A Large-Scale Video Dataset for Moving Camouflaged Animals Understanding

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

[1]Hong Kong University of Science and Technology [2]CFAR & IHPC, A*STAR

Abstract

*In recent years, there has been a growing interest in applying Computer Vision (CV), Artificial Intelligence (AI), and Deep Learning (DL) to the study of animals. However, only a few studies have specifically targeted camouflaged animals. Training DL models is often resource-intensive, requiring large amounts of accurately annotated data. Enhancing data diversity and scale is essential to mitigate algorithmic biases. To address these challenges, we present **CamoVid40K**, a diverse, large-scale, and accurately annotated video dataset of camouflaged animals. This dataset comprises **149** videos with **44,270** finely annotated frames, covering **69** animal categories. **CamoVid40K** is designed for various downstream tasks in CV, such as camouflaged animal classification, detection, and task-specific segmentation,* etc.

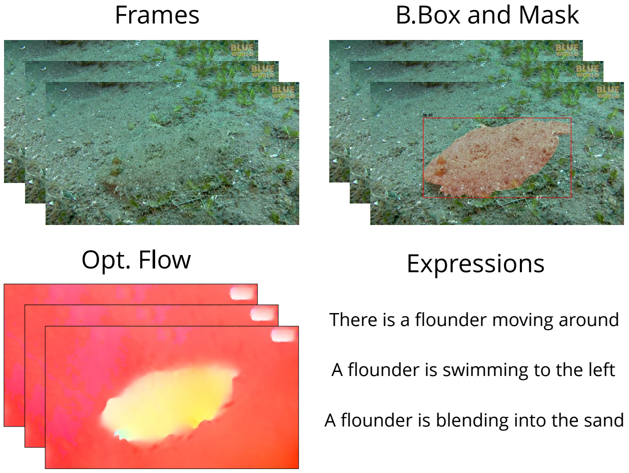
Frames
B.Box and Mask
Opt. Flow
Expressions
There is a flounder moving around
A flounder is swimming to the left
A flounder is blending into the sand

Figure 1. Example from our proposed **CamoVid40K** dataset with bounding box, mask, coarse optical flow, and expressions.

1. Introduction

Camouflage is a powerful biological mechanism for avoiding detection and identification. In nature, camouflage tactics are employed to deceive the sensory and cognitive processes of both preys and predators. Wild animals utilize these tactics in various ways, ranging from blending themselves into the surrounding environment to employing disruptive patterns and colouration [27]. Identifying camouflage is pivotal in many wildlife surveillance applications [11], as it assists in locating hidden individuals for study and protection.

Concealed 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 [10]. CSU tasks can be divided into image-level and video-level categories. Image-level CSU tasks include five main types: concealed object counting (COC) [31], concealed object localization (COL) [22, 23], concealed object segmentation (COS) [9, 12, 14], concealed instance ranking (CIR) [22, 23], and concealed instance segmentation (CIS) [20, 29]. These tasks can be further categorized based on their semantic focus: object-level (including COS and COL) and instance-level (comprising CIR, COC, and CIS). Object-level tasks focus on identifying objects, while instance-level tasks aim to differentiate various entities. Additionally, COC 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 concealed object segmentation (VCOS) [4, 36] and video concealed object detection (VCOD) [16–18, 26, 37, 38]. 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.

In recent years, various datasets have been collected for both image-level (CAMO-COCO [19], NC4K [22], CAMO++ [20], COD10K [9], CAM-LDR [23], S-COD [13], and IOCfish5K [31]) and video-level (CAD [30], MoCA [17], MoCA-Mask [4], MVK [34]) CSU tasks.

The MoCA dataset is the most extensive compilation of videos featuring camouflaged objects, yet it only provides detection labels. Consequently, researchers [37, 38] often evaluate the efficacy of sophisticated segmentation

Dataset	Venue	# videos / frames	# species	Frequency	Class.	B.Box	Mask	Coarse OF	Expres.
CAD [30]	ECCV'16	9 / 839	6	every 5 frames	✓		✓		
MoCA [17]	ACCV'20	141 / 37,250	67	every 5 frames	✓	✓			
MoCA-Mask [4]	CVPR'22	87 / 22,939	44	every 5 frames	✓		✓		
CamoVid40K	**-**	**149 / 44,270**	**69**	**every frames**	✓	✓	✓	✓	✓

Table 1. Comparison with existing video camouflaged animal datasets. Class.: Classification Label, B.Box: Bounding Box, Coarse OF: Coarse Optical Flow, Expres.: Expression.

models by transforming segmentation masks into detection bounding boxes. With the recent advent of MoCA-Mask, there's been a shift towards video segmentation in concealed scenes. However, despite these advancements, the data annotations remain insufficient in both volume and accuracy for developing a reliable video model capable of effectively handling complex concealed situations.

Therefore, to advance the research and development of Camouflaged Animal Understanding (CAU) in real-world scenarios, we present **CamoVid40K**, a comprehensive video dataset dedicated to the understanding of camouflaged animals. It comprises **149** videos with **44,270** finely annotated frames, covering **69** animal categories. Table 1 compares our proposed dataset with previous ones, showing that **CamoVid40K** *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 CAU tasks. This 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*.

2. Our CamoVid40K dataset

2.1. Data Curation

We built our dataset from Camouflaged Animals Dataset (CAD) [30], Moving Camouflaged Animals (MoCA) [17], MoCA-Mask [4], Marine Video Kit (MVK) [34] and crawled video from internet.

The CAD dataset is a concise collection including 9 short video sequences. The sequences were obtained from YouTube videos, and for **every fifth frame**, hand-labeled ground truth masks were provided.

The MoCA Dataset comprises around 37,000 frames extracted from 141 YouTube video sequences. Most videos are presented at a resolution of 720 × 1280 and have a frame rate of 24fps. 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 1379 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. 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.

Preprocessing and Filtering. We collected around 1,929 videos and manually checked and filtered blurry, irrelevant videos with obvious animals. We then extracted every frames (instead of every 5 frames of others) of each video before annotating them. At the end, our dataset comprises **149** videos with **44,270** frames of **69** animals.

Box and Mask Annotation. We utilized annotation tool from [39] which heavily based on Segment Anything Model (SAM) [15] for mask initialization and XMem [3] for mask propagation. Then, we manually check and refine every frame to provide accurate bounding boxes and segmentation masks. ***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 in the next version.

Coarse Optical Flow Annotation. Previous optical flow datasets, including Flying Chair [6], KITTI [25], 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

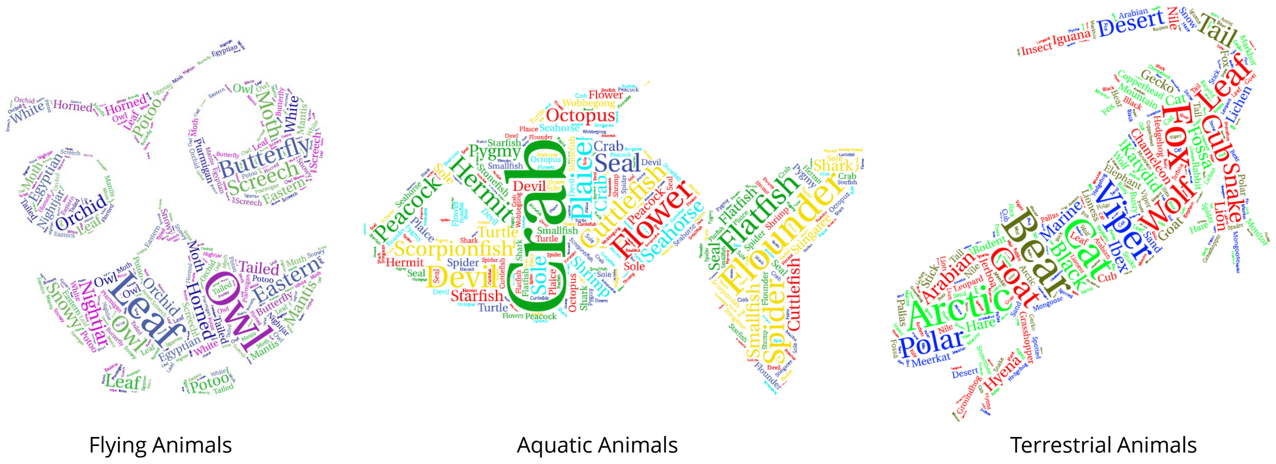

Figure 2. Word cloud of category distribution of camouflaged animals.

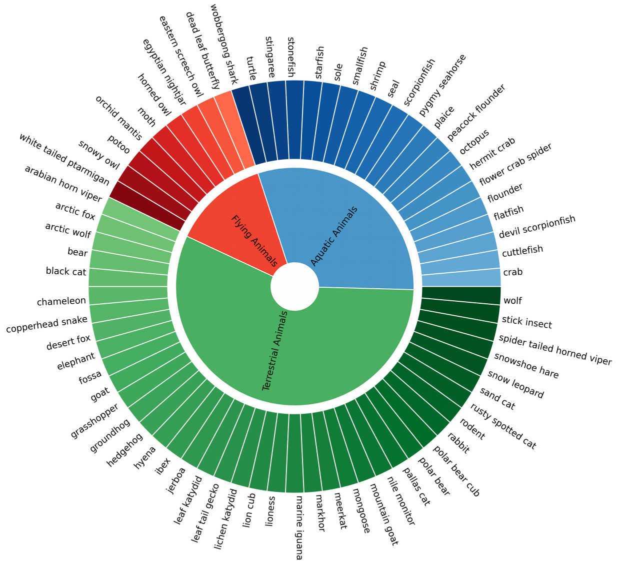

Figure 3. Taxonomic structure of our dataset.

the development of deep learning (DL) techniques, many methods [5, 21, 33, 35] can produce accurate estimated optical flow. Therefore, we utilized these DL methods (in our case, we used RAFT) to compute all pairwise optical flow fields (for each pair, we use the previous flow as initialization to compute the current flow). We filter the estimated optical flow using cycle consistency and appearance consistency checks using DINO features [2, 28] (only keep correspondences for occluded pixels if the correspondences are inconsistent in the first cycle but consistent in the second cycle and if the two frames are adjacent enough, *e.g*. interval < 3). Finally, we applied chain cycle consistent correspondences to create denser correspondences. The chaining rule is as follows: only chain cycle consistent flows between adjacent frames, and if the direct cycle consistent flow exists, the chained flows will be overwritten by the direct flows, which are considered to be more reliable, and the procedure then continues iteratively. The chained flows will go through a final appearance check in both feature and RGB space to reduce spurious correspondences. One can think of this process as augmenting the original direct optical flows (which are unchanged) with some chained ones. Using the chaining rule will help optimize sequences where the number of valid flows is very imbalanced across regions and will help handle sequences with rapid motion and large displacements. ***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 to use as *additional input* for other tasks such as motion segmentation task. It is not recommended to use it as ground truth for evaluation.

Motion Annotation. We manually labeled our dataset by their types of motion, same as [17] as below.

- Locomotion: when the animal has movement that significantly changes its location within the scene.
- 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 [32] 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 [40], a first vision-language model specially designed for the marine domain for aquatic animals. After the initial annotation, we verified

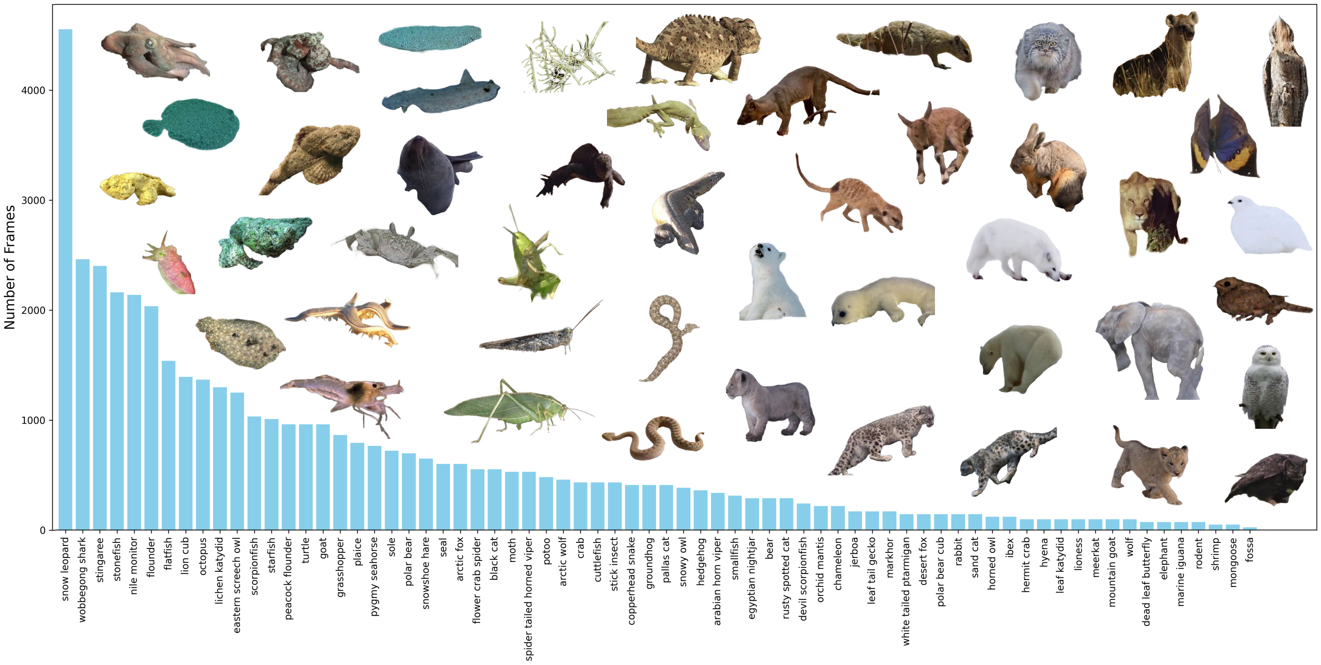

Figure 4. Category distribution and some visual examples (extracted animal masks) of our dataset.

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.

2.2. Dataset Specifications and Statistics

Category Diversity. The distributions of camouflaged animals by category within three super groups are visually represented through word clouds in Figure 2. Additionally, the biology-inspired hierarchical categorization is illustrated in Figure 3. Our dataset, **CamoVid40K**, encompasses a diverse array of animals, spanning 69 categories across flying, terrestrial, and aquatic groups. This comprehensive coverage highlights the varied strategies and adaptations these animals employ to achieve camouflage in their respective environments.

Video Lengths. The distribution of the number of samples in different categories is shown in Figure 4, ranging from 100 to 4,500 frames. This variation ensures a broad representation of different camouflaged animals and allows for comprehensive evaluation across diverse scenarios.

Dataset splits. We split our dataset into training and testing sets, which consist of **120** videos with **35,991** frames and **29** videos with **8,279** frames, respectively.

Data Organization. We split our dataset based on displacement into two subsets: Small displacement (including all data for every single frame) and Large displacement (including selected data for every fifth frame). This division is beneficial for evaluating motion segmentation methods, as it provides a robust framework for analyzing the performance of algorithms under varying conditions of motion and displacement.

Naming Convention. We name every image as follows: ‘SuperClass-SubClass-SubNumber-MotionType-FrameNumber‘. For example, an image with name ‘Aquatic-Crab-3-Loco-015.jpg‘ indicates that this image is the *15th frame* with *locomotion* from the *3rd video* of the *Crab* species within the *Aquatic Animals* superclass. This systematic naming convention ensures clarity and ease of reference within the dataset.

Evaluation Protocol. We utilized the following metrics for our benchmark. **Bounding Box:** We measure the Intersection Over Union (IOU) between the ground truth bounding box and the minimal bounding box that encompasses the predicted segmentation mask. **Mask:** We utilize the same evaluation metrics as those referenced in [4] to assess the pixel-wise masks:

- Mean Absolute Error (MAE) evaluates the pixel-level precision between the predicted and labeled masks.
- Enhanced-alignment measure [8], which assesses both the pixel-level and image-level statistical alignment to effectively evaluate the accuracy of camouflaged object detection results, both overall and in localized areas.
- Structure-measure [7], which examines the structural similarity, emphasizing region-aware and object-aware

aspects.

- Weighted F-measure [24], offering more dependable results than the conventional F_β by accounting for the importance of different evaluation aspects.
- Mean Intersection Over Union (meanIoU) calculates the overlap between two masks.

3. Conclusion and Future Works

In this paper, we introduced a large-scale video dataset for camouflaged animal understanding, named **CamoVid40K**, aimed at fostering further research on animals. This dataset provides a comprehensive benchmark for Camouflaged Animal Understanding (CAU) 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. Future extensions of our dataset will include additional tasks such as novel view synthesis, reconstruction, and more.

New Benchmark. The **CamoVid40K** is a comprehensive and diverse benchmark meticulously curated from publicly accessible datasets and the internet. This benchmark is introduced to enhance the assessment and exploration of algorithmic generalization capabilities for the specified task. It includes a wide range of camouflaged animals across various environments, providing a robust framework for testing and developing new models.

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.

Impact on Animal Study. By providing detailed and varied data on camouflaged animals, the **CamoVid40K** 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.

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

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 3

[3] Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, 2022. 2

[4] 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. 1, 2, 4

[5] Qiaole Dong and Yanwei Fu. MemFlow: Optical flow estimation and prediction with memory. In *CVPR*, 2024. 3

[6] 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. 2

[7] 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. 4

[8] 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. 4

[9] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE T-PAMI*, 2022. 1

[10] 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. 1

[11] Peter J. S. Fleming, Paul D. Meek, Guy Ballard, Peter B. Banks, Andrew W. Claridge, James G. Sanderson, and Don E. Swann. *Camera Trapping: Wildlife Management and Research*. CSIRO Publishing, 2014. 1

[12] 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. 1

[13] Ruozhen He, Qihua Dong, Jiaying Lin, and Rynson WH Lau. Weakly-supervised camouflaged object detection with scribble annotations. In *AAAI*, 2023. 1

[14] 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. 1

[15] 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. 2

[16] 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. 1

[17] Hala Lamdouar, Charig Yang, Weidi Xie, and Andrew Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. *ACCV*, 2020. 1, 2, 3

[18] Hala Lamdouar, Weidi Xie, and Andrew Zisserman. Segmenting invisible moving objects. In *BMVC*, 2021. 1

[19] Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *CVIU*, 2019. 1

[20] 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. 1

[21] Guillaume Le Moing, Jean Ponce, and Cordelia Schmid. Dense optical tracking: Connecting the dots. In *CVPR*, 2024. 3

[22] 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. 1

[23] 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. 1

[24] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *CVPR*, 2014. 5

[25] Moritz Menze and Andreas Geiger. Object scene flow for autonomous vehicles. In *CVPR*, 2015. 2

[26] Etienne Meunier, Anaïs Badoual, and Patrick Bouthemy. Em-driven unsupervised learning for efficient motion segmentation. *IEEE T-PAMI*, 2022. 1

[27] Thi Thu Thuy Nguyen, Anne C. Eichholtzer, Don A. Driscoll, Nathan I. Semianiw, Dean M. Corva, Abbas Z. Kouzani, Thanh Thi Nguyen, and Duc Thanh Nguyen. Sawit: A small-sized animal wild image dataset with annotations. *Multimedia Tools and Applications*, 2023. 1

[28] 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. 3

[29] 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. 1

[30] Erik Learned-Miller Pia Bideau. It's moving! a probabilistic model for causal motion segmentation in moving camera videos. In *ECCV*, 2016. 1, 2

[31] 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. 1

[32] OpenAI team. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[33] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 3

[34] 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. 1, 2

[35] 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. 3

[36] Christopher Xie, Yu Xiang, Zaid Harchaoui, and Dieter Fox. Object discovery in videos as foreground motion clustering. In *CVPR*, 2019. 1

[37] Junyu Xie, Weidi Xie, and Andrew Zisserman. Segmenting moving objects via an object-centric layered representation. *NeurIPS*, 2022. 1

[38] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *ICCV*, 2021. 1

[39] 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. 2

[40] 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. 3