stage.** We strictly adhere to the four major criteria (*i.e.*, redundant object, missing object, occlusion, and refined annotation.) for rejection or acceptance, nearing the ceiling of annotation accuracy.

(a) CAMO-CAM [8]

(b) CAMO-NonCAM [8]

(c) COD10K-CAM (OUR)

(d) COD10K-NonCAM (OUR)

Fig. 3. **Visualizations of the two large-scale COD datasets generated by t-SNE[7].** Please refer to Section 1 for detailed discussions. Zoom-in for a clearer version of this visualization.

Fig. 4. **Extraction of individual samples from 29 sub-classes of our *COD10K* (1/5)–Terrestrial animals.**

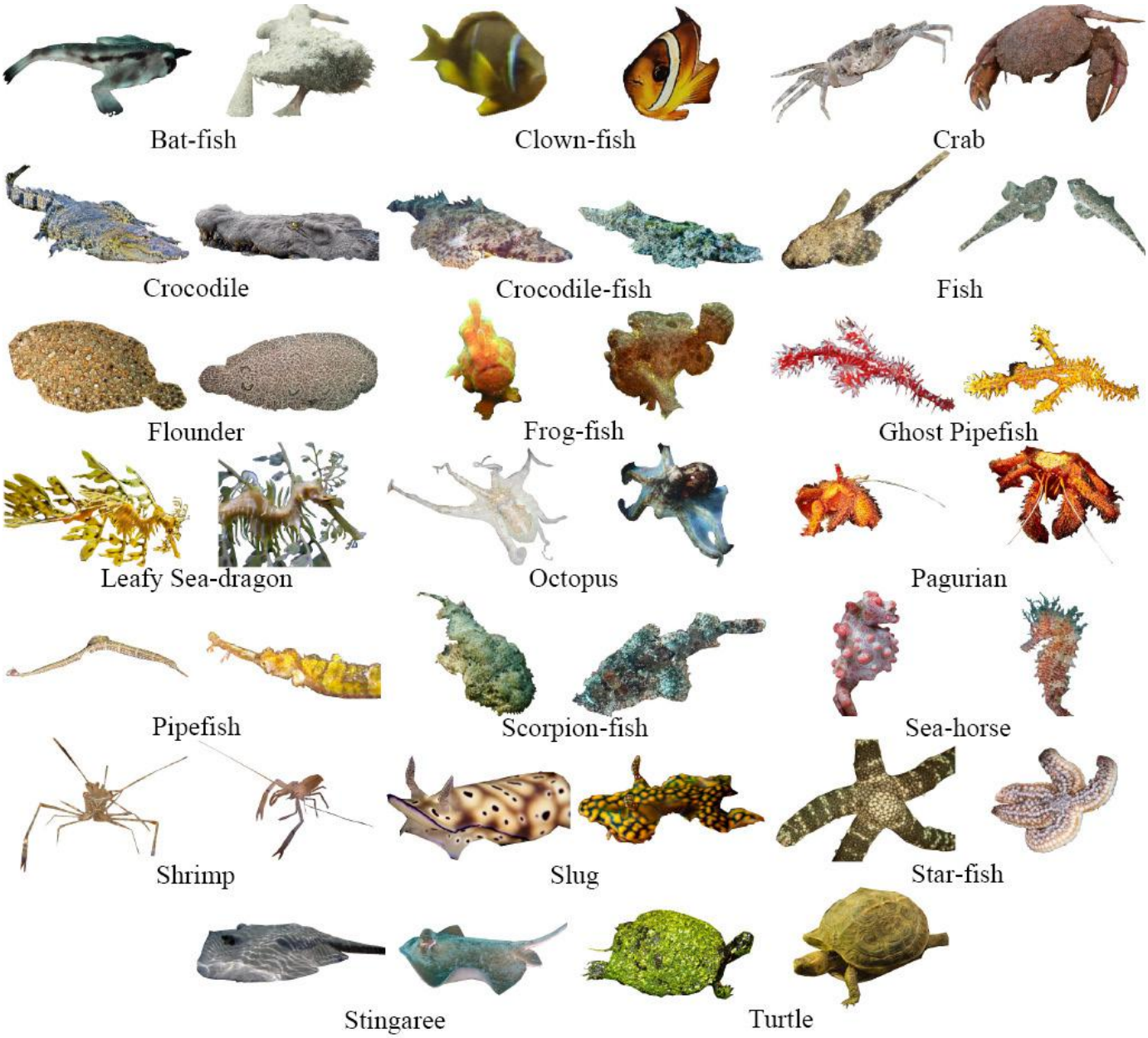

Fig. 5. **Extraction of individual samples from 20 sub-classes of our *COD10K* (2/5)–Aquatic animals.**

Fig. 6. **Extraction of individual samples from two sub-classes of our *COD10K* (3/5)–Amphibians.**

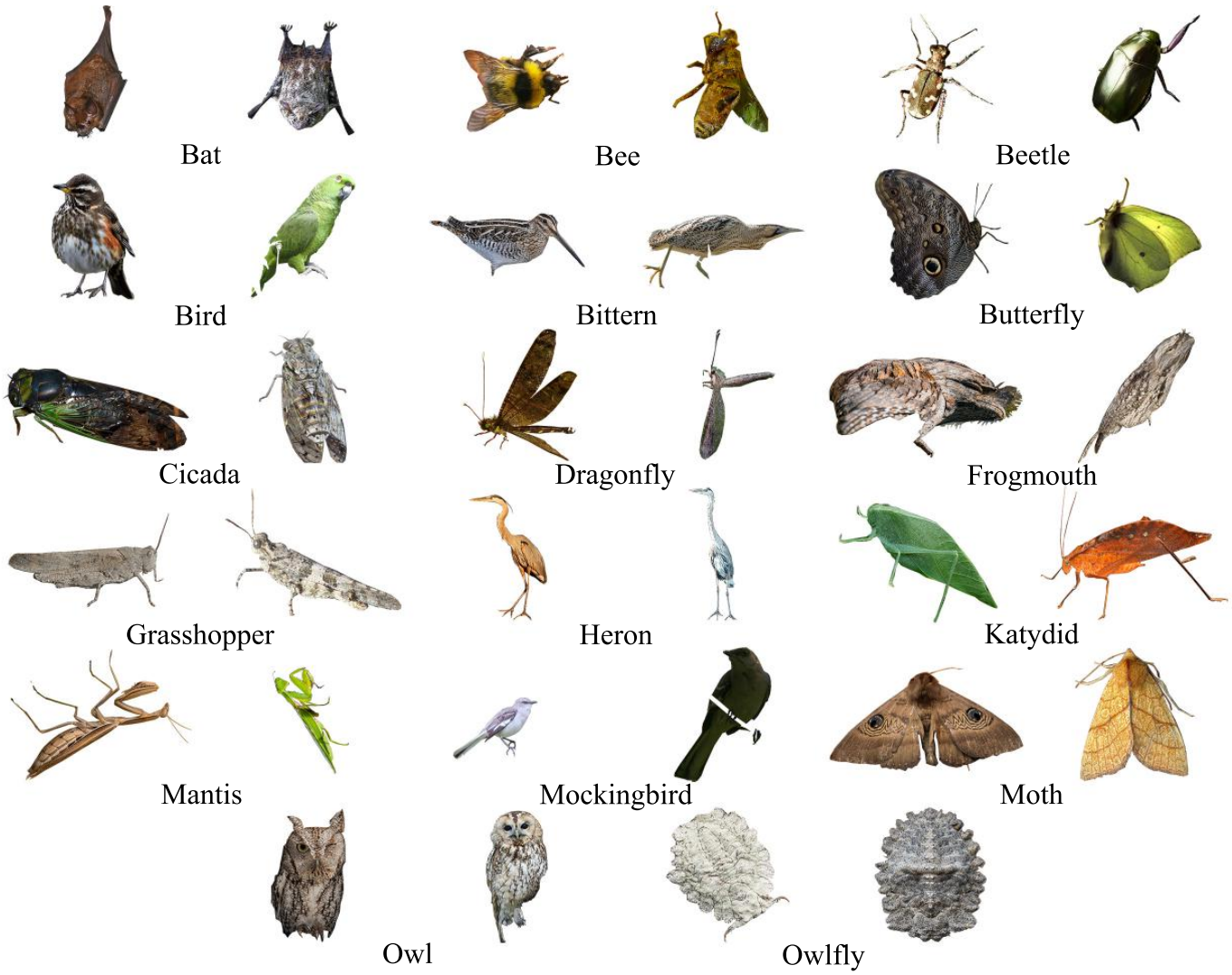

Fig. 7. **Extraction of individual samples from 17 sub-classes of our *COD10K* (4/5)–Flying animals.**

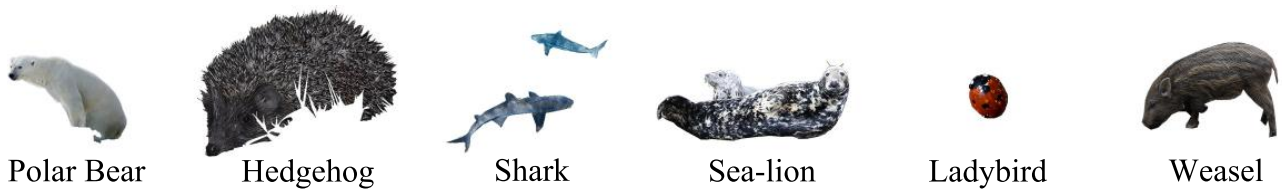

Fig. 8. **Extraction of individual samples from one sub-classes of our *COD10K* (5/5)–Other animals.** Note that we merge 21 classes (*e.g.*, bear, elephant, fox, mouse, shark, sea lion, *etc.*) into a single sub-class because they do not have sufficient images (less than 5).

Fig. 9. **Additional qualitative results of *SINet* and 12 baseline models on *COD10K* (1/4)-Terrestrial animals.** We evaluate our framework on different animals, *e.g.*, *spider*, *lizard*, and *gecko*. Our model is able to capture the concealed objects under different circumstances, *e.g.*, similar color (1^{st} and 4^{th} row) and low illumination (7^{th} row), and produce results similar to the GTs.

Fig. 10. **Additional qualitative results of *SINet* and 12 baseline models on *COD10K* (2/4)-Flying animals.** Note that the texture of the moth in the first row is very similar to its surroundings, making the existing generic object detection or salient object detection models fail. In contrast, our model based on the visual perception mechanism, which is implemented by the search and identification module, is able to produce visually appealing results.

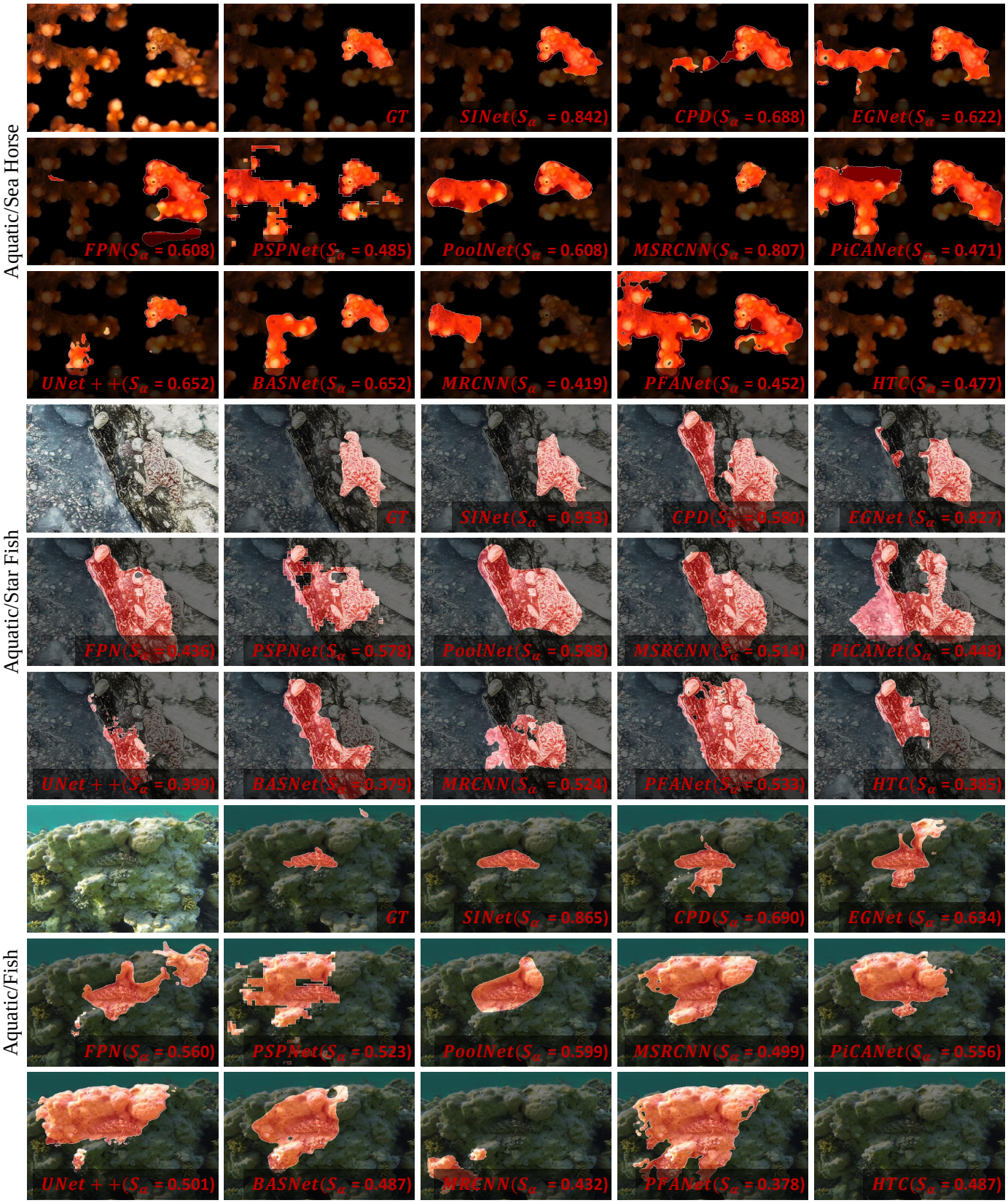

Fig. 11. **Additional qualitative results of *SINet* and 12 baseline models on *COD10K* (3/4)-Aquatic animals.** As can be seen here, the results of our model on sub-classes, *e.g.*, *sea horse*, *star fish*, and *fish*, are very close to the GTs. In contrast, other competitors generate relatively inaccurate for these challenging animals.

Fig. 12. **Additional qualitative results of *SINet* and 12 baseline models on *COD10K* (4/4)-Amphibian & Other animals.** We can see that our general framework *SINet* achieves the best results on sub-classes, *e.g.*, *other* and *toad*. Thanks to the search and identification strategy, our model can infer the real concealed object with fine details. In contrast, the compared models either miss the fine details of objects or only locate the concealed objects.