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

Tuan-Anh Vu[1] **Ziqiang Zheng**[1] **Chengyang Song**[2] **Qing Guo**[3]
Ivor Tsang[3] **Sai-Kit Yeung**[1]
[1]The Hong Kong University of Science and Technology, Hong Kong SAR
[2]Ocean University of China, China [3]CFAR & IHPC, A*STAR, Singapore

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. Performing robust and precise camouflaged animal segmentation is not trivial even for domain experts because of their consistent appearance with backgrounds. Even though several efforts were made to perform camouflaged animal image segmentation, there is only some work on camouflaged animal video segmentation to the best of the author's knowledge. Biologists usually favor videos with redundant information and temporal consistencies to perform biological monitoring and understanding of the behavior and events of animals. The scarcity of such labeled video data is the most hindering issue. 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 CV, 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 into future research directions. Our dataset stands as a novel and challenging testing set to stimulate more powerful camouflaged animal video segmentation algorithms, and there is still a large room for further improvement.

1 Introduction

The continuously evolving neural networks (*e.g.* CNNs [16] and ViTs [8]) provide a powerful and efficient tool to perform visual understanding based on the captured imagery/videos. Enhancing both *data* and *algorithm* has achieved promising success and progress. Large-scale datasets (*e.g.* COCO [27], ADE20K [56] and Object365 [40]) with supervisions serve as the essential stimulus to foster various powerful visual perception algorithms [50] and benchmark various algorithms for revealing further research directions. Most existing datasets mainly contain our everyday objects (*e.g.* 80 categories in COCO). This work focuses on the camouflaged animals, which are currently less concern and explored. Monitoring and understanding camouflaged animals is crucial for biodiversity conservation [38, 42], as it helps protect species that are otherwise difficult to detect and at risk of unnoticed population declines. Furthermore, studying camouflaged animals provides insights into

Preprint. Under review.

Table 1: Comparison with existing video animal datasets. Class.: Classification Label, B.Box: Bounding Box, Motion: Motion of Animal, Coarse OF: Coarse Optical Flow, Expres.: Expression. ***Note that***, MVK [46] dataset mostly consists of normal marine animals with only some camouflaged animals. The frequency of annotations refers to how often each frame is annotated. For instance, MoCA-Mask provides annotations for **every five frames**, resulting in 4,691 annotated frames. In contrast, our CamoVid60K dataset offers a significantly larger volume of data with more frequent annotations and a wider variety of annotation types.

Dataset	Venue	# videos / frames	# species	Frequency	Class.	B.Box	Mask	Motion	Coarse OF	Expres.
CAD [35]	ECCV'16	9 / 839	6	every 5 frames	✓		✓			
MoCA [22]	ACCV'20	141 / 37,250	67	**every frames**	✓	✓		✓		
MoCA-Mask [5]	CVPR'22	87 / 22,939	44	every 5 frames	✓		✓			
MVK [46]	MMM'23	1379 / 992,880	-	every 30 frames	✓					✓
CamoVid60K	**-**	**218 / 62,774**	**70**	**every frames**	✓	✓	✓	✓	✓	✓

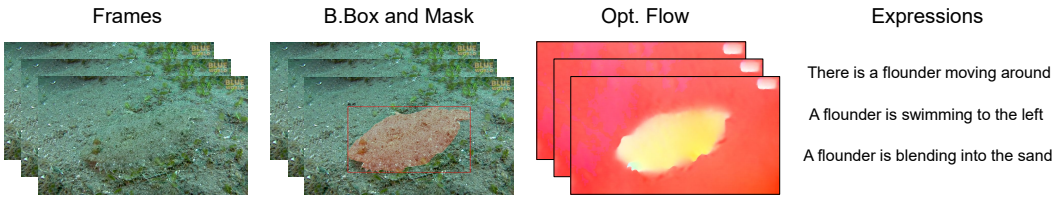

Figure 1: Example from our proposed **CamoVid60K** dataset with bounding box, mask, coarse optical flow, and expressions.

evolutionary biology and adaptation mechanisms, enriching our scientific understanding of natural selection.

However, unlike everyday objects, collecting imagery/videos of camouflaged animals is more challenging, and further annotation procedures usually involve domain experts. Segmentation, generating precise masks for objects' interest, is the 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 yielded masks aid in gathering precise data on their behavior, habitat, and population dynamics, enhancing our overall understanding of their ecology. There are several efforts [5, 24, 47, 50] achieved to perform the camouflaged animal segmentation. Specifically, camouflage is a powerful biological mechanism for avoiding detection and identification, making it more challenging to perform precise segmentation.

Various datasets (*e.g.* CAMO-COCO [25], COD10K [12], CAM-LDR [29], S-COD [17]) have been collected for both image-level camouflaged animal segmentation. However, the image-level camouflaged animal segmentation cannot well satisfy biological monitoring and surveying purposes, where the activity and behavior [51] should be recorded. The MoCA dataset [22] is the most extensive compilation of videos featuring camouflaged objects, yet it only provides detection labels. We argue that only the bounding box annotations cannot well delineate the camouflaged animals, especially those with irregular boundaries, poses, and patterns (*e.g.* the transparent fins of the fish). Furthermore, despite the shifting from image to video, 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 perform camouflaged animal video understanding (**CAVU**) in real-world scenarios, we present **CamoVid60K**, a comprehensive video dataset dedicated to the understanding of 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, showing that **CamoVid60K** *surpasses* all previous datasets in terms of the number of videos/frames and species included. Unlike previous datasets that annotated every fifth frame, our dataset offers annotations for *every single frame*. Additionally, we provide *a wider variety of annotation types* (animal categories, bounding box, annotated mask, coarse optical flow, expression), 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 boundary of these advanced algorithms to perform robust and precise video understanding. Our **CamoVid60K** stands as

a novel and important testing set for both computer vision and wildlife animal research communities. Our main contributions are summarized as follows:

- We present a **large-scale**, **comprehensive** video dataset dedicated to the understanding of camouflaged animals, with **a significantly larger** data and annotation types than the existing datasets.
- We propose a **simple pipeline** for camouflaged animal detection and segmentation with comparable performance.
- We **benchmark various** camouflaged animal video understanding tasks, including image classification, object detection, and motion segmentation based on 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 [13]. CSU tasks can be divided into image-level and video-level categories. Image-level CSU tasks include five main types: camouflaged object counting [43], camouflaged object localization [28, 29], camouflaged object segmentation [12, 15, 18], camouflaged instance ranking [28, 29], and camouflaged instance segmentation [26, 34]. 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 task includes video camouflaged object segmentation [5, 19, 49] and video camouflaged object detection [21–23, 32, 50, 51]. 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 [21–23, 32, 50, 51] and segmentation [5, 19, 24, 49]. The former video camouflaged object detection (VCOD) yields BBOX sequences for the camouflaged animals, while the latter video camouflaged object segmentation (VCOS) generates dense pixel-level masks. MoCA [22] proposed the first large-scale moving camouflaged animals video dataset with BBOX annotations and additional optical flows to boost the detection of camouflaged animals. Further work [23] incorporated visual appearance from a static scene as additional clues to promote the ability of the model to detect camouflaged animals. However, the BBOX annotations could not accurately describe camouflaged objects' pose, appearance, and patterns. To address this issues, Xie *et al*. [49] proposed a novel pixel-trajectory RNN to cluster fore-ground pixels and generate dense segmentation masks for object discovery in videos. MoCA-Mask [5] 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 fifth frame**, 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 for camouflaged animals is quite challenging, even without focusing on long-form videos. This is because manually collecting, observing, and annotating videos with several annotation types is labor-intensive, time-consuming, and expensive. In addition to the costs, ensuring visual data diversity and high-quality annotations adds to the difficulty. This section proposes a staged data collection and processing pipeline in Figure 2. Associated datasheets [14] and data cards [36] for our **CamoVid60K** dataset are provided in the supplementary.

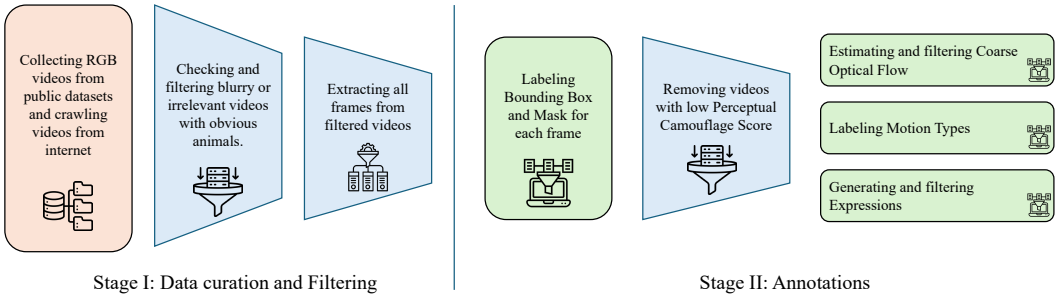

Figure 2: **CamoVid60K** data pipeline. Stage I includes data curation, filtering irrelevant videos, and extracting all frames. Stage II includes data annotation, generation, and filtering.

3.1 Data Construction and Processing

Data sources and Pre-Processing. We built our dataset from previous datasets (Table 1) and crawled videos from the internet to cover various camouflaged animals. We collected 1,929 videos and manually checked and filtered blurry, irrelevant videos with obvious animals. We then extracted every frame (instead of every five frames proposed in existing datasets, see Table 1) of each video before annotating them. At the end, our dataset comprises **218** videos with **62,774** frames of **70** animals.

BBOX and Mask Annotation. We utilized annotation tool from [54] which heavily based on Segment Anything Model (SAM) [20] for mask initialization and bounding box and XMem [4] for mask and bounding box propagation. Then, we manually check and refine every frame to provide accurate bounding boxes and segmentation masks. In addition, we adopt the perceptual camouflage score ($\mathcal{S}_p$) from [24] to quantify the effectiveness of animals' camouflage, *i.e*. how successfully an animal blends into its background. Based on the perceptual camouflage score, we will keep the videos that are higher than the threshold ($\mathcal{S}_p > 0.5$).

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

Coarse Optical Flow Annotation. Previous optical flow datasets, including Flying Chair [7], KITTI [31], Sintel [1] utilized either simulation software or real images with other heavy sensors information (depth, LiDAR, *etc*.) and algorithms to create optical flow ground-truth. It is time-consuming and requires extreme effort. In addition, with the development of deep learning techniques, many methods [45, 48] can produce accurate estimated optical flow. Therefore, we utilized these methods for our coarse optical flow annotation with the pseudo algorithm shown in Algorithm 1.

Note that, even though our processing pipeline for optical flow annotation will produce accurate and dense optical flow, it is still **estimated** optical flow, so it is reasonable and capable of use as *additional input* to boost performance for other tasks such as motion segmentation task. It is **not recommended** to use it as ground truth for evaluation.

Motion Annotation. Following [22], we manually labeled our dataset by their types of motion, as shown below. Based on the motion types, We can further annotate the camouflage methods of animals, which we plan to provide in the next version.

- *Locomotion*: when the animal has movement(s) that significantly changes its location.
- *Deformation*: when the animal engages in a more delicate movement that only changes its pose while remaining in the same location.
- *Still*: when the animal remains still.

Expression Annotation. We first utilized GPT-4V [44] to create a concise description within 30 words that accurately represents the target object for every frame. However, we found that the captions of aquatic animals are less accurate; therefore, we utilized MarineGPT [55], a first vision-language

Algorithm 1 Optical Flow Computation and Filtering

Input: Sequence of frames

Output: Sequence of computed optical flows

1: **for** each pair of frames (i, j) **do**
2: Computing all pairwise optical flows using RAFT [45]
3: Computing DINO features [33] for each frame
4: Filtering flows using cycle consistency and appearance consistency check
5: Applying chain cycle consistent correspondences to create denser correspondences
6: **end for**

model specially designed for the marine domain for aquatic animals. After the initial annotation, we verified and refined all annotations and chose the best three captions for each video sequence. Objects that could not be localized using language expressions were removed.

3.2 Dataset Specifications and Statistics

Data Organization. As shown in Figure 3, we split our dataset based on displacement into two subsets: small displacement (every single frame) and large displacement (every fifth frame). 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 and testing sets with images, pre-computed optical flows, and annotations. We name every image as follows: 'SuperClass-SubClass-SubNumber-MotionType-FrameNumber.' This systematic naming convention ensures clarity and ease of reference within the dataset.

```
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
```

Figure 3: Data organization of our dataset.

Category Diversity. The distributions of camouflaged animals by the biology-inspired hierarchical categorization within three super groups are visually represented through word clouds in Figure 4-Top and Figure 5. Additionally, Figure 4-Bottom showcases some examples with different animals' positions and the total sum of normalized bounding boxes across the entire dataset.

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 [5] to assess the pixel-wise masks: Mean Absolute Error (M), Enhanced-alignment measure (E_ϕ) [10], Structure-measure (S_α) [9], Weighted F-measure (F_β^w) [30], 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 Mask2Former architecture [3] for both object detection and motion segmentation tasks. As shown in Figure 6, our proposed simple pipeline processes a sequence of images or videos by employing any off-the-shelf flow estimation

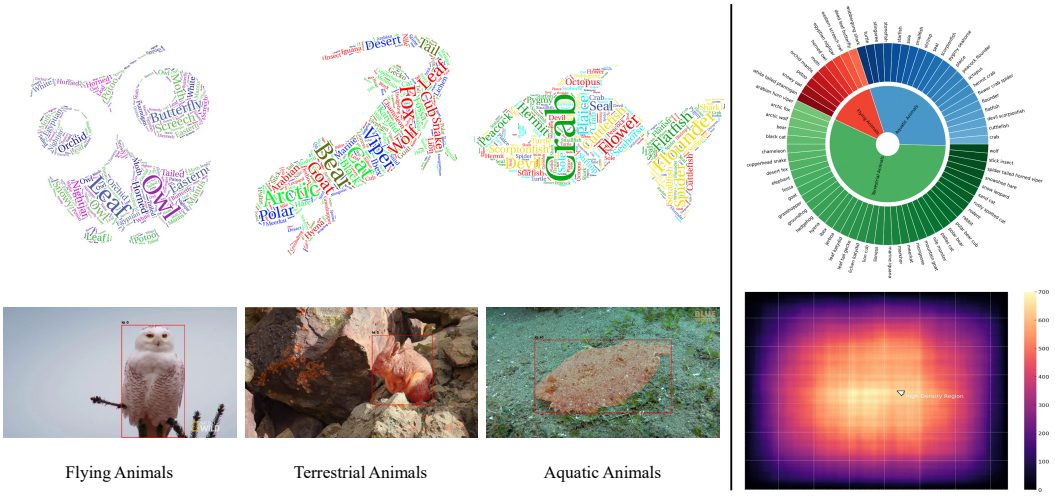

Figure 4: **Left-Top:** Word cloud of category distribution of camouflaged animals. **Right-Top:** Taxonomic structure of our dataset by their biology-inspired hierarchical categorization. It encompasses various animals, spanning 70 categories across flying, terrestrial, and aquatic groups. **Left-Bottom:** Some examples with different animals' positions. **Right-Bottom:** 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.

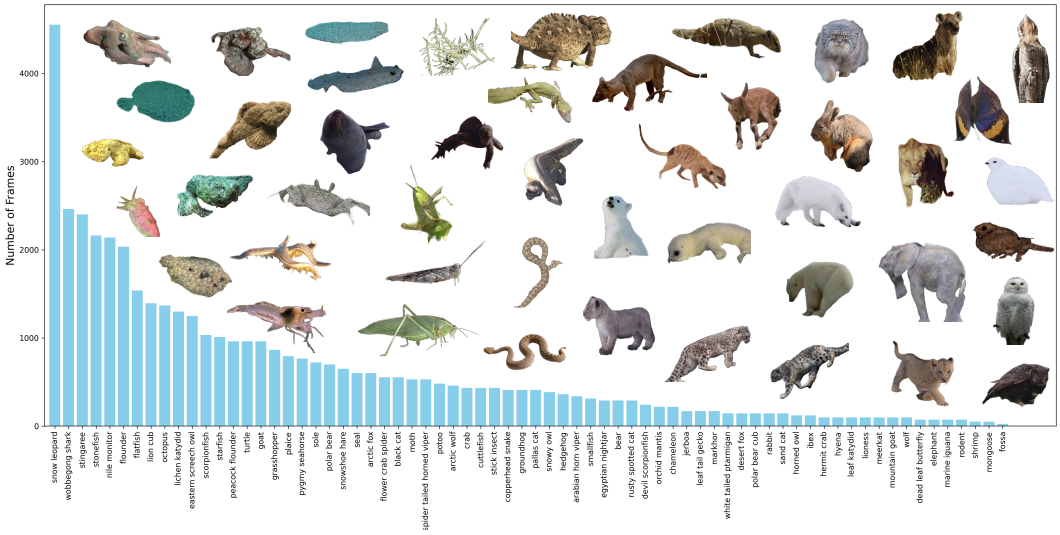

Figure 5: Category distribution (ranging from 100 to 4,500 frames) and some visual examples (extracted animal masks) of our dataset. The variety ensures a wide range of camouflaged animals, allowing for comprehensive evaluation across various scenarios. We will keep adding more data to balance the distribution.

methods. In our case, we directly take the refined optical flow in our dataset instead of utilizing the RAFT method [45] to estimate raw optical flow as [24]. The images and associated estimated flows are passed into two separated encoders for feature extraction. Subsequently, each timestamp's image and flow features are aggregated before going through the decoder to predict the segmentation mask.

Visual Encoder. We adopt the SINet-v2 [12] architecture that takes 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 frame, $h\&w$ are the height & width and outputs visual features $\{f_1^v, f_2^v, \ldots, f_n^v\} = \Phi_{visual}(I^v)$.

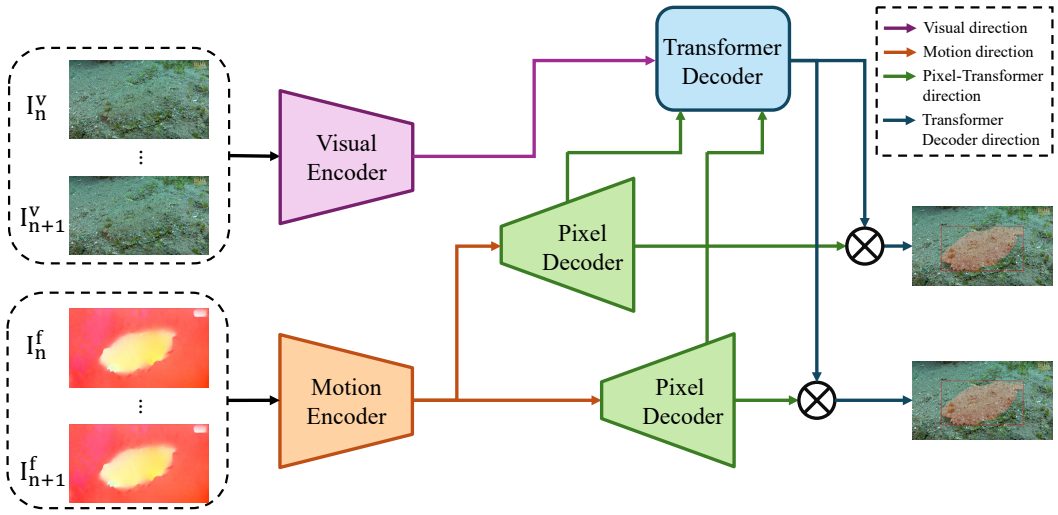

Figure 6: Our simple pipeline takes a sequence of images/video and the associated optical flow as input. They are fed into separated encoders for feature extraction. Then, the motion features with spatial and temporal positional encoding are passed to Pixel Decoders to produce a set of enriched motion features. Next, the Transformer Decoder takes the visual features and enriched motion features to produce mask embedding for the moving object and bounding box.

Motion Encoder. Inspired by the motion segmentation architecture [23], we use light-weight convNet that takes as input a sequence of optical flows $I^f = \{I^f_1, I^f_2, \ldots, I^f_n\} \in \mathbb{R}^{n \times d_f \times h \times w}$, where d_f is the dimension of flow field and output motion features $\{f^m_1, f^m_2, \ldots, f^m_n\} = \Phi_{motion}(I^m)$. Then, concatenating motion features with learned spatial and temporal positional encodings to output a set of enriched motion features.

Decoder. We adopt Mask2Former [3] architecture, which includes Transformer and Pixel Decoders. The Transformer decoder combines a trainable query for mask embedding with the results of the motion encoder and visual features. Like Mask2Former, this query focuses on multi-scale motion features and visual features, resulting in mask embedding for the moving object. In addition, similar to the pixel decoder in Mask2Former, a ConvNet decoder with low computational complexity utilizes skip-connections to generate high-resolution segmentation masks and bounding boxes from the motion features and mask embedding.

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

$$\mathcal{L} = \mathcal{L}_{\text{bce}} + \mathcal{L}_{\text{L1}} + \mathcal{L}_{\text{ce}}, \tag{1}$$

5 Experiments

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

5.1 Baselines

For motion segmentation task, we selected recent SOTAs to compare, including two frame-based methods (PraNet [11], SINet-v2 [12]) and two video-based methods (MG [51], SLT-Net [5]). For a fair comparison, we utilize the implementations provided by the authors and train all methods using the same training set.

For object detection task, we compare with four well-known detection methods, such as Faster-RCNN [39], DETR [2], DINO [53]. We followed the so-called 1× setting (12-epoch setting) for training and used the same ResNet50 [16] as the backbone for all methods.

Table 2: Quantitative results of motion segmentation on CamoVid60K dataset. Our model consistently achieves better performance than other competitors on all metrics.

Methods		$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$M \downarrow$	mDic ↑	mIoU ↑
Image	PraNet [11]	0.526	0.161	0.547	0.045	0.198	0.144
	SINet-v2 [12]	0.529	0.166	0.553	0.042	0.206	0.149
Video	MG [51]	0.522	0.153	0.541	0.043	0.197	0.141
	SLT-Net [5]	**0.576**	**0.253**	**0.591**	**0.039**	0.268	0.249
	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	$AP \uparrow$
F-RCNN [39]	28.71
DETR [2]	37.56
DINO [53]	**39.84**
Ours	38.39

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**

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

	CLIP [37]	UniCL [52]	K-LITE [41]
mAcc	30.06	30.89	**31.44**

For zero-shot image classification task, we tested recent three methods, including CLIP [37], UniCL [52] and K-LITE [41]. We used the Swin-T model for both UniCL and K-LITE (pre-trained on ImageNet-21K dataset [6]) and the ViT-B/32 pre-trained model from OpenAI CLIP.

All methods are trained and tested on the same NVIDIA 3090 GPU, except the pre-trained models in the zero-shot image classification task.

5.2 Benchmarks and Discussions

Comparison with image-based and video-based motion segmentation methods. We report the performance of our method with other methods in Table 2. Compared to image-based approaches, our method demonstrates significantly superior performance thanks to the incorporation of temporal information. When evaluated against video-based approaches, our method also surpasses MG [51], which relies solely on estimated optical flows as input. However, compared to the recent state-of-the-art method SLT-Net [5], our method performs better on certain metrics. This is because SLT-Net 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.

Comparison with object detection methods. As shown in Table 3, the proposed model demonstrates performance comparable to other specialized methods, owing to its dual capability in object detection and motion segmentation. Specifically, our method significantly outperforms CNN-like 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, our approach shows mixed results. It surpasses the standard DETR model [2] yet falls short of DINO, an advanced variant of DETR. DINO [53] enhances performance through several innovative techniques: it employs contrastive denoising training to refine one-to-one matching, a mixed query selection method to initialize the queries better, and a 'look forward twice' method that utilizes gradients from subsequent layers to adjust parameters more accurately.

Additional Analysis and Discussions. As shown in Table 4, optical flow plays a crucial role in the motion segmentation of camouflaged animals because optical flow can detect subtle movements by analyzing the motion vectors between frames, distinguishing moving animals from static backgrounds, which is particularly useful in identifying the slight movements of camouflaged animals.

State-of-the-art methods, including foundation models (trained on large datasets), struggle with zero-shot image classification of camouflaged animals, as shown in Table 5. This is due to their subtle and complex patterns, lack of specific training data, and difficulty in generalizing across different backgrounds and lighting conditions. Improving these methods involves curating specialized training data (or fine-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 a large-scale video dataset for camouflaged animal understanding, named **CamoVid60K**, to foster further research on animals. This dataset provides a considerable 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 a foundational model for studying camouflaged animals.

Limitations and Future Works. 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.

New Benchmark. The **CamoVid60K** 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 various environments, providing a robust framework for testing and developing new models.

Impact on Animal Study. 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 utilize 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.

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.

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

Acknowledgement This work is supported by an internal grant from HKUST (R9429). This work is partially done when Tuan-Anh Vu was a research resident at CFAR & IHPC, A*STAR, Singapore.

References

[1] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[3] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022.

[4] Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, 2022.

[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 *CVPR*, 2022.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[7] A. Dosovitskiy, P. Fischer, E. Ilg, P. Häusser, C. Hazırbaş, V. Golkov, P. v.d. Smagt, D. Cremers, and T. Brox. FlowNet: Learning optical flow with convolutional networks. In *ICCV*, 2015.

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A New Way to Evaluate Foreground Maps. In *ICCV*, 2017.

[10] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In *IJCAI*, 2018.

[11] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 263–273. Springer, 2020.

[12] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE T-PAMI*, 2022.

[13] Deng-Ping Fan, Ge-Peng Ji, Peng Xu, Ming-Ming Cheng, Christos Sakaridis, and Luc Van Gool. Advances in deep concealed scene understanding. *Visual Intelligence (VI)*, 2023.

[14] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92, 2021.

[15] Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *CVPR*, 2023.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[17] Ruozhen He, Qihua Dong, Jiaying Lin, and Rynson WH Lau. Weakly-supervised camouflaged object detection with scribble annotations. In *AAAI*, 2023.

[18] Ge-Peng Ji, Deng-Ping Fan, Yu-Cheng Chou, Dengxin Dai, Alexander Liniger, and Luc Van Gool. Deep gradient learning for efficient camouflaged object detection. *Machine Intelligence Research*, 2023.

[19] Pan Ji, Yiran Zhong, Hongdong Li, and Mathieu Salzmann. Null space clustering with applications to motion segmentation and face clustering. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 283–287, 2014.

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

[21] Matthew Kowal, Mennatullah Siam, Md Amirul Islam, Neil DB Bruce, Richard P Wildes, and Konstantinos G Derpanis. A deeper dive into what deep spatiotemporal networks encode: Quantifying static vs. dynamic information. In *CVPR*, 2022.

[22] Hala Lamdouar, Charig Yang, Weidi Xie, and Andrew Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. *ACCV*, 2020.

[23] Hala Lamdouar, Weidi Xie, and Andrew Zisserman. Segmenting invisible moving objects. In *BMVC*, 2021.

[24] Hala Lamdouar, Weidi Xie, and Andrew Zisserman. The making and breaking of camouflage. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 832–842, 2023.

[25] Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *CVIU*, 2019.

[26] 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 T-IP*, 2021.

[27] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[28] 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 *CVPR*, 2021.

[29] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Nick Barnes, and Deng-Ping Fan. Towards deeper understanding of camouflaged object detection. *IEEE T-CSVT*, 2023.

[30] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *CVPR*, 2014.

[31] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015.

[32] Etienne Meunier, Anaïs Badoual, and Patrick Bouthemy. Em-driven unsupervised learning for efficient motion segmentation. *IEEE T-PAMI*, 2022.

[33] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[34] Jialun Pei, Tianyang Cheng, Deng-Ping Fan, He Tang, Chuanbo Chen, and Luc Van Gool. Osformer: One-stage camouflaged instance segmentation with transformers. In *ECCV*, 2022.

[35] Erik Learned-Miller Pia Bideau. It's moving! a probabilistic model for causal motion segmentation in moving camera videos. In *ECCV*, 2016.

[36] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1776–1826, 2022.

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021.

[38] Michael RW Rands, William M Adams, Leon Bennun, Stuart HM Butchart, Andrew Clements, David Coomes, Abigail Entwistle, Ian Hodge, Valerie Kapos, Jörn PW Scharlemann, et al. Biodiversity conservation: challenges beyond 2010. *science*, 329(5997):1298–1303, 2010.

[39] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[40] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *IEEE/CVF international conference on computer vision (CVPR)*, pages 8430–8439, 2019.

[41] Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Zhe Gan, Lijuan Wang, Lu Yuan, Ce Liu, et al. K-lite: Learning transferable visual models with external knowledge. *Advances in Neural Information Processing Systems*, 35:15558–15573, 2022.

[42] Mahmood Soofi, Sandeep Sharma, Barbod Safaei-Mahroo, Mohammad Sohrabi, Moosa Ghorbani Organli, and Matthias Waltert. Lichens and animal camouflage: some observations from central asian ecoregions. *Journal of Threatened Taxa*, 14(2):20672–20676, 2022.

[43] Guolei Sun, Zhaochong An, Yun Liu, Ce Liu, Christos Sakaridis, Deng-Ping Fan, and Luc Van Gool. Indiscernible object counting in underwater scenes. In *CVPR*, 2023.

[44] OpenAI team. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[45] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020.

[46] Quang-Trung Truong, Tuan-Anh Vu, Tan-Sang Ha, Jakub Lokoč, Yue Him Wong Tim, Ajay Joneja, and Sai-Kit Yeung. Marine Video Kit: A new marine video dataset for content-based analysis and retrieval. In *MultiMedia Modeling - 29th International Conference, MMM 2023*. Springer, 2023.

[47] Tuan-Anh Vu, Duc Thanh Nguyen, Qing Guo, Binh-Son Hua, Nhat Minh Chung, Ivor W Tsang, and Sai-Kit Yeung. Leveraging open-vocabulary diffusion to camouflaged instance segmentation. *arXiv preprint arXiv:2312.17505*, 2023.

[48] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. In *ICCV*, 2023.

[49] Christopher Xie, Yu Xiang, Zaid Harchaoui, and Dieter Fox. Object discovery in videos as foreground motion clustering. In *CVPR*, 2019.

[50] Junyu Xie, Weidi Xie, and Andrew Zisserman. Segmenting moving objects via an object-centric layered representation. *NeurIPS*, 2022.

[51] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *ICCV*, 2021.

[52] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19163–19173, 2022.

[53] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2023.

[54] Ziqiang Zheng, Yaofeng Xie, Haixin Liang, Zhibin Yu, and Sai-Kit Yeung. CoralVOS: Dataset and benchmark for coral video segmentation. *arXiv preprint arXiv:2310.01946*, 2023.

[55] Ziqiang Zheng, Jipeng Zhang, Tuan-Anh Vu, Shizhe Diao, Yue Him Wong Tim, and Sai-Kit Yeung. MarineGPT: Unlocking secrets of ocean to the public. *arXiv preprint arXiv:2310.13596*, 2023.

[56] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*, pages 633–641, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The paper's content is consistent with the contribution, experimental results, and limitations presented in the abstract and introduction.
   (b) Did you describe the limitations of your work? [Yes] We discussed the limitations of our work in Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We discussed the potential broader impact of our work in Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We read the ethics review guidelines and ensured that our paper conforms to them.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [No] We do not have assumptions about the theoretical results.
   (b) Did you include complete proofs of all theoretical results? [No] We do not include the complete proofs since this work is mainly about video camouflaged animal segmentation dataset and benchmark.
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We have already included most instructions to reproduce our results in the main paper and will include more details in the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We elaborate the experimental details in Section 5.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We did not provide the error bars due to limited computational resources but will add this in the future.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We include the total amount of computational resources and the type of resources in Section 5.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] We have cited the related papers, datasets, models, and websites.
(b) Did you mention the license of the assets? [Yes] We mention the license of the assets in Section 6.
(c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We include new assets in the supplemental material to provide more detailed information and support our research.
(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] We discuss the data collection in Section 3.
(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] We have discussed the data sources of our dataset, and we have performed the human review to ensure there is no personally identifiable information or offensive content within the images in Section 3 and Section 6.

5. If you used crowdsourcing or conducted research with human subjects...
(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No] Our dataset and benchmark do not contain any risks about human research.
(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No] We build our dataset by ourselves, so we don't have any wage paid or compensation.

A CamoVid60K Description

A.1 Data Curation

We built our dataset from published datasets (Camouflaged Animals Dataset (CAD) [35], Moving Camouflaged Animals (MoCA) [22], MoCA-Mask [5], Marine Video Kit (MVK) [46]) and crawled video from internet.

The CAD dataset includes nine short video sequences obtained from YouTube videos. Hand-labeled ground truth masks were provided for every fifth frame.

The MoCA dataset comprises around 37,000 frames extracted from 141 YouTube video sequences. Most videos are presented at an image resolution of 1280×720 and the FPS of these videos is 24. This dataset has 67 distinct species of animals in locomotion within their native habitats, although it contains a few occurrences of animals with less camouflaged characteristics.

The MoCA-Mask dataset is built upon the MoCA dataset with some modifications. Therefore, their new subset consists of 87 video sequences with 22,939 frames. It offers precise human-labeled segmentation masks for every fifth frame. Consequently, their 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 in 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 various videos recorded in different conditions, such as variable light levels, points of view, water clarity, and environmental conditions. They also offer roughly 40k frames (extracted at one fps or every 30 frames) with associated expression annotation.

To crawl videos from the internet, we curated a list of animals' names that potentially have camouflage abilities. Then, we created a template for searching and downloading videos '*video of camouflage + animals' name*'. Combining with videos from the above datasets, we collected 1,929 videos.

A.2 Data Filtering

At first, we manually checked and removed blurry, irrelevant (with obvious animals) videos to get 218 videos for annotation. To further check and filter the images and annotations with less camouflaged characteristics, we adopt the perceptual camouflage score ($\mathcal{S}_p$) from [24] to quantify the effectiveness

of animals' camouflage, *i.e*. how successfully an animal blends into its background. Based on the perceptual camouflage score, we will keep the videos that are higher than the threshold ($\mathcal{S}_p > 0.5$). In detail, we explain how to adopt the perceptual camouflage score $\mathcal{S}_p$ as follows:

$$\mathcal{S}_p = (1-\alpha)\mathcal{S}_\mathcal{R} + \alpha\mathcal{S}_\mathcal{B} \tag{2}$$

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

In detail, given an image $\mathcal{I}$ and segmentation mask m_S, the reconstruction fidelity score $\mathcal{S}_\mathcal{R}$ is computed by assessing the difference value between the foreground region and its reconstruction. Specifically, it counts the number of foreground pixels ($\mathcal{I}_{\text{fg}} = I \odot \text{erode}(m_s)$) that have been successfully reconstructed from the background ($\mathcal{I}_{\text{bg}} = I \odot (1 - \text{dilate}(m_s))$):

$$\mathcal{S}_\mathcal{R}(I, m_s) = \frac{1}{N_{\text{fg}}} \sum_{(i,j)\in I_{\text{fg}}} \mathcal{R}(i,j) \tag{3}$$

$$\mathcal{R}(i,j) = \begin{cases} 1 & \text{if } ||I_{\text{fg}} - \Psi_{I_{\text{bg}}}(I_{\text{fg}})||_2 < \lambda ||I_{\text{fg}}||_2 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where $\Psi_{I_{\text{bg}}}(.)$ 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 aims to measure the animal's boundary properties (or contour visibility) by penalizing the boundary pixels that are predicted as contour in both images' contour ($\mathcal{C}$) and the ground truth animal's contour ($\mathcal{C}_{\text{gt}}$) with F1 metric:

$$S_\mathcal{B}(I, m_s) = 1 - \text{F1}(m_b \odot C_{\text{gt}},\ m_b \odot C) \tag{5}$$

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

We used the same values for parameters as in [24], such as $\alpha = 0.35, \lambda = 0.2$.

A.3 Visualization

We show some samples in alphabetical order of our CamoVid60K dataset in the attachment (index.html).

B 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 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 Sections 3 and 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, coarse optical flow, and three expressions.

How many instances are there in total (of each type, if appropriate)?

CamoVid60K has a total of 218 instances, each containing frames, associated bounding box, mask, motion type, coarse optical flow, one category, and three 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 640x360 to 1920x1080.

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, coarse optical flow, one category, and three 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 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 fifth frame). 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/mislabelled 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. 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

Collection Process

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) [35], Moving Camouflaged Animals (MoCA) [22], MoCA-Mask [5], Marine Video Kit (MVK) [46] and crawled video from internet) as shown in Table 1 and Section 3 in the main paper. We utilized an annotation tool from [54], which is heavily based on Segment Anything Model (SAM) [20] for mask initialization and bounding box and XMem [4] for mask and bounding box propagation. We utilized RAFT method [45] to produce an accurate estimated optical flow and refined it using Alg. 1. To construct expression annotations, we utilized GPT-4V [44] to create a concise description for flying and terrestrial animals, and MarineGPT [55] 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 build from source are available in 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 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 along with the guideline 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 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. 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.