# CAMotion: A High-Quality Dataset for Camouflaged Motion Object Detection in the Wild

**Siyuan Yao**[1]*, **Hao Sun**[1]*, **Hai Long**[2], **Ruiqi Yu**[2],
**Jiehong Li**[2], **Xiwei Jiang**[1], **Yanzhao Su**[2], **Wenqi Ren**[2]†, **Xiaochun Cao**
[1] Beijing University Of Posts and Telecommunications; [2] Shenzhen Campus, Sun Yat-sen University

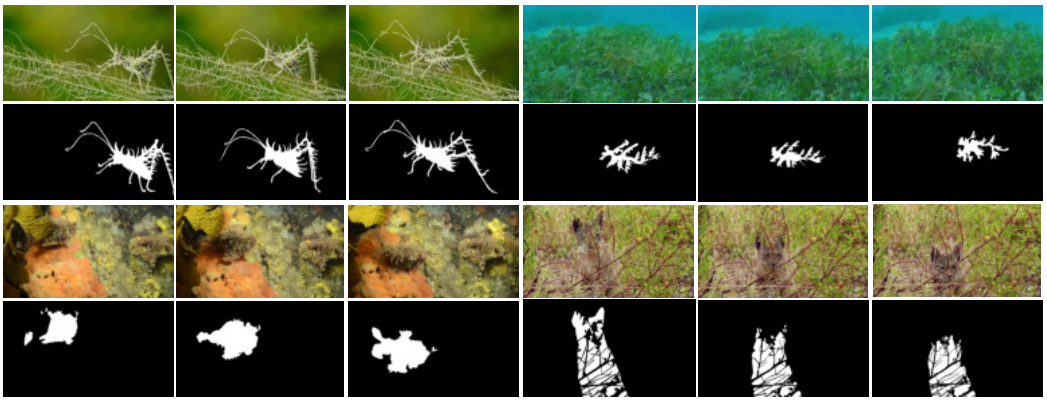


Figure 1: Examples of our CAMotion dataset with corresponding pixel-level annotations. The first and third rows contain original images; the second and last rows contain corresponding pixel-wise ground truth annotations.

## Abstract

Discovering camouflaged objects is a challenging task in computer vision due to the high similarity between camouflaged objects and their surroundings. While the problem of camouflaged object detection over sequential video frames has received increasing attention, the scale and diversity of existing video camouflaged object detection (VCOD) datasets are greatly limited, which hinders the deeper analysis and broader evaluation of recent deep learning-based algorithms with data-hungry training strategy. To break this bottleneck, in this paper, we construct CAMotion, a high-quality dataset covers a wide range of species for camouflaged motion object detection in the wild. CAMotion comprises various sequences with multiple challenging attributes such as uncertain edge, occlusion, motion blur and shape complexity, etc. The sequence annotation details and statistical distribution are presented from various perspectives, allowing CAMotion to provide in-depth analyses on the camouflaged object's motion characteristics in different challenging scenarios. Additionally, we conduct a comprehensive evaluation of existing SOTA models on CAMotion dataset, and discuss the major challenges in VCOD tasks. We will release our dataset soon, hoping our dataset leads to further advancements in the research community.

*Co-first authors
†Corresponding author

Preprint. Under review.

# 1 Introduction

Camouflage is a widespread defensive behavior in natural scenarios that disguises the appearance to blend with the surroundings for deception and paralysis purposes. To distinguish the camouflaged objects in various challenging environments, the Camouflaged Object Detection (COD) has become a prevalent topic in the computer vision community. Different from traditional dense prediction tasks, where objects typically exhibit distinct boundaries, camouflaged objects often share similar colors and textures with the background, making the objects difficult to perceive. This task becomes even more challenging for video sequences due to the dynamic appearance changes of objects and background over time.

With the advent of deep learning-based techniques for camouflaged object detection in recent years, various datasets have been established for comprehensive analyses. The prevailing datasets, such as CAMO [1], COD10K [2] and NC4K [3], have been proposed and have received extensive studies. Concurrently, some researchers also explore to discover the moving camouflaged objects in consecutive video sequences, the representative works like CAD [4] and MoCA-Mask [5] datasets provide pixel-wise annotations and construct comprehensive VCOD benchmark to facilitate research. However, several issues are remained in the assessment of the VCOD algorithms. First, although deep learning-based models have dominated the research field, the scale of existing VCOD dataset is greatly limited, which hinders to investigate the potential of recent deep learning-based algorithms with data-hungry training strategy. Second, since VCOD requires conducting pixel-wise camouflaged objects prediction in unconstraint environment, the data diversity is thus of vital importance for the fair evaluation. Nevertheless, existing VCOD datasets suffer from the small number of scenes and categories. As a result, the generalization capabilities of existing VCOD algorithms are obscure. In addition, as numerous attributes e.g., complex shape and occlusion) may be involved in the video frames, the effectiveness of existing camouflaged object detectors in these challenging attributes is still unclear.

Table 1: Statistics of camouflage datasets. * indicates that the #Cat. is not reported in the original paper and is estimated by us.

| Dataset | Year | Publication | Type | #Img. | #Ann. Img. | #Cat. | Bbox. GT | Mask GT |
|---|---|---|---|---|---|---|---|---|
| CAD[4] | 2016 | ECCV | Video | 839 | 181 | 6 | ✗ | ✓ |
| CHAMELEON[6] | 2018 | - | Image | 76 | 76 | 27* | ✗ | ✓ |
| CAMO[1] | 2019 | CVIU | Image | 2,500 | 1,250 | 97* | ✗ | ✓ |
| COD10K[2] | 2020 | CVPR | Image | 10,000 | 5,066 | 69 | ✓ | ✓ |
| MoCA[7] | 2020 | ACCV | Video | 37,250 | 7,617 | 67 | ✓ | ✗ |
| CAMO++[8] | 2021 | TIP | Image | 5,500 | 5,500 | 93 | ✓ | ✓ |
| NC4K[3] | 2021 | CVPR | Image | 4,121 | 4,121 | 85* | ✗ | ✓ |
| MoCA-Mask[5] | 2022 | CVPR | Video | 22,939 | 4,691 | 44 | ✗ | ✓ |
| **CAMotion** | 2025 | - | Video | 156,218 | 30,007 | 151 | ✓ | ✓ |

To address these issues, in this paper, we construct CAMotion, a high-quality dataset covers a wide range of species for camouflaged motion object detection in the wild. CAMotion consists of more than 150K video frames categorized into 151 species, wherein 30,007 frames has been carefully annotated. Each video frame in this dataset is annotated with multiple challenging attributes such as uncertain edge, occlusion, and shape complexity, etc. The high-quality annotations and category diversity make CAMotion suitable for presenting comprehensive analyses of different camouflaged manners in complex scenarios. As shown in Table 1, our CAMotion is the largest VCOD dataset that is significantly larger than existing COD and VCOD datasets. Moreover, we further conduct a comprehensive evaluation on CAMotion dataset using the most recent SOTA models and investigate the major challenges in VCOD tasks. To the best of our knowledge, this is the largest VCOD dataset with diverse categories in the research community.

In conclusion, the contributions of this paper are summarized below:

• We construct a high-quality VCOD dataset CAMotion, which comprises various sequences with multiple challenging attributes and a wide range of species for camouflaged motion object detection in the wild.

• We present annotation details and statistical distribution of the collected dataset from various perspectives, allowing CAMotion to provide in-depth analyses on the camouflaged object's motion characteristics in different challenging scenarios.

• We conduct a comprehensive evaluation on CAMotion dataset using recent SOTA COD/VCOD models, and discuss the major challenges in the VCOD task.

## 2 Related Work

**Camouflaged object detection.** Camouflaged object detection (COD) aims to discover camouflaged objects from a single RGB image. Inspired by the concealment strategy in biology, some work [2, 9, 10] simulate the behavior process of predators to search and locate camouflaged objects. For example, SINet [2] utilizes a searching module and an identification module to locate and detect objects with similar background distractions. ZoomNet [11] imitates human vision by zooming in and out the imperceptible camouflaged objects with mixed scales. Another strategy is the multi-task joint learning-based approach [12, 13, 14, 15, 16, 17, 18, 19]. These methods typically utilize auxiliary tasks to segment the target object. For instance, in [14, 15, 17], the boundary-aware priors are introduced to extract features that highlight the structural details of the object. Additionally, PUENet [20] respectively models epistemic uncertainty and aleatoric uncertainty for effective segmentation with less model and data bias. Other researchers also [21] introduce visual cues in the frequency domain to capture the subtle details of camouflaged objects from the background.

**Video camouflaged object detection.** In contrast to static COD tasks, video camouflaged object detection (VCOD) leverages both appearance and temporal information between video frames to break camouflage. Early works [7, 22, 23, 24] handle VCOD as a motion segmentation problem which utilizes the predicted optical flow to explicitly model the spatio-temporal correlation between frames. Compared with optical flow based methods, Cheng *et al.* [5] proposes a transformer-based model to implicitly model both short and long-term temporal consistency between frames. Besides, they propose MoCA-Mask, a dataset which selects 87 camouflaged video sequences from MoCA with pixel-level handcrafted labeling. ZoomNeXt [25] imitates human vision by zooming in and out frames to perceive camouflaged objects and utilizes temporal shift to propagate inter-frame differences. With the emergence of visual foundation models, several methods [26, 27] take advantage of the exceptional segmentation performance of SAM [28] to segment camouflaged objects in videos by injecting temporal information into the prompt and SAM features. However, due to the limitation posed by the low diversity of MoCA-Mask, most VCOD methods require pre-training on image datasets, e.g., COD10K, and more importantly, this constraint impedes the further advancement of this task.

## 3 CAMotion Dataset

### 3.1 Video Collection

The limited scale of existing VCOD dataset seriously hinders the comprehensive evaluation of recent VCOD algorithms. To address this issue, we build a large-scale VCOD dataset CAMotion with high-quality pixel-wise annotations. The whole dataset is collected from the viewpoint of biology-inspired hierarchical categorization. We retrieve from the Internet using the keywords *camouflaged mammals*, *concealed insects*, *camouflaged fishes*, etc. Consequently, we obtain 151 representative camouflaged categories, which significantly enrich the diversity of existing VCOD datasets that contain less than 50 categories. Details of the dataset categories can be found in Sec. B.1.

After determining the biology-inspired object categories, we collect more than 4,000 videos as the initial camouflaged videos. Then we evaluate the quality of these camouflaged videos and filter out the unrelated contents in each video and retain the usable clip containing camouflaged objects. As a result, we construct CAMotion consisting of 468 video sequences with more than 150K video frames. We split the total of 468 sequences into 360 sequences as training set and the other 108 sequences as testing set. In this dataset, the length of the video sequences varies from 114 frames to 1,051 frames. Similar to MoCA-Mask, we provide both mask and bounding box annotations with an interval of five frames per sequence, accounting for 30,007 frames in total.

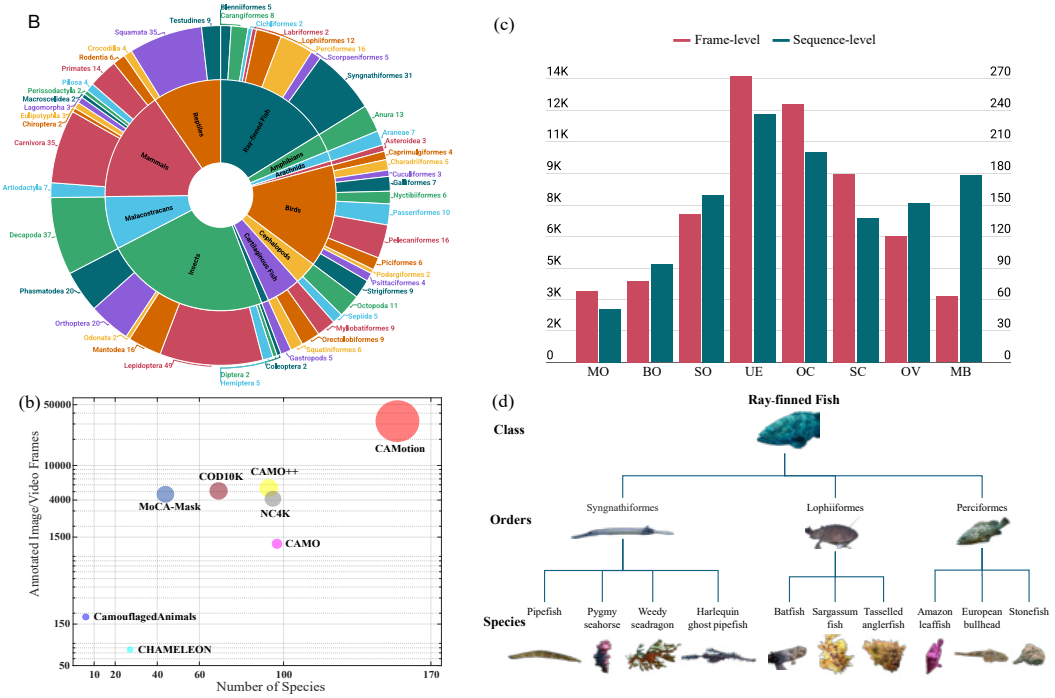



Figure 2: Dataset features and category examples from CAMotion dataset. (a) Taxonomic structure of CAMotion. (b) The scale and species comparison between existing COD dataset and CAMotion. (c) The attributes distribution in frame-level and sequence-level. (d) Examples of the species in CAMotion. Please zoom in for details.

Table 2: List and description of the eight attributes that characterize videos in CAMotion.

| Attr | Description |
|---|---|
| **MO** | Multiple Objects: image contains at least two objects. |
| **BO** | Big Object: ratio between object area and image area $\geq$0.15. |
| **SO** | Small Object: ratio between object area and image area $\leq$0.02. |
| **UE** | Uncertain Edge: the foreground and background areas around object have similar colors and textures. |
| **OC** | Occlusion: the object is partially occluded. |
| **SC** | Shape Complexity: object contains thin parts (e.g., animal foot). |
| **OV** | Out-of-View: some portion of the object leaves the camera field of view. |
| **MB** | Motion Blur: the object region is blurred due to the motion of object or camera. |

### 3.2 Sequence Annotation

The quality of the annotation plays a crucial role in dense prediction task. To this end, we present high-quality pixel-wise annotation in CAMotion, which is significantly larger than existing COD datasets, *e.g.* COD10K, NC4K and VCOD dataset, *e.g.* MoCA-Mask.

**Quality control.** We make great effort to present precise annotations on the collected videos, and conduct feedback error correction to ensure the annotation quality. Specifically, we ask five annotators to identify the camouflaged instances in each image and use an interactive segmentation tool to annotate them via pixel-wise masks. It takes each annotator 3 to 20 minutes to annotate an image depending on its complexity. The annotator manually draws/edits the camouflaged object's boundary in each frame, and two other annotators inspect the results and adjust them if necessary. Afterwards, the annotation results are reviewed by two experts with professional knowledge on VCOD task. If an annotation result is not unanimously agreed by the experts, it will be sent back to the original annotators to revise. To improve the annotation quality as much as possible, the annotators are required to annotate these challenging video frames very carefully and revise them frequently. More than 30% of the initial annotations fail in the first round of validation. some crucial video frames are revised more than three times. We present some challenging frames that are initially

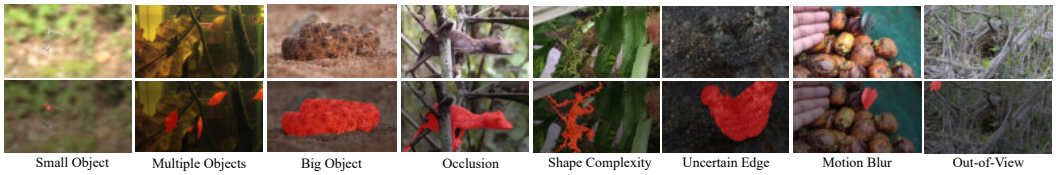



Figure 3: Visualization of the challenging attributes in CAMotion. Best viewed in color and zoom in for details.

labeled inaccurately in Fig. 5. With all these efforts, we finally construct CAMotion dataset with high-quality dense annotation.

**Categories.** As shown in Fig. 2(a), the camouflaged videos in our dataset follow a biology-inspired hierarchical categorization. All of the video sequences are firstly divided into 12 classes, including *mammals, insects, birds, ray-finned fish, etc*. Then these videos are further classified into 50 subclasses, which can be regarded as the biological orders like *carnivora, primates, and lepidoptera, etc*. To present more detailed analyses, we further categorize these data into 151 species such as *polar bears, dragonflies, tigers, cats, and batfishes, etc*. A representative taxonomic hierarchy tree of *Ray-finned Fish* is demonstrated in Fig. 2(d). To the best of our knowledge, our CAMotion is the largest VCOD dataset with diverse categories in the research community.

**Attributes.** To present deep analyses of the camouflaged videos in various challenging scenes, we label each camouflaged frame with 8 attributes, including uncertain edge (UE), big object (BO), multiple objects (MO), small object (SO), occlusions (OC), shape complexity (SC), out-of-view (OV) and motion blur (MB). The details of each attribute description are provided in Table 2. We provide attribute annotations for all the video frames in our dataset.

From Fig. 2(c), we observe that the most common challenge factors in CAMotion are uncertain edge (UE), occlusions (OC), shape complexity (SC) and small object (SO). Such observations align with the intuitive reality that camouflaged objects in local region are seamlessly blend into the surrounding backgrounds, thereby making the camouflaged objects imperceptible in these challenging scenes. Compared to MoCA [7] and MoCA-Mask [5] that simply categorized into three types of motion, *i.e.* static, locomotion and deformation, CAMotion can provide more comprehensive attributes for camouflaged behavior analyses. More detailed analyses of the aforementioned attributes are presented in Sec. B.2.

### 3.3 Dataset Specification and Statistics

**Object size.** Fig. 4(a) shows the object size distribution of the proposed CAMotion, the size distribution mainly ranges from 0.01 to 0.1, indicating that most of the camouflaged objects in CAMotion are tiny/small objects. CAMotion also contains a certain number of camouflaged objects with the size ranging from $[0.1, 0.35]$, making it possible to provide comprehensive analyses on the size distribution of VCOD.

**Duration.** To evaluate the temporal adaptability of the VCOD algorithm, we ensure that each sequence comprises at least four seconds with more than 114 frames, and the average sequence length in CAMotion is around 340 frames, see Fig. 4(b). The longest videos in CAMotion persist nearly 35s, including more than 1000 frames in a clip. Consequently, the durations of the videos are significantly longer than the previous MoCA-Mask dataset.

**Global and local contrast.** We adopt global and local color contrast distributions to measure the difficulty of camouflaged object detection in CAMotion. As shown in Fig. 4(c), the camouflaged objects in most video frames have low local contrast, indicating that the objects are highly similar to the local surroundings. Besides, the higher global contrast verifies that CAMotion has abundant species/scene diversity.

**Motion statistics.** Fig. 4(d) shows the motion statistics of the camouflaged objects in CAMotion dataset, most of the objects share the locomotion or deformation while only $6.0\%$ objects are still without obvious motion or appearance changes. Compared to MoCA-Mask, the camouflaged objects in CAMotion demonstrate more informative motion cues because of the frequent camera pose variations, body-part movements and environmental dynamics.

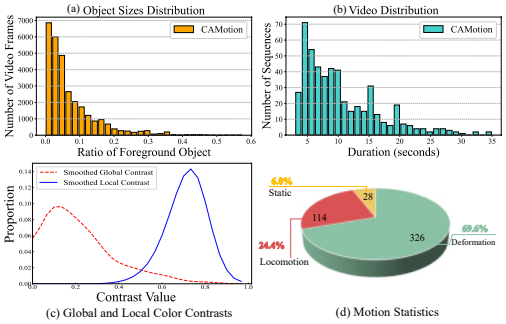



Figure 4: Statistics for CAMotion dataset. (a) Object sizes Distribution. (b) The distribution of video durations. (c) Global/local contrast distribution. (d) Motion statistics of the camouflaged objects.

Figure 5: Examples of fine-tuning initial annotations. White denotes unchanged areas; red and green indicate the original and refined annotations, respectively. Please zoom in for details.

Table 3: Quantitative comparison with 15 cutting-edge methods on CAMotion and MoCA-Mask testing datasets. Notes ↑ / ↓ denotes the higher/lower the better, and the best and second best are **bolded** and underlined for highlighting, respectively. * indicates the model requires the first frame ground-truth mask.

| Methods | CAMotion | | | | | | MoCA-Mask | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $S_\alpha \uparrow$ | $F_\beta^w \uparrow$ | $E_\phi^m \uparrow$ | $\mathcal{M} \downarrow$ | mDic↑ | mIoU↑ | $S_\alpha \uparrow$ | $F_\beta^w \uparrow$ | $E_\phi^m \uparrow$ | $\mathcal{M} \downarrow$ | mDic↑ | mIoU↑ |
| MGL-R[13] | 0.584 | 0.298 | 0.627 | 0.078 | 0.269 | 0.195 | 0.493 | 0.034 | 0.519 | 0.059 | 0.048 | 0.033 |
| PFNet[29] | 0.699 | 0.472 | 0.784 | 0.047 | 0.510 | 0.406 | 0.558 | 0.142 | 0.633 | 0.026 | 0.172 | 0.118 |
| UGTR[30] | 0.693 | 0.432 | 0.765 | 0.043 | 0.477 | 0.482 | 0.493 | 0.048 | 0.459 | 0.088 | 0.078 | 0.049 |
| SegMaR[31] | 0.662 | 0.415 | 0.720 | 0.045 | 0.436 | 0.342 | 0.542 | 0.129 | 0.544 | 0.024 | 0.139 | 0.093 |
| ZoomNet[11] | 0.672 | 0.419 | 0.701 | 0.045 | 0.237 | 0.358 | 0.582 | 0.201 | 0.682 | 0.026 | 0.236 | 0.197 |
| SINet-v2[32] | 0.713 | 0.489 | 0.800 | 0.046 | 0.533 | 0.425 | 0.571 | 0.175 | 0.608 | 0.035 | 0.211 | 0.153 |
| FSPNet[33] | 0.734 | 0.575 | 0.759 | 0.039 | 0.553 | 0.457 | 0.565 | 0.186 | 0.610 | 0.044 | 0.238 | 0.167 |
| PUENet[20] | 0.743 | 0.560 | 0.816 | 0.048 | 0.607 | 0.498 | 0.594 | 0.204 | 0.619 | 0.037 | 0.300 | 0.212 |
| PopNet[34] | 0.688 | 0.459 | 0.758 | 0.048 | 0.492 | 0.394 | 0.613 | 0.317 | 0.694 | 0.035 | 0.307 | 0.219 |
| HGINet[35] | **0.774** | **0.623** | **0.840** | **0.035** | **0.658** | **0.555** | 0.677 | 0.403 | 0.744 | 0.010 | 0.441 | 0.357 |
| CamoDiffusion[36] | 0.756 | 0.581 | 0.829 | 0.043 | 0.609 | 0.516 | 0.676 | 0.382 | 0.747 | 0.012 | 0.410 | 0.340 |
| SLT-Net[5] | 0.711 | 0.466 | 0.776 | 0.050 | 0.553 | 0.443 | 0.631 | 0.311 | 0.759 | 0.027 | 0.360 | 0.272 |
| SAM2 w/o prompt[37] | 0.474 | 0.071 | 0.436 | 0.069 | 0.069 | 0.045 | 0.495 | 0.056 | 0.487 | 0.023 | 0.057 | 0.035 |
| SAM-PM*[27] | 0.551 | 0.342 | 0.668 | 0.085 | 0.492 | 0.327 | 0.728 | **0.567** | **0.813** | **0.009** | **0.594** | **0.502** |
| ZoomNeXt[25] | 0.765 | 0.593 | 0.818 | 0.043 | 0.625 | 0.528 | **0.734** | 0.476 | 0.736 | 0.010 | 0.497 | 0.422 |

# 4 Experiment

## 4.1 Experiment Settings

**Datasets.** We use two VCOD datasets, MoCA-Mask [5] and our CAMotion, and an image COD dataset, COD10K [2] to conduct the experiments. MoCA-Mask is reorganized from MoCA [7], which contains 71 sequences with 3,946 frames for training and 16 sequences with 745 frames for testing. Our proposed CAMotion dataset includes 360 sequences with 23,502 frames for training and 108 sequences with 6,505 frames for testing. COD10K contains 3,040 training and 2,026 testing camouflaged images. Following the previous setting [5, 25], training conducted on MoCA-Mask is pretrained on COD10K and fine-tuned on MoCA-Mask.

**Evaluation metrics.** Following [5], we use six common evaluation metrics for CAMotion, including S-measure ($S_\alpha$) [38], weighted F-measure ($F_\beta^w$) [39], mean E-measure ($E_\phi^m$) [40], mean absolute error ($\mathcal{M}$), mean dice (mDic) and mean IoU (mIoU).

## 4.2 Benchmarks

**Baseline.** We select 15 cutting-edge baselines, including **(i)** 11 COD methods, *i.e.*, MGL-R [13], PFNet [29], UGTR [30], SegMaR [31], ZoomNet [11], SINet-v2 [32], FSPNet [33], PUENet

[20], PopNet [34], HGINet [35], and CamoDiffusion [36], **(ii)** four VCOD methods, *i.e.*, SLT-Net [5], SAM2 [37], SAM-PM [27], and ZoomNeXt [25]. These baselines are selected according to the following criteria [2]: (1) classical architectures, (2) recently published, (3) achieve SOTA performance in the specific fields, *i.e.*, COD and VCOD.

**Quantitative comparison.** We evaluate 15 selected state-of-the-art methods on CAMotion and MoCA-Mask testing datasets, and present the quantitative performance in Table 3. Due to the variations in network architectures, input resolutions, modalities, as well as pre-processing techniques, we make the best effort to ensure a fair comparison on both datasets. Regarding CAMotion, we surprisingly observe that the image-level COD method HGINet [35] outperforms all other methods in all six metrics, even surpassing video-based methods like ZoomNeXt [25]. Specifically, it achieves performance gains of 1.2%, 5.1%, 2.7%, 18.6%, 5.3%, and 5.1% in terms of $S_\alpha$, $F_\beta^w$, $E_\phi^m$, $\mathcal{M}$, mDic, and mIoU, respectively, compared to the current state-of-the-art VCOD method ZoomNeXt. However, ZoomNeXt achieves better performance against HGINet on MoCA-Mask and demonstrates more balanced performance across multiple datasets, which suggests that ZoomNeXt can leverage temporal cues more effectively while still exhibiting limited capability in camouflaged object discrimination. This performance discrepancy highlights the limitations of current approaches and thereby leaves substantial room for further research on our dataset. Additionally, thanks to the diversity of object sizes in our dataset, the evaluation results of the SOTA methods on CAMotion are more stable, particularly $\mathcal{M}$ versus other metrics. In contrast, MoCA-Mask tends to exhibit extremely low $\mathcal{M}$ but significantly worse performance on the remaining five metrics. This imbalance can be attributed to the fact that its test set is composed of almost entire small objects, indicating it lacks scale diversity and the results are greatly biased. The outstanding performance of HGINet on CAMotion further reflects the limitations of existing VCOD methods in balancing camouflaged discovery and temporal consistency. Moreover, the significant performance gap between existing COD datasets and CAMotion, along with the ineffectiveness of SAM2 [37], indicates the difficulties of detecting camouflaged objects from video sequences. We believe that CAMotion opens up a broad and meaningful research space, and we strongly encourage the community to conduct further research in the underexplored areas.

**Qualitative comparison.** As shown in Fig. 6, we perform the visual comparison of HGINet [35] and ZoomNeXt [25] in two typical scenarios, shape complexity (Row 1-4) and uncertain edge (Row 5-8). As shown in Row 1-4, HGINet possesses superior discriminative ability in locating and segmenting camouflaged objects from distracting backgrounds against ZoomNeXt. In contrast, ZoomNeXt tends to propagate distractive information across subsequent frames because of the lack of discriminative ability. However, HGINet fails to maintain consistent object localization, even though the camouflaged object was well-identified in previous frames (see Row 7). In contrast, Row 8 demonstrates the superior results obtained by ZoomNeXt, as it leverages the temporal information to enhance the temporal consistency. Similar to Table 3, Fig. 6 further illustrates that current methods struggle to balance the camouflaged discriminative capability and temporal consistency.

### 4.3 Dataset Analysis

**Attribute-based performances.** To investigate how varying challenging scenes affect the results, we visualize the performance of HGINet [35], CamoDiffusion [36], ZoomNeXt [25], and SLT-Net [5] in eight challenging attributes in terms of $S_\alpha$ and $\mathcal{M}$, see Fig. 7. The four methods exhibit consistent performance trends across these two metrics and eight challenging attributes, which validates the reliability of our dataset. Notably, we observe that the sequences involving multiple objects and complex shapes are significantly more difficult. In contrast, sequences characterized by uncertain edge, occlusion, out-of-view, and motion blur tend to yield relatively better performance. Furthermore, an evident difference is observed between big and small objects, attributed to the intrinsic properties of the evaluation metrics: big objects typically result in higher $S_\alpha$ but poorer $\mathcal{M}$, whereas small objects display the opposite behavior. Details of other metrics can be found in Sec. C.4.

**Cross-dataset generalization.** Since the generalization ability and difficulty of datasets play a significant role in both training and evaluation, we study these two aspects on COD10K, our CAMotion, and MoCA-Mask datasets, using the cross-dataset analysis method [41], *i.e.*, train a model on one dataset and test on all selected datasets. For a fair comparison, we adopt the recently proposed ZoomNeXt as the model and train it on each dataset until the loss becomes stable. Table 4 shows the results of $S_\alpha$ and $\mathcal{M}$. Each row presents performance trained on a specific dataset and evaluated on all selected datasets, *i.e.*, COD10K, CAMotion, and MoCA-Mask, reflecting the generalization capability of the

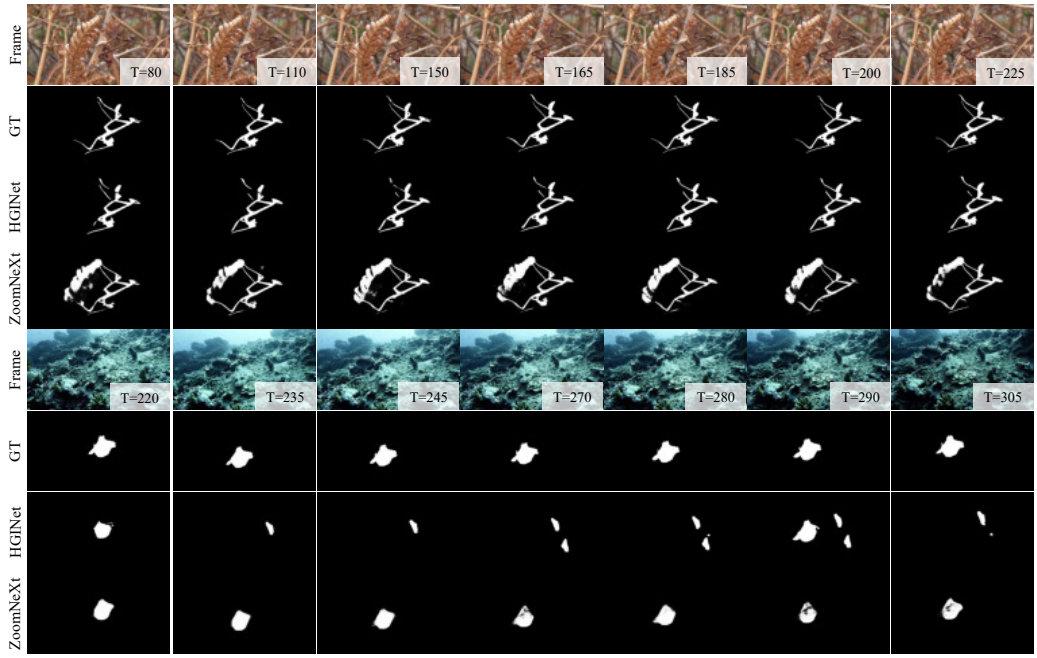
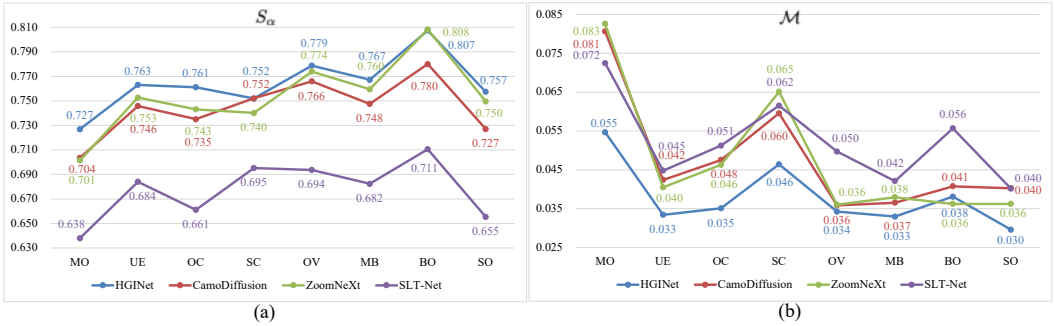



Figure 6: Visual comparison with state-of-the-art methods in challenging scenarios, *i.e.*, shape complexity (Row 1-4) and uncertain edge (Row 5-8). Please zoom in for details.




(a) (b)

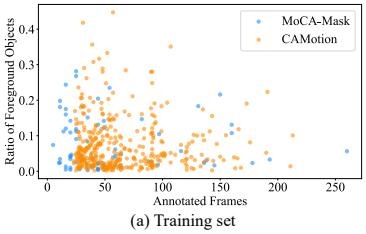
Figure 7: Visualization of SOTA method performances on different challenging attributes in terms of (a) $S_\alpha \uparrow$ and (b) $\mathcal{M} \downarrow$.

training dataset. Each column shows the performance of ZoomNeXt tested on a particular dataset, highlighting the difficulty of each dataset. As expected, we observe that CAMotion is more difficult and achieves better generalization capability against the largest COD benchmark, COD10K, on both $S_\alpha$ and $\mathcal{M}$. Take $S_\alpha$ as an example, the "Percent Drop" for CAMotion is 13% lower than that for COD10K, indicating a stronger generalization capability on CAMotion. In addition, the average $S_\alpha$ scores on COD10K and CAMotion are 0.695 and 0.625, respectively, further confirming the increased difficulty of CAMotion. Furthermore, the model trained on CAMotion outperforms the others on the MoCA-Mask testing set in terms of both metrics, demonstrating the generalization ability and diversity of our CAMotion. We also notice that the models trained on COD10K and CAMotion obtain "Percent Drop" in $\mathcal{M}$ on MoCA-Mask dataset, this is because most of the camouflaged objects in MoCA-Mask are very small, yielding a significant distribution gap compared to COD10K and CAMotion. The inconsistency between $S_\alpha$ and $\mathcal{M}$ on MoCA-Mask supports our analyses.

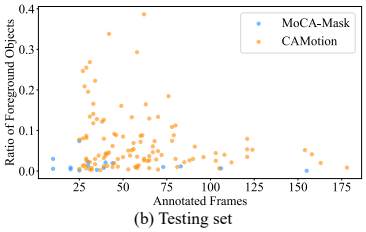
**Scale distribution comparison.** To illustrate the rationale for the mismatch between $S_\alpha$ and $\mathcal{M}$ in Table 4, we present the comparison of the scale distribution between CAMotion and MoCA-Mask in the training and testing datasets, see Fig. 8. As illustrated in Fig. 8(b), the MoCA-Mask testing set consists of 16 video clips, all of which predominantly feature small objects, with most ratios of foreground objects falling between 0 and 0.03. In contrast, our CAMotion testing set contains 108 clips and exhibits a well-balanced distribution between small and large objects. This discrepancy may partially explain why most models perform poorly in terms of most metrics on MoCA-Mask

Table 4: $S_\alpha$ and $\mathcal{M}$ results for cross-dataset generalization. The selected ZoomNeXt is trained on one (rows) dataset and tested on all datasets (columns). "Self" refers to training and testing on the same dataset (same as diagonal), and "Mean Others" refers to averaging performance on all except self.

| Metrics | Trained on \ Tested on | COD10K | CAMotion | MoCA-Mask | Self | Mean Others | Percent Drop |
|---|---|---|---|---|---|---|---|
| $S_\alpha \uparrow$ | COD10K | **0.717** | 0.532 | 0.512 | 0.717 | 0.522 | 27% |
| | CAMotion | 0.693 | **0.749** | 0.588 | 0.749 | 0.641 | 14% |
| | MoCA-Mask | 0.676 | 0.593 | **0.563** | 0.563 | 0.635 | -13% |
| | Mean others | 0.685 | 0.563 | 0.550 | 0.676 | 0.599 | 11% |
| $\mathcal{M} \downarrow$ | COD10K | **0.037** | 0.049 | 0.049 | 0.037 | 0.049 | 32% |
| | CAMotion | 0.030 | **0.044** | 0.034 | 0.044 | 0.032 | -27% |
| | MoCA-Mask | 0.030 | 0.054 | **0.037** | 0.037 | 0.042 | 14% |
| | Mean others | 0.030 | 0.052 | 0.042 | 0.039 | 0.041 | 5% |

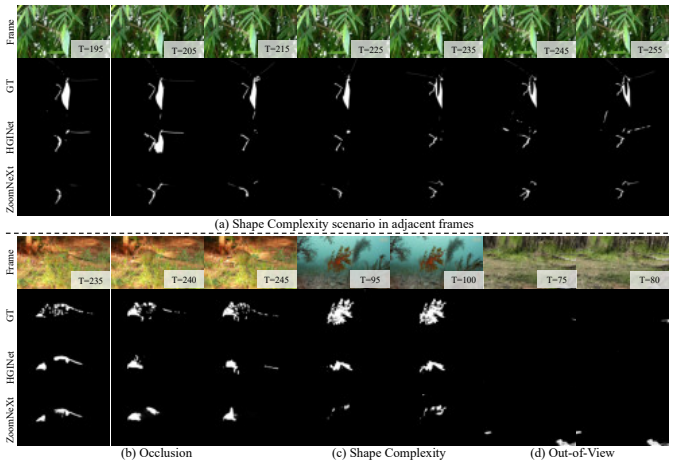


(a) Training set



(b) Testing set

Figure 8: Scale distribution of CAMotion and MoCA-Mask.



(a) Shape Complexity scenario in adjacent frames

(b) Occlusion (c) Shape Complexity (d) Out-of-View

Figure 9: Failure cases on both HGINet and ZoomNeXt in several challenging scenarios. Please zoom in for details.

testing set, but achieve superior performance in term of $\mathcal{M}$. On the other hand, the CAMotion testing set contains 108 clips with a broader distribution of object sizes, it offers a more comprehensive and balanced benchmark that reflecting real-world camouflaged objects with various scales. Fig. 8(a) also shows the scale diversity for the CAMotion training set. In addition, MoCA-Mask contains excessive sequences with fewer than 20 annotated frames, which hinders effective model training. From a broader perspective, we can observe that CAMotion offers more consistent distribution between the training and testing sets, leading to more reliable performance assessments.

**Failure cases.** We further present representative failure cases on both HGINet and ZoomNeXt in several challenging scenarios. As depicted in Fig. 9, Row 3 reveals that HGINet lacks the guidance of temporal cues to segment camouflaged objects across consecutive video frames. Row 4 indicates ZoomNeXt lacks sufficient discriminative ability to break the camouflage and therefore passes the distractive cues to the subsequent frames. Furthermore, Fig. 9(b), (c), and (d) illustrate failure cases under occlusion, shape complexity, and out-of-view scenarios, respectively.

## 5 Conclusion

In this paper, we construct CAMotion, a high-quality dataset covers a wide range of species for camouflaged motion object detection in the wild. CAMotion comprises various sequences with multiple challenging attributes such as uncertain edge, occlusion, fast motion and shape complexity, etc. Then we present annotation details and statistical distributions of the dataset, allowing CAMotion to analyze motion characteristics of camouflaged objects across diverse challenging scenarios. Finally, we conduct a comprehensive evaluation of SOTA models on CAMotion dataset and investigate the major challenges in the VCOD task.

## Reference


[1] Trung-Nghia Le, Tam V. Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer Vision and Image Understanding*, 184:45–56, 2019.

[2] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2774–2784, 2020.

[3] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11591–11601, 2021.

[4] Pia Bideau and Erik G. Learned-Miller. It's moving! A probabilistic model for causal motion segmentation in moving camera videos. In *European Conference on Computer Vision*, pages 433–449, 2016.

[5] Xuelian Cheng, Huan Xiong, Deng-Ping Fan, Yiran Zhong, Mehrtash Harandi, Tom Drummond, and Zongyuan Ge. Implicit motion handling for video camouflaged object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 13854–13863, 2022.

[6] Przemysław Skurowski, Hassan Abdulameer, Jakub Błaszczyk, Tomasz Depta, Adam Kornacki, and Przemysław Kozieł. Animal camouflage analysis: Chameleon database. *Unpublished manuscript*, 2(6):7, 2018.

[7] Hala Lamdouar, Charig Yang, Weidi Xie, and Andrew Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. In *Asian Conference on Computer Vision*, pages 488–503, 2020.

[8] Trung-Nghia Le, Yubo Cao, Tan-Cong Nguyen, Minh-Quan Le, Khanh-Duy Nguyen, Thanh-Toan Do, Minh-Triet Tran, and Tam V. Nguyen. Camouflaged instance segmentation in-the-wild: Dataset, method, and benchmark suite. *IEEE Transactions on Image Processing*, 31:287–300, 2022.

[9] Yujia Sun, Geng Chen, Tao Zhou, Yi Zhang, and Nian Liu. Context-aware cross-level fusion network for camouflaged object detection. In *International Joint Conference on Artificial Intelligence*, pages 1025–1031, 2021.

[10] Miao Zhang, Shuang Xu, Yongri Piao, Dongxiang Shi, Shusen Lin, and Huchuan Lu. Preynet: Preying on camouflaged objects. In *Proceedings of the ACM International Conference on Multimedia*, pages 5323–5332, 2022.

[11] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2150–2160, 2022.

[12] Aixuan Li, Jing Zhang, Yunqiu Lv, Bowen Liu, Tong Zhang, and Yuchao Dai. Uncertainty-aware joint salient object and camouflaged object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10071–10081, 2021.

[13] Qiang Zhai, Xin Li, Fan Yang, Chenglizhao Chen, Hong Cheng, and Deng-Ping Fan. Mutual graph learning for camouflaged object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12997–13007, 2021.

[14] Peng Li, Xuefeng Yan, Hongwei Zhu, Mingqiang Wei, Xiao-Ping Zhang, and Jing Qin. Findnet: Can you find me? boundary-and-texture enhancement network for camouflaged object detection. *IEEE Transactions on Image Processing*, 31:6396–6411, 2022.

[15] Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Ccamouflaged object detection with feature decomposition and edge reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 22046–22055, 2023.

[16] Cunhan Guo and Heyan Huang. Cofinet: Unveiling camouflaged objects with multi-scale finesse. *arXiv preprint arXiv:2402.02217*, 2024.

[17] Chunming He, Kai Li, Yachao Zhang, Yulun Zhang, Zhenhua Guo, Xiu Li, Martin Danelljan, and Fisher Yu. Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects. In *International Conference on Learning Representations*, 2024.

[18] Wenda Zhao, Shigeng Xie, Fan Zhao, You He, and Huchuan Lu. Nowhere to disguise: Spot camouflaged objects via saliency attribute transfer. *IEEE Transactions on Image Processing*, 32:3108–3120, 2023.

[19] Tao Zhou, Yi Zhou, Chen Gong, Jian Yang, and Yu Zhang. Feature aggregation and propagation network for camouflaged object detection. *IEEE Transactions on Image Processing*, 31:7036–7047, 2022.

[20] Yi Zhang, Jing Zhang, Wassim Hamidouche, and Olivier Déforges. Predictive uncertainty estimation for camouflaged object detection. *IEEE Transactions on Image Processing*, 32:3580–3591, 2023.

[21] Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4494–4503, 2022.

[22] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *IEEE International Conference on Computer Vision*, pages 7157–7168, 2021.

[23] Junyu Xie, Weidi Xie, and Andrew Zisserman. Segmenting moving objects via an object-centric layered representation. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Conference on Neural Information Processing Systems*, 2022.

[24] Etienne Meunier, Anaïs Badoual, and Patrick Bouthemy. EM-Driven unsupervised learning for efficient motion segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4462–4473, 2023.

[25] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. ZoomNeXt: A unified collaborative pyramid network for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):9205–9220, 2024.

[26] Wenjun Hui, Zhenfeng Zhu, Shuai Zheng, and Yao Zhao. Endow SAM with keen eyes: Temporal-spatial prompt learning for video camouflaged object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 19058–19067, 2024.

[27] Muhammad Nawfal Meeran, Gokul Adethya T, and Bhanu Pratyush Mantha. SAM-PM: enhancing video camouflaged object detection using spatio-temporal attention. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1857–1866, 2024.

[28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[29] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8772–8781, 2021.

[30] Fan Yang, Qiang Zhai, Xin Li, Rui Huang, Ao Luo, Hong Cheng, and Deng-Ping Fan. Uncertainty-guided transformer reasoning for camouflaged object detection. In *IEEE International Conference on Computer Vision*, pages 4126–4135, 2021.

[31] Qi Jia, Shuilian Yao, Yu Liu, Xin Fan, Risheng Liu, and Zhongxuan Luo. Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4703–4712, 2022.

[32] 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.

[33] 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 *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5557–5566, 2023.

[34] Zongwei Wu, Danda Pani Paudel, Deng-Ping Fan, Jingjing Wang, Shuo Wang, Cédric Demonceaux, Radu Timofte, and Luc Van Gool. Source-free depth for object pop-out. In *IEEE International Conference on Computer Vision*, pages 1032–1042, 2023.

[35] Siyuan Yao, Hao Sun, Tian-Zhu Xiang, Xiao Wang, and Xiaochun Cao. Hierarchical graph interaction transformer with dynamic token clustering for camouflaged object detection. *IEEE Transactions on Image Processing*, 33:5936–5948, 2024.

[36] Ke Sun, Zhongxi Chen, Xianming Lin, Xiaoshuai Sun, Hong Liu, and Rongrong Ji. Conditional diffusion models for camouflaged and salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–16, 2025.

[37] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[38] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *IEEE International Conference on Computer Vision*, pages 4558–4567, 2017.

[39] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2014.

[40] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *International Joint Conference on Artificial Intelligence*, pages 698–704, 2018.

[41] Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1521–1528, 2011.

[42] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:13941–13958, 2023.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We provide a summary of our dataset in the abstract and introduction, describe its main contribution, and contrast it with existing datasets.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We provide a comprehensive review of our dataset in Section D. We also conduct the direction of our future work and the limitations of evaluation metrics.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: Our dataset provides a wide range of species for video camouflaged object detection, which is no need to contain any theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: In this paper, we present a new dataset, CAMotion, and report the performance of classical and recently published methods on our dataset. The training setting and dataset train/test split are also included in the paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The dataset homepage and GitHub repository links will be made available upon camera-ready submission.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We report the performance of existing models on our dataset, following the original settings and publicly available implementations provided by the respective papers. The data splits are also provided in our paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We conduct plenty of experiments on our dataset in Section 4 and Section C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Since we conduct the experiments on various existing models, the settings we use follow the original settings in their codebase. We also report the implementation details in Section C.2.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We carefully read the NeurIPS Code of Ethics, and the research conducted in the paper conforms to the ethics. Besides, all annotators are fairly compensated in accordance with the hiring standards.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the potential societal impact in Section E.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: After collecting the data from the Internet, we conducted a thorough manual inspection to ensure that videos do not contain any unsafe content, such as violent, sexual, or otherwise disturbing imagery. To the best of our knowledge, our dataset does not pose safety or ethical risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We collect video data from publicly accessible sources on the Internet. Before including any video in our dataset, we verify that it is either (1) released under a Creative Commons Attribution (CC-BY) license, which permits reuse with proper credit, or (2) accompanied by explicit terms of use that allow data collection for research purposes.

    Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The license will be involved in the GitHub repository, which will be made available upon camera-ready submission.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: This paper does not involve any usage of LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

# CAMotion: A High-Quality Dataset for Camouflaged Motion Object Detection in the Wild

# Appendix

## Contents

# A Datasets/Benchmarks

## A.1 COD Benchmarks

**CHAMELEON.** CHAMELEON is the smallest COD dataset consisting of only 76 images. The images are collected from the Internet using the keyword "camouflaged animal" via Google search. Each image is manually annotated with pixel-level ground truth. Due to its limited size and the absence of formal peer review, CHAMELEON is typically utilized for preliminary validation or as a supplementary benchmark rather than for model training.

**CAMO.** CAMO is the first officially released COD dataset, containing a total of 2,500 images. It is composed of two equal subsets: the CAMO subset, which includes 1,250 images featuring at least one camouflaged object, and the MS-COCO subset, which includes 1,250 non-camouflaged images used as negative samples. The dataset spans eight super-categories to ensure category diversity. A standard 80%-20% split is applied for training and testing.

**COD10K.** COD10K consists of 10,000 high-resolution images sourced from various photography platforms, encompassing 10 super-categories and 78 sub-categories. The dataset is divided into three types: 5,066 images containing camouflaged objects, 3,000 background-only images, and 1,934 images featuring non-camouflaged objects. All camouflaged instances are annotated in a hierarchical manner, including category labels, bounding boxes, object-level masks, and instance-level segmentation, thus enabling a wide range of downstream tasks. In most research settings, only the 5,066 images with camouflaged objects are utilized for training and evaluation of COD models.

**NC4K.** The NC4K dataset contains 4,121 images collected from online resources, all of which are annotated with both object-level and instance-level ground truth masks. Unlike other datasets primarily used for training, NC4K is typically employed as a dedicated test set to assess the generalization performance of COD models. As the largest test-only dataset currently available in the COD field, it provides a comprehensive and diverse benchmark for evaluating the robustness of COD algorithms in real-world scenarios.

**CAMO++.** CAMO++ is an extended version of the CAMO dataset, expanding both scale and diversity of the dataset. It comprises 5,500 images featuring humans and over 90 distinct animal species, including 2,700 camouflaged images and 2,800 non-camouflaged counterparts. Each image is annotated with a hierarchical labeling scheme that includes meta-category and fine-grained category labels, bounding boxes, and instance-level segmentation masks. Additionally, all instances are annotated with pixel-level ground truth masks through manual labeling. CAMO++ serves as a valuable benchmark not only for camouflaged instance segmentation but also for broader tasks such as semantic camouflage segmentation and video-based camouflaged object detection.

## A.2 VCOD Benchmarks

**CAD.** The Camouflaged Animal Dataset (CAD) is a small-scale dataset specifically designed for camouflaged motion object segmentation. It consists of nine short video sequences collected from publicly available YouTube content. To support temporal analysis, manual pixel-level annotations are provided for every fifth frame within each sequence. Despite its limited size, CAD offers valuable benchmarks for evaluating the temporal consistency and robustness of COD models in dynamic, real-world environments.

**MoCA-Mask.** The original Moving Camouflaged Animals (MoCA) dataset comprises approximately 37,000 frames extracted from 141 YouTube video sequences, predominantly recorded at a resolution of $720 \times 1280$ pixels and a frame rate of 24 fps. This dataset includes 67 animal species captured in natural environments, though not all of these animals are camouflaged. The original annotations provide only bounding box labels rather than pixel-level segmentation masks, limiting its effectiveness for evaluating VCOD segmentation tasks. To address this limitation, MoCA was reprocessed into the MoCA-Mask dataset, which contains 87 high-quality video sequences totaling 22,939 frames with dense, pixel-wise annotations on every fifth frame.

# B CAMotion Details

## B.1 Categories and Species

We present a detailed breakdown of the total frame count according to the taxonomic hierarchy tree, categorized into 12 distinct classes, which are *Ray-finned Fish, Amphibians, Arachnids, Asteroidea, Birds, Cephalopods, Cartilaginous Fish, Gastropods, Insects, Malacostracans, Mammals, and Reptiles* and 151 species (see Fig. 10).

As shown in Fig. 11, we present examples from our CAMotion dataset, showing the original video frames alongside their corresponding optical flow generated using GMFlow [42] and pixel-level annotations. These visualizations highlight the quality and consistency of motion cues alongside semantic labels.

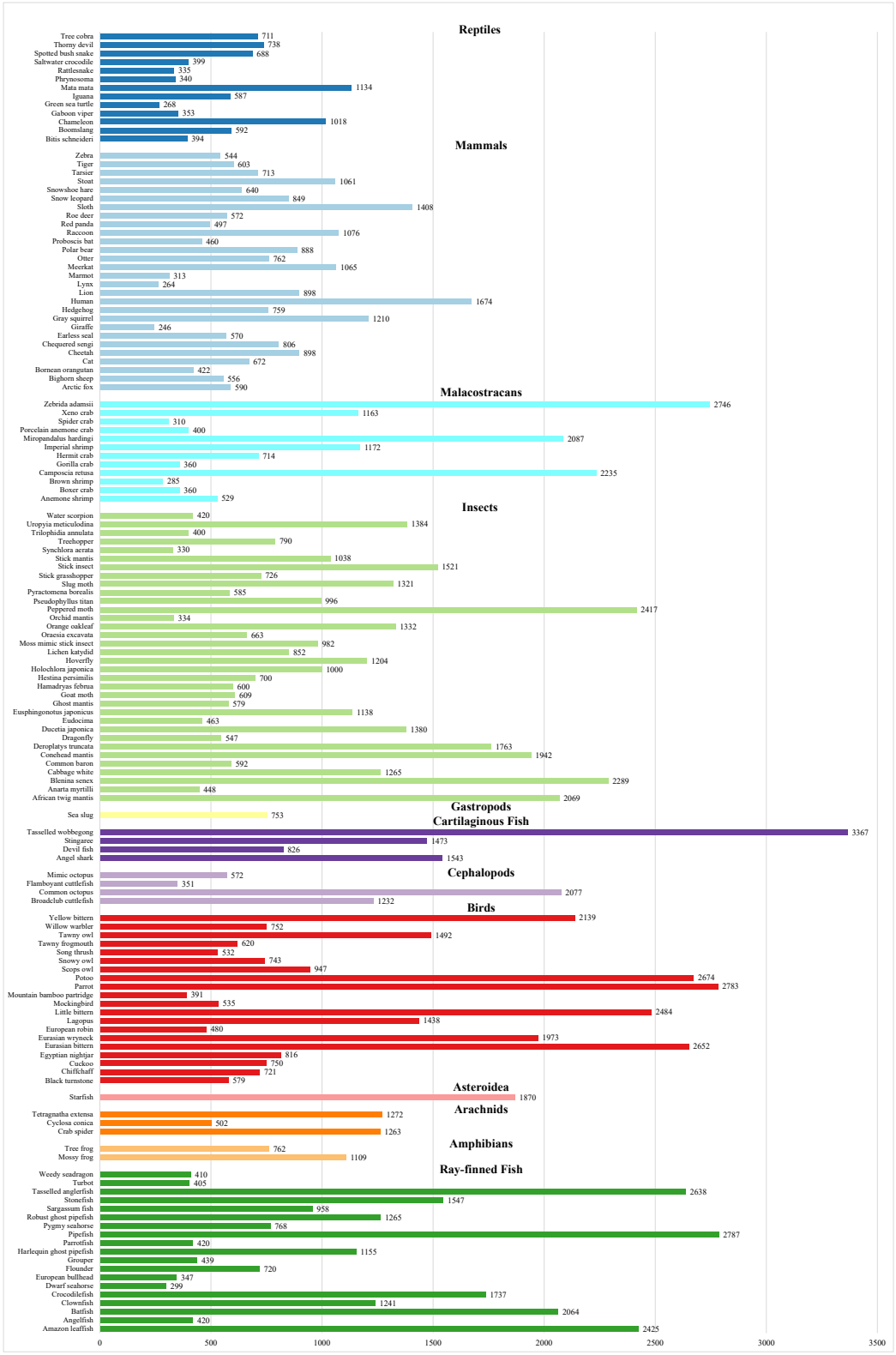



Figure 10: Detailed classification of the total frame count across different species. Please zoom in for details.

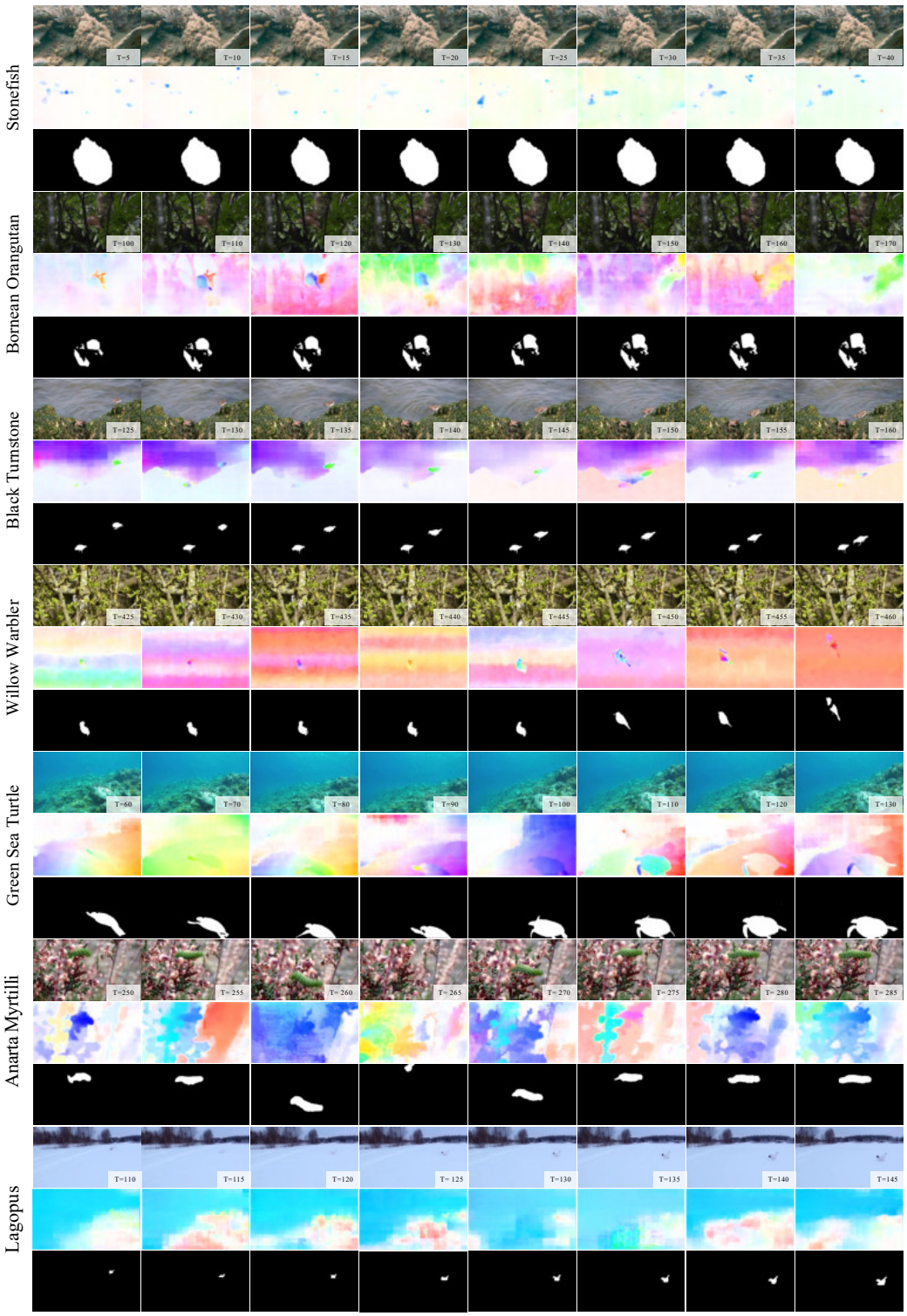



Figure 11: Examples of our CAMotion dataset with corresponding optical flow and pixel-level annotations. Each group is arranged as the original image, optical flow and pixel-level annotation. Please zoom in for details.

### B.2 Attributes

As an extension of Fig. 3, Fig. 12 presents more examples of eight challenging attributes in CAMotion, in terms of *Small Object (SO), Big Object (BO), Multiple Objects (MO), Occlusion (OC), Shape Complexity (SC), Uncertain Edge (UE), Motion Blur (MB), and Out-of-View (OV)*.

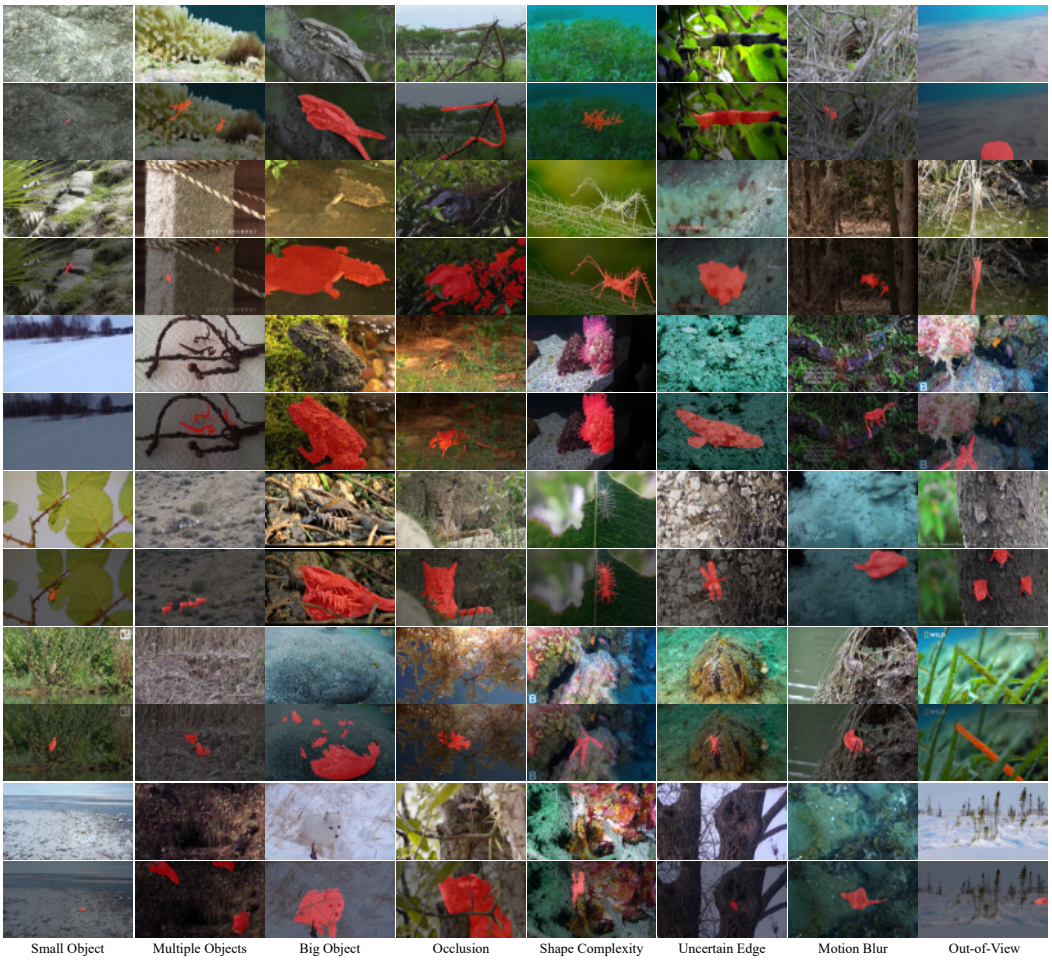



Figure 12: Visualization of the challenging attributes in CAMotion. Best viewed in color and zoom in for details.

### B.3 Annotations

As an extension of Fig. 5, we present more challenging frames that are initially labeled inaccurately and corrected during review and validation; see Fig. 13.

Figure 13: Examples of fine-tuning initial annotations. White denotes unchanged areas; red and green indicate the original and refined annotations, respectively. Please zoom in for details.

## C Experiments

### C.1 Evaluation Metrics

To comprehensively evaluate the performance of SOTA methods on our proposed CAMotion dataset, we adopt six widely-used quantitative metrics[5]: structure measure ($S_\alpha$)[38], weighted F-measure ($F_\beta^w$)[39], enhanced-alignment measure ($E_\phi^m$)[40], Mean absolute error ($\mathcal{M}$), mean dice coefficient (mDic), and mean intersection over union (mIoU). These metrics assess prediction quality from the aspects of structure similarity, pixel-wise accuracy, and spatial overlap.

**Structure measure ($S_\alpha$).** Considering that camouflaged objects have complex shapes, $S_\alpha$ is used for evaluating structural similarity between prediction and ground truth by combining region-aware ($S_r$) and object-aware ($S_o$) components, which is defined as:

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

where $\alpha \in [0,1]$ is the balance parameter and is set to 0.5 in the experiments. Note that the region-aware structural similarity ($S_r$) is designed to assess the object-part structure similarity against the GT masks. It recursively divides each of the predicted and GT masks into four blocks using horizontal and vertical cutoff lines that intersect at the centroid of the GT foreground, and calculates the structural similarity measure (SSIM) of each block using:

$$S_r = \sum_{k=1}^{K} w_k \cdot \mathrm{SSIM}(k), \tag{2}$$

where $w_k$ is the assigned weight of each block proportional to the GT foreground region this block covers, and $K$ is the total number of blocks. The object-aware structural similarity ($S_o$) is designed mainly to capture the sharp foreground-background contrast ($S_{FG}$) and uniform distribution ($S_{BG}$)

between predicted and GT masks, which is defined as:

$$S_o = \mu \cdot S_{FG} + (1 - \mu) \cdot S_{BG} \tag{3}$$

$$S_{FG} = \frac{2\overline{x}_{FG}}{(\overline{x}_{FG})^2 + 1 + 2\lambda \cdot \sigma_{x_{FG}}} \tag{4}$$

$$S_{BG} = \frac{2\overline{x}_{BG}}{(\overline{x}_{BG})^2 + 1 + 2\lambda \cdot \sigma_{x_{BG}}} \tag{5}$$

where $\sigma_x$ and $\overline{x}$ denote the standard deviation and mean of the predicted binary mask, respectively. $\mu$ is the ratio of foreground area in GT to image area, $\lambda$ is a constant to balance the two terms, and $FG$ and $BG$ represent foreground and background area, respectively.

**Weighted F-measure ($F_\beta^w$)** improves the traditional F-measure by assigning different weights to precision and recall based on spatial importance. It is more sensitive to errors near object boundaries and provides a more reliable segmentation evaluation, which is defined as:

$$F_\beta^w = \frac{(1 + \beta^2) \cdot \text{Precision}^w \cdot \text{Recall}^w}{\beta^2 \cdot \text{Precision}^w + \text{Recall}^w}, \tag{6}$$

where $\beta^2$ is typically set to 0.3 and thereby placing more emphasis on precision than recall.

**Enhanced-alignment measure ($E_\phi$)** integrates both pixel-level matching and image-level statistics, simulating human visual perception. It measures alignment quality from both structural and statistical perspectives, especially suitable for binary segmentation tasks, which is defined as:

$$E_\phi = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \phi\left(\frac{2\varphi_G \circ \varphi_P}{\varphi_G \circ \varphi_G + \varphi_P \circ \varphi_P}\right) \tag{7}$$

where $\phi(\cdot)$ denotes the enhanced alignment matrix, $\circ$ is the Hadamard product. $\varphi$ is the deviation matrix, which represents the distance between each pixel value of the input binary mask and its global mean. $P$ and $G$ denote the predicted binary mask and GT, respectively.

**Mean absolute error ($\mathcal{M}$)** computes the average absolute difference between the predicted map and the ground truth mask at the pixel level. It provides a straightforward measure of prediction error with smaller values indicating better performance, which is defined as:

$$\mathcal{M} = \frac{1}{N} \sum_{i=1}^{N} |P_i - G_i|. \tag{8}$$

**Mean dice coefficient** (mDic) quantifies the spatial overlap between predicted and ground truth masks using the Dice formula. It emphasizes the consistency of foreground extraction and is widely adopted in segmentation evaluation. Dice coefficient is defined as:

$$\text{Dice} = \frac{2|P \cap G|}{|P| + |G|} = \frac{2TP}{2TP + FP + FN}. \tag{9}$$

**Mean intersection over union** (mIoU) calculates the ratio of the intersection and union of predicted and ground truth regions. As a stricter metric, it is highly representative in segmentation tasks for evaluating spatial accuracy and generalization. IoU is defined as:

$$\text{IoU} = \frac{TP}{TP + FP + FN}. \tag{10}$$

### C.2 Implementation Details

To ensure comprehensive and fair evaluation, we conduct experiments on two VCOD datasets MoCA-Mask [5] and our proposed CAMotion, as well as one image COD dataset, COD10K [2]. Specifically, MoCA-Mask is reorganized from MoCA[7], consisting of 71 training sequences (3,946 frames) and 16 testing sequences (745 frames). CAMotion significantly expands the data scale, comprising 360 sequences (23,502 frames) for training and 108 sequences (6,505 frames) for testing. COD10K includes 3,040 training images and 2,026 testing images.

Given the diversity in network designs, input resolutions, modalities, and preprocessing strategies among baselines, we carefully follow the original settings specified in each method's official implementation to ensure fair comparisons. We use input resolutions as the original setups: 473 $\times$ 473 for MGL-R [13] and UGTR [30]; 416 $\times$ 416 for PFNet [29]; 352 $\times$ 352 for SegMaR [31], SINet-v2 [32], PopNet [34], CamoDiffusion [36], and SLT-Net [5]; 384 $\times$ 384 for ZoomNet [11], FSPNet [33], and ZoomNeXt [25]; 512 $\times$ 512 for PUENet [20] and HGINet [35]; and 1024 $\times$ 1024 for SAM2 [37] and SAM-PM [27]. All experiments are conducted using four NVIDIA RTX L40 GPUs.

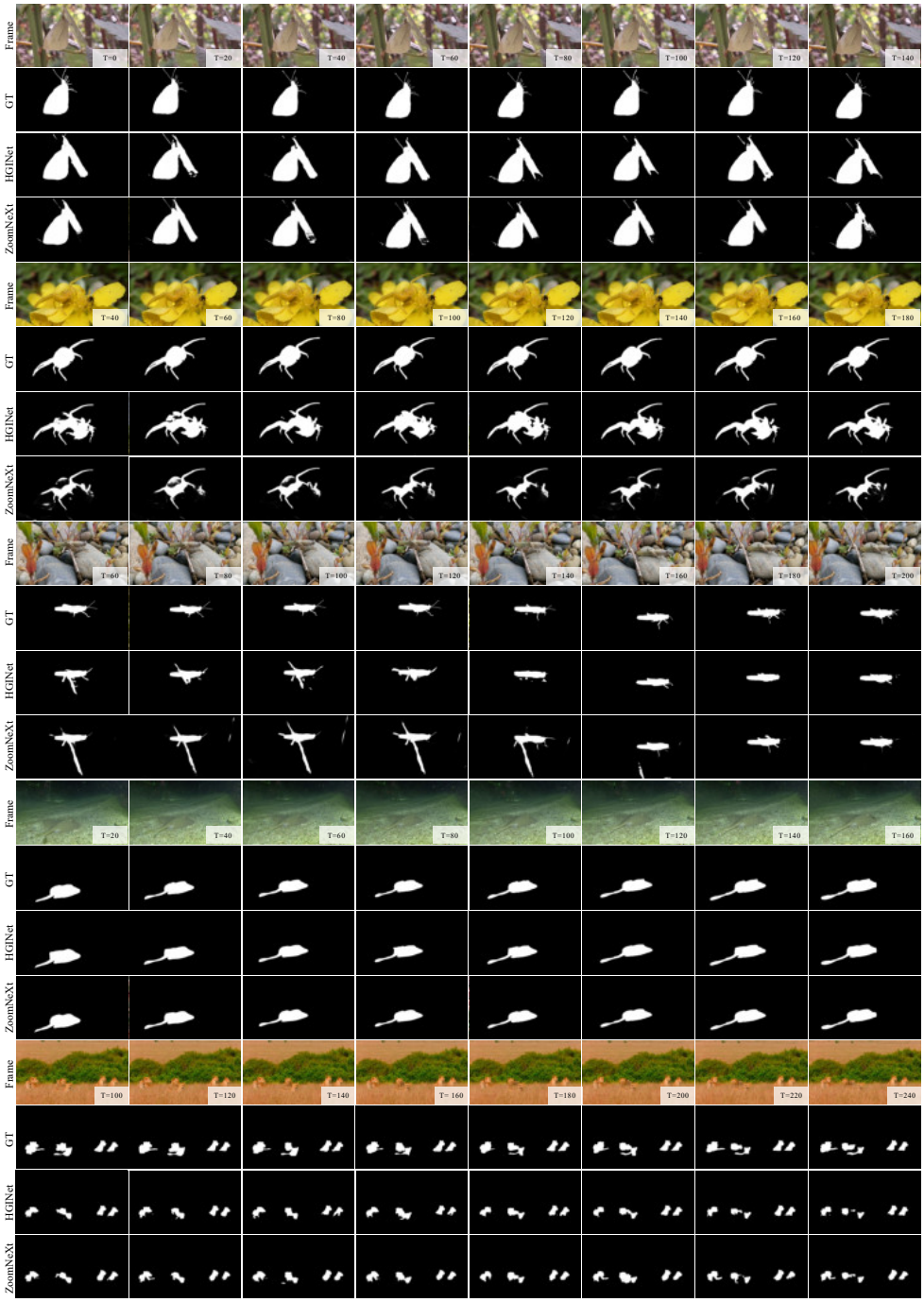



Figure 14: Visual comparison with state-of-the-art methods on CAMotion test dataset. Please zoom in for details.

### C.3 Qualitative Comparison

As shown in Fig. 14, we perform visual comparison of HGINet [35] and ZoomNeXt [25] in general scenarios. Overall, both methods can identify the location and shapes of camouflaged objects in a subset of specific video frames. However, they still suffer from the presence of highly confusing and distracting surrounding backgrounds, which degrade the segmentation performance. As shown in Row 1-4, the leaves surrounding the camouflaged objects exhibit high similarity in both appearance and texture, misguiding the results. A similar issue is observed in Row 5, where the presence of flower stamens around the camouflaged object also confuses HGINet to incorrect segmentation. This problem persists in Row 12, where ZoomNeXt propagates distracting information across subsequent frames due to its limited discriminative ability. In contrast, for scenes with less confusing or distractive context, like those shown in Row 13–16 and Row 17–20, both methods demonstrate strong discriminative capabilities in the single-object and multi-object scenarios. Nevertheless, incomplete segmentation also occurs in the occlusion scenario, as illustrated in Row 19 and 20.

### C.4 Dataset Analysis

**Category-based performances.** In Fig. 15, we visualize the performance of four SOTA methods HGINet[35], CamoDiffusion[36], ZoomNeXt[25], and SLT-Net[5] across different biological camouflaged object categories, in terms of $S_\alpha$, $F_\beta^w$, $E_\phi^m$, $\mathcal{M}$, $\mathrm{mIoU}$ and $\mathrm{mDice}$. Overall, the methods exhibit relatively higher performance on categories such as *Amphibia* and *Gastropoda*. This is because these categories are more distinguishable shape and their texture cues would be more perceptible. In contrast, all models perform worse on categories like *Actinopterygii* and *Malacostraca*, where the high visual similarity with the background significantly challenges detection. Notably, SLT-Net exhibits large performance fluctuations across different categories, indicating weaker generalization capability. In contrast, the other three models demonstrate more consistent trends across all metrics within the twelve categories.

**Attribute-based performances.** To further analyze the robustness of SOTA methods under various challenging conditions, we visualize the attribute-based performance of HGINet[35], CamoDiffusion[36], ZoomNeXt[25], and SLT-Net[5] across eight typical challenging attributes: multiple objects (MO), uncertain edge (UE), occlusion (OC), shape complexity (SC), out-of-view (OV), motion blur (MB), big object (BO), and small object (SO), as shown in Fig. 16. Overall, the four methods exhibit consistent performance trends across all attributes. Notably, sequences with *OV* and *BO* generally yield higher scores across all metrics, suggesting that the models are better at handling large or partially visible objects. In contrast, attributes like *MO*, *SC*, and *SO* are especially challenges. We also notice that compared to big objects (BO), small objects consistently yield lower scores across the metrics, emphasizing the difficulty of accurate segmentation in such cases.

**Failure cases.** As shown in Fig. 17, we perform a visual comparison of HGINet [35] and ZoomNeXt [25] in challenging scenarios with diverse backgrounds. In the first sequence (Row 1–4), both methods struggle to locate and identify the camouflaged object due to the highly confusing similarity between the object and its surrounding background. In the second sequence (Row 5–8), both methods tend to focus on visually salient regions, indicating that both methods may misguide by the salient regions and resulting in incorrect segmentation. Row 9–12 shows that the segmentation results remain fragmented and imprecise due to the highly similar color and texture patterns between the object and its surrounding background, although the camouflaged object is partially detected, indicating a lack of semantic understanding in both methods. Similar challenges are observed in the single-frame examples presented in Row 13–16. These results demonstrate the diversity and difficulty of our proposed CAMotion dataset, emphasizing its value as a benchmark for advancing research in video camouflaged object detection.

## D Limitation and Future Work

Despite significant advances in COD and VCOD, our experiments reveal a notable trade-off between camouflaged object discrimination and temporal consistency. Current image COD models, such as HGINet and ZoomNeXt (image version), demonstrate impressive spatial discrimination on static COD benchmarks, accurately identifying subtle texture and color differences. However, when applied to VCOD datasets, image COD models struggle to maintain consistent predictions across consecutive frames, even though the camouflaged objects are well-identified in the previous frame. Its per-frame

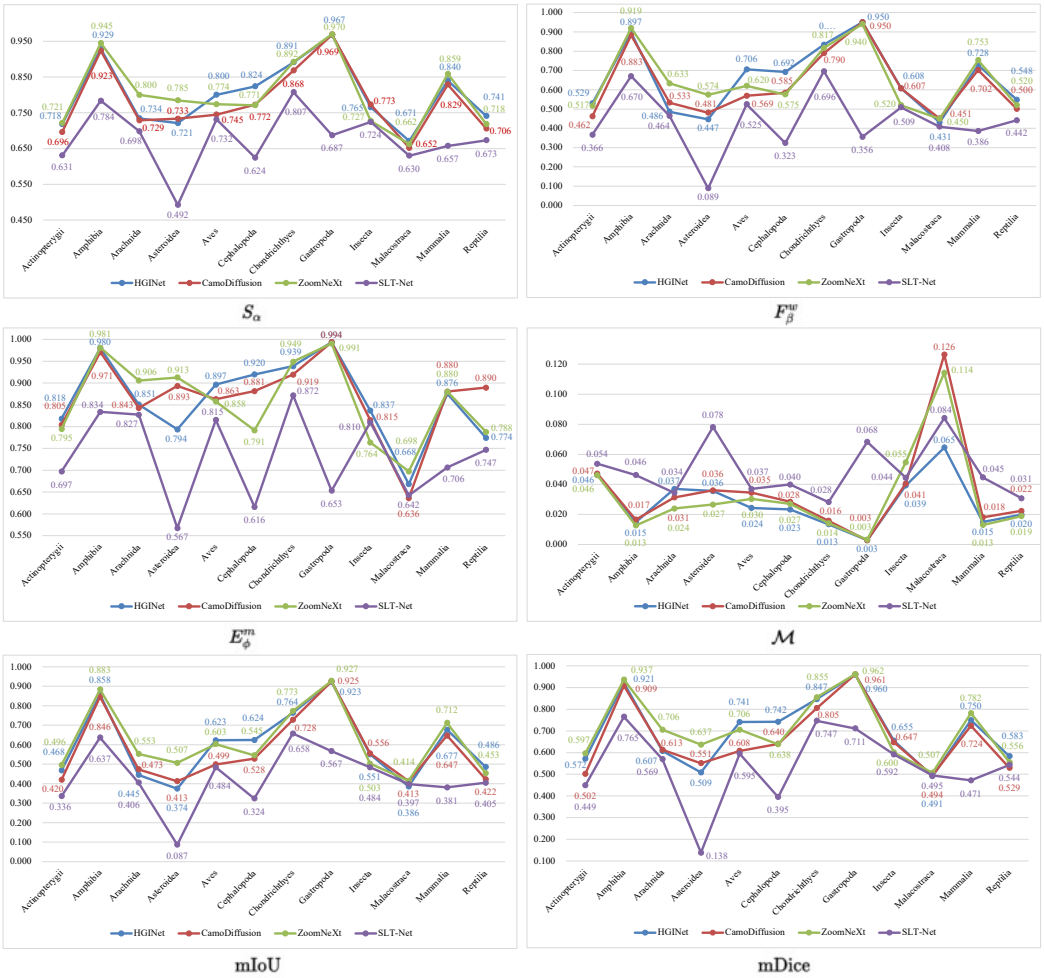



Figure 15: Visualization of SOTA method performances on different categories in terms of $S_\alpha \uparrow$, $F_\beta^w \uparrow$, $E_\phi^m \uparrow$, $\mathcal{M} \downarrow$, mIoU $\uparrow$ and mDice $\uparrow$. Please zoom in for details.

architecture fails to capture temporal cues. Conversely, temporal-aware methods like ZoomNeXt (video version) excel at capturing temporal cues. These models produce temporally coherent masks and handle occlusions or sudden camera motions more robustly. However, models like ZoomNeXt tend to sacrifice the spatial discrimination capability and fail to detect camouflaged objects that closely match the background. As results, no existing model can simultaneously maintain strong discrimination capability and temporal consistency. The static COD models ignore the temporal cues, while VCOD algorithms struggle with challenging camouflaged objects discrimination. Bridging this gap is essential for real-world applications, where both precise localization and stable tracking are required. To address these shortcomings, our future work seeks a new architecture that seamlessly integrates complementary strengths from both COD and VCOD approaches. By unifying camouflaged discrimination with temporal reasoning in a single end-to-end framework, we aim to set a new standard for VCOD, paving the way for practical COD applications in dynamic environments.

## E Potential Societal Impact

Our CAMotion dataset is developed to tackle the limitations of existing VCOD datasets in terms of scale and diversity, providing a rich and challenging benchmark for camouflaged motion object detection in the wild. By offering extensive category coverage and high-quality annotations, CAMotion can significantly accelerate research progress in video object detection, video understanding, and biological behavior analysis. However, the dataset also opens the door to potential misuse like illegal

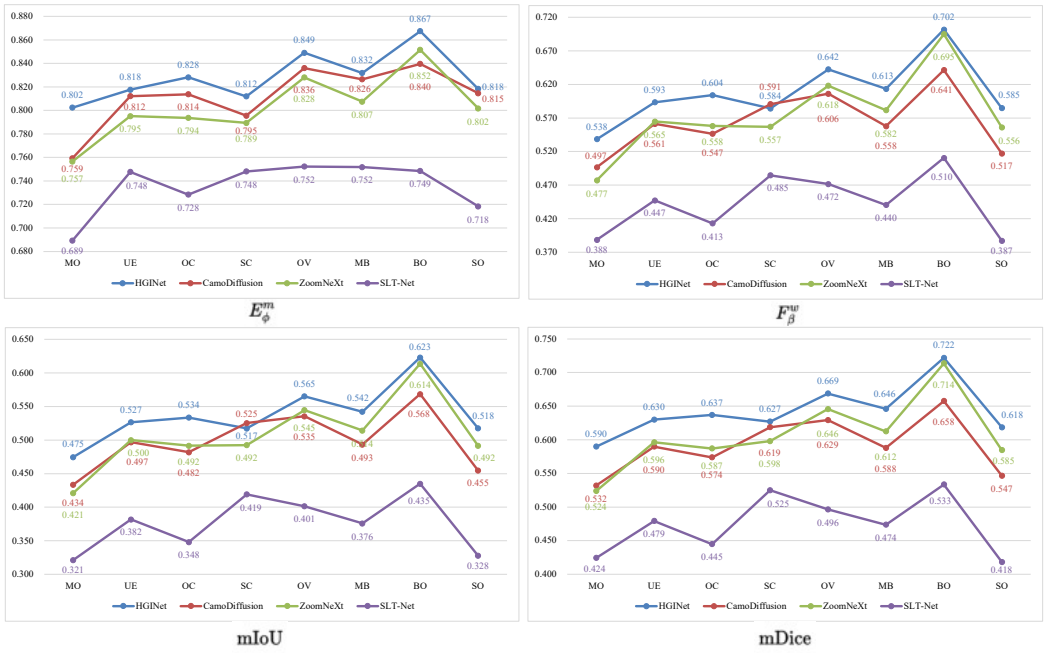



Figure 16: Visualization of SOTA method performances on different challenging attribute in terms of $E_\phi^m \uparrow$, $F_\beta^w \uparrow$, mIoU $\uparrow$ and mDice $\uparrow$. Please zoom in for details.

hunting or poaching, invasive monitoring of animal behavior in protected habitats, military reconnaissance and targeting, and other ethically concerning surveillance activities. Although CAMotion is intended for research and conservation-driven purposes, we acknowledge these risks and encourage the community to consider responsible usage in future research and applications.


Figure 17: Failure cases on both HGINet and ZoomNeXt. Row 1–12 shows failure cases on sequences, while Row 13–16 presents single-frame failure examples. Please zoom in for details.