International Journal of Computer Vision _####################_
https://doi.org/10.1007/s11263-026-02765-8

CamoVid60K: A Large-Scale Video Dataset for Moving Camouflaged Animals Understanding

Tuan-Anh Vu[1,2] · Ziqiang Zheng[1] · Chengyang Song[3] · Qing Guo[2,4] · Ivor W. Tsang[2] · Sai-Kit Yeung[1]

Received: 16 October 2025 / Accepted: 24 January 2026

Abstract

We have been witnessing remarkable success led by the power of neural networks driven by a significant scale of training data in handling various computer vision tasks. However, less attention has been paid to monitoring the camouflaged animals, the masters of hiding themselves in the background. Robust and precise segmentation of camouflaged animals is challenging even for domain experts due to their similarity to the environment. Although several efforts have been made in camouflaged animal image segmentation, to the best of our knowledge, limited work exists on camouflaged animal video understanding (CAVU). Biologists often prefer videos for monitoring and understanding animal behaviors, as videos provide redundant information and temporal consistency. However, the scarcity of labeled video data significantly hinders progress in this area. To address these challenges, we present **CamoVid60K**, a diverse, large-scale, and accurately annotated video dataset of camouflaged animals. This dataset comprises **218** videos with **62,774** finely annotated frames, covering **70** animal categories, which *surpasses* all previous datasets in terms of the number of videos/frames and species included. **CamoVid60K** also offers more diverse downstream tasks in computer vision, such as camouflaged animal classification, detection, and task-specific segmentation (semantic, referring, motion),*etc.*We have benchmarked several state-of-the-art algorithms on the proposed **CamoVid60K** dataset, and the experimental results provide valuable insights for future research directions. Our dataset serves as a novel and challenging benchmark to stimulate the development of more powerful camouflaged animal video segmentation algorithms, with substantial room for further improvement.

Keywords Moving camouflaged animal · Optical flow · Video object segmentation · Video object detection

1 Introduction

The continuous evolution of neural networks (*e.g.*, Convolutional Neural Networks (CNNs) (He et al., 2016) and Vision Transformers (ViTs) (Dosovitskiy et al., 2020)) has provided powerful and efficient tools for visual understanding based on captured images and videos. Enhancements in both *data* and *algorithm* have led to significant progress and success in the field. Large-scale datasets (*e.g.*, COCO (Lin et al., 2014), ADE20K (Zhou et al., 2017) and Object365 (Shao et al., 2019)) with supervised annotations serve as essential stimuli for developing powerful visual perception algorithms (Xie et al., 2022) and benchmarking them to reveal future research directions. However, most existing datasets mainly contain everyday objects (*e.g.*, 80 categories in COCO). This work focuses on camouflaged animals, a less explored area of research. In addition, monitoring and understanding camou-

Communicated by Urs Waldmann.

✉ Tuan-Anh Vu
tavu@connect.ust.hk

Ziqiang Zheng
zzhengaw@connect.ust.hk

Chengyang Song
songchengyang@stu.ouc.edu.cn

Qing Guo
tsingqguo@ieee.org

Ivor W. Tsang
ivor_tsang@cfar.a-star.edu.sg

Sai-Kit Yeung
saikit@ust.hk

[1] Hong Kong University of Science and Technology, Hong Kong, Hong Kong

[2] CFAR & IHPC, A*STAR, Singapore, Singapore

[3] Ocean University of China, Qingdao, China

[4] Nankai University, Tianjin, China

flaged animals is crucial for biodiversity conservation (Soofi et al., 2022; Rands et al., 2010), as it helps protect species that are otherwise difficult to detect and are at risk of unnoticed population declines. Furthermore, studying camouflaged animals provides insights into evolutionary biology and adaptation mechanisms, enriching our scientific understanding of natural selection. [1]

However, unlike everyday objects, collecting images and videos of camouflaged animals is more challenging, and annotation procedures usually involve domain experts. *Segmentation*, which involves generating precise masks for objects of interest, is a fundamental task in computer vision. Camouflaged animal segmentation helps accurately identify and isolate these animals from their backgrounds in images, facilitating detailed study and analysis. The resulting masks aid in gathering precise data on their behavior, habitat, and population dynamics, enhancing our overall understanding of their ecology (Lv et al., 2021; Troscianko et al., 2017). Recently, several efforts Xie et al. (2022); Cheng et al. (2022); Lamdouar et al. (2023); Vu et al. (2023) have been made to perform camouflaged animal segmentation. Specifically, camouflage is a powerful biological mechanism for avoiding detection and identification, making it more challenging to achieve precise segmentation. [2]

Various datasets (*e.g.*, CAMO-COCO Le et al. (2019), COD10K Fan et al. (2022), CAM-LDR Lv et al. (2023), S-COD He et al. (2023)) have been collected for image-level camouflaged animal segmentation. However, image-level camouflaged animal segmentation cannot fully satisfy biological monitoring and surveying purposes, where the activity and behavior (Yang et al., 2021) should be recorded. For video level, the MoCA dataset Lamdouar et al. (2020) is the most extensive compilation of videos featuring camouflaged objects, yet it only provides detection labels. We argue that bounding box annotations alone cannot adequately delineate camouflaged animals, especially those with irregular boundaries, poses, and patterns (*e.g.*, the transparent fins of fish). Furthermore, despite the shift from images to videos, the data annotations remain insufficient in both volume and accuracy for developing a reliable video understanding model capable of effectively handling complex camouflaged situations.

To fill this gap and advance camouflaged animal video understanding (CAVU) in real-world scenarios, we present **CamoVid60K**, a comprehensive video dataset dedicated to studying camouflaged animals. It comprises **218** videos with **62,774** finely annotated frames, covering **70** animal categories. Table 1 compares our proposed dataset with previous ones (CAD (Pia Bideau, 2016), MoCA (Lamdouar et al., 2020), MoCA-Mask (Cheng et al., 2022), MVK (Truong et al., 2023), WATB Wang et al. (2022), and Animal-Track Zhang et al. (2022)), showing that **CamoVid60K** *surpasses* all previous datasets in terms of the number of videos, frames, and species included. Unlike previous datasets that annotated every 5 frames, our dataset offers annotations for every single frame. Additionally, we provide **a wider variety of annotation types** (animal categories, bounding boxes, annotated masks, pseudo-label optical flow, referring expressions), making it a more effective benchmark for CAVU tasks. Our dataset supports **a broad range of downstream tasks**, as shown in Figure 1, including classification, detection, segmentation (semantic, referring, motion), and optical flow estimation, *etc.*

We propose baselines for each task and corresponding benchmarks to explore the capabilities of advanced algorithms in performing robust and precise video understanding. Our **CamoVid60K** serves as a novel and important testing set for both the computer vision and wildlife research communities.

Our main contributions are summarized as follows:

- We present a **large-scale** and **comprehensive** video dataset dedicated to the understanding of camouflaged animals, featuring **significantly more** data and annotation types than existing datasets.
- We propose a **simple pipeline** for camouflaged animal detection and segmentation that achieves performance comparable to state-of-the-art methods.
- We benchmark **various** camouflaged animal video understanding tasks, including image classification, object detection, and motion segmentation using several state-of-the-art models.

2 Related Works

2.1 Camouflaged Scene Understanding

Camouflaged scene understanding (CSU) is a hot computer vision topic aiming to learn discriminative features that can be used to discern camouflaged target objects from their surroundings (Fan et al., 2023). CSU tasks can be divided into image-level and video-level categories. Image-level CSU tasks include five main types: camouflaged object counting (Sun et al., 2023), camouflaged object localization (Lv et al., 2021, 2023), camouflaged object segmentation (Ji et al., 2023; He et al., 2023; Fan et al., 2022), camouflaged instance ranking (Lv et al., 2021, 2023), and camouflaged instance segmentation (Le et al., 2021; Pei et al., 2022). These tasks can be further categorized based on their semantic focus: object-level and instance-level. Object-level tasks focus on identifying objects, while instance-level tasks aim to differentiate various entities. Additionally, camouflaged object counting is considered a sparse prediction task due to its nature, while the other tasks are classified as dense prediction tasks. In addition, CSU video-level tasks include video camouflaged object segmentation (Ji et al., 2014; Cheng et al., 2022; Xie et al., 2019) and video camouflaged object detection (Kowal et al., 2022; Lamdouar et al., 2020; Meunier et al., 2022; Lamdouar et al., 2021; Xie et al., 2022; Yang et al., 2021). Overall, the progress of video-level CSU has been somewhat slower than image-level CSU, primarily because the process of collecting and labeling video data is labor-intensive and time-consuming.

2.2 Video Camouflaged Object Detection and Segmentation

We review two kinds of perception tasks for camouflaged animal videos: detection (Kowal et al., 2022; Lamdouar et al., 2020; Meunier et al., 2022; Lamdouar et al., 2021; Xie et al., 2022; Yang et al., 2021) and segmentation (Lamdouar et al., 2023; Ji et al., 2014; Cheng et al., 2022; Xie et al., 2019). The former video camouflaged object detection (VCOD) yields bounding box sequences for the camouflaged animals, while the latter video camouflaged object segmentation (VCOS) generates dense pixel-level masks. MoCA Lamdouar et al. (2020) proposed the first large-scale moving camouflaged animals video dataset, featuring bounding box annotations and additional optical flows to enhance the detection of camouflaged animals. Further work Lamdouar et al. (2021) incorporated visual appearance from a static scene as additional clues to promote the ability of the model to detect camouflaged animals. However, the bounding box annotations could not accurately describe camouflaged objects' pose, appearance, and patterns. To address this issue, Xie Xie et al. (2019) proposed a novel pixel-trajectory RNN to cluster foreground pixels and generate dense segmentation masks for object discovery in videos. MoCA-Mask Cheng et al. (2022) proposed the first large-scale dataset and benchmark with pixel-level handcrafted ground truth masks for camouflaged animal videos. However, MoCA-Mask provides bounding boxes and pixel-wise masks for **only every 5 frames**, totaling just 4,691 frames, which is insufficient for deep learning approaches. In contrast, our dataset offers annotations for **every frame**, resulting in 62,774 annotated frames (**13 times larger**). This substantial increase can significantly enhance the performance of various downstream tasks. Our dataset and benchmark pave the way for future exploration and a deeper understanding of camouflaged animal analysis.

3 CamoVid60K Dataset

Collecting video datasets of camouflaged animals is quite challenging, even without focusing on long-form videos. Manually collecting, observing, and annotating videos with multiple annotation types is labor-intensive, time-consuming, and expensive. In addition to these costs, ensuring visual data diversity and high-quality annotations adds to the dif-

Table 1 Comparison with existing video animal datasets. Class.: Classification Label, B.Box: Bounding Box, Motion: Motion of Animal, PseudoOF: Pseudo-label Optical Flow, Expres.: Referring Expression. The frequency of annotations refers to how often each frame is annotated. For instance, MoCA-Mask provides annotations for **every 5 frames**, resulting in only 4,691 annotated frames out of a total of 22,939 frames. In contrast, our CamoVid60K dataset offers a significantly larger volume of data with more frequent annotations and a wider variety of annotation types. ∗ ***Note that***, MVK Truong et al. (2023) dataset mostly consists of *normal* marine animals with only some camouflaged animals.

Dataset	Venue	# videos / frames	# species	Frequency	Class.	B.Box	Mask	Motion	PseudoOF	Expres.
CAD	ECCV'16	9 / 839	6	every 5 frames	✓		✓			
MoCA	ACCV'20	141 / 37,250	67	**every frames**	✓	✓		✓		
MoCA-Mask	CVPR'22	87 / 22,939	44	every 5 frames	✓		✓			
MVK∗	MMM'23	1379 / ∼ 992,880	-	every 30 frames	✓					✓
WATB	IJCV'23	206 / ∼ 203,000	-	**every frames**	✓	✓				
AnimalTrack	IJCV'23	58 / ∼ 247,000	-	**every frames**	✓	✓				
CamoVid60K	**-**	**218 / 62,774**	**70**	**every frames**	✓	✓	✓	✓	✓	✓

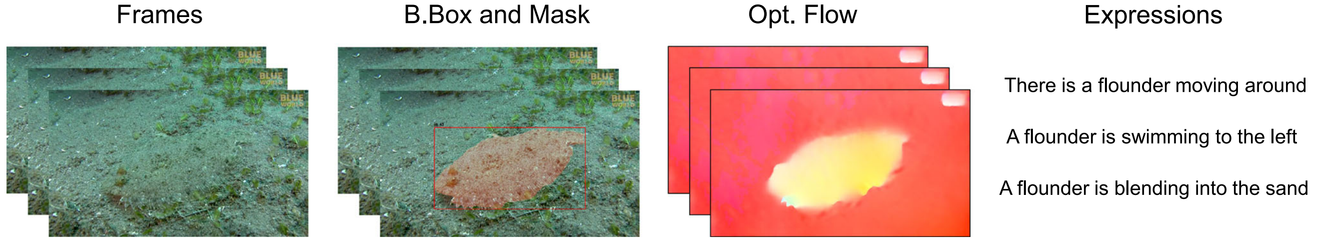

Frames
B.Box and Mask
Opt. Flow
Expressions
There is a flounder moving around
A flounder is swimming to the left
A flounder is blending into the sand

Fig. 1 Example from our proposed **CamoVid60K** dataset with bounding box, mask, pseudo-label optical flow, and referring expressions

ficulty. In this section, we propose a staged data collection and processing pipeline, as shown in Figure 2. Associated datasheets (Gebru et al., 2021) and data cards (Pushkarna et al., 2022) for our **CamoVid60K** dataset are provided in Appendix A.

3.1 Data Construction and Processing

3.1.1 Data Sources and Pre-Processing

We built our dataset by incorporating previous published datasets (Camouflaged Animals Dataset (CAD) (Pia Bideau, 2016), Moving Camouflaged Animals (MoCA) (Lamdouar et al., 2020), MoCA-Mask (Cheng et al., 2022), Marine Video Kit (MVK) (Truong et al., 2023)) and crawling additional videos from the internet to cover a variety of camouflaged animals.

The CAD dataset includes nine short video sequences obtained from YouTube videos. Hand-labeled ground truth masks are provided for every 5 frames.

The MoCA dataset comprises approximately 37,000 frames extracted from 141 YouTube video sequences. Most videos are presented at an image resolution of 1280 × 720 and 3840 × 2160 pixels, and the videos have a frame rate of 24 FPS. This dataset includes 67 distinct species of animals in locomotion within their native habitats, although it contains a few instances of animals with less camouflaged characteristics.

The MoCA-Mask dataset is built upon the MoCA dataset. This new subset consists of 87 video sequences with 22,939 frames. It offers human-labeled segmentation masks for every 5 frames. Consequently, the ground truth (GT) is available in two formats: 4,691 bounding box annotations and 4,691 pixel-level masks.

The MVK dataset comprises 1,379 underwater videos recorded at 36 unique geographical sites during various seasons. These videos exhibit a broad duration spectrum, ranging from as short as 2 seconds to almost 5 minutes, with a total duration slightly above 12 hours. On average, the videos are roughly 29.9 seconds long, with a median length of around 25.4 seconds. Notably, the dataset presents videos recorded under different conditions, such as variable light levels, points of view, water clarity, and environmental conditions. They also offer approximately 40,000 frames (extracted at one FPS or every 30 frames) with associated referring expression annotations.

To crawl videos from the internet, we curated a list of animal names that potentially have camouflage abilities. We then created a template for searching and downloading videos: *“video of camouflaged/concealed + animal’s name"*. By combining these with the videos from the above datasets, we initially collected 1,929 videos. We then manually checked and filtered out any blurry or irrelevant videos, retaining those with clear depictions of animals. Next, we extracted every frame of each video (instead of every 5 frames as proposed in existing datasets, see Table 1) before annotating them. At the end, our dataset comprises **218** videos with **62,774** frames of **70** animal species.

3.1.2 Bounding Box and Mask Annotation.

We utilized the annotation tool from (Zheng et al., 2023), which is heavily based on the Segment Anything Model (SAM) (Kirillov et al., 2023) for mask initialization and bounding box creation, and XMem (Cheng & Schwing, 2022) for mask and bounding box propagation. We then manually checked and refined every frame to provide accurate bounding boxes and segmentation masks. In addition, we adopted the perceptual camouflage score ($\mathcal{S}_p$) from (Lamdouar et al., 2023) to quantify the effectiveness of animals’ camouflage, *i.e.*, how successfully an animal blends into its background. Based on the perceptual camouflage score, we retained videos with a score higher than the threshold ($\mathcal{S}_p > 0.5$). Below, we explain how to compute the perceptual camouflage score $\mathcal{S}_p$:

$$\mathcal{S}_p = (1 - \alpha)\mathcal{S}_\mathcal{R} + \alpha\mathcal{S}_\mathcal{B} \quad (1)$$

where $\mathcal{S}_\mathcal{R}$ is the reconstruction fidelity score, $\mathcal{S}_\mathcal{B}$ is the boundary score, and α is the weighting parameter.

In detail, given an image $\mathcal{I}$ and a segmentation mask $m_\mathcal{S}$, the reconstruction fidelity score $\mathcal{S}_\mathcal{R}$ is computed by assessing the difference between the foreground region and its reconstruction. Specifically, it counts the number of foreground

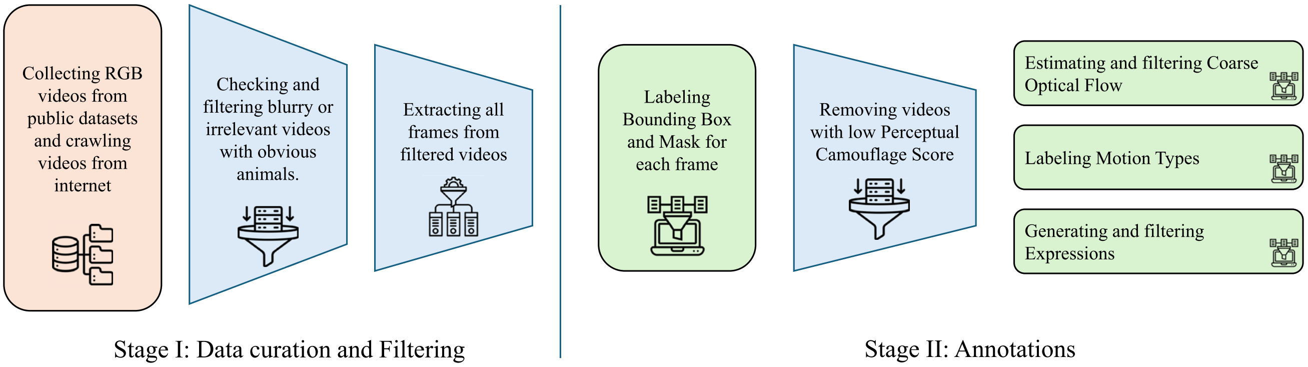

Collecting RGB videos from public datasets and crawling videos from internet
Checking and filtering blurry or irrelevant videos with obvious animals.
Extracting all frames from filtered videos
Labeling Bounding Box and Mask for each frame
Removing videos with low Perceptual Camouflage Score
Estimating and filtering Coarse Optical Flow
Labeling Motion Types
Generating and filtering Expressions
Stage I: Data curation and Filtering
Stage II: Annotations

Fig. 2 **CamoVid60K** data pipeline. Stage I includes data curation, filtering irrelevant videos, and extracting all frames. Stage II includes data annotation, generation, and filtering

pixels ($\mathcal{I}_{\text{fg}} = \mathcal{I} \odot \text{erode}(m_S)$) that have been successfully reconstructed from the background ($\mathcal{I}_{\text{bg}} = \mathcal{I} \odot (1 - \text{dilate}(m_S))$):

$$\mathcal{S}_{\mathcal{R}}(\mathcal{I}, m_S) = \frac{1}{N_{\text{fg}}} \sum_{(i,j)\in\mathcal{I}_{\text{fg}}} \mathcal{R}(i,j), \tag{2}$$

$$\mathcal{R}(i,j) = \begin{cases} 1, & \text{if } \left\|\mathcal{I}_{\text{fg}}(i,j) - \Psi_{\mathcal{I}_{\text{bg}}}(\mathcal{I}_{\text{fg}}(i,j))\right\|_2 < \lambda \left\|\mathcal{I}_{\text{fg}}(i,j)\right\|_2, \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where $\Psi_{\mathcal{I}_{\text{bg}}}(\cdot)$ denotes the reconstruction operation, $N_{\text{fg}} = |\text{erode}(m_S)|$ is the total number of pixels in the foreground region, and λ is a threshold.

Then, the boundary visibility score $\mathcal{S}_{\mathcal{B}}$ aims to measure the animal's boundary properties (or contour visibility) by penalizing the boundary pixels that are predicted as contours in both the image's contour ($\mathcal{C}$) and the ground truth animal's contour ($\mathcal{C}_{\text{gt}}$) using the F1 metric:

$$\mathcal{S}_{\mathcal{B}}(\mathcal{I}, m_S) = 1 - \text{F1}(m_b \odot \mathcal{C}_{\text{gt}},\ m_b \odot \mathcal{C}), \tag{4}$$

where $m_b = \text{dilate}(m_S) - \text{erode}(m_S)$.

We used the same parameter values as in (Lamdouar et al., 2023), specifically $\alpha = 0.35$ and $\lambda = 0.2$. In addition, we illustrate the difference between low-ranking and high-ranking camouflage in Figure 3 and Figure 4.

Note that, due to the nature and characteristics of camouflaged animals and also the low resolution of videos, some frames or videos may contain errors or mislabeling at the boundaries between animals and the background. We will continue improving the quality of the mask annotations and also provide rotated bounding boxes (RBbox) in the next version. RBbox excels over traditional axis-aligned bounding boxes in three main areas: better localization (accurate fit for elongated and rotated objects), reduced overlap of different objects or instances, and improved isolation of objects (capturing the proper aspect ratio and containing fewer background pixels).

3.1.3 Pseudo-label Optical Flow Annotation.

Previous optical flow datasets, such as Flying Chairs (Dosovitskiy et al., 2015), KITTI (Menze & Geiger, 2015), Sintel (Butler et al., 2012), and FlyingThings3D (Mayer et al., 2016), utilized either simulation software or real images with additional heavy sensor information (depth, LiDAR, *etc.*) and algorithms to create optical flow ground truth. This process is time-consuming and requires significant effort. Recently, with the development of deep learning techniques, many methods (Wang et al., 2023; Teed & Deng, 2020) can produce accurately estimated optical flow. Therefore, we utilized these methods for our pseudo-label optical flow annotation, using the algorithm shown in Algorithm 1. We used the pre-trained model of RAFT on FlyingThings3D (Mayer et al., 2016) and the pre-trained DINO model of ViT-B architecture.

Note that, even though our processing pipeline for optical flow annotation produces relatively accurate and dense optical flow, it is still **estimated** optical flow. Therefore, we provide pseudo-flow **as an auxiliary motion cue** (e.g., as additional input to motion segmentation) rather than as ground-truth supervision, and we **do not** use pseudo-flow as the target for any primary benchmark evaluation. Instead, all benchmark metrics reported in this paper are computed against **human annotations** (e.g., masks/boxes/motion labels) following our evaluation protocol.

3.1.4 Motion Annotation.

Following Lamdouar et al. (2020), we manually labeled our dataset according to the types of motion, as shown below. Based on these motion types, we can further annotate the camouflage methods of animals (concealing coloration, disruptive coloration, disguise, mimicry, transparency, and

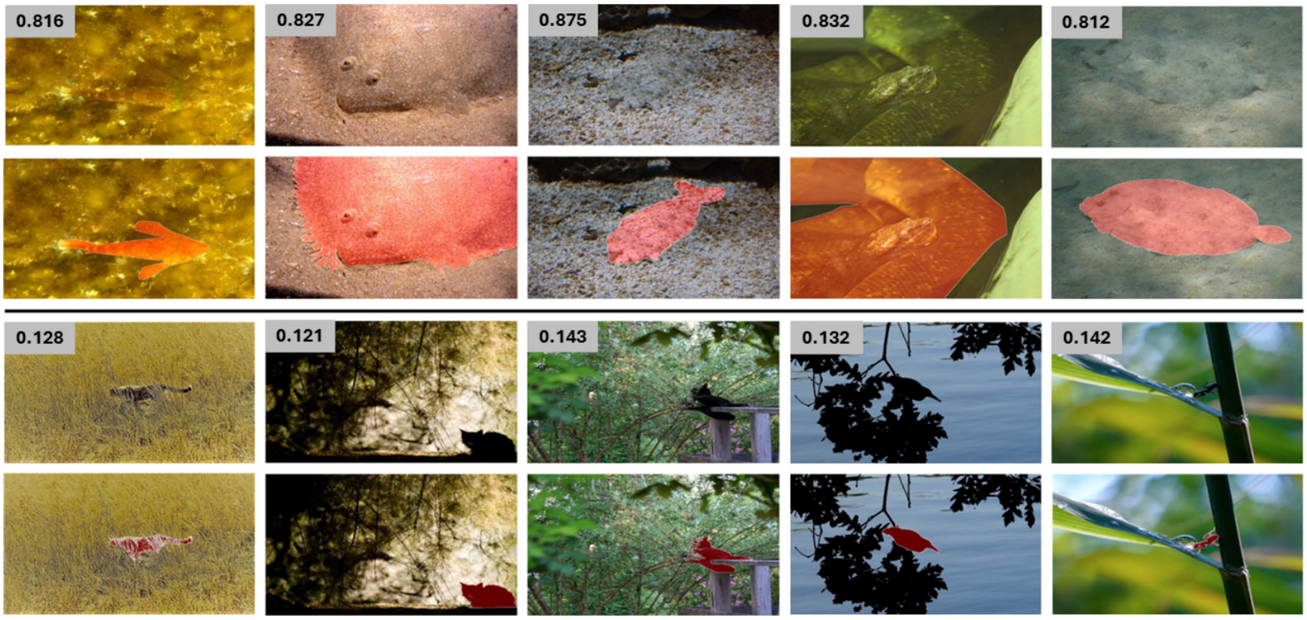


Fig. 3 The example of low-ranking and high-ranking camouflage of a single frame

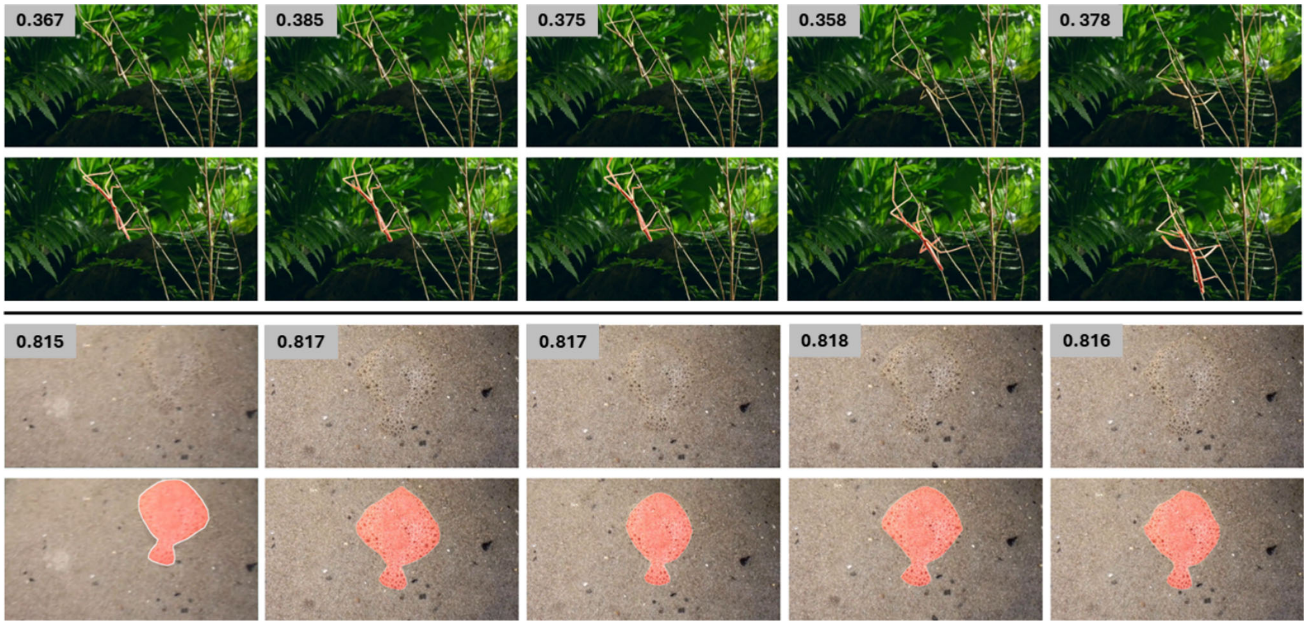


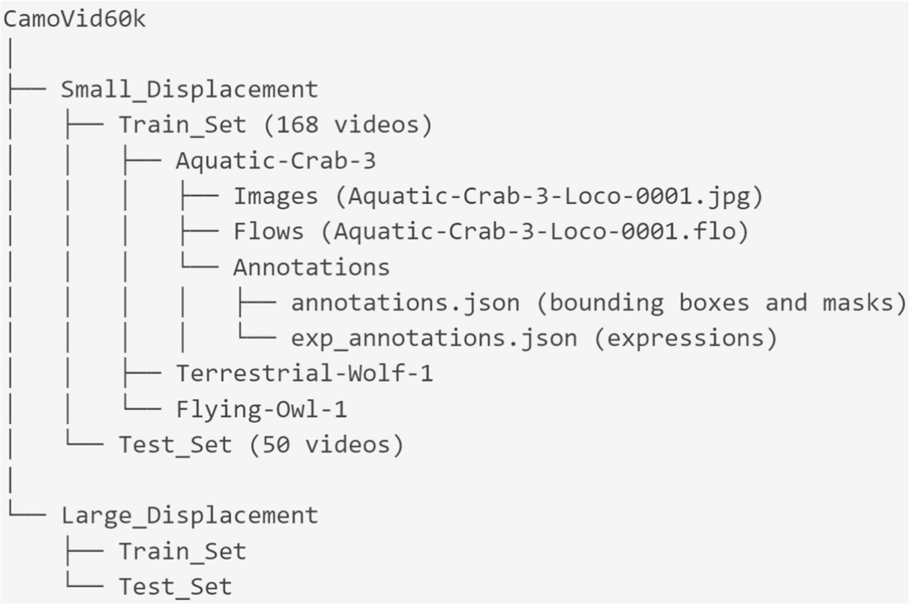
Fig. 4 The example of low-ranking and high-ranking camouflage of consecutive frames

counter-shading), which we plan to provide in the next version.

- *Locomotion*: when the animal makes movements that significantly change its location.
- *Deformation*: when the animal engages in more subtle movements that only change its pose while remaining in the same location.
- *Still*: when the animal remains stationary.

3.1.5 Referring Expression Annotation.

Referring expression annotations are used for the Referring Video Object Segmentation (RVOS) task. RVOS differs from traditional Video Object Segmentation (VOS), where a mask is provided for the first frame, and the model predicts the segmentation for the remaining video frames. In RVOS, the initial frame mask is replaced with a referring expression (*i.e.*a sentence) that accurately describes the target object throughout the entire video, *e.g.* "*the yellow fish swim-*



```
CamoVid60k
│
├── Small_Displacement
│   ├── Train_Set (168 videos)
│   │   ├── Aquatic-Crab-3
│   │   │   ├── Images (Aquatic-Crab-3-Loco-0001.jpg)
│   │   │   ├── Flows (Aquatic-Crab-3-Loco-0001.flo)
│   │   │   └── Annotations
│   │   │   │   ├── annotations.json (bounding boxes and masks)
│   │   │   │   └── exp_annotations.json (expressions)
│   │   ├── Terrestrial-Wolf-1
│   │   └── Flying-Owl-1
│   └── Test_Set (50 videos)
│
└── Large_Displacement
    ├── Train_Set
    └── Test_Set
```

Fig. 5 **Data organization** of our dataset. It includes a small and a large displacement subset

Algorithm 1 Optical Flow Computation and Filtering

```
Require: Sequence of frames
Ensure: Sequence of computed optical flows
for each pair of frames $(i, j)$ do
    Compute all pairwise optical flows using RAFT (Teed & Deng, 2020)
    Compute DINO features (Oquab et al., 2024; Caron et al., 2021) for each frame
    Filter flows using cycle consistency and appearance consistency check
    Apply chain cycle consistent correspondences to create denser correspondences
end for
```

ming toward the camera." This approach also differs from Referring Image Segmentation (RIS), which uses different referring expressions for each image. Referring expression annotations can be utilized for various video understanding tasks, such as RVOS (Seo et al., 2020; Yang et al., 2024), video retrieval systems with semantic understanding (Ha et al., 2023), video grounding (Mu et al., 2024), *etc.*

To generate referring expressions, we first employed GPT-4V (xxxx, 2023) to produce concise captions (within 30 words) describing the target object for each frame. We observed reduced caption accuracy for aquatic animals; therefore, for aquatic videos, we used MarineGPT (Zheng et al., 2023), a vision-language model designed for the marine domain, to generate the initial captions. From these frame-level captions, we formed multiple candidate referring expressions for each video sequence and then applied a human verification and refinement process to obtain the final annotations.

Specifically, two validators were presented with each video and its candidate expressions, and independently ranked the candidates based on their relevance and utility for RVOS. The ranking criteria emphasized (i) correctness with respect to the target object, and (ii) discriminativeness via spatiotemporal cues, such as the target's relative position (e.g., moving left/right or up/down) and motion pattern (e.g., swimming, approaching, blending into the background). We then selected the final three expressions per video by taking the top-ranked candidates agreed upon by both validators (e.g., overlapping selections such as #1/#3/#4 from two ranked lists). Finally, we refined the selected expressions to improve clarity and directness (e.g., resolving ambiguous references and adding the object category name when identifiable) to reduce linguistic ambiguity, given the additional difficulty posed by camouflaged targets. Video instances whose targets could not be reliably localized using language were removed from the referring-expression subset.

3.2 Dataset Specifications and Statistics

3.2.1 Data Organization.

As shown in Figure 5, we split our dataset into two subsets based on the degree of displacement between frames: small displacement (every single frame) and large displacement (every 5 frames). This division is beneficial for evaluating motion segmentation methods, as it provides a robust framework for analyzing algorithms' performance under varying motion and displacement conditions. Each subset includes training and testing sets with images, pre-computed optical flows, and annotations. Importantly, the train/test split is performed at the **video level** to prevent data leakage. Specifically, our dataset contains **218 videos** in total, partitioned into **168 videos** for training and **50 videos** for testing, and

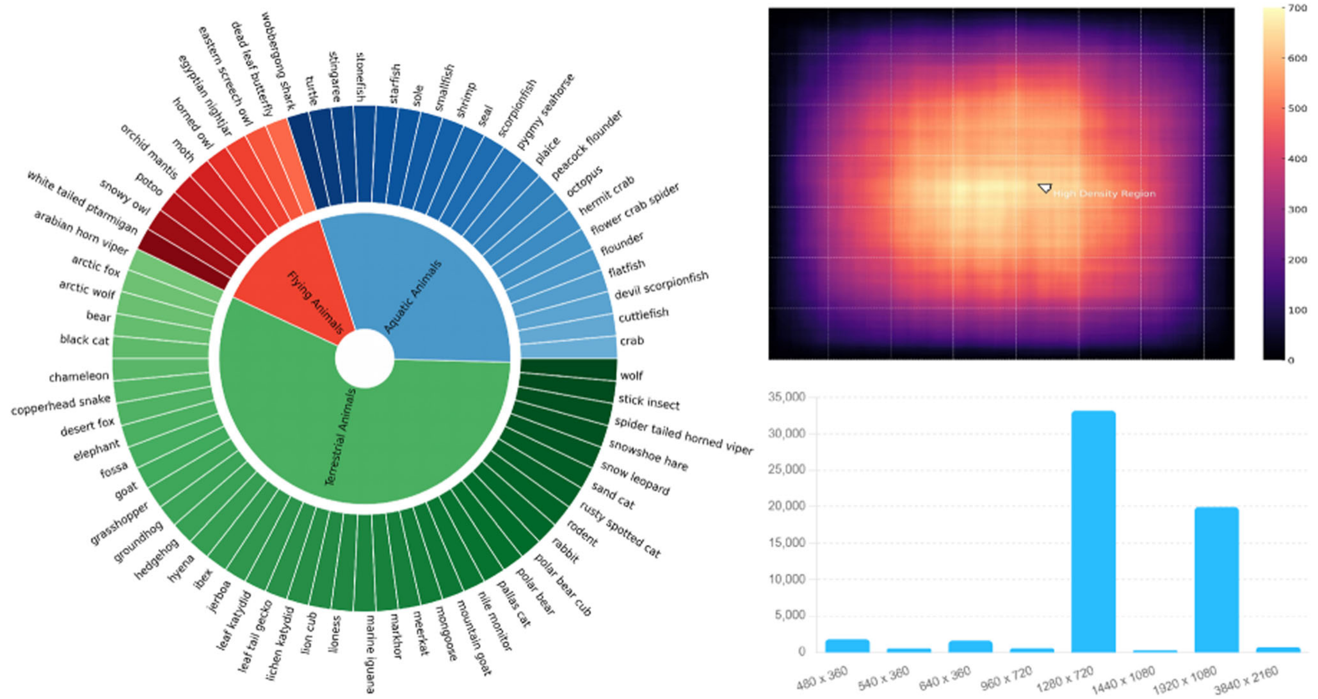


Fig. 6 **Left:** Taxonomic structure of our dataset by their biology-inspired hierarchical categorization. It encompasses various animals, spanning 70 categories across flying, terrestrial, and aquatic groups. **Right-Top:** Spatial distribution of animals' position based on bounding box. It reveals that annotations are more densely concentrated in the central region, while there is a comparatively lower density of annotations towards the edges. **Right-Bottom:** The distribution of our CamoVid60K dataset w.r.t resolution ranging from 480×360 to 3840×2160

the split is **video-disjoint**, *i.e.*no original video appears in both sets. The large-displacement subset is obtained by **temporal subsampling** frames from the same original videos; thus, it inherits the same video-disjoint train/test partition, ensuring that frames from a test video cannot appear in the training set even across the two subsets. The dataset covers **70 species**, and both the training and testing sets include examples from the **flying, terrestrial, and aquatic** superclasses. We name each image using the following format: "`SuperClass-SubClass-SubNumber-MotionType`-FrameNumber". This systematic naming convention ensures clarity and ease of reference within the dataset.

3.2.2 Dataset Features and Statistics.

We now discuss the proposed dataset and provide some statistics.

- *Category diversity:* The distributions of camouflaged animals, categorized hierarchically based on biology within three supergroups (flying, terrestrial, and aquatic), are visually represented through taxonomic structures in Figure 6 (Left) and word clouds in Figure 7. Subsequently, we describe the 70 prevalent subclass groups derived from our collected data in Figure 8.
- *Spatial distribution of animals' positions:* Figure 6 (Right-Top) and Figure 7 (Bottom) showcase examples with different animal positions and present the total sum of normalized bounding boxes across the entire dataset.
- *Resolution distribution:* Using high-resolution data is beneficial as it offers more detailed object boundary information for model training, thereby improving performance during testing (Zeng et al., 2019). In Figure 6 (Right-Bottom), the resolution distribution of **CamoVid60K** is displayed, highlighting the inclusion of numerous HD (720p) and Full HD (1080p) resolution videos.(see Fig. 9)

3.2.3 Evaluation Protocol.

Our dataset supports a broad range of downstream tasks. Therefore, we will evaluate each task using different metrics.

- *Motion Segmentation:* we adopt the same metrics as in (Cheng et al., 2022) to assess the pixel-wise masks: Mean Absolute Error (M), Enhanced-alignment measure (E_ϕ) (Fan et al., 2018), Structure-measure (S_α) (Fan et al., 2017), Weighted F-measure (F_β^w) (Margolin et al.,

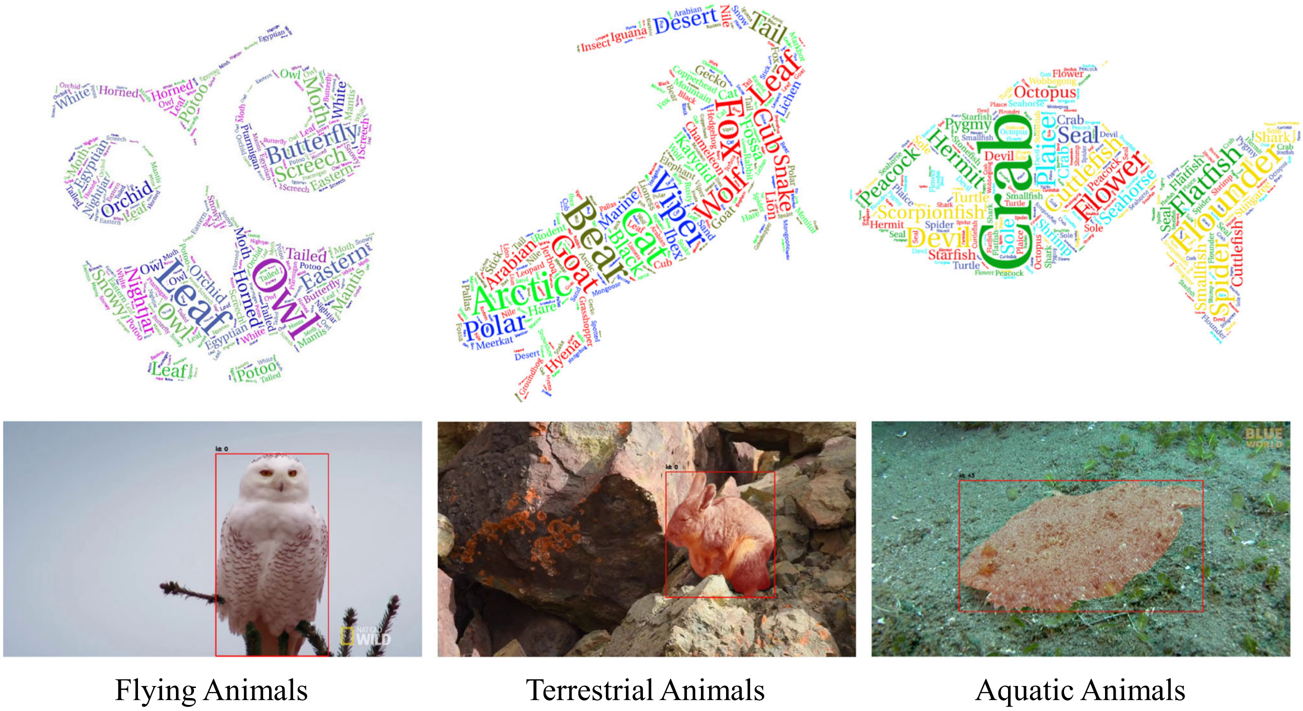


Fig. 7 Word cloud of category distribution of camouflaged animals with corresponding examples showing bounding box, segmentation mask (bottom)

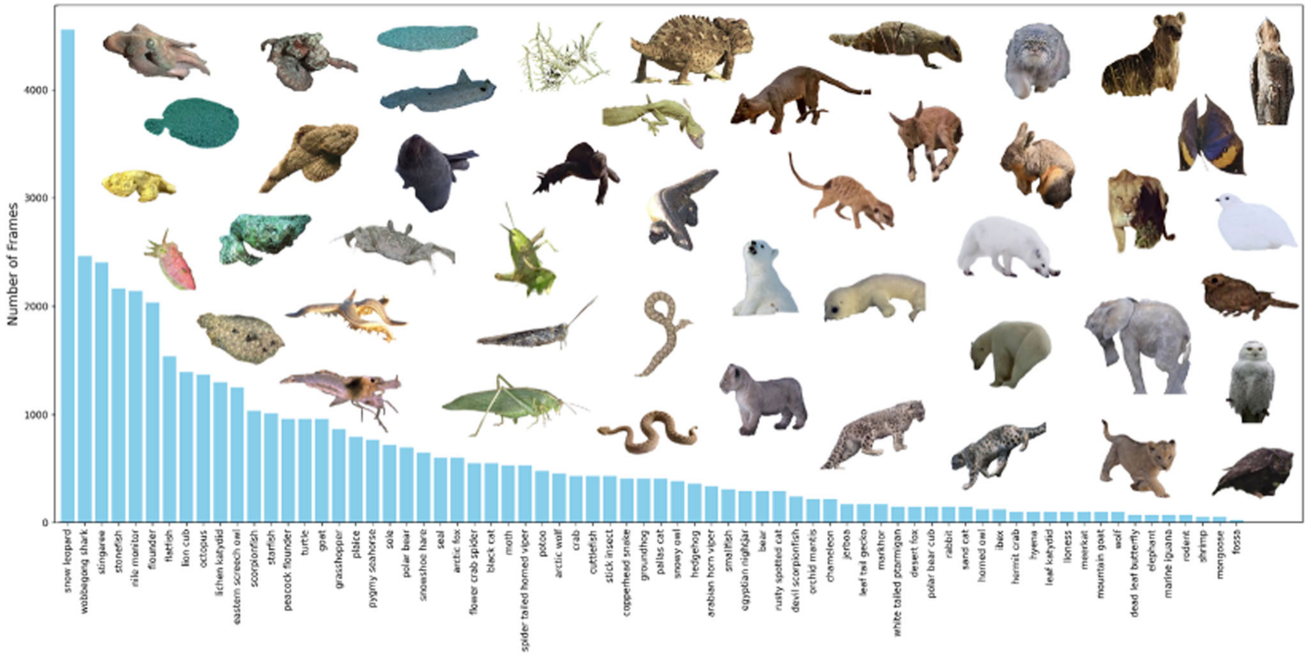


Fig. 8 Category distribution (ranging from 100 to 4,500 frames) and some visual examples (extracted animal masks) of our dataset. The variety ensures a wide range of camouflaged animals, allowing for comprehensive evaluation across various scenarios

2014), mean Intersection Over Union (mIoU), mean Dice (mDic).

- *Object Detection:* we use the mean Average Precision (mAP).
- *Image Classification:* we use the mean Accuracy (mAcc).
- *Referring Segmentation:* we utilize the mIoU, region similarity $\mathcal{J}$ and contour accuracy $\mathcal{F}$, and their average $\mathcal{J}\&\mathcal{F}$ for video object segmentation.

4 A simple pipeline to discern camouflaged animals

After constructing the dataset, we propose a simple pipeline based on the Mask2Former architecture (Lamdouar et al., 2023; Cheng et al., 2022) for both object detection and motion segmentation tasks. The goal of this baseline is to provide a **transparent starting point** and a consistent implementation for benchmarking on CamoVid60K, rather than to propose a heavily engineered state-of-the-art model. In our case, we directly use the refined optical flow provided in our dataset instead of utilizing the RAFT method (Teed & Deng, 2020) to estimate raw optical flow, as done in (Lamdouar et al., 2023). The images and associated estimated flows are passed into two separate encoders for feature extraction. Subsequently, the image and flow features at each timestamp are aggregated before being fed into the decoder to predict the segmentation mask. While this design is intentionally lightweight, an important direction for future work is to replace precomputed flow with learnable motion representations trained end-to-end, which may further improve both practicality and performance.

4.1 Visual Encoder.

We adopt the SINet-v2 (Fan et al., 2022) architecture, which takes an RGB sequence as input $I^v = \{I_1^v, I_2^v, \ldots, I_n^v\} \in \mathbb{R}^{n \times d_v \times h \times w}$, where n is the number of frames, d_v is the dimension of each frame, and h and w are the height and width, respectively. The visual encoder outputs visual features $\{f_1^v, f_2^v, \ldots, f_n^v\} = \Phi_{\text{visual}}(I^v)$.

4.2 Motion Encoder.

Inspired by the motion segmentation architecture (Lamdouar et al., 2021), we use a lightweight ConvNet that takes as input a sequence of optical flows $I^f = \{I_1^f, I_2^f, \ldots, I_n^f\} \in \mathbb{R}^{n \times d_f \times h \times w}$, where d_f is the dimension of the flow field, and outputs motion features $\{f_1^m, f_2^m, \ldots, f_n^m\} = \Phi_{\text{motion}}(I^f)$. We then concatenate the motion features with learned spatial and temporal positional encodings to produce a set of enriched motion features.

4.3 Decoder.

We adopt the Mask2Former (Cheng et al., 2022) architecture, which includes Transformer and Pixel decoders. The Transformer decoder combines a trainable query for mask embedding with the outputs of the motion encoder and visual features. Similar to Mask2Former, this query attends to multi-scale motion features and visual features, resulting in mask embedding for the moving object. Additionally, similar to the pixel decoder in Mask2Former, a ConvNet decoder with low computational complexity uses skip connections to generate high-resolution segmentation masks and bounding boxes from the motion features and mask embedding.

4.4 Training and Loss.

To optimize our pipeline, we utilize the L1 loss for bounding box regression, cross-entropy loss for the confidence score, and binary cross-entropy (BCE) loss for motion segmentation. The total loss for training our pipeline is defined as follows:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{L1}} + \mathcal{L}_{\text{CE}}, \tag{5}$$

where $\mathcal{L}_{\text{BCE}}$ is the binary cross-entropy loss for motion segmentation, $\mathcal{L}_{\text{L1}}$ is the L1 loss for bounding box regression, and $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss for the confidence score.

5 Experiments

This section introduces the baselines and training details for each task. We thoroughly analyze each task in our experiments and discuss the effectiveness of each method, including ours.

5.1 Baselines

For the motion segmentation task, we selected recent state-of-the-art (SOTA) methods for comparison, including two frame-based methods (PraNet (Fan et al., 2020) and SINet-v2 (Fan et al., 2022)) and two video-based methods (MG (Yang et al., 2021) and SLT-Net (Cheng et al., 2022)). For a fair comparison, we utilized the implementations provided by the authors and trained all methods using the same training set.

For the object detection task, we compared our approach with three well-known detection methods: Faster R-CNN (Ren et al., 2015), DETR (Carion et al., 2020), and DINO (Zhang et al., 2023). We followed the 1× (12-epoch) training setting and used the same ResNet50 (He et al., 2016) backbone for all methods.

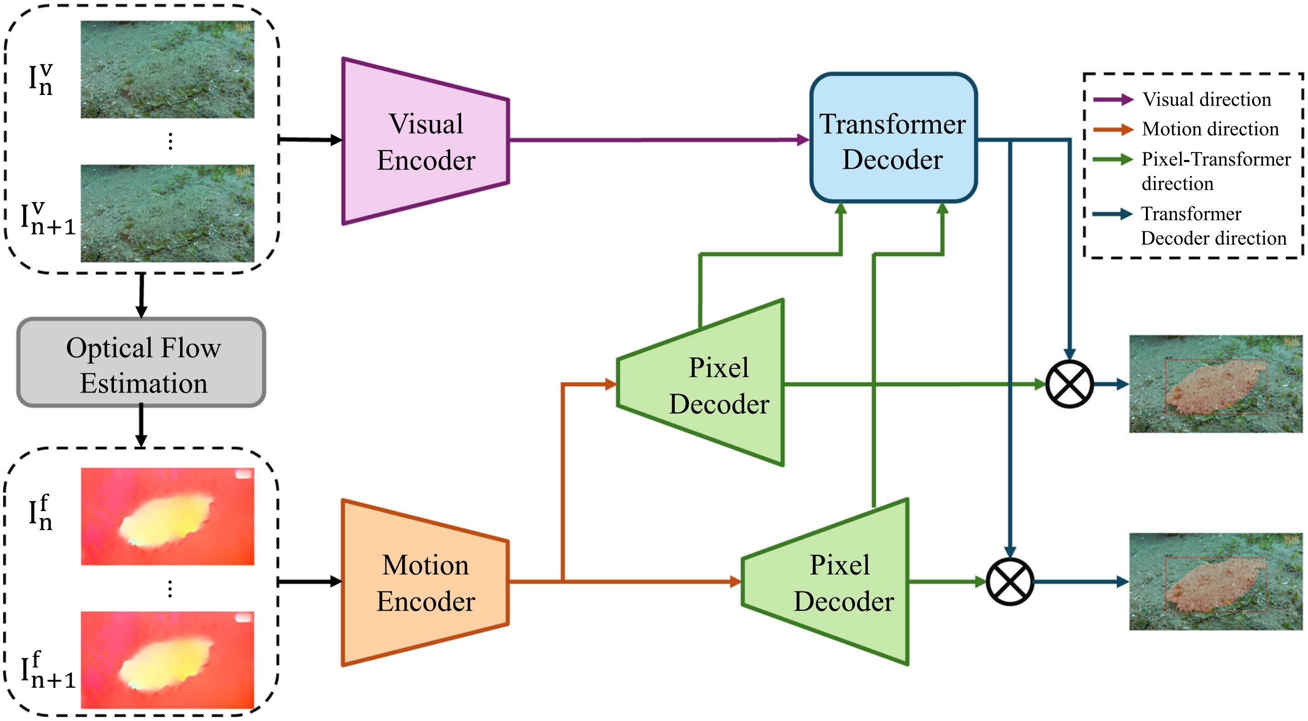


Fig. 9 Our simple pipeline takes a sequence of images (or a video) and the associated pre-computed optical flow (provided in our dataset) as input. They are fed into separate encoders for feature extraction. Then, the motion features with spatial and temporal positional encoding are passed to Pixel Decoders to produce a set of enriched motion features. Next, the Transformer Decoder takes the visual features and enriched motion features to produce mask embedding for the moving object and bounding box

Table 2 Quantitative results of motion segmentation on our CamoVid60K dataset. Our simple pipeline achieves performance comparable to that of other SOTAs on certain metrics.

Methods		$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$MAE \downarrow$	mDice ↑	mIoU ↑
Image–based	PraNet (Fan et al., 2020)	0.526	0.161	0.547	0.045	0.198	0.144
	SINet-v2 (Fan et al., 2022)	0.529	0.166	0.553	0.042	0.206	0.149
	FSEL (Sun et al., 2024)	0.542	0.184	0.565	0.048	0.213	0.154
	Ours (RGB-only)	0.536	0.178	0.557	0.046	0.206	0.151
Video–based	MG (Yang et al., 2021)	0.522	0.153	0.541	0.043	0.197	0.141
	SLT-Net (Cheng et al., 2022)	0.576	0.253	0.591	0.039	0.268	0.249
	EMIP (Zhang et al., 2025)	0.587	0.269	0.598	0.033	0.281	0.254
	Ours	0.566	0.249	0.589	0.041	0.270	0.252

Table 3 Quantitative results of object detection on our CamoVid60K dataset.

Methods	F-RCNN (Ren et al., 2015)	DETR (Carion et al., 2020)	DINO (Zhang et al., 2023)	RT-DETR (Zhao et al., 2024)	Ours
$mAP \uparrow$	28.72	37.56	39.84	40.95	38.39

For the zero-shot image classification task, we tested three recent methods: CLIP (Radford et al., 2021), UniCL (Yang et al., 2022), and K-LITE (Shen et al., 2022). We used the Swin-T model for both UniCL and K-LITE (pre-trained on the ImageNet-21K dataset (Deng et al., 2009)) and the ViT-B/32 pre-trained model from OpenAI's CLIP.

All methods were trained and tested on the same NVIDIA RTX 3090 GPU, except for the pre-trained models used in the zero-shot image classification task, where we utilized the pre-trained models provided by the authors. For our ablation studies, we utilized a small subset of the dataset to evaluate the impact of various components, thereby facilitating a rapid assessment of computational efficiency.

5.2 Metrics

Following previous methods (Cheng et al., 2022; Fan et al., 2022), we evaluate pixel-level masks and bounding box using the following metrics:

- MAE (MAE), which quantifies the absolute per-pixel discrepancy between predictions and ground-truth masks.
- Enhanced-alignment measure (E_ϕ) Fan et al. (2018), which jointly reflects pixel-wise correspondence and image-level statistics; it is well suited to assessing both global and local accuracy in camouflaged object detection. We report the mean E_ϕ in our experiments.
- S-measure (S_α) Fan et al. (2017), capturing region-aware and object-aware structural similarity.
- Weighted F-measure F_β^w Margolin et al. (2014), which typically offers more reliable evaluation than the conventional F_β.
- mean Dice ($mDice$), measuring similarity between two sets.
- meanIoU ($mIOU$), measuring the overlap between two masks.
- mean Average Precision (mAP), is the average of the AP scores across all classes, and AP is a key metric for evaluating object detection models, calculated as the area under the precision-recall curve for a single class (rather than focusing on object proposal proxy metrics).

5.3 Benchmarks and Discussions

5.3.1 Comparison with Image-Based and Video-Based Motion Segmentation Methods.

Table 2 compares the performance of our method with other approaches. Compared to image-based methods, our approach demonstrates significantly superior performance due to the incorporation of temporal information. When evaluated against video-based methods, our approach also surpasses MG (Yang et al., 2021), which relies solely on estimated optical flows as input. However, compared to the recent state-of-the-art method SLT-Net (Cheng et al., 2022) and EMIP (Zhang et al., 2025), our method performs worse on certain metrics. This is because SLT-Net (Cheng et al., 2022) excels at modeling both short-term dynamics and long-term temporal consistency from videos, allowing for joint optimization of motion and camouflaged object segmentation through a single optimization target. While EMIP (Zhang et al., 2025) utilizes intermediate features from one stream as interactive prompts to incorporate additional information into the other stream, it simultaneously conducts camouflaged segmentation and optical flow estimation.

Table 4 Ablation study on the impact of flow information on our method.

	no OF	raw OF	refined OF
mIoU	28.34	32.16	**32.81**

5.3.2 Comparison with Object Detection Methods.

As shown in Table 3, our proposed model demonstrates performance comparable to other specialized methods, owing to its dual capabilities in object detection and motion segmentation. Specifically, our method significantly outperforms conventional CNN-based methods. This advantage stems from dual optimizations in the detection and segmentation streams, along with the integration of additional optical flow information. However, when compared to DETR-like methods (Zhang et al., 2023; Carion et al., 2020), our approach shows mixed results. It surpasses the standard DETR model (Carion et al., 2020), yet falls short of DINO (Zhang et al., 2023) and RT-DETR (Zhao et al., 2024), advanced variants of DETR. DINO (Zhang et al., 2023) enhances performance through several innovative techniques: it employs contrastive denoising training to refine one-to-one matching, a mixed query selection method to better initialize the queries, and a ‘look forward twice’ method that utilizes gradients from subsequent layers to adjust parameters more accurately. RT-DETR (Zhao et al., 2024) introduces two main improvements: a hybrid encoder that efficiently handles multiscale features, and a query selection method that reduces uncertainty, enhancing the quality of the initial object queries.

5.3.3 Additional Analysis and Discussions.

As shown in Table 4, optical flow plays a crucial role in the motion segmentation of camouflaged animals. By analyzing the motion vectors between frames, optical flow can detect subtle movements, distinguishing moving animals from static backgrounds. This capability is particularly useful in identifying the slight movements of camouflaged animals.

State-of-the-art methods, including foundation models trained on large datasets such as CLIP (Radford et al., 2021), UniCL (Yang et al., 2022), and K-LITE (Shen et al., 2022), struggle with zero-shot image classification of camouflaged animals, as shown in Table 5. This is due to the subtle and complex patterns of camouflaged animals, the lack of specific training data, and the difficulty in generalizing across different backgrounds and lighting conditions. Improving these methods involves curating specialized training data (or fine-

Table 5 Zero-shot Image Classification performance on our CamoVid60K dataset.

	CLIP (Radford et al., 2021)	UniCL (Yang et al., 2022)	K-LITE (Shen et al., 2022)
mAcc	30.06	30.89	31.44

tuning on our dataset), using enhanced techniques like data augmentation, few-shot learning, and developing context-aware models.

6 Conclusion

In this paper, we introduced **CamoVid60K**, a large-scale video dataset for camouflaged animal understanding, aiming to foster further research on camouflaged animals. This dataset provides a significant benchmark for camouflaged animal video understanding tasks, enabling the evaluation of various algorithms and methods. We also plan to scale up our dataset and utilize it to build foundational models for studying camouflaged animals.

6.1 Limitations and Future Work.

As mentioned in Section 3, the annotation quality in some cases is suboptimal. We plan to enhance these annotations and introduce more types of annotations in the future. Additionally, our current pipeline requires images and pre-computed optical flow as inputs, which restricts the generation of new data due to the necessity of pre-computed optical flow. To address this limitation, we will propose a learnable module to estimate the implicit optical flow field.

6.2 New Benchmark.

Our **CamoVid60K** dataset is a diverse and comprehensive benchmark curated from publicly accessible datasets and the internet to enhance the assessment and exploration of camouflaged animal understanding. It includes various camouflaged animals across different environments, providing a robust framework for testing and developing new models.

6.3 Impact on Animal Studies.

By providing detailed and varied data on camouflaged animals, the **CamoVid60K** dataset significantly contributes to studying animal behavior, ecology, and evolution. Researchers can utilize this dataset to explore how different species employ camouflage in their natural habitats, leading to deeper insights into predator-prey interactions and survival strategies. Furthermore, this dataset can aid conservation efforts by improving the detection and monitoring of endangered species in their natural environments (Beery et al., 2018; Norouzzadeh et al., 2018; Simões et al., 2023; Troscianko et al., 2017).

6.4 Broader Impact.

The study of camouflaged objects has several important applications, such as identifying and safeguarding rare animal species, preventing wildlife trafficking, detecting medical conditions like polyps or lung infections, and aiding in search-and-rescue operations. Our dataset deliberately excludes any military or sensitive scenes, ensuring its focus remains on benign and beneficial applications. Besides the significant applications mentioned, our work advances the understanding of video content in the presence of distorted motion information, contributing to the broader field of video analysis and computer vision.

6.5 Licenses.

We built our dataset from previous datasets and crawled online videos. Therefore, we will follow their Terms of Use or Licenses (MoCA, MVK) for our dataset, which is under the CC-BY-4.0 license. The copyright remains with the original owners of the videos. In addition, the dataset shall be used only for non-commercial research and educational purposes.

A CamoVid60K Datasheet

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.
There are some studies about camouflaged animal segmentation, and most of them are image-based methods. While some prior works have proposed video datasets for camouflaged animal understanding, they have only provided a small amount of data with limited annotation types. To address those challenges and promote more studies on biological monitoring and understanding of animals' behavior, we introduce our CamoVid60K dataset and related benchmarks for a broad range of video understanding tasks. Please see Section 3 and Section 5 in the main paper for more details.

Who created this dataset (*e.g.*which team, research group) and on behalf of which entity (*e.g.*company, institution, organization)?

The authors created the dataset from the XXX and YYY Institutions. The authors created it for the public at large without reference to any particular organization or institution.

Composition

What do the instances that comprise the dataset represent (*e.g.*documents, photos, people, countries)? Are there multiple types of instances (*e.g.*movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.
Each instance in the dataset represents a sequence of extracted frames from a video with different annotations (category, bounding box, mask, motion type, pseudo-label optical flow, and three referring expressions.

How many instances are there in total (of each type, if appropriate)?
CamoVid60K has a total of 218 instances, each containing frames, an associated bounding box, a mask, a motion type, pseudo-label optical flow, one category, and three referring expressions. You can see further statistics on the whole data in Section 3 of the main paper.

Does the dataset contain all possible instances, or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (*e.g.*geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (*e.g.*to cover a more diverse range of instances because instances were withheld or unavailable).
The dataset contains all instances from previous datasets with additional new data that are crawled from the internet to provide a larger volume of data with more frequent annotations and types, and cover a wider variety of species, ranging from flying to terrestrial and aquatic animals. The detailed statistics are shown in Table 1 and Section 3 of the main paper.

What data does each instance consist of? "Raw" data (*e.g.*unprocessed text or images) or features? In either case, please provide a description.
Each instance in our dataset comprises raw MP4 video data, captured at 24-30 frames per second and with resolution from 480×360 to 3840×2160.

Is there a label or target associated with each instance?If so, please provide a description.
Each instance is associated with a bounding box, mask, motion type, pseudo-label optical flow, one category, and three referring expressions.

Is any information missing from individual instances?If so, please provide a description, explaining why this information is missing (*e.g.*, **because it was unavailable). This does not include intentionally removed information but might include, *e.g.*redacted text.**
All instances are complete.

Are relationships between individual instances made explicit (*e.g.*users' movie ratings, social network links)?If so, please describe how these relationships are made explicit.
Some instances may have the same category name and similar referring expressions because they belong to the same category. However, each instance will have its unique ID.

Are there recommended data splits (*e.g.*training, development/validation, testing)?If so, please provide a description of these splits, explaining the rationale behind them.
CamoVid60K is explicitly designed for learning both small and large motion displacement of camouflaged animals. Therefore, it is split into two subsets: small displacement (every single frame) and large displacement (every 5 frames). This division is beneficial for evaluating motion segmentation methods, as it provides a robust framework for analyzing algorithms' performance under varying motion and displacement conditions. Each subset will include training (168 instances) and testing sets (50 instances), as mentioned in Section 3.2 of the main paper.

Are there any errors, sources of noise, or redundancies in the dataset?If so, please provide a description.
The dataset was carefully manually curated to mitigate any errors within the questions and answers. However, due to the nature and characteristics of camouflaged animals and their resolution, some frames will contain errors/mislabeling at the boundary between the animals and the background. We will keep improving the quality of the mask annotations in the next version.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (*e.g.*websites, tweets, other datasets)?If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (*i.e.***including the external resources as they existed at the time the dataset was created); c) are there any restrictions (*e.g.*licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.**
Entirety of the dataset will be made publicly available at our CamodVid60K website (we will update our website later). CamoVid60K will be publicly released under the CC-BY-4.0 license, which allows public use of the video and annotation data for both research and commercial purposes.

Does the dataset contain data that might be considered confidential (*e.g.*data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)?If so, please provide a description.
No

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?If so, please describe why.

No

Does the dataset relate to people?If not, you may skip the remaining questions in this section.

No, CamoVid60K only contains animals.

Does the dataset identify any subpopulations (*e.g.*by age, gender)?If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.

No

Is it possible to identify individuals (*i.e.*one or more natural persons), either directly or indirectly (*i.e.*in combination with other data) from the dataset? If so, please describe how.

No

Does the dataset contain data that might be considered sensitive in any way (*e.g.*data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.

No

CollectionProcess

How was the data associated with each instance acquired? Was the data directly observable (*e.g.*raw text, movie ratings), reported by subjects (*e.g.*survey responses), or indirectly inferred/derived from other data (*e.g.*part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The raw video data, which is directly observable, was procured from the publicly accessible datasets (Camouflaged Animals Dataset (CAD) (Pia Bideau, 2016), Moving Camouflaged Animals (MoCA) (Lamdouar et al., 2020), MoCA-Mask (Cheng et al., 2022), Marine Video Kit (MVK) (Truong et al., 2023) and crawled video from internet) as shown in Table 1 and Section 3 in the main paper. We utilized an annotation tool from (Zheng et al., 2023), which is heavily based on Segment Anything Model (SAM) (Kirillov et al., 2023) for mask initialization and bounding box, and XMem (Cheng & Schwing, 2022) for mask and bounding box propagation. We utilized the RAFT method (Teed & Deng, 2020) to produce an accurate estimated optical flow and refined it using Algorithm 1. To construct referring expression annotations, we utilized GPT-4V (xxxx, 2023) to create a concise description for flying and terrestrial animals, and MarineGPT (Zheng et al., 2023) for aquatic animals.

What mechanisms or procedures were used to collect the data (*e.g.*hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The videos were downloaded in accordance with the official guidelines for data access of other datasets. For additional videos, we manually curated from the internet. See Section 3 in the main paper for a more detailed explanation.

If the dataset is a sample from a larger set, what was the sampling strategy (*e.g.*deterministic, probabilistic with specific sampling probabilities)?

We used all samples from the published datasets. So, there is no sampling strategy.

Who was involved in the data collection process (*e.g.*students, crowd-workers, contractors) and how were they compensated (*e.g.*how much were crowd-workers paid)?

The authors were involved in the data collection process. No crowd-workers were involved during the data collection process.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (*e.g.*recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

The original videos within the published datasets were collected across various occasions spanning from 2011 to 2022. As for the CamoVid60K, the new videos were collected over several sprints during the first half of 2024.

Were any ethical review processes conducted (*e.g.*by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

No

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

No

Did you collect the data from the individuals in question directly or obtain it via third parties or other sources (*e.g.*websites)?

NA

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

NA

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other

access point to, or otherwise reproduce, the exact language to which the individuals consented.
NA

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).
NA

Has an analysis of the potential impact of the dataset and its use on data subjects (*e.g.*a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.
NA

Preprocessing/cleaning/labeling

Was any preprocessing/cleaning/labeling of the data done (*e.g.*discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.
There was no preprocessing done on the videos, and we only did the frame extraction from the videos.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (*e.g.*to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.
The raw data in our CamoVid60K dataset is video. However, all methods will extract videos into frames, so we only provide the extracted frames in our CamoVid60K dataset.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.
We used the FFmpeg library to extract the frames. The packages, executable files, and sources for Windows, macOS, Linux, or building from source are available on their official website.

Distribution

Will the dataset be distributed to third parties outside of the entity (*e.g.*company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.
The dataset will be made publicly available and can be used for non-commercial research and educational purposes under the CC-BY-4.0 license.

How will the dataset be distributed (*e.g.*tarball on website, API, GitHub) Does the dataset have a digital object identifier (DOI)?
The dataset will be distributed at our CamodVid60K website (we will update our website later) upon acceptance to preserve anonymization.

When will the dataset be distributed?
The complete dataset will be made available upon the acceptance of the paper before the camera-ready deadline.

Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.
CamoVid60K dataset will be publicly released under the CC-BY-4.0 license, which allows direct public use of the video/frames and annotation data for non-commercial research and educational purposes.

Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.
No

Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions and provide a link or other access point to, or otherwise reproduce, any supporting documentation.
No

Maintenance

Who will be supporting/hosting/maintaining the dataset?
The authors of the paper will be maintaining the dataset, pointers to which will be hosted on our CamodVid60K website (we will update our website later), along with the guidelines for download and preprocessing if needed.

How can the owner/curator/manager of the dataset be contacted (*e.g.*email address)?
We will post the contact information on our website, primarily contact through email.

Is there an erratum? If so, please provide a link or other access point.
In the future, we will host an erratum on our CamodVid60K website (we will update our website later) to host any approved errata suggested by the authors or the video research community.

Will the dataset be updated (*e.g.*to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to users (*e.g.*mailing list, GitHub)?
Yes, we plan to host an erratum publicly. There are no specific plans for a v2 version, but there seem to be plenty of opportunities for exciting future dataset work based on CamoVid60K.

If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances

(*e.g.*were individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.

No.

Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to users.

N/A There are no older versions at the current moment. All updates regarding the current version will be communicated via our website.

If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description. Contributions will be made possible using comment functions in our CamodVid60K website (we will update our website later). The CamoVid60K team will verify any new contributions before publishing them on our website, and then we will host any approved errata suggested by the video research community.

Acknowledgements This work is supported by an internal grant from HKUST (R9429), the National Research Foundation, Singapore, and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-GC-2023-008), Career Development Fund (CDF) of Agency for Science, Technology and Research (A*STAR) (No.: C233312028), National Research Foundation, Singapore and Infocomm Media Development Authority under its Trust Tech Funding Initiative (No. DTC-RGC-04). This work was partially done when Tuan-Anh Vu was a research resident at CFAR & IHPC, A*STAR, Singapore.

Data Availability All data, codes, and related materials are available via our CamodVid60K website.

References

Beery, S., Van Horn, G., Perona, P.: Recognition in terra incognita. In: ECCV, pp. 456–473 (2018)

Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: ECCV (2012)

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)

Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR, pp. 1290–1299 (2022)

Cheng, H.K., Schwing, A.G.: XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In: ECCV (2022)

Cheng, X., Xiong, H., Fan, D.-P., Zhong, Y., Harandi, M., Drummond, T., Ge, Z.: Implicit motion handling for video camouflaged object detection. In: CVPR (2022)

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR, pp. 248–255 (2009)

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)

Dosovitskiy, A., Fischer, P., Ilg, E., Häusser, P., Hazırbaş, C., Golkov, V., Smagt, P., Cremers, D., Brox, T.: FlowNet: Learning optical flow with convolutional networks. In: ICCV (2015)

Fan, D.-P., Cheng, M.-M., Liu, Y., Li, T., Borji, A.: Structure-measure: A New Way to Evaluate Foreground Maps. In: ICCV (2017)

Fan, D.-P., Ji, G.-P., Xu, P., Cheng, M.-M., Sakaridis, C., Van Gool, L.: Advances in deep concealed scene understanding. Visual Intelligence (2023)

Fan, D.-P., Ji, G.-P., Xu, P., Cheng, M.-M., Sakaridis, C., & Van Gool, L. (2023). Advances in deep concealed scene understanding. *Visual Intelligence, 1*, 1–16.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J.W., Wallach, H., Iii, H.D., Crawford, K.: Datasheets for datasets. Communications of the ACM 64(12), 86–92 (2021)

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92.

He, R., Dong, Q., Lin, J., Lau, R.W.: Weakly-supervised camouflaged object detection with scribble annotations. In: AAAI (2023)

He, C., Li, K., Zhang, Y., Tang, L., Zhang, Y., Guo, Z., Li, X.: Camouflaged object detection with feature decomposition and edge reconstruction. In: CVPR (2023)

He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR, pp. 770–778 (2016)

Ji, G.-P., Fan, D.-P., Chou, Y.-C., Dai, D., Liniger, A., Van Gool, L.: Deep gradient learning for efficient camouflaged object detection. MIR (2023)

Ji, P., Zhong, Y., Li, H., Salzmann, M.: Null space clustering with applications to motion segmentation and face clustering. In: ICIP, pp. 283–287 (2014)

Ji, G.-P., Fan, D.-P., Chou, Y.-C., Dai, D., Liniger, A., & Van Gool, L. (2023). Deep gradient learning for efficient camouflaged object detection. *Machine Intelligence Research, 20*, 92–108.

Kowal, M., Siam, M., Islam, M.A., Bruce, N.D., Wildes, R.P., Derpanis, K.G.: A deeper dive into what deep spatiotemporal networks encode: Quantifying static vs. dynamic information. In: CVPR (2022)

Lamdouar, H., Xie, W., Zisserman, A.: Segmenting invisible moving objects. In: BMVC (2021)

Lamdouar, H., Xie, W., Zisserman, A.: The making and breaking of camouflage. In: ICCV, pp. 832–842 (2023)

Lamdouar, H., Yang, C., Xie, W., Zisserman, A.: Betrayed by motion: Camouflaged object discovery via motion segmentation. ACCV (2020)

Le, T.-N., Nguyen, T.V., Nie, Z., Tran, M.-T., Sugimoto, A.: Anabranch network for camouflaged object segmentation. CVIU (2019)

Le, T.-N., Cao, Y., Nguyen, T.-C., Le, M.-Q., Nguyen, K.-D., Do, T.-T., Tran, M.-T., & Nguyen, T. V. (2021). Camouflaged instance segmentation in-the-wild: Dataset, method, and benchmark suite. *IEEE Transactions on Image Processing, 31*, 287–300.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV, pp. 740–755 (2014). Springer

Lv, Y., Zhang, J., Dai, Y., Li, A., Liu, B., Barnes, N., Fan, D.-P.: Simultaneously localize, segment and rank the camouflaged objects. In: CVPR (2021)

Margolin, R., Zelnik-Manor, L., Tal, A.: How to evaluate foreground maps? In: CVPR (2014)

Mayer, N., Ilg, E., Häusser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2016)

Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: CVPR (2015)

Meunier, E., Badoual, A., Bouthemy, P.: Em-driven unsupervised learning for efficient motion segmentation. IEEE T-PAMI (2022)

Meunier, E., Badoual, A., & Bouthemy, P. (2022). Em-driven unsupervised learning for efficient motion segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*, 4462–4473.

Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C., Clune, J.: Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. PNAS 115(25), 5716–5725 (2018)

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M. S., Packer, C., & Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *PNAS, 115*(25), 5716–5725.

OpenAI, t.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)

Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.-Y., Li, S.-W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. TMLR (2024)

Pei, J., Cheng, T., Fan, D.-P., Tang, H., Chen, C., Van Gool, L.: Osformer: One-stage camouflaged instance segmentation with transformers. In: ECCV (2022)

Pia Bideau, E.L.-M.: It's moving! a probabilistic model for causal motion segmentation in moving camera videos. In: ECCV (2016)

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML, pp. 8748–8763 (2021)

Rands, M.R., Adams, W.M., Bennun, L., Butchart, S.H., Clements, A., Coomes, D., Entwistle, A., Hodge, I., Kapos, V., Scharlemann, J.P., and others: Biodiversity conservation: challenges beyond 2010. Science 329(5997), 1298–1303 (2010)

Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. NeurIPS **28** (2015)

Seo, S., Lee, J.-Y., Han, B.: Urvos: Unified referring video object segmentation network with a large-scale benchmark. In: ECCV, pp. 208–223 (2020). Springer

Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., Sun, J.: Objects365: A large-scale, high-quality dataset for object detection. In: ICCV, pp. 8430–8439 (2019)

Shen, S., Li, C., Hu, X., Xie, Y., Yang, J., Zhang, P., Gan, Z., Wang, L., Yuan, L., Liu, C., and others: K-lite: Learning transferable visual models with external knowledge. NeurIPS 35, 15558–15573 (2022)

Simões, F., Bouveyron, C., Precioso, F.: Deepwild: Wildlife identification, localisation and estimation on camera trap videos using deep learning. Ecological Informatics 75, 102095 (2023)

Simões, F., Bouveyron, C., & Precioso, F. (2023). Deepwild: Wildlife identification, localisation and estimation on camera trap videos using deep learning. *Ecological Informatics, 75*, Article 102095.

Soofi, M., Sharma, S., Safaei-Mahroo, B., Sohrabi, M., Organli, M.G., Waltert, M.: Lichens and animal camouflage: some observations from central asian ecoregions. Journal of Threatened Taxa 14(2), 20672–20676 (2022)

Soofi, M., Sharma, S., Safaei-Mahroo, B., Sohrabi, M., Organli, M. G., & Waltert, M. (2022). Lichens and animal camouflage: some observations from central asian ecoregions. *Journal of Threatened Taxa, 14*(2), 20672–20676.

Sun, G., An, Z., Liu, Y., Liu, C., Sakaridis, C., Fan, D.-P., Van Gool, L.: Indiscernible object counting in underwater scenes. In: CVPR (2023)

Teed, Z., Deng, J.: RAFT: Recurrent all-pairs field transforms for optical flow. In: ECCV (2020)

Troscianko, J., Skelhorn, J., Stevens, M.: Quantifying camouflage: how to predict detectability from appearance. BMC Evolutionary Biology 17, 1–13 (2017)

Troscianko, J., Skelhorn, J., & Stevens, M. (2017). Quantifying camouflage: how to predict detectability from appearance. *BMC Evolutionary Biology, 17*, 1–13.

Truong, Q.-T., Vu, T.-A., Ha, T.-S., Lokoč, J., Tim, Y.H.W., Joneja, A., Yeung, S.-K.: Marine Video Kit: A new marine video dataset for content-based analysis and retrieval. In: MMM. Springer, ??? (2023)

Vu, T.-A., Nguyen, D.T., Guo, Q., Hua, B.-S., Chung, N.M., Tsang, I.W., Yeung, S.-K.: Leveraging open-vocabulary diffusion to camouflaged instance segmentation. arXiv preprint arXiv:2312.17505 (2023)

Wang, Q., Chang, Y.-Y., Cai, R., Li, Z., Hariharan, B., Holynski, A., Snavely, N.: Tracking everything everywhere all at once. In: ICCV (2023)

Wang, F., Cao, P., Li, F., Wang, X., He, B., & Sun, F. (2022). Watb: Wild animal tracking benchmark. *International Journal of Computer Vision, 131*(4), 899–917. https://doi.org/10.1007/s11263-022-01732-3

Xie, C., Xiang, Y., Harchaoui, Z., Fox, D.: Object discovery in videos as foreground motion clustering. In: CVPR (2019)

Xie, J., Xie, W., Zisserman, A.: Segmenting moving objects via an object-centric layered representation. NeurIPS (2022)

Xie, J., Xie, W., & Zisserman, A. (2022). Segmenting moving objects via an object-centric layered representation. *Advances in neural information processing systems, 35*, 28023–28036.

Yang, C., Lamdouar, H., Lu, E., Zisserman, A., Xie, W.: Self-supervised video object segmentation by motion grouping. In: ICCV (2021)

Yang, J., Li, C., Zhang, P., Xiao, B., Liu, C., Yuan, L., Gao, J.: Unified contrastive learning in image-text-label space. In: CVPR, pp. 19163–19173 (2022)

Yang, Z., Wang, J., Ye, X., Tang, Y., Chen, K., Zhao, H., Torr, P.S.: Language-aware vision transformer for referring segmentation. IEEE T-PAMI 00(00), 1–18 (2024)

Zeng, Y., Zhang, P., Zhang, J., Lin, Z., Lu, H.: Towards high-resolution salient object detection. In: ICCV, pp. 7234–7243 (2019)

Zhang, L., Gao, J., Xiao, Z., Fan, H.: Animaltrack: A benchmark for multi-animal tracking in the wild. International Journal of Computer Vision 131(2), 496–513 (2022) https://doi.org/10.1007/s11263-022-01711-8

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L., Shum, H.-Y.: DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In: ICLR (2023)

Zhang, L., Gao, J., Xiao, Z., & Fan, H. (2022). Animaltrack: A benchmark for multi-animal tracking in the wild. *International Journal of Computer Vision, 131*(2), 496–513. https://doi.org/10.1007/s11263-022-01711-8

Zhang, X., Xiao, T., Ji, G.-P., Wu, X., Fu, K., & Zhao, Q. (2025). Explicit motion handling and interactive prompting for video camouflaged object detection. *IEEE Transactions on Image Processing, 34*, 2853–2866.

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: Detrs beat yolos on real-time object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16965–16974 (2024)

Zheng, Z., Xie, Y., Liang, H., Yu, Z., Yeung, S.-K.: CoralVOS: Dataset and benchmark for coral video segmentation. arXiv preprint arXiv:2310.01946 (2023)

Zheng, Z., Zhang, J., Vu, T.-A., Diao, S., Tim, Y.H.W., Yeung, S.-K.: MarineGPT: Unlocking secrets of ocean to the public. arXiv preprint arXiv:2310.13596 (2023)

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: CVPR, pp. 633–641 (2017)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.