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

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

[1]Hong Kong University of Science and Technology [2]CFAR & IHPC, A*STAR
[3]University of California, Los Angeles [4]Ocean University of China

Abstract

Neural networks have shown significant success in computer vision tasks, but less attention has been given to monitoring camouflaged animals, the masters of hiding themselves in the background. Despite efforts in camouflaged animal image segmentation, limited work exists on camouflaged animal video understanding (CAVU). Biologists prefer videos for monitoring animal behaviors due to their redundant information and temporal consistency. However, the limited labeled video data significantly hinders progress in this area. To address these challenges, we present ***CamoVid60K****, a diverse, large-scale, and accurately annotated video dataset of camouflaged animals. This dataset comprises* ***218*** *videos with* ***62,774*** *annotated frames, covering* ***70*** *animal categories, which surpasses all previous datasets in terms of the number of videos/frames and species included.* ***CamoVid60K*** *also offers more diverse downstream tasks in computer vision, such as camouflaged animal classification, detection, and task-specific segmentation (semantic, referring, motion),* etc. *We have benchmarked several recent algorithms on the proposed* ***CamoVid60K*** *dataset, and the experimental results provide valuable insights for future research directions.*

1. Introduction

Recent advances in neural networks, including Convolutional Neural Networks (CNNs)[14] and Vision Transformers (ViTs)[10], have significantly improved visual understanding tasks. Large-scale datasets, such as COCO [21], ADE20K [48], and Object365 [33], have facilitated this progress through extensive annotated data. However, these datasets predominantly feature everyday objects, overlooking specialized fields such as camouflaged animal analysis. Studying camouflaged animals is critical for biodiversity conservation [31, 36] and provides insights into evolutionary biology and adaptation mechanisms.

Camouflaged animal segmentation, the precise extraction of object masks, allows accurate identification and analysis of animals, aiding ecological studies and population monitoring [22, 38]. Unlike common objects, camouflaged animals pose unique challenges due to their indistinct boundaries, diverse poses, and complex patterns. Although several datasets for camouflaged animal segmentation exist, such as CAMO-COCO [20], COD10K [12], CAM-LDR [23], and S-COD [15], they primarily focus on image-level tasks, insufficient for comprehensive biological monitoring that requires capturing behavior and activity over time [42]. Current video-level datasets, like MoCA [17], only provide bounding box annotations, which inadequately represent detailed animal appearances and behaviors.

To address these limitations, we introduce **CamoVid60K**, a large-scale video dataset designed explicitly for camouflaged animal video understanding (CAVU). Our dataset comprises **218** videos with **62,774** densely annotated frames covering **70** animal categories, surpassing existing datasets such as CAD [28], MoCA [17], MoCA-Mask [7], MVK [39], WATB [40], and AnimalTrack [45]. Unlike prior works annotating every fifth frame, **CamoVid60K** provides precise annotations for every frame, including animal categories, bounding boxes, masks, pseudo-label optical flows, and referring expressions. This extensive annotation supports diverse downstream tasks like classification, detection, semantic and motion segmentation, and optical flow estimation.

We also propose benchmarks using state-of-the-art methods, establishing strong baselines for robust camouflaged animal video analysis. In summary, our main contributions include the following:

- Introducing **CamoVid60K**, a comprehensive video dataset significantly surpassing previous datasets in scale and annotation richness.
- Presenting a **simple yet effective pipeline** achieving competitive results in camouflaged animal detection and segmentation.

Table 1. Comparison with existing video animal datasets. Class.: Classification Label, B.Box: Bounding Box, Motion: Motion of Animal, Pseudo OF: Pseudo-label Optical Flow, Expres.: Referring Expression. ***Note that***, MVK [39] 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 5 frames**, resulting in only 4,691 annotated frames of 22,939 frames. In contrast, our CamoVid60K dataset offers a significantly larger volume of data with more frequent annotations and a wider variety of annotation types.

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

- Benchmarking **multiple video understanding tasks**, establishing foundations for future advancements in computer vision and wildlife research.

2. CamoVid60K Dataset

Collecting video datasets of camouflaged animals is quite challenging, even without focusing on long-form videos. Manually collecting, observing, and annotating videos with multiple annotation types is labor-intensive, time-consuming, and expensive. In addition to these costs, ensuring visual data diversity and high-quality annotations adds to the difficulty. In this section, we propose a staged data collection and processing pipeline, as shown in Figure 1, and then the taxonomic structure of our dataset (with animal names), as shown in Figure 2. Associated datasheets [13] and data cards [29] for our **CamoVid60K** dataset are provided in supplementary materials.

Data Sources and Pre-Processing. We built our dataset by incorporating previous datasets (Table 1) and crawling additional videos from the internet to cover a variety of camouflaged animals. We initially collected 1,929 videos, then manually checked and filtered out any blurry or irrelevant videos, retaining those with clear depictions of animals.

Bounding Box and Mask Annotation. We utilized the annotation tool from [46], which is heavily based on the Segment Anything Model (SAM) [16] for mask initialization and bounding box creation, and XMem [6] for mask and bounding box propagation. We then manually checked and refined every frame to provide accurate bounding boxes and segmentation masks. In addition, we adopted the perceptual camouflage score ($\mathcal{S}_p$) from [19] to quantify the effectiveness of animals' camouflage, *i.e.*, how successfully an animal blends into its background. Based on this score, we retained videos with a score higher than a threshold (*e.g.*, $\mathcal{S}_p > 0.3$).

Previous optical flow datasets, such as Flying Chairs [9], KITTI [25], Sintel [2], and FlyingThings3D [24], utilized either simulation software or real images with additional heavy sensor information (depth, LiDAR, *etc*.) and algorithms to create optical flow ground truth. This process is time-consuming and requires significant effort. Recently, with the development of deep learning techniques, many methods [37, 41] can produce accurately estimated optical flow. Therefore, we utilized these methods for our pseudo-label optical flow annotation, using the algorithm shown in Algorithm 1. We used the pre-trained model of RAFT on FlyingThings3D [24] and the pre-trained DINO model of ViT-B architecture.

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

Motion Annotation. Following Lamdouar et al. [17], we manually labeled our dataset according to the types of motion, as shown below.

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

Referring Expression Annotation. We utilized MarineGPT [47], the first vision-language model specially designed for the marine domain, to generate captions for aquatic animals. After the initial annotation, we verified and refined all captions and selected the best three for each video sequence. Objects that could not be localized using

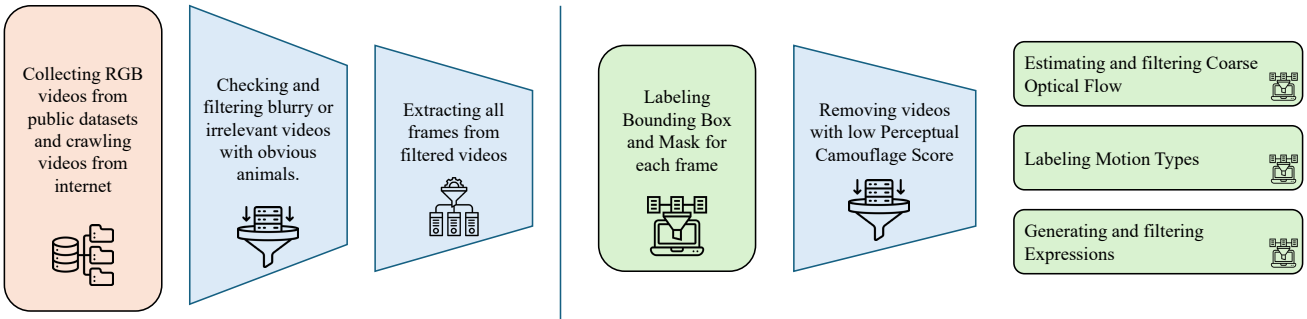

Figure 1. **CamoVid60K** data pipeline. Stage I includes data curation, filtering irrelevant videos, and extracting all frames. Stage II includes data annotation, generation, and filtering.

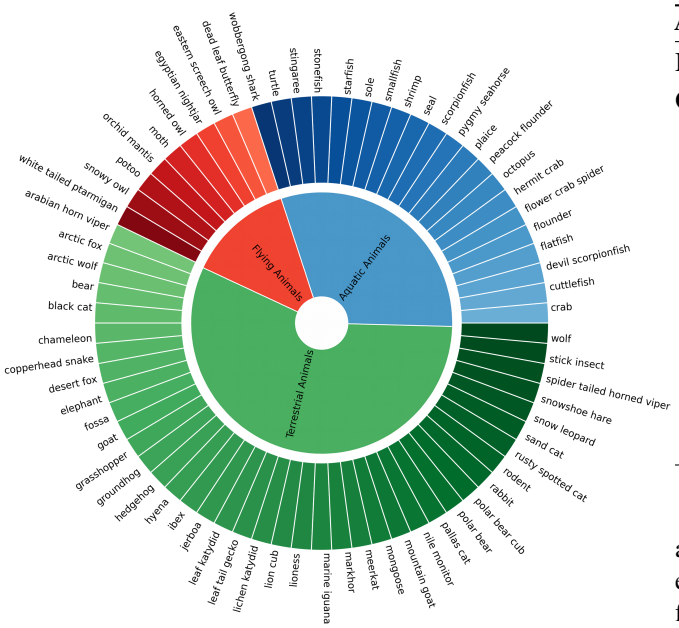

Figure 2. Taxonomic structure of our dataset by their biology-inspired hierarchical categorization. It encompasses various animals, spanning 70 categories across flying, terrestrial, and aquatic groups.

language or referring expressions were removed.

3. A Pipeline to Discern Camouflaged Animals

After constructing our dataset, we propose a simple pipeline based on Mask2Former architecture [5, 19] for both object detection and motion segmentation tasks. As shown in Figure 3, our pipeline processes sequences of images or videos by employing any off-the-shelf flow estimation method. In our case, we directly use the refined optical flow provided in our dataset instead of utilizing the RAFT method [37] to estimate raw optical flow, as done in [19]. The images

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: Estimating all pairwise optical flows using the RAFT method [37]
3: Computing DINO features [4, 27] 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**

and associated estimated flows are passed into two separate encoders for feature extraction. Subsequently, the image and flow features at each timestamp are aggregated before being fed into the decoder to predict the segmentation mask.

Visual Encoder. We adopt the SINet-v2 [12] architecture (utilized ResNet50 as backbone with weights pretrained on ImageNet dataset), which takes an RGB sequence as input $I^v = \{I_1^v, I_2^v, \ldots, I_n^v\} \in \mathbb{R}^{n \times d_v \times h \times w}$, where n is the number of frames, d_v is the dimension of each frame, and h and w are the height and width, respectively. The visual encoder outputs visual features $\{f_1^v, f_2^v, \ldots, f_n^v\} = \Phi_{\text{visual}}(I^v)$.

Motion Encoder. Inspired by the motion segmentation architecture [18], we use a lightweight ConvNet that takes as input a sequence of optical flows $I^f = \{I_1^f, I_2^f, \ldots, I_n^f\} \in \mathbb{R}^{n \times d_f \times h \times w}$, where d_f is the dimension of the flow field, and outputs motion features $\{f_1^m, f_2^m, \ldots, f_n^m\} = \Phi_{\text{motion}}(I^f)$. We then concatenate the motion features with learned spatial and temporal positional encodings to produce a set of enriched motion features.

Decoder. We adopt the Mask2Former [5] architecture, which

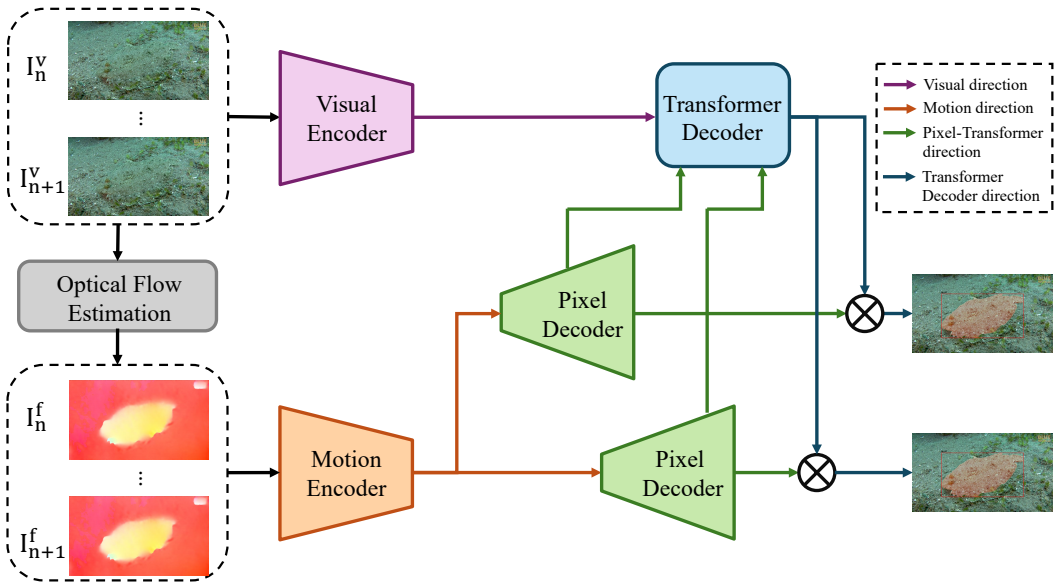

Figure 3. Our simple pipeline takes a sequence of images or a video (and the associated **pre-computed** optical flow if available, we also provided in our dataset) as input. They are fed into separated encoders for feature extraction. Then, the motion features with spatial and temporal positional encoding 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.

includes Transformer and pixel decoders. The Transformer decoder combines a trainable query for mask embedding with the outputs of the motion encoder and visual features. Similar to Mask2Former, this query attends to multi-scale motion features and visual features, resulting in mask embedding for the moving object. Additionally, like 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 utilize the L1 loss for bounding box regression, cross-entropy loss for the confidence score, and binary cross-entropy (BCE) loss for motion segmentation. The total loss for training our pipeline is defined as follows:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{L1}} + \mathcal{L}_{\text{CE}}, \tag{1}$$

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

4. Experiments

Baselines. For motion segmentation, we selected recent state-of-the-art methods for comparison, including frame-based (PraNet [11]) and video-based methods (SLT-Net [7]). For the object detection, we compared our approach with well-known detection methods: Faster R-CNN [32], DETR [3], and DINO [44]. We followed the $1\times$ (12-epoch) training setting and used the same ResNet50 [14] backbone for all methods. For the zero-shot image classification task, we tested three recent methods: CLIP [30], UniCL [43], and K-LITE [34]. We used the Swin-T model for both UniCL and K-LITE (pre-trained on the ImageNet-21K dataset [8]) and the ViT-B/32 pre-trained model from OpenAI's CLIP.

All methods (utilized implementations provided by the authors) were trained and tested on the same NVIDIA RTX 3090 GPU, except for the pre-trained models used in the zero-shot image classification task, where we utilized the pre-trained models provided by the authors.

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

Method	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$M \downarrow$	mDic ↑	mIoU ↑
PraNet	0.526	0.161	0.547	0.045	0.198	0.144
SLT-Net	**0.576**	**0.253**	**0.591**	**0.039**	<u>0.268</u>	<u>0.249</u>
Ours	<u>0.566</u>	<u>0.249</u>	<u>0.589</u>	<u>0.041</u>	**0.270**	**0.252**

Comparison with Image-Based and Video-Based Motion Segmentation Methods. Table 2 compares the performance of our method with other approaches. Compared to image-based methods, our approach demonstrates significantly superior performance due to the incorporation of temporal information. When evaluated against video-based methods, our approach surpasses MG [42], relying solely on estimated

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

Method	F-RCNN	DETR	DINO	Ours
$AP \uparrow$	28.72	37.56	**39.84**	38.39

Table 4. Zero-shot image classification performance.

Method	CLIP	UniCL	K-LITE
mAcc	30.06	30.89	**31.44**

optical flows as input. However, our method performs worse on certain metrics compared to the recent state-of-the-art method SLT-Net [7]. 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, our proposed model demonstrates performance comparable to other specialized methods, owing to its dual capabilities in object detection and motion segmentation. Specifically, our method significantly outperforms conventional CNN-based methods. This advantage stems from dual optimizations in the detection and segmentation streams, along with the integration of additional optical flow information. However, when compared to DETR-like methods [3, 44], our approach shows mixed results. It surpasses the standard DETR model [3], yet falls short of DINO [44], an advanced variant of DETR. DINO enhances performance through several innovative techniques: it employs contrastive denoising training to refine one-to-one matching, a mixed query selection method to better initialize the queries, and a 'look forward twice' method that utilizes gradients from subsequent layers to adjust parameters more accurately.

Additional Analysis and Discussions. State-of-the-art methods, including foundation models trained on large datasets such as CLIP [30], UniCL [43], and K-LITE [34], struggle with zero-shot image classification of camouflaged animals, as shown in Table 4. This is due to the subtle and complex patterns of camouflaged animals, the lack of specific training data, and the difficulty in generalizing across different backgrounds and lighting conditions. Improving these methods involves curating specialized training data (or fine-tuning on our dataset), using enhanced techniques like data augmentation, few-shot learning, and developing context-aware models.

5. Conclusion

We introduced **CamoVid60K**, a large-scale video dataset for camouflaged animal understanding, aiming to foster further research on camouflaged animals and providing a benchmark for camouflaged animal video understanding tasks, enabling the evaluation of various algorithms and methods.

Limitations and Future Work. 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, we will propose a learnable module to estimate the implicit optical flow field.

Impact on Animal Studies. Researchers can utilize our dataset to explore how different species employ camouflage in their natural habitats, leading to deeper insights into predator-prey interactions and survival strategies, aiding conservation efforts by improving detection and monitoring of endangered species in their natural environments [1, 26, 35, 38].

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

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

References

[1] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *ECCV*, pages 456–473, 2018. 5

[2] 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

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 4, 5

[4] 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

[5] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, pages 1290–1299, 2022. 3

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

[7] 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, 4, 5

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 4

[9] 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

[10] 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. 1

[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 *MICCAI*, pages 263–273. Springer, 2020. 4

[12] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE T-PAMI*, 2022. 1, 3

[13] 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. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 4

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

[16] 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

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

[18] Hala Lamdouar, Weidi Xie, and Andrew Zisserman. Segmenting invisible moving objects. In *BMVC*, 2021. 3

[19] Hala Lamdouar, Weidi Xie, and Andrew Zisserman. The making and breaking of camouflage. In *ICCV*, pages 832–842, 2023. 2, 3

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

[21] 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 *ECCV*, pages 740–755. Springer, 2014. 1

[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] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

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

[26] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *PNAS*, 115(25):E5716–E5725, 2018. 5

[27] 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. *TMLR*, 2024. 3

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

[29] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *ACM FAccT*, pages 1776–1826, 2022. 2

[30] 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 *ICML*, pages 8748–8763, 2021. 4, 5

[31] 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. 1

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015. 4

[33] 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 *ICCV*, pages 8430–8439, 2019. 1

[34] 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. *NeurIPS*, 35:15558–15573, 2022. 4, 5

[35] Fanny Simões, Charles Bouveyron, and Frédéric Precioso. Deepwild: Wildlife identification, localisation and estimation on camera trap videos using deep learning. *Ecological Informatics*, 75:102095, 2023. 5

[36] 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. 1

[37] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 2, 3

[38] Jolyon Troscianko, John Skelhorn, and Martin Stevens. Quantifying camouflage: how to predict detectability from appearance. *BMC Evolutionary Biology*, 17:1–13, 2017. 1, 5

[39] 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 *MMM*. Springer, 2023. 1, 2

[40] Fasheng Wang, Ping Cao, Fu Li, Xing Wang, Bing He, and Fuming Sun. Watb: Wild animal tracking benchmark. *International Journal of Computer Vision*, 131(4):899–917, 2022. 1

[41] 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. 2

[42] Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie. Self-supervised video object segmentation by motion grouping. In *ICCV*, 2021. 1, 4

[43] Jianwei Yang, Chunyuan Li, Pengchuan Zhang, Bin Xiao, Ce Liu, Lu Yuan, and Jianfeng Gao. Unified contrastive learning in image-text-label space. In *CVPR*, pages 19163–19173, 2022. 4, 5

[44] 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 *ICLR*, 2023. 4, 5

[45] Libo Zhang, Junyuan Gao, Zhen Xiao, and Heng Fan. Animaltrack: A benchmark for multi-animal tracking in the wild. *International Journal of Computer Vision*, 131(2):496–513, 2022. 1

[46] 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

[47] 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. 2

[48] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017. 1